

UNIVERSIDADE FEDERAL DE MINAS GERAIS

PROGRAMA DE PÓS-GRADUAÇÃO EM ENGENHARIA ELÉTRICA

TESE DE DOUTORADO

**MIXED MESHFREE METHODS IN COMPUTATIONAL ELECTROMAGNETISM:
MATHEMATICAL FOUNDATIONS AND PROBLEMS IN WAVE SCATTERING**

Williams Lara de Nicomedes

22 de Maio de 2015



GAPTEM – Grupo de Antenas, Propagação e Teoria Eletromagnética

**Departamento de Engenharia Eletrônica
Escola de Engenharia
Universidade Federal de Minas Gerais**

Universidade Federal de Minas Gerais
Programa de Pós-Graduação em Engenharia Elétrica
Escola de Engenharia

Mixed Meshfree Methods in Computational Electromagnetism:
Mathematical Foundations and Problems in Wave Scattering

Williams Lara de Nicomedes

Belo Horizonte, May 2015

TESE DE DOUTORADO Nº 201

**MIXED MESHFREE METHODS IN COMPUTATIONAL ELECTROMAGNETISM:
MATHEMATICAL FOUNDATIONS AND PROBLEMS IN WAVE SCATTERING**

Williams Lara de Nicomedes

DATA DA DEFESA: 22/05/2015

**"Mixed Meshfree Methods in Computational Electromagnetism:
Mathematical Foundations and Problems in Wave Scattering"**

Williams Lara de Nicomedes

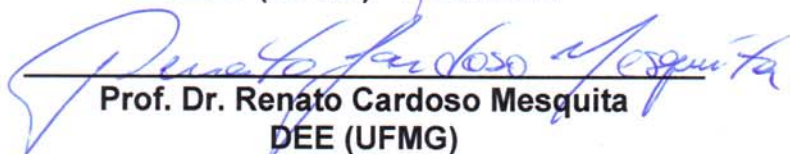
Tese de Doutorado submetida à Banca Examinadora designada pelo Colegiado do Programa de Pós-Graduação em Engenharia Elétrica da Escola de Engenharia da Universidade Federal de Minas Gerais, como requisito para obtenção do grau de Doutor em Engenharia Elétrica.

Aprovada em 22 de maio de 2015.

Por:



Prof. Dr. Fernando José da Silva Moreira
DELT (UFMG) - Orientador



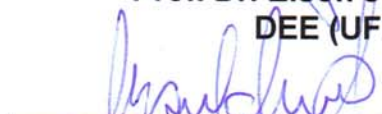
Prof. Dr. Renato Cardoso Mesquita
DEE (UFMG)



Prof. Dr. Cássio Gonçalves do Rego
DELT (UFMG)



Prof. Dr. Elson José da Silva
DEE (UFMG)



Profa. Dra. Ursula do Carmo Resende
Eng. Elétrica (CEFET-MG)



Prof. Dr. Delfim Soares Junior
Eng. de Estruturas (UFJF)

'In that case, my dear Adeimantus,' I said, 'we must certainly not give up, even if the investigation turns up to be rather lengthy.' (376d)*

'I certainly don't know yet; we must let our destination be decided by the winds of the discussion.' (394d)*

Plato, *Republic*

*The two quotes are taken from the Oxford World's Classics edition. Translated by Robin Waterfield. Oxford University Press, 2008.

To my father, my mother and my brother (Head).

Abstract

This thesis is primarily concerned with the extension of nodal meshfree methods to the solution of electromagnetic wave scattering problems in three dimensions. These problems involve vector field quantities, which are usually constrained by a divergence-free condition. The rather innocent addition of such a constraint on the divergence makes the analysis via nodal basis functions particularly challenging. In order to deal with it, we must add a Lagrange multiplier to the discretized weak forms. We are thus led to a mixed formulation which involves two quantities: The electric field and the Lagrange multiplier (also called *pseudopressure*). Next we investigate the conditions under which the aforementioned mixed formulation is well-posed; at this point the so-called *inf-sup* conditions play a fundamental role. After delving deeply on the theorems which comprise the framework of mixed formulations, one observes that the nodal approach we propose is backed by a firm mathematical theory. Finally, our meshfree formulation is put to the test by solving several problems pertaining to the subject of wave scattering.

Resumo

A presente tese versa sobre a extensão dos métodos sem malha ditos ‘nodais’ a problemas de espalhamento eletromagnético em três dimensões. Tais problemas envolvem quantidades vetoriais, sobre as quais geralmente é imposta uma condição de divergente nulo. A simples adição de uma restrição como essa ao divergente torna particularmente difícil a análise via funções de forma nodais. Para lidar com ela de uma maneira adequada, precisamos adicionar um multiplicador de Lagrange à versão discretizada das formas fracas resultantes do problema. Desta forma, somos levados a uma formulação mista que envolve duas quantidades: O campo elétrico e o multiplicador de Lagrange (também chamado de *pseudopressão*). Em seguida, investigamos as condições sob as quais a formulação mista é bem-posta; aqui as chamadas condições *inf-sup* desempenham um papel fundamental. Após uma profunda exploração dos teoremas que dão estrutura às formulações mistas, observa-se que a abordagem nodal proposta é de fato sustentada por uma firme base matemática. Finalmente, a formulação *meshfree* desenvolvida é testada na solução de vários problemas relativos ao espalhamento eletromagnético.

Preface

This work presents a nodal meshfree procedure for solving problems in which the field quantities involved are *vectors*, i.e., quantities which are characterized by a magnitude and a direction in space, as opposed to *scalars*, which are devoid of any sense of direction attached to their meaning. I had the opportunity to deal with meshfree methods and scalar quantities in different circumstances in the past during my Master's work. Thanks to the relative success I obtained, it was decided that the natural path to follow would be the extension of the meshfree approach to scenarios involving vector field quantities, particularly those arising in the analysis of time-harmonic electromagnetic wave propagation and scattering.

The first ideas concerning the application of meshfree techniques to the Maxwell-Helmholtz equation are sketched in the text for the Qualifying Exam I presented to the UFMG Graduate Program in Electrical Engineering in September 2012. (By Maxwell-Helmholtz equation I mean the vector wave equation involving a double curl on $\mathbf{E}$ which one gets from both Faraday's and Ampère's laws written in the frequency domain). It was duly approved by the examining committee, who encouraged me to bring the work to a successful completion. One of the characteristics of this preliminary work is that the discretization process should rely solely on *nodal basis functions* (as opposed to the vector edge and face elements which are standard practice in the finite element literature). The reasons for such a choice is that the underlying meshfree method is a *particle method*, i.e., it is based on *particles* or *nodes* spread throughout the computational domain of interest (denoted as Ω). In doing so, we keep the geometrical structure at a minimum: Just a set of nodes (ordinary points). Edges, faces and tetrahedra should be completely absent. This of course does not preclude the development of different meshfree methods based on objects other than nodes; it only reflects my choice, which is to comply with a minimal geometrical structure.

When certain scalar functions are ascribed to each node in the domain Ω , one gets (under the right conditions) a linear space V^e , spanned by the set of these functions (i.e., formed by all linear combinations of these functions). These scalar functions are the nodal basis functions mentioned in the previous paragraph, and will be described later in the text.

For vectors in the Euclidean space $\mathbb{R}^d$ (such as $\mathbf{E}$), the notions of magnitude and direction can be joined together in order to describe an Euclidean vector at a point $\mathbf{x} \in \Omega \subset \mathbb{R}^d$ as an ordered d -tuple of real (or complex) numbers, also called its *components*, as $\mathbf{E} = [E_1, \dots, E_d]^T$, where $d = 2$ (two dimensions) or $d = 3$ (three dimensions). In a nodal approach, each component of the discretized vector field $\mathbf{E}^h = [E_1^h, \dots, E_d^h]^T$ is taken from V^e , i.e., $E_i^h \in V^e$, $i = 1, \dots, d$, or equivalently, $\mathbf{E}^h \in V^e \times \dots \times V^e$.

The description of what is meant here by a nodal meshfree approach would by now be complete if the governing differential equation were not constrained by some condition on the divergence of the field in question. For the scattered electric field in free-space (a situation with which we will be most concerned here), such a condition reads as $\nabla \cdot \mathbf{E} = 0$. This immediately poses the question: How can we make sure that the discretized field also satisfies this constraint, i.e., how can we guarantee that $\nabla \cdot \mathbf{E}^h = 0$? Moreover, in what sense shall this condition be satisfied? In a pointwise or in a weak sense? The simplicity of the geometrical structure prevents us from embedding the divergence-free condition into the basis functions (as it happens for some vector finite elements).

Roughly speaking, the discretized problem we are trying to solve is formed by two equations: The Maxwell-Helmholtz equation (a vector equation in d variables, namely $E_1^h, \dots, E_d^h$) and the constraint on the divergence. When written componentwise, the first equation produces a system of d differential equations in d variables, whereas the second produces another equation in d variables. We are thus left with a total of $(d + 1)$ equations in d variables. The problem becomes again balanced if we introduce a scalar Lagrange multiplier, or *pseudopressure* p^h , into the system of equations. Now there are $(d + 1)$ equations in $(d + 1)$ variables.

The effect of introducing another quantity p^h is that we get a coupled system of equations, in such a way that we must seek for a solution pair $(\mathbf{E}^h, p^h)$, instead of just solving for $\mathbf{E}^h$ only. The first impression is that the problem becomes more complicated than it should, but all clouds are dissipated when one observes that it fits the structure of a *mixed formulation*, i.e., one which seeks to find approximate solutions for two (or more) quantities simultaneously.

I was presented to the concept of mixed formulations and mixed finite elements during the year of 2013, a period I spent at the Massachusetts Institute of Technology (M.I.T.) Department of Mechanical Engineering conducting the doctoral research as a Visiting Ph.D. student. The fact that our initial development in terms of $\mathbf{E}^h$ and p^h fits the structure of mixed formulations turns out to be a remarkable event, because the theory supporting these formulations has already been given a rigorous mathematical treatment.

The theory of mixed formulations was developed (independently) by I. Babuska and F. Brezzi in the 1970's, and since then, it has provided a basis to assess the well-posedness of finite element discretizations for a number of problems in many branches of computational mechanics. By fitting our meshfree procedure to the structure provided by the general theory of mixed formulations, all the theorems and results necessary for guaranteeing the solvability of our problems are automatically inherited. In doing so, two goals can be reached at once: We not only discover a way to solve (constrained) vector problems through a nodal meshfree method, but we are also provided a means to assess the well-posedness of such problems. All the theoretical development will be presented in the text, of course.

Actually, the theory of mixed formulations relies on deep results from Functional Analysis, namely the Banach Open Mapping and Closed Range theorems, where they are used to study the well-posedness of abstract operator equations. When they are ‘specialized’ to the bilinear forms arising in the mixed formulations, they assume the form of *inf-sup* conditions involving such forms. The role these *inf-sup* conditions play in the analysis of the discretized forms from the scattering problem will be discussed in detail.

Since the idea of approximating a vector field $\mathbf{E}$ by nodal scalar basis functions together with a (scalar) Lagrange multiplier p is not usual in finite element analyses of electromagnetic problems, I resorted to a model in which such approximation proved to be successful: It is the finite element analysis of the steady-state incompressible Navier-Stokes equations from fluid dynamics. There, one usually turns to nodal basis functions in order to discretize the velocity field $\mathbf{u}$, whereas the pressure p automatically plays the role of a Lagrange multiplier in order to enforce the incompressibility condition $\nabla \cdot \mathbf{u} = 0$.

There are many similarities between the mixed formulations for the Maxwell-Helmholtz equations and for the Navier-Stokes equations, or, stated in a better way, I tried to make the formulation of the Maxwell-Helmholtz system to resemble that of the Navier-Stokes system as much as possible. The result may be viewed as some kind of ‘hydrodynamical formulation’ for scattering problems. As odd as it may appear at first, it worked pretty well, as attested by the examples, and it seems that this formulation finally provided a satisfactory answer to the problem of how to address vector problems in electromagnetism through meshfree methods.

Due to the nature of the subject explored in this thesis, i.e., the analysis of the well-posedness of discretized mixed formulations – a large portion of the text is devoted to it – the inclusion of many mathematical statements is unavoidable. The very nature of the problem I proposed myself to solve asks for it. At some points I was obliged to include proofs and derivations in the text. Nevertheless, it should be clear that this is an engineering thesis, not a standard mathematics monograph. Therefore I strived to find a balance between mathematical rigor and engineering pragmatism. I hope I succeeded in this task.

Survey of the chapters

The thesis is organized in six chapters, as follows:

Chapter 1 – Introduction

A brief account of some meshfree methods developed so far. Maxwell’s equations and scattering by conducting objects. Inclusion of the pseudopressure into the system of equations. The Navier-Stokes system from fluid dynamics. Tensor algebra.

Chapter 2 – Variational formulations

The Navier-Stokes and Maxwell-Helmholtz systems in weak form. The theoretical basis that will ultimately support the well-posedness of abstract problems is introduced. Numerous ideas from Functional Analysis.

Chapter 3 – Mixed formulations

The functional analytic results from the previous chapter are specialized to the function spaces from the Navier-Stokes and Maxwell-Helmholtz systems. Well-posedness of the mixed formulations associated to these systems. More ideas from Functional Analysis.

Chapter 4 – The discretization process

Analysis of the mixed formulations in finite-dimensional subspaces. The global linear system of algebraic equations. The meshfree method we develop is presented in detail. The discretized weak forms from the scattering problem are embedded into the structure developed in Chapter 3.

Chapter 5 – Experimental studies

The well-posedness of the discretized problems is assessed through numerical *inf-sup* tests. Numerical integration of the weak forms. Solution of the boundary value problems from electromagnetic wave scattering. Preconditioning for saddle-point problems. Far-fields and calculation of the radar cross section (RCS).

Chapter 6 – Conclusions

Concluding remarks and future works.

Acknowledgements

Of course, due to the opportunity I was given to carry out this doctoral work, I am in debt to many individuals and institutions.

Prof. Fernando Moreira, the thesis advisor, allowed to me to join his antennas and propagation group (GAPTEM), where I began the research works in applied electromagnetism and wave scattering. He has arranged me scholarships and nice computers since I was an undergraduate, and I am grateful for that.

Prof. Renato Mesquita, the thesis co-advisor, introduced me to the field of meshfree simulation during the Master's course. It caused an unexpected shift in the line of research I originally intended to follow, but fortunately an agreement between

Profs. Fernando and Renato was reached, which allowed a powerful partnership to be formed.

Prof. Klaus-Jürgen Bathe, the foreign advisor, opened the doors of M.I.T. to me, where I conducted research at his finite element group. I was introduced to the theory of mixed formulations, and the privilege of discussing many topics with a man of such a stature is beyond measure. Moreover, I am very grateful for the support with the M.I.T. tuition costs he provided, and for the books he gave me.

I was also greatly benefitted by the knowledge derived from the courses taken at the UFMG Graduate Program in Electrical Engineering (PPGEE). I had a good time when attending the classes of the following professors: Prof. Cássio G. Rego (high-frequency methods), Prof. Odilon Maroja (time-harmonic fields), Prof. Elson J. Silva (finite element methods in electromagnetism), Prof. Jaime A. Ramírez (finite difference methods in electromagnetism), and Prof. Rodney R. Saldanha (optimization in electrical engineering).

The staff at PPGEE was instrumental in dealing with all the bureaucratic issues, particularly in what concerns the documentation and the regularization of the monthly stipends I received from the funding agencies.

Going to M.I.T. was a challenge: At some point in the process, everything ultimately depended on dealings with FAPEMIG (Fundação de Amparo à Pesquisa do Estado de Minas Gerais). Prof. Reinaldo M. Palhares (then Chairman at PPGEE) solved all issues with skill and a friendly disposition to help.

Upon arrival at the United States for the 2012/2013 research year, more help: Yin Jin Lee (an M.I.T. graduate student) found me a nice apartment, where I spent the first part of my stay. Mayoka Takemori from the M.I.T. International Students Office provided all the orientation regarding the U.S. immigration issues, and Sucharita Ghosh from the M.I.T. Dept. of Mechanical Engineering took care of the registration procedures at each academic term. In what regards the second part of my stay, I am grateful to Kevin O’Keefe and his wife Anne-Marie for the solution to the problem of housing. Many thanks to Albe Simenas and his wife Nanette who provided me an excellent accommodation in their comfortable house in Cambridge, MA.

One needs money in order to carry out research activities abroad, and the funding was provided by FAPEMIG, to which I am thankful. The financial resources were managed by FUNDEP, and all the assistance I received from Claudia M. Alves, Bernardo Lima and Tatiana Vigato must be mentioned.

In the end, I am grateful to my family, who put up with the long periods of time during which my attention was almost exclusively dedicated to this work.

Williams L. Nicomedes, Belo Horizonte, MG, May 2015

Resumo estendido

Introdução

Os métodos sem malha (*meshfree* ou *meshless*) têm sido aplicados a problemas provenientes do eletromagnetismo computacional com relativo sucesso. Trabalhos como [Maréchal, 1998], [Parreira *et al.*, 2006], [Manzin and Bottauscio, 2008], [Yu and Chen, 2010], [Nicomedes *et al.*, 2012], [Lima and Mesquita, 2013], entre outros, mostraram como as técnicas *meshless* podem ser consideradas como uma alternativa ao tradicional método de elementos finitos (FEM) na solução de problemas em eletromagnetismo.

Entretanto, o grande desafio posto aos métodos sem malha é a sua aplicação a problemas envolvendo grandezas vetoriais em três dimensões. Essa classe de problemas geralmente é resultante de modelos que representam situações de grande interesse prático em vários domínios da engenharia elétrica.

Provavelmente um dos primeiros trabalhos a tentar aplicar um método sem malha a problemas vetoriais em três dimensões é [Yu and Chen, 2009]. Os resultados são interessantes, mas esse trabalho desvia dos nossos interesses em pelo menos dois pontos: Primeiro, ele necessita de diagramas de Voronoi em algum ponto do processo, o que os torna ‘não totalmente sem malha’. Segundo, o método proposto é baseado em colocação, o que o torna muito parecido com o método de diferenças finitas (FDTD).

Estamos a procurar um método que seja baseado em formulações variacionais, como o tradicional FEM. Em síntese, queremos um ‘FEM sem malha’. O próximo candidato a tentar resolver problemas vetoriais em três dimensões é [Lu and Shanker, 2007]. O método proposto por eles é baseado numa formulação variacional, e os autores apresentam uma maneira de construir funções de forma vetoriais, similar aos elementos de aresta do FEM. O método foi aplicado a problemas simples, mas os resultados são bons. Entretanto, o procedimento é aplicável apenas a geometrias retangulares, e além disso há um problema com o fato de que essas funções de forma vetoriais não são linearmente independentes.

O método sem malha que temos em mente também precisa se adaptar a problemas com geometrias curvas, e deve ser testado em situações um pouco mais realísticas. Decidimos então concentrar nossa atenção no espalhamento de ondas eletromagnéticas por objetos condutores perfeitos (PEC). Além de ser uma área de interesse prático, somos automaticamente levados a problemas vetoriais em três dimensões. Se pudermos conceber um método sem malha baseado em formulação variacional e que funcione corretamente nesse cenário, então nosso objetivo terá sido alcançado.

Felizmente, conseguimos desenvolver tal método. O método proposto e a ser estudado nesse trabalho é *inteiramente nodal*, i.e., não depende de funções de forma vetoriais. Para desenvolvê-lo, tivemos que nos afastar um pouco do eletromagnetismo e explorar a hidrodinâmica (mecânica dos fluidos). Tomamos como inspiração métodos destinados à solução da famosa equação de Navier-Stokes e, após várias modificações, construímos uma adaptação apta a ser utilizada em nossos problemas de espalhamento eletromagnético.

A característica fundamental do nosso método é que ele depende de duas variáveis simultaneamente: o campo elétrico e a *pseudopressão*, que é apenas um artifício que deve ser empregado de modo a forçar a condição do divergente nulo. Chegamos assim a um exemplo de *formulação mista* (ou híbrida), que, como é sabido, depende de algumas sutilezas no que diz respeito à solvabilidade dos problemas.

Uma delas é a condição *inf-sup* (ou Babuska-Brezzi), que especifica condições que os espaços de aproximação para o campo elétrico e a pseudopressão devem satisfazer de modo que o problema seja bem-posto.

O método apresentado neste trabalho funciona bem quando aplicado aos problemas de espalhamento os quais originalmente tínhamos em mente, o que representa um avanço. Entretanto, o maior empecilho é que ele é baseado numa matemática não muito simples. O problema de Navier-Stokes tem uma teoria matemática sólida e bem desenvolvida, que foi parcialmente aproveitada na análise do problema de espalhamento. Dizemos parcialmente, e não totalmente, porque *esses dois problemas são similares, mas não idênticos*. Alguns pontos tiveram de ser modificados de modo a acomodar as diferenças. O mais evidente deles é a incorporação da Alternativa de Fredholm, uma vez que a forma sesquilinear proveniente da equação de Helmholtz não é coerciva.

Os desenvolvimentos teóricos formam a base dos Capítulos 2 e 3, e todo o ferramental matemático é introduzido na medida em que se faz necessário. Dedicamos um certo esforço em identificar a ordem correta na qual os argumentos devem ser apresentados, de maneira a tornar o desenvolvimento mais lógico e coerente.

No restante desse resumo, vamos citar os principais pontos de cada capítulo, lembrando que a sua compreensão depende da leitura do texto da tese, onde tivemos um grande cuidado em explicar detalhadamente tudo o que está ocorrendo.

O problema a ser resolvido

Após uma cadeia de raciocínio que se origina com as equações de Maxwell, pode-se mostrar que o problema de espalhamento eletromagnético pode ser modelado, de uma maneira preliminar, pelo sistema de equações:

Encontre $(\mathbf{E}^s, p)$ tal que

$$\nabla^2 \mathbf{E}^s + k_0^2 \mathbf{E}^s + \nabla p = \mathbf{0}, \quad \text{in } \Omega \quad (0.1.a)$$

$$\nabla \cdot \mathbf{E}^s = 0, \quad \text{in } \Omega \quad (0.1.b)$$

$$\hat{\mathbf{n}}_i \times \mathbf{E}^s = -\hat{\mathbf{n}}_i \times \mathbf{E}^{inc}, \quad \text{at } \Gamma_i, \quad i = 1, 2, \dots \quad (0.1.c)$$

$$\hat{\mathbf{n}}_o \times \mathbf{E}^s = \mathbf{0}, \quad \text{at } \Gamma_o, \quad (0.1.d)$$

onde $\mathbf{E}^s$ é o campo elétrico espalhado, p é a pseudopressão, $\mathbf{E}^{inc}$ é o campo elétrico incidente (conhecido) e Ω é a região na qual o problema deve ser resolvido. A fronteira $\partial\Omega = \Gamma$ é composta de duas partes: A superfície dos ‘espalhadores’, ou seja, dos objetos metálicos Γ_i , e a fronteira exterior Γ_o .

É interessante comparar (0.1.a) – (0.1.d) com o sistema de Navier-Stokes para meios homogêneos:

Encontre $(\mathbf{u}, p)$ tal que

$$-\nu \nabla^2 \mathbf{u} + (\mathbf{u} \cdot \nabla) \mathbf{u} + \nabla p = \mathbf{f}, \quad \text{in } \Omega \quad (0.2.a)$$

$$\nabla \cdot \mathbf{u} = 0, \quad \text{in } \Omega \quad (0.2.b)$$

$$\mathbf{u} = \mathbf{g}, \quad \text{at } \Gamma, \quad (0.2.c)$$

em que $\mathbf{u}$ é a velocidade do fluido, ν é a viscosidade cinemática (uma constante), p é a pressão, $\mathbf{f}$ representa a ação de forças em atuação no fluido, e $\mathbf{g}$ especifica o vetor de velocidades na fronteira da região Ω .

Os dois sistemas acima são muito semelhantes, principalmente porque ambos possuem a restrição de divergente nulo. Cabe a pergunta: Poderiam procedimentos empregados na solução de (0.2.a) – (0.2.c) ser adaptados e em seguida empregados na solução de (0.1.a) – (0.1.d)?

No Capítulo 1, decidimos incluir a dedução do sistema (0.2.a) – (0.2.c) a partir de primeiros princípios como uma maneira de iniciar a discussão sobre a álgebra de tensores, que será necessária nas explorações matemáticas do sistema de espalhamento (0.1.a) – (0.1.d), particularmente no que diz respeito à incorporação da PML (*perfectly matched layer*) e também ao espaço de funções que construímos para a aproximação *meshfree* do campo espalhado $\mathbf{E}^s$.

Camada absorvente: PML

O sistema (0.1.a) – (0.1.d) precisa ser modificado de modo a simular ondas que se propagam somente no sentido de se afastar do objeto espalhador. Quando somente um único objeto é considerado, o sistema se torna

Encontre $(\mathbf{E}^s, p)$ tal que

$$\nabla \cdot \bar{\bar{\Lambda}} \cdot \nabla \mathbf{E}^s + k_0^2 \mathbf{E}^s + \nabla p = \mathbf{0}, \quad \text{in } \Omega \quad (0.3.a)$$

$$\nabla \cdot \mathbf{E}^s = 0, \quad \text{in } \Omega \quad (0.3.b)$$

$$\hat{\mathbf{n}}_1 \times \mathbf{E}^s = -\hat{\mathbf{n}}_1 \times \mathbf{E}^{inc}, \quad \text{at } \Gamma_1 \quad (0.3.c)$$

$$\hat{\mathbf{n}}_o \times \mathbf{E}^s = \mathbf{0}, \quad \text{at } \Gamma_o, \quad (0.3.d)$$

onde o tensor PML é descrito por

$$\bar{\bar{\Lambda}} = \Lambda_x \hat{\mathbf{x}} \otimes \hat{\mathbf{x}} + \Lambda_y \hat{\mathbf{y}} \otimes \hat{\mathbf{y}} + \Lambda_z \hat{\mathbf{z}} \otimes \hat{\mathbf{z}}. \quad (0.3.e)$$

A versão do tensor $\bar{\bar{\Lambda}}$ empregada aqui foi originalmente desenvolvida para problemas de propagação de ondas acústicas em mecânica [Bermúdez *et al.*, 2007], e não uma das versões tradicionalmente aplicadas em problemas de eletromagnetismo, como a PML anisotrópica [Sacks *et al.*, 1995]. Uma das razões é que a PML ‘acústica’ é mais adequada para formulações baseadas no Laplaciano, enquanto a PML anisotrópica é muito bem empregada em formulações baseadas no rotacional duplo (ou *curl-curl*).

Entretanto, antes de aplicar a PML acústica a problemas de espalhamento eletromagnético, precisamos realizar alguns ajustes, descritos na Seção 3.3.6.6.

Formulação variacional: Formas fracas

O campo elétrico é primeiramente decomposto como

$$\mathbf{E}^s = \mathbf{e}^0 + \mathbf{u}^g, \quad (0.4.a)$$

em que $\mathbf{u}^g$ é a função de *lifting* relativa às condições de contorno (0.3.c) e (0.3.d). A Seção 2.2.3.5 traz uma discussão considerável acerca da função de *lifting*. A função $\mathbf{e}^0$ é tal que suas componentes tangenciais são nulas em toda a fronteira do domínio Ω , i.e., $\hat{\mathbf{n}} \times \mathbf{e}^0 = \mathbf{0}$ em Γ_o e Γ_1 . Uma vez que a função de *lifting* $\mathbf{u}^g$ é conhecida, $\mathbf{e}^0$ se torna a verdadeira incógnita do problema, juntamente com a pseudopressão p . Observações nos levam a concluir que o espaço de funções no qual $\mathbf{e}^0$ deve ser procurado é $\mathbb{V}_\tau(\Omega)$, definido como

$$\mathbb{V}_\tau(\Omega) = \{\mathbf{v} \in H^1(\Omega)^3 \mid \hat{\mathbf{n}} \times \mathbf{v}|_\Gamma = \mathbf{0}\} \quad (0.4.b)$$

O espaço de funções para p é simplesmente $L^2(\Omega)$. Desta forma, a formulação variacional para o sistema (0.3.a) – (0.3.d) é

Encontre $(\mathbf{e}^0, p) \in \mathbb{V}_\tau(\Omega) \times L^2(\Omega)$ tal que

$$\int_\Omega (\bar{\bar{\Lambda}} \cdot \nabla \mathbf{e}^0) : \nabla \mathbf{v}^* - \int_\Omega k_0^2 \mathbf{e}^0 \cdot \mathbf{v}^* - \int_\Omega p \nabla \cdot \mathbf{v}^* =$$

$$-\int_{\Omega} (\bar{\mathbf{\Lambda}} \cdot \nabla \mathbf{u}^g) : \nabla \mathbf{v}^* + \int_{\Omega} k_0^2 \mathbf{u}^g \cdot \mathbf{v}^*, \quad \forall \mathbf{v} \in \mathbb{V}_{\tau}(\Omega) \quad (0.4.c)$$

$$-\int_{\Omega} q^* \nabla \cdot \mathbf{e}^0 = \int_{\Omega} q^* \nabla \cdot \mathbf{u}^g, \quad \forall q \in L^2(\Omega), \quad (0.4.d)$$

O sistema (0.4.c) – (0.4.d) é uma instância do que se chama de *formulação mista ou híbrida*. Em termos abstratos (i.e., em termos de espaços de funções cuja natureza não é especificada, mas que assume formas diferentes de acordo com cada problema), ela é escrita como

$$\begin{aligned} & \text{Encontre } (u, p) \in \mathcal{X} \times \mathcal{Y} \text{ tal que} \\ & a(u, x) + b(x, p) = \langle f^*, x \rangle_{\mathcal{X}^*, \mathcal{X}} \quad \forall x \in \mathcal{X} \\ & b(u, y) = \langle g^*, y \rangle_{\mathcal{Y}^*, \mathcal{Y}} \quad \forall y \in \mathcal{Y}, \end{aligned} \quad (0.4.e)$$

onde $\mathcal{X}$ e $\mathcal{Y}$ são dois espaços de Hilbert, e f^* e g^* são elementos dos duais $\mathcal{X}^*$ e $\mathcal{Y}^*$.

O sistema (0.4.e) serve como modelo para diversos problemas em mecânica, entre eles problemas em hidrodinâmica governados pela equação de Navier-Stokes [Girault and Raviart, 1986]. A teoria que especifica as condições sob as quais a solução de (0.4.e) existe, é única e limitada (i.e., finita), foi desenvolvida independentemente por I. Babuska e F. Brezzi [Ern and Guermond, 2004]. Entre essas condições, a chamada condição *inf-sup* ou condição de Babuska-Brezzi, [Brezzi and Fortin, 1991] desempenha um papel fundamental. Ela é expressa como:

$$\exists \beta_b > 0 \quad \text{tal que} \quad \inf_{y \in \mathcal{Y} \setminus \{0\}} \sup_{x \in \mathcal{X} \setminus \{0\}} \frac{|b(x, y)|}{\|x\|_{\mathcal{X}} \|y\|_{\mathcal{Y}}} \geq \beta_b. \quad (0.4.f)$$

Uma das idéias deste trabalho é procurar uma formulação para o problema de espalhamento que possa ser ‘embutida’ no *framework* (0.4.e). Mas esse é o caso do sistema (0.4.c) – (0.4.d), como pode ser observado. Ao se fazer essa ‘especialização’, a condição *inf-sup* a ser satisfeita se torna:

$$\exists \beta_b > 0 \quad \text{tal que} \quad \inf_{q \in L^2(\Omega) \setminus \{0\}} \sup_{\mathbf{v} \in \mathbb{V}_{\tau}(\Omega) \setminus \{0\}} \frac{\left| -\int_{\Omega} q \nabla \cdot \mathbf{v} \right|}{\|\mathbf{v}\|_{H^1(\Omega)^3} \|q\|_{L^2(\Omega)}} \geq \beta_b. \quad (0.4.g)$$

A condição (0.4.g) é estudada com profundidade na Seção 3.3.6.5.

Formulação variacional: Espaços de dimensão finita

Ao se considerar a aproximação numérica das grandezas $\mathbf{e}^0$ e p , introduzimos subespaços de $\mathbb{V}_{\tau}(\Omega)$ e $L^2(\Omega)$ de dimensão finita, i.e., gerados a partir de combinações lineares de um número finito de funções de base. Esses subespaços são representados por $\mathbb{V}_{\tau}^h(\Omega)$ e $\mathbb{P}^h(\Omega)$.

Agora não mais estamos interessados em encontrar soluções $\mathbf{e}^0 \in \mathbb{V}_\tau(\Omega)$ e $p \in L^2(\Omega)$; a nossa atenção se volta para as soluções ‘discretizadas’ $\mathbf{e}_h^0 \in \mathbb{V}_\tau^h(\Omega)$ e $p_h \in \mathbb{P}^h(\Omega)$. O problema em subespaços de dimensão finita se torna

Encontre $(\mathbf{e}_h^0, p_h) \in \mathbb{V}_\tau^h(\Omega) \times \mathbb{P}^h(\Omega)$ tal que

$$\begin{aligned} \int_{\Omega} (\bar{\mathbf{A}} \cdot \nabla \mathbf{e}_h^0) : \nabla \mathbf{v}_h^* - \int_{\Omega} k_0^2 \mathbf{e}_h^0 \cdot \mathbf{v}_h^* - \int_{\Omega} p_h \nabla \cdot \mathbf{v}_h^* = \\ - \int_{\Omega} (\bar{\mathbf{A}} \cdot \nabla \mathbf{u}_h^g) : \nabla \mathbf{v}_h^* + \int_{\Omega} k_0^2 \mathbf{u}_h^g \cdot \mathbf{v}_h^*, \quad \forall \mathbf{v}_h \in \mathbb{V}_\tau^h(\Omega) \end{aligned} \quad (0.5.a)$$

$$- \int_{\Omega} q_h^* \nabla \cdot \mathbf{e}_h^0 = \int_{\Omega} q_h^* \nabla \cdot \mathbf{u}_h^g, \quad \forall q_h \in \mathbb{P}^h(\Omega). \quad (0.5.b)$$

O problema (0.5.a) – (0.5.b) também se encaixa no *framework* (0.4.e). Desta forma, esse sistema de equações só será bem-posto se a seguinte condição *inf-sup* for satisfeita:

$$\exists \beta_b^h > 0 \quad \text{tal que} \quad \inf_{q_h \in \mathbb{P}^h(\Omega) \setminus \{0\}} \sup_{\mathbf{v}_h \in \mathbb{V}_\tau^h(\Omega) \setminus \{0\}} \frac{\left| - \int_{\Omega} q_h \nabla \cdot \mathbf{v}_h \right|}{\|\mathbf{v}_h\|_{H^1(\Omega)^3} \|q_h\|_{L^2(\Omega)}} \geq \beta_b^h. \quad (0.5.c)$$

A dificuldade é que mesmo que a condição (0.4.g) valha para os espaços de dimensão infinita $\mathbb{V}_\tau(\Omega)$ e $L^2(\Omega)$, isso não implica que a condição (0.5.c) valha para *quaisquer* subespaços $\mathbb{V}_\tau^h(\Omega) \subset \mathbb{V}_\tau(\Omega)$ e $\mathbb{P}^h(\Omega) \subset L^2(\Omega)$.

Esta talvez seja a principal questão a que o presente trabalho busca responder: Como construir subespaços de dimensão finita $\mathbb{V}_\tau^h(\Omega)$ e $\mathbb{P}^h(\Omega)$, a partir de uma abordagem *meshfree* puramente nodal, de modo que (0.5.c) seja satisfeita?

Subespaços *meshfree*

Os espaços $\mathbb{V}_\tau^h(\Omega)$ e $\mathbb{P}^h(\Omega)$ são construídos a partir dos nós espalhados pelo domínio computacional Ω .

Precisamos de dois subespaços: Um dedicado à aproximação das três componentes escalares do vetor campo elétrico $\mathbf{e}_h^0$, e outro à aproximação da pseudopressão p_h . Esses dois espaços devem ter características diferentes; ao se variar essas características, obtemos diferentes pares de espaços $\mathbb{V}_\tau^h(\Omega)$ e $\mathbb{P}^h(\Omega)$. Alguns satisfazem (0.5.c), e outros não. Aqueles que porventura não satisfaçam (0.5.c) devem ser sumariamente excluídos.

Uma maneira bastante flexível de se obter subespaços *meshfree* consiste em associar um *patch* (uma região cúbica) a cada nó, e em seguida definir um conjunto de funções de base nesse *patch*. Combinações lineares dessas funções dão origem a um espaço

localmente definido no *patch*. Isso deve ser feito para todos os nós espalhados pelo domínio, e em seguida, o conjunto de espaços locais é ‘conectado’ por uma família de funções que tem o atributo da *partição da unidade* (PU), como por exemplo as funções de Shepard.

Espaços com características distintas são construídos na medida em que diferentes funções de base são consideradas em cada espaço local. Todo o raciocínio que leva à construção dos espaços *meshfree* é amplamente discutido na Seção 4.3.

Desenvolvemos um procedimento bastante interessante para a construção de espaços *meshfree* destinados à aproximação do campo elétrico $\mathbf{e}_h^0$ em geometrias curvas. A técnica é baseada no conceito que resolvemos chamar de ‘direções elementais’, que na verdade é uma base ortonormal local para $\mathbb{R}^3$ associada a cada um dos nós, cujos vetores variam de acordo com as direções normais e tangencias associadas ao nó em questão. As derivadas são obtidas com o auxílio de produtos tensoriais.

Estudos experimentais

Uma vez que tenhamos consolidado o domínio sobre o processo de construção de espaços *meshfree* que satisfaçam (0.5.c), podemos aplicá-los seguramente à solução do problema de espalhamento (0.5.a) – (0.5.b).

No Capítulo 5, resolvemos vários problemas de espalhamento em duas e três dimensões. Além disso, apresentamos uma discussão acerca do condicionador que deve ser aplicado juntamente com um método iterativo durante a solução do sistema linear global.

A experimentação termina com um pós-processamento: Estudamos as seções de radar (RCS – *Radar Cross Section*) relativas a certos objetos PEC e estabelecemos uma comparação com resultados provenientes da óptica física.

Conclusões

De maneira geral, pode-se dizer que a tese consiste em duas partes: A obtenção de um método *meshfree* para a solução de problemas de espalhamento eletromagnético e a análise matemática do mesmo.

Acreditamos que o trabalho serviu para consolidar a linha de pesquisa à qual temos nos dedicado há algum tempo: As técnicas sem malha realmente podem ser empregadas na solução de problemas de interesse prático em engenharia elétrica (pelo menos no que diz respeito a problemas de espalhamento e alta frequência).

Obviamente, não estamos a dizer que o trabalho está concluído; pelo contrário, a presente tese abre muitos tópicos para pesquisa futura. Esperamos considerá-los em breve.

Contents

Preface	iii
Resumo estendido	viii
Contents	xv
Chapter 1 – Introduction	1
1.1 Historical information	1
1.2 A brief account on Maxwell’s equations	4
1.3 Wave scattering by PEC objects	7
1.3.1 Scattering boundary value problems	7
1.3.2 The vector Laplacian is more suitable than the double curl	11
1.4 The pseudopressure	12
1.4.1 Scattering and radiation problems are similar	12
1.4.2 The Lagrange multiplier	13
1.4.3 The equations from fluid mechanics	14
1.4.4 Incompressibility	21
Chapter 2 – Variational Formulations	24
2.1 The Navier-Stokes system in weak form	24
2.1.1 Weak derivatives	24
2.1.2 Function spaces: $L^2(\Omega)$ and $H^1(\Omega)$	29
2.1.3 Function spaces: $L^2(\Omega)^d$ and $H^1(\Omega)^d$	31
2.1.4 Function spaces: Density and trace theory	33
2.1.5 Navier-Stokes: Weak forms and weak solutions	37
2.1.5.1 The problem in classical form	39
2.1.5.2 Testing functions	39
2.1.5.3 Relaxing the requirements	39
2.1.5.4 Lifting on the boundary data	43
2.1.5.5 The G map	44
2.1.5.6 Enlarging the space of testing functions	47
2.1.5.7 Weak solutions	49
2.2 The scattering system in weak form	50
2.2.1 Scattering equations	50

2.2.2 PML I: Incorporating the PML	50
2.2.3 The scattering system: Weak forms and weak solutions	51
2.2.3.1 The problem in classical form	52
2.2.3.2 Testing functions	53
2.2.3.3 Relaxing the requirements	55
2.2.3.4 Interlude 1: The space $H(\mathbf{curl}; \Omega)$	58
2.2.3.5 Lifting on the boundary data	62
2.2.3.6 The G map	64
2.2.3.7 Enlarging the space of testing functions	67
2.2.3.8 Weak solutions	70
Chapter 3 – Mixed Formulations	71
3.1 Mixed formulations in abstract form	71
3.1.1 Mixed variational formulations	71
3.1.2 Well-posedness	72
3.2 Mixed formulation for the Navier-Stokes system	74
3.2.1 Continuity and coercivity must be checked	74
3.2.2 The inf-sup condition must be checked	77
3.3 Mixed formulation for the scattering system	84
3.3.1 Determining the structure of the problem	84
3.3.2 Well-posedness	86
3.3.3 The Fredholm Alternative	87
3.3.4 Embeddings	88
3.3.5 Well-posedness of non-coercive problems	90
3.3.6 Back to the scattering system	93
3.3.6.1 Functionals I	93
3.3.6.2 Functionals II	99
3.3.6.3 Theorem 3.9, Hypotheses (i) and (ii)	100
3.3.6.4 Theorem 3.9, Hypotheses (iii), (iv), (viii) and (ix)	101
3.3.6.5 Theorem 3.9, Hypothesis (vi)	104
3.3.6.6 PML II: The PML tensor	106
3.3.6.7 Theorem 3.9, Hypothesis (vii)	111
3.3.6.8 Theorem 3.9, Hypothesis (v)	113
3.3.7 Concluding remarks	117
Chapter 4 – The discretization process	118

4.1 The problem in finite-dimensional subspaces	118
4.1.1 The key theorem: Specialization to the scattering system	118
4.1.1.1 Hypothesis (i)	122
4.1.1.2 Hypothesis (ii)	123
4.1.1.3 Hypotheses (iii) and (viii)	123
4.1.1.4 Hypothesis (ix)	123
4.1.1.5 Hypothesis (iv)	124
4.1.1.6 Hypothesis (v)	124
4.1.1.7 Hypothesis (vi)	125
4.1.1.8 Hypothesis (vii)	126
4.1.1.9 Concluding remarks	127
4.2 The linear system	127
4.2.1 The matrix system: Preliminary form	127
4.2.2 The matrix system: Uniqueness of the solution	130
4.2.3 The matrix system: The inf-sup condition	133
4.3 Meshfree subspaces	135
4.3.1 Nodes and patches	135
4.3.2 Geometrical considerations	141
4.3.3 The spaces $\mathbb{V}_\tau^h(\Omega)$ and $\mathbb{P}^h(\Omega)$	145
4.3.4 Numbering schemes and the assembly process	150
4.3.5 Final comments	159
Chapter 5 – Experimental Studies	160
5.1 Numerical integration	160
5.1.1 Basic integrals	160
5.1.2 Acceleration technique	161
5.1.3 Numerical quadrature	167
5.2 The inf-sup stability test	169
5.3 Preconditioning	176
5.4 Case studies	177
5.4.1 Free-space: Error	177
5.4.2 Scattering of a TE^z plane wave by a circular cylinder	178
5.4.3 Scattering of a TE^z plane wave by a conducting strip	182
5.4.4 The spherical cavity	186
5.4.5 Scattering by PEC plates	195

5.4.6 Radar cross sections	199
5.4.6.1 Three dimensions	201
5.4.6.2 Two dimensions	207
5.4.6.3 Physical Optics	211
Chapter 6 – Conclusions	215
6.1 Concluding remarks	215
6.2 Future work	216
6.2.1 The tangential trace operator	216
6.2.2 Complex eigenvalues	217
6.2.3 Preconditioning	217
Appendix 1 – Theorem 3.8	219
Appendix 2 – Theorem 3.9	232
Appendix 3 – List of Symbols	242
Bibliography	246

Chapter 1

Introduction

In this chapter, we first present some historical information on the development of meshfree methods. The account is by no means extensive, and we concentrate on those works from computational mechanics and computational electromagnetism which in some way influenced the development of this thesis.

After the historical survey we present a brief discussion about the Maxwell's equations, followed by the general description of electromagnetic wave scattering problems.

We proceed by introducing the Lagrange multiplier and the role it shall play in connection with the enforcement of the divergence-free constraint. Also, we present a concise discussion about the Navier-Stokes equations, with a focus on the mathematical form of the problem (i.e., with no regard to the physics these equations describe).

Finally, we assemble both the wave scattering and the Navier-Stokes problems into systems of partial differential equations and point out the similarities and differences between them.

1.1 Historical information

By ‘meshfree’ or ‘meshless’ one actually refers to a family of methods aimed at the numerical solution of differential equations. They were (and have been) developed for a variety of purposes, and may become very different from each other. Nevertheless, they must all share a basic characteristic: In order to be termed ‘meshfree’, a method should not employ any kind of mesh or grid, as opposed to the finite element (FEM) and finite difference methods.

The motivation behind the development of meshfree methods is basically an answer to the difficulties in handling a mesh, particularly in what concerns the automatic mesh generation in three dimensions and also in remeshing procedures, i.e., in problems whose geometry changes with time (and also in adaptive refinement) [Li and Liu, 2007].

Meshfree methods began to be consistently considered as a choice in the early 1990's. Since then, the methods continue to evolve and significant improvements have been made [Liu, 2010].

Among the first meshfree methods to be introduced is the Smoothed Particle Hydrodynamics (SPH) [Gingold and Monaghan, 1977], [Liu and Liu, 2003]. It is a particle method based on collocation [Liu and Liu, 2010]; in order to do so, it relies on certain smooth approximations to the Dirac delta function (the Dirac functional). It has been applied successfully to problems in a number of areas, such as mechanics [Zhang and Batra, 2009] and swarm robotics [Pimenta *et al.*, 2013].

Another collocation method is that based on Radial Point Interpolation (RPIM) basis functions [Yu and Chen, 2009]. In general, collocation methods deal with a particular differential equation in strong form; they are simpler to implement, but may suffer from instabilities. Moreover, sometimes they resort to Voronoi decompositions, which makes them not fully meshfree [Yu and Chen, 2010].

A different category of meshfree methods is that based on weak forms, i.e., these methods are employed in conjunction with some variational expression associated with the differential equation in question. The Element Free Galerkin (EFG) reached a prominent position among these [Belytschko *et al.*, 1994], [Maréchal, 1998], [Cingoski *et al.*, 1998], [Parreira *et al.*, 2006], [Bottauscio *et al.*, 2006], [Manzin and Bottauscio, 2008]. Despite the fact EFG has found a relative acceptance among some authors, it is not regarded as a full meshfree method, since background cells are required for the numerical integration of the weak forms.

A method which also deserves attention is the Meshless Local Petrov-Galerkin (MLPG) method. It remedies the issue of background cells from EFG by introducing certain local domains, in which the numerical integrations are performed. MLPG has a number of variants, and has also found a relatively wide acceptance among the authors [Atluri and Shen, 2002], [Li *et al.*, 2003], [Dehgan and Mirzaei, 2008], [Soares Jr., 2009], [Vavourakis, 2009], [Soares *et al.*, 2014].

The MLPG method constitutes the basis of our previous works [Nicomedes *et al.*, 2011], [Nicomedes *et al.*, 2012], [Nicomedes *et al.*², 2012].

The MLPG worked pretty well in all these examples, but it also suffers from some drawbacks. When it is used together with the Moving Least Squares (MLS) basis functions, it performs poorly when imposing essential boundary conditions. Moreover, the MLS basis functions require relatively large influence domains. The reason is that the basis function associated with a given node requires the participation of neighboring nodes in order to be calculated. These neighboring nodes must also be disposed ‘nicely’, in order to avoid singular local matrices [Liu, 2010].

In order to accommodate better the structure required for dealing with vector problems, we decided to change the underlying meshfree method. We now turn our attention to the Method of Finite Spheres (MFS) [De and Bathe, 2000]. The basis functions from MFS have smaller influence domains (they can be made as small as possible, insofar as the union of all influence domains forms a covering for the computational domain Ω). Through a little change in the way the boundary conditions

are treated in [De and Bathe, 2000], essential conditions can be imposed easily, thanks to a trick to make the basis functions satisfy the Kronecker delta property. The MFS shows a good performance when applied to problems in mechanics [De and Bathe, 2001], [De and Bathe², 2001], [De *et al.*, 2003], [Ham *et al.*, 2014].

The MFS shares some characteristics with the generalized finite element methods (GFEM) based on a partition of unity [Melenk and Babuska, 1996], [Babuska and Melenk, 1997], [Strouboulis *et al.*, 2001]. The GFEM covers the computational domain with overlapping patches, and allows for the inclusion of different sets of basis functions defined on each patch. The advantage is that, in order to attain better approximation properties, information about the unknown solution may be included via proper selection of basis functions on a given patch (for example, when solving a wave problem, one could include sines and cosines in the set of basis functions). These ideas have been shown to work in scalar problems from electromagnetism [Proekt and Tsukerman, 2002].

The works in electromagnetism which deal with meshfree methods based on weak forms listed thus far are all concerned with scalar problems. As far as our knowledge goes, Lu and Shanker's work [Lu and Shanker, 2007] is the only one to try to address vector problems in electromagnetism (in variational form) through a meshfree procedure. They employ the aforementioned generalized finite element method, and define certain vector basis functions on the patches. Despite the fact their method is shown to work only for relatively simple problems, the results obtained are very promising.

However, there are drawbacks in Lu and Shanker's work. First, the method they propose has not been tested on problems with curvilinear geometries. Second, the vector basis functions defined on the patches are not interpolative, and these patches do not conform to the global boundary. As a consequence, the imposition of essential boundary conditions becomes nontrivial, and the authors apply Nitsche's method in order to impose the essential boundary conditions. Nitsche's method works by adding an extra term to the weak forms [Embar *et al.*, 2010]. This extra term depends on some stability parameters, and the overall performance of the method depends on the correct choice for these parameters. Of course, this is very unattractive. Third, the vector basis functions defined on a given patch are not orthogonal to each other, and may even be linearly dependent, which leads to serious issues with the condition number of the global matrix. In order to overcome this, the basis functions must be redefined through some kind of orthogonalization procedure. Apparently a singular value decomposition (SVD) must be performed for each patch in the problem in order to get the new (orthogonal) vector basis functions. This unfortunately increases the total computational cost of the method.

In this thesis, we present a work which provides an answer to the problem of how to solve three dimensional vector electromagnetic problems through a meshfree procedure. The method we propose can be naturally applied to curvilinear geometries,

and the imposition of essential boundary conditions is very easy, similar to the way they are imposed in the standard FEM. Since we rely on nodal basis functions only, the problem of linearly dependent vector basis functions is naturally absent.

Our meshfree method is also based on a formalism similar to that of GFEM, but it is employed for a different purpose. Whereas in the GFEM one includes certain terms as basis functions in order to get better approximation properties, we on the other hand add different basis functions for the components of the electric field $\mathbf{E}^h$ and the Lagrange multiplier p^h in order to get global approximation spaces with distinct characteristics. Since the theory underlying the mixed formulations determines that these spaces should be compatible in some sense, we arrive at a question: What terms are to be included as basis functions in the local spaces for $\mathbf{E}^h$ and p^h in order for the global spaces to be compatible?

Questions such as this one will occupy us for a while. But they will all be addressed in due time, as we progress in our work and as the concepts necessary for their proper understanding are gradually introduced. By now, let us begin our journey from the very principle: The Maxwell's equations.

1.2 A brief account on Maxwell's equations

The dynamics of the electromagnetic fields is governed by the Maxwell's equations (in SI units):

$$\nabla \times \mathcal{E}(\mathbf{x}, t) = -\frac{\partial}{\partial t} \mathcal{B}(\mathbf{x}, t) \quad (1.1)$$

$$\nabla \times \mathcal{H}(\mathbf{x}, t) = \mathcal{J}(\mathbf{x}, t) + \frac{\partial}{\partial t} \mathcal{D}(\mathbf{x}, t) \quad (1.2)$$

$$\nabla \cdot \mathcal{D}(\mathbf{x}, t) = \rho(\mathbf{x}, t) \quad (1.3)$$

$$\nabla \cdot \mathcal{B}(\mathbf{x}, t) = 0 \quad (1.4)$$

where $\mathcal{E}$ is the electric field intensity (volts/meter), $\mathcal{H}$ is the magnetic field intensity (amperes/meter), $\mathcal{D}$ is the electric flux density (coulombs/square meter), $\mathcal{B}$ is the magnetic flux density (webers/square meter), $\mathcal{J}$ is the total electric current density (amperes/square meter) and ρ is the electric charge density (coulombs/cubic meter). All the quantities depend on the position $\mathbf{x} \in \mathbb{R}^3$ and on the time $t \in \mathbb{R}$.

In the course of this thesis, we shall be interested in fields in homogeneous regions, particularly in the free-space. Under these conditions, equations (1.1) – (1.4) may be written as

$$\nabla \times \mathcal{E}(\mathbf{x}, t) = -\mu_0 \mu_r \frac{\partial}{\partial t} \mathcal{H}(\mathbf{x}, t) \quad (1.5)$$

$$\nabla \times \mathcal{H}(\mathbf{x}, t) = (\mathcal{J}_s(\mathbf{x}, t) + \sigma \mathcal{E}(\mathbf{x}, t)) + \varepsilon_0 \varepsilon_r \frac{\partial}{\partial t} \mathcal{E}(\mathbf{x}, t) \quad (1.6)$$

$$\nabla \cdot \mathcal{E}(\mathbf{x}, t) = \frac{p(\mathbf{x}, t)}{\varepsilon_0 \varepsilon_r} \quad (1.7)$$

$$\nabla \cdot \mathcal{H}(\mathbf{x}, t) = 0, \quad (1.8)$$

thanks to the constitutive relations which hold in homogeneous media

$$\mathcal{E}(\mathbf{x}, t) = \varepsilon_0 \varepsilon_r \mathcal{E}(\mathbf{x}, t) \quad (1.9)$$

$$\mathcal{B}(\mathbf{x}, t) = \mu_0 \mu_r \mathcal{H}(\mathbf{x}, t), \quad (1.10)$$

and to the separation of the total current density $\mathcal{J}(\mathbf{x}, t)$ into a source current density $\mathcal{J}_s(\mathbf{x}, t)$ (given) and an induced current density $\sigma \mathcal{E}(\mathbf{x}, t)$. The multiplicative constants appearing in (1.5) – (1.10) are the *relative electric permittivity* ε_r (dimensionless), the *relative magnetic permeability* μ_r (dimensionless) and the *electric conductivity* σ (siemens/meter). The *free-space electric permittivity* is $\varepsilon_0 = 8.854 \times 10^{-12}$ farads/meter and the *free-space magnetic permeability* is $\mu_0 = 4\pi \times 10^{-7}$ henrys/meter. In the examples we are going to study, there will be perfect electric conductors (PEC), which are characterized by an infinite value for the conductivity σ . Since no field can exist inside such a material, these PEC materials essentially define the limits of the computational domain (in the sense that the boundaries are usually PEC surfaces). Therefore the term corresponding to the induced current $\sigma \mathcal{E}(\mathbf{x}, t)$ will be neglected from now on, i.e., $\sigma = 0$ at all points from the domain Ω .

The meshfree method we intend to develop is dependent on a single field, the electric field $\mathcal{E}$. In order to eliminate $\mathcal{H}$ from the system (1.5) – (1.8), we apply the $\nabla \times$ operator to (1.5) and substitute (1.6) in the resulting expression, in order to get a system in $\mathcal{E}$ only:

$$\nabla \times \nabla \times \mathcal{E}(\mathbf{x}, t) + \mu_0 \varepsilon_0 \mu_r \varepsilon_r \frac{\partial^2}{\partial t^2} \mathcal{E}(\mathbf{x}, t) = -\mu_0 \mu_r \frac{\partial}{\partial t} \mathcal{J}_s(\mathbf{x}, t) \quad (1.11)$$

$$\nabla \cdot \mathcal{E}(\mathbf{x}, t) = \frac{p(\mathbf{x}, t)}{\varepsilon_0 \varepsilon_r} \quad (1.12)$$

Equations (1.11) and (1.12) form a system of partial differential equations on the unknown $\mathcal{E}$. The system must be complemented by specific conditions $\mathcal{E}$ must satisfy at the boundary of the domain. Assuming the system (1.11) – (1.12) is to be solved in a domain $\Omega \subset \mathbb{R}^3$, let $\partial\Omega = \Gamma$ denote its surface boundary. If such a surface is an interface PEC/free-space (or other homogeneous material), then the boundary conditions for $\mathcal{E}$ are

$$\forall \mathbf{x} \in \Gamma \quad \hat{\mathbf{n}}(\mathbf{x}) \times \mathcal{E}(\mathbf{x}, t) = \mathbf{0}, \quad (1.13)$$

which means that, for any point $\mathbf{x}$ on the boundary Γ , the outward-pointing unit normal vector $\hat{\mathbf{n}}$ at $\mathbf{x}$ and the electric field vector $\mathcal{E}$ at $\mathbf{x}$ are collinear, or in other words, $\mathcal{E}$ has no tangential component along the surface Γ . If $\mathcal{E}$ should ever be different from zero on the boundary, then it is limited to being parallel to the normal direction at any point on the boundary.

Throughout this work, the analysis will be restricted to electromagnetic fields whose temporal dependency is characterized by a sinusoidal behavior. They oscillate with a *frequency* f (in Hertz), which means that the fields come back to their original configuration every $T = 1/f$ seconds. Under these conditions, the functions describing the fields are separable, i.e., they can be written as a product of two terms, the first of which depends on the spatial coordinates $\mathbf{x}$ only, whereas the second depends on t only. The term governing the temporal dependency is given by $e^{j\omega t}$, where $\omega = 2\pi f$ is the *angular frequency* (radians/second) and $j = \sqrt{-1}$. The quantities $\mathcal{E}$, $\mathcal{J}_S$ and $\mathcal{p}$ in (1.11) – (1.12) therefore reads as:

$$\mathcal{E}(\mathbf{x}, t) = \text{Re}\{\mathbf{E}(\mathbf{x})e^{j\omega t}\} \quad (1.14)$$

$$\mathcal{J}_S(\mathbf{x}, t) = \text{Re}\{\mathbf{J}_S(\mathbf{x})e^{j\omega t}\} \quad (1.15)$$

$$\mathcal{p}(\mathbf{x}, t) = \text{Re}\{\rho(\mathbf{x})e^{j\omega t}\} \quad (1.16)$$

If we substitute (1.14) – (1.16) in (1.11) – (1.12) and manipulate the real part $\text{Re}\{\cdot\}$ and time-derivative $\partial\{\cdot\}/\partial t$ operators, we get a new set of equations, whose form is the same as that from (1.11) – (1.12), and in which the quantities $\mathcal{E}$, $\mathcal{J}_S$ and $\mathcal{p}$ are replaced by $\mathbf{E}$, $\mathbf{J}_S$ and ρ , respectively, whereas the time-derivative is replaced by the product $j\omega$. The new set of equations is said to be in the *frequency domain*, and is written as

$$\nabla \times \nabla \times \mathbf{E}(\mathbf{x}) - \omega^2 \mu_0 \varepsilon_0 \mu_r \varepsilon_r \mathbf{E}(\mathbf{x}) = -j\omega \mu_0 \mu_r \mathbf{J}_S(\mathbf{x}) \quad (1.17)$$

$$\nabla \cdot \mathbf{E}(\mathbf{x}) = \frac{\rho(\mathbf{x})}{\varepsilon_0 \varepsilon_r}. \quad (1.18)$$

The boundary condition in (1.13) becomes

$$\forall \mathbf{x} \in \Gamma \quad \hat{\mathbf{n}}(\mathbf{x}) \times \mathbf{E}(\mathbf{x}) = \mathbf{0}. \quad (1.19)$$

The system (1.17) – (1.19) is now complete, in the sense that the data necessary for its solution (the sources $\mathbf{J}_S$ and ρ , and the boundary conditions) are specified. However, the sources $\mathbf{J}_S$ and ρ are not independent from each other. If we apply the $\nabla \cdot$ operator to (1.6), we get (as $\nabla \cdot \nabla \times = 0$):

$$0 = \nabla \cdot \mathcal{J}_S(\mathbf{x}, t) + \varepsilon_0 \varepsilon_r \frac{\partial}{\partial t} \nabla \cdot \mathcal{E}(\mathbf{x}, t). \quad (1.20)$$

(We assumed that $\sigma = 0$). Applying (1.14) and (1.15), the equivalent expression for (1.20) in the frequency domain becomes

$$0 = \nabla \cdot \mathbf{J}_S(\mathbf{x}) + j\omega\epsilon_0\epsilon_r \nabla \cdot \mathbf{E}(\mathbf{x}). \quad (1.21)$$

Substituting (1.18) into (1.21) we get the relationship between the sources $\mathbf{J}_S$ and ρ :

$$\nabla \cdot \mathbf{J}_S(\mathbf{x}) + j\omega\rho(\mathbf{x}) = 0. \quad (1.22)$$

In this way, ρ can be eliminated from (1.18), which becomes

$$\nabla \cdot \mathbf{E}(\mathbf{x}) = \frac{j}{\omega\epsilon_0\epsilon_r} \nabla \cdot \mathbf{J}_S(\mathbf{x}) \quad (1.23)$$

The system of differential equations to be solved can be summarized as

Find $\mathbf{E}(\mathbf{x})$ such that

$$\nabla \times \nabla \times \mathbf{E}(\mathbf{x}) - k_0^2 \mu_r \epsilon_r \mathbf{E}(\mathbf{x}) = -j\omega\mu_0\mu_r \mathbf{J}_S(\mathbf{x}), \quad \mathbf{x} \in \Omega \quad (1.24.a)$$

$$\nabla \cdot \mathbf{E}(\mathbf{x}) = \frac{j}{\omega\epsilon_0\epsilon_r} \nabla \cdot \mathbf{J}_S(\mathbf{x}), \quad \mathbf{x} \in \Omega \quad (1.24.b)$$

$$\hat{\mathbf{n}}(\mathbf{x}) \times \mathbf{E}(\mathbf{x}) = \mathbf{0}, \quad \mathbf{x} \in \Gamma, \quad (1.24.c)$$

after a further modification in the first equation in which $k_0 = \omega\sqrt{\mu_0\epsilon_0}$ is the *free-space wavenumber* (radians/meter), also defined as $k_0 = 2\pi/\lambda_0$, where λ_0 is the *free-space wavelength* (meters).

1.3 Wave scattering by PEC objects

1.3.1 Scattering boundary value problems

The main category of problems we will be concerned with in this thesis is that related to the scattering of waves by perfect conductors. Even though the method we are going to develop is still applicable to problems in which the current source is different from zero (radiation problems), we decided to concentrate on problems in which $\mathbf{J}_S(\mathbf{x}) = 0$. These are the *scattering problems* [Peterson *et al.*, 1998], [Balanis, 1989], [van Bladel, 2007].

In this class of problems, the excitation is not provided by current sources, but by a preexistent field, called the *incident field* and represented as $\mathbf{E}^{inc}(\mathbf{x})$. The incident field is generally known, i.e., it is a function of the position $\mathbf{x} \in \Omega$ that must be defined prior to the solution of the problem.

In a general scattering problem, conducting objects of arbitrary geometry, called the *scatterers*, are immersed in free-space, as in Fig.1. Let each scatterer occupy a volume Ω_i (rigorously speaking, a subset from $\mathbb{R}^3$), whose boundary is denoted by Γ_i . In *exterior* problems, such as the scattering problems described here, one is generally interested in the behavior of fields at very large distances from the scatterers.

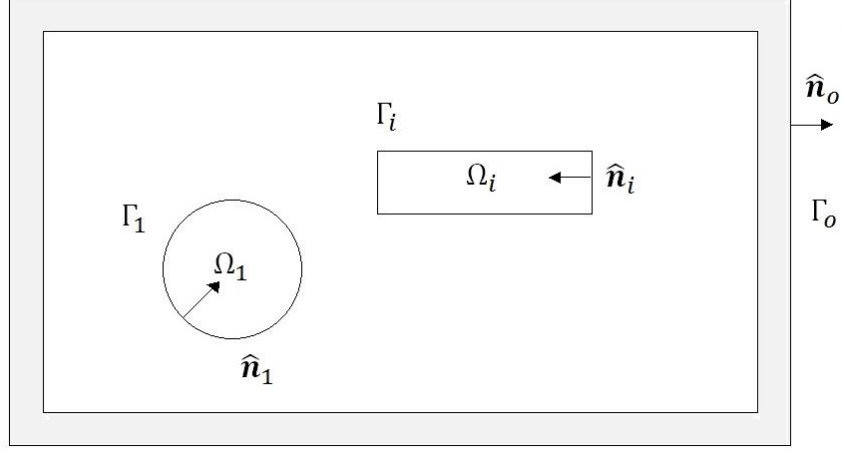


Fig.1. The outer surface Γ_o must lie relatively close to the scatterers, which are characterized by regions Ω_i within Ω_o . These ‘subregions’ are not part of the problem, and we are then left with ‘holes’. Consequently, the normal vector at the surface of the scatterers points inwards. The gray layer close to Γ_o is filled by a fictitious absorbing material (the PML).

However, it is not feasible to carry out the discretization process over these distances, so we must set a limit to the problem by placing an imaginary outer surface Γ_o encompassing all the scatterers. The surface Γ_o must be closed. By ‘encompassing all the scatterers’ we mean that $\bar{\Omega}_i \subset \Omega_o$ for any scatterer i , where the *closure* $\bar{\Omega}_i$ is given by $\bar{\Omega}_i = \Omega_i \cup \Gamma_i$, and Ω_o is nothing else than the interior of Γ_o .

In order to set up the problem, we must first define the domain Ω over which it is posed. Since the fields inside the PEC scatterers are zero, the volumes they occupy shall be excluded from Ω . So we can define the domain Ω as

$$\Omega = \Omega_o \setminus \bigcup_i \bar{\Omega}_i \quad (1.25)$$

i.e., Ω consists of the set difference between Ω_o and the union of all $\bar{\Omega}_i$. It means that if $\mathbf{x} \in \Omega$, then $\mathbf{x} \in \Omega_o$, but $\mathbf{x} \notin \bar{\Omega}_i$, for any i . The boundary of Ω then becomes

$$\Gamma = \Gamma_o \cup \Gamma_1 \cup \dots \cup \Gamma_i \cup \dots \quad (1.26)$$

i.e., Γ consists of the union of the boundaries of all scatterers Γ_i , together with the outer surface Γ_o . According to Fig.1, the domain Ω has holes left by the scatterers Ω_i , as they have been ‘carved out’ from the total volume Ω_o . In other words, Ω is not *simply connected* [Munkres, 2000], [Searcoid, 2007], [Crossley, 2005].

Next, the boundary conditions concerning the electric field $\mathbf{E}$ at Γ must be specified. Since all boundaries Γ_i are conductor surfaces, the conditions are just those from (1.19):

$$\forall \mathbf{x} \in \Gamma_i \quad \hat{\mathbf{n}}_i(\mathbf{x}) \times \mathbf{E}(\mathbf{x}) = \mathbf{0}, \quad i = 1, 2, 3 \dots \quad (1.27)$$

where $\hat{\mathbf{n}}_i(\mathbf{x})$ is the outward-pointing unit vector normal to the boundary Γ_i at $\mathbf{x}$. Now one may ask: What are the boundary conditions to be imposed at the outer surface Γ_o ? We claim that they are the same as those in (1.27), i.e., we set

$$\forall \mathbf{x} \in \Gamma_o \quad \hat{\mathbf{n}}_o(\mathbf{x}) \times \mathbf{E}(\mathbf{x}) = \mathbf{0} \quad (1.28)$$

where $\hat{\mathbf{n}}_o(\mathbf{x})$ is the normal vector at the outer surface.

Condition (1.28) may appear as a rather odd choice, since it is clearly a condition to be satisfied by the electric field at PEC boundaries, not in the free space, as it happens for the outer surface Γ_o (which is just an imaginary surface in the free space encompassing all scatterers). The reason behind the choice of (1.28) is that in order to simulate outward-propagating scattered fields, a layer of reflectionless absorbing material (of a certain thickness) will be placed along Γ_o . When the scattered fields penetrate this layer, hopefully they will be damped, so that their amplitude just before reaching the outer surface Γ_o will become negligible. This is the principle behind the Perfectly Matched Layer (PML) approach to scattered waves [Sacks *et al.*, 1995]. Since $\mathbf{E}$ is essentially zero at Γ_o , there is no harm in choosing the boundary conditions there to be (1.28), which means that the PML is backed by a PEC surface, as it is generally done in the literature [Sacks *et al.*, 1995]. (Actually, there is another deeper reason why we choose PEC conditions for Γ_o . It is related to the stability of the meshfree method we will develop, and it will become clearer in Chapter 3). Of course, when such PML layer is introduced, we are no longer dealing with homogeneous media. However, the discussion about the PML will be postponed to a more convenient time, in Chapter 3.

Since the conditions to be satisfied by the electric field at the boundaries of the scatterers and at the outer boundary are the same, as attested by (1.27) – (1.28), we can write:

$$\forall \mathbf{x} \in \Gamma \quad \hat{\mathbf{n}}(\mathbf{x}) \times \mathbf{E}(\mathbf{x}) = \mathbf{0}, \quad (1.29)$$

where Γ is now given by (1.26).

As we said at the beginning of this section, in the problems we are going to investigate, $\mathbf{J}_S = \mathbf{0}$ for all points $\mathbf{x}$ in Ω . So our system of differential equations (1.24) becomes

Find $\mathbf{E}(\mathbf{x})$ such that

$$\nabla \times \nabla \times \mathbf{E}(\mathbf{x}) - k_0^2 \mu_r \varepsilon_r \mathbf{E}(\mathbf{x}) = \mathbf{0}, \quad \mathbf{x} \in \Omega \quad (1.30.a)$$

$$\nabla \cdot \mathbf{E}(\mathbf{x}) = 0, \quad \mathbf{x} \in \Omega \quad (1.30.b)$$

$$\hat{\mathbf{n}}(\mathbf{x}) \times \mathbf{E}(\mathbf{x}) = \mathbf{0}, \quad \mathbf{x} \in \Gamma, \quad (1.30.c)$$

which in principle looks awkward because it has homogeneous data (neither sources nor boundary conditions are able to ‘excite’ the problem).

However, there is a way out if we write the total electric field $\mathbf{E}$ in $\bar{\Omega} = \Omega \cup \Gamma$ as

$$\forall \mathbf{x} \in \bar{\Omega} \quad \mathbf{E}(\mathbf{x}) = \mathbf{E}^{inc}(\mathbf{x}) + \mathbf{E}^s(\mathbf{x}) \quad (1.31)$$

where $\mathbf{E}^{inc}$ is the incident field and $\mathbf{E}^s$ is the scattered field. The incident field is known in $\bar{\Omega}$, and is, in a way, the field that would exist in Ω if all scatterers were absent, i.e., if all the volume encircled by Γ_o consisted of a homogeneous medium. The incident field is just an ordinary field produced by sources located outside $\bar{\Omega}$ and therefore satisfies the system of equations

$$\nabla \times \nabla \times \mathbf{E}^{inc}(\mathbf{x}) - k_0^2 \mu_r \varepsilon_r \mathbf{E}^{inc}(\mathbf{x}) = \mathbf{0}, \quad \mathbf{x} \in \bar{\Omega} \quad (1.32.a)$$

$$\nabla \cdot \mathbf{E}^{inc}(\mathbf{x}) = 0, \quad \mathbf{x} \in \bar{\Omega} \quad (1.32.b)$$

When the sources of the incident field are located in a region far outside Ω , it is generally the case that $\mathbf{E}^{inc}$ assumes the form of plane waves [Balanis, 1989]. After the substitution of (1.31) – (1.32) in (1.30), we arrive at the system

$$\nabla \times \nabla \times \mathbf{E}^s(\mathbf{x}) - k_0^2 \mu_r \varepsilon_r \mathbf{E}^s(\mathbf{x}) = \mathbf{0}, \quad \mathbf{x} \in \Omega \quad (1.33.a)$$

$$\nabla \cdot \mathbf{E}^s(\mathbf{x}) = 0, \quad \mathbf{x} \in \Omega \quad (1.33.b)$$

$$\hat{\mathbf{n}}(\mathbf{x}) \times \mathbf{E}^s(\mathbf{x}) = -\hat{\mathbf{n}}(\mathbf{x}) \times \mathbf{E}^{inc}(\mathbf{x}), \quad \mathbf{x} \in \Gamma \quad (1.33.c)$$

From (1.33), we discover that $\mathbf{E}^s$ is the true unknown, and that the problem is ‘excited’ by the boundary conditions. But one must be careful at this point. The boundary conditions in (1.33) imply that

$$\forall \mathbf{x} \in \Gamma_i \quad \hat{\mathbf{n}}_i(\mathbf{x}) \times \mathbf{E}^s(\mathbf{x}) = -\hat{\mathbf{n}}_i(\mathbf{x}) \times \mathbf{E}^{inc}(\mathbf{x}), \quad i = 1, 2, \dots \quad (1.34)$$

and that

$$\forall \mathbf{x} \in \Gamma_o \quad \hat{\mathbf{n}}_o(\mathbf{x}) \times \mathbf{E}^s(\mathbf{x}) = -\hat{\mathbf{n}}_o(\mathbf{x}) \times \mathbf{E}^{inc}(\mathbf{x}) \quad (1.35)$$

However, according to the PML approach, the scattered field is zero at the global boundary Γ_o , i.e., by the time it reaches Γ_o , it will be damped to negligible values. So condition (1.35) must be modified to

$$\forall \mathbf{x} \in \Gamma_o \quad \hat{\mathbf{n}}_o(\mathbf{x}) \times \mathbf{E}^s(\mathbf{x}) = \mathbf{0}. \quad (1.36)$$

(Despite the fact we ‘know’ that $\mathbf{E}^s = \mathbf{0}$ at Γ_o , we do not impose it. We must rather impose $\hat{\mathbf{n}}_o \times \mathbf{E}^s = \mathbf{0}$. In the first of these conditions, all components of $\mathbf{E}^s$ satisfy a Dirichlet boundary condition, whereas in the second, just the tangential components satisfy such a condition. These two conditions give rise to different discrete spaces, which by their turn play different roles in the stability of mixed formulations. Chapter 3 brings further discussion on this topic.)

The boundary value problem to be solved changes from (1.30) into

Find $\mathbf{E}^s(\mathbf{x})$ such that

$$\nabla \times \nabla \times \mathbf{E}^s(\mathbf{x}) - k_0^2 \mu_r \varepsilon_r \mathbf{E}^s(\mathbf{x}) = \mathbf{0}, \quad \mathbf{x} \in \Omega \quad (1.37.a)$$

$$\nabla \cdot \mathbf{E}^s(\mathbf{x}) = 0, \quad \mathbf{x} \in \Omega \quad (1.37.b)$$

$$\hat{\mathbf{n}}_i(\mathbf{x}) \times \mathbf{E}^s(\mathbf{x}) = -\hat{\mathbf{n}}_i(\mathbf{x}) \times \mathbf{E}^{inc}(\mathbf{x}), \quad \mathbf{x} \in \Gamma_i, \quad i = 1, 2, \dots \quad (1.37.c)$$

$$\hat{\mathbf{n}}_o(\mathbf{x}) \times \mathbf{E}^s(\mathbf{x}) = \mathbf{0}, \quad \mathbf{x} \in \Gamma_o, \quad (1.37.d)$$

in which the excitation is provided by the ‘matching’ of the tangential components of the fields at the surfaces of all PEC scatterers.

1.3.2 The vector Laplacian is more suitable than the double curl

The first two equations from (1.37), i.e., the equations

$$\nabla \times \nabla \times \mathbf{E}^s(\mathbf{x}) - k_0^2 \mu_r \varepsilon_r \mathbf{E}^s(\mathbf{x}) = \mathbf{0} \quad (1.38.a)$$

$$\nabla \cdot \mathbf{E}^s(\mathbf{x}) = 0, \quad (1.38.b)$$

may be called, by obvious reasons, the *double curl* approach to the vector wave equation. If we recall the vector identity

$$\nabla \times \nabla \times \mathbf{A}(\mathbf{x}) = \nabla \nabla \cdot \mathbf{A}(\mathbf{x}) - \nabla^2 \mathbf{A}(\mathbf{x}), \quad (1.39)$$

where $\mathbf{A}$ is any vector function (i.e., a function $\mathbf{A}: \mathbb{R}^3 \supset \Omega \rightarrow \mathbb{R}^3$) which meets the required differentiability criteria, then the two equations in (1.38) imply that

$$\nabla^2 \mathbf{E}^s(\mathbf{x}) + k_0^2 \mu_r \varepsilon_r \mathbf{E}^s(\mathbf{x}) = \mathbf{0}, \quad (1.40)$$

also called the *vector Helmholtz equation* (as it employs the vector Laplacian instead of the double curl). It should be emphasized that (1.38.a) and (1.38.b) imply (1.40), as we have just shown, but the converse is not true, i.e., (1.40) alone does not imply the two equations in (1.38). On the other hand, the system

$$\nabla^2 \mathbf{E}^s(\mathbf{x}) + k_0^2 \mu_r \varepsilon_r \mathbf{E}^s(\mathbf{x}) = \mathbf{0} \quad (1.41.a)$$

$$\nabla \cdot \mathbf{E}^s(\mathbf{x}) = 0 \quad (1.41.b)$$

is equivalent to (1.38.a) – (1.38.b) [Harrington, 2001].

In this thesis, we stick to (1.41) not only because it is simpler than (1.38), but because it is also less prone to instabilities. It has been shown [Lynch and Paulsen, 1991] that the double curl approach is flawed in the sense that it produces spurious solutions. The authors in [Lynch and Paulsen, 1991] apply a dispersion analysis to the double curl and to the vector Laplacian operators, and show that the cross-derivative terms in the double curl (such as $\partial^2/\partial x \partial y$) are the root cause of numerical parasites. Finally they conclude that the vector Laplacian (or Helmholtz) operator is free of

parasites when discretized with conventional scalar elements, provided that the boundary conditions are divergence-free. As the meshfree formalism is also based on scalar basis functions, and as the incident field $\mathbf{E}^{inc}$ which occurs in the boundary conditions (1.37) is also divergence-free [it is produced by sources located far away from the computational domain Ω , so the incident field is divergence-free not only in the interior of Ω , but at all PEC boundaries Γ_i as well, according to (1.32.b)], we are justified in making such a choice.

So the system of equations changes once again from (1.37) into

Find $\mathbf{E}^s(\mathbf{x})$ such that

$$\nabla^2 \mathbf{E}^s(\mathbf{x}) + k_0^2 \mu_r \varepsilon_r \mathbf{E}^s(\mathbf{x}) = \mathbf{0}, \quad \mathbf{x} \in \Omega \quad (1.42.a)$$

$$\nabla \cdot \mathbf{E}^s(\mathbf{x}) = 0, \quad \mathbf{x} \in \Omega \quad (1.42.b)$$

$$\hat{\mathbf{n}}_i(\mathbf{x}) \times \mathbf{E}^s(\mathbf{x}) = -\hat{\mathbf{n}}_i(\mathbf{x}) \times \mathbf{E}^{inc}(\mathbf{x}), \quad \mathbf{x} \in \Gamma_i, \quad i = 1, 2, \dots \quad (1.42.c)$$

$$\hat{\mathbf{n}}_o(\mathbf{x}) \times \mathbf{E}^s(\mathbf{x}) = \mathbf{0}, \quad \mathbf{x} \in \Gamma_o. \quad (1.42.d)$$

1.4 The pseudopressure

1.4.1 Scattering and radiation problems are similar

Despite the fact this thesis is primarily concerned with scattering problems, it is worth noting that scattering problems and radiation problems have a similar structure. In the former, one is interested in the scattered field $\mathbf{E}^s$, which is a disturbed field caused by the interaction of the incident field $\mathbf{E}^{inc}$ with the conducting objects that happen to be in the domain Ω . In the latter, one is interested in the total field $\mathbf{E}$, produced by a current source $\mathbf{J}_s$ in a region Ω , which may also contain conducting objects.

From now on we shall concentrate on the free-space, so we make $\mu_r = \varepsilon_r = 1$. Also, we shall drop the dependence on position $\mathbf{x}$ from the quantities involved in the equations. The scattering problem (1.42) is summarized in Chart 1.1 below.

Chart 1.1: The scattering problem

Find $\mathbf{E}^s$ such that

$$\nabla^2 \mathbf{E}^s + k_0^2 \mathbf{E}^s = \mathbf{0}, \quad \text{in } \Omega \quad (1.43.a)$$

$$\nabla \cdot \mathbf{E}^s = 0, \quad \text{in } \Omega \quad (1.43.b)$$

$$\hat{\mathbf{n}}_i \times \mathbf{E}^s = -\hat{\mathbf{n}}_i \times \mathbf{E}^{inc}, \quad \text{at } \Gamma_i, \quad i = 1, 2, \dots \quad (1.43.c)$$

$$\hat{\mathbf{n}}_o \times \mathbf{E}^s = \mathbf{0}, \quad \text{at } \Gamma_o \quad (1.43.d)$$

The steps required in going from (1.24) to (1.37) merely reflect the fact that the boundary value problem associated with scattering problems (1.37) and with radiation problems (1.24) have the same mathematical form. If we replace the double curl in (1.24) by the vector Laplacian [via (1.39)], and consider $\mu_r = \varepsilon_r = 1$, we get the system

Find $\mathbf{E}$ such that

$$\nabla^2 \mathbf{E} + k_0^2 \mathbf{E} = j\omega\mu_0 \left(\bar{\mathbf{I}} + \frac{\nabla\nabla \cdot}{k_0^2} \right) \mathbf{J}_S, \quad \text{in } \Omega \quad (1.44.a)$$

$$\nabla \cdot \mathbf{E} = \frac{j}{\omega\varepsilon_0} \nabla \cdot \mathbf{J}_S, \quad \text{in } \Omega \quad (1.44.b)$$

$$\hat{\mathbf{n}}_i \times \mathbf{E} = \mathbf{0}, \quad \text{at } \Gamma_i, \quad i = 1, 2, \dots \quad (1.44.c)$$

$$\hat{\mathbf{n}}_o \times \mathbf{E} = \mathbf{0}, \quad \text{at } \Gamma_o, \quad (1.44.d)$$

which describes a radiation problem. In (1.44.a), $\bar{\mathbf{I}}$ is the *identity tensor*, a mathematical object that maps a vector to itself [Hanson and Yakovlev, 2002].

Problems (1.43) and (1.44) are pretty similar to one another. They are both based on the vector Helmholtz equation, and are both governed by some type of Dirichlet boundary conditions. Differences lie in the fact that (1.43) is driven by a non-homogeneous Dirichlet condition at the PEC surfaces Γ_i , whereas (1.44) is driven by a source term $\mathbf{J}_S$. In what regards the meshfree analysis of these problems, the same spaces can be used in the discretization processes related to (1.43) and to (1.44). We concentrate in (1.43) because scattering phenomena often give rise to more interesting problems than radiation phenomena. Radiation problems such as (1.44) (in which the unknown is the total field $\mathbf{E}$, and not the scattered field $\mathbf{E}^s$), will be addressed only once in this work; they will be briefly considered in connection with eigenvalue problems in Chapter 5. All subsequent developments from this point on shall be related to problem (1.43).

1.4.2 The Lagrange multiplier

In order to enforce the divergence-free condition in (1.43.b), we add the gradient of a scalar potential p , or a Lagrange multiplier, to (1.43.a), motivated by some formulations concerning discontinuous Galerkin methods [Nguyen *et al.*, 2011], [Perugia *et al.*, 2002], [Houston *et al.*, 2005]. The new system is in Chart 1.2.

Chart 1.2: The modified scattering problem

Find $(\mathbf{E}^s, p)$ such that

$$\nabla^2 \mathbf{E}^s + k_0^2 \mathbf{E}^s + \nabla p = \mathbf{0}, \quad \text{in } \Omega \quad (1.45.a)$$

$$\nabla \cdot \mathbf{E}^s = 0, \quad \text{in } \Omega \quad (1.45.b)$$

$$\hat{\mathbf{n}}_i \times \mathbf{E}^s = -\hat{\mathbf{n}}_i \times \mathbf{E}^{inc}, \quad \text{at } \Gamma_i, \quad i = 1, 2, \dots \quad (1.45.c)$$

$$\hat{\mathbf{n}}_o \times \mathbf{E}^s = \mathbf{0}, \quad \text{at } \Gamma_o \quad (1.45.d)$$

The Lagrange multiplier p , also called *pseudopressure*, is included as a means to provide another unknown to the system in order to accommodate the requirement regarding the divergence-free condition.

This condition becomes problematic at the numerical level when the basis functions used in the discretization process are not solenoidal. Since our meshfree method is based on scalar basis functions, there is no way for them to be solenoidal. If one tries to solve (1.43) numerically by some method based on scalar basis functions, one discovers that the system has more equations than unknowns, i.e., three unknowns corresponding to the three components of $\mathbf{E}^s$ and four equations: Three provided by (1.43.a) and one by the divergence-free condition (1.43.b).

The inclusion of an extra unknown p in (1.45.a) makes the system balanced again: There are now four equations and four unknowns. The pseudopressure p is a kind of ‘glue’ which links the vector Helmholtz equation and the divergence-free condition together in a coupled system of differential equations.

The problem (1.45) seems to be well-structured, but a careful observation reveals that the boundary conditions to be satisfied by p are missing. In order to discover these conditions, we need to turn our attention to the weak formulation of the Navier-Stokes problem. But before doing it, a quick introduction to the equations of hydrodynamics will be provided.

1.4.3 The equations from fluid mechanics

In this subsection, we provide a concise presentation of the equations from fluid dynamics, whose solution process will ultimately lead us to a model for the solution of the electromagnetic problem (1.45). A straightforward derivation of these equations from first principles can be found in [Gross and Reusken, 2011] and [Gerbeau *et al.*, 2006]. The authors in [Boyer and Fabrie, 2012], on the other hand, are particularly rigorous in such a task.

The purpose of this subsection (and the next) is twofold. First, this is a thesis in electrical engineering, aimed at solving a problem from electromagnetism through a method which has its roots in the solution of problems from hydrodynamics. Therefore we felt that a minimal familiarity with the equations from fluid dynamics is necessary for our progress. Second, the derivation of these equations makes extensive references to tensor products, which will appear later in the weak forms for the scattering problem (1.45) and in the meshfree spaces we propose for approximating vector fields by scalar basis functions. So this is the right point for introducing them.

The complete derivation of the Navier-Stokes system involves balance equations, namely the conservation of mass and energy, linear momentum and angular momentum principles, and some thermodynamical considerations [Boyer and Fabrie, 2012]. The flow equations are usually written in *Eulerian coordinates*, which are just the coordinates of the fixed reference frame in which the experiment takes place [Boyer and Fabrie, 2012]. The approach consists in considering each point $\mathbf{x} \in \Omega$ and in writing the balance equations at $\mathbf{x}$.

The conservation of mass provides us:

$$\frac{\partial \rho}{\partial t} + \nabla \cdot (\rho \mathbf{u}) = 0, \quad (1.46)$$

where the function ρ is the *density* of the fluid at the point $\mathbf{x}$ and time t (SI units: kilograms/cubic meter), and $\mathbf{u}$ is the *velocity vector* at $(\mathbf{x}, t)$ (SI units: meters/second). In other words, at point $\mathbf{x}$ and time t , the velocity of the fluid is given by the vector $\mathbf{u}$. The density ρ is a scalar function $(\mathbf{x}, t) \mapsto \rho \in \mathbb{R}^+$, whereas the velocity $\mathbf{u}$ is a vector function $(\mathbf{x}, t) \mapsto \mathbf{u} \in \mathbb{R}^d$, $d = 2$ (two dimensions) or $d = 3$ (three dimensions). The reason why the density ρ cannot vanish is that, if it did, it would violate the continuous medium assumption [Boyer and Fabrie, 2012].

The conservation of linear momentum together with the Cauchy stress tensor theorem gives:

$$\frac{\partial(\rho \mathbf{u})}{\partial t} + \nabla \cdot (\rho \mathbf{u} \otimes \mathbf{u}) - \nabla \cdot \bar{\bar{\sigma}} = \rho \mathbf{f}, \quad (1.47)$$

where $\bar{\bar{\sigma}}$ is the *Cauchy stress tensor*. Cauchy's stress $\bar{\bar{\sigma}}$ is a tensor-valued function; it means that at any time t and at any point $\mathbf{x} \in \Omega$ there is a tensor $\bar{\bar{\sigma}}(\mathbf{x}, t)$. This tensor by its turn is a function which maps vectors to vectors: $\bar{\bar{\sigma}}(\mathbf{x}, t)$ 'receives' any unit vector $\mathbf{v}$ and 'returns' another vector, represented by $\bar{\bar{\sigma}}(\mathbf{x}, t) \cdot \mathbf{v}$. In (1.47), $\rho \mathbf{f}$ is the *body force density* at $(\mathbf{x}, t)$ (SI units: newtons/cubic meter), which means that the total body force experienced by the fluid is the volume integral of $\rho \mathbf{f}$. The term $\mathbf{f}$ alone represents the *mass density of forces* (SI units: meters/second squared). The symbol ' $\otimes$ ' is the tensor product operator [Abraham *et al.*, 1988], [Irgens, 2008]. The conservation principles of angular momentum and linear momentum together with the Cauchy stress tensor theorem imply that $\bar{\bar{\sigma}}$ is a symmetric tensor [Boyer and Fabrie, 2012].

For fluids in motion, Cauchy's stress tensor $\bar{\bar{\sigma}}$ may be written as

$$\bar{\bar{\sigma}} = \bar{\bar{\mathcal{J}}} - p \bar{\bar{\mathbf{I}}} \quad (1.48)$$

where $\bar{\bar{\mathcal{J}}}$ is a new tensor, called the *viscous stress tensor*, p is the *hydrostatic pressure* of the fluid and $\bar{\bar{\mathbf{I}}}$ is the identity tensor. (The components of the tensors in (1.48) are quantities measured in newtons/square meter.) Another tensor which plays an important role is the *strain rate tensor* $\bar{\bar{\mathbf{D}}}$, defined as:

$$\bar{\mathbf{D}}(\mathbf{u}) := \frac{1}{2}(\nabla \mathbf{u} + (\nabla \mathbf{u})^T) \quad (1.49)$$

Chart 1.3 below brings some information on the gradient of vector fields expressed in (1.49).

Chart 1.3: The gradient operator

The term $\nabla \mathbf{u}$ in (1.49) may lead to some confusion, because the gradient operator ∇ is applied to a vector $\mathbf{u}$ instead of a scalar. What is happening here is some kind of ‘operator overloading’, as the gradient operator may also be applied to a vector. When ∇ is applied to a scalar, the result is a vector. For example, for some scalar function v , we know that in Cartesian coordinates,

$$\nabla v := \left(\frac{\partial}{\partial x} \hat{\mathbf{x}} + \frac{\partial}{\partial y} \hat{\mathbf{y}} + \frac{\partial}{\partial z} \hat{\mathbf{z}} \right) v = \left(\frac{\partial v}{\partial x} \hat{\mathbf{x}} + \frac{\partial v}{\partial y} \hat{\mathbf{y}} + \frac{\partial v}{\partial z} \hat{\mathbf{z}} \right) = v_{,x} \hat{\mathbf{x}} + v_{,y} \hat{\mathbf{y}} + v_{,z} \hat{\mathbf{z}} \quad (1.50)$$

where $\hat{\mathbf{x}}$, $\hat{\mathbf{y}}$ and $\hat{\mathbf{z}}$ are unit vectors along the x , y and z directions, respectively. (A partial derivative with respect to x is denoted by a comma in the subscript before the x , as in $v_{,x}$. The same is true for y and z). On the other hand, when ∇ is applied to a vector, the result is a tensor. We write $\nabla \mathbf{u}$ as

$$\nabla \mathbf{u} = \nabla(u_x \hat{\mathbf{x}} + u_y \hat{\mathbf{y}} + u_z \hat{\mathbf{z}}) = \nabla u_x \otimes \hat{\mathbf{x}} + \nabla u_y \otimes \hat{\mathbf{y}} + \nabla u_z \otimes \hat{\mathbf{z}}, \quad (1.51)$$

expansion of which reveals that

$$\begin{aligned} \nabla \mathbf{u} = & (u_{x,x} \hat{\mathbf{x}} + u_{x,y} \hat{\mathbf{y}} + u_{x,z} \hat{\mathbf{z}}) \otimes \hat{\mathbf{x}} + \\ & (u_{y,x} \hat{\mathbf{x}} + u_{y,y} \hat{\mathbf{y}} + u_{y,z} \hat{\mathbf{z}}) \otimes \hat{\mathbf{y}} + \\ & (u_{z,x} \hat{\mathbf{x}} + u_{z,y} \hat{\mathbf{y}} + u_{z,z} \hat{\mathbf{z}}) \otimes \hat{\mathbf{z}} \end{aligned} \quad (1.52)$$

and consequently that

$$\begin{aligned} \nabla \mathbf{u} = & u_{x,x} \hat{\mathbf{x}} \otimes \hat{\mathbf{x}} + u_{x,y} \hat{\mathbf{y}} \otimes \hat{\mathbf{x}} + u_{x,z} \hat{\mathbf{z}} \otimes \hat{\mathbf{x}} + \\ & u_{y,x} \hat{\mathbf{x}} \otimes \hat{\mathbf{y}} + u_{y,y} \hat{\mathbf{y}} \otimes \hat{\mathbf{y}} + u_{y,z} \hat{\mathbf{z}} \otimes \hat{\mathbf{y}} + \\ & u_{z,x} \hat{\mathbf{x}} \otimes \hat{\mathbf{z}} + u_{z,y} \hat{\mathbf{y}} \otimes \hat{\mathbf{z}} + u_{z,z} \hat{\mathbf{z}} \otimes \hat{\mathbf{z}} \end{aligned} \quad (1.53)$$

This is what is meant by the gradient of a vector. The objects $\hat{\mathbf{x}} \otimes \hat{\mathbf{x}}$, $\hat{\mathbf{y}} \otimes \hat{\mathbf{x}}$, $\dots$, $\hat{\mathbf{z}} \otimes \hat{\mathbf{z}}$ are called *dyads*. The transpose of a dyad is defined as

$$(\hat{\mathbf{x}} \otimes \hat{\mathbf{y}})^T := \hat{\mathbf{y}} \otimes \hat{\mathbf{x}},$$

and so on for the other dyads. The transpose of $\nabla \mathbf{u}$ is denoted by $(\nabla \mathbf{u})^T$. In this way, the information regarding dyads and their transpose gives a meaning to the strain rate tensor $\bar{\mathbf{D}}(\mathbf{u})$ in (1.49). More detailed accounts on tensor algebra can be found in [Irgens, 2008].

The strain rate tensor $\bar{\bar{\mathbf{D}}}(\mathbf{u})$ is also important in connection with Newtonian fluids. A fluid is said to be Newtonian if it satisfies (experimentally) the three properties listed in the chart below (which brings the mathematical equivalent of these properties [Boyer and Fabrie, 2012]).

Chart 1.4: Mathematical properties of Newtonian fluids

Property I: $\bar{\bar{\mathbf{J}}}$ depends only on $\bar{\bar{\mathbf{D}}}(\mathbf{u})$.

Due to the conservation of angular momentum principle, the Cauchy stress tensor $\bar{\bar{\boldsymbol{\sigma}}}$ is symmetric [Boyer and Fabrie, 2012], i.e.,

$$\bar{\bar{\boldsymbol{\sigma}}}^T = \bar{\bar{\boldsymbol{\sigma}}} \quad (1.54)$$

The transpose of expression (1.48) is

$$\bar{\bar{\boldsymbol{\sigma}}}^T = \bar{\bar{\mathbf{J}}}^T - p\bar{\bar{\mathbf{I}}}^T. \quad (1.55)$$

From (1.54) and from the obvious fact that $\bar{\bar{\mathbf{I}}}^T = \bar{\bar{\mathbf{I}}}$, (1.55) becomes

$$\bar{\bar{\boldsymbol{\sigma}}} = \bar{\bar{\mathbf{J}}}^T - p\bar{\bar{\mathbf{I}}}. \quad (1.56)$$

A comparison between (1.48) and (1.56) allows us to conclude that $\bar{\bar{\mathbf{J}}}^T = \bar{\bar{\mathbf{J}}}$, i.e., that the viscous stress tensor $\bar{\bar{\mathbf{J}}}$ is symmetric. The transpose of expression (1.49) is

$$\left(\bar{\bar{\mathbf{D}}}(\mathbf{u})\right)^T = \frac{1}{2}((\nabla \mathbf{u})^T + ((\nabla \mathbf{u})^T)^T) = \frac{1}{2}((\nabla \mathbf{u})^T + \nabla \mathbf{u}) = \bar{\bar{\mathbf{D}}}(\mathbf{u}), \quad (1.57)$$

since $((\nabla \mathbf{u})^T)^T = \nabla \mathbf{u}$. The strain rate tensor $\bar{\bar{\mathbf{D}}}(\mathbf{u})$ is therefore also symmetric.

Let the set of all symmetric tensors be denoted by $\mathcal{S}$ (of course, we are referring to second-order tensors in three dimensions). Then, $\bar{\bar{\mathbf{D}}}(\mathbf{u}) \in \mathcal{S}$ and $\bar{\bar{\mathbf{J}}} \in \mathcal{S}$. Property I actually means that $\bar{\bar{\mathbf{J}}}$ is determined from $\bar{\bar{\mathbf{D}}}(\mathbf{u})$ by an operator $L: \mathcal{S} \rightarrow \mathcal{S}$, i.e., $\bar{\bar{\mathbf{J}}} = L(\bar{\bar{\mathbf{D}}}(\mathbf{u}))$.

Property II: The dependence of $\bar{\bar{\mathbf{J}}}$ on $\bar{\bar{\mathbf{D}}}(\mathbf{u})$ is given by a linear operator.

This property says that the operator L which relates $\bar{\bar{\mathbf{J}}}$ to $\bar{\bar{\mathbf{D}}}(\mathbf{u})$ is linear. According to the definition of linear operators [Kreyszig, 1989], [Rynne and Youngson, 2007], it means that for any two elements s_1 and s_2 in $\mathcal{S}$, and for any two real numbers α_1 and α_2 , it is true that

$$L(\alpha_1 s_1 + \alpha_2 s_2) = \alpha_1 L(s_1) + \alpha_2 L(s_2) \quad (1.58)$$

Property III: The relation between $\bar{\bar{\mathbf{J}}}$ and $\bar{\bar{\mathbf{D}}}(\mathbf{u})$ is isotropic.

This property is linked to the invariance of some fluid properties when the orthonormal frame is changed. Mathematically, it means this: Let P be an arbitrary orthogonal

matrix (i.e., a real 3×3 matrix for which $PP^T = I$, the identity matrix). Next represent P as a tensor, i.e., from the matrix

$$P = \begin{bmatrix} p_{11} & p_{12} & p_{13} \\ p_{21} & p_{21} & p_{21} \\ p_{31} & p_{31} & p_{31} \end{bmatrix} \quad (1.59)$$

we construct the associated tensor:

$$\bar{\mathbf{P}} = p_{11}\hat{\mathbf{x}} \otimes \hat{\mathbf{x}} + p_{12}\hat{\mathbf{x}} \otimes \hat{\mathbf{y}} + p_{13}\hat{\mathbf{x}} \otimes \hat{\mathbf{z}} + \dots \quad (1.60)$$

(The coordinate directions are represented either as x, y, z or as x_1, x_2, x_3 , and the occasion usually dictates which of the two forms is chosen. Notwithstanding the choice of any representation, they are equivalent to each other: x is associated with ‘1’, y with ‘2’ and z with ‘3’.)

At this point we need to define the *dot product* between tensors. Let $\bar{\mathbf{A}} = \mathbf{a} \otimes \mathbf{b}$ be a tensor formed by the vectors $\mathbf{a}$ and $\mathbf{b}$ (via the tensor product). Likewise, let $\bar{\mathbf{C}} = \mathbf{c} \otimes \mathbf{d}$ be formed by the vectors $\mathbf{c}$ and $\mathbf{d}$. The dot product between $\bar{\mathbf{A}}$ and $\bar{\mathbf{C}}$ is defined here as

$$\bar{\mathbf{A}} \cdot \bar{\mathbf{C}} = (\mathbf{a} \otimes \mathbf{b}) \cdot (\mathbf{c} \otimes \mathbf{d}) = \mathbf{a} \otimes \mathbf{b} \cdot \mathbf{c} \otimes \mathbf{d} := (\mathbf{b} \cdot \mathbf{c})\mathbf{a} \otimes \mathbf{d} \quad (1.61)$$

i.e., $\bar{\mathbf{A}} \cdot \bar{\mathbf{C}}$ is another tensor formed by the tensor $\mathbf{a} \otimes \mathbf{d}$ multiplied by the scalar $\mathbf{b} \cdot \mathbf{c}$. Under these circumstances, the product between two matrices becomes replaced by the dot product between the associated tensors.

Consider now an arbitrary symmetric tensor $s \in \mathcal{S}$. The operator L is called isotropic if it is true that [Boyer and Fabrie, 2012]:

$$L(\bar{\mathbf{P}}^T \cdot (s \cdot \bar{\mathbf{P}})) = \bar{\mathbf{P}}^T \cdot (L(s) \cdot \bar{\mathbf{P}}), \quad (1.62)$$

where $\bar{\mathbf{P}}$ is the arbitrary tensor from (1.60) and (1.59).

In Newtonian fluids, the relation between the viscous stress tensor $\bar{\mathbf{J}}$ and the strain rate tensor $\bar{\mathbf{D}}(\mathbf{u})$ – which must satisfy the three required properties from Chart 1.4 – is given by [Gerbeau *et al.*, 2006], [Boyer and Fabrie, 2012]:

$$\bar{\mathbf{J}} = 2\mu\bar{\mathbf{D}}(\mathbf{u}) + \lambda(\nabla \cdot \mathbf{u})\bar{\mathbf{I}}, \quad (1.63)$$

where μ and λ are real coefficients. At first sight, it seems that (1.63) violates the first property a Newtonian fluid must satisfy ($\bar{\mathbf{J}}$ depends on $\nabla \cdot \mathbf{u}$, so it no longer depends on $\bar{\mathbf{D}}(\mathbf{u})$ only). This difficulty is apparent because of the identity

$$\text{Tr}(\bar{\mathbf{D}}(\mathbf{u})) = \nabla \cdot \mathbf{u}, \quad (1.64)$$

where $\text{Tr}(\cdot)$ denotes the *trace* of a tensor. The trace of a matrix is defined as the sum of the entries from its main diagonal. In order to carry this definition to tensors, we need

the definition of the *double dot product between tensors*. Let again $\bar{\bar{\mathbf{A}}} = \mathbf{a} \otimes \mathbf{b}$ and $\bar{\bar{\mathbf{C}}} = \mathbf{c} \otimes \mathbf{d}$ be tensors formed by the vectors $\mathbf{a}, \mathbf{b}, \mathbf{c}$ and $\mathbf{d}$ (via the appropriate tensor products). The double dot product between $\bar{\bar{\mathbf{A}}}$ and $\bar{\bar{\mathbf{C}}}$ is defined as

$$\bar{\bar{\mathbf{A}}} : \bar{\bar{\mathbf{C}}} = (\mathbf{a} \otimes \mathbf{b}) : (\mathbf{c} \otimes \mathbf{d}) := (\mathbf{a} \cdot \mathbf{c})(\mathbf{b} \cdot \mathbf{d}), \quad (1.65)$$

i.e., $\bar{\bar{\mathbf{A}}} : \bar{\bar{\mathbf{C}}}$ is a scalar formed by the product of the ordinary dot products between vectors $\mathbf{a} \cdot \mathbf{c}$ and $\mathbf{b} \cdot \mathbf{d}$. If we represent the Cartesian basis (i.e., the set of basis unit vectors) $\{\hat{\mathbf{x}}, \hat{\mathbf{y}}, \hat{\mathbf{z}}\}$ as $\{\hat{\mathbf{e}}_1, \hat{\mathbf{e}}_2, \hat{\mathbf{e}}_3\}$, i.e., if we make the identification

$$\begin{aligned} \hat{\mathbf{x}} &\rightarrow \hat{\mathbf{e}}_1 \\ \hat{\mathbf{y}} &\rightarrow \hat{\mathbf{e}}_2 \\ \hat{\mathbf{z}} &\rightarrow \hat{\mathbf{e}}_3 \end{aligned} \quad (1.66)$$

then the trace of an arbitrary tensor $\bar{\bar{\mathbf{T}}}$ may be defined as

$$\text{Tr}(\bar{\bar{\mathbf{T}}}) = \sum_{i=1}^3 \bar{\bar{\mathbf{T}}} : (\hat{\mathbf{e}}_i \otimes \hat{\mathbf{e}}_i). \quad (1.67)$$

After the notion of the double dot product has been introduced, (1.64) follows from (1.49), (1.53), (1.66) and (1.67). The relation in (1.63) can be written in terms of $\bar{\bar{\mathbf{D}}}(\mathbf{u})$ alone as

$$\bar{\bar{\mathbf{J}}} = 2\mu \bar{\bar{\mathbf{D}}}(\mathbf{u}) + \lambda \text{Tr}(\bar{\bar{\mathbf{D}}}(\mathbf{u})) \bar{\bar{\mathbf{I}}}. \quad (1.68)$$

The trace is a linear mapping; the operator L from (1.58) assumes the form

$$L(s) = 2\mu s + \lambda \text{Tr}(s) \bar{\bar{\mathbf{I}}}, \quad (1.69)$$

from which the linear dependence on s becomes evident. Expressions (1.48) and (1.63) allows the Cauchy stress tensor to be written as

$$\bar{\bar{\boldsymbol{\sigma}}} = 2\mu \bar{\bar{\mathbf{D}}}(\mathbf{u}) + (\lambda \nabla \cdot \mathbf{u} - p) \bar{\bar{\mathbf{I}}}. \quad (1.70)$$

The two real numbers μ and λ are called the *Lamé coefficients* [Gerbeau *et al.*, 2006]. After an extensive discussion involving arguments from thermodynamics (associated with the fact that the viscous stresses are dissipative) and an analysis of the evolution equation for the entropy, one concludes that the coefficients μ and λ must be such that [Boyer and Fabrie, 2012]:

$$\mu \geq 0 \quad (1.71.a)$$

$$2\mu + 3\lambda \geq 0. \quad (1.71.b)$$

The coefficient μ is termed the *dynamic viscosity* of the flow, whereas the quantity $(2/3)\mu + \lambda$ is the *bulk viscosity* of the flow (SI units: newtons-second/square meter).

According to the kinetic theory of the monatomic gas [Gerbeau *et al.*, 2006], the relation

$$\lambda = -\frac{2}{3}\mu \quad (1.72)$$

holds true for most fluids in practice. This is also termed the *Stokes' assumption* [Boyer and Fabrie, 2012], which means that the bulk viscosity can be neglected. The expression for the viscous stress tensor $\bar{\bar{\mathcal{J}}}$ in (1.63) can therefore be simplified to

$$\bar{\bar{\mathcal{J}}} = 2\mu \left(\bar{\bar{\mathcal{D}}}(\mathbf{u}) - \frac{1}{3}(\nabla \cdot \mathbf{u})\bar{\bar{\mathbf{I}}} \right). \quad (1.73)$$

Thanks to (1.64), one discovers that $\text{Tr}(\bar{\bar{\mathcal{J}}}) = 0$, since $\text{Tr}(\bar{\bar{\mathbf{I}}}) = 3$. Moreover, from (1.48) and (1.73) the Cauchy stress tensor $\bar{\bar{\sigma}}$ assumes its final form as

$$\bar{\bar{\sigma}} = 2\mu \left(\bar{\bar{\mathcal{D}}}(\mathbf{u}) - \frac{1}{3}(\nabla \cdot \mathbf{u})\bar{\bar{\mathbf{I}}} \right) - p\bar{\bar{\mathbf{I}}}. \quad (1.74)$$

We are now prepared to go back to the conservation of linear momentum principle (1.47). It can be shown [Boyer and Fabrie, 2012] that the following identities involving the divergence of tensors hold true:

$$\nabla \cdot \left(\bar{\bar{\mathcal{D}}}(\mathbf{u}) \right) = \frac{1}{2}(\nabla^2 \mathbf{u} + \nabla \nabla \cdot \mathbf{u}), \quad (1.75.a)$$

$$\nabla \cdot \left((\nabla \cdot \mathbf{u})\bar{\bar{\mathbf{I}}} \right) = \nabla \nabla \cdot \mathbf{u}, \quad (1.75.b)$$

$$\nabla \cdot (p\bar{\bar{\mathbf{I}}}) = \nabla p. \quad (1.75.c)$$

The equations of fluid dynamics relevant to us reduce to the principles of conservation of mass (1.46) and conservation of linear momentum (1.47), which assumes a new form after considering (1.74) and (1.75.a) – (1.75.c). The result is summarized in Chart 1.5 below.

Chart 1.5: Equations of isothermal fluid dynamics

Conservation of mass:

$$\frac{\partial \rho}{\partial t} + \nabla \cdot (\rho \mathbf{u}) = 0. \quad (1.77)$$

Conservation of linear momentum:

$$\frac{\partial(\rho \mathbf{u})}{\partial t} + \nabla \cdot (\rho \mathbf{u} \otimes \mathbf{u}) - \mu \nabla^2 \mathbf{u} - \frac{1}{3}\mu \nabla \nabla \cdot \mathbf{u} + \nabla p = \rho \mathbf{f}. \quad (1.78)$$

The system of equations formed by (1.77) and (1.78) together with an extra *equation of state* which relates the thermodynamical variables (usually the pressure p

and the density ρ), is called the *Navier-Stokes system for isothermal Newtonian fluids* [Glowinski *et al.*, 2003]. We remark that the dynamical viscosity μ is a function of the temperature and the pressure. When the temperature is not constant, another differential equation must be considered in addition to (1.77) and (1.78), namely, the equation for the evolution of total energy [Boyer and Fabrie, 2012]. The aforementioned extra equation of state will also involve the temperature, and the whole system, also called the *Navier-Stokes-Fourier system* [Zeytounian, 2012], becomes more complicated. In applications for which changes in temperature are irrelevant (hence the name *isothermal*), (1.77) and (1.78) together with an equation of state relating p and ρ are sufficient to adequately describe the flow of Newtonian fluids.

1.4.4 Incompressibility

We say that a flow is *incompressible* if it satisfies one of the three equivalent characteristics listed below:

1. Given an arbitrary fluid element, its volume remains constant as the time evolves.
2. The velocity field $\mathbf{u}$ is divergence-free, i.e., for any $\mathbf{x} \in \Omega$ and for any t , it is true that

$$\nabla \cdot \mathbf{u} = 0. \quad (1.79)$$

3. The density ρ is constant along the trajectories associated with $\mathbf{u}$.

For incompressible models, the pressure is no longer related to the other thermodynamical variables. The extra equation of state becomes unnecessary, as the pressure has become an independent variable [Boyer and Fabrie, 2012]. The pressure gradient in (1.78) plays the role of a Lagrange multiplier related to the divergence-free constraint (1.79) [Boyer and Fabrie, 2012]. The Navier-Stokes system for isothermal and incompressible Newtonian fluids reads, after substituting (1.79) in (1.78):

$$\frac{\partial(\rho \mathbf{u})}{\partial t} + \nabla \cdot (\rho \mathbf{u} \otimes \mathbf{u}) - \mu(\rho) \nabla^2 \mathbf{u} + \nabla p = \rho \mathbf{f}, \quad (1.80.a)$$

$$\frac{\partial \rho}{\partial t} + \nabla \cdot (\rho \mathbf{u}) = 0, \quad (1.80.b)$$

$$\nabla \cdot \mathbf{u} = 0, \quad (1.80.c)$$

where in (1.80.a) the dependence of μ on ρ is made explicit.

In the sequel, the following identity will be useful [Gerbeau *et al.*, 2006]:

$$\nabla \cdot (\rho \mathbf{u} \otimes \mathbf{u}) = \mathbf{u} \nabla \cdot (\rho \mathbf{u}) + \rho (\mathbf{u} \cdot \nabla) \mathbf{u}. \quad (1.81)$$

Moreover, there is one last simplification to be made: The fluid shall be *homogeneous*, i.e., the density ρ shall be constant. As a consequence, the dynamic viscosity μ will also be constant. Expression (1.81) then becomes

$$\nabla \cdot (\rho \mathbf{u} \otimes \mathbf{u}) = \mathbf{u} \rho \nabla \cdot \mathbf{u} + \rho (\mathbf{u} \cdot \nabla) \mathbf{u} = \rho (\mathbf{u} \cdot \nabla) \mathbf{u}, \quad (1.82)$$

thanks to (1.80.c). The equation for the conservation of mass (1.80.b) reduces to $\nabla \cdot \mathbf{u} = 0$, identical to (1.80.c). In other words, homogeneity implies incompressibility [Glowinski *et al.*, 2003]. The Navier-Stokes system for isothermal, incompressible and homogeneous Newtonian fluids, called simply the *incompressible Navier-Stokes system* is summarized in Chart 1.6.

Chart 1.6: Incompressible Navier-Stokes equations

$$\rho \left(\frac{\partial \mathbf{u}}{\partial t} + (\mathbf{u} \cdot \nabla) \mathbf{u} \right) - \mu \nabla^2 \mathbf{u} + \nabla p = \rho \mathbf{f}, \quad (1.83.a)$$

$$\nabla \cdot \mathbf{u} = 0. \quad (1.83.b)$$

The solution process of the system (1.83) via mixed finite elements will provide the basis for the meshfree method developed in this thesis. Since we are not directly interested in the solution of (1.83), we can simplify it further. We can neglect the time derivative and divide the first equation by the density ρ , thus arriving at the steady state incompressible system:

$$-\frac{\mu}{\rho} \nabla^2 \mathbf{u} + (\mathbf{u} \cdot \nabla) \mathbf{u} + \nabla \left(\frac{p}{\rho} \right) = \mathbf{f}, \quad (1.84.a)$$

$$\nabla \cdot \mathbf{u} = 0. \quad (1.84.b)$$

In (1.84), the dynamic viscosity μ divided by the density ρ is called the *kinematic viscosity* ν . Moreover, since ρ is constant, once one determines the quotient p/ρ at a point, the real pressure p can be retrieved. From now on, we commit an abuse of notation by referring to the real pressure divided by the density (i.e., to p/ρ) simply as ‘pressure’ p . The equations (1.84) become

$$-\nu \nabla^2 \mathbf{u} + (\mathbf{u} \cdot \nabla) \mathbf{u} + \nabla p = \mathbf{f}, \quad (1.85.a)$$

$$\nabla \cdot \mathbf{u} = 0. \quad (1.85.b)$$

The system (1.85) gives the dynamics of the velocity field $\mathbf{u}$ at a given point. If we are interested in studying the flow on a region Ω , in addition to requiring (1.85) to be valid at all points $\mathbf{x} \in \Omega$, we also need suitable conditions prescribed on the boundary $\Gamma = \partial\Omega$. We shall consider only one kind of boundary condition, that in which $\mathbf{u}$ is known at all points from Γ :

$$\mathbf{u} = \mathbf{g} \quad \text{at } \Gamma, \quad (1.86)$$

i.e., we shall consider *Dirichlet conditions* for the velocity field $\mathbf{u}$ ($\mathbf{g}$ is a known function). Other types of boundary conditions for the steady state Navier-Stokes system are discussed in [Quarteroni, 2009], [Quarteroni and Valli, 1994], [Glowinski *et al.*, 2003].

When we put (1.85) and (1.86) together we get the final form of the Navier-Stokes system, stated in Chart 1.7 below.

Chart 1.7: Steady-state Incompressible Navier-Stokes equations

Find $(\mathbf{u}, p)$ such that

$$-\nu \nabla^2 \mathbf{u} + (\mathbf{u} \cdot \nabla) \mathbf{u} + \nabla p = \mathbf{f} \quad \text{in } \Omega, \quad (1.87.a)$$

$$\nabla \cdot \mathbf{u} = 0 \quad \text{in } \Omega, \quad (1.87.b)$$

$$\mathbf{u} = \mathbf{g} \quad \text{at } \Gamma. \quad (1.87.c)$$

The equations for the scattering problem from Chart 1.2 are rewritten in the Chart 1.8 below for convenience:

Chart 1.8: The modified scattering problem

Find $(\mathbf{E}^s, p)$ such that

$$\nabla^2 \mathbf{E}^s + k_0^2 \mathbf{E}^s + \nabla p = \mathbf{0} \quad \text{in } \Omega, \quad (1.88.a)$$

$$\nabla \cdot \mathbf{E}^s = 0 \quad \text{in } \Omega, \quad (1.88.b)$$

$$\hat{\mathbf{n}}_i \times \mathbf{E}^s = -\hat{\mathbf{n}}_i \times \mathbf{E}^{inc} \quad \text{at } \Gamma_i, \quad i = 1, 2, \dots \quad (1.88.c)$$

$$\hat{\mathbf{n}}_o \times \mathbf{E}^s = \mathbf{0} \quad \text{at } \Gamma_o. \quad (1.88.d)$$

A comparison between Charts 1.7 and 1.8 reveals that the scattering and the Navier-Stokes problems have a similar structure. So the idea of applying solution processes aimed at solving (1.87) to the solution of (1.88) is not meaningless. The motivation is that (1.87) can be solved by nodal finite elements, which (at least in principle) suggests that (1.88) also can. But we must go a step further: We solve (1.88) also by nodal finite elements, but we must take the mesh away. The result is that (1.88) shall be solved by a nodal meshfree method.

Of course, there are differences between (1.87) and (1.88). In (1.87.a), the pressure p is a real meaningful quantity, whereas p in (1.88.a) is just a mathematical artifact used to enforce the divergence-free condition. (A careful observation reveals that in both systems the boundary conditions for p are missing.) In (1.87.c), all components of the velocity field $\mathbf{u}$ are known at the boundary, whereas in (1.88.c) – (1.88.d) just the tangential components of the scattered electric field $\mathbf{E}^s$ are prescribed. There are other differences that will gradually be revealed as the process unfolds, particularly in what concerns the variational formulations of the aforementioned problems, which are the subject of the next chapter.

Chapter 2

Variational formulations

This chapter has two sections. The variational formulation of the steady-state incompressible Navier-Stokes system is discussed in the first section.

In the second, an analogous development is made in what concerns the wave scattering system.

The mathematical ideas necessary for assessing the variational formulations are spread throughout the text, and are introduced as they become necessary.

2.1 The Navier-Stokes system in weak form

2.1.1 Weak derivatives

In order to proceed with the variational formulations, we need some terminology first.

Let Ω be a *domain* in $\mathbb{R}^d$, i.e., an open and connected subset of $\mathbb{R}^d$. In this thesis, we shall be concerned with bounded domains only. We say that Ω is *bounded* if it can be placed within a ball of finite radius, i.e., if there is a point $x_0 \in \mathbb{R}^d$ and a positive number R such that $\Omega \subset B(x_0, R)$. If $d = 2$, $B(x_0, R)$ is just a circle of radius R centered at x_0 , whereas if $d = 3$, $B(x_0, R)$ is a sphere of radius R centered at x_0 . The definition of connectedness is more intricate [Searcoid, 2007], but for our purposes it suffices to say that a connected set cannot be represented as the union of two or more disjoint, nonempty, and open subsets.

In the subsequent development, the notion of *compact subsets* in $\mathbb{R}^d$ is needed. Although the true definition of compactness is also intricate [Searcoid, 2007], we will not need to work with the notion of compactness directly. We only need to know when a given subset of $\mathbb{R}^d$ is compact. A subset $S \subset \mathbb{R}^d$ is compact if and only if S is closed and bounded [Searcoid, 2007], [Kreyszig, 1989].

A subset $S \subset \mathbb{R}^d$ is called *closed* if it contains all its limit points. We say that $x \in \mathbb{R}^d$ is a *limit point* of S if we can find a sequence of points in S which converge to x . An arbitrary limit point x need not be in S ; if all of them happen to be in S , then S is closed. The set formed by the union of S and all its limit points is called the *closure* of S , and represented as $\bar{S}$.

A point $x \in \bar{S}$ is said to be on the *boundary* of S if it does not belong to the interior of S , i.e., if every neighborhood of x contains at least one point in S and at least one point not in S . The boundary of S is represented by ∂S .

Let V be a subset from our domain Ω . We say that V is *compactly contained* in Ω if two requirements are met: First, the closure $\bar{V}$ is contained in Ω . Second, the closure $\bar{V}$ is compact. This is sometimes represented as $V \subset\subset \Omega$. (Informally, it means that no point from either V or from its boundary ∂V touch the boundary $\partial\Omega$ of Ω .)

The space $C_0^\infty(\Omega)$ comprises all infinitely differentiable functions $\varphi: \Omega \rightarrow \mathbb{R}$ whose support is compactly contained in Ω . The support of φ is defined as:

$$\text{supp}(\varphi) := \overline{\{x \in \Omega \mid \varphi(x) \neq 0\}}. \quad (2.1)$$

So if $\varphi \in C_0^\infty(\Omega)$ then $\text{supp}(\varphi) \subset\subset \Omega$.

Until the end of this subsection, we shall assume that $d = 3$, i.e., the results will be stated in three dimensions. The same ideas apply when $d = 2$. The space $C^1(\Omega)$ comprises all functions from Ω into $\mathbb{R}$ that admit first order classical derivatives, i.e., if $u \in C^1(\Omega)$, then $\partial u / \partial x$, $\partial u / \partial y$ and $\partial u / \partial z$ are continuous at all points $x \in \Omega$.

Let $\mathbf{F}$ be an arbitrary 3×1 vector whose components are elements of $C^1(\Omega)$. We write it as $\mathbf{F} \in C^1(\Omega)^3$. Let also $\varphi \in C_0^\infty(\Omega)$ be an arbitrary *test function*. Take the identity

$$\nabla \cdot (\varphi \mathbf{F}) = \nabla \varphi \cdot \mathbf{F} + \varphi \nabla \cdot \mathbf{F}, \quad (2.2)$$

and integrate over Ω . After the Divergence Theorem and observing that φ is zero at $\partial\Omega$, (because $\text{supp}(\varphi) \subset\subset \Omega$) we conclude that

$$\int_{\Omega} \nabla \varphi \cdot \mathbf{F} \, d\Omega = - \int_{\Omega} \varphi \nabla \cdot \mathbf{F} \, d\Omega. \quad (2.3)$$

Since $\mathbf{F}$ is arbitrary, we can allow it to assume any form. Consider an arbitrary function $u \in C^1(\Omega)$. In the first choice, make $\mathbf{F} = [u, 0, 0]^T$. In the second, make $\mathbf{F} = [0, u, 0]^T$, and in the third, make $\mathbf{F} = [0, 0, u]^T$. When considering these three particular choices, (2.3) allows us to conclude that

$$\int_{\Omega} \frac{\partial \varphi}{\partial x} u \, d\Omega = - \int_{\Omega} \varphi \frac{\partial u}{\partial x} \, d\Omega \quad (2.4.a)$$

$$\int_{\Omega} \frac{\partial \varphi}{\partial y} u \, d\Omega = - \int_{\Omega} \varphi \frac{\partial u}{\partial y} \, d\Omega \quad (2.4.b)$$

$$\int_{\Omega} \frac{\partial \varphi}{\partial z} u \, d\Omega = - \int_{\Omega} \varphi \frac{\partial u}{\partial z} \, d\Omega. \quad (2.4.c)$$

Observation 2.1: From now on, whenever an integral is written in this thesis, the volume element $d\Omega$ will be omitted from the volume integrals. Analogously, the surface element $d\Gamma$ will be omitted from all surface (boundary) integrals. This allows the expressions to be written in a cleaner way, particularly when long integrands are considered. So

$$\int_{\Omega} f \quad \text{and} \quad \int_{\partial\Omega} g \quad (2.5)$$

will be written as

$$\int_{\Omega} f \quad \text{and} \quad \int_{\partial\Omega} g. \quad (2.6)$$

Identification of whether a given integral is either a volume or a boundary integral may be done by observing the proper symbol which indicates the region where the integration is performed.

The expressions in (2.4) make perfect sense. Since $u \in C^1(\Omega)$, the first derivatives in the right side of (2.4) are continuous, and these integrals are therefore well-defined, i.e., they assume finite values. There is no risk of any of them going to infinite.

We now ask if expressions (2.4) may still be meaningful if u is no longer in $C^1(\Omega)$. Particularly, we are interested in the validity of (2.4) when u belongs to another space in which the first derivatives are not well-defined. In order to proceed, we need the notion of L^p spaces.

The Lebesgue space $L^p(\Omega)$ is defined as:

$$L^p(\Omega) := \{v: \Omega \rightarrow \mathbb{R} \mid v \text{ is Lebesgue measurable and } \|v\|_{L^p(\Omega)} < \infty\}. \quad (2.7)$$

The proper clarification of the term ‘Lebesgue measurable functions’ needs introduction of a technical machinery which falls outside the scope of this thesis [Tao, 2011], [Cheney, 2001], [Rynne and Youngson, 2007]. It suffices for us to know that by restricting our attention to measurable functions we will not be dealing with functions which are ‘nonconventional’ in a sense. So we must concentrate on the second requirement in (2.7), which means

$$\|v\|_{L^p(\Omega)} = \left(\int_{\Omega} |v|^p \, d\Omega \right)^{\frac{1}{p}} < \infty, \quad (2.8)$$

where $1 \leq p < \infty$. (The Lebesgue spaces are traditionally spelled as L^p , and the same is done here. No confusion should be made between the index p in (2.8) and the pressure or the pseudopressure p presented in Chapter 1).

Another space that will be mentioned is the space of all *locally summable* functions, defined as

$$L^1_{loc}(\Omega) := \{v: \Omega \rightarrow \mathbb{R} \mid \forall S \subset\subset \Omega \ v \in L^1(S)\}, \quad (2.9)$$

i.e., we say that $v \in L^1_{loc}(\Omega)$ if, for any subset S compactly contained in Ω , it is true that v restricted to S is summable. According to the terminology from [Evans, 2010], a function is called *integrable* if it has an integral (which may assume infinite values). When the integral is finite, the function is called *summable*.

In order to relax the requirement that $u \in C^1(\Omega)$ in (2.4), we begin by noticing that, since the arbitrary test function φ and its derivatives are different from zero only at the support $\text{supp}(\varphi)$, we rewrite (2.4) as

$$\int_{\text{supp}(\varphi)} \frac{\partial \varphi}{\partial x} u \, d\Omega = - \int_{\text{supp}(\varphi)} \varphi \frac{\partial u}{\partial x} \, d\Omega \quad (2.10.a)$$

$$\int_{\text{supp}(\varphi)} \frac{\partial \varphi}{\partial y} u \, d\Omega = - \int_{\text{supp}(\varphi)} \varphi \frac{\partial u}{\partial y} \, d\Omega \quad (2.10.b)$$

$$\int_{\text{supp}(\varphi)} \frac{\partial \varphi}{\partial z} u \, d\Omega = - \int_{\text{supp}(\varphi)} \varphi \frac{\partial u}{\partial z} \, d\Omega. \quad (2.10.c)$$

Since all derivatives of φ are continuous, we see that the integrals from the left side in (2.10) will still be meaningful if $u \in L^1_{loc}(\Omega)$, according to the definition (2.9). When we assume that $u \in L^1_{loc}(\Omega)$, then it is true that

$$u \in L^1(\text{supp}(\varphi)), \quad (2.11)$$

as $\text{supp}(\varphi) \subset\subset \Omega$. Different test functions from $C_0^\infty(\Omega)$ have different supports, but they are all compactly contained in Ω . Hence the requirement for u to be summable on all such subsets, i.e., $u \in L^1_{loc}(\Omega)$.

In the left side of (2.10), u is no longer required to be continuous; it only needs to be summable on all subsets compactly contained in Ω (subsets S such that no point from either S or from its boundary ∂S touch $\partial\Omega$). But what about the right side of (2.10)? The problem is that, since we ‘replaced’ $C^1(\Omega)$ by $L^1_{loc}(\Omega)$, u may not be differentiable at all points from Ω . The space $L^1_{loc}(\Omega)$ admits discontinuous functions, which may risk the integrability of the right side of (2.10).

At this point it comes the definition of *weak derivatives*. Suppose that for any arbitrary test function φ we are able to find functions v^x, v^y and v^z in $L^1_{loc}(\Omega)$ such that

$$\int_{\Omega} \frac{\partial \varphi}{\partial x} u \, d\Omega = - \int_{\Omega} \varphi v^x \, d\Omega \quad (2.12.a)$$

$$\int_{\Omega} \frac{\partial \varphi}{\partial y} u \, d\Omega = - \int_{\Omega} \varphi v^y \, d\Omega \quad (2.12. b)$$

$$\int_{\Omega} \frac{\partial \varphi}{\partial z} u \, d\Omega = - \int_{\Omega} \varphi v^z \, d\Omega. \quad (2.12. c)$$

When that is the case, we say that v^x, v^y and v^z are the weak derivatives of u .

These functions do not need to be continuous. All that is required from them is that they are locally summable. It may happen that u belongs to $L^1_{loc}(\Omega)$, and at the same time be so badly discontinuous that no functions v^x, v^y and v^z can be found so that the right side of (2.12) makes sense. When this is the case, we say that u does not possess weak derivatives. So now we can define weak derivatives.

Definition: Weak derivatives – Let u, v^x, v^y and v^z be elements of $L^1_{loc}(\Omega)$. If for all $\varphi \in C_0^\infty(\Omega)$ it is true that

$$\int_{\Omega} \frac{\partial \varphi}{\partial x} u \, d\Omega = - \int_{\Omega} \varphi v^x \, d\Omega \quad (2.13. a)$$

$$\int_{\Omega} \frac{\partial \varphi}{\partial y} u \, d\Omega = - \int_{\Omega} \varphi v^y \, d\Omega \quad (2.13. b)$$

$$\int_{\Omega} \frac{\partial \varphi}{\partial z} u \, d\Omega = - \int_{\Omega} \varphi v^z \, d\Omega \quad (2.13. c)$$

we say that v^x, v^y and v^z are the weak partial derivatives of u with respect to x, y and z , respectively.

The weak derivatives and the classical (pointwise) derivatives are distinct objects. There may be circumstances in which they coincide, e.g. if $u \in C^1(\Omega)$ [Salsa, 2008]. In order to make this distinction apparent, the weak derivatives are sometimes written differently, as

$$D^x u, D^y u, D^z u, \quad (2.14)$$

which represents the weak partial derivatives of u with respect to x, y , and z , respectively.

The advantage of employing weak derivatives is twofold. First, they extend the notion of derivatives to functions which are not continuous. In a sense, classical derivatives may be represented as operators from $C^1(\Omega)$ into $C^0(\Omega)$. [Actually, from $C^m(\Omega)$ into $C^{m-1}(\Omega)$, $m \geq 1$. But since $C^m(\Omega) \subset C^1(\Omega)$ and $C^{m-1}(\Omega) \subset C^0(\Omega)$ for all $m \geq 1$, we concentrate on the supersets $C^1(\Omega)$ and $C^0(\Omega)$]. On the other hand, weak derivatives may be represented as operators from $L^1_{loc}(\Omega)$ into $L^1_{loc}(\Omega)$. Since in general $C^1(\Omega) \subset L^1_{loc}(\Omega)$, there are functions in $L^1_{loc}(\Omega)$ which do not possess classical

derivatives, but do possess weak derivatives. Second, weak derivatives allow a reduction in the order of the derivatives appearing in the differential equations. For example, in the variational formulations, instead of dealing with classical second-order derivatives of $\mathbf{E}^s$ in (1.88) and $\mathbf{u}$ in (1.87), we can deal with first-order weak derivatives of the same quantities.

The notion of weak derivatives is central to the finite element method, and consequently to meshfree methods as well. More details can be found in [Salsa, 2008], [Evans, 2010], [Brezis, 2010].

2.1.2 Function spaces: $L^2(\Omega)$ and $H^1(\Omega)$

Before proceeding to the Navier-Stokes system in weak form, we need some more notions, like that concerning a particular Sobolev space, which will appear over and again in the course of this work.

Definition: The space $H^1(\Omega)$ – The Sobolev space $W^{1,2}(\Omega)$ is defined as

$$W^{1,2}(\Omega) = \{v \in L^1_{loc}(\Omega) \mid v \in L^2(\Omega) \text{ and } D^i v \in L^2(\Omega), \ i = x, y, z\}. \quad (2.15)$$

The space $W^{1,2}(\Omega)$ is often written as $H^1(\Omega)$.

If a function u belongs to $H^1(\Omega)$, then u itself and all its weak partial derivatives (of course, they must exist) are square integrable, i.e.,

$$\|u\|_{L^2(\Omega)} < \infty \quad \text{and} \quad \|D^i u\|_{L^2(\Omega)} < \infty, \quad i = x, y, z, \quad (2.16)$$

according to (2.8).

If the domain Ω is bounded (has a finite measure), and if $1 \leq p_1 < p_2 \leq \infty$, then it is true that $L^{p_2}(\Omega) \subset L^{p_1}(\Omega)$ [Salsa, 2008]. So we can conclude that $L^2(\Omega) \subset L^1(\Omega)$. Moreover, if $1 \leq p \leq \infty$, then $L^p(\Omega) \subset L^1_{loc}(\Omega)$ [Salsa, 2008], which implies that $L^1(\Omega) \subset L^1_{loc}(\Omega)$. The spaces referred to so far are related as

$$L^2(\Omega) \subset L^1(\Omega) \subset L^1_{loc}(\Omega). \quad (2.17)$$

Since all functions from $H^1(\Omega)$ are also in $L^2(\Omega)$, then $H^1(\Omega) \subset L^2(\Omega)$.

The space $H^1(\Omega)$ is a Hilbert space [Brezis, 2010], [Cheney, 2001] when endowed with the inner product:

$$(u, v)_{H^1(\Omega)} := \int_{\Omega} uv + \int_{\Omega} (D^x u D^x v + D^y u D^y v + D^z u D^z v), \quad \forall u, v \in H^1(\Omega) \quad (2.18)$$

From now on, we shall commit an abuse of notation and represent the weak first derivatives as components of a gradient vector, i.e., as long as the weak derivatives of u exist, they can be represented as

$$\nabla u = [D^x u, D^y u, D^z u]^T. \quad (2.19)$$

Whenever we write a gradient such as ∇u , the context will make it clear whether we will be referring to a vector of weak derivatives as in (2.19) or to a vector of classical derivatives. The inner product in (2.18) then becomes

$$(u, v)_{H^1(\Omega)} := \int_{\Omega} uv + \int_{\Omega} \nabla u \cdot \nabla v, \quad \forall u, v \in H^1(\Omega). \quad (2.20)$$

As it happens in Hilbert spaces, the inner product in (2.18) induces a norm $\|\cdot\|_{H^1(\Omega)}$, given by

$$\|u\|_{H^1(\Omega)} := \sqrt{(u, u)_{H^1(\Omega)}} = \left(\int_{\Omega} |u|^2 + \int_{\Omega} |\nabla u|^2 \right)^{\frac{1}{2}}, \quad \forall u \in H^1(\Omega). \quad (2.21)$$

In $H^1(\Omega)$ we can also define a seminorm $|\cdot|_{H^1(\Omega)}$, expressed as

$$|u|_{H^1(\Omega)} := \left(\int_{\Omega} |\nabla u|^2 \right)^{\frac{1}{2}}, \quad \forall u \in H^1(\Omega). \quad (2.22)$$

The space $L^2(\Omega)$, of which $H^1(\Omega)$ is a subspace, is also a Hilbert space when endowed with the inner product [Brezis, 2010]:

$$(u, v)_{L^2(\Omega)} := \int_{\Omega} uv, \quad \forall u, v \in L^2(\Omega). \quad (2.23)$$

The norm induced by the inner product in $L^2(\Omega)$ is just the expression (2.8) evaluated when $p = 2$, i.e.,

$$\|u\|_{L^2(\Omega)} := \sqrt{(u, u)_{L^2(\Omega)}} = \left(\int_{\Omega} |u|^2 \right)^{\frac{1}{2}}, \quad \forall u \in L^2(\Omega). \quad (2.24)$$

From (2.18), (2.21), and (2.23), we observe that

$$\|u\|_{H^1(\Omega)}^2 = \|u\|_{L^2(\Omega)}^2 + \|D^x u\|_{L^2(\Omega)}^2 + \|D^y u\|_{L^2(\Omega)}^2 + \|D^z u\|_{L^2(\Omega)}^2, \quad \forall u \in H^1(\Omega), \quad (2.25)$$

since $D^x u$, $D^y u$ and $D^z u$ are in $L^2(\Omega)$, according to (2.15).

2.1.3 Function spaces: $L^2(\Omega)^d$ and $H^1(\Omega)^d$

When dealing with vectors whose components are elements of $L^2(\Omega)$ or $H^1(\Omega)$, it is useful to review the notion of norm extended to product spaces.

An abstract normed space consists of a linear space U together with a norm $\|\cdot\|_U$ defined on elements of U [Kreyszig, 1989], [Conway, 1994]. Let such a normed space be represented as the pair $\{U, \|\cdot\|_U\}$. Suppose we are given $d = 3$ normed spaces $\{U, \|\cdot\|_U\}$, $\{V, \|\cdot\|_V\}$ and $\{W, \|\cdot\|_W\}$. We can define a new linear space $U \times V \times W$ formed by the Cartesian product of the three linear spaces U, V, W in the following way:

$$\forall u \in U \quad \forall v \in V \quad \forall w \in W \quad [u, v, w]^T \in U \times V \times W. \quad (2.26)$$

The question is: How does the norm on $U \times V \times W$ relate to the norm on the individual spaces U, V, W ? In other words, can the norm $\|\cdot\|_{U \times V \times W}$ be written as a function of the norms on the individual spaces $\|\cdot\|_U$, $\|\cdot\|_V$ and $\|\cdot\|_W$? The answer is yes, and generally there is more than one way to accomplish that [Searcoid, 2007]. For our purposes, it will be suitable to set

$$\|\cdot\|_{U \times V \times W} = (\|\cdot\|_U^2 + \|\cdot\|_V^2 + \|\cdot\|_W^2)^{\frac{1}{2}}, \quad (2.27)$$

i.e., for arbitrary elements $u \in U$, $v \in V$ and $w \in W$ that happen to be the ‘components’ of the object $[u, v, w]^T$, the norm in $U \times V \times W$ is given by

$$\|[u, v, w]^T\|_{U \times V \times W} = (\|u\|_U^2 + \|v\|_V^2 + \|w\|_W^2)^{\frac{1}{2}}. \quad (2.28)$$

The conclusion thus far is: given $d = 3$ arbitrary normed spaces $\{U, \|\cdot\|_U\}$, $\{V, \|\cdot\|_V\}$ and $\{W, \|\cdot\|_W\}$, we can form a new normed space whose associated linear space is formed by d -dimensional column vectors whose components are elements of the individual linear spaces U, V, W , and whose associated norm is given by (2.28).

When we consider $U = V = W = L^2(\Omega)$ and $\|\cdot\|_U = \|\cdot\|_V = \|\cdot\|_W = \|\cdot\|_{L^2(\Omega)}$, we get the normed space $L^2(\Omega)^3$ formed by triples:

$$L^2(\Omega)^3 := \{[u, v, w]^T \mid u \in L^2(\Omega), v \in L^2(\Omega), w \in L^2(\Omega)\}. \quad (2.29)$$

The space $L^2(\Omega)^3$ is a Hilbert space when equipped with the inner product

$$([u_1, v_1, w_1]^T, [u_2, v_2, w_2]^T)_{L^2(\Omega)^3} := \int_{\Omega} u_1 u_2 + v_1 v_2 + w_1 w_2, \quad (2.30.a)$$

valid for all vectors $[u_1, v_1, w_1]^T$ and $[u_2, v_2, w_2]^T$ in $L^2(\Omega)^3$. The inner product (2.30.a) induces a norm:

$$(2.30.b)$$

$$\| [u, v, w]^T \|_{L^2(\Omega)^3} := \sqrt{([u, v, w]^T, [u, v, w]^T)_{L^2(\Omega)^3}} = \left(\int_{\Omega} |u|^2 + |v|^2 + |w|^2 \right)^{\frac{1}{2}}$$

that coincides with (2.27). So the norm induced by the inner product is a valid norm when we see $L^2(\Omega)^3$ as the Cartesian product of 3 spaces.

Analogously, we can define the space $H^1(\Omega)^3$:

$$H^1(\Omega)^3 := \{ [u, v, w]^T \mid u \in H^1(\Omega), v \in H^1(\Omega), w \in H^1(\Omega) \}, \quad (2.31)$$

which is a Hilbert space when endowed with the inner product

$$\begin{aligned} ([u_1, v_1, w_1]^T, [u_2, v_2, w_2]^T)_{H^1(\Omega)^3} &:= (u_1, u_2)_{H^1(\Omega)} + (v_1, v_2)_{H^1(\Omega)} + (w_1, w_2)_{H^1(\Omega)} \\ &= \int_{\Omega} u_1 u_2 + \nabla u_1 \cdot \nabla u_2 + v_1 v_2 + \nabla v_1 \cdot \nabla v_2 + w_1 w_2 + \nabla w_1 \cdot \nabla w_2. \end{aligned} \quad (2.32)$$

The norm in $H^1(\Omega)^3$ becomes

$$\begin{aligned} \| [u, v, w]^T \|_{H^1(\Omega)^3} &:= \sqrt{([u, v, w]^T, [u, v, w]^T)_{H^1(\Omega)^3}} \\ &= \left(\int_{\Omega} |u|^2 + |\nabla u|^2 + |v|^2 + |\nabla v|^2 + |w|^2 + |\nabla w|^2 \right)^{\frac{1}{2}}, \end{aligned} \quad (2.33)$$

whereas the seminorm is given by

$$| [u, v, w]^T |_{H^1(\Omega)^3} := \left(\int_{\Omega} |\nabla u|^2 + |\nabla v|^2 + |\nabla w|^2 \right)^{\frac{1}{2}}. \quad (2.34)$$

In favor of a more compact notation, let us represent the elements of either $L^2(\Omega)^3$ or $H^1(\Omega)^3$ as vectors, i.e., let us make $\mathbf{u} = [u, v, w]^T$, and so on. Then the inner product in $L^2(\Omega)^3$ (2.30.a) becomes

$$(\mathbf{u}, \mathbf{v})_{L^2(\Omega)^3} := \int_{\Omega} \mathbf{u} \cdot \mathbf{v}, \quad \forall \mathbf{u}, \mathbf{v} \in L^2(\Omega)^3, \quad (2.35)$$

and the norm (2.30.b) simplifies to

$$\| \mathbf{u} \|_{L^2(\Omega)^3} := \sqrt{(\mathbf{u}, \mathbf{u})_{L^2(\Omega)^3}} = \left(\int_{\Omega} \mathbf{u} \cdot \mathbf{u} \right)^{\frac{1}{2}}, \quad \forall \mathbf{u} \in L^2(\Omega)^3. \quad (2.36)$$

The inner product in $H^1(\Omega)^3$ (2.32) becomes

$$(\mathbf{u}, \mathbf{v})_{H^1(\Omega)^3} := \int_{\Omega} \mathbf{u} \cdot \mathbf{v} + \nabla \mathbf{u} : \nabla \mathbf{v}, \quad \forall \mathbf{u}, \mathbf{v} \in H^1(\Omega)^3, \quad (2.37)$$

if we recall the notion of double dot product from (1.65). The norm (2.33) and the seminorm (2.34) simplify to

$$\|\mathbf{u}\|_{H^1(\Omega)^3} := \sqrt{(\mathbf{u}, \mathbf{u})_{H^1(\Omega)^3}} = \left(\int_{\Omega} \mathbf{u} \cdot \mathbf{u} + \nabla \mathbf{u} : \nabla \mathbf{u} \right)^{\frac{1}{2}}, \quad \forall \mathbf{u} \in H^1(\Omega)^3 \quad (2.38)$$

and

$$|\mathbf{u}|_{H^1(\Omega)^3} := \left(\int_{\Omega} \nabla \mathbf{u} : \nabla \mathbf{u} \right)^{\frac{1}{2}}, \quad \forall \mathbf{u} \in H^1(\Omega)^3, \quad (2.39)$$

respectively.

In (2.26) – (2.39), the development has been carried out for the three-dimensional case $d = 3$. Similar results hold for the two-dimensional case $d = 2$.

We end this section by noticing some important relations, summarized in Chart 2.1 below.

Chart 2.1: Function spaces and norms

From (2.21), (2.22) and (2.24):

$$\|u\|_{H^1(\Omega)}^2 = \|u\|_{L^2(\Omega)}^2 + |u|_{H^1(\Omega)}^2, \quad \forall u \in H^1(\Omega) \quad (2.40)$$

From (2.40), (2.22) and (2.36):

$$\|u\|_{H^1(\Omega)}^2 = \|u\|_{L^2(\Omega)}^2 + \|\nabla u\|_{L^2(\Omega)^3}^2, \quad \forall u \in H^1(\Omega) \quad (2.41)$$

From (2.36), (2.38) and (2.39):

$$\|\mathbf{u}\|_{H^1(\Omega)^3}^2 = \|\mathbf{u}\|_{L^2(\Omega)^3}^2 + |\mathbf{u}|_{H^1(\Omega)^3}^2, \quad \forall \mathbf{u} \in H^1(\Omega)^3 \quad (2.42)$$

2.1.4 Function spaces: Density and trace theory

In order to establish the weak forms associated with the Navier-Stokes system (1.87), the notion of density will prove to be very useful.

Let it be an abstract normed space $\{U, \|\cdot\|_U\}$. Suppose S is a subset of U , i.e., $S \subset U$. We say that S is *dense* in U if its closure is equal to U [Kreyszig, 1989], [Conway, 1994], [Rynne and Youngson, 2007], i.e., if

$$\bar{S} = U. \quad (2.43)$$

By this, we mean that the union of S and all its limit points is equal to U . We can clarify further the notion of limit point: We say that $x_0 \in U$ is a *limit point* of S if, for every ball centered at x_0 , no matter how small, there is at least one point $y \in S$ such that y is distinct from x_0 . This can be expressed symbolically as

$$\forall \varepsilon > 0 \quad \exists y \in B(x_0, \varepsilon) \quad y \in S \quad \text{and} \quad y \neq x_0. \quad (2.44)$$

If we make ε successively smaller, e.g., $\varepsilon = 1/n$, $n \in \mathbb{N}$, we get the more useful equivalent result: $x_0 \in U$ is a limit point of the subset S if there is a sequence $s_1, \dots, s_n, \dots$ of elements from S such that s_n converges to x_0 . Symbolically,

$$\exists \{s_n\}_{n=1}^{\infty} \subset S \quad s_n \rightarrow x_0 \quad (2.45)$$

A sequence in S is just a map from the natural numbers into the subset S , i.e., a map $s: \mathbb{N} \rightarrow S$. In (2.45), the $\{s_n\}_{n=1}^{\infty}$ represents the range of the map s , which evidently is a subset of S .

The idea of density roughly represents this: Given an arbitrary point x_0 from U which is not necessarily in S , it can nonetheless be ‘approximated’ by a sequence of elements which are in S . The most interesting case happens when x_0 is not an element from S . The density hypothesis says that, despite the fact x_0 is not in S , there are other elements from S that are infinitely close to x_0 . But how is this ‘closeness’ actually measured? It is measured by the norm of the superspace U , i.e., by $\|\cdot\|_U$. The ball in (2.44) means

$$B(x_0, \varepsilon) = \{z \in U \mid \|z - x_0\|_U < \varepsilon\}, \quad (2.46)$$

so that convergence in (2.45) is indeed the convergence in the $\|\cdot\|_U$ norm, i.e.,

$$\forall \varepsilon > 0 \quad \exists N \in \mathbb{N} \quad \forall n \geq N \quad \|s_n - x_0\|_U < \varepsilon. \quad (2.47)$$

In order to make clear that the convergence is in the $\|\cdot\|_U$ norm, (2.43) is often written as

$$\overline{S}^U = U. \quad (2.48)$$

The notion of density has been introduced in a rather abstract way; in order for it to be useful, it should be specialized to some of the spaces introduced thus far.

The space $C_0^\infty(\Omega)$ together with all its limit points regarding the $\|\cdot\|_{L^p(\Omega)}$ norm in (2.8) is the space $L^p(\Omega)$ itself, when $1 \leq p < \infty$ [Salsa, 2008], [Brezis, 2010]. When $p = 2$, we may write

$$\overline{C_0^\infty(\Omega)}^{L^2(\Omega)} = L^2(\Omega) \quad (2.49)$$

where convergence is measured in the $\|\cdot\|_{L^2(\Omega)}$ norm from (2.24).

The space $C_0^\infty(\Omega)$ together with all its limit points regarding the $\|\cdot\|_{H^1(\Omega)}$ norm in (2.21) is also very special. It is a subspace of $H^1(\Omega)$ [Salsa, 2008], [Brezis, 2010] and it is denoted by $H_0^1(\Omega)$. It will occur frequently in the course of this work. Formally,

$$H_0^1(\Omega) := \overline{C_0^\infty(\Omega)}^{H^1(\Omega)}. \quad (2.50)$$

The space $H_0^1(\Omega)$ is defined by (2.50), whose meaning is: Given an arbitrary $u \in H_0^1(\Omega)$, there is a sequence of elements from $C_0^\infty(\Omega)$ which converges to u in the $\|\cdot\|_{H^1(\Omega)}$ norm. Specializing (2.45),

$$\exists \{\varphi_n\}_{n=1}^\infty \subset C_0^\infty(\Omega) \quad \varphi_n \rightarrow u. \quad (2.51)$$

The support of any function φ in $C_0^\infty(\Omega)$ is compactly contained in Ω , i.e., $\text{supp}(\varphi) \subset\subset \Omega$. So φ is zero at the boundary $\Gamma = \partial\Omega$, i.e., $\varphi|_\Gamma = 0$. This characteristic is somehow inherited by the functions in $H_0^1(\Omega)$, i.e., any function u in $H_0^1(\Omega)$ is ‘somehow’ zero at Γ . It is said that it has *zero trace* on Γ . (It is not quite correct to say that u assumes the value zero at all points from Γ . The reason is that elements from the Lebesgue spaces are not defined pointwise. Proper explanation of this fact requires ideas from measure theory that are outside the scope of this work. [Tao, 2011])

In order to clarify the idea of trace, some more notions are required. Spaces whose elements are functions which admit continuous derivatives up to order m are represented by $C^m(\Omega)$:

$$C^m(\Omega) = \{u: \Omega \rightarrow \mathbb{R} \mid u \text{ is } m \text{ times continuously differentiable}\}. \quad (2.52)$$

The space $C^1(\Omega)$ from the beginning of Section 2.1.1 is just (2.52) specialized to the case $m = 1$. Let us concentrate on the case when $m = \infty$ and $\Omega = \mathbb{R}^d$, i.e., the whole space. The space $C^\infty(\mathbb{R}^d)$ comprises those functions which admit continuous derivatives of all orders at all points from $\mathbb{R}^d$. If $u \in C^\infty(\mathbb{R}^d)$, then u is well-defined and admits continuous derivatives at all points from $\mathbb{R}^d$, particularly at those which lie inside the domain Ω and at those on the boundary $\Gamma = \partial\Omega$ as well. Form now the space which consists of the restrictions to $\bar{\Omega}$ of functions in $C^\infty(\mathbb{R}^d)$, i.e., the space

$$C^\infty(\bar{\Omega}) := \{u: \bar{\Omega} \rightarrow \mathbb{R} \mid u = \psi|_{\bar{\Omega}}, \psi \in C^\infty(\mathbb{R}^d)\}. \quad (2.53)$$

There is a very important theorem, which summarizes the notion of trace [Evans, 2010], [Salsa, 2008], [Boffi *et al.*, 2013], [Boyer and Fabrie, 2012], [Leoni, 2009], [Girault and Raviart, 1986], [Galdi, 2011].

Theorem 2.1: The Trace Theorem – *Let Ω be a bounded and Lipschitz domain in $\mathbb{R}^d$. Then there exists a linear operator (the trace operator) $\gamma_0: H^1(\Omega) \rightarrow L^2(\Gamma)$ such that:*

1. *If $\varphi \in C^\infty(\bar{\Omega})$, then $\gamma_0\varphi = \varphi|_\Gamma$.*
2. *There is a constant $C > 0$ such that $\|\gamma_0 u\|_{L^2(\Gamma)} \leq C\|u\|_{H^1(\Omega)}$ for all $u \in H^1(\Omega)$.*

The notion of Lipschitz domain is rather technical [Galdi, 2011], but it suffices to say here that ordinary domains such as squares, rectangles, triangles, circles, cubes and spheres are Lipschitz. Moreover, the constant C in the theorem above depends on the domain Ω and on the dimension d , sometimes being represented as $C(\Omega, d)$ [Salsa, 2008]. Of course, it is *independent* of u .

Theorem (2.1) concerns the existence of an operator that ascribes functions from $L^2(\Gamma)$ – functions which are defined at the boundary Γ – to functions from $H^1(\Omega)$. When the function φ is in $C^\infty(\overline{\Omega})$, which is obviously a subspace of $H^1(\Omega)$, it is well-behaved enough to be associated with its restriction to the boundary $\varphi|_\Gamma$. When a function u is in $H^1(\Omega)$ but not in $C^\infty(\overline{\Omega})$, it is associated to the function $\gamma_0 u$. This function $\gamma_0 u$ is not defined pointwise (due to the technicalities from measure theory [Tao, 2011]), but on the other hand its norm in $L^2(\Gamma)$ is related to the norm of the original function u in $H^1(\Omega)$.

The trace operator γ_0 is not surjective, i.e., there are functions from $L^2(\Gamma)$ which are not in the range of γ_0 . It is proved that the range of γ_0 is surjective on the space $H^{1/2}(\Gamma)$, a Sobolev space of fractional order, and whose characterization is not trivial [Leoni, 2009]. In order to find out if a given function defined on the boundary Γ is a trace from another function in $H^1(\Omega)$, the following result from [Boffi et al., 2013] is useful:

$$H^1(\Gamma) \subset \gamma_0(H^1(\Omega)) \subset L^2(\Gamma), \quad (2.54)$$

where $\gamma_0(H^1(\Omega)) = H^{1/2}(\Gamma)$ is the range (or image) of γ_0 . Expression (2.54) says that if a function g defined on the boundary is in $H^1(\Gamma)$, then it is guaranteed to be in the range of the trace operator, i.e, $g \in \gamma_0(H^1(\Omega))$, which implies that there is a v in $H^1(\Omega)$ such that $\gamma_0 v = g$.

As the notion of trace has been clarified, one may ask about those functions u from $H^1(\Omega)$ which have zero trace on Γ , i.e., functions such that $\|\gamma_0 u\|_{L^2(\Gamma)} = 0$. It can be proved [Boyer and Fabrie, 2012] that these functions form a space, which is precisely the space $H_0^1(\Omega)$ defined in (2.50):

$$\text{Ker } \gamma_0 = H_0^1(\Omega), \quad (2.55)$$

i.e., the kernel (or null space) of the trace operator is precisely the space $H_0^1(\Omega)$.

The results introduced so far concerning density and traces can be extended to the product spaces $H^1(\Omega)^3$.

$$H_0^1(\Omega) \times H_0^1(\Omega) \times H_0^1(\Omega) = H_0^1(\Omega)^3 \subset H^1(\Omega)^3 \quad (2.56.a)$$

$$C_0^\infty(\Omega) \times C_0^\infty(\Omega) \times C_0^\infty(\Omega) = C_0^\infty(\Omega)^3 \quad (2.56.b)$$

$$H_0^1(\Omega)^3 := \overline{C_0^\infty(\Omega)^3}^{H^1(\Omega)^3}. \quad (2.56.c)$$

Expression (2.56.c) says that a function $\mathbf{v}$ from $H_0^1(\Omega)^3$ can be approximated by a sequence in $C_0^\infty(\Omega)^3$ which converges in the $\|\cdot\|_{H^1(\Omega)^3}$ norm (2.33) to $\mathbf{v}$. Moreover, the product version of the space in (2.52) becomes:

$$C^\infty(\bar{\Omega}) \times C^\infty(\bar{\Omega}) \times C^\infty(\bar{\Omega}) = C^\infty(\bar{\Omega})^3. \quad (2.57)$$

If we define the multidimensional trace operator $\boldsymbol{\gamma}_0^d: H^1(\Omega)^3 \rightarrow H^{1/2}(\Gamma)^3$ as

$$\boldsymbol{\gamma}_0^d \mathbf{u} = \boldsymbol{\gamma}_0^d [u_x, u_y, u_z]^T := [\gamma_0 u_x, \gamma_0 u_y, \gamma_0 u_z]^T, \quad (2.58)$$

then the trace theorem applied to each of the components of $\mathbf{u}$ allows us to conclude that

$$\boldsymbol{\varphi} \in C^\infty(\bar{\Omega})^3 \Rightarrow \boldsymbol{\gamma}_0^d \boldsymbol{\varphi} = \boldsymbol{\varphi}|_\Gamma, \quad (2.59.a)$$

$$\exists C(\Omega, d) > 0 \quad \forall \mathbf{u} \in H^1(\Omega)^3 \quad \|\boldsymbol{\gamma}_0^d \mathbf{u}\|_{L^2(\Gamma)^3} \leq C \|\mathbf{u}\|_{H^1(\Omega)^3}, \quad (2.59.b)$$

where the ' $\Rightarrow$ ' arrow is the implication connective (if... then). The norm $\|\cdot\|_{L^2(\Gamma)^3}$ is the same as that from (2.31) or (2.36). Also,

$$\text{Ker } \boldsymbol{\gamma}_0^d = H_0^1(\Omega)^3. \quad (2.60)$$

The reasoning (2.56) – (2.58) applies also to the bidimensional case $d = 2$.

2.1.5 Navier-Stokes: Weak forms and weak solutions

It is now time to return to the Navier-Stokes system (1.87), rewritten below for convenience:

Find $(\mathbf{u}, p)$ such that

$$-\nu \nabla^2 \mathbf{u} + (\mathbf{u} \cdot \nabla) \mathbf{u} + \nabla p = \mathbf{f} \quad \text{in } \Omega, \quad (2.61.a)$$

$$\nabla \cdot \mathbf{u} = 0 \quad \text{in } \Omega, \quad (2.61.b)$$

$$\mathbf{u} = \mathbf{g} \quad \text{at } \Gamma. \quad (2.61.c)$$

We call $(\mathbf{u}, p)$ a *classical solution* if all derivatives appearing in (2.61) are defined pointwise. The 'classical' velocity field $\mathbf{u}$ belongs to the space $C^2(\Omega)^3$, in which $C^2(\Omega)$ has been defined in (2.52). In a classical solution, generally it is required that $\mathbf{u}$ be well-behaved close to the boundary $\partial\Omega$; one then adds the requirement that $\mathbf{u}$ must also belong to the space $C(\bar{\Omega})^3$, where

$$C(\bar{\Omega}) = \{u \in C(\Omega) \mid u \text{ is uniformly continuous}\}. \quad (2.62)$$

Thus if $u \in C(\bar{\Omega})$ then u can be continuously extended to the boundary $\partial\Omega$, i.e., when going from the interior to the boundary $\partial\Omega$, one experiences no discontinuity. So $\mathbf{u} \in C^2(\Omega)^3 \cap C(\bar{\Omega})^3$. In the same way, the 'classical' pressure p belongs to $C^1(\Omega) \cap$

$C(\bar{\Omega})$. Finally, the ‘classical picture’ is completed by requiring the excitation $\mathbf{f}$ to be in $C(\Omega)^3$ and the boundary condition $\mathbf{g}$ to be in $C(\Gamma)^3$. We can rewrite problem (2.61) as

Find $(\mathbf{u}, p) \in C^2(\Omega)^3 \cap C(\bar{\Omega})^3 \times C^1(\Omega) \cap C(\bar{\Omega})$ such that

$$-\nu \nabla^2 \mathbf{u} + (\mathbf{u} \cdot \nabla) \mathbf{u} + \nabla p = \mathbf{f} \quad \text{in } \Omega, \quad (2.63.a)$$

$$\nabla \cdot \mathbf{u} = 0 \quad \text{in } \Omega, \quad (2.63.b)$$

$$\mathbf{u} = \mathbf{g} \quad \text{at } \Gamma. \quad (2.63.c)$$

In problem (2.63), all derivatives are the classical (pointwise) derivatives.

In order to devise a strategy to solve (2.63), one must first show that the problem is well-posed, i.e., that the solution to (2.63) exists, is unique and depends continuously on the data $\mathbf{f}$ and $\mathbf{g}$. However, such a task may prove to be very difficult, if not impossible. Moreover, there may be situations of physical interest in which the data $\mathbf{f}$ and $\mathbf{g}$ are not continuous. The question is that requiring everything to be continuous is a fairly restrictive hypothesis, and the solution to our problem may not exist.

A reasonable idea is to ‘relax’ the requirements on the solution we are seeking. Hopefully, since we have somehow widened the search space of our solution, it may become easier to find out if the problem in this new setting is well-posed. Roughly speaking, this new ‘relaxed solution’ is the *weak solution* to our problem. It usually happens that the enlarged search space has a richer structure, the exploration of which is greatly enhanced by the tools and inequalities available from functional analysis. In this way it becomes easier to establish the well-posedness in the new setting.

After the existence of the weak solution has been established, one may begin to inquire about its *smoothness*. At this point one tries to show that the weak solution is more regular than expected. For example, one initially shows that a weak solution exists in $H^1(\Omega)$; thereafter he may be able to show that this solution happens to be in the more regular space $H^2(\Omega)$, and so on. In general, given a weak solution $u \in H^1(\Omega)$, one may try to solve the problem:

$$\text{Find } \max \{s \mid u \in H^s(\Omega)\}. \quad (2.63.d)$$

A solution u that happens to be in $H^s(\Omega)$ for $s > 1$ is usually termed a *strong solution*. If the solution is found to be regular enough, then one may study if it qualifies as a classical solution. Such questions are addressed by the *regularity theory*, which is a very advanced branch in the study of partial differential equations and is outside the scope of this work. The book by [Evans, 2010] brings more discussions about the concept of weak solutions and the problem of regularity.

In this thesis, we shall be concerned with the weak solutions only. As will become clearer later, the finite element method (and consequently our meshfree method) seeks for approximations of the weak solutions. We devised a neat way to present the process of going from the classical form (2.63.a) – (2.63.c) to the weak

form, which explores all the notions introduced so far. It will be applied to the Navier-Stokes system first and to the scattering system later.

2.1.5.1 The problem in classical form

The problem should be stated in its classical form, as in (2.63). Write the residuals of (2.63.a) and (2.63.b), as below.

$$\text{Find } (\mathbf{u}, p) \in C^2(\Omega)^3 \cap C(\bar{\Omega})^3 \times C^1(\Omega) \cap C(\bar{\Omega}) \text{ such that}$$

$$-\nu \nabla^2 \mathbf{u} + (\mathbf{u} \cdot \nabla) \mathbf{u} + \nabla p - \mathbf{f} = \mathbf{0} \quad \text{in } \Omega, \quad (2.64.a)$$

$$\nabla \cdot \mathbf{u} = 0 \quad \text{in } \Omega, \quad (2.64.b)$$

$$\mathbf{u} = \mathbf{g} \quad \text{at } \Gamma. \quad (2.64.c)$$

2.1.5.2 Testing functions

The first equation (2.64.a) is multiplied by an arbitrary testing function $\boldsymbol{\varphi} \in C_0^\infty(\Omega)^3$ and (2.64.b) by another arbitrary testing function $\varphi \in C_0^\infty(\Omega)$. The result is integrated over the domain Ω . After application of successive vector identities, one arrives at the expressions

$$\text{Find } (\mathbf{u}, p) \in C^2(\Omega)^3 \cap C(\bar{\Omega})^3 \times C^1(\Omega) \cap C(\bar{\Omega}) \text{ such that}$$

$$\int_{\Omega} \nu \nabla \mathbf{u} : \nabla \boldsymbol{\varphi} + \int_{\Omega} [(\mathbf{u} \cdot \nabla) \mathbf{u}] \cdot \boldsymbol{\varphi} - \int_{\Omega} p \nabla \cdot \boldsymbol{\varphi} - \int_{\Omega} \mathbf{f} \cdot \boldsymbol{\varphi} - \oint_{\Omega} \left(\frac{\partial \mathbf{u}}{\partial n} - p \hat{\mathbf{n}} \right) \cdot \boldsymbol{\varphi} = 0, \quad \forall \boldsymbol{\varphi} \in C_0^\infty(\Omega)^3 \quad (2.65.a)$$

$$\int_{\Omega} \varphi \nabla \cdot \mathbf{u} = 0, \quad \forall \varphi \in C_0^\infty(\Omega) \quad (2.65.b)$$

$$\mathbf{u} = \mathbf{g}, \quad \text{at } \Gamma. \quad (2.65.c)$$

Since $\boldsymbol{\varphi}|_{\Gamma} = \mathbf{0}$, as all components are elements from $C_0^\infty(\Omega)$ (a space whose functions are compactly contained in Ω), the surface integral in (2.65.a) is disregarded.

2.1.5.3 Relaxing the requirements

Let us write $\boldsymbol{\varphi} = [\varphi^x, \varphi^y, \varphi^z]^T$. The first integral in (2.65.a) is a sum like

$$\int_{\Omega} \nu \left(\frac{\partial u_x}{\partial x} \frac{\partial \varphi^x}{\partial x} + \frac{\partial u_x}{\partial y} \frac{\partial \varphi^x}{\partial y} + \dots + \frac{\partial u_z}{\partial z} \frac{\partial \varphi^z}{\partial z} \right). \quad (2.66)$$

Since all terms involving the test functions in (2.66) are compactly supported, the integral in (2.66) still makes sense if we require that $\partial u_x/\partial x, \dots, \partial u_z/\partial z$ are in $L^1_{loc}(\Omega)$. This is equivalent to saying that all weak partial derivatives of u must exist. However, requiring only that all components of $\mathbf{u}$ and its weak derivatives are in $L^1_{loc}(\Omega)$ adds too much freedom to the ‘relaxed’ solution. For reasons that will become apparent as we progress, it is better to restrict it a little bit and require that all components of $\mathbf{u}$ and its weak derivatives are in $L^2(\Omega) \subset L^1_{loc}(\Omega)$. In other words, the initial space $C^2(\Omega)$ is too restrictive, and $L^1_{loc}(\Omega)$ is too permissive. The intermediary space $L^2(\Omega)$ looks as a promising choice.

Requiring that all components of $\mathbf{u}$ and its weak derivatives are in $L^2(\Omega)$ is the same as requiring that $\mathbf{u} \in H^1(\Omega)^3$.

The second integral in (2.65.a) is a sum like

$$\int_{\Omega} \left(u_x \frac{\partial u_x}{\partial x} \varphi^x + u_y \frac{\partial u_x}{\partial y} \varphi^x + \dots + u_z \frac{\partial u_z}{\partial z} \varphi^z \right) \quad (2.67)$$

In order to verify if (2.67) is summable, let us evaluate how its individual terms behave. There is a result which will prove to be very useful. It will be stated in the form of a theorem, whose proof is in [Brezis, 2010].

Theorem 2.2: The Hölder inequality – Let $f \in L^p(\Omega)$ and $g \in L^q(\Omega)$ with $1 \leq p \leq \infty$ and $1/p + 1/q = 1$. Then $fg \in L^1(\Omega)$ and

$$\|fg\|_{L^1(\Omega)} = \int_{\Omega} |fg| \leq \|f\|_{L^p(\Omega)} \|g\|_{L^q(\Omega)}. \quad (2.68)$$

Let us concentrate on the first term from (2.67). Since $u_x \in L^2(\Omega)$ and $\partial u_x/\partial x \in L^2(\Omega)$, then $u_x \partial u_x/\partial x \in L^1(\Omega)$ due to the Hölder inequality for $p = 2$. According to (2.17), $L^1(\Omega) \subset L^1_{loc}(\Omega)$, so we see that $u_x \partial u_x/\partial x \in L^1_{loc}(\Omega)$. Finally, since $\text{supp}(\varphi^x) \subset\subset \Omega$, it can be concluded that the first term in (2.67) is summable. The same analysis can be extended to all the remaining terms from (2.67), and the conclusion is the same: They are also summable. Therefore, the whole expression (2.67) is summable, i.e., the integral is finite.

In the third integral from (2.65.a), $\nabla \cdot \boldsymbol{\varphi} = \partial \varphi^x/\partial x + \partial \varphi^y/\partial y + \partial \varphi^z/\partial z$. Since φ^x, φ^y and φ^z are compactly contained in Ω , the same is also true for their partial derivatives. It follows that $\text{supp}(\nabla \cdot \boldsymbol{\varphi}) \subset\subset \Omega$. So the third term in (2.65.a) makes sense if $p \in L^1_{loc}(\Omega)$. But we have already concluded that $L^2(\Omega) \subset L^1_{loc}(\Omega)$ is a better choice, so we demand that $p \in L^2(\Omega)$.

According to what is said at the end of Section 2.5.1.2, the surface integral is disregarded, so the only instance in which p appears in the problem is in the third integral from (2.65.a). One can observe that p is determined up to a constant. In order to see this, suppose $(\mathbf{u}, p)$ is a solution to (2.65). Will $(\mathbf{u}, p + c)$, $c \in \mathbb{R}$, also be a

solution to (2.65)? When we replace p by $p + c$ in (2.65), the whole expression (2.65.a) remains the same, except for the extra term

$$\int_{\Omega} c \nabla \cdot \boldsymbol{\varphi}. \quad (2.69)$$

When the Divergence theorem is applied to (2.69), one gets

$$\int_{\Omega} c \nabla \cdot \boldsymbol{\varphi} = c \int_{\Omega} \nabla \cdot \boldsymbol{\varphi} = c \oint_{\Gamma} \boldsymbol{\varphi} \cdot \hat{\mathbf{n}} = 0, \quad (2.70)$$

since $\boldsymbol{\varphi} \in C_0^\infty(\Omega)^3$ is zero at the boundary Γ .

So if p is a solution to the problem, then $p + c$ will be also. So the solution space for p seems to be $L^2(\Omega)$ divided into equivalence classes (subsets) in such a way that the elements of a class are precisely those functions p which differ from each other by a constant. In order to make the solution p unique, one usually proceeds by choosing a single representative from each class. The representative element of each class is chosen as that one which has *zero average* over Ω . By restricting p to be the zero average representative of each class, the right space for searching p is [Boyer and Fabrie, 2012], [Galdi, 2011], [Girault and Raviart, 1986], [Ern and Guermond, 2004], [Glowinski *et al.*, 2003]:

$$L_0^2(\Omega) := \left\{ p \in L^2(\Omega) \mid \int_{\Omega} p = 0 \right\}. \quad (2.71)$$

The fourth integral from (2.65.a) is a sum like

$$\int_{\Omega} f_x \varphi^x + f_y \varphi^y + f_z \varphi^z, \quad (2.72)$$

where the excitation vector $\mathbf{f}$ has been represented as $[f_x, f_y, f_z]^T$. As φ^x , φ^y and φ^z are compactly supported in Ω , the integral in (2.72) still makes sense if all components of $\mathbf{f}$ are in $L_{loc}^1(\Omega)$. Again, we simply demand that $\mathbf{f} \in L^2(\Omega)^3$.

In order to evaluate the integral in (2.65.b), we need to inquire about the divergence $\nabla \cdot \mathbf{u}$. We have already required that $\mathbf{u} \in H^1(\Omega)^3$, which implies that the weak derivatives of all components of $\mathbf{u}$ are in $L^2(\Omega)$. Particularly, $\partial u_x / \partial x \in L^2(\Omega)$, $\partial u_y / \partial y \in L^2(\Omega)$ and $\partial u_z / \partial z \in L^2(\Omega)$. Since $\text{supp}(\boldsymbol{\varphi}) \subset\subset \Omega$, (2.65.b) makes sense if we show that $\nabla \cdot \mathbf{u} \in L_{loc}^1(\Omega)$. Another very useful result is the following inequality, whose proof can be found in [Brezis, 2010].

Theorem 2.3: The Minkowski inequality in L^p spaces – Assume $1 \leq p \leq \infty$, $u \in L^p(\Omega)$ and $v \in L^p(\Omega)$. Then

$$\|u + v\|_{L^p(\Omega)} \leq \|u\|_{L^p(\Omega)} + \|v\|_{L^p(\Omega)}. \quad (2.73)$$

Applying (2.73) to $\partial u_x/\partial x$ and $\partial u_y/\partial y$, we get that

$$\left\| \frac{\partial u_x}{\partial x} + \frac{\partial u_y}{\partial y} \right\|_{L^2(\Omega)} \leq \left\| \frac{\partial u_x}{\partial x} \right\|_{L^2(\Omega)} + \left\| \frac{\partial u_y}{\partial y} \right\|_{L^2(\Omega)}, \quad (2.74)$$

and we conclude that $\partial u_x/\partial x + \partial u_y/\partial y$ is in $L^2(\Omega)$, since $\partial u_x/\partial x \in L^2(\Omega)$ and $\partial u_y/\partial y \in L^2(\Omega)$. Next, we apply (2.73) again to $\partial u_x/\partial x + \partial u_y/\partial y$ and $\partial u_z/\partial z$. We get

$$\left\| \left(\frac{\partial u_x}{\partial x} + \frac{\partial u_y}{\partial y} \right) + \frac{\partial u_z}{\partial z} \right\|_{L^2(\Omega)} \leq \left\| \frac{\partial u_x}{\partial x} + \frac{\partial u_y}{\partial y} \right\|_{L^2(\Omega)} + \left\| \frac{\partial u_z}{\partial z} \right\|_{L^2(\Omega)}, \quad (2.75)$$

and conclude that the left side of (2.75), which nothing else than $\nabla \cdot \mathbf{u}$, is in $L^2(\Omega)$. Consequently, by (2.17), $\nabla \cdot \mathbf{u} \in L^1_{loc}(\Omega)$.

The only term left to analyze is the boundary condition (2.65.c). Initially, we demanded that $\mathbf{u} = \mathbf{g} \in C(\Gamma)^3$. However, since now we require that $\mathbf{u} \in H^1(\Omega)^3$, there is no sense in asking $\mathbf{u}$ to be equal to $\mathbf{g}$ pointwise at Γ . We must relax it a little and require that $\mathbf{u}$ be equal to $\mathbf{g}$ in the sense of the traces, i.e., we require that

$$\gamma_0^d \mathbf{u} = \mathbf{g}. \quad (2.76)$$

So the new requirement for $\mathbf{g}$ is that it should be in the range of the trace operator γ_0^d , i.e., we must require that $\mathbf{g} \in H^{1/2}(\Gamma)^3$.

We have now analyzed (2.65) term by term, and concluded that it is safe to relax the requirements in order to enlarge the search space. The conclusions are summarized in the table below.

TABLE 2.1 – REQUIREMENTS ON THE QUANTITIES IN THE NAVIER-STOKES SYSTEM

Quantity	Classical solution	‘Relaxed’ solution
$\mathbf{u}$	$C^2(\Omega)^3 \cap C(\bar{\Omega})^3$	$H^1(\Omega)^3$
p	$C^1(\Omega) \cap C(\bar{\Omega})$	$L^2_0(\Omega)$
$\mathbf{f}$	$C(\Omega)^3$	$L^2(\Omega)^3$
$\mathbf{g}$	$C(\Gamma)^3$	$H^{1/2}(\Gamma)^3$

The ‘relaxed’ problem thus becomes:

$$\text{Find } (\mathbf{u}, p) \in H^1(\Omega)^3 \times L^2_0(\Omega) \text{ such that} \quad (2.77.a)$$

$$\int_{\Omega} \nu \nabla \mathbf{u} : \nabla \boldsymbol{\varphi} + \int_{\Omega} [(\mathbf{u} \cdot \nabla) \mathbf{u}] \cdot \boldsymbol{\varphi} - \int_{\Omega} p \nabla \cdot \boldsymbol{\varphi} - \int_{\Omega} \mathbf{f} \cdot \boldsymbol{\varphi} = 0, \quad \forall \boldsymbol{\varphi} \in C_0^\infty(\Omega)^3$$

$$\int_{\Omega} \varphi \nabla \cdot \mathbf{u} = 0, \quad \forall \varphi \in C_0^\infty(\Omega) \quad (2.77.b)$$

$$\boldsymbol{\gamma}_0^d \mathbf{u} = \mathbf{g}. \quad (2.77.c)$$

All derivatives which appear in system (2.77) are weak derivatives.

2.1.5.4 Lifting on the boundary data

The trace operator $\boldsymbol{\gamma}_0^d: H^1(\Omega)^3 \rightarrow H^{1/2}(\Gamma)^3$ is surjective, but not injective; according to (2.60), its kernel is the whole space $H_0^1(\Omega)^3$, and therefore different from zero.

From (2.77.c), we learn that $\mathbf{g}$ is in the range of $\boldsymbol{\gamma}_0^d$. Since this operator is not injective, there is more than one function from $H^1(\Omega)^3$ associated to $\mathbf{g}$. In order to see this, let $\mathbf{v} \in H_0^1(\Omega)^3$ be arbitrary. The trace operator is linear, so $\boldsymbol{\gamma}_0^d(\mathbf{u} + \mathbf{v}) = \boldsymbol{\gamma}_0^d \mathbf{u} + \boldsymbol{\gamma}_0^d \mathbf{v} = \boldsymbol{\gamma}_0^d \mathbf{u} = \mathbf{g}$. So the trace of $\mathbf{u} + \mathbf{v}$ is equal to trace of $\mathbf{u}$, but obviously $\mathbf{u} + \mathbf{v}$ is different from $\mathbf{u}$, since $\mathbf{v}$ can be anything in $H_0^1(\Omega)^3$.

So there must be another function $\mathbf{u}^g$ in $H^1(\Omega)^3$, different from $\mathbf{u}$, such that $\boldsymbol{\gamma}_0^d \mathbf{u}^g$ is also $\mathbf{g}$. Let us take this particular $\mathbf{u}^g$ and set

$$\mathbf{u} = \mathbf{u}^0 + \mathbf{u}^g. \quad (2.78)$$

Applying the trace operator to both sides of (2.78), one readily concludes that $\boldsymbol{\gamma}_0^d \mathbf{u}^0 = \mathbf{0}$.

The function $\mathbf{u}^g$ is called the *lifting* on the original Dirichlet boundary condition $\boldsymbol{\gamma}_0^d \mathbf{u} = \mathbf{g}$. The idea is that it is a somehow known function: Once we are given the boundary condition $\mathbf{g}$, we can find a particular function in $H^1(\Omega)^3$ such that its trace is $\mathbf{g}$, because the trace operator is surjective. For example, let $\mathbf{u}^g$ be the simplest function in $H^1(\Omega)^3$ we can imagine such that $\boldsymbol{\gamma}_0^d \mathbf{u}^g = \mathbf{g}$. Despite the fact that finding such an $\mathbf{u}^g$ here at the continuous level is not a straightforward task, it turns out to be very easy at the finite element level. More discussions on the lifting procedure can be found in [Girault and Raviart, 1986], [Boyer and Fabrie, 2012], [Quarteroni, 2009], [Ern and Guermond, 2004].

After $\mathbf{u}^g$ has been determined, when we insert it in (2.78), it becomes clear that $\mathbf{u}^0$ is the true unknown. Substituting (2.78) in (2.77), we get a new problem:

Find $(\mathbf{u}^0, p) \in H^1(\Omega)^3 \times L_0^2(\Omega)$ such that

$$\int_{\Omega} \nu \nabla \mathbf{u}^0 : \nabla \boldsymbol{\varphi} + \int_{\Omega} \nu \nabla \mathbf{u}^g : \nabla \boldsymbol{\varphi} + \int_{\Omega} [(\mathbf{u}^0 \cdot \nabla) \mathbf{u}^0] \cdot \boldsymbol{\varphi} + \int_{\Omega} [(\mathbf{u}^0 \cdot \nabla) \mathbf{u}^g] \cdot \boldsymbol{\varphi} +$$

$$\int_{\Omega} [(\mathbf{u}^g \cdot \nabla) \mathbf{u}^0] \cdot \boldsymbol{\varphi} + \int_{\Omega} [(\mathbf{u}^g \cdot \nabla) \mathbf{u}^g] \cdot \boldsymbol{\varphi} - \int_{\Omega} p \nabla \cdot \boldsymbol{\varphi} - \int_{\Omega} \mathbf{f} \cdot \boldsymbol{\varphi} = 0, \quad \forall \boldsymbol{\varphi} \in C_0^\infty(\Omega)^3 \quad (2.79.a)$$

$$\int_{\Omega} \boldsymbol{\varphi} \nabla \cdot \mathbf{u}^0 + \int_{\Omega} \boldsymbol{\varphi} \nabla \cdot \mathbf{u}^g = 0, \quad \forall \boldsymbol{\varphi} \in C_0^\infty(\Omega) \quad (2.79.b)$$

$$\boldsymbol{\gamma}_0^d \mathbf{u}^0 = \mathbf{0}. \quad (2.79.c)$$

The advantage of the lifting is that we no longer have to worry about non-homogeneous Dirichlet boundary conditions: They enter the problem through suitable integrals involving a known quantity, namely, $\mathbf{u}^g$. In the new problem (2.79), homogeneous Dirichlet boundary conditions are to be imposed, since $\boldsymbol{\gamma}_0^d \mathbf{u}^0 = \mathbf{0}$, according to (2.79.c). But this amounts to saying that $\mathbf{u}^0 \in H_0^1(\Omega)^3$, so we may rewrite (2.79) as

Find $(\mathbf{u}^0, p) \in H_0^1(\Omega)^3 \times L_0^2(\Omega)$ such that

$$\begin{aligned} & \int_{\Omega} \nu \nabla \mathbf{u}^0 : \nabla \boldsymbol{\varphi} + \int_{\Omega} \nu \nabla \mathbf{u}^g : \nabla \boldsymbol{\varphi} + \int_{\Omega} [(\mathbf{u}^0 \cdot \nabla) \mathbf{u}^0] \cdot \boldsymbol{\varphi} + \int_{\Omega} [(\mathbf{u}^0 \cdot \nabla) \mathbf{u}^g] \cdot \boldsymbol{\varphi} + \\ & \int_{\Omega} [(\mathbf{u}^g \cdot \nabla) \mathbf{u}^0] \cdot \boldsymbol{\varphi} + \int_{\Omega} [(\mathbf{u}^g \cdot \nabla) \mathbf{u}^g] \cdot \boldsymbol{\varphi} - \int_{\Omega} p \nabla \cdot \boldsymbol{\varphi} - \int_{\Omega} \mathbf{f} \cdot \boldsymbol{\varphi} = 0, \\ & \forall \boldsymbol{\varphi} \in C_0^\infty(\Omega)^3 \end{aligned} \quad (2.80.a)$$

$$\int_{\Omega} \boldsymbol{\varphi} \nabla \cdot \mathbf{u}^0 + \int_{\Omega} \boldsymbol{\varphi} \nabla \cdot \mathbf{u}^g = 0, \quad \forall \boldsymbol{\varphi} \in C_0^\infty(\Omega). \quad (2.80.b)$$

The homogeneous Dirichlet boundary conditions have been embedded in the search space for $\mathbf{u}^0$, which now becomes $H_0^1(\Omega)^3$.

2.1.5.5 The G map

Expressions (2.80.a) and (2.80.b) can be summed together into a single expression as

Find $(\mathbf{u}^0, p) \in H_0^1(\Omega)^3 \times L_0^2(\Omega)$ such that

$$\begin{aligned} & \int_{\Omega} \nu \nabla \mathbf{u}^0 : \nabla \boldsymbol{\varphi} + \int_{\Omega} \nu \nabla \mathbf{u}^g : \nabla \boldsymbol{\varphi} + \int_{\Omega} [(\mathbf{u}^0 \cdot \nabla) \mathbf{u}^0] \cdot \boldsymbol{\varphi} + \int_{\Omega} [(\mathbf{u}^0 \cdot \nabla) \mathbf{u}^g] \cdot \boldsymbol{\varphi} + \\ & \int_{\Omega} [(\mathbf{u}^g \cdot \nabla) \mathbf{u}^0] \cdot \boldsymbol{\varphi} + \int_{\Omega} [(\mathbf{u}^g \cdot \nabla) \mathbf{u}^g] \cdot \boldsymbol{\varphi} - \int_{\Omega} p \nabla \cdot \boldsymbol{\varphi} - \int_{\Omega} \mathbf{f} \cdot \boldsymbol{\varphi} + \int_{\Omega} \boldsymbol{\varphi} \nabla \cdot \mathbf{u}^0 + \end{aligned}$$

$$\int_{\Omega} \varphi \nabla \cdot \mathbf{u}^g = 0, \quad \forall \varphi \in C_0^\infty(\Omega)^3 \quad \forall \varphi \in C_0^\infty(\Omega). \quad (2.81)$$

Let us introduce the map

$$G: H^1(\Omega)^3 \times L^2(\Omega) \times H^1(\Omega)^3 \times L^2(\Omega) \rightarrow \mathbb{R}, \quad (2.82)$$

defined by

$$\begin{aligned} G(\mathbf{v}_1, q_1, \mathbf{v}_2, q_2) = & \int_{\Omega} \nu \nabla \mathbf{v}_1 : \nabla \mathbf{v}_2 + \int_{\Omega} \nu \nabla \mathbf{u}^g : \nabla \mathbf{v}_2 + \int_{\Omega} [(\mathbf{v}_1 \cdot \nabla) \mathbf{v}_1] \cdot \mathbf{v}_2 + \\ & \int_{\Omega} [(\mathbf{v}_1 \cdot \nabla) \mathbf{u}^g] \cdot \mathbf{v}_2 + \int_{\Omega} [(\mathbf{u}^g \cdot \nabla) \mathbf{v}_1] \cdot \mathbf{v}_2 + \int_{\Omega} [(\mathbf{u}^g \cdot \nabla) \mathbf{u}^g] \cdot \mathbf{v}_2 - \int_{\Omega} q_1 \nabla \cdot \mathbf{v}_2 - \\ & \int_{\Omega} \mathbf{f} \cdot \mathbf{v}_2 + \int_{\Omega} q_2 \nabla \cdot \mathbf{v}_1 + \int_{\Omega} q_2 \nabla \cdot \mathbf{u}^g, \end{aligned} \quad (2.83)$$

where $\mathbf{u}^g$ and $\mathbf{f}$ are given functions (already known from the previous subsections). Since $H_0^1(\Omega)^3$ and $C_0^\infty(\Omega)^3$ are subsets of $H^1(\Omega)^3$, and $L_0^2(\Omega)$ and $C_0^\infty(\Omega)$ are subsets of $L^2(\Omega)$, problem (2.81) can be recast as:

$$\begin{aligned} & \text{Find } (\mathbf{u}^0, p) \in H_0^1(\Omega)^3 \times L_0^2(\Omega) \text{ such that} \\ & G(\mathbf{u}^0, p, \varphi, \varphi) = 0, \quad \forall \varphi \in C_0^\infty(\Omega)^3 \quad \forall \varphi \in C_0^\infty(\Omega). \end{aligned} \quad (2.84)$$

According to this definition, the map G is linear in the last two arguments. In order to determine if G is also bounded with respect to the last two arguments, we need some results. The following two inequalities hold [Quarteroni, 2009]:

$$\left| \int_{\Omega} \nu \nabla \mathbf{v} : \nabla \mathbf{w} \right| \leq \nu |\mathbf{v}|_{H^1(\Omega)^3} |\mathbf{w}|_{H^1(\Omega)^3} \quad \forall \mathbf{v}, \mathbf{w} \in H^1(\Omega)^3 \quad (2.85.a)$$

$$\left| \int_{\Omega} q \nabla \cdot \mathbf{v} \right| \leq \|q\|_{L^2(\Omega)} |\mathbf{v}|_{H^1(\Omega)^3} \quad \forall q \in L^2(\Omega) \quad \forall \mathbf{v} \in H^1(\Omega)^3, \quad (2.85.b)$$

where $|\cdot|_{H^1(\Omega)^3}$ is the seminorm from (2.39). Relation (2.42) allows the seminorms in (2.85) to be replaced by norms, i.e.,

$$\left| \int_{\Omega} \nu \nabla \mathbf{v} : \nabla \mathbf{w} \right| \leq \nu \|\mathbf{v}\|_{H^1(\Omega)^3} \|\mathbf{w}\|_{H^1(\Omega)^3} \quad \forall \mathbf{v}, \mathbf{w} \in H^1(\Omega)^3, \quad (2.86.a)$$

$$\left| \int_{\Omega} q \nabla \cdot \mathbf{v} \right| \leq \|q\|_{L^2(\Omega)} \|\mathbf{v}\|_{H^1(\Omega)^3} \quad \forall q \in L^2(\Omega) \quad \forall \mathbf{v} \in H^1(\Omega)^3. \quad (2.86.b)$$

Boundedness of the nonlinear term is provided by the following theorem (stated as a lemma and proved in [Girault and Raviart, 1986]):

Theorem 2.4: Boundedness of the nonlinear term – For $d \leq 4$, the form

$$\int_{\Omega} [(\mathbf{w}_1 \cdot \nabla) \mathbf{w}_2] \cdot \mathbf{w}_3 \quad (2.87)$$

is continuous on $H^1(\Omega)^d$, i.e., there is a positive constant C such that for all $\mathbf{w}_1, \mathbf{w}_2$ and $\mathbf{w}_3$ in $H^1(\Omega)^d$,

$$\left| \int_{\Omega} [(\mathbf{w}_1 \cdot \nabla) \mathbf{w}_2] \cdot \mathbf{w}_3 \right| \leq C \|\mathbf{w}_2\|_{H^1(\Omega)^d} \|\mathbf{w}_3\|_{H^1(\Omega)^3} \|\mathbf{w}_1\|_{H^1(\Omega)^3} \quad (2.88)$$

Since in our case $d = 2$ or $d = 3$, theorem 2.4 holds true. Moreover, thanks to (2.42), the seminorm in (2.88) can be replaced by a norm, i.e., for all $\mathbf{w}_1, \mathbf{w}_2$ and $\mathbf{w}_3$ in $H^1(\Omega)^3$,

$$\left| \int_{\Omega} [(\mathbf{w}_1 \cdot \nabla) \mathbf{w}_2] \cdot \mathbf{w}_3 \right| \leq C \|\mathbf{w}_2\|_{H^1(\Omega)^3} \|\mathbf{w}_3\|_{H^1(\Omega)^3} \|\mathbf{w}_1\|_{H^1(\Omega)^3}. \quad (2.89)$$

The last result we need is an extension of the Hölder inequality (2.68) to $L^2(\Omega)^3$. We recall the Cauchy-Schwarz inequality in $\mathbb{R}^3$, which states that, for two vectors $\mathbf{a}$ and $\mathbf{b}$ in $\mathbb{R}^3$,

$$|\mathbf{a} \cdot \mathbf{b}| \leq (\mathbf{a} \cdot \mathbf{a})^{\frac{1}{2}} (\mathbf{b} \cdot \mathbf{b})^{\frac{1}{2}}. \quad (2.90)$$

Let $\mathbf{v}$ and $\mathbf{w}$ be arbitrary elements from $L^2(\Omega)^3$. We may write

$$\left| \int_{\Omega} \mathbf{v} \cdot \mathbf{w} \right| \leq \int_{\Omega} |\mathbf{v} \cdot \mathbf{w}| \leq \int_{\Omega} (\mathbf{v} \cdot \mathbf{v})^{\frac{1}{2}} (\mathbf{w} \cdot \mathbf{w})^{\frac{1}{2}}. \quad (2.91)$$

If we make $f = (\mathbf{v} \cdot \mathbf{v})^{1/2}$ and $g = (\mathbf{w} \cdot \mathbf{w})^{1/2}$ in (2.68) with $p = 2$, we get

$$\int_{\Omega} (\mathbf{v} \cdot \mathbf{v})^{\frac{1}{2}} (\mathbf{w} \cdot \mathbf{w})^{\frac{1}{2}} \leq \left(\int_{\Omega} \mathbf{v} \cdot \mathbf{v} \right)^{\frac{1}{2}} \left(\int_{\Omega} \mathbf{w} \cdot \mathbf{w} \right)^{\frac{1}{2}} = \|\mathbf{v}\|_{L^2(\Omega)^3} \|\mathbf{w}\|_{L^2(\Omega)^3}. \quad (2.92)$$

From (2.91) and (2.92),

$$\left| \int_{\Omega} \mathbf{v} \cdot \mathbf{w} \right| \leq \|\mathbf{v}\|_{L^2(\Omega)^3} \|\mathbf{w}\|_{L^2(\Omega)^3}, \quad \forall \mathbf{v}, \mathbf{w} \in L^2(\Omega)^3. \quad (2.93)$$

Consequently, if we assume further that $\mathbf{w} \in H^1(\Omega)^3 \subset L^2(\Omega)^3$, then $\|\mathbf{w}\|_{L^2(\Omega)^3} \leq \|\mathbf{w}\|_{H^1(\Omega)^3}$, due to (2.42). Finally,

$$\left| \int_{\Omega} \mathbf{v} \cdot \mathbf{w} \right| \leq \|\mathbf{v}\|_{L^2(\Omega)^3} \|\mathbf{w}\|_{H^1(\Omega)^3}, \quad \forall \mathbf{v} \in L^2(\Omega)^3 \quad \forall \mathbf{w} \in H^1(\Omega)^3. \quad (2.94)$$

We are now at a position to evaluate if G is bounded with respect to the last two arguments, or, equivalently, if G depends continuously on its last two arguments. We explore the fact that the absolute value of a sum of terms is smaller than or equal to the sum of the absolute value of each term, and then apply (2.86.a), (2.86.b), (2.89) and (2.94) whenever it is necessary. Then,

$$\begin{aligned}
|G(\mathbf{v}_1, q_1, \mathbf{v}_2, q_2)| &\leq ([\nu \|\mathbf{v}_1\|_{H^1(\Omega)^3} + \nu \|\mathbf{u}^g\|_{H^1(\Omega)^3} + C \|\mathbf{v}_1\|_{H^1(\Omega)^3}^2 + \\
&2C \|\mathbf{v}_1\|_{H^1(\Omega)^3} \|\mathbf{u}^g\|_{H^1(\Omega)^3} + C \|\mathbf{u}^g\|_{H^1(\Omega)^3}^2 + \|q_1\|_{L^2(\Omega)} + \|\mathbf{f}\|_{L^2(\Omega)^3}) \|\mathbf{v}_2\|_{H^1(\Omega)^3} + \\
&(\|\mathbf{v}_1\|_{H^1(\Omega)^3} + \|\mathbf{u}^g\|_{H^1(\Omega)^3}) \|q_2\|_{L^2(\Omega)}.
\end{aligned} \tag{2.95}$$

In this way it becomes clear to us that G is bounded with respect to $\mathbf{v}_2$ and q_2 .

2.1.5.6 Enlarging the space of testing functions

According to (2.84), the solution $(\mathbf{u}^0, p)$ to our problem can be given a new meaning: When we insert $\mathbf{u}^0$ and p as the first two arguments, the G map assumes the value zero as the third argument varies over $C_0^\infty(\Omega)^3$ and the fourth varies over $C_0^\infty(\Omega)$.

But one may ask: What happens if the third and fourth arguments vary over spaces larger than $C_0^\infty(\Omega)^3$ and $C_0^\infty(\Omega)$, respectively? The question is that such spaces are too regular, and their elements are not that easy to obtain. In practice, it would be good if the third and fourth arguments could vary over other spaces $\mathcal{X} \supset C_0^\infty(\Omega)^3$ and $\mathcal{Y} \supset C_0^\infty(\Omega)$, while at the same time keeping the G map equal to zero. If we are successful in showing that such spaces $\mathcal{X}$ and $\mathcal{Y}$ exist, then the solution to our problem is still $(\mathbf{u}^0, p)$, but it now allows less regular candidates as testing functions.

We claim that such spaces $\mathcal{X}$ and $\mathcal{Y}$ exist: They are $\mathcal{X} = H_0^1(\Omega)^3$ and $\mathcal{Y} = L_0^2(\Omega)$. In order to show this, let $(\mathbf{u}^0, p)$ be the solution to problem (2.84). We need to prove that

$$G(\mathbf{u}^0, p, \mathbf{v}, q) = 0, \quad \forall \mathbf{v} \in H_0^1(\Omega)^3 \text{ and } \forall q \in L_0^2(\Omega) \tag{2.96}$$

Proof: Let $\mathbf{v} \in H_0^1(\Omega)^3$ and $q \in L_0^2(\Omega) \subset L^2(\Omega)$ be arbitrary. According to the density results from (2.56.c) and (2.49), respectively,

$$\exists \{\boldsymbol{\varphi}_n\}_{n=1}^\infty \subset C_0^\infty(\Omega)^3 \quad \|\mathbf{v} - \boldsymbol{\varphi}_n\|_{H^1(\Omega)^3} \rightarrow 0 \tag{2.97.a}$$

$$\exists \{\varphi_n\}_{n=1}^\infty \subset C_0^\infty(\Omega) \quad \|q - \varphi_n\|_{L^2(\Omega)} \rightarrow 0. \tag{2.97.b}$$

Since all elements from the sequence $\{\boldsymbol{\varphi}_n\}_{n=1}^\infty$ are in $C_0^\infty(\Omega)^3$, and all elements from the sequence $\{\varphi_n\}_{n=1}^\infty$ are in $C_0^\infty(\Omega)$, we can employ them as testing functions in (2.84). The G map is zero, so we write

$$\forall n \in \mathbb{N} \quad G(\mathbf{u}^0, p, \boldsymbol{\varphi}_n, \varphi_n) = 0. \tag{2.98}$$

The map $G: H^1(\Omega)^3 \times L^2(\Omega) \times H^1(\Omega)^3 \times L^2(\Omega) \rightarrow \mathbb{R}$ is linear in the last two arguments, so we write:

$$G(\mathbf{u}^0, p, \mathbf{v}, q) - G(\mathbf{u}^0, p, \boldsymbol{\varphi}_n, \varphi_n) = G(\mathbf{u}^0, p, \mathbf{v} - \boldsymbol{\varphi}_n, q - \varphi_n), \quad (2.99)$$

where (2.99) holds for all $n \in \mathbb{N}$. Of course,

$$|G(\mathbf{u}^0, p, \mathbf{v}, q) - G(\mathbf{u}^0, p, \boldsymbol{\varphi}_n, \varphi_n)| = |G(\mathbf{u}^0, p, \mathbf{v} - \boldsymbol{\varphi}_n, q - \varphi_n)|. \quad (2.100)$$

But since G is bounded with respect to the two last arguments, from (2.95) we get:

$$\begin{aligned} |G(\mathbf{u}^0, p, \mathbf{v} - \boldsymbol{\varphi}_n, q - \varphi_n)| &\leq ([\nu \|\mathbf{u}^0\|_{H^1(\Omega)^3} + \nu \|\mathbf{u}^g\|_{H^1(\Omega)^3} + C \|\mathbf{u}^0\|_{H^1(\Omega)^3}^2 + \\ &2C \|\mathbf{u}^0\|_{H^1(\Omega)^3} \|\mathbf{u}^g\|_{H^1(\Omega)^3} + C \|\mathbf{u}^g\|_{H^1(\Omega)^3}^2 + \|p\|_{L^2(\Omega)} + \|\mathbf{f}\|_{L^2(\Omega)^3}) \|\mathbf{v} - \boldsymbol{\varphi}_n\|_{H^1(\Omega)^3} + \\ &(\|\mathbf{u}^0\|_{H^1(\Omega)^3} + \|\mathbf{u}^g\|_{H^1(\Omega)^3}) \|q - \varphi_n\|_{L^2(\Omega)}. \end{aligned} \quad (2.101)$$

We have already verified that $\mathbf{u}^0 \in H_0^1(\Omega)^3$, $\mathbf{u}^g \in H^1(\Omega)^3$, $p \in L_0^2(\Omega)$ and $\mathbf{f} \in L^2(\Omega)^3$. So all the norms within parentheses in (2.101) are finite; for the sake of clarity, let us rewrite (2.101) as

$$|G(\mathbf{u}^0, p, \mathbf{v} - \boldsymbol{\varphi}_n, q - \varphi_n)| \leq M_1 \|\mathbf{v} - \boldsymbol{\varphi}_n\|_{H^1(\Omega)^3} + M_2 \|q - \varphi_n\|_{L^2(\Omega)}, \quad (2.102)$$

where the constants M_1 and M_2 are finite and depend on $\mathbf{u}^0$, $\mathbf{u}^g$, p and $\mathbf{f}$.

We now let $n \rightarrow \infty$. The right side of (2.102) goes to zero, thanks to (2.97). Naturally,

$$|G(\mathbf{u}^0, p, \mathbf{v} - \boldsymbol{\varphi}_n, q - \varphi_n)| \rightarrow 0. \quad (2.103)$$

From (2.100) and (2.103),

$$|G(\mathbf{u}^0, p, \mathbf{v}, q) - G(\mathbf{u}^0, p, \boldsymbol{\varphi}_n, \varphi_n)| \rightarrow 0. \quad (2.104)$$

But $G(\mathbf{u}^0, p, \boldsymbol{\varphi}_n, \varphi_n) = 0$ for all n , according to (2.98). Expression (2.104) therefore is true only if $G(\mathbf{u}^0, p, \mathbf{v}, q) = 0$. So we are allowed to conclude that

$$G(\mathbf{u}^0, p, \mathbf{v}, q) = 0. \quad (2.105)$$

Since $\mathbf{v}$ and q are arbitrary, we are able to see that indeed

$$G(\mathbf{u}^0, p, \mathbf{v}, q) = 0, \quad \forall \mathbf{v} \in H_0^1(\Omega)^3 \text{ and } \forall q \in L_0^2(\Omega), \quad (2.106)$$

as we have set ourselves to prove in (2.96). ■

The G map is zero when we consider the enlarged spaces $H_0^1(\Omega)^3$ and $L_0^2(\Omega)$; problem (2.84) then assumes a new form:

$$\begin{aligned} & \text{Find } (\mathbf{u}^0, p) \in H_0^1(\Omega)^3 \times L_0^2(\Omega) \text{ such that} \\ & G(\mathbf{u}^0, p, \mathbf{v}, q) = 0, \quad \forall \mathbf{v} \in H_0^1(\Omega)^3 \quad \forall q \in L_0^2(\Omega). \end{aligned} \quad (2.107)$$

When we consider the definition of the G map in (2.83), we get

$$\begin{aligned} & \text{Find } (\mathbf{u}^0, p) \in H_0^1(\Omega)^3 \times L_0^2(\Omega) \text{ such that} \\ & \int_{\Omega} \nu \nabla \mathbf{u}^0 : \nabla \mathbf{v} + \int_{\Omega} \nu \nabla \mathbf{u}^g : \nabla \mathbf{v} + \int_{\Omega} [(\mathbf{u}^0 \cdot \nabla) \mathbf{u}^0] \cdot \mathbf{v} + \int_{\Omega} [(\mathbf{u}^0 \cdot \nabla) \mathbf{u}^g] \cdot \mathbf{v} + \\ & \int_{\Omega} [(\mathbf{u}^g \cdot \nabla) \mathbf{u}^0] \cdot \mathbf{v} + \int_{\Omega} [(\mathbf{u}^g \cdot \nabla) \mathbf{u}^g] \cdot \mathbf{v} - \int_{\Omega} p \nabla \cdot \mathbf{v} - \int_{\Omega} \mathbf{f} \cdot \mathbf{v} + \int_{\Omega} q \nabla \cdot \mathbf{u}^0 + \\ & \int_{\Omega} q \nabla \cdot \mathbf{u}^g = 0, \quad \forall \mathbf{v} \in H_0^1(\Omega)^3 \quad \forall q \in L_0^2(\Omega). \end{aligned} \quad (2.108)$$

By first making $q = 0$ and $\mathbf{v}$ arbitrary, and by making $\mathbf{v} = 0$ and q arbitrary, we are able to recover the Navier-Stokes system (2.80):

$$\begin{aligned} & \text{Find } (\mathbf{u}^0, p) \in H_0^1(\Omega)^3 \times L_0^2(\Omega) \text{ such that} \\ & \int_{\Omega} \nu \nabla \mathbf{u}^0 : \nabla \mathbf{v} + \int_{\Omega} \nu \nabla \mathbf{u}^g : \nabla \mathbf{v} + \int_{\Omega} [(\mathbf{u}^0 \cdot \nabla) \mathbf{u}^0] \cdot \mathbf{v} + \int_{\Omega} [(\mathbf{u}^0 \cdot \nabla) \mathbf{u}^g] \cdot \mathbf{v} + \\ & \int_{\Omega} [(\mathbf{u}^g \cdot \nabla) \mathbf{u}^0] \cdot \mathbf{v} + \int_{\Omega} [(\mathbf{u}^g \cdot \nabla) \mathbf{u}^g] \cdot \mathbf{v} - \int_{\Omega} p \nabla \cdot \mathbf{v} - \int_{\Omega} \mathbf{f} \cdot \mathbf{v} = 0, \\ & \quad \quad \quad \forall \mathbf{v} \in H_0^1(\Omega)^3 \end{aligned} \quad (2.109.a)$$

$$\int_{\Omega} q \nabla \cdot \mathbf{u}^0 + \int_{\Omega} q \nabla \cdot \mathbf{u}^g = 0, \quad \forall q \in L_0^2(\Omega), \quad (2.109.b)$$

but now with the testing functions in the enlarged spaces $H_0^1(\Omega)^3$ and $L_0^2(\Omega)$.

2.1.5.7 Weak solutions

The system (2.109) is said to be in *weak form*, due to the fact that ‘relaxed’ assumptions have been made for the derivatives and function spaces. After getting $(\mathbf{u}^0, p)$ from (2.109), one adds the known lifting function $\mathbf{u}^g$ to $\mathbf{u}^0$ and then gets the solution $\mathbf{u} = \mathbf{u}^0 + \mathbf{u}^g$, in the spirit of (2.78). The final solution $(\mathbf{u}, p)$ thus obtained is said to be the *weak solution* associated with the original problem (2.61). Said in another way, (2.109) is the *variational formulation* of problem (2.61).

The work regarding the variational formulation of the Navier-Stokes system is by now essentially done. We have covered all the details from the process, and more importantly, we have managed to put all information concerning the procedure together in a single place. The idea of the G map we introduced here happened to be particularly

useful in clarifying how the extension of testing functions to less regular spaces is actually carried out.

We must now go through the same process again in order to study the scattering system, whose solution is the main topic of this thesis. Fortunately, since all the machinery has already been introduced, the progress will be swift.

2.2 The scattering system in weak form

2.2.1 Scattering equations

We begin by rewriting below the equations (1.88) which describe the scattering problem:

Find $(\mathbf{E}^s, p)$ such that

$$\nabla^2 \mathbf{E}^s + k_0^2 \mathbf{E}^s + \nabla p = \mathbf{0} \quad \text{in } \Omega, \quad (2.110.a)$$

$$\nabla \cdot \mathbf{E}^s = 0 \quad \text{in } \Omega, \quad (2.110.b)$$

$$\hat{\mathbf{n}}_i \times \mathbf{E}^s = -\hat{\mathbf{n}}_i \times \mathbf{E}^{inc} \quad \text{at } \Gamma_i, \quad i = 1, 2, \dots \quad (2.110.c)$$

$$\hat{\mathbf{n}}_o \times \mathbf{E}^s = \mathbf{0} \quad \text{at } \Gamma_o. \quad (2.110.d)$$

In the course of this thesis, we shall be concerned with the scattering of electromagnetic waves by a single object only, i.e., we shall focus on the surroundings of a single scatterer. So there is only one PEC surface, denoted by Γ_1 . After this simplifying assumption, problem (2.110) becomes:

Find $(\mathbf{E}^s, p)$ such that

$$\nabla^2 \mathbf{E}^s + k_0^2 \mathbf{E}^s + \nabla p = \mathbf{0} \quad \text{in } \Omega, \quad (2.111.a)$$

$$\nabla \cdot \mathbf{E}^s = 0 \quad \text{in } \Omega, \quad (2.111.b)$$

$$\hat{\mathbf{n}}_1 \times \mathbf{E}^s = -\hat{\mathbf{n}}_1 \times \mathbf{E}^{inc} \quad \text{at } \Gamma_1, \quad (2.111.c)$$

$$\hat{\mathbf{n}}_o \times \mathbf{E}^s = \mathbf{0} \quad \text{at } \Gamma_o. \quad (2.111.d)$$

Let us proceed to derive the variational form associated with (2.111).

2.2.2 PML I: Incorporating the PML

The system in form (2.111) actually models an ‘irradiating surface’ Γ_1 which acts as a source for scattered waves that are simply reflected back by the PEC surface at Γ_o . In order to correctly model outward propagating waves, these waves ‘irradiated’ by the surface Γ_1 must be attenuated in such a way that they become essentially zero by the

time they reach the PEC surface Γ_o . The idea of the PML (discussed in Section 1.3) is to place a layer of an artificial absorbing reflectionless material covering some distance from the exterior PEC surface Γ_o . Therefore some material parameter must enter the system (2.111).

The PML type to be employed in this work requires the domain Ω to be a rectangular parallelepiped surrounding the three-dimensional scatterer (or a rectangle surrounding a two-dimensional scatterer). It is a *rectangular PML*. In other words, Γ_o must be the surface of a box (or the contour of a rectangle). A given scatterer is characterized by a hole within the domain, and of course, it can have any shape. In what regards the mathematical aspect, incorporation of the PML introduces certain functions which act on the higher derivatives of the electrical field $\mathbf{E}^s$ in (2.111). The vector Laplacian $\nabla^2 \mathbf{E}^s$ in (2.111.a) must be replaced by

$$\nabla \cdot \bar{\bar{\mathbf{A}}} \cdot \nabla \mathbf{E}^s, \quad (2.112)$$

where $\bar{\bar{\mathbf{A}}}$ is a tensor whose components assume the form

$$\bar{\bar{\mathbf{A}}} = \Lambda_x \hat{\mathbf{x}} \otimes \hat{\mathbf{x}} + \Lambda_y \hat{\mathbf{y}} \otimes \hat{\mathbf{y}} + \Lambda_z \hat{\mathbf{z}} \otimes \hat{\mathbf{z}}, \quad (2.113)$$

and ∇ is the nabla operator (vector). The components Λ_x , Λ_y and Λ_z assume complex values, and will be presented later in Section 3.3.6.6. Incorporation of the PML modifies the system (2.111) into

Find $(\mathbf{E}^s, p)$ such that

$$\nabla \cdot \bar{\bar{\mathbf{A}}} \cdot \nabla \mathbf{E}^s + k_0^2 \mathbf{E}^s + \nabla p = \mathbf{0} \quad \text{in } \Omega, \quad (2.114.a)$$

$$\nabla \cdot \mathbf{E}^s = 0 \quad \text{in } \Omega, \quad (2.114.b)$$

$$\hat{\mathbf{n}}_1 \times \mathbf{E}^s = -\hat{\mathbf{n}}_1 \times \mathbf{E}^{inc} \quad \text{at } \Gamma_1, \quad (2.114.c)$$

$$\hat{\mathbf{n}}_o \times \mathbf{E}^s = \mathbf{0} \quad \text{at } \Gamma_o. \quad (2.114.d)$$

2.2.3 The scattering system: Weak forms and weak solutions

Before we begin investigating the function spaces pertinent to the scattering problem, it should be noticed that, since the components of the tensor $\bar{\bar{\mathbf{A}}}$ and the incident field $\mathbf{E}^{inc}$ assume complex values, our solution $\mathbf{E}^s$ is going to be complex. So the function spaces describing the quantities should also allow complex-valued functions. In what regards the spaces introduced so far, it suffices to consider their complex versions, e.g., in (2.7) where one reads:

$$L^p(\Omega) := \{v: \Omega \rightarrow \mathbb{R} \mid v \text{ is Lebesgue measurable and } \|v\|_{L^p(\Omega)} < \infty\}, \quad (2.115)$$

one must now read:

$$L^p(\Omega) := \{v: \Omega \rightarrow \mathbb{C} \mid v \text{ is Lebesgue measurable and } \|v\|_{L^p(\Omega)} < \infty\}, \quad (2.116)$$

and so on for the other spaces. In what regards inner products, as in (2.23), where one reads

$$(u, v)_{L^2(\Omega)} := \int_{\Omega} uv, \quad \forall u, v \in L^2(\Omega), \quad (2.117)$$

one must now read

$$(u, v)_{L^2(\Omega)} := \int_{\Omega} uv^*, \quad \forall u, v \in L^2(\Omega), \quad (2.118)$$

where v^* is the complex conjugate of v . In the course of the text, we shall occasionally indicate particular situations in which complex values must be taken into account.

2.2.3.1 The problem in classical form

The classical electrical field $\mathbf{E}^s$ belongs to $C^2(\Omega)^3 \cap C(\bar{\Omega})^3$, i.e., the second derivatives of each component are continuous throughout the domain Ω , and there should be no jumps when going from the interior of Ω to the boundary $\partial\Omega = \Gamma = \Gamma_o \cup \Gamma_1$. The pseudopressure p is treated as in the Navier-Stokes system (2.63), i.e., we assume that $p \in C^1(\Omega) \cap C(\bar{\Omega})$. In what regards the boundary conditions at the scatterer surface Γ_1 , we demand that the tangential components of the incident field be continuous, i.e., that $-\hat{\mathbf{n}}_1 \times \mathbf{E}^{inc} \in C(\Gamma_1)^3$. So the classical problem is written as

Find $(\mathbf{E}^s, p) \in C^2(\Omega)^3 \cap C(\bar{\Omega})^3 \times C^1(\Omega) \cap C(\bar{\Omega})$ such that

$$\nabla \cdot \bar{\bar{\Lambda}} \cdot \nabla \mathbf{E}^s + k_0^2 \mathbf{E}^s + \nabla p = \mathbf{0} \quad \text{in } \Omega, \quad (2.119.a)$$

$$\nabla \cdot \mathbf{E}^s = 0 \quad \text{in } \Omega, \quad (2.119.b)$$

$$\hat{\mathbf{n}}_1 \times \mathbf{E}^s = -\hat{\mathbf{n}}_1 \times \mathbf{E}^{inc} \quad \text{at } \Gamma_1, \quad (2.119.c)$$

$$\hat{\mathbf{n}}_o \times \mathbf{E}^s = \mathbf{0} \quad \text{at } \Gamma_o. \quad (2.119.d)$$

In order for (2.119.a) to be differentiable in the classical sense (pointwise), some requirements on the PML tensor are needed. Since the divergence operator acts on $\bar{\bar{\Lambda}} \cdot \nabla \mathbf{E}^s$, the resulting terms from this expression should be at least in $C^1(\Omega)$. If $\bar{\bar{\Lambda}} \cdot \nabla \mathbf{E}^s$ is expanded in terms of the components of $\mathbf{E}^s$, it becomes a sum of terms like $(\Lambda_x \partial E_x^s / \partial x) \hat{\mathbf{x}} \otimes \hat{\mathbf{x}} + (\Lambda_y \partial E_x^s / \partial y) \hat{\mathbf{y}} \otimes \hat{\mathbf{x}} + (\Lambda_z \partial E_x^s / \partial z) \hat{\mathbf{z}} \otimes \hat{\mathbf{x}} + \dots$. So each term as $\Lambda_x \partial E_x^s / \partial x$ should be in $C^1(\Omega)$. Because $E_x^s \in C^2(\Omega)$, then $\partial E_x^s / \partial x \in C^1(\Omega)$. One learns that the individual term $\Lambda_x \partial E_x^s / \partial x$ is in $C^1(\Omega)$ if Λ_x is also in $C^1(\Omega)$. The same analysis is extended to the other terms of the expansion, and one discovers that in order for all classical derivatives in (2.119.a) to be meaningful, one must require that the components of the PML tensor be in $C^1(\Omega)$.

2.2.3.2 Testing functions

Let us recall the space $C^\infty(\bar{\Omega})^3$ from (2.57) and introduce the subspace

$$\mathcal{D}_\tau(\Omega) := \{\mathbf{v} \in C^\infty(\bar{\Omega})^3 \mid \hat{\mathbf{n}} \times \mathbf{v}|_\Gamma = \mathbf{0}\}. \quad (2.120)$$

The space $\mathcal{D}_\tau(\Omega)$ comprises all functions from $C^\infty(\bar{\Omega})^3$ whose tangential components vanish (pointwise) at all points from the boundary $\Gamma = \Gamma_o \cup \Gamma_1$.

Since there are no sources in (2.119), there is no need to form the residuals. The first equation (2.119.a) is multiplied by an arbitrary testing function $\boldsymbol{\varphi}^* \in \mathcal{D}_\tau(\Omega)$, and (2.119.b) by another testing function $\varphi^* \in C_0^\infty(\Omega)$. After integration over the domain Ω and application of vector and tensor identities, we get

Find $(\mathbf{E}^s, p) \in C^2(\Omega)^3 \cap C(\bar{\Omega})^3 \times C^1(\Omega) \cap C(\bar{\Omega})$ such that

$$\begin{aligned} \int_{\Omega} (\bar{\bar{\mathbf{A}}} \cdot \nabla \mathbf{E}^s) : \nabla \boldsymbol{\varphi}^* - \int_{\Omega} k_0^2 \mathbf{E}^s \cdot \boldsymbol{\varphi}^* - \int_{\Omega} p \nabla \cdot \boldsymbol{\varphi}^* - \\ - \oint_{\Gamma} \left((\bar{\bar{\mathbf{A}}} \cdot \nabla \mathbf{E}^s) \cdot \hat{\mathbf{n}} - p \hat{\mathbf{n}} \right) \cdot \boldsymbol{\varphi}^* = 0, \quad \forall \boldsymbol{\varphi}^* \in \mathcal{D}_\tau(\Omega) \end{aligned} \quad (2.121.a)$$

$$\int_{\Omega} \varphi^* \nabla \cdot \mathbf{E}^s = 0, \quad \forall \varphi^* \in C_0^\infty(\Omega) \quad (2.121.b)$$

$$\hat{\mathbf{n}}_1 \times \mathbf{E}^s = -\hat{\mathbf{n}}_1 \times \mathbf{E}^{inc}, \quad \text{at } \Gamma_1 \quad (2.121.c)$$

$$\hat{\mathbf{n}}_o \times \mathbf{E}^s = \mathbf{0}, \quad \text{at } \Gamma_o. \quad (2.121.d)$$

By now, we are not concerned with the specific form assumed by the components of the PML tensor $\bar{\bar{\mathbf{A}}}$. More information about them will be introduced gradually, as dictated by necessity. At this point, it suffices to know two points. The first states that

$$\bar{\bar{\mathbf{A}}} = \bar{\mathbf{I}} \quad \text{in } \Omega \setminus \Omega_{PML}, \quad (2.122)$$

where Ω_{PML} is the region occupied by the PML, which is nothing more than a layer of thickness w_{PML} , usually small, measured from the outer surface Γ_o . Consequently, (2.122) holds in the bulk of the domain Ω , and particularly at the PEC scatterer surface Γ_1 . The second will be stated in the form of a conjecture.

Conjecture 2.1: Nullity of $\mathbf{E}^s$ at the outer surface Γ_o – All components of the scattered electric field $\mathbf{E}^s$ and its derivatives are zero at Γ_o .

Conjecture (2.1) above means that, if the PML works as it should, all components of $\mathbf{E}^s$ are attenuated in such a way that they are zero by the time they reach the outer surface Γ_o . The amplitude of $\mathbf{E}^s$ goes to zero, and $\mathbf{E}^s$ essentially disappears

(together with all its derivatives, of course) before reaching Γ_o . We have no formal proof for this hypothesis, hence it is stated in the form of a conjecture. Nevertheless, it is very reasonable and has been verified over and again in the experiments.

It can be observed that, since the functions from $\mathcal{D}_\tau(\Omega)$ do not have all their components equal to zero at the boundaries Γ_o and Γ_1 , the boundary integral in (2.121.a) does not automatically vanish as it happened for the Navier-Stokes system (2.65.a). Moreover, the pseudopressure p is devoid of a physical meaning here; in our formulation, it is just a Lagrange multiplier used to enforce the divergence-free condition. The model also does not state any boundary condition that p must satisfy. Therefore we just discard the boundary integral in which p figures, i.e., we make:

$$\oint_{\Gamma} \left((\bar{\mathbf{A}} \cdot \nabla \mathbf{E}^s) \cdot \hat{\mathbf{n}} - p \hat{\mathbf{n}} \right) \cdot \boldsymbol{\varphi}^* = 0, \quad \forall \boldsymbol{\varphi}^* \in \mathcal{D}_\tau(\Omega). \quad (2.123)$$

In order to study the consequences of (2.123), we first break it into two boundary integrals over Γ_o and Γ_1 . So for all $\boldsymbol{\varphi}^* \in \mathcal{D}_\tau(\Omega)$,

$$\int_{\Gamma_o} \left((\bar{\mathbf{A}} \cdot \nabla \mathbf{E}^s) \cdot \hat{\mathbf{n}}_o - p \hat{\mathbf{n}}_o \right) \cdot \boldsymbol{\varphi}^* + \int_{\Gamma_1} \left((\bar{\mathbf{A}} \cdot \nabla \mathbf{E}^s) \cdot \hat{\mathbf{n}}_1 - p \hat{\mathbf{n}}_1 \right) \cdot \boldsymbol{\varphi}^* = 0. \quad (2.124)$$

According to (2.120), if $\boldsymbol{\varphi}^* \in \mathcal{D}_\tau(\Omega)$, then $\hat{\mathbf{n}} \times \boldsymbol{\varphi}^*|_{\Gamma} = \mathbf{0}$, which means that $\hat{\mathbf{n}} \times \boldsymbol{\varphi}^*|_{\Gamma_o} = \mathbf{0}$ and $\hat{\mathbf{n}} \times \boldsymbol{\varphi}^*|_{\Gamma_1} = \mathbf{0}$, i.e., it has no tangential components along Γ_o and along Γ_1 . Let us form the subspace of $\mathcal{D}_\tau(\Omega)$ whose elements have all components equal to zero at Γ_1 , i.e., let

$$\mathcal{D}_\tau^1(\Omega) = \{ \mathbf{v} \in \mathcal{D}_\tau(\Omega) \mid \mathbf{v}|_{\Gamma_1} = \mathbf{0} \}. \quad (2.125)$$

Due to Conjecture 2.1, $\nabla \mathbf{E}^s = \mathbf{0}$ at Γ_o , so the first integral in (2.124) implies that

$$\int_{\Gamma_o} p \hat{\mathbf{n}}_o \cdot \boldsymbol{\varphi}^* = 0, \quad \forall \boldsymbol{\varphi}^* \in \mathcal{D}_\tau^1(\Omega). \quad (2.126)$$

Analogously, let us form the subspace of $\mathcal{D}_\tau(\Omega)$ whose elements have all components equal to zero at Γ_o :

$$\mathcal{D}_\tau^o(\Omega) = \{ \mathbf{v} \in \mathcal{D}_\tau(\Omega) \mid \mathbf{v}|_{\Gamma_o} = \mathbf{0} \}. \quad (2.127)$$

According to (2.122), the second integral in (2.124) implies that

$$\int_{\Gamma_1} \left(\frac{\partial \mathbf{E}^s}{\partial n} - p \hat{\mathbf{n}}_1 \right) \cdot \boldsymbol{\varphi}^* = 0, \quad \forall \boldsymbol{\varphi}^* \in \mathcal{D}_\tau^o(\Omega). \quad (2.128)$$

The consequences of discarding the boundary integral in (2.121.a) are: The pseudopressure p is in a certain sense equal to zero along the outer boundary Γ_o [according to (2.126)], and also in a certain sense related to the normal derivatives of $\mathbf{E}^s$

along the scatterer surface Γ_1 [according to (2.128)]. But the values assumed by p at the boundaries are immaterial to our analysis, and so we are safe to ignore the boundary integral in (2.121.a).

The problem (2.121) is therefore rewritten as

Find $(\mathbf{E}^s, p) \in C^2(\Omega)^3 \cap C(\bar{\Omega})^3 \times C^1(\Omega) \cap C(\bar{\Omega})$ such that

$$\int_{\Omega} (\bar{\mathbf{A}} \cdot \nabla \mathbf{E}^s) : \nabla \boldsymbol{\varphi}^* - \int_{\Omega} k_0^2 \mathbf{E}^s \cdot \boldsymbol{\varphi}^* - \int_{\Omega} p \nabla \cdot \boldsymbol{\varphi}^* = 0, \quad \forall \boldsymbol{\varphi}^* \in \mathcal{D}_{\tau}(\Omega) \quad (2.129.a)$$

$$\int_{\Omega} \boldsymbol{\varphi}^* \nabla \cdot \mathbf{E}^s = 0, \quad \forall \boldsymbol{\varphi}^* \in C_0^{\infty}(\Omega) \quad (2.129.b)$$

$$\hat{\mathbf{n}}_1 \times \mathbf{E}^s = -\hat{\mathbf{n}}_1 \times \mathbf{E}^{inc} \quad \text{at } \Gamma_1, \quad (2.129.c)$$

$$\hat{\mathbf{n}}_o \times \mathbf{E}^s = \mathbf{0} \quad \text{at } \Gamma_o. \quad (2.129.d)$$

The system (2.129) and the Navier-Stokes system (2.65) (after the removal of the boundary integral) show a remarkable symmetry involving the divergence terms in the first and second equations. This symmetry plays a key role in the mixed formulation, which will be the topic of Chapter 3. But now, let us concentrate on relaxing the function spaces associated with problem (2.129).

2.2.3.3 Relaxing the requirements

Let us write $\boldsymbol{\varphi}^* = [\varphi_x^*, \varphi_y^*, \varphi_z^*]^T$. The first integral in (2.129.a), when expanded, is a sum like

$$\begin{aligned} & \int_{\Omega} \Lambda_x \frac{\partial E_x^s}{\partial x} \frac{\partial \varphi_x^*}{\partial x} + \Lambda_y \frac{\partial E_x^s}{\partial y} \frac{\partial \varphi_x^*}{\partial y} + \Lambda_z \frac{\partial E_x^s}{\partial z} \frac{\partial \varphi_x^*}{\partial z} + \\ & \int_{\Omega} \Lambda_x \frac{\partial E_y^s}{\partial x} \frac{\partial \varphi_y^*}{\partial x} + \Lambda_y \frac{\partial E_y^s}{\partial y} \frac{\partial \varphi_y^*}{\partial y} + \Lambda_z \frac{\partial E_y^s}{\partial z} \frac{\partial \varphi_y^*}{\partial z} + \\ & \int_{\Omega} \Lambda_x \frac{\partial E_z^s}{\partial x} \frac{\partial \varphi_z^*}{\partial x} + \Lambda_y \frac{\partial E_z^s}{\partial y} \frac{\partial \varphi_z^*}{\partial y} + \Lambda_z \frac{\partial E_z^s}{\partial z} \frac{\partial \varphi_z^*}{\partial z}. \end{aligned} \quad (2.130)$$

The components of $\boldsymbol{\varphi}^*$ are not compactly supported – they are in $C^{\infty}(\bar{\Omega})$, not in $C_0^{\infty}(\Omega)$ – but they are still very smooth. In order to verify if (2.130) remains finite when the function space for $\mathbf{E}^s$ is modified, let us remember the basic triangle inequality for complex numbers

$$|a + b| \leq |a| + |b|, \quad \forall a, b \in \mathbb{C}, \quad (2.131)$$

which may be extended to a sum of terms as

$$\left| \sum_{i=1}^N a_i \right| \leq \sum_{i=1}^N |a_i|, \quad a_i \in \mathbb{C}. \quad (2.132)$$

According to (2.132), expression (2.130) is finite if the absolute value of each term is also finite. So let us concentrate on the first term from (2.130). It is true that

$$\left| \int_{\Omega} \Lambda_x \frac{\partial E_x^s}{\partial x} \frac{\partial \varphi_x^*}{\partial x} \right| \leq \int_{\Omega} \left| \Lambda_x \frac{\partial E_x^s}{\partial x} \frac{\partial \varphi_x^*}{\partial x} \right| \quad (2.133.a)$$

$$= \int_{\Omega} \left| \Lambda_x \frac{\partial E_x^s}{\partial x} \right| \left| \frac{\partial \varphi_x^*}{\partial x} \right| \quad (2.133.b)$$

$$\leq \max_{x \in \bar{\Omega}} \left| \frac{\partial \varphi_x^*}{\partial x} \right| \int_{\Omega} \left| \Lambda_x \frac{\partial E_x^s}{\partial x} \right|. \quad (2.133.c)$$

Since $\varphi_x^* \in C^\infty(\bar{\Omega})$, then $\partial \varphi_x^* / \partial x \in C^\infty(\bar{\Omega})$ also. It means that $\partial \varphi_x^* / \partial x$ is continuous and well defined up to the boundary, and therefore assumes a finite maximum value at some point $\mathbf{x}_M \in \bar{\Omega}$ [which justifies (2.133.c)]. From (2.133.c), we can conclude that the first term in (2.130) remains finite if

$$\int_{\Omega} \left| \Lambda_x \frac{\partial E_x^s}{\partial x} \right| < \infty, \quad (2.134)$$

which is the same as saying that

$$\Lambda_x \frac{\partial E_x^s}{\partial x} \in L^1(\Omega). \quad (2.135)$$

If we demand that $\Lambda_x \in L^\infty(\Omega)$ and $\partial E_x^s / \partial x \in L^1(\Omega)$, then the Hölder inequality (2.68) tells us that

$$\int_{\Omega} \left| \Lambda_x \frac{\partial E_x^s}{\partial x} \right| = \left\| \Lambda_x \frac{\partial E_x^s}{\partial x} \right\|_{L^1(\Omega)} \leq \|\Lambda_x\|_{L^\infty(\Omega)} \left\| \frac{\partial E_x^s}{\partial x} \right\|_{L^1(\Omega)} < \infty. \quad (2.136)$$

Since $L^2(\Omega)$ is ‘nicer’ than $L^1(\Omega)$, and since moreover $L^2(\Omega) \subset L^1(\Omega)$ according to (2.17), we demand that $\partial E_x^s / \partial x \in L^2(\Omega)$. Applying the same analysis to the other terms in (2.130), we conclude that the first integral in (2.129.a) remains bounded if the first derivatives of all components of $\mathbf{E}^s$ are in $L^2(\Omega)$ and the components of the PML tensor Λ_x , Λ_y and Λ_z are in $L^\infty(\Omega)$. These derivatives are no longer classical (pointwise) derivatives, but weak derivatives.

The second integral in (2.129.a) is a sum like

$$\int_{\Omega} E_x^s \varphi_x^* + E_y^s \varphi_y^* + E_z^s \varphi_z^*. \quad (2.137)$$

(The squared wavenumber k_0^2 has been removed from (2.137), as it is a constant term and has no bearing in the analysis.) Applying (2.132) to (2.137) and concentrating on the first term, it can be seen that

$$\left| \int_{\Omega} E_x^s \varphi_x^* \right| \leq \int_{\Omega} |E_x^s \varphi_x^*| \quad (2.138.a)$$

$$\leq \max_{x \in \bar{\Omega}} |\varphi_x^*| \int_{\Omega} |E_x^s|. \quad (2.138.b)$$

The justification for (2.138.b) comes from the fact that $\varphi_x^* \in C^\infty(\bar{\Omega})$, and therefore assumes a finite maximum value at some point in $\bar{\Omega}$. So if we demand that $E_x^s \in L^1(\Omega)$, then the first term in (2.137) is finite. When the same analysis is extended to the other terms, we conclude that the second integral in (2.129.a) is bounded if all components of $\mathbf{E}^s$ are in $L^1(\Omega)$. But for our purposes the space $L^2(\Omega)$ is better to work with than $L^1(\Omega)$, and then we demand that $\mathbf{E}^s \in L^2(\Omega)^3$.

Demanding that all components of $\mathbf{E}^s$ and all their derivatives be in $L^2(\Omega)$ is the same as demanding that $\mathbf{E}^s \in H^1(\Omega)^3$.

It is not difficult to see that the divergence from the third integral in (2.129.a), which is a term like $\nabla \cdot \boldsymbol{\varphi}^* = \partial \varphi_x^* / \partial x + \partial \varphi_y^* / \partial y + \partial \varphi_z^* / \partial z$, is in $C^\infty(\bar{\Omega})$ and therefore assumes a maximum at some point in $\bar{\Omega}$. So

$$\left| \int_{\Omega} p \nabla \cdot \boldsymbol{\varphi}^* \right| \leq \int_{\Omega} |p \nabla \cdot \boldsymbol{\varphi}^*| \quad (2.139.a)$$

$$\leq \max_{x \in \bar{\Omega}} |\nabla \cdot \boldsymbol{\varphi}^*| \int_{\Omega} |p|, \quad (2.139.b)$$

which allows us to conclude that if $p \in L^1(\Omega)$, then the third integral in (2.129.a) is bounded. As usual, we just require that $p \in L^2(\Omega) \subset L^1(\Omega)$.

In addition to not have to deal with boundary integrals, (2.123) brings one more advantage. Expressions (2.126) and (2.128), which are a consequence of (2.123), somehow ‘fix’ the values assumed by p at the boundary. So p is no longer determined up to a constant as in the Navier-Stokes system. By this, we mean that if $(\mathbf{E}^s, p)$ is a solution to (2.129), then $(\mathbf{E}^s, p + c)$ is not a solution for $c \neq 0$. In order to see it, we just replace p by $p + c$ in (2.129). The combination of all terms but one amounts to zero due to the fact that $(\mathbf{E}^s, p)$ is a solution. The only remaining term is

$$\int_{\Omega} c \nabla \cdot \boldsymbol{\varphi}^* = c \int_{\Omega} \nabla \cdot \boldsymbol{\varphi}^* \quad (2.140.a)$$

$$= c \oint_{\Omega} \boldsymbol{\varphi}^* \cdot \hat{\mathbf{n}}, \quad (2.140.b)$$

which is guaranteed to be zero only if $c = 0$, as the arbitrary testing function $\boldsymbol{\varphi}^*$ belongs to $\mathcal{D}_\tau(\Omega)$ in (2.120), a function space whose elements may possess nonzero normal components. Consequently, the space chosen for p is simply $L^2(\Omega)$, and not $L_0^2(\Omega)$ as in (2.71).

The only integral left to analyze is (2.129.b). Since $\mathbf{E}^s \in H^1(\Omega)^3$, $\partial E_x^s/\partial x$, $\partial E_y^s/\partial y$ and $\partial E_z^s/\partial z$ are in $L^2(\Omega)$. Due to (2.75) – just a consequence of Minkowski's inequality (2.73) – the divergence $\nabla \cdot \mathbf{E}^s$ is also in $L^2(\Omega)$. In order for (2.129.b) to make sense, $\nabla \cdot \mathbf{E}^s$ should be in $L_{loc}^1(\Omega)$, as the arbitrary test function $\varphi^* \in C_0^\infty(\Omega)$ is compactly contained in Ω . But it is of course true that $\nabla \cdot \mathbf{E}^s \in L_{loc}^1(\Omega)$, since $L^2(\Omega) \subset L_{loc}^1(\Omega)$ according to (2.17).

The analysis concerning the relaxed requirements on the function spaces needs to be completed by the study of the boundary conditions (2.129.c) and (2.129.d). However, in order to proceed, we need some more definitions that are peculiar to the electromagnetic problem. They will be explored next.

2.2.3.4 Interlude 1: The space $H(\mathbf{curl}; \Omega)$

In the sequel, the space $H(\mathbf{curl}, \Omega)$ will play an important role. It is defined as [Girault and Raviart, 1986], [Boyer and Fabrie, 2012], [Boffi *et al.*, 2013]:

$$H(\mathbf{curl}; \Omega) := \{\mathbf{v} \in L^2(\Omega)^3 \mid \nabla \times \mathbf{v} \in L^2(\Omega)^3\}. \quad (2.141)$$

The norm in this space is given by

$$\|\mathbf{v}\|_{H(\mathbf{curl}; \Omega)} := \left(\|\mathbf{v}\|_{L^2(\Omega)^3}^2 + \|\nabla \times \mathbf{v}\|_{L^2(\Omega)^3}^2 \right)^{\frac{1}{2}}. \quad (2.142)$$

It is not difficult to verify that

$$H^1(\Omega)^3 \subset H(\mathbf{curl}; \Omega). \quad (2.143)$$

Proof: Let $\mathbf{v} \in H^1(\Omega)^3$ be arbitrary. It is obvious that $\mathbf{v} \in L^2(\Omega)^3$, since $H^1(\Omega)^3 \subset L^2(\Omega)^3$. The curl of $\mathbf{v}$ is given by the traditional result:

$$\nabla \times \mathbf{v} = \left(\frac{\partial v_z}{\partial y} - \frac{\partial v_y}{\partial z} \right) \hat{\mathbf{x}} + \left(\frac{\partial v_x}{\partial z} - \frac{\partial v_z}{\partial x} \right) \hat{\mathbf{y}} + \left(\frac{\partial v_y}{\partial x} - \frac{\partial v_x}{\partial y} \right) \hat{\mathbf{z}}. \quad (2.144)$$

From (2.31),

$$\|\nabla \times \mathbf{v}\|_{L^2(\Omega)^3}^2 = \left\| \frac{\partial v_z}{\partial y} - \frac{\partial v_y}{\partial z} \right\|_{L^2(\Omega)}^2 + \left\| \frac{\partial v_x}{\partial z} - \frac{\partial v_z}{\partial x} \right\|_{L^2(\Omega)}^2 + \left\| \frac{\partial v_y}{\partial x} - \frac{\partial v_x}{\partial y} \right\|_{L^2(\Omega)}^2. \quad (2.145)$$

The Minkowski inequality (2.73) tells us that

$$\left\| \frac{\partial v_z}{\partial y} - \frac{\partial v_y}{\partial z} \right\|_{L^2(\Omega)} \leq \left\| \frac{\partial v_z}{\partial y} \right\|_{L^2(\Omega)} + \left\| \frac{\partial v_y}{\partial z} \right\|_{L^2(\Omega)}, \quad (2.146)$$

and likewise for the other two terms in (2.145). Consequently,

$$\left\| \frac{\partial v_z}{\partial y} - \frac{\partial v_y}{\partial z} \right\|_{L^2(\Omega)}^2 \leq \left(\left\| \frac{\partial v_z}{\partial y} \right\|_{L^2(\Omega)} + \left\| \frac{\partial v_y}{\partial z} \right\|_{L^2(\Omega)} \right)^2. \quad (2.147)$$

Since $\mathbf{v} \in H^1(\Omega)^3$, the derivatives of all components are in $L^2(\Omega)$, in particular $\partial v_z/\partial y$ and $\partial v_y/\partial z$. So the right side from (2.147) remains finite. The same conclusion is reached in what concerns the other two terms in (2.145). The final result is that $\nabla \times \mathbf{v}$ is square summable, i.e., that $\nabla \times \mathbf{v} \in L^2(\Omega)^3$.

We have just showed that $\mathbf{v} \in L^2(\Omega)^3$ and $\nabla \times \mathbf{v} \in L^2(\Omega)^3$. By (2.141), $\mathbf{v} \in H(\mathbf{curl}; \Omega)$. Since $\mathbf{v}$ is arbitrary, we have just concluded that $\forall \mathbf{v} \in H^1(\Omega)^3$ $\mathbf{v} \in H(\mathbf{curl}; \Omega)$, or, equivalently, that $H^1(\Omega)^3 \subset H(\mathbf{curl}; \Omega)$. ■

The space $H(\mathbf{curl}; \Omega)$ plays an important role in the functional analytic treatment of Maxwell's equations [Boffi *et al.*, 2013], [Monk, 2003]. It serves as the theoretical basis for the so called *edge elements*, which occupy a prominent position in the finite element analysis of vector problems in electromagnetism [Ern and Guermond, 2004], [Bossavit, 1997]. The functional analytic treatment of the Navier-Stokes problem, on the other hand, is largely based on the $H^1(\Omega)^3$ space [Girault and Raviart, 1986]. The space $H^1(\Omega)^3$ is amenable to discretization via nodal elements, and hence, via the nodal basis functions from the traditional meshfree methods. In this work, we consider a vector problem in electromagnetism and, instead of providing a formulation based on $H(\mathbf{curl}; \Omega)$, we provide another one based on $H^1(\Omega)^3$. In doing so, we are in a sense treating the electromagnetic wave scattering problem as a hydrodynamic problem.

An important subspace of $H(\mathbf{curl}; \Omega)$, denoted by $H_0(\mathbf{curl}; \Omega)$, is defined via density as [Boyer and Fabrie, 2012], [Monk, 2003]:

$$H_0(\mathbf{curl}; \Omega) := \overline{C_0^\infty(\Omega)^3}^{H(\mathbf{curl}; \Omega)}, \quad (2.148)$$

i.e., $H_0(\mathbf{curl}; \Omega)$ is the closure of $C_0^\infty(\Omega)^3$ in the norm (2.142). Another very useful density result will be stated as a theorem, whose proof can be found in [Boyer and Fabrie, 2012], [Girault and Raviart, 1986], [Monk, 2003]:

Theorem 2.5: The space $H(\mathbf{curl}; \Omega)$ – Suppose Ω is a bounded and Lipschitz domain in $\mathbb{R}^3$. Then it is true that

$$\overline{C^\infty(\bar{\Omega})}^{H(\mathbf{curl}; \Omega)} = H(\mathbf{curl}; \Omega), \quad (2.149)$$

where $C^\infty(\bar{\Omega})$ is defined in (2.53).

The space $H(\mathbf{curl}; \Omega)$ is also endowed with the notion of traces. Notwithstanding the fact that traces in $H(\mathbf{curl}; \Omega)$ are still an object of research [Boffi

et al., 2013], there are some basic notions concerning them that will be useful to us. They will be stated as a theorem here, and there are proofs in [Monk, 2003] and [Boyer and Fabrie, 2012]:

Theorem 2.6: Tangential traces – Let Ω be a bounded and Lipschitz domain in $\mathbb{R}^3$. Then there exists a linear operator $\gamma_t: H(\mathbf{curl}; \Omega) \rightarrow H^{-1/2}(\Gamma)^3$ such that:

1. If $\boldsymbol{\varphi} \in C^\infty(\overline{\Omega})^3$, then $\gamma_t \boldsymbol{\varphi} = \hat{\mathbf{n}} \times \boldsymbol{\varphi}|_\Gamma$.
2. There is a constant $C > 0$ such that $\|\gamma_t \mathbf{u}\|_{H^{-1/2}(\Gamma)^3} \leq C \|\mathbf{u}\|_{H(\mathbf{curl}; \Omega)}$ for all $\mathbf{u} \in H(\mathbf{curl}; \Omega)$.

Some clarification is in order. The space $H^{1/2}(\Gamma)$ is the range of the trace operator γ_0 , discussed in (2.54). This space has its dual $H^{-1/2}(\Gamma)$, which is the space of all functionals on $H^{1/2}(\Gamma)$ (i.e., bounded linear operators which act on the elements of $H^{1/2}(\Gamma)$ and return a real or complex a number). The original space $H^{1/2}(\Gamma)$ is a Hilbert space [Boffi *et al.*, 2013].

The interpretation of Theorem 2.6 is as follows: If $\boldsymbol{\varphi} \in C^\infty(\overline{\Omega})^3$, it is well-behaved enough so that $\gamma_t \boldsymbol{\varphi}$ is just the tangential component $\hat{\mathbf{n}} \times \boldsymbol{\varphi}$ at the boundary Γ . On the other hand, when the only information we possess about $\mathbf{u}$ is that it is in $H(\mathbf{curl}; \Omega)$ – be it in $C^\infty(\overline{\Omega})^3$ or not – one deduces the existence of a functional $\gamma_t \mathbf{u}$ whose norm is related to the norm of $\mathbf{u}$ via the second conclusion from Theorem 2.6. The quantity $\gamma_t \mathbf{u}$ is some kind of ‘tangential component’ of $\mathbf{u}$; hence the name *tangential trace*.

At this point, one may ask: What are the functions from $H(\mathbf{curl}; \Omega)$ which have zero tangential trace, i.e., what are those $\mathbf{u}$ in $H(\mathbf{curl}; \Omega)$ for which $\gamma_t \mathbf{u} = \mathbf{0}$? The answer is given by [Monk, 2003], [Girault and Raviart, 1986], and [Boyer and Fabrie, 2012]:

$$\text{Ker } \gamma_t = H_0(\mathbf{curl}; \Omega), \quad (2.150)$$

i.e., the kernel of γ_t is exactly the space $H_0(\mathbf{curl}; \Omega)$ [defined via density in (2.148)].

The trace operator γ_t is not surjective onto $H^{-1/2}(\Gamma)^3$, i.e., there are elements in $H^{-1/2}(\Gamma)^3$ which are not traces of elements from $H(\mathbf{curl}; \Omega)$. Symbolically, it means that

$$\gamma_t(H(\mathbf{curl}; \Omega)) = Y(\Gamma) \subset H^{-1/2}(\Gamma)^3, \quad (2.151)$$

i.e., that the range of the trace operator γ_t is a subspace of $H^{-1/2}(\Gamma)^3$, denoted by $Y(\Gamma)$. A proper characterization of $Y(\Gamma)$ falls outside the scope of this thesis, but the following results will be useful for us later. For Lipschitz domains, the space $Y(\Gamma)$ is given by [Monk, 2003]:

$$Y(\Gamma) = \{\mathbf{f} \in \mathbf{H}_t^{-1/2}(\Gamma) \mid \nabla_\Gamma \cdot \mathbf{f} \in H^{-1/2}(\Gamma)\}, \quad (2.152)$$

where $\nabla_\Gamma \cdot$ is the surface divergence, defined for any $\mathbf{v} \in H(\mathbf{curl}; \Omega)$ as

$$\nabla_\Gamma \cdot (\hat{\mathbf{n}} \times \mathbf{v}) = -\hat{\mathbf{n}} \cdot (\nabla \times \mathbf{v}) \quad \text{in } H^{-1/2}(\Gamma). \quad (2.153)$$

The space $\mathbf{H}_t^{-1/2}(\Gamma)$ is defined as

$$\mathbf{H}_t^{-1/2}(\Gamma) = \{\mathbf{u} \in H^{-1/2}(\Gamma)^3 \mid \mathbf{u} \cdot \hat{\mathbf{n}} = 0 \text{ a.e. on } \Gamma\}, \quad (2.154)$$

where ‘a.e.’ means ‘almost everywhere’, and is a technicality from measure theory [Tao, 2011]. More details about $Y(\Gamma)$ can be found in [Monk, 2003] and [Boffi et al., 2013].

We are now at a position to state the new ‘relaxed’ requirements on the non-homogeneous boundary conditions (2.129.c) and (2.129.d). In the analysis of the classical solution at Subsection 2.2.3.1, we had originally demanded that $-\hat{\mathbf{n}}_1 \times \mathbf{E}^{inc} \in C(\Gamma_1)^3$. We concluded from Subsection 2.2.3.2 that the scattered electric field $\mathbf{E}^s$ should now be in $H^1(\Omega)^3$; which implies that $\mathbf{E}^s \in H(\mathbf{curl}; \Omega)$, by (2.143). In this new setting, $\hat{\mathbf{n}} \times \mathbf{E}^s$ is no longer defined pointwise at Γ . So we must therefore resort to the notion of tangential traces from Theorem 2.6 and demand that

$$\boldsymbol{\gamma}_t \mathbf{E}^s = \begin{cases} \mathbf{0}, & \text{at } \Gamma_o \\ -\hat{\mathbf{n}}_1 \times \mathbf{E}^{inc}, & \text{at } \Gamma_1 \end{cases} \quad (2.155)$$

By this, we require that (2.155) should define a functional which is in the range of the operator $\boldsymbol{\gamma}_t$, or equivalently, that (2.155) be an element from $Y(\Gamma)$.

The system (2.129) has been analyzed term by term, and we relaxed the requirements in order to enlarge the search space of solutions. The conclusions are summarized in Table 2.2 below.

TABLE 2.2 – REQUIREMENTS ON THE QUANTITIES IN THE SCATTERING SYSTEM

<i>Quantity</i>	<i>Classical solution</i>	<i>‘Relaxed’ solution</i>
$\mathbf{E}^s$	$C^2(\Omega)^3 \cap C(\bar{\Omega})^3$	$H^1(\Omega)^3$
p	$C^1(\Omega) \cap C(\bar{\Omega})$	$L^2(\Omega)$
$\Lambda_x, \Lambda_y, \Lambda_z$	$C^1(\Omega)$	$L^\infty(\Omega)$

The relaxed problem becomes:

Find $(\mathbf{E}^s, p) \in H^1(\Omega)^3 \times L^2(\Omega)$ such that

$$\int_{\Omega} (\bar{\mathbf{\Lambda}} \cdot \nabla \mathbf{E}^s) : \nabla \boldsymbol{\varphi}^* - \int_{\Omega} k_0^2 \mathbf{E}^s \cdot \boldsymbol{\varphi}^* - \int_{\Omega} p \nabla \cdot \boldsymbol{\varphi}^* = 0, \quad \forall \boldsymbol{\varphi}^* \in \mathcal{D}_\tau(\Omega) \quad (2.156.a)$$

$$\int_{\Omega} \varphi^* \nabla \cdot \mathbf{E}^s = 0, \quad \forall \varphi^* \in C_0^\infty(\Omega) \quad (2.156.b)$$

$$\boldsymbol{\gamma}_t \mathbf{E}^s = \begin{cases} \mathbf{0}, & \text{at } \Gamma_o \\ -\hat{\mathbf{n}}_1 \times \mathbf{E}^{inc}, & \text{at } \Gamma_1. \end{cases} \quad (2.156.c)$$

Of course, all derivatives in (2.156) are meaningful if they are understood in the weak sense (i.e., they are weak derivatives).

2.2.3.5 Lifting on the boundary data

We now assume that $-\hat{\mathbf{n}}_1 \times \mathbf{E}^{inc}$ is such that the boundary function

$$\mathbf{g} = \begin{cases} \mathbf{0}, & \text{at } \Gamma_o \\ -\hat{\mathbf{n}}_1 \times \mathbf{E}^{inc}, & \text{at } \Gamma_1. \end{cases} \quad (2.157)$$

defines a functional which belongs to $Y(\Gamma)$. (Rigorously speaking, a functional and a function are different objects. In this context, the function $\mathbf{g}$, when seen in isolation, is just a function. It may be discontinuous. On the other hand, when it operates on testing functions from $H^{1/2}(\Gamma)^3$, it defines a functional).

Since $\boldsymbol{\gamma}_t$ is surjective on $Y(\Gamma)$, there are elements from $H(\mathbf{curl}; \Omega)$ whose tangential trace is exactly the $\mathbf{g}$ from (2.157), among which it figures our solution $\mathbf{E}^s$. Let us pick up a particular element $\mathbf{u}^g$, different from $\mathbf{E}^s$. Such an element exists. Indeed, if $\boldsymbol{\gamma}_t \mathbf{E}^s = \mathbf{g}$, then $\boldsymbol{\gamma}_t(\mathbf{E}^s + \mathbf{v}) = \boldsymbol{\gamma}_t \mathbf{E}^s = \mathbf{g}$ for all $\mathbf{v} \in H_0(\mathbf{curl}; \Omega)$, because $\boldsymbol{\gamma}_t$ is linear and because $H_0(\mathbf{curl}; \Omega)$ is the nullspace of $\boldsymbol{\gamma}_t$, according to (2.150). We may choose, for example, the $\mathbf{u}^g$ that looks ‘easier’ to construct. (Here at the continuous level it suffices to know that such a particular $\mathbf{u}^g$ exists. On the other hand, at the numerical level, this particular $\mathbf{u}^g$ can be found in a remarkably easy way.) After it has been chosen, the function $\mathbf{u}^g$ is termed the *lifting* on the Dirichlet boundary condition (2.157).

However, there is a problem lurking behind our choice for $\mathbf{u}^g$. The tangential trace theorem says that if $\mathbf{g} \in Y(\Gamma)$, then we can find a particular $\mathbf{u}^g \in H(\mathbf{curl}; \Omega)$ such that $\boldsymbol{\gamma}_t \mathbf{u}^g = \mathbf{g}$. But we are working on $H^1(\Omega)^3$, which is a subspace of $H(\mathbf{curl}; \Omega)$, according to (2.143). Our nodal meshfree formulation is based on $H^1(\Omega)^3$, and we are looking for solutions $\mathbf{E}^s$ that are in $H^1(\Omega)^3$. The problem becomes evident when one makes the question: What if this $\mathbf{u}^g$ belongs to $H(\mathbf{curl}; \Omega)$, but not to $H^1(\Omega)^3$? In other words, the problem is that the trace theorem says that $\mathbf{u}^g$ is in $H(\mathbf{curl}; \Omega)$, and does not guarantee that $\mathbf{u}^g$ is in the more regular subspace $H^1(\Omega)^3 \subset H(\mathbf{curl}; \Omega)$.

We must find a remedy for this situation. Once we know our functional $\mathbf{g} \in Y(\Gamma)$, there are only two cases.

Case 1: We can find a lifting $\mathbf{u}^g$ in $H^1(\Omega)^3 \subset H(\mathbf{curl}; \Omega)$. In this case, nothing needs to be done. We have already found a function which is in $H^1(\Omega)^3$ and whose trace is $\mathbf{g}$, namely, $\mathbf{u}^g$.

Case 2: We cannot find a lifting $\mathbf{u}^g$ in $H^1(\Omega)^3$. So $\mathbf{u}^g$ is in $H(\mathbf{curl}; \Omega)$, but not in $H^1(\Omega)^3$. In this case, we may recall the density result (2.149) and conclude that

$$\exists \{\boldsymbol{\varphi}_n\}_{n=1}^\infty \subset C^\infty(\bar{\Omega})^3 \quad \|\mathbf{u}^g - \boldsymbol{\varphi}_n\|_{H(\mathbf{curl}; \Omega)} \rightarrow 0, \quad (2.158)$$

i.e., there is a sequence of elements in $C^\infty(\bar{\Omega})^3$ which converges to $\mathbf{u}^g$ in the $H(\mathbf{curl}; \Omega)$ norm. According to the first conclusion from Theorem 2.6, for all $n \in \mathbb{N}$, $\boldsymbol{\gamma}_t \boldsymbol{\varphi}_n$ is just $\hat{\mathbf{n}} \times \boldsymbol{\varphi}_n|_\Gamma$, and this quantity defines a functional in $Y(\Gamma)$. Since of course $\boldsymbol{\varphi}_n \in H(\mathbf{curl}; \Omega)$, which is a linear space, the second conclusion from Theorem 2.6 allows us to write

$$\forall n \in \mathbb{N} \quad \|\boldsymbol{\gamma}_t(\mathbf{u}^g - \boldsymbol{\varphi}_n)\|_{H^{-1/2}(\Gamma)^3} \leq C \|\mathbf{u}^g - \boldsymbol{\varphi}_n\|_{H(\mathbf{curl}; \Omega)}. \quad (2.159)$$

The trace operator is also linear, so (2.159) is modified into

$$\forall n \in \mathbb{N} \quad \|\boldsymbol{\gamma}_t \mathbf{u}^g - \boldsymbol{\gamma}_t \boldsymbol{\varphi}_n\|_{H^{-1/2}(\Gamma)^3} \leq C \|\mathbf{u}^g - \boldsymbol{\varphi}_n\|_{H(\mathbf{curl}; \Omega)}. \quad (2.160)$$

As $n \rightarrow \infty$, (2.158) says that the right side of (2.160) approaches zero. So we conclude that

$$\lim_{n \rightarrow \infty} \|\mathbf{g} - \boldsymbol{\gamma}_t \boldsymbol{\varphi}_n\|_{H^{-1/2}(\Gamma)^3} = 0. \quad (2.161)$$

Expression (2.161) means that, given any number $\varepsilon > 0$, no matter how small, one can find an element $\boldsymbol{\varphi}_N$ such that $\|\mathbf{g} - \boldsymbol{\gamma}_t \boldsymbol{\varphi}_N\|_{H^{-1/2}(\Gamma)^3} < \varepsilon$. Given that $\|\cdot\|_{H^{-1/2}(\Gamma)^3}$ is a norm, it obviously satisfies the norm axioms [Conway, 1994], [Kreyszig, 1989], [Rynne and Youngson, 2007], one of which states that if the norm of an element is zero, then this element is zero. (The specific form assumed by the aforementioned norm does not interest us at this moment.) Since the norm of the difference $\mathbf{g} - \boldsymbol{\gamma}_t \boldsymbol{\varphi}_n$ tends to zero, so the difference itself tends to zero, i.e., $\boldsymbol{\gamma}_t \boldsymbol{\varphi}_n$ gets in a sense arbitrarily close to $\mathbf{g}$.

Now pick up an $\varepsilon > 0$ extravagantly small. There is an $\boldsymbol{\varphi}_N \in C^\infty(\bar{\Omega})^3$ such that $\|\mathbf{g} - \boldsymbol{\gamma}_t \boldsymbol{\varphi}_N\|_{H^{-1/2}(\Gamma)^3} < \varepsilon$. It is not difficult to see that $C^\infty(\bar{\Omega})^3 \subset H^1(\Omega)^3$, as the elements from $C^\infty(\bar{\Omega})^3$ and their derivatives are all continuous and well-behaved up to the boundary, and therefore square summable over Ω . So we have managed to find an element from $H^1(\Omega)^3$ whose trace is infinitely close to $\mathbf{g}$, namely, $\boldsymbol{\varphi}_N$.

To summarize: When Case 1 happens, we can find an element from $H^1(\Omega)^3$ whose trace is exactly $\mathbf{g}$, and when Case 2 happens, we can find an element from $H^1(\Omega)^3$ whose trace is arbitrarily close to $\mathbf{g}$.

This point is a delicate feature in the theory we are constructing, and we assume situations in which Case 1 always happens. In our future research, we will look for

restrictions on the domain Ω and on the admissible functions $\mathbf{g}$ such that we can find a lifting $\mathbf{u}^g$ which is guaranteed to be in $H^1(\Omega)^3$.

Despite the fact that just being able to find an element $\boldsymbol{\varphi}_N$ whose trace is very close to $\mathbf{g}$ does not seem a very relevant issue at the numerical level (where we can simply make an approximation and assume that $\boldsymbol{\gamma}_t \boldsymbol{\varphi}_N = \mathbf{g}$, which could at most produce a small error), at the continuous level there may be consequences which are more difficult to assess. So from now on, we shall deal with Case 1 only.

We write the scattered electric field as

$$\mathbf{E}^s = \mathbf{e}^0 + \mathbf{u}^g. \quad (2.162)$$

When applying $\boldsymbol{\gamma}_t$ to both sides of (2.162), we get that $\boldsymbol{\gamma}_t \mathbf{E}^s = \boldsymbol{\gamma}_t \mathbf{e}^0 + \boldsymbol{\gamma}_t \mathbf{u}^g$. Since $\boldsymbol{\gamma}_t \mathbf{u}^g = \boldsymbol{\gamma}_t \mathbf{E}^s = \mathbf{g}$, we conclude that $\boldsymbol{\gamma}_t \mathbf{e}^0 = \mathbf{0}$. Moreover, since $\mathbf{E}^s$ and $\mathbf{u}^g$ are in $H^1(\Omega)^3$, then $\mathbf{e}^0$ is in $H^1(\Omega)^3$ also. Let us introduce the space:

$$\mathbb{V}_\tau(\Omega) = \{\mathbf{v} \in H^1(\Omega)^3 \mid \boldsymbol{\gamma}_t \mathbf{v} = \mathbf{0}\}, \quad (2.163)$$

which is just a more formal way of representing the space

$$\{\mathbf{v} \in H^1(\Omega)^3 \mid (\hat{\mathbf{n}} \times \mathbf{v})|_\Gamma = \mathbf{0} \text{ in the sense of traces}\}. \quad (2.164)$$

It is clear that $\mathbf{e}^0 \in \mathbb{V}_\tau(\Omega)$. We substitute (2.162) in (2.156) and write a new problem in which $\mathbf{e}^0$ is the new unknown:

Find $(\mathbf{e}^0, p) \in \mathbb{V}_\tau(\Omega) \times L^2(\Omega)$ such that

$$\begin{aligned} \int_\Omega (\bar{\Lambda} \cdot \nabla \mathbf{e}^0) : \nabla \boldsymbol{\varphi}^* + \int_\Omega (\bar{\Lambda} \cdot \nabla \mathbf{u}^g) : \nabla \boldsymbol{\varphi}^* - \int_\Omega k_0^2 \mathbf{e}^0 \cdot \boldsymbol{\varphi}^* - \int_\Omega k_0^2 \mathbf{u}^g \cdot \boldsymbol{\varphi}^* - \\ \int_\Omega p \nabla \cdot \boldsymbol{\varphi}^* = 0, \quad \forall \boldsymbol{\varphi} \in \mathcal{D}_\tau(\Omega) \end{aligned} \quad (2.165.a)$$

$$\int_\Omega \boldsymbol{\varphi}^* \nabla \cdot \mathbf{e}^0 + \int_\Omega \boldsymbol{\varphi}^* \nabla \cdot \mathbf{u}^g = 0, \quad \forall \boldsymbol{\varphi} \in \mathcal{C}_0^\infty(\Omega). \quad (2.165.b)$$

The nonhomogeneous Dirichlet boundary condition $\mathbf{g}$ in (2.157) has been embedded into a suitable lifting function $\mathbf{u}^g$, so that now the new unknown $\mathbf{e}^0$ must be sought in the space (2.163), whose elements have zero tangential components along the boundary Γ .

2.2.3.6 The G map

Expressions (2.165.a) and (2.165.b) can be summed together, which allows us to rewrite the problem as:

Find $(\mathbf{e}^0, p) \in \mathbb{V}_\tau(\Omega) \times L^2(\Omega)$ such that

$$\begin{aligned} & \int_{\Omega} (\bar{\Lambda} \cdot \nabla \mathbf{e}^0) : \nabla \boldsymbol{\varphi}^* + \int_{\Omega} (\bar{\Lambda} \cdot \nabla \mathbf{u}^g) : \nabla \boldsymbol{\varphi}^* - \int_{\Omega} k_0^2 \mathbf{e}^0 \cdot \boldsymbol{\varphi}^* - \int_{\Omega} k_0^2 \mathbf{u}^g \cdot \boldsymbol{\varphi}^* - \\ & \int_{\Omega} p \nabla \cdot \boldsymbol{\varphi}^* + \int_{\Omega} \boldsymbol{\varphi}^* \nabla \cdot \mathbf{e}^0 + \int_{\Omega} \boldsymbol{\varphi}^* \nabla \cdot \mathbf{u}^g = 0, \quad \forall \boldsymbol{\varphi} \in \mathcal{D}_\tau(\Omega) \quad \forall \boldsymbol{\varphi} \in C_0^\infty(\Omega) \end{aligned} \quad (2.166)$$

We introduce the map

$$G : H^1(\Omega)^3 \times L^2(\Omega) \times H^1(\Omega)^3 \times L^2(\Omega) \rightarrow \mathbb{C} \quad (2.167)$$

defined by

$$\begin{aligned} G(\mathbf{v}_1, q_1, \mathbf{v}_2, q_2) &= \int_{\Omega} (\bar{\Lambda} \cdot \nabla \mathbf{v}_1) : \nabla \mathbf{v}_2^* + \int_{\Omega} (\bar{\Lambda} \cdot \nabla \mathbf{u}^g) : \nabla \mathbf{v}_2^* - \int_{\Omega} k_0^2 \mathbf{v}_1 \cdot \mathbf{v}_2^* \\ &- \int_{\Omega} k_0^2 \mathbf{u}^g \cdot \mathbf{v}_2^* - \int_{\Omega} q_1 \nabla \cdot \mathbf{v}_2^* + \int_{\Omega} q_2^* \nabla \cdot \mathbf{v}_1 + \int_{\Omega} q_2^* \nabla \cdot \mathbf{u}^g, \end{aligned} \quad (2.168)$$

where $\mathbf{u}^g \in H^1(\Omega)^3$ is known from the previous section. Since $\mathbb{V}_\tau(\Omega)$ and $\mathcal{D}_\tau(\Omega)$ are subsets of $H^1(\Omega)^3$, and $C_0^\infty(\Omega)$ is a subset of $L^2(\Omega)$, problem (2.166) can be reset as

Find $(\mathbf{e}^0, p) \in \mathbb{V}_\tau(\Omega) \times L^2(\Omega)$ such that

$$G(\mathbf{e}^0, p, \boldsymbol{\varphi}, \varphi) = 0, \quad \forall \boldsymbol{\varphi} \in \mathcal{D}_\tau(\Omega) \quad \forall \varphi \in C_0^\infty(\Omega). \quad (2.169)$$

According to (2.168), the G map is linear in $\mathbf{v}_2$ and q_2 . We must now investigate if G is also continuous with respect to the two last arguments. Let us concentrate on the first two terms from (2.168), which share the same form

$$\int_{\Omega} (\bar{\Lambda} \cdot \nabla \mathbf{v}) : \nabla \mathbf{w}^*, \quad (2.170)$$

where $\mathbf{v}$ and $\mathbf{w}$ are elements from $H^1(\Omega)^3$. When expanded, (2.170) reveals its form as

$$\begin{aligned} & \int_{\Omega} \Lambda_x \frac{\partial v_x}{\partial x} \frac{\partial w_x^*}{\partial x} + \Lambda_y \frac{\partial v_x}{\partial y} \frac{\partial w_x^*}{\partial y} + \Lambda_z \frac{\partial v_x}{\partial z} \frac{\partial w_x^*}{\partial z} + \\ & \int_{\Omega} \Lambda_x \frac{\partial v_y}{\partial x} \frac{\partial w_y^*}{\partial x} + \Lambda_y \frac{\partial v_y}{\partial y} \frac{\partial w_y^*}{\partial y} + \Lambda_z \frac{\partial v_y}{\partial z} \frac{\partial w_y^*}{\partial z} + \\ & \int_{\Omega} \Lambda_x \frac{\partial v_z}{\partial x} \frac{\partial w_z^*}{\partial x} + \Lambda_y \frac{\partial v_z}{\partial y} \frac{\partial w_z^*}{\partial y} + \Lambda_z \frac{\partial v_z}{\partial z} \frac{\partial w_z^*}{\partial z}. \end{aligned} \quad (2.171)$$

From (2.132), we learn that

$$\left| \int_{\Omega} (\bar{\Lambda} \cdot \nabla \mathbf{v}) : \nabla \mathbf{w}^* \right| \leq \left| \int_{\Omega} \Lambda_x \frac{\partial v_x}{\partial x} \frac{\partial w_x^*}{\partial x} \right| + \left| \int_{\Omega} \Lambda_y \frac{\partial v_x}{\partial y} \frac{\partial w_x^*}{\partial y} \right| + \dots \quad (2.172.a)$$

$$\leq \int_{\Omega} \left| \Lambda_x \frac{\partial v_x}{\partial x} \frac{\partial w_x^*}{\partial x} \right| + \int_{\Omega} \left| \Lambda_y \frac{\partial v_x}{\partial y} \frac{\partial w_x^*}{\partial y} \right| + \dots \quad (2.172.b)$$

Since $\mathbf{v}$ and $\mathbf{w}$ are in $H^1(\Omega)^3$, $\partial v_x/\partial x$ and $\partial w_x^*/\partial x$ are in $L^2(\Omega)$. (And likewise for the other terms). Consequently, the product of these quantities is in $L^1(\Omega)$, due to the Hölder inequality (2.68) for $p = 2$, i.e.,

$$\left\| \frac{\partial v_x}{\partial x} \frac{\partial w_x^*}{\partial x} \right\|_{L^1(\Omega)} \leq \left\| \frac{\partial v_x}{\partial x} \right\|_{L^2(\Omega)} \left\| \frac{\partial w_x^*}{\partial x} \right\|_{L^2(\Omega)}. \quad (2.173)$$

Since $\Lambda_x \in L^\infty(\Omega)$, we apply the Hölder inequality again and verify that

$$\int_{\Omega} \left| \Lambda_x \frac{\partial v_x}{\partial x} \frac{\partial w_x^*}{\partial x} \right| \leq \|\Lambda_x\|_{L^\infty(\Omega)} \left\| \frac{\partial v_x}{\partial x} \frac{\partial w_x^*}{\partial x} \right\|_{L^1(\Omega)}. \quad (2.174)$$

Expressions (2.173) and (2.174) together say that

$$\int_{\Omega} \left| \Lambda_x \frac{\partial v_x}{\partial x} \frac{\partial w_x^*}{\partial x} \right| \leq \|\Lambda_x\|_{L^\infty(\Omega)} \left\| \frac{\partial v_x}{\partial x} \right\|_{L^2(\Omega)} \left\| \frac{\partial w_x^*}{\partial x} \right\|_{L^2(\Omega)} \quad (2.175.a)$$

$$\leq \Lambda_M \left\| \frac{\partial v_x}{\partial x} \right\|_{L^2(\Omega)} \left\| \frac{\partial w_x^*}{\partial x} \right\|_{L^2(\Omega)}, \quad (2.175.b)$$

where

$$\Lambda_M = \max \left\{ \|\Lambda_x\|_{L^\infty(\Omega)}, \|\Lambda_y\|_{L^\infty(\Omega)}, \|\Lambda_z\|_{L^\infty(\Omega)} \right\}. \quad (2.176)$$

Similar conclusions are valid for all the other terms from (2.171). Inequality (2.172) is modified into

$$\left| \int_{\Omega} (\bar{\mathbf{\Lambda}} \cdot \nabla \mathbf{v}) : \nabla \mathbf{w}^* \right| \leq \quad (2.177)$$

$$\begin{aligned} & \Lambda_M \left(\left\| \frac{\partial v_x}{\partial x} \right\|_{L^2(\Omega)} \left\| \frac{\partial w_x^*}{\partial x} \right\|_{L^2(\Omega)} + \left\| \frac{\partial v_x}{\partial y} \right\|_{L^2(\Omega)} \left\| \frac{\partial w_x^*}{\partial y} \right\|_{L^2(\Omega)} + \left\| \frac{\partial v_x}{\partial z} \right\|_{L^2(\Omega)} \left\| \frac{\partial w_x^*}{\partial z} \right\|_{L^2(\Omega)} + \right. \\ & \left. \left\| \frac{\partial v_y}{\partial x} \right\|_{L^2(\Omega)} \left\| \frac{\partial w_y^*}{\partial x} \right\|_{L^2(\Omega)} + \left\| \frac{\partial v_y}{\partial y} \right\|_{L^2(\Omega)} \left\| \frac{\partial w_y^*}{\partial y} \right\|_{L^2(\Omega)} + \left\| \frac{\partial v_y}{\partial z} \right\|_{L^2(\Omega)} \left\| \frac{\partial w_y^*}{\partial z} \right\|_{L^2(\Omega)} + \right. \\ & \left. \left\| \frac{\partial v_z}{\partial x} \right\|_{L^2(\Omega)} \left\| \frac{\partial w_z^*}{\partial x} \right\|_{L^2(\Omega)} + \left\| \frac{\partial v_z}{\partial y} \right\|_{L^2(\Omega)} \left\| \frac{\partial w_z^*}{\partial y} \right\|_{L^2(\Omega)} + \left\| \frac{\partial v_z}{\partial z} \right\|_{L^2(\Omega)} \left\| \frac{\partial w_z^*}{\partial z} \right\|_{L^2(\Omega)} \right) \end{aligned}$$

If the define the norm in the complex $L^2(\Omega)$ space (2.118) as in (2.24), then the complex conjugate may be removed from all components of $\mathbf{w}$ in (2.177). According to (2.34), it is true that

$$|\mathbf{v}|_{H^1(\Omega)^3}^2 = \|\nabla v_x\|_{L^2(\Omega)^3}^2 + \|\nabla v_y\|_{L^2(\Omega)^3}^2 + \|\nabla v_z\|_{L^2(\Omega)^3}^2 = \quad (2.178.a)$$

$$\begin{aligned} & \left\| \frac{\partial v_x}{\partial x} \right\|_{L^2(\Omega)}^2 + \left\| \frac{\partial v_x}{\partial y} \right\|_{L^2(\Omega)}^2 + \left\| \frac{\partial v_x}{\partial z} \right\|_{L^2(\Omega)}^2 + \\ & \left\| \frac{\partial v_y}{\partial x} \right\|_{L^2(\Omega)}^2 + \left\| \frac{\partial v_y}{\partial y} \right\|_{L^2(\Omega)}^2 + \left\| \frac{\partial v_y}{\partial z} \right\|_{L^2(\Omega)}^2 + \\ & \left\| \frac{\partial v_z}{\partial x} \right\|_{L^2(\Omega)}^2 + \left\| \frac{\partial v_z}{\partial y} \right\|_{L^2(\Omega)}^2 + \left\| \frac{\partial v_z}{\partial z} \right\|_{L^2(\Omega)}^2, \end{aligned}$$

whereas a similar result holds for $|\mathbf{w}|_{H^1(\Omega)^3}^2$. Introduce now two 9×1 vectors given by

$$\mathbf{F} = \left[\left\| \frac{\partial v_x}{\partial x} \right\|_{L^2(\Omega)}, \left\| \frac{\partial v_x}{\partial y} \right\|_{L^2(\Omega)}, \left\| \frac{\partial v_x}{\partial z} \right\|_{L^2(\Omega)}, \dots, \left\| \frac{\partial v_z}{\partial y} \right\|_{L^2(\Omega)}, \left\| \frac{\partial v_z}{\partial z} \right\|_{L^2(\Omega)} \right]^T, \quad (2.178.b)$$

$$\mathbf{G} = \left[\left\| \frac{\partial w_x}{\partial x} \right\|_{L^2(\Omega)}, \left\| \frac{\partial w_x}{\partial y} \right\|_{L^2(\Omega)}, \left\| \frac{\partial w_x}{\partial z} \right\|_{L^2(\Omega)}, \dots, \left\| \frac{\partial w_z}{\partial y} \right\|_{L^2(\Omega)}, \left\| \frac{\partial w_z}{\partial z} \right\|_{L^2(\Omega)} \right]^T. \quad (2.178.c)$$

In this way, (2.177) can be rewritten as

$$\left| \int_{\Omega} (\bar{\mathbf{A}} \cdot \nabla \mathbf{v}) : \nabla \mathbf{w}^* \right| \leq \Lambda_M \mathbf{F} \cdot \mathbf{G}. \quad (2.179.a)$$

The Cauchy-Schwarz inequality for vectors tell us that $|\mathbf{F} \cdot \mathbf{G}| \leq (\mathbf{F} \cdot \mathbf{F})^{1/2} (\mathbf{G} \cdot \mathbf{G})^{1/2}$. Also, from (2.178.a) and (2.178.b) we can see that

$$|\mathbf{v}|_{H^1(\Omega)^3}^2 = \mathbf{F} \cdot \mathbf{F}. \quad (2.179.b)$$

Analogously, it is true that

$$|\mathbf{w}|_{H^1(\Omega)^3}^2 = \mathbf{G} \cdot \mathbf{G}. \quad (2.179.c)$$

Back to (2.179.a),

$$\left| \int_{\Omega} (\bar{\mathbf{A}} \cdot \nabla \mathbf{v}) : \nabla \mathbf{w}^* \right| \leq \Lambda_M \mathbf{F} \cdot \mathbf{G} \leq \Lambda_M |\mathbf{F} \cdot \mathbf{G}| \leq (\mathbf{F} \cdot \mathbf{F})^{1/2} (\mathbf{G} \cdot \mathbf{G})^{1/2}. \quad (2.179.d)$$

Inserting (2.179.b) and (2.179.c) in (2.179.d) allows us to conclude that

$$\left| \int_{\Omega} (\bar{\mathbf{A}} \cdot \nabla \mathbf{v}) : \nabla \mathbf{w}^* \right| \leq \Lambda_M |\mathbf{v}|_{H^1(\Omega)^3} |\mathbf{w}|_{H^1(\Omega)^3} \quad \forall \mathbf{v}, \mathbf{w} \in H^1(\Omega)^3, \quad (2.180)$$

which is related to (2.85.a). By (2.42), the seminorms in (2.180) can be replaced by norms, and thus we get the final result we need:

$$\left| \int_{\Omega} (\bar{\Lambda} \cdot \nabla v) : \nabla w^* \right| \leq \Lambda_M \|v\|_{H^1(\Omega)^3} \|w\|_{H^1(\Omega)^3} \quad \forall v, w \in H^1(\Omega)^3, \quad (2.181)$$

We can now inquire about the continuity of (2.168) in what regards its last two arguments. The inequalities (2.86.b), (2.94) and (2.181) when applied to (2.168) reveals that

$$|G(v_1, q_1, v_2, q_2)| \leq (\Lambda_M \|v_1\|_{H^1(\Omega)^3} + \Lambda_M \|u^g\|_{H^1(\Omega)^3} + k_0^2 \|v_1\|_{L^2(\Omega)^3} + k_0^2 \|u^g\|_{L^2(\Omega)^3} + \|q_1\|_{L^2(\Omega)}) \|v_2\|_{H^1(\Omega)^3} + (\|v_1\|_{H^1(\Omega)^3} + \|u^g\|_{H^1(\Omega)^3}) \|q_2\|_{L^2(\Omega)} \quad (2.182)$$

The continuity of the G map with respect to v_2 and q_2 is now evident.

2.2.3.7 Enlarging the space of testing functions

Our problem (2.165) can be given a new interpretation in terms of the G map as in (2.169). This expression says that, if we insert the solution (e^0, p) in the first two arguments, the map G assumes the value 0 whenever we consider arbitrary elements from $\mathcal{D}_\tau(\Omega)$ and $C_0^\infty(\Omega)$ as the last two arguments, respectively. Suppose we happen find other spaces $\mathcal{X} \supset \mathcal{D}_\tau(\Omega)$ and $\mathcal{Y} \supset C_0^\infty(\Omega)$ such that $G(e^0, p, \psi, \xi) = 0$ for all $\psi \in \mathcal{X}$ and for all $\xi \in \mathcal{Y}$. Since the (2.169) is just the traditional problem (2.166) written in a different form, it means that functions from these new spaces $\mathcal{X}$ and $\mathcal{Y}$ qualify as testing functions as well.

Elements of $\mathcal{D}_\tau(\Omega)$ may be particularly difficult to build, so we are better off if we find another ‘enlarged’ space $\mathcal{X}$ which contains $\mathcal{D}_\tau(\Omega)$ as a subspace and also allows less regular functions [which may be easier to construct than the infinitely differentiable elements from $\mathcal{D}_\tau(\Omega)$]. The same reasoning applies to $C_0^\infty(\Omega)$.

Fortunately, such spaces exist: They are $\mathcal{X} = \mathbb{V}_\tau(\Omega)$ and $\mathcal{Y} = L^2(\Omega)$. In order to proceed with the demonstration that such spaces qualify as testing spaces, we need the following density result which is stated in [Monk, 2003] (with a different notation, though):

$$\overline{\mathcal{D}_\tau(\Omega)}^{H^1(\Omega)^3} = \mathbb{V}_\tau(\Omega) \quad (2.183)$$

i.e., $\mathcal{D}_\tau(\Omega)$ is dense in $\mathbb{V}_\tau(\Omega)$ with respect to the $\|\cdot\|_{H^1(\Omega)^3}$ norm (2.38).

Let (e^0, p) be the solution to problem (2.169). We need to prove that

$$G(e^0, p, v, q) = 0, \quad \forall v \in \mathbb{V}_\tau(\Omega) \text{ and } \forall q \in L^2(\Omega) \quad (2.184)$$

Proof: Let $v \in \mathbb{V}_\tau(\Omega)$ and $q \in L^2(\Omega)$ be arbitrary. According to the density results from (2.183) and (2.49), respectively,

$$\exists \{\varphi_n\}_{n=1}^\infty \subset \mathcal{D}_\tau(\Omega) \quad \|v - \varphi_n\|_{H^1(\Omega)^3} \rightarrow 0 \quad (2.185.a)$$

$$\exists \{\varphi_n\}_{n=1}^\infty \subset C_0^\infty(\Omega) \quad \|q - \varphi_n\|_{L^2(\Omega)} \rightarrow 0 \quad (2.185.b)$$

Since all elements from the sequence $\{\boldsymbol{\varphi}_n\}_{n=1}^\infty$ are in $\mathcal{D}_\tau(\Omega)$, and all elements from the sequence $\{\varphi_n\}_{n=1}^\infty$ are in $C_0^\infty(\Omega)$, we can employ them as testing functions in (2.169). Consequently,

$$\forall n \in \mathbb{N} \quad G(\mathbf{e}^0, p, \boldsymbol{\varphi}_n, \varphi_n) = 0. \quad (2.186)$$

The map G is linear in the last two arguments, so we write:

$$G(\mathbf{e}^0, p, \mathbf{v}, q) - G(\mathbf{e}^0, p, \boldsymbol{\varphi}_n, \varphi_n) = G(\mathbf{e}^0, p, \mathbf{v} - \boldsymbol{\varphi}_n, q - \varphi_n), \quad (2.187)$$

where (2.187) holds for all $n \in \mathbb{N}$. Of course,

$$|G(\mathbf{e}^0, p, \mathbf{v}, q) - G(\mathbf{e}^0, p, \boldsymbol{\varphi}_n, \varphi_n)| = |G(\mathbf{e}^0, p, \mathbf{v} - \boldsymbol{\varphi}_n, q - \varphi_n)|. \quad (2.188)$$

But since G is bounded with respect to the two last arguments, from (2.182) we get:

$$\begin{aligned} |G(\mathbf{e}^0, p, \mathbf{v} - \boldsymbol{\varphi}_n, q - \varphi_n)| &\leq (\Lambda_M \|\mathbf{e}^0\|_{H^1(\Omega)^3} + \Lambda_M \|\mathbf{u}^g\|_{H^1(\Omega)^3} + k_0^2 \|\mathbf{e}^0\|_{L^2(\Omega)^3} + \\ &\quad k_0^2 \|\mathbf{u}^g\|_{L^2(\Omega)^3} + \|p\|_{L^2(\Omega)}) \|\mathbf{v} - \boldsymbol{\varphi}_n\|_{H^1(\Omega)^3} + \\ &\quad (\|\mathbf{e}^0\|_{H^1(\Omega)^3} + \|\mathbf{u}^g\|_{H^1(\Omega)^3}) \|q - \varphi_n\|_{L^2(\Omega)}. \end{aligned} \quad (2.189)$$

We have already assumed that $\mathbf{e}^0 \in \mathbb{V}_\tau(\Omega) \subset H^1(\Omega)^3$, $\mathbf{u}^g \in H^1(\Omega)^3$ and $p \in L^2(\Omega)$. So all the norms within parentheses in (2.189) are finite; for the sake of clarity, let us rewrite it as

$$|G(\mathbf{e}^0, p, \mathbf{v} - \boldsymbol{\varphi}_n, q - \varphi_n)| \leq M_1 \|\mathbf{v} - \boldsymbol{\varphi}_n\|_{H^1(\Omega)^3} + M_2 \|q - \varphi_n\|_{L^2(\Omega)}, \quad (2.190)$$

where the constants M_1 and M_2 are finite and depend on $\mathbf{e}^0$, $\mathbf{u}^g$ and p .

By letting $n \rightarrow \infty$, we conclude from (2.185) that

$$|G(\mathbf{e}^0, p, \mathbf{v} - \boldsymbol{\varphi}_n, q - \varphi_n)| \rightarrow 0. \quad (2.191)$$

From (2.188) and (2.191),

$$|G(\mathbf{e}^0, p, \mathbf{v}, q) - G(\mathbf{e}^0, p, \boldsymbol{\varphi}_n, \varphi_n)| \rightarrow 0. \quad (2.192)$$

But $G(\mathbf{e}^0, p, \boldsymbol{\varphi}_n, \varphi_n) = 0$ for all n , according to (2.186). Expression (2.192) therefore is true only if $G(\mathbf{e}^0, p, \mathbf{v}, q) = 0$. So we are allowed to conclude that:

$$G(\mathbf{e}^0, p, \mathbf{v}, q) = 0 \quad (2.193)$$

Since $\mathbf{v}$ and q are arbitrary, we are able to see that indeed

$$G(\mathbf{e}^0, p, \mathbf{v}, q) = 0, \quad \forall \mathbf{v} \in \mathbb{V}_\tau(\Omega) \text{ and } \forall q \in L^2(\Omega). \quad (2.194)$$

■

The G map is zero when we consider the enlarged spaces $\mathbb{V}_\tau(\Omega)$ and $L^2(\Omega)$; problem (2.169) then assumes a new form:

$$\begin{aligned} & \text{Find } (\mathbf{e}^0, p) \in \mathbb{V}_\tau(\Omega) \times L^2(\Omega) \text{ such that} \\ & G(\mathbf{e}^0, p, \mathbf{v}, q) = 0, \quad \forall \mathbf{v} \in \mathbb{V}_\tau(\Omega) \text{ and } \forall q \in L^2(\Omega). \end{aligned} \quad (2.195)$$

When we consider the definition of the G map in (2.168), we get

$$\begin{aligned} & \text{Find } (\mathbf{e}^0, p) \in \mathbb{V}_\tau(\Omega) \times L^2(\Omega) \text{ such that} \\ & \int_{\Omega} (\bar{\mathbf{A}} \cdot \nabla \mathbf{e}^0) : \nabla \mathbf{v}^* + \int_{\Omega} (\bar{\mathbf{A}} \cdot \nabla \mathbf{u}^g) : \nabla \mathbf{v}^* - \int_{\Omega} k_0^2 \mathbf{e}^0 \cdot \mathbf{v}^* - \int_{\Omega} k_0^2 \mathbf{u}^g \cdot \mathbf{v}^* - \\ & \int_{\Omega} p \nabla \cdot \mathbf{v}^* + \int_{\Omega} q^* \nabla \cdot \mathbf{e}^0 + \int_{\Omega} q^* \nabla \cdot \mathbf{u}^g = 0, \quad \forall \mathbf{v} \in \mathbb{V}_\tau(\Omega) \quad \forall q \in L^2(\Omega) \end{aligned} \quad (2.196)$$

When we first make $q = 0$ and $\mathbf{v}$ arbitrary, and then make $\mathbf{v} = \mathbf{0}$ and q arbitrary, we are able to recover the scattering system (2.165)

$$\begin{aligned} & \text{Find } (\mathbf{e}^0, p) \in \mathbb{V}_\tau(\Omega) \times L^2(\Omega) \text{ such that} \\ & \int_{\Omega} (\bar{\mathbf{A}} \cdot \nabla \mathbf{e}^0) : \nabla \mathbf{v}^* + \int_{\Omega} (\bar{\mathbf{A}} \cdot \nabla \mathbf{u}^g) : \nabla \mathbf{v}^* - \int_{\Omega} k_0^2 \mathbf{e}^0 \cdot \mathbf{v}^* - \int_{\Omega} k_0^2 \mathbf{u}^g \cdot \mathbf{v}^* - \\ & \int_{\Omega} p \nabla \cdot \mathbf{v}^* = 0, \quad \forall \mathbf{v} \in \mathbb{V}_\tau(\Omega) \quad (2.197.a) \\ & \int_{\Omega} q^* \nabla \cdot \mathbf{e}^0 + \int_{\Omega} q^* \nabla \cdot \mathbf{u}^g = 0, \quad \forall q \in L^2(\Omega), \quad (2.197.b) \end{aligned}$$

but now with the testing functions in the enlarged spaces $\mathbb{V}_\tau(\Omega)$ and $L^2(\Omega)$.

2.2.3.8 Weak solutions

The system (2.197) is essentially the scattering problem in weak form. After we get $(\mathbf{e}^0, p)$ from (2.197), we add the known particular lifting function to $\mathbf{e}^0$ and finally get the scattered field $\mathbf{E}^s = \mathbf{e}^0 + \mathbf{u}^g$, according to (2.162). The pair $(\mathbf{E}^s, p)$ thus obtained is the *weak solution* associated with the original problem (2.114). Or, equivalently, (2.197) is the variational formulation of problem (2.114).

We have now finished the study of the variational formulation associated with the scattering problem. The right spaces for $\mathbf{E}^s$ and p have been identified; by ‘right’ we mean that they both agree with the theoretical development and are amenable to a discretization via nodal elements. In the next chapter, we will introduce the concept of *mixed formulations* and show that the scattering system (2.197) is indeed an example of such.

Chapter 3

Mixed formulations

In this chapter, we will introduce the notion of mixed formulation, on which rests the concept of mixed finite elements.

In the first section, the idea of mixed formulations will be presented in the abstract setting, i.e., in terms of bilinear forms acting on abstract spaces (whose nature is left unspecified).

The second section specializes the notion to the case of the stationary incompressible Navier-Stokes system. These results are traditional, and have been explored in the literature for a while. It is presented here as a means for clarifying what is going on, and at the same time it is the departure point for the analysis of our scattering system.

In the third chapter, we specialize the notion of mixed formulation to the scattering system. The problem at this point can be summarized as follows. The well-posedness of the mixed formulations depends, among other things, on a property of the bilinear forms called coercivity. But it is known that the bilinear forms associated with time-harmonic wave problems (and hence the scattering problem) are not coercive. In this scenario, well-posedness is proved through another way, called the Fredholm Alternative. This alternative has been used to assess the well-posedness of wave problems ‘in isolation’, i.e., when there is only one unknown involved (for example, in the propagation of scalar waves). But our scattering system depends on two unknowns: the electric field and the pseudopressure. Our work in this chapter is to find a way to embed the Fredholm Alternative within the traditional framework of mixed formulations.

3.1 Mixed formulations in abstract form

3.1.1 Mixed variational formulations

Let $\mathcal{X}$ and $\mathcal{Y}$ be two Hilbert spaces. We say that θ is a *sesquilinear form* if θ is a map

$$\theta: \mathcal{X} \times \mathcal{Y} \rightarrow \mathbb{K}, \quad (3.1)$$

which obeys the two properties below:

$$\theta(\alpha_1 x_1 + \alpha_2 x_2, y) = \alpha_1 \theta(x_1, y) + \alpha_2 \theta(x_2, y) \quad (3.2.a)$$

$$\theta(x, \beta_1 y_1 + \beta_2 y_2) = \beta_1^* \theta(x, y_1) + \beta_2^* \theta(x, y_2) \quad (3.2.b)$$

for any x, x_1, x_2 in $\mathcal{X}$, any y, y_1, y_2 in $\mathcal{Y}$ and any $\alpha_1, \alpha_2, \beta_1, \beta_2$ in $\mathbb{K}$. The field $\mathbb{K}$ is either $\mathbb{R}$ or $\mathbb{C}$. (Of course, the complex conjugation in (3.2.b) makes sense only if $\mathbb{K} = \mathbb{C}$.) A sesquilinear form θ is *bounded* or *continuous* if there is a positive constant C such that

$$|\theta(x, y)| \leq C \|x\|_{\mathcal{X}} \|y\|_{\mathcal{Y}}, \quad \forall x \in \mathcal{X} \quad \forall y \in \mathcal{Y}, \quad (3.3)$$

where $\|\cdot\|_{\mathcal{X}}$ and $\|\cdot\|_{\mathcal{Y}}$ are the norms in the spaces $\mathcal{X}$ and $\mathcal{Y}$, respectively.

Suppose $a: \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{K}$ and $b: \mathcal{X} \times \mathcal{Y} \rightarrow \mathbb{K}$ are two given continuous sesquilinear forms. Moreover, let f^* be an element from the dual space $\mathcal{X}^*$ i.e., f^* is a bounded and linear functional acting on the elements from $\mathcal{X}$. This is represented as $f^* \in \mathcal{X}^*$. In the same way, let $g^* \in \mathcal{Y}^*$. We say a problem is cast in a *mixed variational formulation* (simply mixed formulation, or mixed form) if it assumes the form:

Find $(u, p) \in \mathcal{X} \times \mathcal{Y}$ such that

$$a(u, x) + b(x, p) = \langle f^*, x \rangle_{\mathcal{X}^*, \mathcal{X}} \quad \forall x \in \mathcal{X} \quad (3.4)$$

$$b(u, y) = \langle g^*, y \rangle_{\mathcal{Y}^*, \mathcal{Y}} \quad \forall y \in \mathcal{Y}.$$

In (3.4) above, $\langle f^*, x \rangle_{\mathcal{X}^*, \mathcal{X}}$ is the duality pairing between the functional f^* and the particular element x , i.e., it is just the action of f^* on x [sometimes represented as $f^*(x)$]. The same applies to $\langle g^*, y \rangle_{\mathcal{Y}^*, \mathcal{Y}}$.

3.1.2 Well-posedness

After we get the variational expression for our problem and discover that it fits the mixed form (3.4), the next step is to inquire if this form leads to a well-posed problem, i.e., a problem whose solution exists, is unique and depends continuously on the data (or is bounded in some sense).

The theory which investigates the conditions under which the system (3.4) is well-posed was developed independently by I. Babuska and F. Brezzi, and achieved tremendous success over the years. At the most abstract level, it is a rephrasing of Banach's Closed Range and Open Mapping Theorems [Brezis, 2010], which are used as tools to investigate operator equations in functional analysis. History has it that Necas [Necas, 1962] developed a theoretical work in which these theorems were recast in terms of *inf-sup* conditions, and that Babuska and Brezzi did further work concerning these inf-sup conditions in connection with finite element methods. Information about this theory can be found in the classical book [Brezzi and Fortin, 1991], and also in [Boffi *et al.*, 2013], [Roberts and Thomas, 1991], [Ern and Guermond, 2004], [Quarteroni and Valli, 1994], [Brezzi and Bathe, 1990], [Chapelle and Bathe, 2011].

In this work, we will just state the final result, as the formal proof is quite intricate.

Theorem 3.1: Well-posedness of mixed formulations – Let $\mathcal{X}$ and $\mathcal{Y}$ be two Hilbert spaces, and let $a: \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{K}$ and $b: \mathcal{X} \times \mathcal{Y} \rightarrow \mathbb{K}$ be two continuous sesquilinear forms, i.e., there are positive constants α_a and α_b such that

$$|a(x, v)| \leq \alpha_a \|x\|_{\mathcal{X}} \|v\|_{\mathcal{X}}, \quad \forall x, v \in \mathcal{X} \quad (3.5.a)$$

$$|b(x, y)| \leq \alpha_b \|x\|_{\mathcal{X}} \|y\|_{\mathcal{Y}}, \quad \forall x \in \mathcal{X} \quad \forall y \in \mathcal{Y} \quad (3.5.b)$$

Moreover, let $\mathcal{X}^0$ be the kernel of the sesquilinear form b i.e.,

$$\mathcal{X}^0 = \text{Ker } b = \{x \in \mathcal{X} \mid b(x, y) = 0, \quad \forall y \in \mathcal{Y}\}. \quad (3.5.c)$$

Suppose the sesquilinear form a is coercive on $\mathcal{X}^0$, i.e., there is a positive constant β_a such that

$$|a(x, x)| \geq \beta_a \|x\|_{\mathcal{X}}^2, \quad \forall x \in \mathcal{X}^0, \quad (3.5.d)$$

and that the sesquilinear form b satisfies the inf-sup condition, which says that there exists a constant $\beta_b > 0$ such that

$$\inf_{y \in \mathcal{Y} \setminus \{0\}} \sup_{x \in \mathcal{X} \setminus \{0\}} \frac{|b(x, y)|}{\|x\|_{\mathcal{X}} \|y\|_{\mathcal{Y}}} \geq \beta_b, \quad (3.5.e)$$

when $\mathbb{K} = \mathbb{C}$, or

$$\inf_{y \in \mathcal{Y} \setminus \{0\}} \sup_{x \in \mathcal{X} \setminus \{0\}} \frac{b(x, y)}{\|x\|_{\mathcal{X}} \|y\|_{\mathcal{Y}}} \geq \beta_b, \quad (3.5.f)$$

when $\mathbb{K} = \mathbb{R}$. Then, for each $f^* \in \mathcal{X}^*$ and $g^* \in \mathcal{Y}^*$, there is a unique solution to the problem

Find $(u, p) \in \mathcal{X} \times \mathcal{Y}$ such that

$$a(u, x) + b(x, p) = \langle f^*, x \rangle_{\mathcal{X}^*, \mathcal{X}} \quad \forall x \in \mathcal{X} \quad (3.5.g)$$

$$b(u, y) = \langle g^*, y \rangle_{\mathcal{Y}^*, \mathcal{Y}} \quad \forall y \in \mathcal{Y}$$

Moreover, the following estimate holds:

$$\|u\|_{\mathcal{X}} + \|p\|_{\mathcal{Y}} \leq C(\alpha_a, \alpha_b, \beta_a, \beta_b)(\|f^*\|_{\mathcal{X}^*} + \|g^*\|_{\mathcal{Y}^*}), \quad (3.5.h)$$

i.e., the solution depends continuously on the data.

This result is central to our work. In order to show that a mixed formulation in a pair of Hilbert spaces is well-posed, one needs to verify the four hypotheses (3.5.a), (3.5.b), (3.5.d) and (3.5.e) [or (3.5.f)]. Given that the sesquilinear forms are usually continuous, one actually needs to concentrate on verifying (3.5.d) and (3.5.e) [or

(3.5.f)]. The conditions (3.5.e) and (3.5.f) are particularly important, since they establish some kind of compatibility criterion between the two Hilbert spaces under consideration. They are also called the *Babuska-Brezzi conditions*, due to the fathers of the theory. In (3.5.h), C is a constant whose values depend on the other constants appearing in hypotheses (3.5.a), (3.5.b), (3.5.d) and (3.5.e). Of course, C does not depend on either x or y .

3.2 Mixed formulation for the Navier-Stokes system

3.2.1 Continuity and coercivity must be checked

When the Navier-Stokes system (2.109) is rewritten so as to transfer all information about the excitation source $\mathbf{f}$ and the lifting function $\mathbf{u}^g$ to the right side, it assumes the form:

$$\text{Find } (\mathbf{u}^0, p) \in H_0^1(\Omega)^3 \times L_0^2(\Omega) \text{ such that} \quad (3.6)$$

$$\begin{aligned} & \int_{\Omega} \nu \nabla \mathbf{u}^0 : \nabla \mathbf{v} + \int_{\Omega} [(\mathbf{u}^0 \cdot \nabla) \mathbf{u}^0] \cdot \mathbf{v} + \int_{\Omega} [(\mathbf{u}^0 \cdot \nabla) \mathbf{u}^g] \cdot \mathbf{v} + \int_{\Omega} [(\mathbf{u}^g \cdot \nabla) \mathbf{u}^0] \cdot \mathbf{v} \\ & - \int_{\Omega} p \nabla \cdot \mathbf{v} = \int_{\Omega} \mathbf{f} \cdot \mathbf{v} - \int_{\Omega} \nu \nabla \mathbf{u}^g : \nabla \mathbf{v} - \int_{\Omega} [(\mathbf{u}^g \cdot \nabla) \mathbf{u}^g] \cdot \mathbf{v}, \quad \forall \mathbf{v} \in H_0^1(\Omega)^3 \\ & - \int_{\Omega} q \nabla \cdot \mathbf{u}^0 = \int_{\Omega} q \nabla \cdot \mathbf{u}^g, \quad \forall q \in L_0^2(\Omega). \end{aligned}$$

The Navier-Stokes system is clearly nonlinear, due to the convective term. Rigorously speaking, the best way to account for the nonlinearity is to insert a *trilinear* form a in (3.5.g), instead of a bilinear form [careful observation reveals that there are three ‘slots’ in each of the second, third and fourth integrals from (3.6)].

Since in the Navier-Stokes system the quantities are real, sesquilinear forms automatically become bilinear forms. In other words, for the Navier-Stokes system, $\mathbb{K} = \mathbb{R}$.

So let it be the trilinear form $a(\cdot; \cdot, \cdot): H^1(\Omega)^3 \times H^1(\Omega)^3 \times H^1(\Omega)^3 \rightarrow \mathbb{R}$ be defined as

$$\begin{aligned} a(\mathbf{u}^0; \mathbf{w}, \mathbf{v}) &= \int_{\Omega} \nu \nabla \mathbf{w} : \nabla \mathbf{v} + \int_{\Omega} [(\mathbf{u}^0 \cdot \nabla) \mathbf{w}] \cdot \mathbf{v} + \\ & \int_{\Omega} [(\mathbf{w} \cdot \nabla) \mathbf{u}^g] \cdot \mathbf{v} + \int_{\Omega} [(\mathbf{u}^g \cdot \nabla) \mathbf{w}] \cdot \mathbf{v} \end{aligned} \quad (3.7)$$

In (3.7), the unknown $\mathbf{u}^0$ is fixed in the first slot from the second integral, so there are only two ‘free’ slots in each integral, precisely those occupied by $\mathbf{w}$ and $\mathbf{v}$. One observes that (3.6) assumes the form (3.5.f), if we make the following identifications:

$$H_0^1(\Omega)^3 \rightarrow \mathcal{X} \quad (3.8.a)$$

$$H^{-1}(\Omega)^3 \rightarrow \mathcal{X}^* \quad (3.8.b)$$

$$L_0^2(\Omega) \rightarrow \mathcal{Y} \quad (3.8.c)$$

$$L_0^2(\Omega) \rightarrow \mathcal{Y}^* \quad (3.8.d)$$

$$\begin{aligned} \{\mathbf{w}, \mathbf{v}\} \rightarrow & \left(\int_{\Omega} \nu \nabla \mathbf{w} : \nabla \mathbf{v} + \int_{\Omega} [(\mathbf{u}^0 \cdot \nabla) \mathbf{w}] \cdot \mathbf{v} + \right. \\ & \left. \int_{\Omega} [(\mathbf{w} \cdot \nabla) \mathbf{u}^g] \cdot \mathbf{v} + \int_{\Omega} [(\mathbf{u}^g \cdot \nabla) \mathbf{w}] \cdot \mathbf{v} \right) \rightarrow a(\mathbf{u}^0; \cdot, \cdot) \end{aligned} \quad (3.8.e)$$

$$\{\mathbf{v}, p\} \rightarrow \left(- \int_{\Omega} p \nabla \cdot \mathbf{v} \right) \rightarrow b(\cdot, \cdot) \quad (3.8.f)$$

$$\left(\int_{\Omega} \mathbf{f} \cdot (\quad) - \int_{\Omega} \nu \nabla \mathbf{u}^g : \nabla (\quad) - \int_{\Omega} [(\mathbf{u}^g \cdot \nabla) \mathbf{u}^g] \cdot (\quad) \right) \rightarrow f^* \quad (3.8.g)$$

$$\left(\int_{\Omega} (\quad) \nabla \cdot \mathbf{u}^g \right) \rightarrow g^*. \quad (3.8.h)$$

In (3.8.b), the dual space of $H_0^1(\Omega)$ is traditionally represented as $H^{-1}(\Omega)$, instead of $H_0^1(\Omega)^*$. Since $L_0^2(\Omega)$ is a Hilbert space, its dual is $L_0^2(\Omega)$ itself. (Hilbert spaces and their duals may be identified with each other, via Riesz’s representation theorem [Conway, 1994], [Kreyszig, 1989], [Rynne and Youngson, 2007].) In (3.8.e), $\{\mathbf{w}, \mathbf{v}\}$ means ‘consider $\{\mathbf{w}, \mathbf{v}\}$ as unknowns to be inserted as the arguments for $a(\mathbf{u}^0; \cdot, \cdot)$ ’, whereas in (3.8.f) $\{\mathbf{v}, p\}$ means ‘consider $\{\mathbf{v}, p\}$ as unknowns to be inserted as the arguments for $b(\cdot, \cdot)$ ’. In (3.8.g) and (3.8.h), the empty parentheses are to be filled with elements from $H_0^1(\Omega)^3$ and $L_0^2(\Omega)$, respectively.

According to the identification (3.8), problem (3.6) can be rewritten as

$$\begin{aligned} & \text{Find } (\mathbf{u}^0, p) \in H_0^1(\Omega)^3 \times L_0^2(\Omega) \text{ such that} \\ & a(\mathbf{u}^0; \mathbf{u}^0, \mathbf{v}) + b(\mathbf{v}, p) = \langle f^*, \mathbf{v} \rangle_{H^{-1}(\Omega)^3, H_0^1(\Omega)^3} \quad \forall \mathbf{v} \in H_0^1(\Omega)^3 \\ & b(\mathbf{u}^0, q) = \langle g^*, q \rangle_{L_0^2(\Omega)^*, L_0^2(\Omega)} \quad \forall q \in L_0^2(\Omega). \end{aligned} \quad (3.9)$$

In order to apply Theorem 3.1 to (3.9), some observations are in order. We notice that a slight modification had to be done in order to make the identification (3.8) fit the model from Theorem 3.1, namely, that a trilinear form a should be used instead

of a bilinear form a . In a sense, this reflects the power of Theorem 3.1: It is advisable to always try to reduce a system to the form (3.5.f), in order to enjoy the conclusions it provides.

If the trilinear form a with its first argument fixed as $\mathbf{u}^0$ has the same properties as those of a bilinear form, i.e., if it satisfies (3.5.a) and (3.5.d), then it is shown that the conclusions of Theorem 3.1 are automatically transferred to the system (3.9) [Girault and Raviart, 1986].

However, since we will not be concerned with the solution of the Navier-Stokes system in this work, (its presentation being just a means to guide our reasoning in what concerns the scattering system), we will no longer dwell on these details.

According to [Girault and Raviart, 1986], the forms a from (3.8.e) and b from (3.8.f) are continuous. Moreover, still according to [Girault and Raviart, 1986], the following relation holds:

$$\begin{aligned} a(\mathbf{u}^0; \mathbf{v}, \mathbf{v}) &= \int_{\Omega} \nu \nabla \mathbf{v} : \nabla \mathbf{v} + \int_{\Omega} [(\mathbf{u}^0 \cdot \nabla) \mathbf{v}] \cdot \mathbf{v} + \\ &\quad \int_{\Omega} [(\mathbf{v} \cdot \nabla) \mathbf{u}^g] \cdot \mathbf{v} + \int_{\Omega} [(\mathbf{u}^g \cdot \nabla) \mathbf{v}] \cdot \mathbf{v} \\ &\geq \gamma |\mathbf{v}|_{H^1(\Omega)^3}^2 \quad \forall \mathbf{v} \in V, \end{aligned} \quad (3.10)$$

where

$$V = \{\mathbf{v} \in H_0^1(\Omega)^3 \mid \nabla \cdot \mathbf{v} = 0\}, \quad (3.11)$$

and γ is a positive constant. In (3.11), equality is understood in the L^2 sense, i.e.,

$$V = \left\{ \mathbf{v} \in H_0^1(\Omega)^3 \mid \int_{\Omega} q \nabla \cdot \mathbf{v} = 0, \quad \forall q \in L_0^2(\Omega) \right\}. \quad (3.12)$$

But V is precisely the kernel of the bilinear form b in (3.5.c) after the identification (3.8), i.e.,

$$\text{Ker } b = \left\{ \mathbf{v} \in H_0^1(\Omega)^3 \mid \int_{\Omega} q \nabla \cdot \mathbf{v} = 0, \quad \forall q \in L_0^2(\Omega) \right\}. \quad (3.13)$$

From (3.10) and (3.13),

$$a(\mathbf{u}^0; \mathbf{v}, \mathbf{v}) \geq \gamma |\mathbf{v}|_{H^1(\Omega)^3}^2, \quad \forall \mathbf{v} \in \text{Ker } b. \quad (3.14)$$

In order to show that the seminorm in the right of (3.14) can be replaced by a norm, we need the following result, stated as a theorem [Quarteroni and Valli, 1994], [Ern and Guermond, 2004], [Salsa, 2008]:

Theorem 3.2: Poincaré inequality – Let Ω be a bounded and connected open set of $\mathbb{R}^d$, together with its boundary $\Gamma = \partial\Omega$. Suppose that $S \subset \Gamma$ is a Lipschitz-continuous subset of non-zero measure. Then there is a constant $c_\Omega > 0$ such that

$$\|v\|_{L^2(\Omega)}^2 \leq c_\Omega \|\nabla v\|_{L^2(\Omega)}^2, \quad \forall v \in H_S^1(\Omega), \quad (3.15.a)$$

where $H_S^1(\Omega)$ is the space

$$H_S^1(\Omega) = \{v \in H^1(\Omega) \mid v|_S = 0 \text{ in the sense of traces}\}. \quad (3.15.b)$$

When S corresponds to the whole boundary Γ , it means that $H_S^1(\Omega) = H_\Gamma^1(\Omega)$. But according to (3.15.b), $H_\Gamma^1(\Omega) = \{v \in H^1(\Omega) \mid v|_\Gamma = 0 \text{ in the sense of traces}\}$, which is exactly the traditional space $H_0^1(\Omega)$ from (2.55).

Let now $\mathbf{v} \in H_0^1(\Omega)^3$. If we apply (3.15.a) to all components of $\mathbf{v}$, we get that

$$\|\mathbf{v}\|_{L^2(\Omega)^3}^2 \leq c_\Omega |\mathbf{v}|_{H^1(\Omega)^3}^2. \quad (3.16)$$

Inequality (3.16) implies that $|\mathbf{v}|_{H^1(\Omega)^3}^2 + \|\mathbf{v}\|_{L^2(\Omega)^3}^2 \leq (1 + c_\Omega) |\mathbf{v}|_{H^1(\Omega)^3}^2$, whose left side is precisely $\|\mathbf{v}\|_{H^1(\Omega)^3}^2$, according to (2.42). So we get

$$|\mathbf{v}|_{H^1(\Omega)^3}^2 \geq \frac{1}{(1 + c_\Omega)} \|\mathbf{v}\|_{H^1(\Omega)^3}^2. \quad (3.17)$$

From (3.14) and (3.17) we conclude that

$$a(\mathbf{u}^0; \mathbf{v}, \mathbf{v}) \geq \frac{\gamma}{(1 + c_\Omega)} \|\mathbf{v}\|_{H^1(\Omega)^3}^2, \quad \forall \mathbf{v} \in \text{Ker } b, \quad (3.18)$$

i.e., the form $a(\mathbf{u}^0; \cdot, \cdot)$ is coercive in the kernel of b , and therefore satisfies requirement (3.5.d).

The last step to be shown in order for all requirements from Theorem 3.1 to be satisfied is the inf-sup condition (3.5.e).

3.2.2 The inf-sup condition must be checked

Let us get back to the sesquilinear/bilinear forms in abstract Hilbert spaces $\mathcal{X}$ and $\mathcal{Y}$. Suppose we are given a continuous sesquilinear form $a: \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{K}$. Its action is such that

$$a(x_1, x_2) = \text{a scalar}, \quad \forall x_1, x_2 \in \mathcal{X}. \quad (3.19)$$

If we fix the first argument x_1 , then the map

$$a(x_1; \cdot): \mathcal{X} \rightarrow \mathbb{K}, \quad (3.20)$$

defines a functional on $\mathcal{X}$. It can be proved [Chapelle and Bathe, 2011] that (3.20) defines a bounded and linear functional on $\mathcal{X}$, an hence, an element from the dual space $\mathcal{X}^*$. Since this functional depends on the fixed choice for x_1 , it is represented as $Ax_1 \in \mathcal{X}^*$. We therefore write

$$a(x_1, x_2) =: \langle Ax_1, x_2 \rangle_{\mathcal{X}^*, \mathcal{X}}, \quad \forall x_1, x_2 \in \mathcal{X}. \quad (3.21)$$

The operator A in (3.21) is sometimes referred to as ‘induced by the sesquilinear form a ’. So A maps elements from $\mathcal{X}$ (e.g. x_1) into elements from $\mathcal{X}^*$ [e.g. $Ax_1 = a(x_1; \cdot)$], i.e.,

$$A: \mathcal{X} \rightarrow \mathcal{X}^* \quad (3.22)$$

Again, it can be shown that the operator A is linear is bounded in the operator norm [Chapelle and Bathe, 2011].

Similar conclusions are reached concerning a form b which operates on two distinct spaces. Suppose we are given a continuous sesquilinear form $b: \mathcal{X} \times \mathcal{Y} \rightarrow \mathbb{K}$. So

$$b(x, y) = \text{a scalar}, \quad \forall x \in \mathcal{X} \quad \forall y \in \mathcal{Y}. \quad (3.23)$$

If we fix the element $x \in \mathcal{X}$, then the map

$$b(x; \cdot): \mathcal{Y} \rightarrow \mathbb{K} \quad (3.24)$$

defines an element from the dual space $\mathcal{Y}^*$, which is represented as $Bx \in \mathcal{Y}^*$. The operator B is also ‘induced by the sesquilinear form b ’:

$$b(x, y) =: \langle Bx, y \rangle_{\mathcal{Y}^*, \mathcal{Y}}, \quad \forall x \in \mathcal{X} \quad \forall y \in \mathcal{Y}. \quad (3.25)$$

This operator B maps elements from $\mathcal{X}$ into elements from $\mathcal{Y}^*$, i.e.,

$$B: \mathcal{X} \rightarrow \mathcal{Y}^* \quad (3.26)$$

Moreover, B is bounded in the operator norm.

In (3.24), we could have fixed $y \in \mathcal{Y}$ instead. The map

$$b(\cdot; y): \mathcal{X} \rightarrow \mathbb{K} \quad (3.27)$$

then defines an element from $\mathcal{X}^*$, which is represented as $B^T y \in \mathcal{X}^*$. The operator B^T maps elements from $\mathcal{Y}$ into $\mathcal{X}^*$, i.e.,

$$B^T: \mathcal{Y} \rightarrow \mathcal{X}^*, \quad (3.28)$$

and its operation is characterized by

$$b(x, y) =: \langle x, B^T y \rangle_{\mathcal{X}, \mathcal{X}^*}, \quad \forall x \in \mathcal{X} \quad \forall y \in \mathcal{Y}. \quad (3.29)$$

The operators B and B^T are adjoints, and are induced by the same sesquilinear form b , as (3.25) and (3.29) reveals.

The problem in mixed form (3.4), which is repeated below for convenience,

$$\begin{aligned} & \text{Find } (u, p) \in \mathcal{X} \times \mathcal{Y} \text{ such that} \\ & a(u, x) + b(x, p) = \langle f^*, x \rangle_{\mathcal{X}^*, \mathcal{X}} \quad \forall x \in \mathcal{X} \\ & b(u, y) = \langle g^*, y \rangle_{\mathcal{Y}^*, \mathcal{Y}} \quad \forall y \in \mathcal{Y}, \end{aligned} \tag{3.30}$$

can be recast in terms of operators if (3.21), (3.25) and (3.29) are employed. We begin by rewriting (3.30) as

$$\begin{aligned} & \text{Find } (u, p) \in \mathcal{X} \times \mathcal{Y} \text{ such that} \\ & \langle Au, x \rangle_{\mathcal{X}^*, \mathcal{X}} + \langle x, B^T p \rangle_{\mathcal{X}, \mathcal{X}^*} = \langle f^*, x \rangle_{\mathcal{X}^*, \mathcal{X}} \quad \forall x \in \mathcal{X} \\ & \langle Bu, y \rangle_{\mathcal{Y}^*, \mathcal{Y}} = \langle g^*, y \rangle_{\mathcal{Y}^*, \mathcal{Y}} \quad \forall y \in \mathcal{Y}. \end{aligned} \tag{3.31}$$

In the first equation from (3.31), the duality pairing $\langle x, B^T p \rangle_{\mathcal{X}, \mathcal{X}^*}$ is obviously equal to $\langle B^T p, x \rangle_{\mathcal{X}^*, \mathcal{X}}$. Since all duality pairings are linear, the system (3.31) can be written as

$$\begin{aligned} & \text{Find } (u, p) \in \mathcal{X} \times \mathcal{Y} \text{ such that} \\ & Au + B^T p = f^*, \text{ in } \mathcal{X}^* \\ & Bu = g^*, \text{ in } \mathcal{Y}^* \end{aligned} \tag{3.32}$$

The system (3.32) is an operator equation, i.e., the equations represent relations valid within the dual spaces $\mathcal{X}^*$ (the first) and $\mathcal{Y}^*$ (the second). Since elements from dual spaces are characterized by their actions on the elements from the original spaces, by ‘testing’ the functionals from system (3.32) on arbitrary functions from $\mathcal{X}$ and $\mathcal{Y}$, one recovers system (3.31).

The goal of this section is to show that the inf-sup condition (3.5.e) holds for the bilinear form b from (3.8). But we need first a very important result concerning the inf-sup conditions, stated as a theorem [Girault and Raviart, 1986], [Quarteroni and Valli, 1994], [Gerbeau et al., 2006], [Boffi et al., 2013]:

Theorem 3.3: On the inf-sup condition – Suppose $\mathcal{X}$ and $\mathcal{Y}$ are Hilbert spaces. Let $b: \mathcal{X} \times \mathcal{Y} \rightarrow \mathbb{K}$ be a continuous sesquilinear form. Then assertions (i), (ii), and (iii) below are equivalent to each other

(i) When $\mathbb{K} = \mathbb{C}$, it holds the inf-sup condition, i.e., there is a positive constant β_b such that

$$\inf_{y \in \mathcal{Y} \setminus \{0\}} \sup_{x \in \mathcal{X} \setminus \{0\}} \frac{|b(x, y)|}{\|x\|_{\mathcal{X}} \|y\|_{\mathcal{Y}}} \geq \beta_b, \tag{3.33. a}$$

which may also be written in its equivalent form as

$$\forall y \in \mathcal{Y} \quad \exists x \in \mathcal{X} \setminus \{0\} \quad \text{such that} \quad |b(x, y)| \geq \beta_b \|x\|_x \|y\|_y. \quad (3.33.b)$$

Analogously, when $\mathbb{K} = \mathbb{R}$, it holds the inf-sup condition, i.e., there is a positive constant β_b such that

$$\inf_{y \in \mathcal{Y} \setminus \{0\}} \sup_{x \in \mathcal{X} \setminus \{0\}} \frac{b(x, y)}{\|x\|_x \|y\|_y} \geq \beta_b, \quad (3.33.c)$$

which may also be written as

$$\forall y \in \mathcal{Y} \quad \exists x \in \mathcal{X} \setminus \{0\} \quad \text{such that} \quad b(x, y) \geq \beta_b \|x\|_x \|y\|_y. \quad (3.33.d)$$

(ii) The operator $B^T: \mathcal{Y} \rightarrow \mathcal{X}^*$ is injective and has a closed range.

(iii) The operator $B: \mathcal{X} \rightarrow \mathcal{Y}^*$ is surjective.

Moreover, if (3.33.a) or (3.33.c) holds, then it can be shown that

(iv)

$$\sup_{x \in \mathcal{X} \setminus \{0\}} \frac{\langle x, B^T y \rangle_{\mathcal{X}, \mathcal{X}^*}}{\|x\|_x} \geq \beta_b \|y\|_y, \quad \forall y \in \mathcal{Y} \quad (3.33.e)$$

(v)

$$\sup_{y \in \mathcal{Y} \setminus \{0\}} \frac{\langle Bx, y \rangle_{\mathcal{Y}^*, \mathcal{Y}}}{\|y\|_y} \geq \beta_b \|x\|_x, \quad \forall x \in (\mathcal{X}^0)^\perp, \quad (3.33.f)$$

where the Hilbert space $\mathcal{X}$ is decomposed as $\mathcal{X} = \mathcal{X}^0 \oplus (\mathcal{X}^0)^\perp$.

In (3.33), $\mathcal{X} \setminus \{0\}$ is the space $\mathcal{X}$ with the zero element removed, and likewise for $\mathcal{Y} \setminus \{0\}$. The subspace $\mathcal{X}^0$ is just the kernel of the sesquilinear form b , defined in (3.5.c).

In order to show that (3.33a) [or (3.33.c)] holds, which may be very difficult, trying to prove one of the equivalent assertions (ii) or (iii) is a good strategy.

In what regards the Navier-Stokes system (for which $\mathbb{K} = \mathbb{R}$), the identification (3.8) reveals that the bilinear form b is characterized by

$$b(\mathbf{v}, q) = - \int_{\Omega} q \nabla \cdot \mathbf{v}, \quad (3.34)$$

for which the underlying spaces $\mathcal{X}$ and $\mathcal{Y}$ are $H_0^1(\Omega)^3$ and $L_0^2(\Omega)$, respectively. If we take an arbitrary $\mathbf{v} \in H_0^1(\Omega)^3$ and write, as in (3.25),

$$b(\mathbf{v}, q) = \langle B\mathbf{v}, q \rangle_{L_0^2(\Omega)^*, L_0^2(\Omega)} = - \int_{\Omega} q \nabla \cdot \mathbf{v}, \quad \forall q \in L_0^2(\Omega), \quad (3.35)$$

we see that $B\mathbf{v} = -\nabla \cdot \mathbf{v}$, as the integral is a representation of the duality pairing between $L_0^2(\Omega)$ and its dual. In the specific case of the Navier-Stokes system the operator B is the negative of the divergence operator, i.e., $B = -\nabla \cdot$. So $B: \mathcal{X} \rightarrow \mathcal{Y}^*$ from (3.26) becomes

$$-\nabla \cdot : H_0^1(\Omega)^3 \rightarrow L_0^2(\Omega)^*. \quad (3.36)$$

At this point it would be remarkable if one could just show that $-\nabla \cdot$ is surjective. If one succeeded in showing it, by Theorem 3.3 it is implied that the inf-sup condition (3.33.c) also holds. Before we proceed to verifying if such a proof exists or not, we need to clarify some points concerning the space $L_0^2(\Omega)$.

Observation 3.1: The space $L_0^2(\Omega)$ – Originally, for bounded domains in $\mathbb{R}^d$, the space $L_0^2(\Omega)$ is defined as $L^2(\Omega)/\mathbb{R}$, i.e., the spaces of classes of functions of $L^2(\Omega)$ which differ (a.e.) by a constant. (Rigorously speaking, an element of $L_0^2(\Omega)$ is a subset, not a single function). Let $L^2(\Omega)$ be divided into non-overlapping subsets, called classes. Each class (a subset) is formed by all elements from $L^2(\Omega)$ which differ from each other by a constant. For example, if $u \in L^2(\Omega)$ belongs to a class, then all other elements of the type $u + c$, $c \in \mathbb{R}$, belong to the same class.

When equipped with the inner product

$$(u, v)_{L_0^2(\Omega)} = \int_{\Omega} (u - u_{av})(v - v_{av}), \quad (3.37.a)$$

where the average (or mean) of any $u \in L^2(\Omega)$ is

$$u_{av} = \frac{1}{|\Omega|} \int_{\Omega} u \quad (3.37.b)$$

it can be proved that $L_0^2(\Omega)$ is a Hilbert space [Boyer and Fabrie, 2012]. Moreover, it can also be shown that $L_0^2(\Omega)$ is isomorphic to the closed subspace of $L^2(\Omega)$ whose functions have zero average. This means that, instead of working with subsets of functions (classes), we can work with individual functions by choosing a specific representative of each subset. This representative happens to be precisely those whose average is zero. So in a sense, $L_0^2(\Omega)$ can be identified with the subspace

$$\left\{ v \in L^2(\Omega) \mid \frac{1}{|\Omega|} \int_{\Omega} v = 0 \right\}, \quad (3.37.c)$$

already introduced in (2.71). By restricting attention only to those elements whose average is zero, the expression for the inner product in (3.37.a) becomes similar to the expression for the standard inner product in $L^2(\Omega)$.

Since under these circumstances $L_0^2(\Omega)$ is a Hilbert space by itself, it can be identified with its dual, i.e., $L_0^2(\Omega) = L_0^2(\Omega)^*$. More discussion about the structure of the $L_0^2(\Omega)$ space can be found in [Boyer and Fabrie, 2012].

With this new information, (3.36) is modified into

$$-\nabla \cdot : H_0^1(\Omega)^3 \rightarrow L_0^2(\Omega). \quad (3.37)$$

The main result of this section is: The divergence operator (3.37) is surjective. The result comes from a powerful theorem, due to Bogovskii [Bogovskii, 1980], [Boyer and Fabrie, 2012], [Galdi, 2011].

Theorem 3.4: Surjectivity of the divergence operator – *Let Ω be a connected, bounded and Lipschitz domain of $\mathbb{R}^d$. Then there exists a continuous linear operator Π from $L_0^2(\Omega)$ into $H_0^1(\Omega)^3$ such that, for all $q \in L_0^2(\Omega)$, the function $\mathbf{v} = \Pi(q)$ satisfies*

$$\nabla \cdot \mathbf{v} = q. \quad (3.38)$$

In order to show that Theorem 3.4 implies the surjectivity of the divergence, let us first state what it is meant by surjectivity. The operator $-\nabla \cdot$ in (3.37) is surjective if we can show that

$$\forall q \in L_0^2(\Omega) \quad \exists \mathbf{v} \in H_0^1(\Omega)^3 \quad -\nabla \cdot \mathbf{v} = q. \quad (3.39)$$

Indeed, by letting $q \in L_0^2(\Omega)$ be arbitrary, it is obviously true that its negative $-q$ also belongs to $L_0^2(\Omega)$. According to Theorem (3.4), there is an element $\mathbf{v} = \Pi(-q)$ from $H_0^1(\Omega)^3$ such that $\nabla \cdot \mathbf{v} = -q$. Of course, this last equation is equivalent to $-\nabla \cdot \mathbf{v} = q$. We have just showed that, for any q in $L_0^2(\Omega)$, we are able to find a $\mathbf{v}$ in $H_0^1(\Omega)^3$ such that $\nabla \cdot \mathbf{v} = -q$, which is nothing else than (3.39). Therefore, $-\nabla \cdot$ is surjective.

The surjectivity of $-\nabla \cdot$ being proved, according to Theorem 3.3, the following inf-sup condition holds:

$$\inf_{q \in L_0^2(\Omega)} \sup_{\mathbf{v} \in H_0^1(\Omega)^d} \frac{-\int_{\Omega} q \nabla \cdot \mathbf{v}}{\|\mathbf{v}\|_{H_0^1(\Omega)^d} \|q\|_{L_0^2(\Omega)}} \geq \beta_b > 0. \quad (3.40)$$

When dealing with real function spaces, as it is generally the case regarding the Navier-Stokes, expression (3.40) assumes the equivalent form

$$\inf_{q \in L_0^2(\Omega)} \sup_{\mathbf{v} \in H_0^1(\Omega)^d} \frac{\int_{\Omega} q \nabla \cdot \mathbf{v}}{\|\mathbf{v}\|_{H_0^1(\Omega)^d} \|q\|_{L_0^2(\Omega)}} \geq \beta_b > 0. \quad (3.41)$$

In order to see it, suppose (3.40) is true. According to Theorem 3.3, it is equivalent to

$$\forall q \in L_0^2(\Omega) \quad \exists \mathbf{v} \in H_0^1(\Omega)^d \setminus \{\mathbf{0}\} \quad \text{s.t.} \quad -\int_{\Omega} q \nabla \cdot \mathbf{v} \geq \beta_b \|\mathbf{v}\|_{H_0^1(\Omega)^d} \|q\|_{L_0^2(\Omega)}. \quad (3.42)$$

Let $q \in L_0^2(\Omega)$ be arbitrary. Then there is a $\mathbf{v}_1 \in H_0^1(\Omega)^d \setminus \{\mathbf{0}\}$ such that

$$-\int_{\Omega} q \nabla \cdot \mathbf{v}_1 \geq \beta_b \|\mathbf{v}_1\|_{H_0^1(\Omega)^d} \|q\|_{L_0^2(\Omega)}. \quad (3.43)$$

Introduce the element $\mathbf{v}_2 = -\mathbf{v}_1$. Of course, $\mathbf{v}_2 \in H_0^1(\Omega)^d \setminus \{\mathbf{0}\}$. Since $\mathbf{v}_1 = -\mathbf{v}_2$, substitute this into (3.43) in order to get

$$\int_{\Omega} q \nabla \cdot \mathbf{v}_2 \geq \beta_b \|\mathbf{v}_2\|_{H_0^1(\Omega)^d} \|q\|_{L_0^2(\Omega)}. \quad (3.44)$$

For our choice of q , we have just deduced the existence of an element $\mathbf{v}$ from $H_0^1(\Omega)^d \setminus \{\mathbf{0}\}$, namely, $\mathbf{v}_2$, such that

$$\int_{\Omega} q \nabla \cdot \mathbf{v} \geq \beta_b \|\mathbf{v}\|_{H_0^1(\Omega)^d} \|q\|_{L_0^2(\Omega)}. \quad (3.45)$$

In other words, for our particular choice of q , we showed that

$$\exists \mathbf{v} \in H_0^1(\Omega)^d \setminus \{\mathbf{0}\} \quad \text{s. t.} \quad \int_{\Omega} q \nabla \cdot \mathbf{v} \geq \beta_b \|\mathbf{v}\|_{H_0^1(\Omega)^d} \|q\|_{L_0^2(\Omega)}. \quad (3.46)$$

Since this $q \in L_0^2(\Omega)$ was arbitrary, we conclude that

$$\forall q \in L_0^2(\Omega) \quad \exists \mathbf{v} \in H_0^1(\Omega)^d \setminus \{\mathbf{0}\} \quad \text{s. t.} \quad \int_{\Omega} q \nabla \cdot \mathbf{v} \geq \beta_b \|\mathbf{v}\|_{H_0^1(\Omega)^d} \|q\|_{L_0^2(\Omega)}, \quad (3.47)$$

which, according to Theorem 3.3, is equivalent to (3.41). In order to prove the converse, by a similar reasoning, we begin with (3.41) and show that (3.40) holds.

When specializing the system (3.4) to the Navier-Stokes setting via the identification (3.8), one is able to show that all requirements from Theorem 3.1 are satisfied. In this way, it follows that the (weak) solution to the stationary incompressible Navier-Stokes system exists, is unique and depends continuously on the data.

Now that the long path connecting the original differential equations to the well-posedness of their variational formulations has been established, we will no longer make any reference to the Navier-Stokes system in the course of this thesis. It was a kind of ‘preparatory journey’, and it is time to devote all our attention to the scattering system. We are on our own now. But thanks to the acquired expertise, we do not expect great difficulties.

3.3 Mixed formulation for the scattering system

3.3.1 Determining the structure of the problem

When the scattering system (2.197) is rewritten in such a way that all known information from the lifting function $\mathbf{u}^g$ is moved to the right side, it takes the form:

Find $(\mathbf{e}^0, p) \in \mathbb{V}_\tau(\Omega) \times L^2(\Omega)$ such that

$$\begin{aligned} \int_{\Omega} (\bar{\mathbf{A}} \cdot \nabla \mathbf{e}^0) : \nabla \mathbf{v}^* - \int_{\Omega} k_0^2 \mathbf{e}^0 \cdot \mathbf{v}^* - \int_{\Omega} p \nabla \cdot \mathbf{v}^* = \\ - \int_{\Omega} (\bar{\mathbf{A}} \cdot \nabla \mathbf{u}^g) : \nabla \mathbf{v}^* + \int_{\Omega} k_0^2 \mathbf{u}^g \cdot \mathbf{v}^*, \quad \forall \mathbf{v} \in \mathbb{V}_\tau(\Omega) \end{aligned} \quad (3.48.a)$$

$$- \int_{\Omega} q^* \nabla \cdot \mathbf{e}^0 = \int_{\Omega} q^* \nabla \cdot \mathbf{u}^g, \quad \forall q \in L^2(\Omega), \quad (3.48.b)$$

The scattering system is linear, and $\mathbb{K} = \mathbb{C}$ i.e., the forms are going to assume complex values. If we make the identification:

$$\mathbb{V}_\tau(\Omega) \rightarrow \mathcal{X} \quad (3.49.a)$$

$$\mathbb{V}_\tau(\Omega)^* \rightarrow \mathcal{X}^* \quad (3.49.b)$$

$$L^2(\Omega) \rightarrow \mathcal{Y} \quad (3.49.c)$$

$$L^2(\Omega) \rightarrow \mathcal{Y}^* \quad (3.49.d)$$

$$\{\mathbf{w}, \mathbf{v}\} \rightarrow \left(\int_{\Omega} (\bar{\mathbf{A}} \cdot \nabla \mathbf{w}) : \nabla \mathbf{v}^* - \int_{\Omega} k_0^2 \mathbf{w} \cdot \mathbf{v}^* \right) \rightarrow a(\cdot, \cdot) \quad (3.49.e)$$

$$\{\mathbf{v}, p\} \rightarrow \left(- \int_{\Omega} p^* \nabla \cdot \mathbf{v} \right) \rightarrow b(\cdot, \cdot) \quad (3.49.f)$$

$$\left(- \int_{\Omega} (\bar{\mathbf{A}} \cdot \nabla \mathbf{u}^g) : \nabla (\cdot)^* + \int_{\Omega} k_0^2 \mathbf{u}^g \cdot (\cdot)^* \right) \rightarrow f^* \quad (3.49.g)$$

$$\left(\int_{\Omega} (\cdot)^* \nabla \cdot \mathbf{u}^g \right) \rightarrow g^*. \quad (3.49.h)$$

The basic space for the electric field is $\mathbb{V}_\tau(\Omega)$, defined in (2.163). It is a subspace of $H^1(\Omega)^3$, and when equipped with the inner product of the parental space $H^1(\Omega)^3$, it becomes a Hilbert space on its own. We show this in Chart 3.1 below.

Chart 3.1 – $\mathbb{V}_\tau(\Omega)$ is a Hilbert space.

In order to see it, we need first to show that $\mathbb{V}_\tau(\Omega)$ is closed in the $\|\cdot\|_{H^1(\Omega)^3}$ norm (which happens to be the norm induced by the inner product). Showing that $\mathbb{V}_\tau(\Omega)$ is closed amounts to showing that it contains all its limit points.

So let $\mathbf{v} \in H^1(\Omega)^3$ be an arbitrary limit point of $\mathbb{V}_\tau(\Omega)$. [We need to prove that $\mathbf{v} \in \mathbb{V}_\tau(\Omega)$.] It follows that

$$\exists \{\mathbf{v}_n\}_{n=1}^\infty \subset \mathbb{V}_\tau(\Omega) \quad \|\mathbf{v} - \mathbf{v}_n\|_{H^1(\Omega)^3} \rightarrow 0. \quad (3.50)$$

Since $\mathbf{v}$ and all elements from $\mathbb{V}_\tau(\Omega)$ are in $H^1(\Omega)^3$, they are also in $H(\mathbf{curl}; \Omega)$, as $H^1(\Omega)^3 \subset H(\mathbf{curl}; \Omega)$, from (2.143). We apply Theorem 2.6 to the difference $\mathbf{v} - \mathbf{v}_n$:

$$\forall n \in \mathbb{N} \quad \|\boldsymbol{\gamma}_t(\mathbf{v} - \mathbf{v}_n)\|_{H^{-1/2}(\Gamma)^3} \leq C \|\mathbf{v} - \mathbf{v}_n\|_{H(\mathbf{curl}; \Omega)}. \quad (3.51)$$

The tangential trace $\boldsymbol{\gamma}_t$ is linear, so $\boldsymbol{\gamma}_t(\mathbf{v} - \mathbf{v}_n) = \boldsymbol{\gamma}_t \mathbf{v} - \boldsymbol{\gamma}_t \mathbf{v}_n$. Expression (3.51) becomes

$$\forall n \in \mathbb{N} \quad \|\boldsymbol{\gamma}_t \mathbf{v} - \boldsymbol{\gamma}_t \mathbf{v}_n\|_{H^{-1/2}(\Gamma)^3} \leq C \|\mathbf{v} - \mathbf{v}_n\|_{H(\mathbf{curl}; \Omega)}. \quad (3.52)$$

For any sequence of vectors $\{\mathbf{w}_n\}_{n=1}^\infty$ in $H^1(\Omega)^3$, if $\|\mathbf{w}_n\|_{H^1(\Omega)^3} \rightarrow 0$, then $\|\mathbf{w}_n\|_{H(\mathbf{curl}; \Omega)} \rightarrow 0$. In order to see why, $\|\mathbf{w}_n\|_{H^1(\Omega)^3} \rightarrow 0$ means that all Cartesian components of $\mathbf{w}_n$ go to zero in the L^2 -norm, and also that all derivatives of all Cartesian components also go to zero in the L^2 -norm. In this way, since the components of $\nabla \times \mathbf{w}_n$ are combinations of the derivatives of the components of $\mathbf{w}_n$, then $\|\nabla \times \mathbf{w}_n\|_{L^2(\Omega)^3}$ goes to zero as well.

Since $\{\mathbf{v} - \mathbf{v}_n\}_{n=1}^\infty$ is a sequence in $H^1(\Omega)^3$, and due to (3.50), we can pass to the limit in (3.52) and get

$$\|\boldsymbol{\gamma}_t \mathbf{v} - \boldsymbol{\gamma}_t \mathbf{v}_n\|_{H^{-1/2}(\Gamma)^3} \rightarrow 0. \quad (3.53)$$

But for all $\mathbf{v}_n$ in $\mathbb{V}_\tau(\Omega)$, $\boldsymbol{\gamma}_t \mathbf{v}_n = \mathbf{0}$, due to the very definition of this space. Then,

$$\|\boldsymbol{\gamma}_t \mathbf{v}\|_{H^{-1/2}(\Gamma)^3} \rightarrow 0, \quad (3.54)$$

which is meaningful only if $\|\boldsymbol{\gamma}_t \mathbf{v}\|_{H^{-1/2}(\Gamma)^3} = 0$. By the norm axioms, we conclude that $\boldsymbol{\gamma}_t \mathbf{v} = \mathbf{0}$.

So $\mathbf{v} \in H^1(\Omega)^3$ and $\boldsymbol{\gamma}_t \mathbf{v} = \mathbf{0}$. Consequently, $\mathbf{v} \in \mathbb{V}_\tau(\Omega)$. Since $\mathbf{v}$ was an arbitrary limit point, it can be concluded that $\mathbb{V}_\tau(\Omega)$ is closed.

There is a theorem which lists the circumstances under which a subspace of a Hilbert space is a Hilbert space by itself [Kreyszig, 1989], [Conway, 1994]:

Theorem 3.5: Subspaces of Hilbert spaces – Let $\mathcal{H}$ be a Hilbert space and $\mathcal{Z} \subset \mathcal{H}$ be a subspace of it, i.e., the inner product on $\mathcal{Z}$ is just the inner product on $\mathcal{H}$ restricted to elements from $\mathcal{Z}$. Then $\mathcal{Z}$ is complete (and hence Hilbert) if and only if $\mathcal{Z}$ is closed.

Since $H^1(\Omega)^3$ is a Hilbert space and $\mathbb{V}_\tau(\Omega)$ is a closed subspace of it, Theorem 3.5 allows us to conclude that $\mathbb{V}_\tau(\Omega)$ is a Hilbert space by itself.

In (3.49.e), $\{\mathbf{w}, \mathbf{v}\}$ means ‘consider $\{\mathbf{w}, \mathbf{v}\}$ as unknowns to be inserted as the arguments for $a(\cdot, \cdot)$ ’, whereas in (3.49.f) $\{\mathbf{v}, p\}$ means ‘consider $\{\mathbf{v}, p\}$ as unknowns to be inserted as the arguments for $b(\cdot, \cdot)$ ’. In (3.49.g) and (3.49.h), the empty parentheses are to be filled with elements from $\mathbb{V}_\tau(\Omega)$ and $L^2(\Omega)$, respectively.

According to the identification (3.49), problem (3.48) can be rewritten as

$$\begin{aligned} & \text{Find } (\mathbf{e}^0, p) \in \mathbb{V}_\tau(\Omega) \times L^2(\Omega) \text{ such that} \\ & a(\mathbf{e}^0, \mathbf{v}) + b(\mathbf{v}, p) = \langle f^*, \mathbf{v} \rangle_{\mathbb{V}_\tau(\Omega)^*, \mathbb{V}_\tau(\Omega)} \quad \forall \mathbf{v} \in \mathbb{V}_\tau(\Omega) \\ & b(\mathbf{e}^0, q) = \langle g^*, q \rangle_{L^2(\Omega)^*, L^2(\Omega)} \quad \forall q \in L^2(\Omega). \end{aligned} \quad (3.55)$$

Problem (3.55) fits the framework of Theorem 3.1. In order to show that (3.55) is well-posed, all one needs to do is to verify the four hypotheses (3.5.a), (3.5.b), (3.5.e) and (3.5.f). However, straight from the beginning, there is a serious issue with the sesquilinear form a :

$$a(\mathbf{w}, \mathbf{v}) = \int_{\Omega} (\bar{\mathbf{\Lambda}} \cdot \nabla \mathbf{w}) : \nabla \mathbf{v}^* - \int_{\Omega} k_0^2 \mathbf{w} \cdot \mathbf{v}^*, \quad \forall \mathbf{w}, \mathbf{v} \in \mathbb{V}_\tau(\Omega). \quad (3.56)$$

The sesquilinear form (3.56) and variants thereof result from standard variational formulations associated with the Helmholtz equation, which is one of the pillars in the study of time-harmonic waves. The question is that sesquilinear forms associated with the Helmholtz equation are known to be *not coercive*, i.e., they do not satisfy a condition such as (3.5.d) [Ihlenburg, 1998], [Moiola and Spence, 2014]. This poses a difficulty when assessing the well-posedness of the weak formulations in which they occur. In the next section, the well-posedness of the scattering system (3.55) will be studied in a different way.

3.3.2 Well-posedness

When the sesquilinear form under examination is not coercive, one can resort to other methods to show that the variational problem is well-posed. The Fredholm Alternative is generally employed in the study of the variational formulation resulting from the Helmholtz equation [Evans, 2010], [Salsa, 2008], [Ihlenburg, 1998].

The application of the Fredholm Alternative to the study of the well-posedness of differential equations is generally presented for problems in a single variable. We, on

the other hand, are interested in problems described by *two* variables, namely, the scattered electric field and the Lagrange multiplier (pseudopressure).

What we are going to do is to find a way to merge the Fredholm Alternative and the theory of mixed formulations in order to get a result similar to Theorem 3.1, able to take non-coercive forms into account. The next sections are, in a sense, the most important of this thesis, since they will provide the theoretical basis for the meshfree method to be presented later.

3.3.3 The Fredholm Alternative

In order to state the Fredholm Alternative, we need some more notions.

A sequence of elements $\{x_n\}_{n=1}^{\infty}$ in a normed space $\mathcal{X}$ is *bounded* if there is a positive real number M such that the norm of all elements of the sequence is smaller than or equal to M , i.e.,

$$\forall n \in \mathbb{N} \quad \|x_n\|_{\mathcal{X}} \leq M. \quad (3.57)$$

Among the notions of compactness, the one that suits best our purposes is the *sequential compactness* [Searcoid, 2007], [Conway, 1994], [Kreyszig, 1989]. In relation to operators, a *compact operator* can be characterized as follows. Suppose $\mathcal{X}$ and $\mathcal{Y}$ are normed spaces, and $T: \mathcal{X} \rightarrow \mathcal{Y}$ is a bounded linear operator. We say that T is compact if and only if for any bounded sequence $\{x_n\}_{n=1}^{\infty}$ in $\mathcal{X}$, the image sequence $\{y_n\}_{n=1}^{\infty} = \{Tx_n\}_{n=1}^{\infty}$ in $\mathcal{Y}$ admits a convergent subsequence.

The space of all bounded linear operators between normed spaces $\mathcal{X}$ and $\mathcal{Y}$ is usually represented as $\mathcal{L}(\mathcal{X}, \mathcal{Y})$, whereas the space of all compact operators between $\mathcal{X}$ and $\mathcal{Y}$ is represented as $\mathcal{K}(\mathcal{X}, \mathcal{Y})$. It can be proved that $\mathcal{K}(\mathcal{X}, \mathcal{Y})$ is a closed linear subspace of $\mathcal{L}(\mathcal{X}, \mathcal{Y})$ [Brezis, 2010].

In what regards compact operators, the following result holds true (it is generally not stated as a theorem, but we will call it so here) [Conway, 1994], [Brezis, 2010]:

Theorem 3.5: Composition of operators – *Let E , F and G be three Banach spaces. Suppose that operators T and S are such that either*

- (i) $T \in \mathcal{L}(E, F)$ and $S \in \mathcal{K}(F, G)$ or
- (ii) $T \in \mathcal{K}(E, F)$ and $S \in \mathcal{L}(F, G)$ is the case. Then

$$S \circ T \in \mathcal{K}(E, G). \quad (3.58)$$

We can now state the Fredholm Alternative, in the form of a theorem [Brezis, 2010].

Theorem 3.6: The Fredholm Alternative – Let V be a normed vector space, and let $T \in \mathcal{K}(V, V)$. Then,

- (i) $\text{Ker}(I_V - T)$ is finite dimensional.
- (ii) $R(I_V - T)$ is closed. Moreover, $R(I_V - T) = (\text{Ker}(I_V - T^*))^\perp$.
- (iii) $\text{Ker}(I_V - T) = \{0_V\} \Leftrightarrow R(I_V - T) = V$.
- (iv) $\dim \text{Ker}(I_V - T) = \dim \text{Ker}(I_V - T^*)$.

In the statement of Theorem 3.6 above, Ker denotes the *kernel*, or null space of an operator, and I_V is the *identity operator* on the space V , i.e., I_V maps elements of V to themselves: $\forall x \in V \quad I_V x = x$. Also, ‘ R ’ means the *range* or image of an operator, and T^* is the adjoint operator. (In the course of our work, the notion of adjoint operator will not be necessary, so we will not state the definition here. Standard books on functional analysis discuss it exhaustively.) Finally, 0_V is the zero element of the space V (in order to distinguish it from the real number 0), and ‘ $\dim$ ’ means ‘dimension’.

Theorem 3.6 is stated in very abstract terms, i.e., it expresses relations between kernels and ranges of operators in spaces whose nature is left unspecified. In this thesis, we are concerned with sesquilinear forms ‘acting’ on function spaces, so more specialization is required. Before we move on, we need more definitions.

3.3.4 Embeddings

Let V and H be two Hilbert spaces. We say that V is *continuously embedded* in H , represented as $V \hookrightarrow H$, if two requirements are met. First, there is an injective and structure preserving map

$$I_{V \rightarrow H}: V \rightarrow H. \quad (3.59)$$

For our purposes, this map will be either the *inclusion map* (the case when $V \subset H$, and $I_{V \rightarrow H}$ is just the identity map) or the *Riesz map* (the case when $H = V^*$, and $I_{V \rightarrow V^*}$ is the map which identifies an element from a Hilbert space with a functional in its dual space, according to the Riesz’s representation theorem). The second requirement is that the map $I_{V \rightarrow H}$ is continuous, i.e.,

$$\|I_{V \rightarrow H}(u)\|_H \leq C_e \|u\|_V, \quad \forall u \in V, \quad (3.60)$$

where C_e is a positive constant independent of u . When the linear structure is preserved (as is the case in our applications), (3.59) and (3.60) allows us to conclude that

$$I_{V \rightarrow H} \in \mathcal{L}(V, H). \quad (3.61)$$

In the spaces V and H , the norms are usually different, so u measured by the norm of V is not generally equal to $I_{V \rightarrow H}(u)$ measured by the norm of H . These measurements are related via (3.60), though.

When establishing the existence and uniqueness of solutions to variational problems, the following result will be useful [Böhmer, 2010], [Salsa, 2008].

Theorem 3.7: Embeddings – *Let V and H be two Hilbert spaces, and suppose that $V \hookrightarrow H$. If we define an operator T by*

$$\langle T(w), v \rangle_{V^*, V} := (w, I_{V \rightarrow H}(v))_H, \quad \forall w \in H, \quad \forall v \in V, \quad (3.62)$$

then

$$T \in \mathcal{L}(H, V^*) \quad (3.63)$$

Proof: Fix an arbitrary $w \in H$. Then

$$|\langle T(w), v \rangle_{V^*, V}| = |(w, I_{V \rightarrow H}(v))_H| \leq \|w\|_H \|I_{V \rightarrow H}(v)\|_H, \quad \forall v \in V \quad (3.64.a)$$

according to the Cauchy-Schwarz inequality in the Hilbert space H . From (3.60), we see that

$$|\langle T(w), v \rangle_{V^*, V}| \leq C_e \|w\|_H \|v\|_V, \quad \forall v \in V, \quad (3.64.b)$$

and it becomes evident that

$$\|T(w)\|_{V^*} := \sup_{v \in V \setminus \{0\}} \frac{|\langle T(w), v \rangle_{V^*, V}|}{\|v\|_V} \leq C_e \|w\|_H \quad (3.64.c)$$

and so $T(w)$ is a bounded linear functional on V , i.e., $T(w) \in V^*$. Since $w \in H$ was arbitrary, we get that

$$\|T(w)\|_{V^*} \leq C_e \|w\|_H, \quad \forall w \in H \quad (3.64.d)$$

which implies that

$$\|T\|_{\mathcal{L}(H, V^*)} := \sup_{w \in H \setminus \{0\}} \frac{\|T(w)\|_{V^*}}{\|w\|_H} \leq C_e. \quad (3.64.e)$$

Since C_e is finite, T is a bounded linear operator, i.e., $T \in \mathcal{L}(H, V^*)$. ■

3.3.5 Well-posedness of non-coercive problems

Based on the material we have gathered so far, we can now state and prove a result concerning the well-posedness of problems in which the sesquilinear form a is not coercive.

Theorem 3.8: Non-coercive problems – Suppose the following hypotheses are true:

- (i) V and H are two Hilbert spaces satisfying the requirements of Theorem 3.7, i.e., $V \hookrightarrow H$.
- (ii) The map $I_{V \rightarrow H}$ is compact, i.e., $I_{V \rightarrow H} \in \mathcal{K}(V, H)$.
- (iii) $a(\cdot, \cdot): V \times V \rightarrow \mathbb{C}$ is a continuous sesquilinear form.
- (iv) The sesquilinear form from item (iii) satisfies the property: There exist constants $\eta > 0$ and $\kappa_0 \geq 0$ such that

$$\operatorname{Re}\{a(u, u)\} + \kappa_0 \|I_{V \rightarrow H}(u)\|_H^2 \geq \eta \|u\|_V^2, \quad \forall u \in V \quad (3.65)$$

It can be concluded that if the solution to the homogeneous (zero-data) problem

Find $u \in V$ such that

$$a(u, v) = 0, \quad \forall v \in V \quad (3.66)$$

is the zero element $u = 0_V$, then it is true that:

(a) The solution to the general problem

Find $u \in V$ such that

$$a(u, v) = \langle F, v \rangle_{V^*, V}, \quad \forall v \in V \quad (3.67)$$

exists and is unique for every functional $F \in V^*$.

(b) The solution u from (a) depends continuously on the data, i.e., there exists a positive constant C_{FA} such that

$$\|I_{V \rightarrow H}(u)\|_H \leq C_{FA} \|F\|_{V^*} \quad (3.68)$$

In (3.65), $\operatorname{Re}\{\cdot\}$ means ‘the real part of’. Theorem 3.8 says that uniqueness (the kernel of the form a is the zero element) implies existence. This theorem is so important for the development of our work that we shall prove it. There is a sketch of the proof in [Evans, 2010], restricted to the case when $V = H_0^1(\Omega)$ and $H = L^2(\Omega)$. We, on the other hand, develop a complete proof in the abstract setting, always emphasizing the operators which appear in the course of the development. We provide all details required by our standards, and the consequence is a rather long process, over ten pages long. The proof, which depends on Theorems 3.5, 3.6 and 3.7, has been moved to Appendix 1 in order to keep the continuity of the text.

We will now state the main theorem of this thesis, which deals with mixed formulations in which the sesquilinear form a is not coercive. In a sense, we shall merge Theorems 3.1 and 3.8 together. The challenge is to substitute the coercivity hypothesis (3.5.d) by condition (3.65) at the right place. This needs to be done in order to accommodate the Fredholm Alternative. In a sense, Theorem 3.9 is a mixture between the Fredholm Alternative and the Babuska-Brezzi theory of mixed formulations.

Theorem 3.9: Well-posedness of mixed formulations, non-coercive case – *Let $\mathcal{X}$ and $\mathcal{Y}$ be two Hilbert spaces, and let $a: \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{C}$ and $b: \mathcal{X} \times \mathcal{Y} \rightarrow \mathbb{C}$ be two continuous sesquilinear forms, i.e., there are positive constants α_a and α_b such that:*

(i) a is continuous, i.e.,

$$|a(x, v)| \leq \alpha_a \|x\|_{\mathcal{X}} \|v\|_{\mathcal{X}}, \quad \forall x, v \in \mathcal{X} \quad (3.69.a)$$

(ii) b is continuous, i.e.,

$$|b(x, y)| \leq \alpha_b \|x\|_{\mathcal{X}} \|y\|_{\mathcal{Y}}, \quad \forall x \in \mathcal{X} \quad \forall y \in \mathcal{Y} \quad (3.69.b)$$

Let $\mathcal{X}^0$ be the kernel of the sesquilinear form b i.e.,

$$\mathcal{X}^0 = \text{Ker } b = \{x \in \mathcal{X} \mid b(x, y) = 0, \quad \forall y \in \mathcal{Y}\}. \quad (3.69.c)$$

Consider a third Hilbert space H such that $\mathcal{X}^0$ and H satisfy the requirements of Theorem 3.7, i.e.,

(iii) $\mathcal{X}^0$ is continuously embedded into H , i.e., $\mathcal{X}^0 \hookrightarrow H$.

Moreover, it holds that:

(iv) The map $I_{\mathcal{X}^0 \rightarrow H}$ is compact, i.e., $I_{\mathcal{X}^0 \rightarrow H} \in \mathcal{K}(\mathcal{X}^0, H)$.

(v) The sesquilinear form a satisfies the following property on the kernel $\mathcal{X}^0$: There exist constants $\eta > 0$ and $\kappa_0 \geq 0$ such that

$$\text{Re}\{a(u, u)\} + \kappa_0 \|I_{\mathcal{X}^0 \rightarrow H}(u)\|_H^2 \geq \eta \|u\|_{\mathcal{X}}^2, \quad \forall u \in \mathcal{X}^0. \quad (3.69.d)$$

(vi) The sesquilinear form b satisfies the inf-sup condition, i.e., there is a positive constant $\beta_b > 0$ such that

$$\inf_{y \in \mathcal{Y} \setminus \{0\}} \sup_{x \in \mathcal{X} \setminus \{0\}} \frac{|b(x, y)|}{\|x\|_{\mathcal{X}} \|y\|_{\mathcal{Y}}} \geq \beta_b. \quad (3.69.e)$$

(vii) The solution to the homogeneous (zero-data) problem at the kernel $\mathcal{X}^0$

Find $w \in \mathcal{X}^0$ such that

$$a(w, v) = 0, \quad \forall w \in \mathcal{X}^0 \quad (3.69.f)$$

is the zero element $w = 0$. Furthermore, let us assume that:

(viii) The original space $\mathcal{X}$ is also continuously embedded H , i.e., $\mathcal{X} \hookrightarrow H$.

(ix) The spaces $\mathcal{X}$ and $\mathcal{X}^0$ are subspaces of H , i.e., $\mathcal{X} \subset H$ and $\mathcal{X}^0 \subset H$ (which implies that $I_{\mathcal{X} \rightarrow H}$ and $I_{\mathcal{X}^0 \rightarrow H}$ are inclusion maps).

Then it can be concluded that for each $f^* \in \mathcal{X}^*$ and $g^* \in \mathcal{Y}^*$, there is a unique solution to the mixed problem

$$\begin{aligned} & \text{Find } (u, p) \in \mathcal{X} \times \mathcal{Y} \text{ such that} \\ & a(u, x) + b(x, p) = \langle f^*, x \rangle_{\mathcal{X}^*, \mathcal{X}} \quad \forall x \in \mathcal{X} \\ & b(u, y) = \langle g^*, y \rangle_{\mathcal{Y}^*, \mathcal{Y}} \quad \forall y \in \mathcal{Y} \end{aligned} \tag{3.69.g}$$

It also follows that the solution u depends continuously on the data f^* and g^* in the H norm, i.e., there are positive constants K_1 and K_2 such that

$$\|u\|_H \leq K_1 \|f^*\|_{\mathcal{X}^*} + K_2 \|g^*\|_{\mathcal{Y}^*} \tag{3.69.h}$$

Note: The embedding map in expression (3.69.d) can make things look more complicated than they really are, and some explanation is required. To begin with, the element u belongs to $\mathcal{X}^0$, which is a subspace of the original Hilbert space $\mathcal{X}$, according to (3.69.c). In this way, as an element of $\mathcal{X}$ (because $\mathcal{X}^0 \subset \mathcal{X}$), it is originally measured in the $\|\cdot\|_{\mathcal{X}}$ norm, i.e., its ‘original size’ is $\|u\|_{\mathcal{X}}$.

The embedding $I_{\mathcal{X}^0 \rightarrow H}$ takes this u and maps it to the element $I_{\mathcal{X}^0 \rightarrow H}(u)$, which belongs to the Hilbert space H , different from the original Hilbert space $\mathcal{X}$. The ‘size’ of the element $I_{\mathcal{X}^0 \rightarrow H}(u)$ is therefore given by the norm in H , i.e., by $\|I_{\mathcal{X}^0 \rightarrow H}(u)\|_H$. In principle, $\|u\|_{\mathcal{X}}$ and $\|I_{\mathcal{X}^0 \rightarrow H}(u)\|_H$ are different.

In this work, it will be the case that $\mathcal{X}^0 \subset H$, according to hypothesis (ix). The implication is that the element u will be mapped to itself, i.e., $I_{\mathcal{X}^0 \rightarrow H}(u) = u$. In the end, we will get two ways of assessing the ‘size’ of u : $\|u\|_{\mathcal{X}}$ and $\|u\|_H$.

However, when we want to measure the size of u in the norm of H , as in (3.69.d), we will keep the embedding map and indicate this as $\|I_{\mathcal{X}^0 \rightarrow H}(u)\|_H$ instead of $\|u\|_H$. So this is the role of embeddings (at least in this work): To provide more than one measure for the size of an element.

■

In order to prove this theorem, we need some additional results from functional analysis. The first concept is that of *annihilator*, also called *polar set* [Brezis, 2010], [Quarteroni and Valli, 1994]. Let W be a Banach space, and let U be a subspace of W , i.e., $U \subset W$. The annihilator of U is the set

$$U_{\#} = \{G \in W^* \mid \langle G, u \rangle_{W^*, W} = 0, \quad \forall u \in U\}, \tag{3.70.a}$$

i.e., if a functional $G \in W^*$ is such that its action on all elements from the subspace U is zero, then G belongs to the annihilator of U . The next result we need is the Banach Closed Range Theorem. However, we do not need all its conclusions, so we will state just the two which will be useful to us. The proof and the other conclusions can be found in [Brezis, 2010].

Theorem 3.10: Banach Closed Range Theorem (incomplete) – Let U and W be two Banach spaces, and suppose that L is a bounded and linear operator between L and U , i.e., $L \in \mathcal{L}(U, W)$. Then

$$R(L) = (\text{Ker } L^T)_\# \quad (3.70.b)$$

$$R(L^T) = (\text{Ker } L)_\# \quad (3.70.c)$$

Expression (3.70.b) says that the range of operator L is equal to the annihilator of the kernel of the adjoint L^T . Conversely, (3.70.c) means that the range of the adjoint operator L^T is equal to the annihilator of the kernel of L .

Since the proof of Theorem 3.9 also occupies a number of pages, it has been moved to Appendix 2.

3.3.6 Back to the scattering system

The challenge now is to show that our electromagnetic problem (3.48), together with the identification (3.49), does indeed satisfy all requirements from Theorem 3.9. If we are successful in this task, our object of interest, the electric field $\mathbf{e}^0$, will exist, be unique, and will depend continuously on the data. We will begin by investigating the data, i.e., the functionals from (3.69.g).

3.3.6.1 Functionals I

The true scattered electric field is given by (2.162),

$$\mathbf{E}^s = \mathbf{e}^0 + \mathbf{u}^g, \quad (3.71.a)$$

where $\mathbf{u}^g$ is the lifting function on the boundary conditions (2.157),

$$\mathbf{g} = \begin{cases} \mathbf{0}, & \text{at } \Gamma_o \\ -\hat{\mathbf{n}}_1 \times \mathbf{E}^{inc}, & \text{at } \Gamma_1. \end{cases} \quad (3.71.b)$$

We must now ask if the boundary conditions (3.71.b) originate a lifting function $\mathbf{u}^g$ such that, after it is substituted into the right side of (3.48), it gives rise to functionals acting on elements from $\mathbb{V}_\tau(\Omega)$ and $L^2(\Omega)$. As we discussed in Section 2.2.3.5, if $\mathbf{g}$ defines a functional which is in $Y(\Gamma)$ (the range of the tangential trace operator $\boldsymbol{\gamma}_t$), then the lifting $\mathbf{u}^g$ is in $H(\mathbf{curl}; \Omega)$. Then we discussed two cases. In Case 1, $\mathbf{u}^g$ is smooth enough to be in the subspace $H^1(\Omega)^3 \subset H(\mathbf{curl}; \Omega)$, which is what interests us. In Case

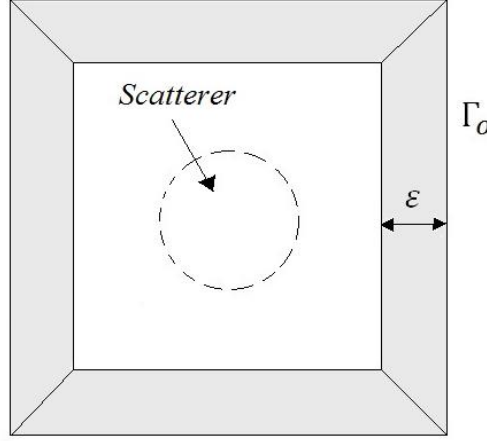


Fig. 3.1. In the scattering problems we are going to investigate, Γ_o is a rectangular contour (or a cubic surface, in 3D). The function v^ε defined in (3.71.f) decays linearly to zero inside the layer of width ε . Outside the layer, it assumes the value 1. The partial derivatives are discontinuous across the four diagonal lines. The scatterer, represented by the dotted curve, must lie outside the ε -layer.

2, $\mathbf{u}^g$ is in $H(\mathbf{curl}; \Omega)$ but not in $H^1(\Omega)^3$. However, by a density argument, we showed that in this case $\mathbf{u}^g$ can be approximated by elements from $H^1(\Omega)^3$.

Anyhow, we need to show that the $\mathbf{g}$ from (3.71.b) is in $Y(\Gamma)$. If we succeed, then we know for sure that there is an $\mathbf{u}^g$ in $H(\mathbf{curl}; \Omega)$ such that $\boldsymbol{\gamma}_t \mathbf{u}^g = \mathbf{g}$. Thereafter, we investigate solutions to the problem

$$\begin{aligned} &\text{Find } \mathbf{u}^g \in H^1(\Omega)^3 \text{ such that} \\ &\boldsymbol{\gamma}_t \mathbf{u}^g = \mathbf{g} \end{aligned} \tag{3.71.c}$$

i.e., if there is an $\mathbf{u}^g$ smooth enough to be qualified as an element from $H^1(\Omega)^3$, a subspace from $H(\mathbf{curl}; \Omega)$. In this thesis, we shall not investigate problem (3.71.c). We assume that the solution to (3.71.c) exists, i.e., we make a conjecture.

Conjecture 3.1: Lifting in $H^1(\Omega)^3$ – Consider the non-homogeneous boundary conditions

$$\mathbf{g} = \begin{cases} \mathbf{0}, & \text{at } \Gamma_o \\ -\hat{\mathbf{n}}_1 \times \mathbf{E}^{inc}, & \text{at } \Gamma_1. \end{cases} \tag{3.71.d}$$

If $\mathbf{g} \in Y(\Gamma)$, then we can find an $\mathbf{u}^g \in H^1(\Omega)^3$ such that $\boldsymbol{\gamma}_t \mathbf{u}^g = \mathbf{g}$.

The space $Y(\Gamma)$ is characterized in (2.152). Let us consider a function $v^\varepsilon: \bar{\Omega} \rightarrow \mathbb{R}$ defined by

$$v^\varepsilon(\mathbf{x}) = \begin{cases} 1, & d(\mathbf{x}, \Gamma_o) > \varepsilon \\ \frac{d(\mathbf{x}, \Gamma_o)}{\varepsilon}, & d(\mathbf{x}, \Gamma_o) \leq \varepsilon, \end{cases} \tag{3.71.e}$$

where $d(\mathbf{x}, \Gamma_o)$ is the distance from the point $\mathbf{x}$ to the outer boundary Γ_o and $\varepsilon > 0$. In

order to illustrate the meaning of (3.71.e), let us consider a two-dimensional domain Ω . The function v^ε is such that it is 1 for those points $\mathbf{x}$ whose distance to Γ_o is larger than ε . If the distance is smaller than ε , then v^ε decays linearly to zero. In other words, there is a layer of width ε ; outside this layer, v^ε is equal to 1. Inside the layer, v^ε decays linearly to zero. The width of the layer must be chosen in such a way that the scatterer is located completely outside the layer (i.e., for all points $\mathbf{x}$ in the scatterer surface Γ_1 , $d(\mathbf{x}, \Gamma_o) > \varepsilon$). The function v^ε is illustrated in Fig. 3.1.

The function v^ε is continuous in $\bar{\Omega}$, and therefore $v^\varepsilon \in L^2(\Omega)$. The derivatives $\partial v^\varepsilon / \partial x$ and $\partial v^\varepsilon / \partial y$, on the other hand, experience discontinuities along the diagonals (Fig. 3.1). But it is not difficult to see that the derivatives are square summable. According to Fig. 3.1, if Γ_o is the surface of a box defined by $X_1 \leq x \leq X_2$ and $Y_1 \leq y \leq Y_2$, then $d(\mathbf{x}, \Gamma_o)$ assumes the form

$$d(\mathbf{x}, \Gamma_o) = \min\{(x - X_1), (X_2 - x), (y - Y_1), (Y_2 - y)\}, \quad (3.71.f)$$

where $\mathbf{x} = [x, y]^T$ is an arbitrary point in $\bar{\Omega}$. From (3.71.f), we see that, within the layer, the derivative of $d(\mathbf{x}, \Gamma_o)$ with respect to x is either +1 or -1. In the same way, the derivative with respect to y is either +1 or -1. Back to (3.71.e), we conclude that, if $d(\mathbf{x}, \Gamma_o) > \varepsilon$, then $\partial v^\varepsilon / \partial x$ and $\partial v^\varepsilon / \partial y$ are zero. If $d(\mathbf{x}, \Gamma_o) \leq \varepsilon$, $|\partial v^\varepsilon / \partial x|$ and $|\partial v^\varepsilon / \partial y|$ are equal to $1/\varepsilon$. (Except at the diagonals (points of discontinuity), which constitute a set of measure zero.) Then $\partial v^\varepsilon / \partial x$ and $\partial v^\varepsilon / \partial y$ are also in $L^2(\Omega)$.

The same reasoning applies to three dimensions in what regards the derivative $\partial v^\varepsilon / \partial z$.

Before we proceed, we need two results which give us conditions under which vector fields in Ω define functionals at the boundary Γ . The proof can be found in [Ern and Guermond, 2004].

Theorem 3.11: ‘Divergence’ functionals – Let Ω be a bounded open set, and let $1 \leq p < \infty$. Suppose $\mathbf{u} \in L^2(\Omega)^3$ is a vector field such that $\nabla \cdot \mathbf{u} \in L^2(\Omega)$. Then

$$\mathbf{u} \cdot \hat{\mathbf{n}} \in W^{-\frac{1}{p}, p}(\Gamma). \quad (3.71.g)$$

Theorem 3.12: ‘Curl’ functionals – Let Ω be a bounded open set, and let $1 \leq p < \infty$. Suppose $\mathbf{u} \in L^2(\Omega)^3$ is a vector field such that $\nabla \times \mathbf{u} \in L^2(\Omega)^3$. Then

$$\mathbf{u} \times \hat{\mathbf{n}} \in W^{-\frac{1}{p}, p}(\Gamma)^3. \quad (3.71.h)$$

Let us consider an incident field $\mathbf{E}^{inc} \in L^2(\Omega)^3$ such that $\nabla \times \mathbf{E}^{inc} \in L^2(\Omega)^3$ and $\nabla \cdot \mathbf{E}^{inc} \in L^2(\Omega)$. Since $v^\varepsilon \in L^2(\Omega)$, we take the component E_x^{inc} and discover that

$$\int_{\Omega} |v^\varepsilon E_x^{inc}|^2 = \int_{\Omega} |v^\varepsilon|^2 |E_x^{inc}|^2 \leq \int_{\Omega} |E_x^{inc}|^2 < \infty, \quad (3.71.i)$$

as $|v^\varepsilon| \leq 1$ from (3.71.e) and $E_x^{inc} \in L^2(\Omega)$. We conclude that $v^\varepsilon E_x^{inc} \in L^2(\Omega)$. The same steps are applied to the other components and we discover that

$$v^\varepsilon \mathbf{E}^{inc} \in L^2(\Omega)^3. \quad (3.71.j)$$

The curl of $v^\varepsilon \mathbf{E}^{inc}$ is given by

$$\nabla \times v^\varepsilon \mathbf{E}^{inc} = \nabla v^\varepsilon \times \mathbf{E}^{inc} + v^\varepsilon \nabla \times \mathbf{E}^{inc} \quad (3.71.k)$$

Let us focus on the x -component of $\nabla \times v^\varepsilon \mathbf{E}^{inc}$ and discover that

$$\left\| (\nabla \times v^\varepsilon \mathbf{E}^{inc})_x \right\|_{L^2(\Omega)} = \left\| (\nabla v^\varepsilon \times \mathbf{E}^{inc})_x + (v^\varepsilon \nabla \times \mathbf{E}^{inc})_x \right\|_{L^2(\Omega)} \quad (3.71.l)$$

$$\leq \left\| (\nabla v^\varepsilon \times \mathbf{E}^{inc})_x \right\|_{L^2(\Omega)} + \left\| (v^\varepsilon \nabla \times \mathbf{E}^{inc})_x \right\|_{L^2(\Omega)} \quad (3.71.m)$$

$$= \left\| \frac{\partial v^\varepsilon}{\partial y} E_z^{inc} - \frac{\partial v^\varepsilon}{\partial z} E_y^{inc} \right\|_{L^2(\Omega)} + \left\| (v^\varepsilon \nabla \times \mathbf{E}^{inc})_x \right\|_{L^2(\Omega)} \quad (3.71.n)$$

$$\leq \left\| \frac{\partial v^\varepsilon}{\partial y} E_z^{inc} \right\|_{L^2(\Omega)} + \left\| \frac{\partial v^\varepsilon}{\partial z} E_y^{inc} \right\|_{L^2(\Omega)} + \left\| (v^\varepsilon \nabla \times \mathbf{E}^{inc})_x \right\|_{L^2(\Omega)}, \quad (3.71.o)$$

where the Minkowski inequality (2.73) has been used in (3.71.m) and (3.71.o). Let us now concentrate on the first term from (3.71.o):

$$\left\| \frac{\partial v^\varepsilon}{\partial y} E_z^{inc} \right\|_{L^2(\Omega)} = \left(\int_{\Omega} \left| \frac{\partial v^\varepsilon}{\partial y} E_z^{inc} \right|^2 \right)^{\frac{1}{2}} \quad (3.71.p)$$

$$= \left(\int_{\Omega} \left| \frac{\partial v^\varepsilon}{\partial y} \right|^2 |E_z^{inc}|^2 \right)^{\frac{1}{2}} \quad (3.71.q)$$

$$= \left(\int_{\Omega^\varepsilon} \left| \frac{\partial v^\varepsilon}{\partial y} \right|^2 |E_z^{inc}|^2 \right)^{\frac{1}{2}}, \quad (3.71.r)$$

where Ω^ε is the portion of the domain Ω in which $\partial v^\varepsilon / \partial y$ is different from zero. As we have seen, if the distance of a point to Γ_o is larger than ε , then $\partial v^\varepsilon / \partial y$ is zero. Also, in Ω^ε , it is true that $|\partial v^\varepsilon / \partial y| = 1/\varepsilon$. From (3.71.r) we get

$$\left(\int_{\Omega^\varepsilon} \left| \frac{\partial v^\varepsilon}{\partial y} \right|^2 |E_z^{inc}|^2 \right)^{\frac{1}{2}} = \left(\int_{\Omega^\varepsilon} \left| \frac{1}{\varepsilon} \right|^2 |E_z^{inc}|^2 \right)^{\frac{1}{2}} \quad (3.71.s)$$

$$= \frac{1}{\varepsilon} \left(\int_{\Omega^\varepsilon} |E_z^{inc}|^2 \right)^{\frac{1}{2}} \quad (3.71.t)$$

$$\leq \frac{1}{\varepsilon} \left(\int_{\Omega} |E_z^{inc}|^2 \right)^{\frac{1}{2}} = \frac{1}{\varepsilon} \|E_z^{inc}\|_{L^2(\Omega)} < \infty, \quad (3.71.u)$$

since $E_z^{inc} \in L^2(\Omega)$. We conclude that the first term in (3.71.o) is finite. The same analysis applied to the second term in (3.71.o) reveals that it is also finite. It is true that $(\nu^\varepsilon \nabla \times \mathbf{E}^{inc})_x = \nu^\varepsilon (\nabla \times \mathbf{E}^{inc})_x$, and also that $(\nabla \times \mathbf{E}^{inc})_x \in L^2(\Omega)$. We apply the same reasoning as that from (3.71.i) and discover that $(\nu^\varepsilon \nabla \times \mathbf{E}^{inc})_x \in L^2(\Omega)$. In this way, all terms from (3.71.o) are finite, which implies that the x -component of $\nabla \times (\nu^\varepsilon \mathbf{E}^{inc})$ is in $L^2(\Omega)$. If we repeat this argument to the y and z -components, we finally find that

$$\nabla \times \nu^\varepsilon \mathbf{E}^{inc} \in L^2(\Omega)^3. \quad (3.71.v)$$

From (3.71.j) and (3.71.v), we make $\mathbf{u} = \nu^\varepsilon \mathbf{E}^{inc}$ and $p = 2$ in Theorem 3.12 and discover that

$$(\nu^\varepsilon \mathbf{E}^{inc}) \times \hat{\mathbf{n}} \in W^{-\frac{1}{2},2}(\Gamma)^3. \quad (3.71.w)$$

Since the Sobolev spaces $W^{m,p}$ are usually represented as H^m when $p = 2$, expression above is equivalent to

$$(\nu^\varepsilon \mathbf{E}^{inc}) \times \hat{\mathbf{n}} \in H^{-1/2}(\Gamma)^3, \quad (3.71.x)$$

which of course implies that

$$-\hat{\mathbf{n}} \times \nu^\varepsilon \mathbf{E}^{inc} \in H^{-1/2}(\Gamma)^3, \quad (3.71.y)$$

Now that we know that $-\hat{\mathbf{n}} \times \nu^\varepsilon \mathbf{E}^{inc}$ defines a functional, we may ask: How does it operate on elements from $H^{1/2}(\Gamma)^3$? The usual duality pairing between elements from $H^{-1/2}(\Gamma)$ and $H^{1/2}(\Gamma)$ is just a boundary integral [Boffi et al., 2013]. If $w' \in H^{-1/2}(\Gamma)$ and $v \in H^{1/2}(\Gamma)$, then

$$\langle w', v \rangle_{H^{-1/2}(\Gamma), H^{1/2}(\Gamma)} := \int_{\Gamma} w' v. \quad (3.71.z)$$

Now let $\mathbf{v} \in H^1(\Omega)^3$ be arbitrary. According to the trace operator $\boldsymbol{\gamma}_0^d$ in (2.58), $\boldsymbol{\gamma}_0^d \mathbf{v} \in H^{1/2}(\Gamma)^3$. The action of $-\hat{\mathbf{n}} \times \nu^\varepsilon \mathbf{E}^{inc}$ on elements from $H^{1/2}(\Gamma)^3$ is given by

$$\langle -\hat{\mathbf{n}} \times \nu^\varepsilon \mathbf{E}^{inc}, \boldsymbol{\gamma}_0^d \mathbf{v} \rangle_{H^{-1/2}(\Gamma)^3, H^{1/2}(\Gamma)^3} := \int_{\Gamma} (-\hat{\mathbf{n}} \times \nu^\varepsilon \mathbf{E}^{inc}) \cdot \mathbf{v} \quad (3.72.a)$$

$$= \int_{\Gamma_0} (-\hat{\mathbf{n}} \times \nu^\varepsilon \mathbf{E}^{inc}) \cdot \mathbf{v} + \int_{\Gamma_1} (-\hat{\mathbf{n}} \times \nu^\varepsilon \mathbf{E}^{inc}) \cdot \mathbf{v} \quad (3.72.b)$$

Since according to (3.71.e) ν^ε is 0 at Γ_0 and 1 at Γ_1 , we see that

$$\langle -\hat{\mathbf{n}} \times \nu^\varepsilon \mathbf{E}^{inc}, \boldsymbol{\gamma}_0^d \mathbf{v} \rangle_{H^{-1/2}(\Gamma)^3, H^{1/2}(\Gamma)^3} = \int_{\Gamma_1} (-\hat{\mathbf{n}} \times \mathbf{E}^{inc}) \cdot \mathbf{v}, \quad (3.72.c)$$

which is precisely what one would expect in what regards the action of the function $\mathbf{g}$ in (3.71.b) on other functions defined at the boundary Γ_1 . In a sense, the functional from (3.71.y) together with its operation (3.72.c) is a more elegant description than just saying “the functional $\mathbf{g}$ from (3.71.b)”.

Now that we have a proper description of a functional induced by the boundary condition $\mathbf{g}$, we must ask if this functional is in the range of the tangential trace operator $\boldsymbol{\gamma}_t$, i.e., if it is an element from $Y(\Gamma)$. In order to give an affirmative answer, we need to show that our functional satisfies the requirements from (2.152). The strategy to follow is: First, to show that our functional is in the space defined in (2.154). Second, to show with the help of (2.153) that the surface divergence of our functional is in $H^{-1/2}(\Gamma)$.

It is true that

$$(-\hat{\mathbf{n}} \times \nu^\varepsilon \mathbf{E}^{inc}) \cdot \hat{\mathbf{n}} = 0 \quad (3.72.d)$$

on all points of Γ (excluding sets of measure zero), since the vector $-\hat{\mathbf{n}} \times \nu^\varepsilon \mathbf{E}^{inc}$ is by definition orthogonal to the normal vector $\hat{\mathbf{n}}$. Then, $-\hat{\mathbf{n}} \times \nu^\varepsilon \mathbf{E}^{inc} \in \mathbf{H}_t^{-1/2}(\Gamma)$, defined in (2.154).

From (3.71.v), we know that $\nabla \times \nu^\varepsilon \mathbf{E}^{inc} \in L^2(\Omega)^3$. Of course, since the divergence of a curl is zero, $\nabla \cdot (\nabla \times \nu^\varepsilon \mathbf{E}^{inc}) = 0 \in L^2(\Omega)$. Therefore, we consider Theorem 3.11 with $\mathbf{u} = \nabla \times \nu^\varepsilon \mathbf{E}^{inc}$ and $p = 2$ and conclude that

$$(\nabla \times \nu^\varepsilon \mathbf{E}^{inc}) \cdot \hat{\mathbf{n}} \in H^{-1/2}(\Gamma), \quad (3.72.e)$$

which is no different than

$$\hat{\mathbf{n}} \cdot (\nabla \times \nu^\varepsilon \mathbf{E}^{inc}) \in H^{-1/2}(\Gamma). \quad (3.72.f)$$

Consider now identity (2.153) with $\mathbf{v} = -\nu^\varepsilon \mathbf{E}^{inc}$ [which belongs to $H(\mathbf{curl}; \Omega)$ due to (3.71.j) and (3.71.v)] and find that

$$\nabla_\Gamma \cdot (-\hat{\mathbf{n}} \times \nu^\varepsilon \mathbf{E}^{inc}) = \hat{\mathbf{n}} \cdot (\nabla \times \nu^\varepsilon \mathbf{E}^{inc}) \quad (3.72.g)$$

From (3.72.f) and (3.72.g), we learn that the surface divergence of our functional $-\hat{\mathbf{n}} \times \nu^\varepsilon \mathbf{E}^{inc}$ is indeed in $H^{-1/2}(\Gamma)$. From (2.152), we are finally able to conclude that

$$-\hat{\mathbf{n}} \times \nu^\varepsilon \mathbf{E}^{inc} \in Y(\Gamma) \quad (3.72.h)$$

According to (3.71.e) ν^ε is 0 at Γ_o and 1 at Γ_1 , so (3.72.h) above is the same as saying that

$$\mathbf{g} \in Y(\Gamma), \quad (3.72.i)$$

where $\mathbf{g}$ has been defined in (3.71.b). Since $\mathbf{g}$ is in the range of the trace operator $\boldsymbol{\gamma}_t$, there are functions $\mathbf{w}$ in $H(\mathbf{curl}; \Omega)$ such that $\boldsymbol{\gamma}_t \mathbf{w} = \mathbf{g}$. There is an infinite number of such functions, as $\boldsymbol{\gamma}_t$ is not injective [its kernel is given by (2.150)]. We may ask: Among these functions in $H(\mathbf{curl}; \Omega)$ whose trace is $\mathbf{g}$, can we find one which is in $H^1(\Omega)^3$? We have not explored the conditions which ultimately assure us that such a function exists. Hence we just conjecture its existence (Conjecture 3.1). So we assume that such a function exists in $H^1(\Omega)^3$, and call it $\mathbf{u}^g$.

3.3.6.2 Functionals II

In the right side of (3.48.a), we define a functional f^* according to identification (3.49) whose action on testing functions from $\mathbb{V}_\tau(\Omega)$ is given by

$$f^*(\mathbf{v}) = - \int_{\Omega} (\bar{\mathbf{\Lambda}} \cdot \nabla \mathbf{u}^g) : \nabla \mathbf{v}^* + \int_{\Omega} k_0^2 \mathbf{u}^g \cdot \mathbf{v}^*, \quad \forall \mathbf{v} \in \mathbb{V}_\tau(\Omega). \quad (3.72.j)$$

It is clearly (anti-)linear; but now we may ask: Is it bounded in order to qualify as an element from $\mathbb{V}_\tau(\Omega)^*$? From (2.181) and (2.94) [adapted to the complex setting], repeated below,

$$\left| \int_{\Omega} (\bar{\mathbf{\Lambda}} \cdot \nabla \mathbf{v}) : \nabla \mathbf{w}^* \right| \leq \Lambda_M \|\mathbf{v}\|_{H^1(\Omega)^3} \|\mathbf{w}\|_{H^1(\Omega)^3} \quad \forall \mathbf{v}, \mathbf{w} \in H^1(\Omega)^3, \quad (3.72.k)$$

$$\left| \int_{\Omega} \mathbf{v} \cdot \mathbf{w}^* \right| \leq \|\mathbf{v}\|_{L^2(\Omega)^3} \|\mathbf{w}\|_{H^1(\Omega)^3}, \quad \forall \mathbf{v} \in L^2(\Omega)^3 \quad \forall \mathbf{w} \in H^1(\Omega)^3, \quad (3.72.l)$$

we observe that

$$\left| \int_{\Omega} (\bar{\mathbf{\Lambda}} \cdot \nabla \mathbf{u}^g) : \nabla \mathbf{v}^* \right| \leq \Lambda_M \|\mathbf{u}^g\|_{H^1(\Omega)^3} \|\mathbf{v}\|_{H^1(\Omega)^3} \quad \forall \mathbf{v} \in H^1(\Omega)^3, \quad (3.72.m)$$

$$\left| \int_{\Omega} k_0^2 \mathbf{u}^g \cdot \mathbf{v}^* \right| \leq k_0^2 \|\mathbf{u}^g\|_{L^2(\Omega)^3} \|\mathbf{v}\|_{H^1(\Omega)^3}, \quad \forall \mathbf{v} \in H^1(\Omega)^3. \quad (3.72.n)$$

Since $\mathbb{V}_\tau(\Omega) \subset H^1(\Omega)^3$, (3.72.m) and (3.72.n) imply that

$$\left| \int_{\Omega} (\bar{\mathbf{\Lambda}} \cdot \nabla \mathbf{u}^g) : \nabla \mathbf{v}^* \right| \leq \Lambda_M \|\mathbf{u}^g\|_{H^1(\Omega)^3} \|\mathbf{v}\|_{H^1(\Omega)^3} \quad \forall \mathbf{v} \in \mathbb{V}_\tau(\Omega), \quad (3.72.o)$$

$$\left| \int_{\Omega} k_0^2 \mathbf{u}^g \cdot \mathbf{v}^* \right| \leq k_0^2 \|\mathbf{u}^g\|_{L^2(\Omega)^3} \|\mathbf{v}\|_{H^1(\Omega)^3}, \quad \forall \mathbf{v} \in \mathbb{V}_\tau(\Omega). \quad (3.72.p)$$

From (3.72.j) and with the help of the triangle inequality, we learn that

$$|f^*(\mathbf{v})| \leq \left| \int_{\Omega} (\bar{\mathbf{\Lambda}} \cdot \nabla \mathbf{u}^g) : \nabla \mathbf{v}^* \right| + \left| \int_{\Omega} k_0^2 \mathbf{u}^g \cdot \mathbf{v}^* \right|, \quad \forall \mathbf{v} \in \mathbb{V}_\tau(\Omega). \quad (3.72.q)$$

When we consider (3.72.o) and (3.72.p), it is not difficult to see that

$$|f^*(\mathbf{v})| \leq (\Lambda_M + k_0^2) \|\mathbf{u}^g\|_{H^1(\Omega)^3} \|\mathbf{v}\|_{H^1(\Omega)^3}, \quad \forall \mathbf{v} \in \mathbb{V}_\tau(\Omega), \quad (3.72.r)$$

and hence that

$$f^* \in \mathbb{V}_\tau(\Omega)^*. \quad (3.72.s)$$

We now concentrate on the right side of (3.48.b), and define a functional g^* whose action on testing functions from $L^2(\Omega)$ is given by

$$g^*(q) = \int_{\Omega} q^* \nabla \cdot \mathbf{u}^g, \quad \forall q \in L^2(\Omega). \quad (3.72.t)$$

It is (anti-)linear, and it remains to verify if it is bounded. From (2.86.b) [adapted to the complex setting] we observe that

$$\left| \int_{\Omega} q^* \nabla \cdot \mathbf{u}^g \right| \leq \|q\|_{L^2(\Omega)} \|\mathbf{u}^g\|_{H^1(\Omega)^3} \quad \forall q \in L^2(\Omega). \quad (3.72.u)$$

Expressions (3.72.t) and (3.72.u) reveal us that

$$|g^*(q)| = \left| \int_{\Omega} q^* \nabla \cdot \mathbf{u}^g \right| \leq \|q\|_{L^2(\Omega)} \|\mathbf{u}^g\|_{H^1(\Omega)^3} \quad \forall q \in L^2(\Omega), \quad (3.72.v)$$

from which it is not difficult to see that g^* is bounded. Therefore,

$$g^* \in L^2(\Omega)^* = L^2(\Omega), \quad (3.72.w)$$

since $L^2(\Omega)$ can be identified with its dual.

Now we are going to study the hypotheses from Theorem 3.9, and show that the scattering system (3.48) satisfies each one of them. The order in which they will be addressed is such that the easier ones will be considered first.

3.3.6.3 Theorem 3.9, Hypotheses (i) and (ii)

The original spaces $\mathbb{V}_\tau(\Omega)$ and $L^2(\Omega)$ are Hilbert spaces. [This question is addressed in Chart 3.1, regarding $\mathbb{V}_\tau(\Omega)$, and in [Brezis, 2010], in what concerns $L^2(\Omega)$]. According to the identification (3.49),

$$a(\mathbf{u}, \mathbf{v}) = \int_{\Omega} (\bar{\Lambda} \cdot \nabla \mathbf{u}) : \nabla \mathbf{v}^* - \int_{\Omega} k_0^2 \mathbf{u} \cdot \mathbf{v}^*, \quad \forall \mathbf{u}, \mathbf{v} \in \mathbb{V}_\tau(\Omega). \quad (3.73.a)$$

We observe that

$$|a(\mathbf{u}, \mathbf{v})| \leq \left| \int_{\Omega} (\bar{\Lambda} \cdot \nabla \mathbf{u}) : \nabla \mathbf{v}^* \right| + \left| \int_{\Omega} k_0^2 \mathbf{u} \cdot \mathbf{v}^* \right|, \quad \forall \mathbf{u}, \mathbf{v} \in \mathbb{V}_\tau(\Omega). \quad (3.73.b)$$

thanks to the triangle inequality. From (2.181) and (2.94), and from the fact that $\mathbb{V}_\tau(\Omega) \subset H^1(\Omega)^3 \subset L^2(\Omega)^3$,

$$|a(\mathbf{u}, \mathbf{v})| \leq (\Lambda_M \|\mathbf{u}\|_{H^1(\Omega)^3} + k_0^2 \|\mathbf{u}\|_{L^2(\Omega)^3}) \|\mathbf{v}\|_{H^1(\Omega)^3}, \quad \forall \mathbf{u}, \mathbf{v} \in \mathbb{V}_\tau(\Omega). \quad (3.73.c)$$

Moreover, since $\|\mathbf{u}\|_{L^2(\Omega)^3} \leq \|\mathbf{u}\|_{H^1(\Omega)^3}$ – as given by (2.142) – we arrive at

$$|a(\mathbf{u}, \mathbf{v})| \leq (\Lambda_M + k_0^2) \|\mathbf{u}\|_{H^1(\Omega)^3} \|\mathbf{v}\|_{H^1(\Omega)^3}, \quad \forall \mathbf{u}, \mathbf{v} \in \mathbb{V}_\tau(\Omega), \quad (3.73.d)$$

which, according to (3.3), allows to conclude that the sesquilinear form a is bounded (or continuous), and that $\alpha_a = \Lambda_M + k_0^2$.

The sesquilinear form b is given by

$$b(\mathbf{v}, q) = - \int_{\Omega} q^* \nabla \cdot \mathbf{v}, \quad \forall q \in L^2(\Omega) \quad \forall \mathbf{v} \in \mathbb{V}_\tau(\Omega) \quad (3.73.e)$$

From (2.86.b),

$$|b(\mathbf{v}, q)| \leq \|q\|_{L^2(\Omega)} \|\mathbf{v}\|_{H^1(\Omega)^3}, \quad \forall q \in L^2(\Omega) \quad \forall \mathbf{v} \in \mathbb{V}_\tau(\Omega), \quad (3.73.f)$$

as $\mathbb{V}_\tau(\Omega) \subset H^1(\Omega)^3$. Again, (3.3) shows that the sesquilinear form b is bounded (or continuous), and that $\alpha_b = 1$.

3.3.6.4 Theorem 3.9, Hypotheses (iii), (iv), (viii) and (ix)

We need the following fact from the theory of Sobolev spaces [Leoni, 2009], [Brezis, 2010], [Salsa, 2008]:

Theorem 3.13: Compact Embeddings – *Let Ω be a bounded and Lipschitz domain in $\mathbb{R}^d$. Then,*

(a) *If $d > 2$, then $H^1(\Omega) \hookrightarrow L^p(\Omega)$ for $2 \leq p \leq 2n/(n-2)$. Moreover, if $2 \leq p \leq 2n/(n-2)$, the embedding of $H^1(\Omega)$ in $L^p(\Omega)$ is compact.*

(b) *If $d = 2$, then $H^1(\Omega) \hookrightarrow L^p(\Omega)$ for $2 \leq p \leq \infty$, with compact embedding.*

We are interested in the case $p = 2$. So from Theorem 3.13 we are able to conclude that in either 2 or 3 dimensions, it is true that

$$H^1(\Omega) \hookrightarrow L^2(\Omega), \quad (3.74.a)$$

in which the embedding map $I_{H^1(\Omega) \rightarrow L^2(\Omega)}$ is compact. From the discussion in Section 3.3.4, (3.74.a) means that there is a positive constant C_e such that

$$\|I_{H^1(\Omega) \rightarrow L^2(\Omega)}(u)\|_{L^2(\Omega)} \leq C_e \|u\|_{H^1(\Omega)}, \quad \forall u \in H^1(\Omega). \quad (3.74.b)$$

Moreover, in the notation from Section 3.3.3,

$$I_{H^1(\Omega) \rightarrow L^2(\Omega)} \in \mathcal{K}(H^1(\Omega), L^2(\Omega)). \quad (3.74.c)$$

Let $\mathbf{u} \in H^1(\Omega)^3$ be arbitrary. (The same analysis applies for the two-dimensional case, so we will stick to the more general three-dimensional case here.) Since each of its components is in $H^1(\Omega)$, they are continuously embedded in $L^2(\Omega)$, i.e.,

$$\|I_{H^1(\Omega) \rightarrow L^2(\Omega)}(u_x)\|_{L^2(\Omega)} \leq C_e \|u_x\|_{H^1(\Omega)} \quad (3.74.d)$$

$$\|I_{H^1(\Omega) \rightarrow L^2(\Omega)}(u_y)\|_{L^2(\Omega)} \leq C_e \|u_y\|_{H^1(\Omega)} \quad (3.74.e)$$

$$\|I_{H^1(\Omega) \rightarrow L^2(\Omega)}(u_z)\|_{L^2(\Omega)} \leq C_e \|u_z\|_{H^1(\Omega)}. \quad (3.74.f)$$

If we consider expressions (3.74.d) – (3.74.f) squared, and also the definition of the $H^1(\Omega)^3$ norm in (2.34), we get

$$\begin{aligned} & \|I_{H^1(\Omega) \rightarrow L^2(\Omega)}(u_x)\|_{L^2(\Omega)}^2 + \|I_{H^1(\Omega) \rightarrow L^2(\Omega)}(u_y)\|_{L^2(\Omega)}^2 + \\ & \|I_{H^1(\Omega) \rightarrow L^2(\Omega)}(u_z)\|_{L^2(\Omega)}^2 \leq C_e^2 \|\mathbf{u}\|_{H^1(\Omega)^3}^2. \end{aligned} \quad (3.74.g)$$

If we define the ‘multidimensional’ embedding map $I_{H^1(\Omega)^3 \rightarrow L^2(\Omega)^3}$ as

$$I_{H^1(\Omega)^3 \rightarrow L^2(\Omega)^3}(\mathbf{u}) := \begin{bmatrix} I_{H^1(\Omega) \rightarrow L^2(\Omega)}(u_x) \\ I_{H^1(\Omega) \rightarrow L^2(\Omega)}(u_y) \\ I_{H^1(\Omega) \rightarrow L^2(\Omega)}(u_z) \end{bmatrix}, \quad (3.74.h)$$

then (3.74.g) becomes

$$\|I_{H^1(\Omega)^3 \rightarrow L^2(\Omega)^3}(\mathbf{u})\|_{L^2(\Omega)^3}^2 \leq C_e^2 \|\mathbf{u}\|_{H^1(\Omega)^3}^2. \quad (3.74.i)$$

Consequently, the embedding defined in (3.74.h) is continuous.

In order to find out if it also compact, we consider an arbitrary bounded sequence $\{\mathbf{v}_n\}_{n=1}^\infty$ in $H^1(\Omega)^3$, i.e., there is a positive constant M such that $\|\mathbf{v}_n\|_{H^1(\Omega)^3} \leq M$ for all n .

This sequence defines three individual sequences in $H^1(\Omega)$, namely, the sequences $\{v_n^x\}_{n=1}^\infty$, $\{v_n^y\}_{n=1}^\infty$ and $\{v_n^z\}_{n=1}^\infty$ formed by the x , y , and z components of $\{\mathbf{v}_n\}_{n=1}^\infty$. Since [with the help of (2.33)] it is true that for all n ,

$$\|\mathbf{v}_n\|_{H^1(\Omega)^3}^2 = \|v_n^x\|_{H^1(\Omega)}^2 + \|v_n^y\|_{H^1(\Omega)}^2 + \|v_n^z\|_{H^1(\Omega)}^2 \leq M^2, \quad (3.74.j)$$

then $\{v_n^x\}_{n=1}^\infty \leq M$, $\{v_n^y\}_{n=1}^\infty \leq M$, and $\{v_n^z\}_{n=1}^\infty \leq M$, i.e., the three individual sequences are bounded.

As $\{v_n^x\}_{n=1}^\infty$ is bounded, $\{I_{H^1(\Omega) \rightarrow L^2(\Omega)}(v_n^x)\}_{n=1}^\infty$ admits a convergent subsequence in $L^2(\Omega)$, i.e.,

$$I_{H^1(\Omega) \rightarrow L^2(\Omega)}(v_{n_j}^x) \rightarrow w^x \text{ as } j \rightarrow \infty. \quad (3.74.k)$$

Of course, since $\{v_n^y\}_{n=1}^\infty$ is bounded, then $\{v_{n_j}^y\}_{j=1}^\infty$ is also bounded. Therefore, $\{I_{H^1(\Omega) \rightarrow L^2(\Omega)}(v_{n_j}^y)\}_{j=1}^\infty$ admits a convergent subsequence in $L^2(\Omega)$, i.e.,

$$I_{H^1(\Omega) \rightarrow L^2(\Omega)}(v_{n_{j_k}}^y) \rightarrow w^y \text{ as } k \rightarrow \infty. \quad (3.74.l)$$

By the same argument, since $\{v_n^z\}_{n=1}^\infty$ is bounded, then $\{v_{n_{j_k}}^z\}_{k=1}^\infty$ is also bounded. Then, $\{I_{H^1(\Omega) \rightarrow L^2(\Omega)}(v_{n_{j_k}}^z)\}_{k=1}^\infty$ admits a convergent subsequence in $L^2(\Omega)$, i.e.,

$$I_{H^1(\Omega) \rightarrow L^2(\Omega)}(v_{n_{j_{k_l}}}^z) \rightarrow w^z \text{ as } l \rightarrow \infty. \quad (3.74.m)$$

Since subsequences of convergent sequences converge to the same limit, from (3.74.k) and (3.74.l) we see that

$$I_{H^1(\Omega) \rightarrow L^2(\Omega)}(v_{n_{j_{k_l}}}^x) \rightarrow w^x \text{ as } l \rightarrow \infty \quad (3.74.n)$$

$$I_{H^1(\Omega) \rightarrow L^2(\Omega)}(v_{n_{j_{k_l}}}^y) \rightarrow w^y \text{ as } l \rightarrow \infty. \quad (3.74.o)$$

If we take into account the embedding defined in (3.74.h),

$$I_{H^1(\Omega)^3 \rightarrow L^2(\Omega)^3}(\mathbf{v}_{n_{j_{k_l}}}) = \begin{bmatrix} I_{H^1(\Omega) \rightarrow L^2(\Omega)}(v_{n_{j_{k_l}}}^x) \\ I_{H^1(\Omega) \rightarrow L^2(\Omega)}(v_{n_{j_{k_l}}}^y) \\ I_{H^1(\Omega) \rightarrow L^2(\Omega)}(v_{n_{j_{k_l}}}^z) \end{bmatrix} \rightarrow \begin{bmatrix} w^x \\ w^y \\ w^z \end{bmatrix} \text{ as } l \rightarrow \infty. \quad (3.74.p)$$

The lesson is that, from an arbitrary bounded sequence $\{\mathbf{v}_n\}_{n=1}^\infty$ in $H^1(\Omega)^3$, its image $\{I_{H^1(\Omega)^3 \rightarrow L^2(\Omega)^3}(\mathbf{v}_n)\}_{n=1}^\infty$ admits a convergent subsequence in $L^2(\Omega)^3$. Therefore, the embedding from $H^1(\Omega)^3$ into $L^2(\Omega)^3$ is compact.

The meaning of (3.74.i) is just

$$\|I_{H^1(\Omega)^3 \rightarrow L^2(\Omega)^3}(\mathbf{u})\|_{L^2(\Omega)^3} \leq C_e \|\mathbf{u}\|_{H^1(\Omega)^3}, \quad \forall \mathbf{u} \in H^1(\Omega)^3 \quad (3.74.q)$$

We now claim that the auxiliary Hilbert space H from Theorem 3.9 is $L^2(\Omega)^3$, i.e., we make

$$H = L^2(\Omega)^3. \quad (3.74.r)$$

Since $\mathbb{V}_\tau(\Omega) \subset H^1(\Omega)^3$, then (3.74.q) remains valid for all $\mathbf{u} \in \mathbb{V}_\tau(\Omega)$. Then,

$$\mathbb{V}_\tau(\Omega) \hookrightarrow L^2(\Omega)^3, \quad (3.74.s)$$

and hypothesis (viii) is checked.

We already know that the kernel $\mathcal{X}^0$ is a subspace of $\mathcal{X}$, according to (3.69.c). Since $\mathbb{V}_\tau(\Omega)$ has been identified with the original Hilbert space $\mathcal{X}$ in (3.49.a), $\mathcal{X}^0$ is a subspace of $\mathbb{V}_\tau(\Omega)$. When we specialize (3.74.q) to functions in $\mathcal{X}^0$ we conclude that

$$\mathcal{X}^0 \hookrightarrow L^2(\Omega)^3, \quad (3.74.t)$$

and hypothesis (iii) is checked.

The following chain of inclusions is valid:

$$\mathcal{X}^0 \subset \mathbb{V}_\tau(\Omega) \subset H^1(\Omega)^3 \subset L^2(\Omega)^3 = H. \quad (3.74.u)$$

From (3.74.u), we observe that hypothesis (ix) is checked. Therefore, $I_{H^1(\Omega)^3 \rightarrow L^2(\Omega)^3}$, $I_{\mathbb{V}_\tau(\Omega) \rightarrow L^2(\Omega)^3}$, and $I_{\mathcal{X}^0 \rightarrow L^2(\Omega)^3}$ are all identity maps. Particularly,

$$I_{\mathcal{X}^0 \rightarrow L^2(\Omega)^3}(\mathbf{u}) = \mathbf{u}, \quad \forall \mathbf{u} \in \mathcal{X}^0. \quad (3.74.v)$$

In order to show that $I_{\mathcal{X}^0 \rightarrow L^2(\Omega)^3}$ is compact, we take an arbitrary bounded sequence $\{\mathbf{t}_n\}_{n=1}^\infty$ in $\mathcal{X}^0$. Since $\mathcal{X}^0 \subset H^1(\Omega)^3$, the same reasoning from (3.74.j) – (3.74.q) can be applied to $\{\mathbf{t}_n\}_{n=1}^\infty$. The result is that the image of this sequence under $I_{\mathcal{X}^0 \rightarrow L^2(\Omega)^3}$ admits a convergent subsequence in $L^2(\Omega)^3$. In this way, $I_{\mathcal{X}^0 \rightarrow L^2(\Omega)^3}$ is compact. Thus, hypothesis (iv) has been checked.

3.3.6.5 Theorem 3.9, Hypothesis (vi)

We need to show that

$$\inf_{q \in L^2(\Omega) \setminus \{0\}} \sup_{\mathbf{v} \in \mathbb{V}_\tau(\Omega) \setminus \{0\}} \frac{\left| - \int_\Omega q \nabla \cdot \mathbf{v} \right|}{\|\mathbf{v}\|_{H^1(\Omega)^3} \|q\|_{L^2(\Omega)}} \geq \beta_b > 0, \quad (3.75.a)$$

In other words, we need to show that the operator $-\nabla \cdot$ is surjective from $\mathbb{V}_\tau(\Omega)$ onto $L^2(\Omega)$. The argument we developed to show that (3.75.a) is indeed the case is vital for our progress. As such, it will be presented as a theorem.

Theorem 3.14: Surjectivity of $-\nabla \cdot$ – *Let the requirements of Theorem 3.4 be satisfied. Then,*

$$\forall q \in L^2(\Omega) \quad \exists \mathbf{v} \in \mathbb{V}_\tau(\Omega) \quad -\nabla \cdot \mathbf{v} = q. \quad (3.75.b)$$

Proof: Let $q \in L^2(\Omega)$ be arbitrary, and make

$$q_0 = q - \frac{1}{|\Omega|} \int_{\Omega} q. \quad (3.75.c)$$

It is clear that $q_0 \in L_0^2(\Omega)$, discussed in (3.37.c). According to (3.39), there is a $\mathbf{v}_0 \in H_0^1(\Omega)^3$ such that

$$-\nabla \cdot \mathbf{v}_0 = q_0. \quad (3.75.d)$$

Now let

$$\alpha = \frac{1}{|\Omega|} \int_{\Omega} q \quad (3.75.e)$$

We are looking for a function $\mathbf{v}_{\alpha} \in H^1(\Omega)^3$ such that

$$\hat{\mathbf{n}} \times \mathbf{v}_{\alpha}|_{\Gamma} = \mathbf{0} \quad (3.75.f)$$

$$\nabla \cdot \mathbf{v}_{\alpha} = \alpha. \quad (3.75.g)$$

Expression (3.75.f) means that $\mathbf{v}_{\alpha}$ has no tangential components at the boundary Γ . In order to find this $\mathbf{v}_{\alpha}$, we claim that $\mathbf{v}_{\alpha}$ is the gradient of some function ϕ , i.e., we make

$$\mathbf{v}_{\alpha} = \nabla \phi. \quad (3.75.h)$$

From (3.75.g) and (3.75.h), we see that $\nabla \cdot \mathbf{v}_{\alpha} = \nabla \cdot \nabla \phi = \nabla^2 \phi = \alpha$. We also claim that ϕ satisfies homogeneous Dirichlet boundary conditions at Γ . Next, we seek for the solution of the problem

$$\begin{cases} \nabla^2 \phi = \alpha \\ \phi|_{\Gamma} = 0 \end{cases} \quad (3.75.i)$$

which is just an ordinary Poisson equation, whose weak solution is smooth enough to guarantee that $\nabla \phi \in H^1(\Omega)^3$, according to Theorem 3 in Chapter 6 from [Evans, 2010].

Since ϕ is constant at the boundary Γ , it defines a level curve there. Therefore, $\nabla \phi$ is normal to Γ , i.e., $\nabla \phi = f \hat{\mathbf{n}}$, where f is a scalar function of the points located on Γ .

We learn that $\hat{\mathbf{n}} \times \nabla \phi|_{\Gamma} = \hat{\mathbf{n}} \times f \hat{\mathbf{n}}|_{\Gamma} = \mathbf{0}$, which validates the choice of $\nabla \phi$ for $\mathbf{v}_{\alpha}$, according to requirement (3.75.f).

We now form the vector

$$\mathbf{v} = \mathbf{v}_0 - \mathbf{v}_{\alpha}. \quad (3.75.j)$$

It is clear that

$$-\nabla \cdot \mathbf{v} = -\nabla \cdot \mathbf{v}_0 + \nabla \cdot \mathbf{v}_{\alpha} = q_0 + \alpha, \quad (3.75.k)$$

according to (3.75.d) and (3.75.g). But from (3.75.c) and (3.75.e), we get that $q_0 = q - \alpha$. Consequently, (3.75.k) implies that

$$-\nabla \cdot \mathbf{v} = q. \quad (3.75.l)$$

Since $\mathbf{v}_\alpha \in H^1(\Omega)^3$ and $\mathbf{v}_0 \in H_0^1(\Omega)^3$, then $\mathbf{v} \in H^1(\Omega)^3$. Moreover,

$$\hat{\mathbf{n}} \times \mathbf{v}|_\Gamma = \hat{\mathbf{n}} \times \mathbf{v}_0|_\Gamma + \hat{\mathbf{n}} \times \mathbf{v}_\alpha|_\Gamma = \mathbf{0}, \quad (3.75.m)$$

because all components of $\mathbf{v}_0$ are zero at Γ and because of (3.75.f). As $\mathbf{v} \in H^1(\Omega)^3$ and $\hat{\mathbf{n}} \times \mathbf{v}|_\Gamma = \mathbf{0}$, then $\mathbf{v} \in \mathbb{V}_\tau(\Omega)$.

So we have been able to show that, given an arbitrary $q \in L^2(\Omega)$, there is a $\mathbf{v} \in \mathbb{V}_\tau(\Omega)$ such that $-\nabla \cdot \mathbf{v} = q$. In other words, $-\nabla \cdot$ is surjective from $\mathbb{V}_\tau(\Omega)$ onto $L^2(\Omega)$.

■

A more or less ‘physical’ interpretation of Theorem 3.14 goes like this: Suppose Ω is a hollow metallic cavity, and let q be a square-summable charge density within Ω . Then, there is a field $\mathbf{E}$ such that $-\nabla \cdot \mathbf{E} = q$.

If we write the action of the sesquilinear form b on arbitrary elements $\mathbf{v} \in \mathbb{V}_\tau(\Omega)$ and $q \in L^2(\Omega)$, we get, after the identification (3.49):

$$b(\mathbf{v}, q) = \langle B\mathbf{v}, q \rangle_{L_0^2(\Omega)^*, L_0^2(\Omega)} = - \int_\Omega q^* \nabla \cdot \mathbf{v}, \quad (3.75.n)$$

i.e., we are able to see that $B\mathbf{v} = -\nabla \cdot \mathbf{v}$. So the operator B induced by the sesquilinear form b is indeed the negative of the divergence operator, i.e., $B = -\nabla \cdot$. Since $-\nabla \cdot$ is surjective, then B is surjective. Theorem 3.3 says that B being surjective is equivalent to the fact that there is a $\beta_b > 0$ such that

$$\inf_{y \in \mathcal{Y} \setminus \{0\}} \sup_{x \in \mathcal{X} \setminus \{0\}} \frac{|b(x, y)|}{\|x\|_{\mathcal{X}} \|y\|_{\mathcal{Y}}} \geq \beta_b. \quad (3.75.o)$$

When we make the identification (3.49), we conclude that for the scattering problem it is true that

$$\inf_{q \in L^2(\Omega) \setminus \{0\}} \sup_{\mathbf{v} \in \mathbb{V}_\tau(\Omega) \setminus \{0\}} \frac{\left| - \int_\Omega q \nabla \cdot \mathbf{v} \right|}{\|\mathbf{v}\|_{H^1(\Omega)^3} \|q\|_{L^2(\Omega)}} \geq \beta_b, \quad (3.75.p)$$

which is nothing else than (3.75.a). Therefore, hypothesis (vi) has been checked.

The two remaining conditions (v) and (vii) are more difficult to check. They depend on the explicit form of the PML tensor $\bar{\mathbf{A}}$.

3.3.6.6 PML II: The PML tensor

Thus far, the only information we have concerning the PML tensor is that it has the form

$$\bar{\bar{\Lambda}} = \Lambda_x \hat{x} \otimes \hat{x} + \Lambda_y \hat{y} \otimes \hat{y} + \Lambda_z \hat{z} \otimes \hat{z}, \quad (3.76.a)$$

presented in (2.113). Moreover, from the discussion in Section 2.2.3.3, in order for the weak solutions to make sense, we discovered that the components of $\bar{\bar{\Lambda}}$ must be elements of $L^\infty(\Omega)$. For our purposes, it means that there are positive constants M_x , M_y and M_z such that

$$|\Lambda_x(\mathbf{x})| \leq M_x, \quad \forall \mathbf{x} \in \Omega \quad (3.76.b)$$

$$|\Lambda_y(\mathbf{x})| \leq M_y, \quad \forall \mathbf{x} \in \Omega \quad (3.76.c)$$

$$|\Lambda_z(\mathbf{x})| \leq M_z, \quad \forall \mathbf{x} \in \Omega. \quad (3.76.d)$$

In this way, Λ_M in (2.176) can be taken as

$$\Lambda_M = \max \{M_x, M_y, M_z\}. \quad (3.76.e)$$

The components of the PML tensor are complex quantities, so we write them as

$$\Lambda_x(\mathbf{x}) = \beta_x(\mathbf{x}) + j\delta_x(\mathbf{x}), \quad (3.76.f)$$

$$\Lambda_y(\mathbf{x}) = \beta_y(\mathbf{x}) + j\delta_y(\mathbf{x}), \quad (3.76.g)$$

$$\Lambda_z(\mathbf{x}) = \beta_z(\mathbf{x}) + j\delta_z(\mathbf{x}). \quad (3.76.h)$$

In this representation, β_x , β_y , β_z and δ_x , δ_y , δ_z are all real functions of the position $\mathbf{x} \in \Omega$.

We now make two extra requirements: There is a positive constant $\beta > 0$ such that

$$\beta_x, \beta_y, \beta_z > \beta, \quad \forall \mathbf{x} \in \Omega; \quad (3.76.i)$$

Moreover, the imaginary part of the components should be nonnegative, i.e.,

$$\delta_x, \delta_y, \delta_z \geq 0, \quad \forall \mathbf{x} \in \Omega. \quad (3.76.j)$$

After we have set up the requirements for the PML, we ask: Is there a rectangular PML obeying the form (3.76.a) which satisfy the conditions (3.76.b) – (3.76.d), (3.76.i) and (3.76.j)?

In a sense, yes. We consider a PML originally developed for scalar waves [Bermúdez *et al.*, 2004], [Bermúdez *et al.*, 2007], [Bermúdez *et al.*, 2010] and which has been successfully applied to the FEM analysis of mechanical waves [Ham and Bathe, 2012]. If we make some adjustments, we discover that the resulting PML can be applied to our electromagnetic scattering problem, while at the same time satisfying all the above requirements.

First of all, suppose the outer boundary Γ_o is the surface of a rectangular region

defined by $X_1 \leq x \leq X_2$, $Y_1 \leq y \leq Y_2$ and $Z_1 \leq z \leq Z_2$. Our computational domain Ω consists of this box with the volume occupied by the PEC scatterer removed. The removal of the scatterer introduces an interior surface Γ_1 , which is obviously the surface of the PEC object.

The PML is just a layer of width w_{PML} , which we assume is the same for the three directions. The value of w_{PML} must be chosen in such a way that the surface of the PEC scatterer lies entirely outside the PML layer. Next, given a point $\mathbf{x} = [x, y, z]^T$ in the domain Ω , it defines three distances to the x , y , and z walls comprising the outer boundary Γ_o . They are given by

$$d_x = \min\{(x - X_1), (X_2 - x)\} \quad (3.76.k)$$

$$d_y = \min\{(y - Y_1), (Y_2 - y)\} \quad (3.76.l)$$

$$d_z = \min\{(z - Z_1), (Z_2 - z)\}. \quad (3.76.m)$$

From the distances above, we calculate three auxiliary quantities as

$$\gamma_x = \begin{cases} 1 - j \frac{1}{k_0 d_x}, & \text{if } d_x < w_{PML} \\ 1, & \text{if } d_x \geq w_{PML} \end{cases} \quad (3.76.n)$$

$$\gamma_y = \begin{cases} 1 - j \frac{1}{k_0 d_y}, & \text{if } d_y < w_{PML} \\ 1, & \text{if } d_y \geq w_{PML} \end{cases} \quad (3.76.o)$$

$$\gamma_z = \begin{cases} 1 - j \frac{1}{k_0 d_z}, & \text{if } d_z < w_{PML} \\ 1, & \text{if } d_z \geq w_{PML}, \end{cases} \quad (3.76.p)$$

where k_0 is the free-space wavenumber. The components of the PML tensor are then calculated as

$$\Lambda_x = \frac{1}{\gamma_x^2}, \quad (3.76.q)$$

$$\Lambda_y = \frac{1}{\gamma_y^2}, \quad (3.76.r)$$

$$\Lambda_z = \frac{1}{\gamma_z^2}. \quad (3.76.s)$$

Let us concentrate on the expression for Λ_x within the PML, i.e., when $d_x < w_{PML}$. When worked out, we see that

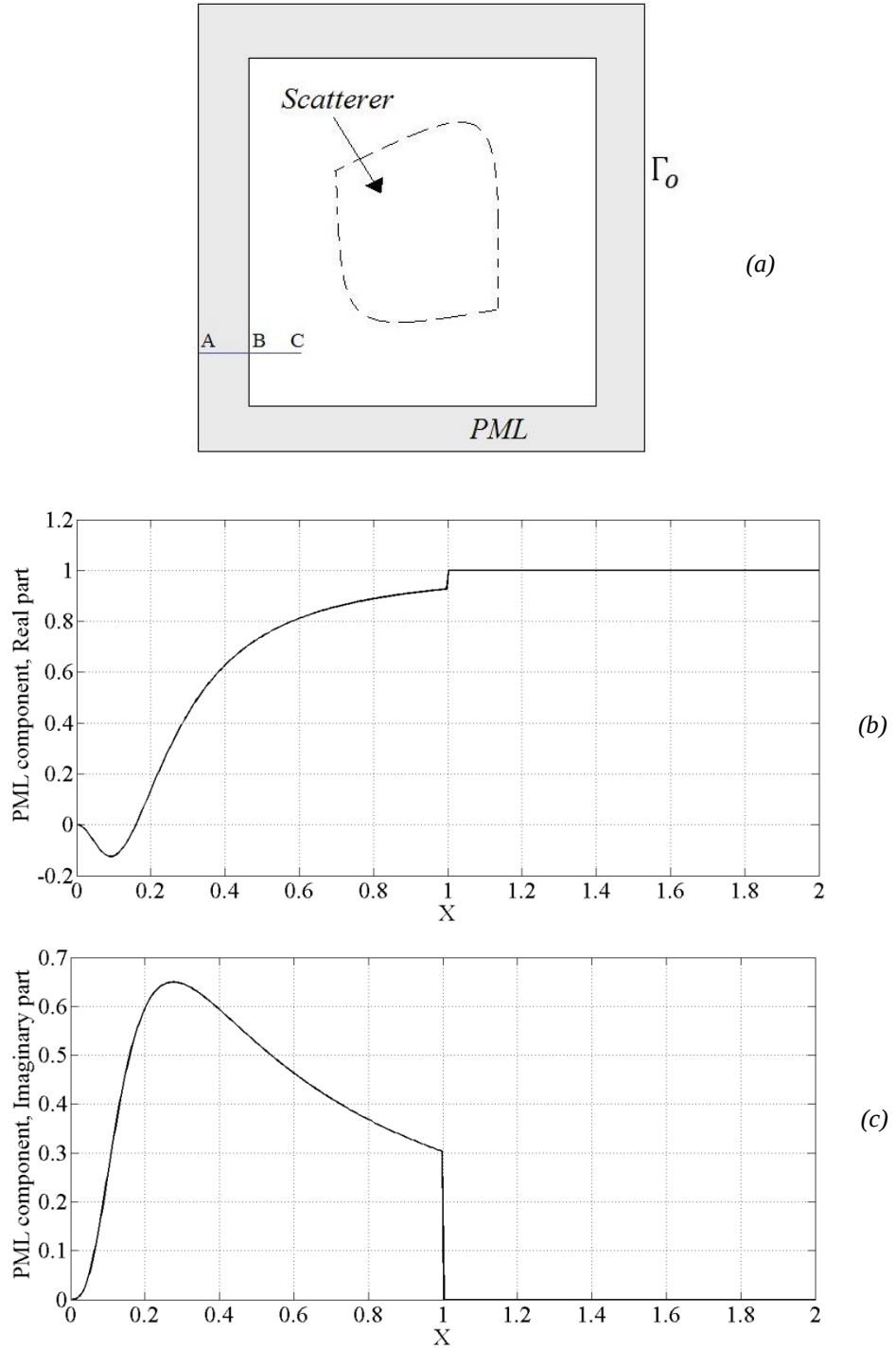


Fig. 3.2. (a) A sample of the computational domain Ω , showing the PML layer in gray. The distances AB and BC are the same, and equal to the PML width w_{PML} . (b) The real part of the component Λ_x of the PML tensor along the path ABC. The point A corresponds to $x = 0$, B to $x = 1$ and C to $x = 2$. (c) The imaginary part. Figures (b) and (c) illustrate the unperturbed component, described by (3.76.q). Notice how the real part of Λ_x becomes negative for small x (i.e., close to the boundary Γ_o).

$$\Lambda_x = \frac{(k_0 d_x)^2 [(k_0 d_x)^2 - 1]}{[(k_0 d_x)^2 - 1]^2 + 4(k_0 d_x)^2} + j \frac{2(k_0 d_x)^3}{[(k_0 d_x)^2 - 1]^2 + 4(k_0 d_x)^2}. \quad (3.76.t)$$

From the expression above, it can be observed that the imaginary part within the PML is always positive, i.e., $\delta_x \geq 0$. Furthermore, since *outside* the PML (when $d_x \geq w_{PML}$)

the factor γ_x is equal to 1, the imaginary part of Λ_x is zero there. Then, it is true that for all points in the domain Ω , $\delta_x \geq 0$. In this way, (3.76.j) is satisfied.

There is an issue with the real part of Λ_x . Expression (3.76.t) reveals that it becomes negative when $k_0 d_x < 1$, i.e., for distances $d_x < 1/k_0$. However, this distance $1/k_0$ is very close to the outer boundary Γ_o . When the fields reach this distance, they will already be very well attenuated, so that their amplitudes will be, for any practical purposes, essentially zero. The behavior of the real and imaginary parts of Λ_x is illustrated in Fig. 3.2.

As the real part of Λ_x becomes negative for some points at the interior of the PML, (3.76.i) cannot be satisfied.

However, there is a way out. Since this limit distance $1/k_0$ is very close Γ_o , the fields will be essentially zero by the time they come this close to the outer boundary. So we argue that there will be no significant trouble if Λ_x is perturbed in such a way that its real part does not become negative for very small distances.

The idea goes as follows: We consider a ‘threshold’ distance

$$d_{th} = \frac{5}{4} \left(\frac{1}{k_0} \right). \quad (3.76.u)$$

The distance d_{th} is slightly larger than $1/k_0$. But, as it can be verified from Fig. 3.2, the real part of Λ_x is positive there. Then we perturb Λ_x according to the rule: If d_x is larger than d_{th} , then the original Λ_x in (3.76.q) is kept. If d_x is smaller than d_{th} , then Λ_x is just the value of Λ_x calculated at d_{th} . In other words, we consider the perturbed version of Λ_x as

$$\Lambda_x(d_x) = \begin{cases} \frac{1}{\gamma_x(d_x)^2}, & \text{if } d_x \geq d_{th} \\ \frac{1}{\gamma_x(d_{th})^2}, & \text{if } d_x < d_{th}, \end{cases} \quad (3.76.v)$$

where $\gamma_x(d_x)$ means “the γ_x from (3.76.n) evaluated for d_x ”, and $\gamma_x(d_{th})$ means “the γ_x from (3.76.n) evaluated for d_{th} ”. The perturbed version (3.76.v) is illustrated in Fig. 3.3.

Figure 3.3 reveals that the real part of the perturbed Λ_x never reaches zero. Therefore, it satisfies (3.76.i). Moreover, the imaginary part of the perturbed Λ_x is always larger than or equal to zero, and so it satisfies (3.76.j). And finally, it is obvious from Fig. 3.3 that both real and imaginary parts of Λ_x are bounded, and so (3.76.b) holds true.

When this reasoning is applied to Λ_y and Λ_z , we arrive at the same conclusions. And in this way, we can answer affirmatively to the question concerning the existence of a rectangular PML which obeys the form (3.76.a) and which satisfy the conditions (3.76.b) – (3.76.d), (3.76.i) and (3.76.j).

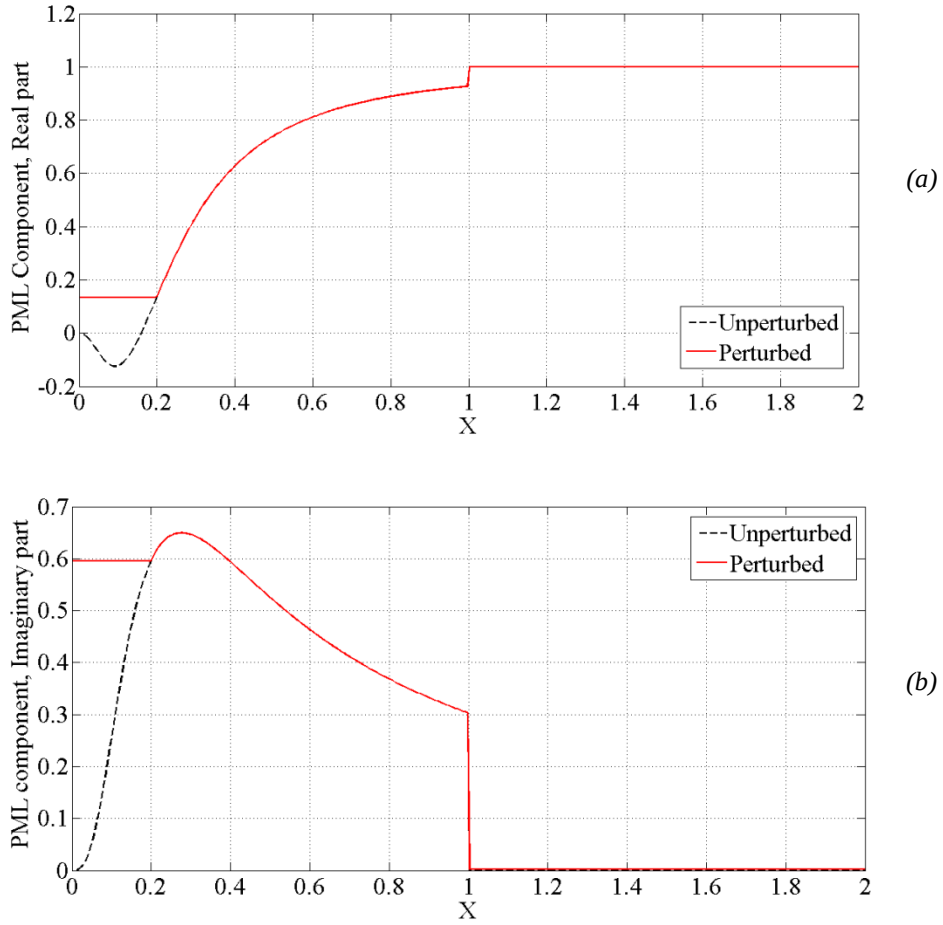


Fig. 3.3. The perturbed PML component in (3.76.v) along the path ABC in Fig. 3.2.a. It is shown together with the unperturbed component, so that a comparison can be made. (a) Real part. (b) Imaginary part.

3.3.6.7 Theorem 3.9, Hypothesis (vii)

In order to check the hypothesis (vii) in Theorem 3.9, we need to show that the solution to the homogeneous problem

$$\begin{aligned} & \text{Find } w \in \mathcal{X}^0 \text{ such that} \\ & a(w, v) = 0, \quad \forall v \in \mathcal{X}^0 \end{aligned} \quad (3.77.a)$$

is the zero element. After the identification (3.49), (3.77.a) becomes

$$\begin{aligned} & \text{Find } \mathbf{w} \in \mathcal{X}^0 \text{ such that} \\ & a(\mathbf{w}, \mathbf{v}) = \int_{\Omega} (\bar{\mathbf{\Lambda}} \cdot \nabla \mathbf{w}) : \nabla \mathbf{v}^* - \int_{\Omega} k_0^2 \mathbf{w} \cdot \mathbf{v}^* = 0, \quad \forall \mathbf{v} \in \mathcal{X}^0, \end{aligned} \quad (3.77.b)$$

where $\mathcal{X}^0 \subset \mathbb{V}_{\tau}(\Omega) \subset H^1(\Omega)^3$.

Suppose that $\mathbf{w}$ is a nonzero solution to (3.77.b). Then it is true that

$$\int_{\Omega} (\bar{\Lambda} \cdot \nabla \mathbf{w}) : \nabla \mathbf{v}^* = k_0^2 \int_{\Omega} \mathbf{w} \cdot \mathbf{v}^*, \quad \forall \mathbf{v} \in \mathcal{X}^0, \quad (3.77.c)$$

i.e., $\mathbf{w}$ is one of the solutions to the eigenproblem

Find $\mathbf{w}_{\sigma} \in \mathcal{X}^0$ such that

$$\int_{\Omega} (\bar{\Lambda} \cdot \nabla \mathbf{w}_{\sigma}) : \nabla \mathbf{v}^* = \sigma \int_{\Omega} \mathbf{w}_{\sigma} \cdot \mathbf{v}^*, \quad \forall \mathbf{v} \in \mathcal{X}^0. \quad (3.77.d)$$

If we denote by Σ the set of all eigenvalues associated to (3.77.d), then we learn that, if (3.77.b) has a nonzero solution, then $k_0^2 \in \Sigma$ and $\mathbf{w}$ is one of the eigenfunctions associated to (3.77.d).

We have showed that, if $\mathbf{w} \neq \mathbf{0}$, then $k_0^2 \in \Sigma$. Conversely, we can conclude that

$$\text{If } k_0^2 \notin \Sigma \text{ then } \mathbf{w} = \mathbf{0}. \quad (3.77.e)$$

In other words, if k_0^2 is not an eigenvalue, then the solution $\mathbf{w}$ to the homogeneous problem (3.77.b) is the zero element $\mathbf{0}$.

So in order to satisfy hypothesis (vii) in Theorem 3.9, we must make sure that k_0^2 is not one of the eigenvalues associated with the problem stated in the kernel $\mathcal{X}^0$. If we want the solution of the Helmholtz equation (and variants thereof) to exist and be unique, then we must stay away from the eigenvalues. Or said in another way, the solution to the Helmholtz equation exists, provided the wavenumber we are interested in is such that k_0^2 is not an eigenvalue. This kind of result is common in the literature [Evans, 2010], [Ihlenburg, 1998]. (This issue plagues the well-posedness of the Helmholtz equation in all scenarios; it is not restricted to the situation described in this thesis.)

The conclusion is that we cannot choose any value for k_0 . We may ask: Does it imply a loss of freedom when working with the Helmholtz equation? How can we find out if k_0^2 is an eigenvalue or not, without having to solve an eigenproblem first?

The fact that the problem (3.77.b) incorporates a PML tensor with complex entries may provide a plausible answer. Suppose we want to solve the eigenproblem (3.77.d). Let σ be one of the eigenvalues, together with its associated eigenfunction $\mathbf{w}_{\sigma}$. Since the testing functions are taken from $\mathcal{X}^0$, and we know that $\mathbf{w}_{\sigma} \in \mathcal{X}^0$, then we make $\mathbf{v} = \mathbf{w}_{\sigma}$ and get

$$\int_{\Omega} (\bar{\Lambda} \cdot \nabla \mathbf{w}_{\sigma}) : \nabla \mathbf{w}_{\sigma}^* = \sigma \int_{\Omega} |\mathbf{w}_{\sigma}|^2. \quad (3.77.f)$$

The first integral in (3.77.f) can be expanded according to (2.171). If we represent $\mathbf{w}_{\sigma}$ as $\mathbf{w}_{\sigma} = [w_{\sigma}^x, w_{\sigma}^y, w_{\sigma}^z]^T$, (3.77.f) becomes:

$$\begin{aligned}
& \int_{\Omega} \Lambda_x \left| \frac{\partial w_{\sigma}^x}{\partial x} \right|^2 + \Lambda_y \left| \frac{\partial w_{\sigma}^x}{\partial y} \right|^2 + \Lambda_z \left| \frac{\partial w_{\sigma}^x}{\partial z} \right|^2 + \\
& \int_{\Omega} \Lambda_x \left| \frac{\partial w_{\sigma}^y}{\partial x} \right|^2 + \Lambda_y \left| \frac{\partial w_{\sigma}^y}{\partial y} \right|^2 + \Lambda_z \left| \frac{\partial w_{\sigma}^y}{\partial z} \right|^2 + \\
& \int_{\Omega} \Lambda_x \left| \frac{\partial w_{\sigma}^z}{\partial x} \right|^2 + \Lambda_y \left| \frac{\partial w_{\sigma}^z}{\partial y} \right|^2 + \Lambda_z \left| \frac{\partial w_{\sigma}^z}{\partial z} \right|^2 = \sigma \int_{\Omega} |w_{\sigma}|^2.
\end{aligned} \tag{3.77.g}$$

When inspecting (3.77.g), one observes that the left side will probably be complex, because the squared derivatives (within bars) are all positive, and also because of (3.76.j), which says that the imaginary parts of Λ_x , Λ_y and Λ_z are positive. On the other hand, the integral

$$\int_{\Omega} |w_{\sigma}|^2 \tag{3.77.h}$$

is a positive real number. So the left side of (3.77.f) may be complex, whereas the integral at the right side in (3.77.f) is a real number. The only way to avoid a contradiction is to allow the eigenvalue σ to be a complex number.

We concluded that, if $\sigma \in \Sigma$, then $\text{Im}\{\sigma\} \neq 0$. Conversely, we can conclude that if $\text{Im}\{\sigma\} = 0$ (i.e., σ is a real number), then $\sigma \notin \Sigma$ (i.e., σ is not an eigenvalue). Since waves in the free-space are described by real wavenumbers, for any choice we make for k_0 , k_0^2 will always be a real number, and therefore, will not be an eigenvalue.

In a sense, we showed that there is a *high probability* that for any choice of k_0 , the solution to (3.77.b) will be the zero element. We say it is probable because in order to make an assertion, we need to investigate the influence of the complex PML tensor on the spectral properties of problem (3.77.b), i.e., we need a formal proof that all eigenvalues of (3.77.d) are complex. Even though it constitutes a very interesting problem, it falls outside the scope of this thesis. However, if we assume from the start that k_0^2 is not an eigenvalue, then hypothesis (vii) in Theorem 3.9 is satisfied.

3.3.6.8 Theorem 3.9, Hypothesis (v)

The only hypothesis to be verified is (v). After the identification (3.49), it concerns the existence of constants $\eta > 0$ and $\kappa_0 \geq 0$ such that

$$\text{Re}\{a(\mathbf{u}, \mathbf{u})\} + \kappa_0 \|I_{\mathcal{X}^0 \rightarrow L^2(\Omega)^3}(\mathbf{u})\|_{L^2(\Omega)^3}^2 \geq \eta \|\mathbf{u}\|_{\mathcal{X}^0}^2, \quad \forall \mathbf{u} \in \mathcal{X}^0. \tag{3.78.a}$$

From (3.74.u) it is true that $\mathcal{X}^0 \subset L^2(\Omega)^3$, and we concluded that $I_{\mathcal{X}^0 \rightarrow L^2(\Omega)^3}$ is just the identity map, i.e., $I_{\mathcal{X}^0 \rightarrow L^2(\Omega)^3}(\mathbf{u}) = \mathbf{u}$. In this way (3.78.a) becomes

$$\text{Re}\{a(\mathbf{u}, \mathbf{u})\} + \kappa_0 \|\mathbf{u}\|_{L^2(\Omega)^3}^2 \geq \eta \|\mathbf{u}\|_{H^1(\Omega)^3}^2, \quad \forall \mathbf{u} \in \mathcal{X}^0, \tag{3.78.b}$$

because since $\mathcal{X} = \mathbb{V}_\tau(\Omega)$ [according to the identification (3.49)], and as $\mathbb{V}_\tau(\Omega) \subset H^1(\Omega)^3$, the norm $\|\cdot\|_{\mathcal{X}}$ is just the norm $\|\cdot\|_{H^1(\Omega)^3}$.

We need to find constants $\eta > 0$ and $\kappa_0 \geq 0$ such that (3.78.b) is true. (In other words, proving (3.78.b) is our goal.)

After the substitution of both $\mathbf{w}$ and $\mathbf{v}$ by an arbitrary $\mathbf{u} \in \mathbb{V}_\tau(\Omega)$, the sesquilinear form $a(\mathbf{w}, \mathbf{v})$ in (3.73.b) becomes

$$a(\mathbf{u}, \mathbf{u}) = \int_{\Omega} (\bar{\bar{\Lambda}} \cdot \nabla \mathbf{u}) : \nabla \mathbf{u}^* - \int_{\Omega} k_0^2 \mathbf{u} \cdot \mathbf{u}^*, \quad \forall \mathbf{u} \in \mathbb{V}_\tau(\Omega). \quad (3.78.c)$$

We can rewrite (3.78.c) as

$$a(\mathbf{u}, \mathbf{u}) + k_0^2 \int_{\Omega} \mathbf{u} \cdot \mathbf{u}^* = \int_{\Omega} (\bar{\bar{\Lambda}} \cdot \nabla \mathbf{u}) : \nabla \mathbf{u}^*, \quad \forall \mathbf{u} \in \mathbb{V}_\tau(\Omega). \quad (3.78.d)$$

When we consider only the real part of (3.78.d), we get

$$\operatorname{Re}\{a(\mathbf{u}, \mathbf{u})\} + k_0^2 \int_{\Omega} \mathbf{u} \cdot \mathbf{u}^* = \operatorname{Re}\left\{ \int_{\Omega} (\bar{\bar{\Lambda}} \cdot \nabla \mathbf{u}) : \nabla \mathbf{u}^* \right\}, \quad \forall \mathbf{u} \in \mathbb{V}_\tau(\Omega), \quad (3.78.e)$$

because the second integral in (3.78.d) is a real number. When we expand the integral in the right side of (3.78.d) as in (3.77.g), its real part

$$\begin{aligned} & \operatorname{Re}\left\{ \int_{\Omega} \Lambda_x \left| \frac{\partial u_x}{\partial x} \right|^2 + \Lambda_y \left| \frac{\partial u_x}{\partial y} \right|^2 + \Lambda_z \left| \frac{\partial u_x}{\partial z} \right|^2 + \right. \\ & \quad \int_{\Omega} \Lambda_x \left| \frac{\partial u_y}{\partial x} \right|^2 + \Lambda_y \left| \frac{\partial u_y}{\partial y} \right|^2 + \Lambda_z \left| \frac{\partial u_y}{\partial z} \right|^2 + \\ & \quad \left. \int_{\Omega} \Lambda_x \left| \frac{\partial u_z}{\partial x} \right|^2 + \Lambda_y \left| \frac{\partial u_z}{\partial y} \right|^2 + \Lambda_z \left| \frac{\partial u_z}{\partial x} \right|^2 \right\} \end{aligned} \quad (3.78.f)$$

is indeed equal to

$$\begin{aligned} & \int_{\Omega} \beta_x \left| \frac{\partial u_x}{\partial x} \right|^2 + \beta_y \left| \frac{\partial u_x}{\partial y} \right|^2 + \beta_z \left| \frac{\partial u_x}{\partial z} \right|^2 + \\ & \quad \int_{\Omega} \beta_x \left| \frac{\partial u_y}{\partial x} \right|^2 + \beta_y \left| \frac{\partial u_y}{\partial y} \right|^2 + \beta_z \left| \frac{\partial u_y}{\partial z} \right|^2 + \\ & \quad \int_{\Omega} \beta_x \left| \frac{\partial u_z}{\partial x} \right|^2 + \beta_y \left| \frac{\partial u_z}{\partial y} \right|^2 + \beta_z \left| \frac{\partial u_z}{\partial x} \right|^2, \end{aligned} \quad (3.78.g)$$

because all quantities between the bars are real numbers. The quantities β_x , β_y and β_z are the real parts of the PML tensor components Λ_x , Λ_y and Λ_z , respectively. Since all terms in (3.78.g) are positive, from (3.76.i) we conclude that

$$\operatorname{Re} \left\{ \int_{\Omega} (\bar{\Lambda} \cdot \nabla \mathbf{u}) : \nabla \mathbf{u}^* \right\} \geq \beta \int_{\Omega} \nabla \mathbf{u} : \nabla \mathbf{u}^*. \quad (3.78.h)$$

From (3.78.e), (3.78.h), (2.36) and (2.39), we get

$$\operatorname{Re}\{a(\mathbf{u}, \mathbf{u})\} + k_0^2 \|\mathbf{u}\|_{L^2(\Omega)^3}^2 \geq \beta \|\mathbf{u}\|_{H^1(\Omega)^3}^2, \quad \forall \mathbf{u} \in \mathbb{V}_{\tau}(\Omega). \quad (3.78.i)$$

Expression (3.78.i) resembles our goal (3.78.b). The difference is that the seminorm figures at the right of (3.78.i), whereas in the right side of (3.78.b) there is a norm.

In order to replace the seminorm $\|\mathbf{u}\|_{H^1(\Omega)^3}$ in (3.78.i) by the norm $\|\mathbf{u}\|_{H^1(\Omega)^3}$, we need the Poincaré inequality from Theorem 3.2.

According to (2.164), the space $\mathbb{V}_{\tau}(\Omega)$ in which the electric field is sought is

$$\mathbb{V}_{\tau}(\Omega) = \{\mathbf{v} \in H^1(\Omega)^3 \mid \boldsymbol{\gamma}_t \mathbf{v} = \mathbf{0}\}, \quad (3.78.j)$$

which means that $(\hat{\mathbf{n}} \times \mathbf{v})|_{\Gamma} = \mathbf{0}$. In other words, the tangential components of the elements from $\mathbb{V}_{\tau}(\Omega)$ are zero at the boundary Γ , which is formed by the outer boundary Γ_o and by the scatterer surface Γ_1 . Therefore it is true that

$$(\hat{\mathbf{n}} \times \mathbf{u})|_{\Gamma_o} = \mathbf{0}, \quad \forall \mathbf{u} \in \mathbb{V}_{\tau}(\Omega). \quad (3.78.k)$$

Moreover, an arbitrary element $\mathbf{u}$ from $\mathbb{V}_{\tau}(\Omega)$ is described by its three Cartesian components as $\mathbf{u} = [u_x, u_y, u_z]^T$.

As stated earlier in section 3.3.6.6, the outer boundary Γ_o is the surface of a rectangular box defined by $X_1 \leq x \leq X_2$, $Y_1 \leq y \leq Y_2$ and $Z_1 \leq z \leq Z_2$. Given an arbitrary $\mathbf{u} \in \mathbb{V}_{\tau}(\Omega)$, let us concentrate first on its component u_x . Since the tangential components of $\mathbf{u}$ are zero on Γ_o , it implies that u_x is zero over the set

$$S = \{\mathbf{x} \mid X_1 \leq x \leq X_2, y = Y_1, Z_1 \leq z \leq Z_2\} \cup \quad (3.78.l)$$

$$\{\mathbf{x} \mid X_1 \leq x \leq X_2, y = Y_2, Z_1 \leq z \leq Z_2\} \cup$$

$$\{\mathbf{x} \mid X_1 \leq x \leq X_2, Y_1 \leq y \leq Y_2, z = Z_1\} \cup$$

$$\{\mathbf{x} \mid X_1 \leq x \leq X_2, Y_1 \leq y \leq Y_2, z = Z_2\}.$$

The set S in (3.78.l) is just the four faces from Γ_o which are parallel to the x -axis. It is clear that S has a positive measure (i.e., its area is different from zero). Therefore, we can say that $u_x \in H^1(\Omega)$ and that u_x vanishes on a non-zero measure subset S of the boundary Γ . According to the terminology of Theorem 3.2, these are just the

requirements for u_x to be an element of $H_S^1(\Omega)$. From the same theorem, we conclude that

$$\|u_x\|_{L^2(\Omega)}^2 \leq c_\Omega \|\nabla u_x\|_{L^2(\Omega)^3}^2. \quad (3.78.m)$$

The same reasoning can be extended to the other components u_y and u_z . (Of course, by considering different subsets S of Γ_o). We get similar conclusions:

$$\|u_y\|_{L^2(\Omega)}^2 \leq c_\Omega \|\nabla u_y\|_{L^2(\Omega)^3}^2, \quad (3.78.n)$$

$$\|u_z\|_{L^2(\Omega)}^2 \leq c_\Omega \|\nabla u_z\|_{L^2(\Omega)^3}^2. \quad (3.78.o)$$

When we sum the last three inequalities, we arrive at

$$(3.78.p)$$

$$\|u_x\|_{L^2(\Omega)}^2 + \|u_y\|_{L^2(\Omega)}^2 + \|u_z\|_{L^2(\Omega)}^2 \leq c_\Omega \left(\|\nabla u_x\|_{L^2(\Omega)^3}^2 + \|\nabla u_y\|_{L^2(\Omega)^3}^2 + \|\nabla u_z\|_{L^2(\Omega)^3}^2 \right)$$

From (2.31), the left side in (3.78.p) is nothing else than $\|\mathbf{u}\|_{L^2(\Omega)^3}^2$. And from (2.34), it is evident that the right side in (3.78.p) is $c_\Omega |\mathbf{u}|_{H^1(\Omega)^3}^2$. Therefore,

$$\|\mathbf{u}\|_{L^2(\Omega)^3}^2 \leq c_\Omega |\mathbf{u}|_{H^1(\Omega)^3}^2. \quad (3.78.q)$$

If we add $|\mathbf{u}|_{H^1(\Omega)^3}^2$ to both sides in (3.78.q), and then consider (2.42), we see that

$$\|\mathbf{u}\|_{H^1(\Omega)^3}^2 = \|\mathbf{u}\|_{L^2(\Omega)^3}^2 + |\mathbf{u}|_{H^1(\Omega)^3}^2 \leq (c_\Omega + 1) |\mathbf{u}|_{H^1(\Omega)^3}^2, \quad (3.78.r)$$

which readily implies that

$$|\mathbf{u}|_{H^1(\Omega)^3}^2 \geq \frac{1}{(c_\Omega + 1)} \|\mathbf{u}\|_{H^1(\Omega)^3}^2. \quad (3.78.s)$$

It is now time to get back to (3.78.i); the information provided by (3.78.s) allows us to rewrite it as

$$\operatorname{Re}\{a(\mathbf{u}, \mathbf{u})\} + k_0^2 \|\mathbf{u}\|_{L^2(\Omega)^3}^2 \geq \frac{\beta}{(c_\Omega + 1)} \|\mathbf{u}\|_{H^1(\Omega)^3}^2, \quad \forall \mathbf{u} \in \mathbb{V}_\tau(\Omega). \quad (3.78.t)$$

The inequality (3.78.t) is of great importance. It means that we have managed to show that the sesquilinear form a obeys some kind of ‘weak’ coercivity in the whole space $\mathbb{V}_\tau(\Omega)$, not just on the kernel $\mathcal{X}^0$. Since $\mathcal{X}^0 \subset \mathbb{V}_\tau(\Omega)$, (3.78.t) implies that

$$\operatorname{Re}\{a(\mathbf{u}, \mathbf{u})\} + k_0^2 \|\mathbf{u}\|_{L^2(\Omega)^3}^2 \geq \frac{\beta}{(c_\Omega + 1)} \|\mathbf{u}\|_{H^1(\Omega)^3}^2, \quad \forall \mathbf{u} \in \mathcal{X}^0, \quad (3.78.u)$$

which is just our goal with κ_0 identified with k_0^2 (which is obviously larger than or equal to zero) and with η identified with $\beta/(c_\Omega + 1)$ [which is larger than zero, due to

(3.76.i)]. Since we know that $I_{\mathcal{X}^0 \rightarrow L^2(\Omega)^3}(\mathbf{u}) = \mathbf{u}$ according to (3.74.v), then it follows that

$$\operatorname{Re}\{a(\mathbf{u}, \mathbf{u})\} + k_0^2 \|I_{\mathcal{X}^0 \rightarrow H}(\mathbf{u})\|_{L^2(\Omega)^3}^2 \geq \frac{\beta}{(c_\Omega + 1)} \|\mathbf{u}\|_{H^1(\Omega)^3}^2, \quad \forall \mathbf{u} \in \mathcal{X}^0, \quad (3.78.v)$$

In this way, hypothesis (v) has been checked.

3.3.7 Concluding remarks

In this section, we provided a theoretical foundation for the well-posedness of the scattering system (3.48). The result is codified into a key theorem (Theorem 3.9), which somehow merges the traditional Babuska-Brezzi theory of mixed formulations and the Fredholm Alternative for non-coercive forms. In order for the theorem to be valid, a total of nine hypotheses need to be satisfied. Fortunately, we have managed to show that each one of them holds true when specialized to the function spaces of our problem.

In what regards the theoretical aspects of this thesis, we are done. Once the theory has been established, the transition to the discrete setting will be very smooth.

Chapter 4

The discretization process

This chapter essentially deals with the discretization process of the scattering system (3.48).

In the first section, we study the extension of Theorem 3.9 to finite-dimensional subspaces. The analysis will be applied to the ‘specialized’ setting of the scattering system.

After all hypotheses are considered, in the second section we shall explore further the notion of finite-dimensional subspaces, which will reveal to us the form assumed by the final linear system.

The third section is concerned with the question: How to construct suitable finite-dimensional subspaces for the Hilbert spaces $\mathbb{V}_\tau(\Omega)$ and $L^2(\Omega)$? At this point we present the meshfree spaces that will be used in the discretization process.

4.1 The problem in finite-dimensional subspaces

4.1.1 The key theorem: *Specialization to the scattering system*

In the development of the final form of the scattering system (3.48), we learned in (2.156) that the scattered field $\mathbf{E}^s$ and the pseudopressure p belong to $H^1(\Omega)^3$ and $L^2(\Omega)$, respectively. Before we look for their discretized counterparts, we now introduce the finite-dimensional subspaces

$$\mathbb{E}^h(\Omega) \subset H^1(\Omega)^3 \quad (4.1.a)$$

$$\mathbb{P}^h(\Omega) \subset L^2(\Omega). \quad (4.1.b)$$

(The meaning of the superscript h will become clear later.) Moreover, according to the standard finite element literature, it is common to include h either as a superscript or a subscript in the representation of the elements from the finite-dimensional subspaces. This is a kind of signature which makes it easier to identify the element as belonging to a subspace.

Because the finite-dimensional subspaces in (4.1) ultimately come from the discretization process, there is no harm in calling them ‘discretized spaces’, and elements from $\mathbb{E}^h(\Omega)$ and $\mathbb{P}^h(\Omega)$ as ‘discretized electric fields’ and ‘discretized pseudopressures’, respectively.

In (2.162), the original scattered electric field $\mathbf{E}^s \in H^1(\Omega)^3$ is split in two parts

$$\mathbf{E}^s = \mathbf{e}^0 + \mathbf{u}^g, \quad (4.1.c)$$

where $\mathbf{e}^0 \in \mathbb{V}_\tau(\Omega)$ and the lifting function $\mathbf{u}^g \in H^1(\Omega)^3$ obeys the boundary conditions

$$\boldsymbol{\gamma}_t \mathbf{u}^g = \begin{cases} \mathbf{0}, & \text{at } \Gamma_o \\ -\hat{\mathbf{n}}_1 \times \mathbf{E}^{inc}, & \text{at } \Gamma_1. \end{cases} \quad (4.1.d)$$

The splitting of $\mathbf{E}^s$ as in (4.1.c) paved the way for the formulation of the scattering problem (3.48) in terms of $\mathbf{e}^0$. After $\mathbf{e}^0$ is found, one just needs to add the known lifting $\mathbf{u}^g$ to it and the total scattered electric field $\mathbf{E}^s$ is recovered.

When working at the discretized level, the original scattering system (3.48) will be specialized to finite-dimensional subspaces. In doing so, we will get a discretized version of $\mathbf{e}^0$, represented by $\mathbf{e}_h^0$. This $\mathbf{e}_h^0$ belongs to a finite-dimensional subspace of $\mathbb{E}^h(\Omega)$ – namely, a space formed by elements in $\mathbb{E}^h(\Omega)$ whose tangential trace is zero, [to be introduced later in (4.3)]. The question is that after we find this finite-dimensional $\mathbf{e}_h^0$, if we add the infinite-dimensional lifting function $\mathbf{u}^g$ to it as in (4.1.c), it may happen that $\mathbf{e}_h^0 + \mathbf{u}^g$ will not be an element from the finite-dimensional subspace $\mathbb{E}^h(\Omega)$. In order to rule out this possibility, we shall consider not $\mathbf{u}^g$, but a finite-dimensional approximation to it in $\mathbb{E}^h(\Omega)$, denoted by $\mathbf{u}_h^g$. In this way,

$$\boldsymbol{\gamma}_t \mathbf{u}_h^g \cong \begin{cases} \mathbf{0}, & \text{at } \Gamma_o \\ -\hat{\mathbf{n}}_1 \times \mathbf{E}^{inc}, & \text{at } \Gamma_1. \end{cases} \quad (4.1.e)$$

i.e., the trace will be approximately equal to that of the continuous lifting function in (4.1.d). Consequently, now we can make sure that $\mathbf{e}_h^0 + \mathbf{u}_h^g$ will be an element of the finite-dimensional subspace $\mathbb{E}^h(\Omega)$. This is nothing else than the discretized scattered field

$$\mathbf{E}_h^s = \mathbf{e}_h^0 + \mathbf{u}_h^g \quad (4.2)$$

The advantage is that $\mathbf{E}_h^s$, $\mathbf{e}_h^0$ and $\mathbf{u}_h^g$ will ultimately belong to the same space $\mathbb{E}^h(\Omega)$.

We can now introduce a discretized version of $\mathbb{V}_\tau(\Omega)$, defined as

$$\mathbb{V}_\tau^h(\Omega) = \{\mathbf{v}_h \in \mathbb{E}^h(\Omega) \mid \boldsymbol{\gamma}_t \mathbf{v}_h = \mathbf{0}\}. \quad (4.3)$$

It can be seen that $\mathbf{e}_h^0 \in \mathbb{V}_\tau^h(\Omega)$.

There is a result in functional analysis which says that finite-dimensional subspaces are always closed [Kreyszig, 1989]. Since $\mathbb{E}^h(\Omega)$ is a finite-dimensional subspace of $H^1(\Omega)^3$, it is closed. As it will become clear later, the space $\mathbb{V}_\tau^h(\Omega)$ from (4.3) is also finite-dimensional (i.e., it is spanned by a set of basis functions). Therefore, $\mathbb{V}_\tau^h(\Omega)$ is closed. When equipped with the inner product of the ‘parental’ space $H^1(\Omega)^3$, it becomes a Hilbert space, due to Theorem 3.5. By the same reasoning, $\mathbb{P}^h(\Omega)$ in

(4.1.b) is a finite-dimensional subspace of $L^2(\Omega)$; when endowed with the inner product of $L^2(\Omega)$, it also becomes a Hilbert space.

After we find a suitable a lifting function $\mathbf{u}_h^g$ in (4.2), the discretized counterpart of problem (3.48) becomes:

Find $(\mathbf{e}_h^0, p_h) \in \mathbb{V}_\tau^h(\Omega) \times \mathbb{P}^h(\Omega)$ such that

$$\begin{aligned} \int_{\Omega} (\bar{\Lambda} \cdot \nabla \mathbf{e}_h^0) : \nabla \mathbf{v}_h^* - \int_{\Omega} k_0^2 \mathbf{e}_h^0 \cdot \mathbf{v}_h^* - \int_{\Omega} p_h \nabla \cdot \mathbf{v}_h^* = \\ - \int_{\Omega} (\bar{\Lambda} \cdot \nabla \mathbf{u}_h^g) : \nabla \mathbf{v}_h^* + \int_{\Omega} k_0^2 \mathbf{u}_h^g \cdot \mathbf{v}_h^*, \quad \forall \mathbf{v}_h \in \mathbb{V}_\tau^h(\Omega) \end{aligned} \quad (4.4.a)$$

$$- \int_{\Omega} q_h^* \nabla \cdot \mathbf{e}_h^0 = \int_{\Omega} q_h^* \nabla \cdot \mathbf{u}_h^g, \quad \forall q_h \in \mathbb{P}^h(\Omega), \quad (4.4.b)$$

We can now make a new identification:

$$\mathbb{V}_\tau^h(\Omega) \rightarrow \mathcal{X} \quad (4.5.a)$$

$$\mathbb{V}_\tau^h(\Omega)^* \rightarrow \mathcal{X}^* \quad (4.5.b)$$

$$\mathbb{P}^h(\Omega) \rightarrow \mathcal{Y} \quad (4.5.c)$$

$$\mathbb{P}^h(\Omega)^* \rightarrow \mathcal{Y}^* \quad (4.5.d)$$

$$\{\mathbf{w}_h, \mathbf{v}_h\} \rightarrow \left(\int_{\Omega} (\bar{\Lambda} \cdot \nabla \mathbf{w}_h) : \nabla \mathbf{v}_h^* - \int_{\Omega} k_0^2 \mathbf{w}_h \cdot \mathbf{v}_h^* \right) \rightarrow a(\cdot, \cdot) \quad (4.5.e)$$

$$\{\mathbf{v}_h, p_h\} \rightarrow \left(- \int_{\Omega} p_h^* \nabla \cdot \mathbf{v}_h \right) \rightarrow b(\cdot, \cdot) \quad (4.5.f)$$

$$\left(- \int_{\Omega} (\bar{\Lambda} \cdot \nabla \mathbf{u}_h^g) : \nabla (\cdot)^* + \int_{\Omega} k_0^2 \mathbf{u}_h^g \cdot (\cdot)^* \right) \rightarrow f^* \quad (4.5.g)$$

$$\left(\int_{\Omega} (\cdot)^* \nabla \cdot \mathbf{u}_h^g \right) \rightarrow g^*. \quad (4.5.h)$$

The identification above is clear enough. Since $\mathbf{u}_h^g \in \mathbb{E}^h(\Omega)$, it is ultimately an element of $H^1(\Omega)^3$. Applying a reasoning similar to that in Section 3.3.6.2, it is not difficult to see that the integrals at the right side of (4.4.a) define a functional f^* on elements of $H^1(\Omega)^3$. As $\mathbb{V}_\tau^h(\Omega)$ is also a subspace of $H^1(\Omega)^3$, when the action of this functional is restricted to elements from $\mathbb{V}_\tau^h(\Omega)$, it defines a functional on $\mathbb{V}_\tau^h(\Omega)$, i.e., $f^* \in \mathbb{V}_\tau^h(\Omega)^*$. In the same way, the integral at the right side of (4.4.b) defines an functional $g^* \in \mathbb{P}^h(\Omega)^*$.

We can now state an extension of Theorem 3.9 which is concerned with the well-posedness of the system (4.4).

Theorem 4.1: Well-posedness of the scattering system, finite-dimensional case – Let it be the finite-dimensional complex-valued Hilbert spaces $\mathbb{V}_\tau^h(\Omega)$ and $\mathbb{P}^h(\Omega)$. Suppose there are two positive constants α_a^h and α_b^h such that:

(i) a is continuous, i.e.,

$$\left| \int_{\Omega} (\bar{\Lambda} \cdot \nabla \mathbf{w}_h) : \nabla \mathbf{v}_h^* - \int_{\Omega} k_0^2 \mathbf{w}_h \cdot \mathbf{v}_h^* \right| \leq \alpha_a^h \|\mathbf{w}_h\|_{H^1(\Omega)^3} \|\mathbf{v}_h\|_{H^1(\Omega)^3},$$

$$\forall \mathbf{w}_h, \mathbf{v}_h \in \mathbb{V}_\tau^h(\Omega) \quad (4.6.a)$$

(ii) b is continuous, i.e.,

$$\left| - \int_{\Omega} q_h^* \nabla \cdot \mathbf{v}_h \right| \leq \alpha_b^h \|\mathbf{v}_h\|_{H^1(\Omega)^3} \|q_h\|_{L^2(\Omega)}, \quad \forall \mathbf{v}_h \in \mathbb{V}_\tau^h(\Omega) \quad \forall q_h \in \mathbb{P}^h(\Omega) \quad (4.6.b)$$

Let $\mathcal{X}_h^0$ be the kernel of the sesquilinear form b i.e.,

$$\mathcal{X}_h^0 = \text{Ker } b = \left\{ \mathbf{v}_h \in \mathbb{V}_\tau^h(\Omega) \mid - \int_{\Omega} q_h^* \nabla \cdot \mathbf{v}_h = 0, \quad \forall q_h \in \mathbb{P}^h(\Omega) \right\}. \quad (4.6.c)$$

Consider a third Hilbert space H such that $\mathcal{X}_h^0$ and H satisfy the requirements of Theorem 3.7, i.e.,

(iii) $\mathcal{X}_h^0$ is continuously embedded into H , i.e., $\mathcal{X}_h^0 \hookrightarrow H$.

Moreover, it holds that:

(iv) The map $I_{\mathcal{X}_h^0 \rightarrow H}$ is compact, i.e., $I_{\mathcal{X}_h^0 \rightarrow H} \in \mathcal{K}(\mathcal{X}_h^0, H)$.

(v) The sesquilinear form a satisfies the following property on the kernel $\mathcal{X}_h^0$: There exist constants $\eta^h > 0$ and $\kappa_0^h \geq 0$ such that

$$\text{Re}\{a(\mathbf{u}_h, \mathbf{u}_h)\} + \kappa_0^h \|I_{\mathcal{X}_h^0 \rightarrow H}(\mathbf{u}_h)\|_H^2 \geq \eta^h \|\mathbf{u}_h\|_{\mathcal{X}}^2, \quad \forall \mathbf{u}_h \in \mathcal{X}_h^0. \quad (4.6.d)$$

(vi) The sesquilinear form b satisfies the *inf-sup* condition, i.e., there is a positive constant $\beta_b^h > 0$ such that

$$\inf_{q_h \in \mathbb{P}^h(\Omega) \setminus \{0\}} \sup_{\mathbf{v}_h \in \mathbb{V}_\tau^h(\Omega) \setminus \{0\}} \frac{\left| - \int_{\Omega} q_h \nabla \cdot \mathbf{v}_h \right|}{\|\mathbf{v}_h\|_{H^1(\Omega)^3} \|q_h\|_{L^2(\Omega)}} \geq \beta_b^h. \quad (4.6.e)$$

(vii) The solution to the homogeneous (zero-data) problem at the kernel $\mathcal{X}_h^0$

Find $\mathbf{w}_h \in \mathcal{X}_h^0$ such that

$$a(\mathbf{w}_h, \mathbf{v}_h) = 0, \quad \forall \mathbf{v}_h \in \mathcal{X}_h^0 \quad (4.6.f)$$

is the zero element $\mathbf{w}_h = \mathbf{0}$. Furthermore, let us assume that:

(viii) The original space $\mathbb{V}_\tau^h(\Omega)$ is also continuously embedded H , i.e., $\mathbb{V}_\tau^h(\Omega) \hookrightarrow H$.

(ix) The spaces $\mathbb{V}_\tau^h(\Omega)$ and $\mathcal{X}_h^0$ are subspaces of H , i.e., $\mathbb{V}_\tau^h(\Omega) \subset H$ and $\mathcal{X}_h^0 \subset H$ (which implies that $I_{\mathbb{V}_\tau^h(\Omega) \rightarrow H}$ and $I_{\mathcal{X}_h^0 \rightarrow H}$ are inclusion maps).

Then it can be concluded that for each $f^* \in \mathbb{V}_\tau^h(\Omega)^*$ and $g^* \in \mathbb{P}^h(\Omega)^*$, there is a unique solution to the mixed problem

Find $(\mathbf{e}_h^0, p_h) \in \mathbb{V}_\tau^h(\Omega) \times \mathbb{P}^h(\Omega)$ such that

$$a(\mathbf{e}_h^0, \mathbf{v}_h) + b(\mathbf{v}_h, p_h) = \langle f^*, \mathbf{v}_h \rangle_{\mathbb{V}_\tau^h(\Omega)^*, \mathbb{V}_\tau^h(\Omega)}, \quad \forall \mathbf{v}_h \in \mathbb{V}_\tau^h(\Omega) \quad (4.6.g)$$

$$b(\mathbf{e}_h^0, q_h) = \langle g^*, q_h \rangle_{\mathbb{P}^h(\Omega)^*, \mathbb{P}^h(\Omega)}, \quad \forall q_h \in \mathbb{P}^h(\Omega)$$

It also follows that the solution $\mathbf{e}_h^0$ depends continuously on the data f^* and g^* in the H norm, i.e., there are positive constants K_1 and K_2 such that

$$\|\mathbf{e}_h^0\|_H \leq K_1 \|f^*\|_{\mathbb{V}_\tau^h(\Omega)^*} + K_2 \|g^*\|_{\mathbb{P}^h(\Omega)^*} \quad (4.6.h)$$

In (4.6.a), (4.6.b), (4.6.d) and (4.6.e), the superscript h has been introduced in the constants in order to indicate that these constants *may* depend on the specific subspaces $\mathbb{V}_\tau^h(\Omega)$ and $\mathbb{P}^h(\Omega)$ under consideration.

What we are going to do next is to verify if all the nine hypotheses of Theorem 4.1 hold true. The results we got in Sections 3.3.6.2 – 3.3.6.8 for will help us considerably.

4.1.1.1 Hypothesis (i)

From (3.72.k) and (3.72.l), it is not difficult to see that

$$(4.7.a)$$

$$\left| \int_{\Omega} (\bar{\mathbf{A}} \cdot \nabla \mathbf{w}) : \nabla \mathbf{v}^* - \int_{\Omega} k_0^2 \mathbf{w} \cdot \mathbf{v}^* \right| \leq (\Lambda_M \|\mathbf{w}\|_{H^1(\Omega)^3} + k_0^2 \|\mathbf{w}\|_{L^2(\Omega)^3}) \|\mathbf{v}\|_{H^1(\Omega)^3},$$

for any $\mathbf{w}, \mathbf{v} \in H^1(\Omega)^3$. Since $\|\mathbf{w}\|_{L^2(\Omega)^3} \leq \|\mathbf{w}\|_{H^1(\Omega)^3}$ by (2.42), we conclude that for any $\mathbf{w}, \mathbf{v} \in H^1(\Omega)^3$,

$$\left| \int_{\Omega} (\bar{\mathbf{A}} \cdot \nabla \mathbf{w}) : \nabla \mathbf{v}^* - \int_{\Omega} k_0^2 \mathbf{w} \cdot \mathbf{v}^* \right| \leq (\Lambda_M + k_0^2) \|\mathbf{w}\|_{H^1(\Omega)^3} \|\mathbf{v}\|_{H^1(\Omega)^3}. \quad (4.7.b)$$

Since $\mathbb{V}_\tau^h(\Omega) \subset \mathbb{E}^h(\Omega) \subset H^1(\Omega)^3$, the elements in (4.7.b) can be restricted to those in $\mathbb{V}_\tau^h(\Omega)$. As a consequence, we get (4.6.a) in which the constant α_a^h is independent of h and is given by $\Lambda_M + k_0^2$. Hypothesis (i) has been checked.

4.1.1.2 Hypothesis (ii)

The inequality (2.86.b), gives us that

$$\left| \int_{\Omega} q_h^* \nabla \cdot \mathbf{v} \right| \leq \|q\|_{L^2(\Omega)} \|\mathbf{v}\|_{H^1(\Omega)^3} \quad \forall q \in L^2(\Omega) \quad \forall \mathbf{v} \in H^1(\Omega)^3. \quad (4.8)$$

As $\mathbb{V}_\tau^h(\Omega) \subset H^1(\Omega)^3$ and $\mathbb{P}^h(\Omega) \subset L^2(\Omega)$, (4.8) can be restricted to these spaces. The result is (4.6.b), in which $\alpha_b^h = 1$. Hypothesis (ii) has been checked.

4.1.1.3 Hypotheses (iii) and (viii)

As in (3.74.r), we make

$$H = L^2(\Omega)^3. \quad (4.9.a)$$

From (3.74.q), $H^1(\Omega)^3 \hookrightarrow L^2(\Omega)^3$. Since $\mathcal{X}_h^0 \subset \mathbb{V}_\tau^h(\Omega) \subset \mathbb{E}^h(\Omega) \subset H^1(\Omega)^3$, then we may conclude that

$$\mathbb{V}_\tau^h(\Omega) \hookrightarrow L^2(\Omega)^3 = H \quad (4.9.b)$$

$$\mathcal{X}_h^0 \hookrightarrow L^2(\Omega)^3 = H \quad (4.9.c)$$

From the two expressions above, we get that hypotheses (iii) and (viii) have been checked.

4.1.1.4 Hypothesis (ix)

The following chain of inclusions is valid:

$$\mathcal{X}_h^0 \subset \mathbb{V}_\tau^h(\Omega) \subset \mathbb{E}^h(\Omega) \subset H^1(\Omega)^3 \subset L^2(\Omega)^3 = H, \quad (4.10.a)$$

from which it becomes evident that $\mathbb{V}_\tau^h(\Omega) \subset H$ and $\mathcal{X}_h^0 \subset H$. In this way, $I_{\mathbb{V}_\tau^h(\Omega) \rightarrow L^2(\Omega)^3}$, and $I_{\mathcal{X}_h^0 \rightarrow L^2(\Omega)^3}$ are identity maps, and thus hypothesis (ix) has been checked. Since these are identity maps, it means that

$$I_{\mathcal{X}_h^0 \rightarrow L^2(\Omega)^3}(\mathbf{u}) = \mathbf{u}, \quad \forall \mathbf{u} \in \mathcal{X}_h^0 \quad (4.10.b)$$

$$I_{\mathbb{V}_\tau^h(\Omega) \rightarrow L^2(\Omega)^3}(\mathbf{u}) = \mathbf{u}, \quad \forall \mathbf{u} \in \mathbb{V}_\tau^h(\Omega) \quad (4.10.c)$$

Despite the fact that elements from $\mathcal{X}_h^0$ and $\mathbb{V}_\tau^h(\Omega)$ are also elements of $L^2(\Omega)^3$, they are measured differently. When seen as elements of $\mathcal{X}_h^0$ and $\mathbb{V}_\tau^h(\Omega)$, they are measured in

the $\|\cdot\|_{H^1(\Omega)^3}$ norm. On the other hand, after the action of the embedding map, they are seen as elements of $L^2(\Omega)^3$, and therefore measured in the $\|\cdot\|_{L^2(\Omega)^3}$ norm.

4.1.1.5 Hypothesis (iv)

According to Section 4.1.1.3, there is an embedding map $I_{\mathcal{X}_h^0 \rightarrow L^2(\Omega)^3}$. We need to show that it is compact. In order to do so, let $\{\mathbf{v}_h^n\}_{n=1}^\infty \subset \mathcal{X}_h^0$ be an arbitrary bounded sequence in $\mathcal{X}_h^0$. Since $\mathcal{X}_h^0 \subset H^1(\Omega)^3$, it also constitutes a bounded sequence in $H^1(\Omega)^3$. As we concluded in Section 3.3.6.4, the embedding of $H^1(\Omega)^3$ into $L^2(\Omega)^3$ is compact. Then the image $\{I_{H^1(\Omega)^3 \rightarrow L^2(\Omega)^3}(\mathbf{v}_h^n)\}_{n=1}^\infty$ admits a convergent subsequence in $L^2(\Omega)^3$. But all elements of the sequence are in $\mathcal{X}_h^0$. From (4.9.c), it follows that $\{I_{\mathcal{X}_h^0 \rightarrow L^2(\Omega)^3}(\mathbf{v}_h^n)\}_{n=1}^\infty$ admits a convergent subsequence in $L^2(\Omega)^3$. Therefore, $I_{\mathcal{X}_h^0 \rightarrow L^2(\Omega)^3}$ is compact, and in this way the hypothesis (iv) has been checked.

4.1.1.6 Hypothesis (v)

Let $\mathbf{u}_h \in \mathbb{V}_\tau^h(\Omega)$ be arbitrary. According to the definition (4.3), $\mathbf{u}_h \in \mathbb{E}^h(\Omega)$ and $\boldsymbol{\gamma}_t \mathbf{u}_h = \mathbf{0}$. But $\mathbb{E}^h(\Omega) \subset H^1(\Omega)^3$, according to (4.1.a). But if an element of $H^1(\Omega)^3$ is such that its tangential trace is zero, then it belongs to $\mathbb{V}_\tau(\Omega)$, according to (2.163). Consequently, $\mathbf{u}_h \in \mathbb{V}_\tau(\Omega)$. Since $\mathbf{u}_h$ was arbitrary, we are allowed to conclude that

$$\mathbb{V}_\tau^h(\Omega) \subset \mathbb{V}_\tau(\Omega). \quad (4.11.a)$$

From (4.10.a) and (3.78.t), get

$$\operatorname{Re}\{a(\mathbf{u}_h, \mathbf{u}_h)\} + k_0^2 \|\mathbf{u}_h\|_{L^2(\Omega)^3}^2 \geq \frac{\beta}{(c_\Omega + 1)} \|\mathbf{u}_h\|_{H^1(\Omega)^3}^2, \quad \forall \mathbf{u}_h \in \mathbb{V}_\tau^h(\Omega). \quad (4.11.b)$$

It may be noticed that, notwithstanding the fact that (4.11.a) is a truth, the space $\mathbb{V}_\tau^h(\Omega)$ is not introduced directly as a subspace of $\mathbb{V}_\tau(\Omega)$ (although it is). It is introduced as a subspace of $\mathbb{E}^h(\Omega)$ in (4.3). The reason is that, as it will become clearer later, after $\mathbb{E}^h(\Omega)$ is constructed from a set of basis functions, the construction of $\mathbb{V}_\tau^h(\Omega)$ follows in a remarkably easy way.

Since $\mathcal{X}_h^0 \subset \mathbb{V}_\tau^h(\Omega)$, (4.11.b) can be restricted to those elements in $\mathcal{X}_h^0$, which allows one to conclude that

$$\operatorname{Re}\{a(\mathbf{u}_h, \mathbf{u}_h)\} + k_0^2 \|\mathbf{u}_h\|_{L^2(\Omega)^3}^2 \geq \frac{\beta}{(c_\Omega + 1)} \|\mathbf{u}_h\|_{H^1(\Omega)^3}^2, \quad \forall \mathbf{u}_h \in \mathcal{X}_h^0. \quad (4.11.c)$$

Given that $I_{\mathcal{X}_h^0 \rightarrow L^2(\Omega)^3}$ is the identity map, $I_{\mathcal{X}_h^0 \rightarrow L^2(\Omega)^3}(\mathbf{u}_h) = \mathbf{u}_h$ for any $\mathbf{u}_h \in \mathbb{V}_\tau^h(\Omega)$, according to (4.10.b). This implies that

$$\left\| I_{\mathcal{X}_h^0 \rightarrow L^2(\Omega)^3}(\mathbf{u}_h) \right\|_{L^2(\Omega)^3} = \|\mathbf{u}_h\|_{L^2(\Omega)^3}. \quad (4.11.d)$$

This allows us to rewrite (4.11.c) as

$$\operatorname{Re}\{a(\mathbf{u}_h, \mathbf{u}_h)\} + k_0^2 \left\| I_{\mathcal{X}_h^0 \rightarrow L^2(\Omega)^3}(\mathbf{u}_h) \right\|_{L^2(\Omega)^3}^2 \geq \frac{\beta}{(c_\Omega + 1)} \|\mathbf{u}_h\|_{H^1(\Omega)^3}^2, \quad \forall \mathbf{u}_h \in \mathcal{X}_h^0, \quad (4.11.e)$$

which is nothing else than (4.6.d). And so, hypothesis (v) has been checked. The constants κ_0^h and η^h in (4.6.d) are such that $\kappa_0^h = k_0^2$ and $\eta^h = \beta/(c_\Omega + 1)$, i.e., they are the same as those occurring in the infinite-dimensional case (and therefore are independent of h).

4.1.1.7 Hypothesis (vi)

According to (3.33.b) in Theorem 3.3, the inf-sup condition in (4.6.e) is equivalent to the fact that there is a positive constant $\beta_b^h > 0$ such that

$$\forall q_h \in \mathbb{P}^h(\Omega) \exists \mathbf{v}_h \in \mathbb{V}_\tau^h(\Omega) \setminus \{\mathbf{0}\} \text{ s. t. } \left| - \int_\Omega q_h \nabla \cdot \mathbf{v}_h \right| \geq \beta_b^h \|\mathbf{v}_h\|_{H^1(\Omega)^3} \|q_h\|_{L^2(\Omega)} \quad (4.12.a)$$

In the same way, the inf-sup condition in (3.75.p), which we proved to be true, is equivalent the fact that there is a positive $\beta_b > 0$ such that

$$\forall q \in L^2(\Omega) \exists \mathbf{v} \in \mathbb{V}_\tau(\Omega) \setminus \{\mathbf{0}\} \text{ s. t. } \left| - \int_\Omega q \nabla \cdot \mathbf{v} \right| \geq \beta_b \|\mathbf{v}\|_{H^1(\Omega)^3} \|q\|_{L^2(\Omega)} \quad (4.12.b)$$

One may ask: Is it true that (4.12.b) implies (4.12.a)? The answer is negative. The question is that the inf-sup condition at the finite-dimensional level (4.12.a) *does not* inherit its validity from its infinite-dimensional counterpart (4.12.b). There is a very subtle argument to show it.

Assume that (4.12.b) is true (which it is, indeed). Now let $q_h \in \mathbb{P}^h(\Omega)$ be arbitrary. According to (4.1.b), $\mathbb{P}^h(\Omega) \subset L^2(\Omega)$, so $q_h \in L^2(\Omega)$. From (4.12.b), it follows that

$$\exists \mathbf{v} \in \boxed{\mathbb{V}_\tau(\Omega) \setminus \{\mathbf{0}\}} \text{ s. t. } \left| - \int_\Omega q_h \nabla \cdot \mathbf{v} \right| \geq \beta_b \|\mathbf{v}\|_{H^1(\Omega)^3} \|q_h\|_{L^2(\Omega)}. \quad (4.12.c)$$

Expression (4.12.c) says that there is an element in $\mathbb{V}_\tau(\Omega) \setminus \{\mathbf{0}\}$ such that the integral inequality at the right is satisfied. However, in order to prove (4.12.a), we need to show that, given $q_h \in L^2(\Omega)$, there must exist an element in $\mathbb{V}_\tau^h(\Omega) \setminus \{\mathbf{0}\}$ such that the inequality is satisfied, i.e., we need to show that

$$\exists \mathbf{v}_h \in \boxed{\mathbb{V}_\tau^h(\Omega) \setminus \{\mathbf{0}\}} \text{ s. t. } \left| - \int_{\Omega} q_h \nabla \cdot \mathbf{v}_h \right| \geq \beta_b^h \|\mathbf{v}_h\|_{H^1(\Omega)^3} \|q_h\|_{L^2(\Omega)} \quad (4.12.d)$$

The point is that the $\mathbf{v}$ from (4.12.c) *may not be in the finite-dimensional subspace* $\mathbb{V}_\tau^h(\Omega)$, as required. Expression (4.12.c) acknowledges the existence of an element in the larger space $\mathbb{V}_\tau(\Omega)$. But we need to be sure that this element belongs to the subspace $\mathbb{V}_\tau^h(\Omega)$. This subtle difference is indicated by the boxes in expressions (4.12.c) and (4.12.d).

In this way, hypothesis (vi) has not been satisfied. In a sense, there is no general proof that (4.12.a) holds true for *any* pair of finite-dimensional subspaces $\mathbb{V}_\tau^h(\Omega)$ and $\mathbb{P}^h(\Omega)$.

The same situation happens when one considers the discretized version of the Navier-Stokes problem. The inf-sup condition at the infinite-dimensional level (3.40) is known to be true. However, at the discretized level, there is no proof that it holds for *any* pair of finite-dimensional subspaces of $H_0^1(\Omega)^d$ and $L_0^2(\Omega)$. When these finite-dimensional subspaces are finite-element spaces (i.e., relying on a mesh), there are certain pairs for which researchers were able to prove that they satisfy the inf-sup condition. There is a list of such pairs in [Girault and Raviart, 1986], [Brezzi and Fortin, 1991], [Glowinski et al., 2003]. This is a very delicate issue; one cannot choose whatever pair he wants, because a pair which does not satisfy the inf-sup condition may lead to an ill-posed problem, which is prone to instabilities. But does it mean that one is doomed to use only those pairs already catalogued in the literature?

Fortunately, no. There is a test to assess if a given pair satisfies the discrete inf-sup condition. In this way, one could develop a pair of finite-element spaces, and then apply the test. If they pass the test, then they lead to a well-posed problem. This test is carried out at the numerical (i.e., matrix) level, and was developed by K. J. Bathe in [Bathe, 2001], [Brezzi and Bathe, 1990]. It will be explained in due time.

It is true that the spaces involved in the Navier-Stokes and in the scattering problem are different. So those pairs from the literature do not apply, as they have been developed for the Navier-Stokes system. Moreover, we are planning to construct *meshfree* finite-dimensional spaces. Of course, to prove that a given pair of finite-dimensional spaces $\mathbb{V}_\tau^h(\Omega)$ and $\mathbb{P}^h(\Omega)$ spanned by meshfree basis functions satisfies (4.12.a) is out of question. The only alternative is to resort to the aforementioned test.

4.1.1.8 Hypothesis (vii)

Similar conclusions from Section 3.3.6.7 are also valid here. As long as the wavenumber k_0^2 is not one of the eigenvalues of the (discretized) problem

Find $\mathbf{w}_h^\sigma \in \mathcal{X}_h^0$ such that

$$\int_{\Omega} (\bar{\mathbf{\Lambda}} \cdot \nabla \mathbf{w}_h^\sigma) : \nabla \mathbf{v}_h^* = \sigma \int_{\Omega} \mathbf{w}_h^\sigma \cdot \mathbf{v}_h^*, \quad \forall \mathbf{v}_h \in \mathcal{X}_h^0, \quad (4.13.a)$$

the solution of the homogeneous (zero-data) problem

Find $\mathbf{w}_h \in \mathcal{X}_h^0$ such that

$$a(\mathbf{w}_h, \mathbf{v}_h) = \int_{\Omega} (\bar{\mathbf{\Lambda}} \cdot \nabla \mathbf{w}_h) : \nabla \mathbf{v}_h^* - \int_{\Omega} k_0^2 \mathbf{w}_h \cdot \mathbf{v}_h^* = 0, \quad \forall \mathbf{w}_h \in \mathcal{X}_h^0 \quad (4.13.b)$$

is the zero element, i.e., $\mathbf{w}_h = \mathbf{0}$.

However, as it was discussed in Section 3.3.6.7, there is a high probability that the eigenvalues associated to problem (4.13.a) are complex. In this way, any real k_0^2 will not be an eigenvalue. So we can say that hypothesis (vii) has been checked.

4.1.1.9 Concluding remarks

The well-posedness of problem (4.4) is thus shown to depend only on the inf-sup condition. All other hypotheses hold true, except the sixth. In the next section, we expect to offer a solution to this issue.

4.2 The linear system

4.2.1 The matrix system: Preliminary form

Let us consider the discretized problem in (4.6.g)

Find $(\mathbf{e}_h^0, p_h) \in \mathbb{V}_\tau^h(\Omega) \times \mathbb{P}^h(\Omega)$ such that

$$a(\mathbf{e}_h^0, \mathbf{v}_h) + b(\mathbf{v}_h, p_h) = \langle f^*, \mathbf{v}_h \rangle_{\mathbb{V}_\tau^h(\Omega)^*, \mathbb{V}_\tau^h(\Omega)}, \quad \forall \mathbf{v}_h \in \mathbb{V}_\tau^h(\Omega) \quad (4.14.a)$$

$$b(\mathbf{e}_h^0, q_h) = \langle g^*, q_h \rangle_{\mathbb{P}^h(\Omega)^*, \mathbb{P}^h(\Omega)}, \quad \forall q_h \in \mathbb{P}^h(\Omega)$$

Suppose that the space $\mathbb{V}_\tau^h(\Omega)$ is spanned by a total of Q basis functions:

$$\mathbb{V}_\tau^h(\Omega) = \text{span}\{\boldsymbol{\psi}_1, \boldsymbol{\psi}_2, \dots, \boldsymbol{\psi}_Q\}, \quad (4.14.b)$$

and suppose also that $\mathbb{P}^h(\Omega)$ is spanned by a total of M basis functions:

$$\mathbb{P}^h(\Omega) = \text{span}\{\theta_1, \theta_2, \dots, \theta_M\}. \quad (4.14.c)$$

Of course, these basis functions are functions of the position $\mathbf{x} \in \Omega$. But instead of writing $\boldsymbol{\psi}_1(\mathbf{x})$, we write just $\boldsymbol{\psi}_1$, for the sake of a cleaner notation. Under these circumstances, it is true that

$$\dim \mathbb{V}_\tau^h(\Omega) = Q \quad (4.14.d)$$

$$\dim \mathbb{P}^h(\Omega) = M. \quad (4.14.e)$$

The quantities $\mathbf{e}_h^0$, $\mathbf{v}_h$, p_h and q_h in (4.14.a) admit expansions of the type

$$\mathbf{e}_h^0 = \sum_{j=1}^Q \boldsymbol{\psi}_j \hat{e}_j \quad (4.14.f)$$

$$\mathbf{v}_h = \sum_{i=1}^Q \boldsymbol{\psi}_i \hat{v}_i \quad (4.14.g)$$

$$p_h = \sum_{j=1}^N \theta_j \hat{p}_j \quad (4.14.h)$$

$$q_h = \sum_{i=1}^N \theta_i \hat{q}_i, \quad (4.14.i)$$

where the $\hat{e}_j$ are the scalar coefficients associated with the basis function $\boldsymbol{\psi}_j$ in a given expansion for $\mathbf{e}_h^0$, and so on for the others. These coefficients are also referred to as *degrees of freedom* (DoF's), and particularly for the scattering system, they are complex numbers.

It is useful to put all DoF's together in a vector, as follows:

$$[\hat{e}_1, \hat{e}_2, \dots, \hat{e}_Q]^T =: \bar{\mathbf{e}} \quad (4.14.j)$$

$$[\hat{v}_1, \hat{v}_2, \dots, \hat{v}_Q]^T =: \bar{\mathbf{v}} \quad (4.14.k)$$

$$[\hat{p}_1, \hat{p}_2, \dots, \hat{p}_M]^T =: \bar{\mathbf{p}} \quad (4.14.l)$$

$$[\hat{q}_1, \hat{q}_2, \dots, \hat{q}_N]^T =: \bar{\mathbf{q}}. \quad (4.14.m)$$

The vectors in (4.14.j) and (4.14.k) are elements of $\mathbb{C}^Q$, whereas those in (4.14.l) and (4.14.m) are in $\mathbb{C}^M$.

When (4.14.f) – (4.14.i) are substituted into the system (4.14.a), after some manipulation, one arrives at the algebraic system:

$$\begin{aligned} & \text{Find } (\bar{\mathbf{e}}, \bar{\mathbf{p}}) \in \mathbb{C}^Q \times \mathbb{C}^M \text{ such that} \\ & \bar{\mathbf{v}}^\dagger \bar{\mathbf{A}} \bar{\mathbf{e}} + \bar{\mathbf{v}}^\dagger \bar{\mathbf{B}}^\dagger \bar{\mathbf{p}} = \bar{\mathbf{v}}^\dagger \bar{\mathbf{f}}, \quad \forall \bar{\mathbf{v}} \in \mathbb{C}^Q \\ & \bar{\mathbf{q}}^\dagger \bar{\mathbf{B}} \bar{\mathbf{e}} = \bar{\mathbf{q}}^\dagger \bar{\mathbf{g}}, \quad \forall \bar{\mathbf{q}} \in \mathbb{C}^M, \end{aligned} \quad (4.14.n)$$

where “ $\dagger$ ” means the conjugate transpose. The first equation in (4.14.n) is rewritten as

$$\bar{\mathbf{v}}^\dagger (\bar{\mathbf{A}}\bar{\mathbf{e}} + \bar{\mathbf{B}}^\dagger \bar{\mathbf{p}} - \bar{\mathbf{f}}) = 0, \quad \forall \bar{\mathbf{v}} \in \mathbb{C}^Q, \quad (4.14.o)$$

which implies that $\bar{\mathbf{A}}\bar{\mathbf{e}} + \bar{\mathbf{B}}^\dagger \bar{\mathbf{p}} - \bar{\mathbf{f}}$ must be orthogonal to all elements from $\mathbb{C}^Q$. The only possibility is that $\bar{\mathbf{A}}\bar{\mathbf{e}} + \bar{\mathbf{B}}^\dagger \bar{\mathbf{p}} - \bar{\mathbf{f}} = \mathbf{0}$ (the zero vector in $\mathbb{C}^Q$), or equivalently, that $\bar{\mathbf{A}}\bar{\mathbf{e}} + \bar{\mathbf{B}}^\dagger \bar{\mathbf{p}} = \bar{\mathbf{f}}$. The same analysis must be applied to the second equation in (4.14.n), and the conclusion is that $\bar{\mathbf{B}}\bar{\mathbf{e}} = \bar{\mathbf{g}}$. We thus arrive at a linear system:

Find $(\bar{\mathbf{e}}, \bar{\mathbf{p}}) \in \mathbb{C}^Q \times \mathbb{C}^M$ such that

$$\bar{\mathbf{A}}\bar{\mathbf{e}} + \bar{\mathbf{B}}^\dagger \bar{\mathbf{p}} = \bar{\mathbf{f}} \quad (4.14.p)$$

$$\bar{\mathbf{B}}\bar{\mathbf{e}} = \bar{\mathbf{g}}.$$

The matrices $\bar{\mathbf{A}}$ and $\bar{\mathbf{B}}$, and the vectors $\bar{\mathbf{f}}$ and $\bar{\mathbf{g}}$ in (4.14.n) are described by their coefficients:

$$[\bar{\mathbf{A}}]_{ij} = a(\boldsymbol{\psi}_j, \boldsymbol{\psi}_i) \quad (4.14.q)$$

$$[\bar{\mathbf{B}}]_{ij} = b(\boldsymbol{\psi}_j, \theta_i) \quad (4.14.r)$$

$$[\bar{\mathbf{f}}]_i = \langle f^*, \boldsymbol{\psi}_i \rangle_{\mathbb{V}_\tau^h(\Omega)^*, \mathbb{V}_\tau^h(\Omega)} \quad (4.14.s)$$

$$[\bar{\mathbf{g}}]_i = \langle g^*, \theta_i \rangle_{\mathbb{P}^h(\Omega)^*, \mathbb{P}^h(\Omega)}. \quad (4.14.t)$$

The equations in (4.14.p) can be assembled together into a matrix system as

Find $(\bar{\mathbf{e}}, \bar{\mathbf{p}}) \in \mathbb{C}^Q \times \mathbb{C}^M$ such that

$$\begin{bmatrix} \bar{\mathbf{A}} & \bar{\mathbf{B}}^\dagger \\ \bar{\mathbf{B}} & \mathbf{0} \end{bmatrix} \begin{bmatrix} \bar{\mathbf{e}} \\ \bar{\mathbf{p}} \end{bmatrix} = \begin{bmatrix} \bar{\mathbf{f}} \\ \bar{\mathbf{g}} \end{bmatrix} \quad (4.14.u)$$

After the identification (4.5), the matrix coefficients in (4.14.q) – (4.14.t) can be expressed in terms of basis functions from (4.14.b) and (4.14.c) as

$$a(\boldsymbol{\psi}_j, \boldsymbol{\psi}_i) = \int_{\Omega} (\bar{\bar{\mathbf{A}}} \cdot \nabla \boldsymbol{\psi}_j) : \nabla \boldsymbol{\psi}_i - \int_{\Omega} k_0^2 \boldsymbol{\psi}_j \cdot \boldsymbol{\psi}_i \quad (4.14.v)$$

$$b(\boldsymbol{\psi}_j, \theta_i) = - \int_{\Omega} \theta_i \nabla \cdot \boldsymbol{\psi}_j \quad (4.14.w)$$

$$\langle f^*, \boldsymbol{\psi}_i \rangle_{\mathbb{V}_\tau^h(\Omega)^*, \mathbb{V}_\tau^h(\Omega)} = - \int_{\Omega} (\bar{\bar{\mathbf{A}}} \cdot \nabla \mathbf{u}_h^g) : \nabla \boldsymbol{\psi}_i + \int_{\Omega} k_0^2 \mathbf{u}_h^g \cdot \boldsymbol{\psi}_i \quad (4.14.x)$$

$$\langle g^*, \theta_i \rangle_{\mathbb{P}^h(\Omega)^*, \mathbb{P}^h(\Omega)} = \int_{\Omega} \theta_i \nabla \cdot \mathbf{u}_h^g. \quad (4.14.y)$$

Some observations are in order. The basis functions in (4.14.b) and (4.14.c) are real functions, i.e., they have no imaginary part. If any of the quantities in (4.14.f) – (4.14.i) are complex, this is due solely to the coefficients (DoF's) being complex.

The coefficients of the matrix $\bar{\mathbf{A}}$ are complex, because the PML tensor $\bar{\bar{\mathbf{A}}}$ enters their calculation, as revealed by (4.14.v). In what regards the matrix $\bar{\mathbf{B}}$, its entries are real, according to (4.14.w). Consequently, $\bar{\mathbf{B}} = \bar{\mathbf{B}}^*$. Since $\bar{\mathbf{B}}^\dagger = (\bar{\mathbf{B}}^*)^T$, then $\bar{\mathbf{B}}^\dagger = \bar{\mathbf{B}}^T$. In this way the system (4.14.u) assumes the standard form

Find $(\bar{\mathbf{e}}, \bar{\mathbf{p}}) \in \mathbb{C}^Q \times \mathbb{C}^M$ such that

$$\begin{bmatrix} \bar{\mathbf{A}} & \bar{\mathbf{B}}^T \\ \bar{\mathbf{B}} & \mathbf{0} \end{bmatrix} \begin{bmatrix} \bar{\mathbf{e}} \\ \bar{\mathbf{p}} \end{bmatrix} = \begin{bmatrix} \bar{\mathbf{f}} \\ \bar{\mathbf{g}} \end{bmatrix} \quad (4.14.z)$$

4.2.2 The matrix system: Uniqueness of the solution

In this subsection we shall investigate the solvability of problem (4.14.z). The analysis will be brief, as much has already been done in the study of Theorem 4.1. The intention is to show how some of the hypotheses actually become ‘manifest’ down here at the matrix level.

Before we proceed, we need two observations regarding the kernel of the sesquilinear form b . First, given arbitrary elements $\mathbf{v}_h, \mathbf{w}_h \in \mathbb{V}_\tau^h(\Omega)$ and $q_h \in \mathbb{P}^h(\Omega)$ together with their expansions in basis functions according to (4.14.f) – (4.14.m), it is true that

$$b(\mathbf{v}_h, q_h) = \bar{\mathbf{q}}^\dagger \bar{\mathbf{B}} \bar{\mathbf{v}}. \quad (4.15.a)$$

If we remember the definition of the kernel (null-space) of the form b in (4.6.c),

$$\text{Ker } b = \{\mathbf{v}_h \in \mathbb{V}_\tau^h(\Omega) \mid b(\mathbf{v}_h, q_h) = 0, \forall q_h \in \mathbb{P}^h(\Omega)\}. \quad (4.15.b)$$

then it is not difficult to conclude that

$$\mathbf{v}_h \in \text{Ker } b \Leftrightarrow \bar{\mathbf{v}} \in \text{Ker } \bar{\mathbf{B}}. \quad (4.15.c)$$

In other words, if $\mathbf{v}_h$ is in $\text{Ker } b$, then the vector of DoF's corresponding to the expansion of $\mathbf{v}_h$ is in $\text{Ker } \bar{\mathbf{B}}$.

Second, let $\bar{\mathbf{q}}$ be a vector of DoF's such that $\bar{\mathbf{q}} \in \text{Ker } \bar{\mathbf{B}}^T = \text{Ker } \bar{\mathbf{B}}^\dagger$ (as the entries are real). This means that $\bar{\mathbf{B}}^\dagger \bar{\mathbf{q}} = \mathbf{0}$, i.e., the zero vector in $\mathbb{C}^Q$. It also is not difficult to see that

$$\bar{\mathbf{q}} \in \text{Ker } \bar{\mathbf{B}}^T \Leftrightarrow q_h \in \text{Ker } B^T \quad (4.15.d)$$

where the operator B^T is defined in (3.28).

We now ask under which conditions the solution to the system (4.14.z) is unique. This amounts to showing that the solution to the homogeneous problem

Find $(\bar{\mathbf{e}}, \bar{\mathbf{p}}) \in \mathbb{C}^Q \times \mathbb{C}^M$ such that

$$\begin{bmatrix} \bar{\mathbf{A}} & \bar{\mathbf{B}}^T \\ \bar{\mathbf{B}} & \mathbf{0} \end{bmatrix} \begin{bmatrix} \bar{\mathbf{e}} \\ \bar{\mathbf{p}} \end{bmatrix} = \begin{bmatrix} \mathbf{0} \\ \mathbf{0} \end{bmatrix} \quad (4.15.e)$$

is $(\bar{\mathbf{e}}, \bar{\mathbf{p}}) = (\mathbf{0}, \mathbf{0})$.

The second equation tells us that $\bar{\mathbf{B}}\bar{\mathbf{e}} = \mathbf{0}$, which implies that $\bar{\mathbf{e}} \in \text{Ker } \bar{\mathbf{B}}$. The first equation is

$$\bar{\mathbf{A}}\bar{\mathbf{e}} + \bar{\mathbf{B}}^T\bar{\mathbf{p}} = \mathbf{0}. \quad (4.15.f)$$

In order to get any information regarding $\bar{\mathbf{e}}$ in (4.15.f), the matrix $\bar{\mathbf{A}}$ must be invertible on the kernel of $\bar{\mathbf{B}}$. Then we get

$$\bar{\mathbf{e}} = -\bar{\mathbf{A}}^{-1}\bar{\mathbf{B}}^T\bar{\mathbf{p}}. \quad (4.15.g)$$

From (4.15.g) and the second equation in (4.15.e), we arrive at

$$\bar{\mathbf{B}}\bar{\mathbf{A}}^{-1}\bar{\mathbf{B}}^T\bar{\mathbf{p}} = \mathbf{0}. \quad (4.15.h)$$

To make sure that $(\bar{\mathbf{e}}, \bar{\mathbf{p}}) = (\mathbf{0}, \mathbf{0})$, we need two conditions:

$$1. \text{Ker } \bar{\mathbf{B}}^T = \{\mathbf{0}\}; \quad (4.15.i)$$

$$2. \text{The matrix } \bar{\mathbf{A}} \text{ is invertible on } \text{Ker } \bar{\mathbf{B}}. \quad (4.15.j)$$

The reasoning goes as follows. It can be seen that if $\text{Ker } \bar{\mathbf{B}}^T = \{\mathbf{0}\}$, then $\text{Ker } \bar{\mathbf{B}}\bar{\mathbf{A}}^{-1}\bar{\mathbf{B}}^T = \{\mathbf{0}\}$. Consequently, the linear mapping described by the matrix $\bar{\mathbf{B}}\bar{\mathbf{A}}^{-1}\bar{\mathbf{B}}^T$ in (4.15.h) is one-to-one. From this, one concludes that $\bar{\mathbf{p}} = \mathbf{0}$. If $\bar{\mathbf{p}} = \mathbf{0}$, then $\bar{\mathbf{B}}^T\bar{\mathbf{p}} = \mathbf{0}$; from (4.15.g) we get $\bar{\mathbf{e}} = -\bar{\mathbf{A}}^{-1}\mathbf{0} = \mathbf{0}$, since $\bar{\mathbf{A}}^{-1}$ exists. In this way, $(\bar{\mathbf{e}}, \bar{\mathbf{p}}) = (\mathbf{0}, \mathbf{0})$.

But now we may ask: How can we guarantee that conditions 1 and 2 hold true? The answer: They are consequences of hypotheses (vi) and (vii) in Theorem 4.1. To see why, let us restate the hypothesis (vii), which says that

$$\text{If } a(\mathbf{w}_h, \mathbf{v}_h) = 0, \quad \forall \mathbf{v}_h \in \text{Ker } b, \quad \text{then } \mathbf{w}_h = \mathbf{0}. \quad (4.15.k)$$

When we consider the expansions of $\mathbf{v}_h$ and $\mathbf{w}_h$ together with (4.15.c), we arrive at the equivalent condition expressed in algebraic terms:

$$\text{If } \bar{\mathbf{v}}^\dagger \bar{\mathbf{A}} \bar{\mathbf{w}} = 0, \quad \forall \bar{\mathbf{v}} \in \text{Ker } \bar{\mathbf{B}}, \quad \text{then } \bar{\mathbf{w}} = \mathbf{0}. \quad (4.15.l)$$

Condition above really means

$$\text{If } \bar{\mathbf{A}} \bar{\mathbf{w}} = \mathbf{0} \text{ in } \text{Ker } \bar{\mathbf{B}}, \quad \text{then } \bar{\mathbf{w}} = \mathbf{0}. \quad (4.15.m)$$

In other words, it says: Take any element $\bar{\mathbf{w}}$ from $\text{Ker } \bar{\mathbf{B}}$. If $\bar{\mathbf{A}}\bar{\mathbf{w}} = \mathbf{0}$, then $\bar{\mathbf{w}} = \mathbf{0}$. But this is nothing else than saying that $\bar{\mathbf{A}}$ is injective in $\text{Ker } \bar{\mathbf{B}}$, i.e., there is an inverse $\bar{\mathbf{A}}^{-1}$ well-defined on $\text{Ker } \bar{\mathbf{B}}$. In this way condition 2 in (4.15.j) has been established.

We can make the notion of “ $\bar{\mathbf{A}}$ being invertible on $\text{Ker } \bar{\mathbf{B}}$ ” more understandable. The original matrix $\bar{\mathbf{A}}$ belongs to $\mathbb{C}^{Q \times Q}$, which means that it maps vectors from $\mathbb{C}^Q$ into vectors of $\mathbb{C}^Q$. Since $\text{Ker } \bar{\mathbf{B}} \subset \mathbb{C}^Q$, then $\dim \text{Ker } \bar{\mathbf{B}} = K \leq \dim \mathbb{C}^Q = Q$. Let us find an orthonormal basis for $\text{Ker } \bar{\mathbf{B}}$. Then take $Q - K$ vectors from $\mathbb{C}^Q$ and complete the basis (through a Gram-Schmidt procedure, for example). We now have a new basis for $\mathbb{C}^Q$. In this new basis every element of $\text{Ker } \bar{\mathbf{B}}$ is such that its last $Q - K$ coefficients are all zero. When we represent the matrix $\bar{\mathbf{A}}$ in this new basis, it assumes the form

$$\bar{\mathbf{A}}^{new} = \begin{bmatrix} A_{kk} & A_{kt} \\ A_{tk} & A_{tt} \end{bmatrix} \quad (4.15.n)$$

where the indices k and t are such that $1 \leq k \leq K$ and $(K + 1) \leq t \leq Q$. By invertibility on the kernel what is really meant is that the submatrix A_{kk} is invertible. The question is that to ask for invertibility in the whole space $\mathbb{C}^Q$ may be too much. If $\bar{\mathbf{A}}$ is invertible on the whole space $\mathbb{C}^Q$, good. If not, then requiring just the invertibility on the kernel is fine. For more on this subject, see [Brezzi and Bathe, 1990].

The hypothesis (vi) is just the inf-sup condition (4.15.e). According to the statement (ii) in Theorem 3.3, we know that it is equivalent to the fact that B^T is injective, i.e., that

$$\text{Ker } B^T = \{0_{\mathbb{P}^h(\Omega)}\}, \quad (4.15.o)$$

where $0_{\mathbb{P}^h(\Omega)}$ is the zero element from the space $\mathbb{P}^h(\Omega)$. So $q_h = 0_{\mathbb{P}^h(\Omega)}$ is the only element from $\text{Ker } B^T$; with the help of (4.15.d), we can conclude that $\bar{\mathbf{q}} = \mathbf{0}$ is the only element from $\text{Ker } \bar{\mathbf{B}}^T$, i.e.,

$$\text{Ker } \bar{\mathbf{B}}^T = \{\mathbf{0}\}, \quad (4.15.p)$$

which is precisely the condition 1 in (4.15.i).

The lesson learned so far is that the validity of hypotheses (vi) and (vii) in Theorem 4.1 entail conditions (4.15.i) and (4.15.j), which in their turn imply that the solutions to the final linear system (4.14.z) is unique. It is interesting to track down this chain of influences. First, hypotheses are made at the very abstract level in Theorem 3.9. Second, the abstract spaces and sesquilinear forms from Theorem 3.9 are specialized to the spaces and forms occurring in the scattering problem, as illustrated in Section 3.3.6. Third, these spaces and forms are specialized further to finite-dimensional subspaces in Theorem 4.1. Fourth, these hypotheses are shown to ultimately influence the solvability of the final linear system (4.14.z).

Hypothesis (vii) holds true, according to Section 4.1.1.8, but we have not been able to show that hypothesis (vi) does also, as discussed in Section 4.1.1.7. As we could conclude from this subsection, its validity is fundamental. The state of affairs is such that everything depends on the inf-sup condition (4.6.e). We shall examine it more closely now.

4.2.3 The matrix system: The inf-sup condition

Let it be the inf-sup condition (4.6.e), restated below for convenience:

$$\inf_{q_h \in \mathbb{P}^h(\Omega) \setminus \{0\}} \sup_{\mathbf{v}_h \in \mathbb{V}_\tau^h(\Omega) \setminus \{0\}} \frac{\left| -\int_{\Omega} q_h \nabla \cdot \mathbf{v}_h \right|}{\|\mathbf{v}_h\|_{H^1(\Omega)^3} \|q_h\|_{L^2(\Omega)}} \geq \beta_b^h. \quad (4.16.a)$$

The spaces in (4.16.a) are complex spaces, i.e., $\mathbb{V}_\tau^h(\Omega)$ and $\mathbb{P}^h(\Omega)$ admit elements which have both real and complex parts. According to statement (iii) in Theorem 3.3, the inf-sup condition above is equivalent to the fact that the operator $-\nabla \cdot : \mathbb{V}_\tau^h(\Omega) \rightarrow \mathbb{P}^h(\Omega)^*$ is surjective. Since $\mathbb{P}^h(\Omega)$ is a subset of $L^2(\Omega)$, and since $L^2(\Omega)$ is identified with its dual, there is no harm in identifying $\mathbb{P}^h(\Omega)$ with its dual.

One must then show that the operator $-\nabla \cdot : \mathbb{V}_\tau^h(\Omega) \rightarrow \mathbb{P}^h(\Omega)$ is surjective. In doing so, one does not need to show surjectivity for the complex versions of $\mathbb{V}_\tau^h(\Omega)$ and $\mathbb{P}^h(\Omega)$. Just the real version needs to be taken into account. The reason is as follows. Suppose that $-\nabla \cdot$ is surjective from the (real) $\mathbb{V}_\tau^h(\Omega)$ onto the (real) $\mathbb{P}^h(\Omega)$.

Consider an arbitrary q_h belonging to the (complex) $\mathbb{P}^h(\Omega)$. It means that q_h can be written as $q_h = q_h^R + j q_h^I$, in which both q_h^R and q_h^I are elements from the (real) $\mathbb{P}^h(\Omega)$. From the surjectivity between the real spaces, it follows that there are elements $\mathbf{v}_h^R, \mathbf{v}_h^I$ in (real) $\mathbb{V}_\tau^h(\Omega)$ such that $-\nabla \cdot \mathbf{v}_h^R = q_h^R$ and $-\nabla \cdot \mathbf{v}_h^I = q_h^I$. If we make $\mathbf{v}_h = \mathbf{v}_h^R + j \mathbf{v}_h^I$, then it is true that $-\nabla \cdot \mathbf{v}_h = q_h$. So from an arbitrary q_h in the (complex) $\mathbb{P}^h(\Omega)$, we were able to find a $\mathbf{v}_h$ in the (complex) $\mathbb{V}_\tau^h(\Omega)$ such that $-\nabla \cdot \mathbf{v}_h = q_h$. In other words, we are able to conclude that $-\nabla \cdot$ is surjective from the (complex) $\mathbb{V}_\tau^h(\Omega)$ onto the (complex) $\mathbb{P}^h(\Omega)$. Once we have shown the surjectivity, the inf-sup condition (4.16.a) follows from the Theorem 3.3. Thus far the reasoning is:

Surjectivity between real spaces $\Rightarrow$ Surjectivity between complex spaces $\Rightarrow$ inf-sup condition in complex spaces.

However, how can we prove surjectivity between real spaces? We may resort again to Theorem 3.3: It is equivalent to the inf-sup condition in real spaces. So the whole argument becomes:

Inf-sup condition in real spaces $\Rightarrow$ Surjectivity between real spaces $\Rightarrow$ Surjectivity between complex spaces $\Rightarrow$ inf-sup condition in complex spaces.

So in order to show that (4.16.a) is true, all we need to do is to prove its real counterpart

$$\inf_{q_h^R \in \mathbb{P}^h(\Omega) \setminus \{0\}} \sup_{\mathbf{v}_h^R \in \mathbb{V}_\tau^h(\Omega) \setminus \{0\}} \frac{- \int_{\Omega} q_h^R \nabla \cdot \mathbf{v}_h^R}{\|\mathbf{v}_h^R\|_{H^1(\Omega)^3} \|q_h^R\|_{L^2(\Omega)}} \geq \beta_b^h. \quad (4.16.b)$$

The spaces $\mathbb{V}_\tau^h(\Omega)$ and $\mathbb{P}^h(\Omega)$ in (4.16.b) now admit only real elements. Since these spaces are Hilbert spaces, the norms are induced by the inner products as in

$$\|\mathbf{v}_h^R\|_{H^1(\Omega)^3}^2 = (\mathbf{v}_h^R, \mathbf{v}_h^R)_{H^1(\Omega)^3} \quad (4.16.c)$$

$$\|q_h^R\|_{L^2(\Omega)}^2 = (q_h^R, q_h^R)_{L^2(\Omega)}. \quad (4.16.d)$$

If $\mathbf{v}_h^R$ and q_h^R are expanded as in (4.14.g) and (4.14.i), respectively:

$$\mathbf{v}_h^R = \sum_{i=1}^Q \boldsymbol{\psi}_i \hat{v}_i \quad (4.16.e)$$

$$q_h^R = \sum_{i=1}^M \theta_i \hat{q}_i, \quad (4.16.f)$$

then (4.16.c) and (4.16.d) may be written as

$$\|\mathbf{v}_h^R\|_{H^1(\Omega)^3}^2 = \bar{\mathbf{v}}^T \bar{\mathbf{S}}_E \bar{\mathbf{v}} \quad (4.16.g)$$

$$\|q_h^R\|_{L^2(\Omega)}^2 = \bar{\mathbf{q}}^T \bar{\mathbf{S}}_p \bar{\mathbf{q}}. \quad (4.16.h)$$

The coefficients of the matrices $\bar{\mathbf{S}}_E$ and $\bar{\mathbf{S}}_p$ are given by

$$[\bar{\mathbf{S}}_E]_{ij} = (\boldsymbol{\psi}_i, \boldsymbol{\psi}_j)_{H^1(\Omega)^3} \quad (4.16.i)$$

$$[\bar{\mathbf{S}}_p]_{ij} = (\theta_i, \theta_j)_{L^2(\Omega)} \quad (4.16.j)$$

Since according to (4.14.w)

$$- \int_{\Omega} q_h^R \nabla \cdot \mathbf{v}_h^R = b(\mathbf{v}_h^R, q_h^R) = \bar{\mathbf{q}}^T \bar{\mathbf{B}} \bar{\mathbf{v}}, \quad (4.16.k)$$

the inf-sup condition (4.16.b) becomes: There should be a $\beta_b^h > 0$ such that

$$\inf_{\bar{\mathbf{q}} \in \mathbb{R}^M \setminus \{0\}} \sup_{\bar{\mathbf{v}} \in \mathbb{R}^N \setminus \{0\}} \frac{\bar{\mathbf{q}}^T \bar{\mathbf{B}} \bar{\mathbf{v}}}{(\bar{\mathbf{v}}^T \bar{\mathbf{S}}_E \bar{\mathbf{v}})^{1/2} (\bar{\mathbf{q}}^T \bar{\mathbf{S}}_p \bar{\mathbf{q}})^{1/2}} \geq \beta_b^h. \quad (4.16.l)$$

It can be proved through a formidable algebra [Brezzi and Fortin, 1991], [Bathe, 1996] that

$$\inf_{\bar{\mathbf{q}} \in \mathbb{R}^M \setminus \{0\}} \sup_{\bar{\mathbf{v}} \in \mathbb{R}^N \setminus \{0\}} \frac{\bar{\mathbf{q}}^T \bar{\mathbf{B}} \bar{\mathbf{v}}}{(\bar{\mathbf{v}}^T \bar{\mathbf{S}}_E \bar{\mathbf{v}})^{1/2} (\bar{\mathbf{q}}^T \bar{\mathbf{S}}_p \bar{\mathbf{q}})^{1/2}} = \mu_{min}, \quad (4.16.m)$$

where μ_{min} is the *smallest eigenvalue* associated with the problem

$$\overline{\mathbf{B}}\overline{\mathbf{S}}_E^{-1}\overline{\mathbf{B}}^T\overline{\mathbf{w}}_i = \mu_i^2\overline{\mathbf{S}}_p\overline{\mathbf{w}}_i \quad (4.16.n)$$

In essence, this is the numerical evaluation of the inf-sup condition we mentioned in Section 4.1.1.7. Given a pair of finite-dimensional subspaces $\mathbb{V}_\tau^h(\Omega)$ and $\mathbb{P}^h(\Omega)$, from their real basis functions in (4.14.b) and (4.14.c), we construct the real matrices $\overline{\mathbf{B}}$, $\overline{\mathbf{S}}_E$ and $\overline{\mathbf{S}}_p$. Thereafter, we look for the smallest eigenvalue μ_{min} of the generalized eigenvalue problem (4.16.n). The quantity in the left side of (4.16.b) is given precisely by this value. Then we must verify: If $\mu_{min} > 0$, then the pair $\mathbb{V}_\tau^h(\Omega)/\mathbb{P}^h(\Omega)$ satisfies the inf-sup condition (4.16.b), and their associate inf-sup constant is therefore $\beta_b^h = \mu_{min}$. On the other hand, if $\mu_{min} = 0$, the pair $\mathbb{V}_\tau^h(\Omega)/\mathbb{P}^h(\Omega)$ does not satisfy the inf-sup condition.

Furthermore, if $\mathbb{V}_\tau^h(\Omega)/\mathbb{P}^h(\Omega)$ satisfies the inf-sup condition for real spaces in (4.16.b), then it follows from the argument presented earlier that it also satisfies the inf-sup condition for the complex spaces in (4.16.a), which is nothing else than the hypothesis (vi) in Theorem 4.1.

So the hypothesis (vi) in Theorem 4.1 is not actually proved; it is *verified* at the numerical level. Of course, different choices for $\mathbb{V}_\tau^h(\Omega)/\mathbb{P}^h(\Omega)$ lead to different inf-sup constants β_b^h ; hence the superscript h , to indicate that it depends on the specific finite-dimensional subspaces considered. The numerical test allows a certain freedom in the construction of $\mathbb{V}_\tau^h(\Omega)$ and $\mathbb{P}^h(\Omega)$. Indeed, before solving the scattering problem, we can construct different pairs and test if they satisfy the inf-sup condition. This test turns out to be the ideal one to deal with meshfree methods. As one knows, the subspaces there are spanned by basis functions generated by clouds of nodes distributed (at least in principle) throughout the domain in a more or less disordered way.

4.3 Meshfree subspaces

4.3.1 Nodes and patches

It is now time to specify the spaces $\mathbb{V}_\tau^h(\Omega)$ and $\mathbb{P}^h(\Omega)$ further. In the sequel, we will look for subspaces generated by meshfree basis functions. The formulation we develop thus leads to a ‘meshfree method’, if by method we mean the way the subspaces are constructed. Interestingly enough, the discussion thus far has not made any reference to something being qualified as ‘meshfree’. The whole formulation, theorems, hypotheses and even the final form of the matrix system do not depend on $\mathbb{V}_\tau^h(\Omega)$ and $\mathbb{P}^h(\Omega)$ being meshfree or not. What does depend is the specific form assumed by these finite-dimensional subspaces and their ability to provide an approximate solution to the scattering problem.

As stated in Chapter 1, the ‘method’ to be used in this work is basically the method of finite spheres (MFS) [De and Bathe, 2000], with some modifications here and there. We begin by describing our computational domain Ω . In principle, it is the same as the domain in which the system of differential equations is stated; even when curved boundaries are considered, it is not approximated by flat faces of triangles/tetrahedra as it happens in FEM. In this way Ω is just an open and connected subset of $\mathbb{R}^d$, where $d = 2$ or $d = 3$.

After the domain has been defined, one spreads *nodes* over Ω and also on its boundary $\partial\Omega = \Gamma$. Nodes are simple points; sometimes they are referred to as *particles*. They are spread freely over Ω ; by freely one means that there is no fixed rules their distribution should follow. (Saying that random distributions are allowed is a little bit nonsensical, but there is nothing wrong with *quasi*-random distributions.)

These nodes must be numbered, or labeled. They are usually ordered according to the natural numbers, so we talk of node 1, node 2, and so on. The *index* of a node is nothing more than the natural number to which it is associated. The number of nodes shall be finite; so in a sense there will be a total of N of them. Each node is described by its Cartesian coordinates; for example, a node with index I is located at position $\mathbf{x}_I = [x_I, y_I, z_I]^T$, $1 \leq I \leq N$.

To each node I we associate an open set Ω_I , also called a *patch*. In this work, each patch is a square ($d = 2$) or a cube ($d = 3$). The node and the patch are arranged in such a way that the node is located at the center of the patch. In these circumstances, the patch Ω_I is just the subset of $\mathbb{R}^d$ given by

$$\begin{aligned} \Omega_I = \{ \mathbf{x} \in \mathbb{R}^d \mid & x_I - r_I < x < x_I + r_I, \\ & y_I - r_I < y < y_I + r_I, \\ & z_I - r_I < z < z_I + r_I \}. \end{aligned} \quad (4.17.a)$$

The number r_I is a measure of the size of the patch Ω_I . According to (4.17.a), the side of Ω_I is given by $2r_I$. The patches can overlap with each other (if nodes I and J are close enough, probably $\Omega_I \cap \Omega_J \neq \emptyset$). Also, some portions of Ω_I may even be outside the computational domain Ω (as it happens for the nodes located at the boundary Γ , for example).

But there are requirements these patches must satisfy. They must form a covering for $\bar{\Omega}$. In other words,

$$\bar{\Omega} \subset \bigcup_{I=1}^N \Omega_I. \quad (4.17.b)$$

Expression above means that, any point x in $\bar{\Omega}$ (i.e., either in the interior Ω or at the boundary Γ) must belong to at least one patch Ω_I . In other words, the patches cover the domain Ω and its boundary Γ in such a way that *no holes* are left behind.

Each patch Ω_I presents itself as a nice environment to define certain functions, i.e., we can construct functions which are defined only in the interior of the patch Ω_I . So lets us represent these local functions as

$$\ell_{I,m}: \Omega_I \rightarrow \mathbb{R}, \quad 1 \leq m \leq \#_I \quad (4.17.c)$$

i.e., these local functions (hence the “ ℓ ”) are real-valued and defined *only within* Ω_I . In a patch Ω_I there are $\#_I$ local functions, labeled as $\ell_{I,1}$, $\ell_{I,2}$, and so on. They must be *linearly independent*, but are not required to be orthogonal to each other in any sense.

We can now introduce a local space V_I , spanned by the ℓ_{Im} ’s as

$$V_I = \text{span}\{\ell_{I,1}, \ell_{I,2}, \dots, \ell_{I,\#_I}\}, \quad 1 \leq I \leq N \quad (4.17.d)$$

So each patch has its corresponding local space. In this way, there will be a total of N local spaces.

As it stands, these local spaces are ‘loose’ in the sense that they do not, at first sight, incorporate information concerning the underlying nodal distribution. In other words, it is not clear how the distribution of neighbor nodes influences the local functions defined on a patch.

In fact, it does not. The functions in (4.17.c) are entirely local, and generally do not incorporate information regarding the neighboring nodes. All the local spaces must be ‘glued together’ in order to form a coherent structure which takes both the local spaces and the nodal distribution into account.

This ‘gluing’ is provided by the partition of unity (PU), which is defined below [De and Bathe, 2000]:

Chart 4.1: Partition of unity (PU)

Let Ω be a bounded domain in $\mathbb{R}^d$. Consider a family of open subsets $\{\Omega_I\}_{I=1}^N$ which forms a covering for Ω , i.e., they are such that

$$\bar{\Omega} \subset \bigcup_{I=1}^N \Omega_I \quad (4.17.e)$$

Then there exists a system of functions $\{\phi_I\}_{I=1}^N \subset C_0^m(\mathbb{R}^d)$, $m \geq 0$ which satisfy the two properties below:

$$\sum_{I=1}^N \phi_I(x) = 1, \quad \forall x \in \bar{\Omega} \quad (4.17.f)$$

$$\text{supp}(\phi_I) \subset \bar{\Omega}_I, \quad 1 \leq I \leq N. \quad (4.17.g)$$

This system of functions $\{\phi_I\}_{I=1}^N$ is called the *partition of unity* subordinate to $\{\Omega_I\}_{I=1}^N$.

We may take the family of open sets $\{\Omega_I\}_{I=1}^N$ as the collection of all cubic patches we defined over Ω , according to (4.17.a). The definition above claims the existence of a certain set of functions in $C_0^m(\mathbb{R}^d)$, a space defined by

$$C_0^m(\mathbb{R}^d) = \{u \in C^m(\Omega) \mid \text{supp}(u) \subset\subset \mathbb{R}^d\}, \quad (4.17.h)$$

where $C^m(\Omega)$ is given in (2.52) and the notion of support is introduced in (2.1). In a sense, every function ϕ_I is m -times continuously differentiable, and its support is a closed subset of $\mathbb{R}^d$. The exact value of m depends on the way the PU is generated; the definition above only acknowledges the existence of a system of continuous functions which satisfy (4.17.f) and (4.17.g).

It is likely that each point $\mathbf{x}$ in the domain Ω is within more than one patch. Property (4.17.f) says that the sum of the functions ϕ_I calculated at $\mathbf{x}$ is always 1. Since $\phi_I \in C_0^m(\mathbb{R}^d)$, its support is a closed subset of $\mathbb{R}^d$. But property (4.17.g) refines this knowledge: It says that the support of ϕ_I is *compactly contained* in the patch Ω_I , i.e., it is a closed subset entirely contained within Ω_I (but it can touch the boundary Γ , though, as it happens for nodes located on or very close to it).

The method of finite spheres is based on a family of non-polynomial PU functions. Let w be a quartic spline weight (or window) function [Duarte and Oden, 1996]:

$$w(s) = \begin{cases} 1 - 6s^2 + 8s^3 - 3s^4, & 0 \leq s < 1 \\ 0, & s \geq 1. \end{cases} \quad (4.17.i)$$

Then a partition of unity can be constructed by tensor-product Shepard functions as

$$\varphi_I^0(\mathbf{x}) = \frac{w\left(\frac{|\mathbf{x} - \mathbf{x}_I|}{r_I}\right) w\left(\frac{|\mathbf{y} - \mathbf{y}_I|}{r_I}\right) w\left(\frac{|\mathbf{z} - \mathbf{z}_I|}{r_I}\right)}{\sum_{J=1}^N w\left(\frac{|\mathbf{x} - \mathbf{x}_J|}{r_J}\right) w\left(\frac{|\mathbf{y} - \mathbf{y}_J|}{r_J}\right) w\left(\frac{|\mathbf{z} - \mathbf{z}_J|}{r_J}\right)}. \quad (4.17.j)$$

An example of a typical Shepard PU function is illustrated in Fig. 4.1.

It can be seen that the system $\{\varphi_I^0\}_{I=1}^N$ thus obtained satisfies (4.17.f) and (4.17.g). The function w in (4.17.i) belongs to $C^1([0,1])$, and so each triple product in (4.17.j) belongs to $C^1(\Omega)$. Consequently, both the numerator and the denominator in (4.17.j) belong to $C^1(\Omega)$. The denominator never blows up, as the weight function w attains a maximum value of 1. In this way, the derivative of $\varphi_I^0(\mathbf{x})$ is also continuous, and therefore we conclude that φ_I^0 is (at least) in $C^1(\Omega)$.

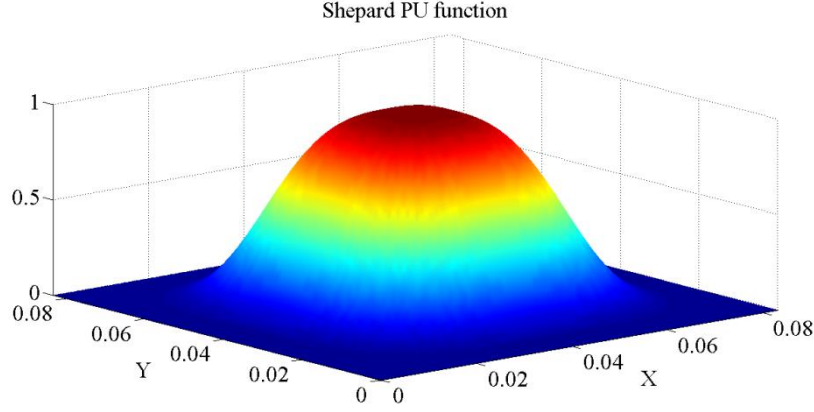


Fig. 4.1. A typical Shepard PU function φ_I^0 over a two-dimensional square patch. According to (4.18.c) in Chart 4.2, φ_I^0 attains the value 1 at the node location.

The Shepard PU functions φ_I^0 are compactly supported, as its support is contained within $\bar{\Omega}_I$. Moreover, they have zero-order consistency, i.e., they can reproduce constant functions exactly (hence the superscript 0). Higher-order consistency is provided by the functions in the local spaces (4.17.d), as will be explained in the next pages.

It is now time to ‘glue’ the local spaces (4.17.d) and the PU together. The result is a global approximation space, constructed as follows. For each local space V_I in (4.17.d), we form its ‘weighted’ version as

$$\varphi_I^0 V_I = \text{span}\{\varphi_I^0 \ell_{I,1}, \varphi_I^0 \ell_{I,2}, \dots, \varphi_I^0 \ell_{I,\#_I}\}, \quad (4.17.k)$$

i.e., the local functions defined in the patch Ω_I are multiplied by its corresponding Shepard PU function φ_I^0 . Of course, the support of the functions in $\varphi_I^0 V$ is the same as the support of φ_I^0 (i.e., the functions in the local space become ‘confined within the patch through multiplication by a function which ‘exists’ only on the patch).

If we consider two weighted local spaces $\varphi_I^0 V_I$ and $\varphi_J^0 V_J$, it is not difficult to see that they are linearly independent, since their elements are functions defined in different regions Ω_I and Ω_J . The global approximation space is just the sum of these weighted subspaces:

$$V = \varphi_1^0 V_1 \oplus \varphi_2^0 V_2 \oplus \dots \oplus \varphi_N^0 V_N \quad (4.17.l)$$

If $v \in V$, then it is represented by the double sum

$$v = \sum_{I=1}^N \sum_{m=1}^{\#_I} \varphi_I^0 \ell_{I,m} \hat{v}_{Im}, \quad (4.17.m)$$

where I runs through all nodes and m runs through all local functions of the local space V_I [whose dimension is $\#_I$, according to (4.17.c)]. The scalars $\hat{v}_{Im}$ are the DoF’s in the expansion. So the true shape or basis functions of our method is

$$h_{Im} := \varphi_I^0 \ell_{I,m} , \quad (4.17.n)$$

formed by the Shepard PU function φ_I^0 multiplied by the local function $\ell_{I,m}$. So we may rewrite (4.17.m) as

$$v(\mathbf{x}) = \sum_{I=1}^N \sum_{m=1}^{\#I} h_{Im}(\mathbf{x}) \hat{v}_{Im} , \quad (4.17.o)$$

which gives the right expansion at a point $\mathbf{x}$.

There is an important result concerning the meshfree spaces V in (4.17.l): They span a subset of $H^1(\Omega)$ [De and Bathe, 2001].

Proposition 4.1: On the global meshfree spaces V – Suppose that a polynomial basis (of order m) is included in every local space V_I , i.e., $Q_m \subset V_I, \forall I$. Then the global space V defined as in (4.17.l), i.e., as

$$V = \varphi_1^0 V_1 \oplus \varphi_2^0 V_2 \oplus \cdots \oplus \varphi_N^0 V_N \quad (4.17.p)$$

is a subspace of $H^1(\Omega)$. In other words,

$$V \subset H^1(\Omega) \quad (4.17.q)$$

In what regards the derivatives of the basis functions in (4.17.n), there is nothing new (provided only differentiable functions are included in the local basis). The ordinary chain rule works fine:

$$\frac{\partial}{\partial x} h_{Im} = \frac{\partial}{\partial x} (\varphi_I^0 \ell_{I,m}) = \frac{\partial \varphi_I^0}{\partial x} \ell_{I,m} + \varphi_I^0 \frac{\partial \ell_{I,m}}{\partial x}. \quad (4.17.r)$$

The same reasoning is extended to the derivatives with respect to y and z .

The meshfree basis functions h_{Im} have nice properties. First, they are compactly supported, which means that a discretization process based on them leads to sparse linear systems. Second, they do not depend on matrix inversions as the Moving Least Squares does [Liu, 2010]. Third, since they do not depend on the distribution of neighbor nodes as the MLS (the PU shape functions are influenced by neighboring nodes, but they do not depend on them in order to be well-defined), the patches can be made as small as possible, just enough to satisfy the covering criterion (4.17.b). This is true regardless of what one decides to include in the local spaces, and is in stark contrast to the MLS, where if one decides to include higher-order terms in the process, then the ‘influence domains’ must be made larger in order to encompass a larger number of neighboring nodes. Fourth, they satisfy the reproducibility/consistency properties below, stated as a theorem [Melenk and Babuska, 1996]:

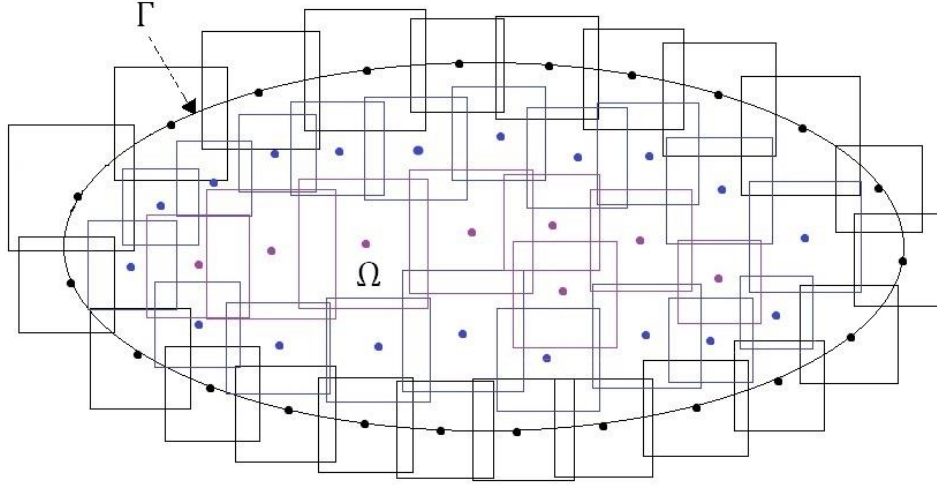


Fig. 4.2. In two dimensions, the set of square patches must form a covering for the computational domain Ω and its boundary Γ . According to consideration 1 in Chart 4.2, there is only one node per patch. The patches can be made as small as possible, just enough to not leave any hole behind. The extension to 3D is straightforward; we just need to substitute squares for cubes.

Theorem 4.2: Reproducibility/Compatibility – If any function $f(\mathbf{x})$ is included in the local bases, it is possible to exactly reproduce it. Moreover, if $Q_m \subset V_I, \forall I$, then $Q_m \in V$.

In Theorem 4.2, Q_m is the space spanned by all polynomials of degree less than or equal to m . The last statement says that if Q_m is a subspace of all local spaces V_I , then it is also a subspace of the global space V . In other words, if we include, for example, the terms $\{1, x, y\}$ in every local space V_I , then the global space V will be able to reproduce exactly any function which is a linear combination of $\{1, x, y\}$, namely, it will reproduce exactly any linear function defined on Ω .

4.3.2 Geometrical considerations

The properties of the basis functions h_{Im} are good, but we can make them even better. In this work, we propose three considerations.

Chart 4.2: On the improvement of the basis functions h_{Im}

1. The size of the square/cubic patch Ω_I is such that

$$\text{If } J \neq I \text{ then } x_I \notin \bar{\Omega}_J, \quad 1 \leq J \leq N. \quad (4.18.a)$$

Expression (4.18.a) says that the only patch in which node I is contained is the patch Ω_I itself. Equivalently, there is only one node per patch. This is illustrated in Fig. 4.2.

Since the patches can be made as small as one desires [but always keeping (4.17.b) in mind], one can decrease the sizes of the other patches Ω_J so that their boundaries $\partial\Omega_J$ become very close to node $\mathbf{x}_I$, but do not need to actually touch $\mathbf{x}_I$.

We know that the partition of unity (4.17.f) holds for any point in $\bar{\Omega}$. Particularly, it holds in the location of node I at $\mathbf{x}_I$:

$$\sum_{j=1}^N \varphi_j^0(\mathbf{x}_I) = 1 \quad (4.18.b)$$

For each node J (4.17.g) says that $\text{supp}(\varphi_J^0) \subset \bar{\Omega}_J$, i.e., that the support of the PU function φ_J^0 is contained in the patch J (in other words, the PU function φ_J^0 only ‘exists’ within the patch J). But according to (4.18.a), if $J \neq I$, then $\mathbf{x}_I \notin \bar{\Omega}_J$. Since $\text{supp}(\varphi_J^0) \subset \bar{\Omega}_J$, it is also true that if $J \neq I$, then $\mathbf{x}_I \notin \text{supp}(\varphi_J^0)$. But if $\mathbf{x}_I$ is not in the support of φ_J^0 , then $\varphi_J^0(\mathbf{x}_I) = 0$. We are thus able to conclude that the sum in (4.18.b) reduces to a single term: that for which $J = I$. Then,

$$\varphi_I^0(\mathbf{x}_I) = 1 \quad (4.18.c)$$

Expression above says that for any I , the PU function φ_I^0 evaluated at $\mathbf{x}_I$ is equal to 1, or that

$$\varphi_J^0(\mathbf{x}_I) = \delta_{IJ} \quad (4.18.d)$$

where δ_{IJ} is the Kronecker delta.

2. For any node I located at the interior of the domain Ω (i.e., not at the global boundary Γ), the patch Ω_I is such that it does not intercept Γ . Symbolically,

$$\text{If } \mathbf{x}_I \notin \Gamma \quad \text{then} \quad \Omega_I \cap \Gamma = \emptyset \quad (4.18.e)$$

In this way, any function, in the course of the meshfree discretization process, has its behavior at the boundary Γ governed by the boundary nodes only.

3. If the node I is located at a portion of the global boundary Γ in which Dirichlet boundary conditions are prescribed, then $\ell_{I,1}(\mathbf{x}) = 1$ is the only term to be included in the local basis. In other words,

$$\text{If } \mathbf{x}_I \in \Gamma_D \quad \text{then} \quad V_I = \text{span}\{1\} \quad (4.18.f)$$

The considerations above have a positive influence when handling Dirichlet boundary conditions. Suppose we are trying to find a meshfree approximation to the solution of a problem in which Dirichlet boundary conditions have been prescribed, as in

$$u|_{\Gamma_D} = g, \quad (4.18.g)$$

where u is a scalar unknown (for example, a component of some vector field $\mathbf{u}$) and g is a known function (the essential condition). Let I be a node in the Dirichlet boundary, i.e., $\mathbf{x}_I \in \Gamma_D$. Since $\mathbf{x}_I \in \Gamma_D$, then

$$u(\mathbf{x}_I) = g(\mathbf{x}_I) \quad (4.18.h)$$

If we expand u at $\mathbf{x}_I$ in terms of basis functions as in (4.17.o), we get

$$u(\mathbf{x}_I) = \sum_{j=1}^N \sum_{m=1}^{\#_j} h_{jm}(\mathbf{x}_I) \hat{u}_{jm}, \quad (4.18.i)$$

According to (4.18.a), the only patch to which the nodal point $\mathbf{x}_I$ belongs is Ω_I , and so the outer sum in (4.18.i) has a single term, namely, I . Then,

$$u(\mathbf{x}_I) = \sum_{m=1}^{\#_I} h_{Im}(\mathbf{x}_I) \hat{u}_{Im} \quad (4.18.j)$$

If we take (4.17.n) into account,

$$u(\mathbf{x}_I) = \sum_{m=1}^{\#_I} \varphi_I^0(\mathbf{x}_I) \ell_{I,1}(\mathbf{x}_I) \hat{u}_{Im} \quad (4.18.k)$$

But $\varphi_I^0(\mathbf{x}_I) = 1$, according to (4.18.c), and (4.18.f) tells us that $\ell_{I,1}(\mathbf{x}) = 1$ is the only term in the local basis for V_I . Therefore,

$$u(\mathbf{x}_I) = \hat{u}_{I1}, \quad (4.18.l)$$

i.e., the DoF $\hat{u}_{I1}$ is the function u evaluated at $\mathbf{x}_I$. When we combine (4.18.l) and (4.18.h), we find that

$$\text{If } \mathbf{x}_I \in \Gamma_D \text{ then } \hat{u}_{I1} = g(\mathbf{x}_I) \quad (4.18.m)$$

To summarize: A node at the Dirichlet boundary has a single term in its local basis, and consequently a single DoF in the meshfree expansion. It turns out that this DoF is precisely the value of the known function g evaluated at the node location.

The conclusion we arrived at (4.18.m) has striking consequences in the construction of the lifting function associated with Dirichlet boundary conditions. Suppose we want to solve a scalar problem

Find $u_h \in U^h$ such that

$$\mathcal{D}(u_h, v_h) = F(v_h), \quad \forall v_h \in V^h \quad (4.18.n)$$

$$u_h|_{\Gamma_D} = g \quad (4.18.o)$$

where u is some scalar unknown and $\mathcal{D}$ is some differential operator in weak form. The solution is to be sought in the set U of admissible functions. Elements of U^h satisfy (4.18.o), whereas $v|_{\Gamma_D} = 0$ for any testing function $v \in V^h$. The philosophy of the lifting procedure is to write the solution u as

$$u_h = u_h^0 + u_h^g, \quad (4.18.p)$$

where $u_h^0 \in V^h$ and u_h^g is any function satisfying $u_h^g|_{\Gamma_D} = g$. Therefore it makes sense to take u_h^g as the *easiest* function to construct. In the discrete level, this easiest function can be constructed according to the following recipe outlined in the Chart 4.3.

Chart 4.3: The lifting function u_h^g

Let it be the N nodes spread throughout the domain Ω and on its boundary Γ . Suppose also that there is a portion Γ_D of the boundary in which the Dirichlet condition (4.18.o) holds. We can construct a numerical lifting function u_h^g as follows:

1. The function u_h^g admits the traditional meshfree expansion

$$u_h^g(\mathbf{x}) = \sum_{l=1}^N \sum_{m=1}^{\#_l} h_{lm}(\mathbf{x}) \hat{u}_{lm}. \quad (4.18.q)$$

2. The DoF's $\hat{u}_{lm}$ are determined according to the rule:

2.a. If l is an interior node, i.e., if $\mathbf{x}_l \notin \Gamma$, then $\hat{u}_{lm} = 0$.

2.b. If l lies at the boundary Γ but not at the Dirichlet boundary Γ_D , i.e., if $\mathbf{x}_l \notin \Gamma_D$, then $\hat{u}_{lm} = 0$.

2.c. If l lies at the Dirichlet boundary Γ_D , i.e., if $\mathbf{x}_l \in \Gamma_D$, then $\hat{u}_{l1} = g(\mathbf{x}_l)$. (Remember that according to the consideration 3 in Chart 4.2, nodes at the Dirichlet boundary have a single DoF.)

After the lifting function u_h^g is found, substitution of (4.18.p) in (4.18.n) produces a new problem

$$\begin{aligned} &\text{Find } u_h^0 \in V^h \text{ such that} \\ &\mathcal{D}(u_h^0, v_h) = F(v_h) - \mathcal{D}(u_h^g, v_h), \quad \forall v_h \in V^h \end{aligned} \quad (4.18.r)$$

where both the solution and testing functions belong to same space V^h (whose elements satisfy homogeneous Dirichlet boundary conditions).

The procedure of finding a lifting function thus becomes a very easy task to do at a numerical level, thanks to the considerations we have made at Chart 4.2.

It turns out that the procedure for finding lifting functions for vector quantities $\mathbf{u}_h$ is equally easy. We just need to apply the procedure just outlined to each of the scalar components of $\mathbf{u}_h$.

4.3.3 The spaces $\mathbb{V}_\tau^h(\Omega)$ and $\mathbb{P}^h(\Omega)$

According to (4.14.b) and (4.14.c), the spaces $\mathbb{V}_\tau^h(\Omega)$ and $\mathbb{P}^h(\Omega)$ are

$$\mathbb{V}_\tau^h(\Omega) = \text{span}\{\boldsymbol{\psi}_1, \boldsymbol{\psi}_2, \dots, \boldsymbol{\psi}_N\}, \quad (4.19.a)$$

$$\mathbb{P}^h(\Omega) = \text{span}\{\theta_1, \theta_2, \dots, \theta_M\}. \quad (4.19.b)$$

A problem of paramount importance to us is this: With our meshfree basis functions, we are able to produce meshfree spaces V given by (4.17.l). How can we construct the meshfree spaces $\mathbb{V}_\tau^h(\Omega)$ and $\mathbb{P}^h(\Omega)$ above out from the spaces V in (4.17.l)? The answer is easy for the space $\mathbb{P}^h(\Omega)$, but is not clear for $\mathbb{V}_\tau^h(\Omega)$.

We can make the discussion a little bit more precise and introduce some distinctions. We would like to point out that in the course of developing the meshfree space $\mathbb{P}^h(\Omega)$, we want to emphasize the fact that the local basis functions reflect the choice of functions concerning the pseudopressure p . So the local spaces will be represented by

$$V_I^p = \text{span}\{\ell_{I,1}^p, \ell_{I,2}^p, \dots, \ell_{I,M_I}^p\}, \quad 1 \leq I \leq N \quad (4.19.c)$$

where M_I denotes the number of basis functions which span the local space V_I^p in (4.19.c). The superscript ‘ p ’ reflects the fact that the local basis functions are related to the pseudopressure p . In this way, the global approximation space becomes

$$V^p = \varphi_1^0 V_1^p \oplus \varphi_2^0 V_2^p \oplus \dots \oplus \varphi_N^0 V_N^p \quad (4.19.d)$$

This space is a subspace of $H^1(\Omega)$, according to (4.17.q). Since $H^1(\Omega)$ is a subspace of $L^2(\Omega)$, it follows that $V^p \subset L^2(\Omega)$, as required by (4.1.b). Therefore it is valid to choose

$$\mathbb{P}^h(\Omega) = V^p. \quad (4.19.e)$$

If $p \in \mathbb{P}^h(\Omega)$, then it admits the meshfree expansion

$$p(\mathbf{x}) = \sum_{I=1}^N \sum_{m=1}^{M_I} h_{Im}^p(\mathbf{x}) \hat{p}_{Im}, \quad (4.19.f)$$

where

$$h_{Im}^p(\mathbf{x}) = \varphi_I^0(\mathbf{x}) \ell_{I,m}^p(\mathbf{x}), \quad (4.19.g)$$

Now we have two representations for $\mathbb{P}^h(\Omega)$: (4.19.b) and (4.19.f). According to (4.19.b), a function p in $\mathbb{P}^h(\Omega)$ has the expansion

$$p = \sum_{j=1}^M \theta_j \hat{\alpha}_j \quad (4.19.h)$$

where the $\hat{\alpha}_j$'s are the DoF's. When we compare (4.18.q) and (4.18.s), we discover that the basis functions θ_j for $\mathbb{P}^h(\Omega)$ are just the (double-indexed) h_{lm}^p . If we write them in order, we are able to see that

$$\theta_1 = h_{11}^p \quad (4.19.i)$$

$$\theta_2 = h_{12}^p$$

$$\theta_3 = h_{13}^p$$

$$\vdots$$

$$\theta_{M_1} = h_{1M_1}^p$$

$$\theta_{M_1+1} = h_{21}^p$$

and so on.

So if $\dim V_1^p = M_1$, $\dim V_2^p = M_2, \dots, \dim V_N^p = M_N$, then the dimension M of the global space $\mathbb{P}^h(\Omega)$ is

$$\dim \mathbb{P}^h(\Omega) = M = M_1 + M_2 + \dots + M_N \quad (4.19.j)$$

The situation is more complicated for the space $\mathbb{V}_\tau^h(\Omega)$. Let us denote by V_I^e the local space whose basis functions are related to a scalar component of the electric field:

$$V_I^e = \text{span}\{\ell_{I,1}^e, \ell_{I,2}^e, \dots, \ell_{I,Q_I}^e\}, \quad 1 \leq I \leq N \quad (4.19.k)$$

So when it comes to a component of the electric field, each patch I has Q_I functions, which span the local space V_I^e . The global approximation space then becomes

$$V^e = \varphi_1^0 V_1^e \oplus \varphi_2^0 V_2^e \oplus \dots \oplus \varphi_N^0 V_N^e. \quad (4.19.l)$$

The dimension of V^e is given by

$$\dim V^e = Q_1 + Q_2 + \dots + Q_N. \quad (4.19.m)$$

According to (4.17.q), it is true that

$$V^e \subset H^1(\Omega). \quad (4.19.n)$$

We would like to say once more that we use superscripts because the terms included in the local spaces for the pseudopressure will be different from those included

in the local spaces for the components of the electric field. This distinction is summarized in the tables below.

TABLE 4.1 – LOCAL BASES AND LOCAL SPACES

	<i>Pseudopressure p</i>	<i>Scalar component of the electric field</i>
Local space	V_I^p	V_I^e
Dimension	M_I	Q_I
Terms	$\ell_{I,1}^p, \ell_{I,2}^p, \dots, \ell_{I,M_I}^p$	$\ell_{I,1}^e, \ell_{I,2}^e, \dots, \ell_{I,Q_I}^e$

TABLE 4.2 – GLOBAL SPACES

	<i>Pseudopressure p</i>	<i>Scalar component of the electric field</i>
Global space	$\varphi_1^0 V_1^p \oplus \dots \oplus \varphi_N^0 V_N^p$	$\varphi_1^0 V_1^e \oplus \dots \oplus \varphi_N^0 V_N^e$
Dimension	$M_1 + M_2 + \dots + M_N$	$Q_1 + Q_2 + \dots + Q_N$

In other words, for a given patch I , the functions $\ell_{I,1}^p, \dots, \ell_{I,M_I}^p$ will be different from the $\ell_{I,1}^e, \dots, \ell_{I,Q_I}^e$.

Now that the meshfree space for $\mathbb{P}^h(\Omega)$ has been defined in (4.19.e), we must turn to the construction of $\mathbb{V}_\tau^h(\Omega)$. The problem is not easy. One could begin by trying to find a basis for $\mathbb{E}^h(\Omega)$ in (4.1.a) as follows. Let the x, y and z components be elements from V^e in (4.19.l). This amounts to making

$$\mathbb{E}^h(\Omega) = V^e \times V^e \times V^e. \quad (4.19.o)$$

Thereafter one could make the DoF's associated with the tangential components equal to zero. In this way, we get a suitable meshfree space for $\mathbb{V}_\tau^h(\Omega)$, introduced in (4.3) and rewritten below:

$$\mathbb{V}_\tau^h(\Omega) = \{\mathbf{v}_h \in \mathbb{E}^h(\Omega) \mid \boldsymbol{\gamma}_t \mathbf{v}_h = \mathbf{0}\}. \quad (4.19.p)$$

For example, suppose that Ω is a cube. In the upper face, the outward normal direction is $\mathbf{z}$. For every node located on this face, we make the DoF's of the x and y components equal to zero. So every element of the resulting space has zero tangential components on this face. The same applies to the other faces of the cube.

The problem is that this approach is limited to domains with ‘rectangular’ boundaries, i.e., boundaries which are described by flat faces. Let us say we are interested in solving a problem in a spherical domain. In the spherical surface, the tangential vectors are not described by one of the Cartesian directions only. So we

cannot get fields which have no tangential components by just making the DoF's associated with either x or y or z equal to zero.

We want a way to get spaces of vectors having no tangential components in *any* geometry, because the PEC surface of the scatterer can have an arbitrary shape. We found a solution to this problem. The ‘discovery’ of a meshfree representation of (4.19.p) is one of the most important achievements of this work. It will be described next.

As we said in Section 4.3.1, we begin by spreading N nodes over the domain Ω and also on its boundary Γ . Each node is associated to a cubic patch, whose construction is detailed in (4.17.a). On each of these patches, we defined local spaces V_I^e as in (4.19.k), which are ‘glued together’ via the PU functions in order to produce the global space V^e (4.19.l). This procedure is able to find a meshfree space for a scalar quantity, which can be a component of the scattered electric field. One may ask: Which component? The x -component? Or the y -component? The z -component, maybe? The answer is: None of these. The electric field will not be expanded in terms of the Cartesian components.

Let us add more structure. To each node I , we will associate *three* directions, called the *elemental directions*. They are just unit vectors in $\mathbb{R}^3$, and will be represented by $\hat{\mathbf{a}}_I$, $\hat{\mathbf{b}}_I$ and $\hat{\mathbf{c}}_I$. We require them to be mutually orthogonal, i.e.:

$$\hat{\mathbf{a}}_I \cdot \hat{\mathbf{a}}_I = \hat{\mathbf{b}}_I \cdot \hat{\mathbf{b}}_I = \hat{\mathbf{c}}_I \cdot \hat{\mathbf{c}}_I = 1, \quad (4.19.q)$$

$$\hat{\mathbf{a}}_I \cdot \hat{\mathbf{b}}_I = \hat{\mathbf{b}}_I \cdot \hat{\mathbf{c}}_I = \hat{\mathbf{c}}_I \cdot \hat{\mathbf{a}}_I = 0. \quad (4.19.r)$$

The elemental directions are determined as follows: If a node I is an interior node, then they are just the Cartesian directions $\hat{\mathbf{x}}$, $\hat{\mathbf{y}}$ and $\hat{\mathbf{z}}$. If, on the other hand, the node I is a boundary node, then they are the normal and tangential directions at $\mathbf{x}_I$. In other words,

$$\text{If } \mathbf{x}_I \in \Omega \quad \text{then} \quad \begin{cases} \hat{\mathbf{a}}_I = \hat{\mathbf{x}} \\ \hat{\mathbf{b}}_I = \hat{\mathbf{y}} \\ \hat{\mathbf{c}}_I = \hat{\mathbf{z}} \end{cases} \quad (4.19.s)$$

$$\text{If } \mathbf{x}_I \in \Gamma \quad \text{then} \quad \begin{cases} \hat{\mathbf{a}}_I = \hat{\mathbf{n}}(\mathbf{x}_I) \\ \hat{\mathbf{b}}_I = \hat{\mathbf{t}}_1(\mathbf{x}_I) \\ \hat{\mathbf{c}}_I = \hat{\mathbf{t}}_2(\mathbf{x}_I) \end{cases} \quad (4.19.t)$$

For a node location $\mathbf{x}_I$ at the boundary Γ , the normal $\hat{\mathbf{n}}$ at this point should be available, since we (of course) know about the geometry we are studying. The tangential vectors $\hat{\mathbf{t}}_1$ and $\hat{\mathbf{t}}_2$ in (4.19.t) are *any* two unit orthogonal vectors such that

$$\hat{\mathbf{t}}_1 \times \hat{\mathbf{t}}_2 = \hat{\mathbf{n}} \quad (4.19.u)$$

$$\hat{\mathbf{t}}_1 \cdot \hat{\mathbf{t}}_2 = 0. \quad (4.19.v)$$

After the set of elemental directions has been determined for all nodes, an element $\mathbf{e} \in \mathbb{E}^h(\Omega)$ is expanded as

$$\mathbf{e}(\mathbf{x}) = \sum_{I=1}^N \sum_{m=1}^{Q_I} h_{Im}^e(\mathbf{x}) (\hat{\alpha}_{Im} \hat{\mathbf{a}}_I + \hat{\beta}_{Im} \hat{\mathbf{b}}_I + \hat{\gamma}_{Im} \hat{\mathbf{c}}_I). \quad (4.19.w)$$

The coefficients (or DoF's) $\hat{\alpha}_{Im}$, in a sense, give the amplitude of the field $\mathbf{e}$ in the $\hat{\mathbf{a}}_I$ direction. The same goes on for $\hat{\beta}_{Im}$ and $\hat{\gamma}_{Im}$, which give the amplitude in the $\hat{\mathbf{b}}_I$ and $\hat{\mathbf{c}}_I$ directions, respectively.

Now it becomes easier to construct a space whose elements have zero tangential components. Since for each node I which happens to be in the boundary Γ the tangential directions are (locally) given by $\hat{\mathbf{b}}_I$ and $\hat{\mathbf{c}}_I$, it suffices to make the coefficients $\hat{\beta}_{Im} = 0$ and $\hat{\gamma}_{Im} = 0$ in (4.19.w). In this way, the resulting field $\mathbf{e}$ will have components only along the normal direction, (locally) given by the $\hat{\mathbf{a}}_I$. In this way, the space $\mathbb{V}_\tau^h(\Omega)$ is easily determined from $\mathbb{E}^h(\Omega)$.

In the interior of Ω , the elemental directions are the ordinary Cartesian directions. But since the interior patches do not intersect the global boundary (due to consideration 2 in Chart 4.2), they have no influence on the normal/tangential components of the resulting field.

If one desires to retrieve the x -component of the electric field $\mathbf{e}$ in (4.19.w), it suffices to take the dot product between $\mathbf{e}$ and $\hat{\mathbf{x}}$:

$$e_x(\mathbf{x}) = \mathbf{e}(\mathbf{x}) \cdot \hat{\mathbf{x}} = \sum_{I=1}^N \sum_{m=1}^{Q_I} h_{Im}^e(\mathbf{x}) (\hat{\alpha}_{Im} \hat{\mathbf{a}}_I \cdot \hat{\mathbf{x}} + \hat{\beta}_{Im} \hat{\mathbf{b}}_I \cdot \hat{\mathbf{x}} + \hat{\gamma}_{Im} \hat{\mathbf{c}}_I \cdot \hat{\mathbf{x}}). \quad (4.19.x)$$

The meshfree basis functions h_{Im}^e are obviously given by

$$h_{Im}^e(\mathbf{x}) = \varphi_I^0(\mathbf{x}) \ell_{I,m}^e(\mathbf{x}), \quad (4.19.y)$$

where the terms in the local basis $\ell_{I,m}^e$ come from (4.19.k). One observes that there are two representations for elements in $\mathbb{V}_\tau^h(\Omega)$: (4.19.a) and (4.19.w). The basis functions $\boldsymbol{\psi}_J$ in (4.19.a) are just the (double-indexed) h_{Im}^e . However, the ordering depends on how one decides to construct the numbering scheme (i.e., on how to put the DoF's in order, and consequently on how to attribute a row in the global matrix to each DoF). This topic will be discussed later.

In what regards the derivatives of the elements in $\mathbb{E}^h(\Omega)$ [and also in $\mathbb{V}_\tau^h(\Omega)$], we can apply the gradient operator to (4.19.w); with the help of the tensor product operator $\otimes$ we get:

$$\nabla \mathbf{e} = \sum_{I=1}^N \sum_{m=1}^{Q_I} \nabla h_{Im}^e \otimes (\hat{\alpha}_{Im} \hat{\mathbf{a}}_I + \hat{\beta}_{Im} \hat{\mathbf{b}}_I + \hat{\gamma}_{Im} \hat{\mathbf{c}}_I) \quad (4.20.a)$$

$$= \sum_{I=1}^N \sum_{m=1}^{Q_I} (\nabla h_{Im}^e \otimes \hat{\mathbf{a}}_I) \hat{\alpha}_{Im} + (\nabla h_{Im}^e \otimes \hat{\mathbf{b}}_I) \hat{\beta}_{Im} + (\nabla h_{Im}^e \otimes \hat{\mathbf{c}}_I) \hat{\gamma}_{Im}, \quad (4.20.b)$$

where the dependence of $\mathbf{e}$ and h_{Im}^e on the position $\mathbf{x} \in \Omega$ has been dropped, for the sake of clarity. (According to the rules of tensor algebra, the gradient of a vector is a tensor [Irgens, 2008]) The gradient ∇h_{Im}^e is calculated in the usual way:

$$\nabla h_{Im}^e = \frac{\partial h_{Im}^e}{\partial x} \hat{\mathbf{x}} + \frac{\partial h_{Im}^e}{\partial y} \hat{\mathbf{y}} + \frac{\partial h_{Im}^e}{\partial z} \hat{\mathbf{z}}. \quad (4.20.c)$$

4.3.4 Numbering schemes and the assembly process

In this work, the numbering scheme is organized in the following way. First, all local spaces V_I^e in (4.19.k) have the same dimension, i.e., we make

$$Q_I = N_e, \quad 1 \leq I \leq N. \quad (4.21.a)$$

Consequently, the global space V^e in (4.19.m) has dimension

$$\dim V^e = NN_e. \quad (4.21.b)$$

Second, all the local spaces V_I^p in (4.19.c) have the same dimension, i.e., we make

$$M_I = N_p, \quad 1 \leq I \leq N, \quad (4.21.c)$$

from which it follows that the global space V^p in (4.19.d) has dimension

$$\dim V^p = NN_p. \quad (4.21.d)$$

Third, the NN_e DoF's associated with the $\hat{\mathbf{a}}_I$'s, the NN_e DoF's associated with the $\hat{\mathbf{b}}_I$'s, the NN_e DoF's associated with the $\hat{\mathbf{c}}_I$'s, and the NN_p DoF's associated with the pseudopressure p are arranged in order. In this way, the α_{Im} in (4.19.w) gets mapped to the global index

$$I(\hat{\alpha}_{Im}) = (I - 1)N_e + m \quad (4.21.e)$$

in the global matrix. The β_{Im} in (4.19.w) gets mapped to the row

$$I(\hat{\beta}_{Im}) = NN_e + (I - 1)N_e + m. \quad (4.21.f)$$

In the same way, the γ_{Im} in (4.19.w) gets mapped to the row

$$I(\hat{\gamma}_{Im}) = 2NN_e + (I - 1)N_e + m. \quad (4.21.g)$$

Finally, the $\hat{p}_{Im}$ in (4.19.f) gets mapped to the row

$$I(\hat{p}_{Im}) = 3NN_e + (I - 1)N_p + m. \quad (4.21.h)$$

The total number of unknowns in the problem thus becomes:

$$3NN_e + NN_p = N(3N_e + N_p). \quad (4.21.i)$$

However, the DoF's corresponding to the tangential components of the fields must be zero. This is easily fixed by just identifying the boundary nodes, going to the global matrix and making the DoF's corresponding to the tangential components (which will be two of the elemental directions) equal to zero. In the end, the total number of DoF's is smaller than that in (4.21.i).

We will now take a closer look on the specific form assumed by the terms in (4.14.v) – (4.14.y). The scattering system, stated in (4.4.a) – (4.4.b), and whose solution we are interested in, is rewritten below for convenience:

Find $(\mathbf{e}_h^0, p_h) \in \mathbb{V}_\tau^h(\Omega) \times \mathbb{P}^h(\Omega)$ such that

$$\begin{aligned} \int_{\Omega} (\bar{\mathbf{A}} \cdot \nabla \mathbf{e}_h^0) : \nabla \mathbf{v}_h^* - \int_{\Omega} k_0^2 \mathbf{e}_h^0 \cdot \mathbf{v}_h^* - \int_{\Omega} p_h \nabla \cdot \mathbf{v}_h^* = \\ - \int_{\Omega} (\bar{\mathbf{A}} \cdot \nabla \mathbf{u}_h^g) : \nabla \mathbf{v}_h^* + \int_{\Omega} k_0^2 \mathbf{u}_h^g \cdot \mathbf{v}_h^*, \quad \forall \mathbf{v}_h \in \mathbb{V}_\tau^h(\Omega) \end{aligned} \quad (4.21.j)$$

$$- \int_{\Omega} q_h^* \nabla \cdot \mathbf{e}_h^0 = \int_{\Omega} q_h^* \nabla \cdot \mathbf{u}_h^g, \quad \forall q_h \in \mathbb{P}^h(\Omega), \quad (4.21.k)$$

Since the lifting $\mathbf{u}_h^g$ can be easily determined from the procedure outlined in Section 4.3.2, we concentrate on the unknowns $\mathbf{e}_h^0$ and p_h . They are expanded as

$$\mathbf{e}_h^0 = \sum_{J=1}^N \sum_{n=1}^{N_e} h_{Jn}^e (\hat{\alpha}_{Jn} \hat{\mathbf{a}}_J + \hat{\beta}_{Jn} \hat{\mathbf{b}}_J + \hat{\gamma}_{Jn} \hat{\mathbf{c}}_J). \quad (4.21.l)$$

$$p_h = \sum_{K=1}^N \sum_{r=1}^{N_p} h_{Kr}^p \hat{p}_{Kr} \quad (4.21.m)$$

The equations (4.21.j) and (4.21.k) hold true for *any* testing function in $\mathbb{V}_\tau^h(\Omega)$ and $\mathbb{P}^h(\Omega)$, respectively. These testing functions are expanded likewise as

$$\mathbf{v}_h = \sum_{I=1}^N \sum_{m=1}^{N_e} h_{Im}^e (\hat{\alpha}_{Im} \hat{\mathbf{a}}_I + \hat{\beta}_{Im} \hat{\mathbf{b}}_I + \hat{\gamma}_{Im} \hat{\mathbf{c}}_I) \quad (4.21.n)$$

$$q_h = \sum_{L=1}^N \sum_{s=1}^{N_p} h_{LS}^p \hat{p}_{LS} \quad (4.21.o)$$

The gradient vectors are provided by (4.20.b), i.e.,

$$\nabla \mathbf{e}_h^0 = \sum_{J=1}^N \sum_{n=1}^{N_e} (\nabla h_{Jn}^e \otimes \hat{\mathbf{a}}_J) \hat{\alpha}_{Jn} + (\nabla h_{Jn}^e \otimes \hat{\mathbf{b}}_J) \hat{\beta}_{Jn} + (\nabla h_{Jn}^e \otimes \hat{\mathbf{c}}_J) \hat{\gamma}_{Jn} \quad (4.21.p)$$

$$\nabla \mathbf{v}_h = \sum_{I=1}^N \sum_{m=1}^{N_e} (\nabla h_{Im}^e \otimes \hat{\mathbf{a}}_I) \hat{\alpha}_{Im} + (\nabla h_{Im}^e \otimes \hat{\mathbf{b}}_I) \hat{\beta}_{Im} + (\nabla h_{Im}^e \otimes \hat{\mathbf{c}}_I) \hat{\gamma}_{Im} \quad (4.21.q)$$

When we substitute (4.21.l) and (4.21.m) in (4.21.j) and (4.21.k), we get:

$$\begin{aligned} & \sum_{J=1}^N \sum_{n=1}^{N_e} \left(\int_{\Omega} \left(\bar{\mathbf{\Lambda}} \cdot ((\nabla h_{Jn}^e \otimes \hat{\mathbf{a}}_J) \hat{\alpha}_{Jn} + (\nabla h_{Jn}^e \otimes \hat{\mathbf{b}}_J) \hat{\beta}_{Jn} + (\nabla h_{Jn}^e \otimes \hat{\mathbf{c}}_J) \hat{\gamma}_{Jn}) \right) : \nabla \mathbf{v}_h^* \right) \\ & - \sum_{J=1}^N \sum_{n=1}^{N_e} \left(\int_{\Omega} k_0^2 (h_{Jn}^e (\hat{\alpha}_{Jn} \hat{\mathbf{a}}_J + \hat{\beta}_{Jn} \hat{\mathbf{b}}_J + \hat{\gamma}_{Jn} \hat{\mathbf{c}}_J)) \cdot \mathbf{v}_h^* \right) = \\ & - \sum_{K=1}^N \sum_{r=1}^{N_p} \left(\int_{\Omega} h_{Kr}^p \nabla \cdot \mathbf{v}_h^* \right) \hat{p}_{Kr} = \\ & - \int_{\Omega} (\bar{\mathbf{\Lambda}} \cdot \nabla \mathbf{u}_h^g) : \nabla \mathbf{v}_h^* + \int_{\Omega} k_0^2 \mathbf{u}_h^g \cdot \mathbf{v}_h^*, \quad \forall \mathbf{v}_h \in \mathbb{V}_{\tau}^h(\Omega) \end{aligned} \quad (4.21.q)$$

$$\begin{aligned} & - \sum_{J=1}^N \sum_{n=1}^{N_e} \left(\int_{\Omega} q_h^* \nabla \cdot (h_{Jn}^e (\hat{\alpha}_{Jn} \hat{\mathbf{a}}_J + \hat{\beta}_{Jn} \hat{\mathbf{b}}_J + \hat{\gamma}_{Jn} \hat{\mathbf{c}}_J)) \right) = \\ & \int_{\Omega} q_h^* \nabla \cdot \mathbf{u}_h^g, \quad \forall q_h \in \mathbb{P}^h(\Omega) \end{aligned} \quad (4.21.r)$$

The testing function $\mathbf{v}_h$ in (4.21.q) is arbitrary. From the representation for $\mathbf{v}_h$ in (4.21.n), the components relative to the elemental directions $\hat{\mathbf{a}}_I$, $\hat{\mathbf{b}}_I$, and $\hat{\mathbf{c}}_I$ are linearly independent. It means that (4.21.q) can be broken down into three expressions, each of them having the same form. In the first, the $\hat{\alpha}_{Im}$ are arbitrary, whereas $\hat{\beta}_{Im} = \hat{\gamma}_{Im} = 0$. In the second, the $\hat{\beta}_{Im}$ are arbitrary, whereas $\hat{\alpha}_{Im} = \hat{\gamma}_{Im} = 0$. And in the third, the $\hat{\gamma}_{Im}$ are arbitrary, whereas $\hat{\alpha}_{Im} = \hat{\beta}_{Im} = 0$. In each case, these scalar coefficients will appear at both sides of the equation, so in the end their effect will be immaterial. These cases will now be examined carefully. After we are done, the equation (4.21.r) shall also be examined.

Case 1: $\hat{\alpha}_{Im} \in \mathbb{C}$, $\hat{\beta}_{Im} = 0$, $\hat{\gamma}_{Im} = 0$

Let us concentrate on a single testing function defined by its double index I and m . The $\mathbf{v}_h$ in (4.21.n) and its derivative in (4.21.q) become

$$\mathbf{v}_h = (h_{Im}^e \hat{\mathbf{a}}_I) \hat{\alpha}_{Im} \quad (4.22.a)$$

$$\nabla \mathbf{v}_h = (\nabla h_{Im}^e \otimes \hat{\mathbf{a}}_I) \hat{\alpha}_{Im} \quad (4.22.b)$$

When we substitute (4.22.a) and (4.22.b) in (4.21.q), we expect to arrive at an equation like

$$\sum_{j=1}^N \sum_{n=1}^{N_e} \mathcal{A}_{ImJn}^a \hat{\alpha}_{Jn} + \mathcal{A}_{ImJn}^b \hat{\beta}_{Jn} + \mathcal{A}_{ImJn}^c \hat{\gamma}_{Jn} + \sum_{K=1}^N \sum_{r=1}^{N_p} \mathcal{A}_{ImKr}^p \hat{p}_{Kr} = \mathcal{F}_{Im}^a \quad (4.22.c)$$

The terms $\mathcal{A}_{ImJn}^a$, $\mathcal{A}_{ImJn}^b$, $\mathcal{A}_{ImJn}^c$, and $\mathcal{A}_{ImKr}^p$ will be mapped to the global matrix according to the Table 4.3 below.

TABLE 4.3 – MAPPING TO THE GLOBAL MATRIX

<i>Term</i>	<i>Row</i>	<i>Column</i>
$\mathcal{A}_{ImJn}^a$	$I(\hat{\alpha}_{Im})$	$I(\hat{\alpha}_{Jn})$
$\mathcal{A}_{ImJn}^b$	$I(\hat{\alpha}_{Im})$	$I(\hat{\beta}_{Jn})$
$\mathcal{A}_{ImJn}^c$	$I(\hat{\alpha}_{Im})$	$I(\hat{\gamma}_{Jn})$
$\mathcal{A}_{ImKr}^p$	$I(\hat{\alpha}_{Im})$	$I(\hat{p}_{Im})$

The index functions are given by (4.21.e) – (4.21.h). The term $\mathcal{F}_{Im}^a$ will be mapped to the position $I(\hat{\alpha}_{Im})$ in the right-side global vector.

We know from (2.113) that the PML tensor has the form

$$\bar{\bar{\mathbf{\Lambda}}} = \Lambda_x \hat{\mathbf{x}} \otimes \hat{\mathbf{x}} + \Lambda_y \hat{\mathbf{y}} \otimes \hat{\mathbf{y}} + \Lambda_z \hat{\mathbf{z}} \otimes \hat{\mathbf{z}}. \quad (4.22.d)$$

In this way, we can apply the definition of scalar product between two tensors in (1.61) and find out that the very first term in the first integral in (4.21.q) can be worked out as

$$\begin{aligned} \bar{\bar{\mathbf{\Lambda}}} \cdot (\nabla h_{Jn}^e \otimes \hat{\mathbf{a}}_J) &= \\ \bar{\bar{\mathbf{\Lambda}}} \cdot \nabla h_{Jn}^e \otimes \hat{\mathbf{a}}_J &= \\ (\Lambda_x \hat{\mathbf{x}} \otimes \hat{\mathbf{x}} + \Lambda_y \hat{\mathbf{y}} \otimes \hat{\mathbf{y}} + \Lambda_z \hat{\mathbf{z}} \otimes \hat{\mathbf{z}}) \cdot \nabla h_{Jn}^e \otimes \hat{\mathbf{a}}_J &= \\ (\Lambda_x \hat{\mathbf{x}} (\hat{\mathbf{x}} \cdot \nabla h_{Jn}^e) + \Lambda_y \hat{\mathbf{y}} (\hat{\mathbf{y}} \cdot \nabla h_{Jn}^e) + \Lambda_z \hat{\mathbf{z}} (\hat{\mathbf{z}} \cdot \nabla h_{Jn}^e)) \otimes \hat{\mathbf{a}}_J &= \end{aligned}$$

$$\Lambda_x \frac{\partial h_{jn}^e}{\partial x} \hat{\mathbf{x}} \otimes \hat{\mathbf{a}}_J + \Lambda_y \frac{\partial h_{jn}^e}{\partial y} \hat{\mathbf{y}} \otimes \hat{\mathbf{a}}_J + \Lambda_z \frac{\partial h_{jn}^e}{\partial z} \hat{\mathbf{z}} \otimes \hat{\mathbf{a}}_J \quad (4.22.e)$$

Moreover, from (4.22.e) and (4.22.b), one finds out with the help of the double dot product defined in (1.65) that

$$\begin{aligned} & \left(\bar{\Lambda} \cdot (\nabla h_{jn}^e \otimes \hat{\mathbf{a}}_J) \right) : \nabla \mathbf{v}_h^* = \\ & \left(\Lambda_x \frac{\partial h_{jn}^e}{\partial x} \hat{\mathbf{x}} \otimes \hat{\mathbf{a}}_J + \Lambda_y \frac{\partial h_{jn}^e}{\partial y} \hat{\mathbf{y}} \otimes \hat{\mathbf{a}}_J + \Lambda_z \frac{\partial h_{jn}^e}{\partial z} \hat{\mathbf{z}} \otimes \hat{\mathbf{a}}_J \right) : ((\nabla h_{lm}^e \otimes \hat{\mathbf{a}}_I) \hat{\alpha}_{lm}^*) = \\ & \left(\Lambda_x \frac{\partial h_{jn}^e}{\partial x} (\hat{\mathbf{x}} \cdot \nabla h_{lm}^e) (\hat{\mathbf{a}}_J \cdot \hat{\mathbf{a}}_I) + \Lambda_y \frac{\partial h_{jn}^e}{\partial y} (\hat{\mathbf{y}} \cdot \nabla h_{lm}^e) (\hat{\mathbf{a}}_J \cdot \hat{\mathbf{a}}_I) + \right. \\ & \quad \left. \Lambda_z \frac{\partial h_{jn}^e}{\partial z} (\hat{\mathbf{z}} \cdot \nabla h_{lm}^e) (\hat{\mathbf{a}}_J \cdot \hat{\mathbf{a}}_I) \right) \hat{\alpha}_{lm}^* = \\ & \left(\Lambda_x \frac{\partial h_{jn}^e}{\partial x} \frac{\partial h_{lm}^e}{\partial x} + \Lambda_y \frac{\partial h_{jn}^e}{\partial y} \frac{\partial h_{lm}^e}{\partial y} + \Lambda_z \frac{\partial h_{jn}^e}{\partial z} \frac{\partial h_{lm}^e}{\partial z} \right) (\hat{\mathbf{a}}_J \cdot \hat{\mathbf{a}}_I) \hat{\alpha}_{lm}^* \end{aligned} \quad (4.22.f)$$

After we substitute (4.22.f) back into (4.21.q), and apply the same reasoning to the other integrals, we discover that the first three terms in Table 4.3 are given by:

$$\begin{aligned} \mathcal{A}_{lmjn}^a &= \int_{\Omega} \left(\left(\Lambda_x \frac{\partial h_{jn}^e}{\partial x} \frac{\partial h_{lm}^e}{\partial x} + \Lambda_y \frac{\partial h_{jn}^e}{\partial y} \frac{\partial h_{lm}^e}{\partial y} + \Lambda_z \frac{\partial h_{jn}^e}{\partial z} \frac{\partial h_{lm}^e}{\partial z} \right) - k_0^2 h_{jn}^e h_{lm}^e \right) \hat{\mathbf{a}}_J \cdot \hat{\mathbf{a}}_I \\ \mathcal{A}_{lmjn}^b &= \int_{\Omega} \left(\left(\Lambda_x \frac{\partial h_{jn}^e}{\partial x} \frac{\partial h_{lm}^e}{\partial x} + \Lambda_y \frac{\partial h_{jn}^e}{\partial y} \frac{\partial h_{lm}^e}{\partial y} + \Lambda_z \frac{\partial h_{jn}^e}{\partial z} \frac{\partial h_{lm}^e}{\partial z} \right) - k_0^2 h_{jn}^e h_{lm}^e \right) \hat{\mathbf{b}}_J \cdot \hat{\mathbf{a}}_I \\ \mathcal{A}_{lmjn}^c &= \int_{\Omega} \left(\left(\Lambda_x \frac{\partial h_{jn}^e}{\partial x} \frac{\partial h_{lm}^e}{\partial x} + \Lambda_y \frac{\partial h_{jn}^e}{\partial y} \frac{\partial h_{lm}^e}{\partial y} + \Lambda_z \frac{\partial h_{jn}^e}{\partial z} \frac{\partial h_{lm}^e}{\partial z} \right) - k_0^2 h_{jn}^e h_{lm}^e \right) \hat{\mathbf{c}}_J \cdot \hat{\mathbf{a}}_I \end{aligned} \quad (4.22.g)$$

The fourth term can be found with the help of the identity

$$\nabla \cdot \mathbf{w} = \nabla \mathbf{w} : \bar{\mathbf{I}}, \quad (4.22.h)$$

where $\bar{\mathbf{I}}$ is the identity tensor given by

$$\bar{\mathbf{I}} = \hat{\mathbf{x}} \otimes \hat{\mathbf{x}} + \hat{\mathbf{y}} \otimes \hat{\mathbf{y}} + \hat{\mathbf{z}} \otimes \hat{\mathbf{z}} \quad (4.22.i)$$

and $\mathbf{w}$ is an arbitrary vector. In this way, from (4.22.b), (4.22.h) and (4.22.i), the pertinent terms in (4.21.q) can be developed as:

$$\int_{\Omega} h_{Kr}^p \nabla \cdot \mathbf{v}_h^* = \quad (4.22.j)$$

$$\int_{\Omega} h_{Kr}^p \nabla \mathbf{v}_h^* : \bar{\mathbf{I}} =$$

$$\begin{aligned} & \int_{\Omega} h_{Kr}^p ((\nabla h_{Im}^e \otimes \hat{\mathbf{a}}_l) \hat{\mathbf{a}}_{Im})^* : (\hat{\mathbf{x}} \otimes \hat{\mathbf{x}} + \hat{\mathbf{y}} \otimes \hat{\mathbf{y}} + \hat{\mathbf{z}} \otimes \hat{\mathbf{z}}) = \\ & \int_{\Omega} h_{Kr}^p \left(\frac{\partial h_{Im}^e}{\partial x} (\hat{\mathbf{a}}_l \cdot \hat{\mathbf{x}}) + \frac{\partial h_{Im}^e}{\partial y} (\hat{\mathbf{a}}_l \cdot \hat{\mathbf{y}}) + \frac{\partial h_{Im}^e}{\partial z} (\hat{\mathbf{a}}_l \cdot \hat{\mathbf{z}}) \right) \hat{\mathbf{a}}_{Im}^*, \end{aligned}$$

which allows us to conclude that

$$\mathcal{A}_{ImKr}^p = - \int_{\Omega} h_{Kr}^p \left(\frac{\partial h_{Im}^e}{\partial x} (\hat{\mathbf{a}}_l \cdot \hat{\mathbf{x}}) + \frac{\partial h_{Im}^e}{\partial y} (\hat{\mathbf{a}}_l \cdot \hat{\mathbf{y}}) + \frac{\partial h_{Im}^e}{\partial z} (\hat{\mathbf{a}}_l \cdot \hat{\mathbf{z}}) \right) \quad (4.22.k)$$

The term $\mathcal{F}_{Im}^a$ in (4.22.c) is found with the help of (4.22.a) and (4.22.b), which are substituted in the right side of (4.21.q). Before we state its final form, we need to take a look at how the lifting function $\mathbf{u}_h^g$ is found. According to (4.1.e), the lifting $\mathbf{u}_h^g$ must be such that

$$\boldsymbol{\gamma}_t \mathbf{u}_h^g \cong \begin{cases} \mathbf{0}, & \text{at } \Gamma_o \\ -\hat{\mathbf{n}}_1 \times \mathbf{E}^{inc}, & \text{at } \Gamma_1. \end{cases} \quad (4.22.l)$$

Since we took that $\mathbf{u}_h^g \in \mathbb{E}^h(\Omega)$, it admits an expansion like (4.19.w):

$$\mathbf{u}_h^g = \sum_{P=1}^N \sum_{t=1}^{N_e} h_{Pt}^e (\hat{\alpha}_{Pt} \hat{\mathbf{a}}_P + \hat{\beta}_{Pt} \hat{\mathbf{b}}_P + \hat{\gamma}_{Pt} \hat{\mathbf{c}}_P) \quad (4.22.m)$$

$$\nabla \mathbf{u}_h^g = \sum_{P=1}^N \sum_{t=1}^{N_e} (\nabla h_{Pt}^e \otimes \hat{\mathbf{a}}_P) \hat{\alpha}_{Pt} + (\nabla h_{Pt}^e \otimes \hat{\mathbf{b}}_P) \hat{\beta}_{Pt} + (\nabla h_{Pt}^e \otimes \hat{\mathbf{c}}_P) \hat{\gamma}_{Pt} \quad (4.22.n)$$

The coefficients $\hat{\alpha}_{Pt}$, $\hat{\beta}_{Pt}$ and $\hat{\gamma}_{Pt}$ can be easily determined, thanks to the procedure outlined in Chart 4.3 extended to vector functions. It will be explained in detail below.

1. We consider all N nodes from the problem, i.e., $1 \leq P \leq N$.

2. If P is an interior node, i.e., if $\mathbf{x}_P \notin \Gamma$, then $\hat{\alpha}_{Pt} = \hat{\beta}_{Pt} = \hat{\gamma}_{Pt} = 0$.

3. Because the role of $\mathbf{u}_h^g$ is to essentially capture the behavior of the tangential components of the scattered field, we can take its normal component to be zero. The normal component of the scattered field will be captured by the $\mathbf{e}_h^0$ in (4.2). Moreover, according to the definition of elemental directions in (4.19.t), the components in the normal direction are controlled by the $\hat{\alpha}_{Pt}$. So we make $\hat{\alpha}_{Pt} = 0$, $1 \leq P \leq N$.

4. If $\mathbf{x}_p \in \Gamma_1$, the tangential directions at the node location will be given precisely by $\hat{\mathbf{b}}_p$ and $\hat{\mathbf{c}}_p$. Then we can make

$$\hat{\beta}_{p1} = (-\hat{\mathbf{n}}_1 \times \mathbf{E}^{inc}) \cdot \hat{\mathbf{b}}_p \quad (4.22.o)$$

$$\hat{\gamma}_{p1} = (-\hat{\mathbf{n}}_1 \times \mathbf{E}^{inc}) \cdot \hat{\mathbf{c}}_p \quad (4.22.p)$$

(According to the consideration 3 in Chart 4.2, nodes at the Dirichlet boundaries have a single DoF.)

5. If $\mathbf{x}_p \in \Gamma_o$, the tangential directions at the node location will be given precisely by $\hat{\mathbf{b}}_p$ and $\hat{\mathbf{c}}_p$. Then we can make

$$\hat{\beta}_{p1} = 0 \quad (4.22.q)$$

$$\hat{\gamma}_{p1} = 0 \quad (4.22.r)$$

As evidenced by the five steps above, the only DoF's able to 'excite' the problem are those associated with the tangential directions along the scatterer surface Γ_1 . Now that we know all the coefficients in the expansion (4.22.m), $\mathbf{u}_h^g$ can be easily determined.

When we substitute (4.22.m), (4.22.n), (4.22.a) and (4.22.b) in the right side of (4.21.q), we find that

$$\mathcal{F}_{lm}^a = - \int_{\Omega} (\bar{\mathbf{A}} \cdot \nabla \mathbf{u}_h^g) : (\nabla h_{lm}^e \otimes \hat{\mathbf{a}}_l) + \int_{\Omega} k_0^2 \mathbf{u}_h^g \cdot (h_{lm}^e \hat{\mathbf{a}}_l). \quad (4.22.s)$$

We believe that the explanation is sufficiently clear, so we will not work out the double dot products between tensors. The steps are very similar to those in (4.22.e) – (4.22.f).

$$\text{Case 2: } \hat{\alpha}_{lm} = 0, \hat{\beta}_{lm} \in \mathbb{C}, \hat{\gamma}_{lm} = 0$$

The $\mathbf{v}_h$ in (4.21.n) and its derivative in (4.21.q) now become

$$\mathbf{v}_h = (h_{lm}^e \hat{\mathbf{b}}_l) \hat{\beta}_{lm} \quad (4.23.a)$$

$$\nabla \mathbf{v}_h = (\nabla h_{lm}^e \otimes \hat{\mathbf{b}}_l) \hat{\beta}_{lm} \quad (4.23.b)$$

When we substitute (4.23.a) and (4.23.b) in (4.21.q), we expect to arrive at an equation like

$$\sum_{j=1}^N \sum_{n=1}^{N_e} \mathcal{B}_{lmjn}^a \hat{\alpha}_{jn} + \mathcal{B}_{lmjn}^b \hat{\beta}_{jn} + \mathcal{B}_{lmjn}^c \hat{\gamma}_{jn} + \sum_{K=1}^N \sum_{r=1}^{N_p} \mathcal{B}_{lmKr}^p \hat{p}_{Kr} = \mathcal{F}_{lm}^b \quad (4.23.c)$$

The terms $\mathcal{B}_{lmjn}^a$, $\mathcal{B}_{lmjn}^b$, $\mathcal{B}_{lmjn}^c$, and $\mathcal{B}_{lmKr}^p$ will be mapped to the global matrix according to the Table 4.4 below.

TABLE 4.4 – MAPPING TO THE GLOBAL MATRIX

<i>Term</i>	<i>Row</i>	<i>Column</i>
$\mathcal{B}_{ImJn}^a$	$I(\hat{\beta}_{Im})$	$I(\hat{\alpha}_{Jn})$
$\mathcal{B}_{ImJn}^b$	$I(\hat{\beta}_{Im})$	$I(\hat{\beta}_{Jn})$
$\mathcal{B}_{ImJn}^c$	$I(\hat{\beta}_{Im})$	$I(\hat{\gamma}_{Jn})$
$\mathcal{B}_{ImKr}^p$	$I(\hat{\beta}_{Im})$	$I(\hat{p}_{Im})$

The index functions are given by (4.21.e) – (4.21.h). The term $\mathcal{F}_{Im}^b$ will be mapped to the position $I(\hat{\beta}_{Im})$ in the right-side global vector. These terms are given by

(4.23. d)

$$\mathcal{B}_{ImJn}^a = \int_{\Omega} \left(\left(\Lambda_x \frac{\partial h_{Jn}^e}{\partial x} \frac{\partial h_{Im}^e}{\partial x} + \Lambda_y \frac{\partial h_{Jn}^e}{\partial y} \frac{\partial h_{Im}^e}{\partial y} + \Lambda_z \frac{\partial h_{Jn}^e}{\partial z} \frac{\partial h_{Im}^e}{\partial z} \right) - k_0^2 h_{Jn}^e h_{Im}^e \right) \hat{\mathbf{a}}_J \cdot \hat{\mathbf{b}}_I$$

$$\mathcal{B}_{ImJn}^b = \int_{\Omega} \left(\left(\Lambda_x \frac{\partial h_{Jn}^e}{\partial x} \frac{\partial h_{Im}^e}{\partial x} + \Lambda_y \frac{\partial h_{Jn}^e}{\partial y} \frac{\partial h_{Im}^e}{\partial y} + \Lambda_z \frac{\partial h_{Jn}^e}{\partial z} \frac{\partial h_{Im}^e}{\partial z} \right) - k_0^2 h_{Jn}^e h_{Im}^e \right) \hat{\mathbf{b}}_J \cdot \hat{\mathbf{b}}_I$$

$$\mathcal{B}_{ImJn}^c = \int_{\Omega} \left(\left(\Lambda_x \frac{\partial h_{Jn}^e}{\partial x} \frac{\partial h_{Im}^e}{\partial x} + \Lambda_y \frac{\partial h_{Jn}^e}{\partial y} \frac{\partial h_{Im}^e}{\partial y} + \Lambda_z \frac{\partial h_{Jn}^e}{\partial z} \frac{\partial h_{Im}^e}{\partial z} \right) - k_0^2 h_{Jn}^e h_{Im}^e \right) \hat{\mathbf{c}}_J \cdot \hat{\mathbf{b}}_I$$

$$\mathcal{B}_{ImKr}^p = - \int_{\Omega} h_{Kr}^p \left(\frac{\partial h_{Im}^e}{\partial x} (\hat{\mathbf{b}}_I \cdot \hat{\mathbf{x}}) + \frac{\partial h_{Im}^e}{\partial y} (\hat{\mathbf{b}}_I \cdot \hat{\mathbf{y}}) + \frac{\partial h_{Im}^e}{\partial z} (\hat{\mathbf{b}}_I \cdot \hat{\mathbf{z}}) \right)$$

$$\mathcal{F}_{Im}^b = - \int_{\Omega} (\bar{\mathbf{\Lambda}} \cdot \nabla \mathbf{u}_h^g) : (\nabla h_{Im}^e \otimes \hat{\mathbf{b}}_I) + \int_{\Omega} k_0^2 \mathbf{u}_h^g \cdot (h_{Im}^e \hat{\mathbf{b}}_I).$$

Case 3: $\hat{\alpha}_{Im} = 0$, $\hat{\beta}_{Im} = 0$, $\hat{\gamma}_{Im} \in \mathbb{C}$

The same all over again. The $\mathbf{v}_h$ in (4.21.n) and its derivative in (4.21.q) now become

$$\mathbf{v}_h = (h_{Im}^e \hat{\mathbf{c}}_I) \hat{\gamma}_{Im} \quad (4.24. a)$$

$$\nabla \mathbf{v}_h = (\nabla h_{Im}^e \otimes \hat{\mathbf{c}}_I) \hat{\gamma}_{Im} \quad (4.24. b)$$

When we substitute (4.24.a) and (4.24.b) in (4.21.q), we expect to arrive at an equation like

$$\sum_{J=1}^N \sum_{n=1}^{N_e} \mathcal{C}_{ImJn}^a \hat{\alpha}_{Jn} + \mathcal{C}_{ImJn}^b \hat{\beta}_{Jn} + \mathcal{C}_{ImJn}^c \hat{\gamma}_{Jn} + \sum_{K=1}^N \sum_{r=1}^{N_p} \mathcal{C}_{ImKr}^p \hat{p}_{Kr} = \mathcal{F}_{Im}^c \quad (4.24. c)$$

The terms $\mathcal{C}_{ImJn}^a$, $\mathcal{C}_{ImJn}^b$, $\mathcal{C}_{ImJn}^c$, and $\mathcal{C}_{ImKr}^p$ will be mapped to the global matrix according to the Table 4.5 below.

TABLE 4.5 – MAPPING TO THE GLOBAL MATRIX

Term	Row	Column
$\mathcal{C}_{ImJn}^a$	$I(\hat{\gamma}_{Im})$	$I(\hat{\alpha}_{Jn})$
$\mathcal{C}_{ImJn}^b$	$I(\hat{\gamma}_{Im})$	$I(\hat{\beta}_{Jn})$
$\mathcal{C}_{ImJn}^c$	$I(\hat{\gamma}_{Im})$	$I(\hat{\gamma}_{Jn})$
$\mathcal{C}_{ImKr}^p$	$I(\hat{\gamma}_{Im})$	$I(\hat{p}_{Im})$

The index functions are given by (4.21.e) – (4.21.h). The term $\mathcal{F}_{Im}^c$ will be mapped to the position $I(\hat{\gamma}_{Im})$ in the right-side global vector. These terms are given by

(4.24. d)

$$\mathcal{C}_{ImJn}^a = \int_{\Omega} \left(\left(\Lambda_x \frac{\partial h_{Jn}^e}{\partial x} \frac{\partial h_{Im}^e}{\partial x} + \Lambda_y \frac{\partial h_{Jn}^e}{\partial y} \frac{\partial h_{Im}^e}{\partial y} + \Lambda_z \frac{\partial h_{Jn}^e}{\partial z} \frac{\partial h_{Im}^e}{\partial z} \right) - k_0^2 h_{Jn}^e h_{Im}^e \right) \hat{\mathbf{a}}_J \cdot \hat{\mathbf{c}}_I$$

$$\mathcal{C}_{ImJn}^b = \int_{\Omega} \left(\left(\Lambda_x \frac{\partial h_{Jn}^e}{\partial x} \frac{\partial h_{Im}^e}{\partial x} + \Lambda_y \frac{\partial h_{Jn}^e}{\partial y} \frac{\partial h_{Im}^e}{\partial y} + \Lambda_z \frac{\partial h_{Jn}^e}{\partial z} \frac{\partial h_{Im}^e}{\partial z} \right) - k_0^2 h_{Jn}^e h_{Im}^e \right) \hat{\mathbf{b}}_J \cdot \hat{\mathbf{c}}_I$$

$$\mathcal{C}_{ImJn}^c = \int_{\Omega} \left(\left(\Lambda_x \frac{\partial h_{Jn}^e}{\partial x} \frac{\partial h_{Im}^e}{\partial x} + \Lambda_y \frac{\partial h_{Jn}^e}{\partial y} \frac{\partial h_{Im}^e}{\partial y} + \Lambda_z \frac{\partial h_{Jn}^e}{\partial z} \frac{\partial h_{Im}^e}{\partial z} \right) - k_0^2 h_{Jn}^e h_{Im}^e \right) \hat{\mathbf{c}}_J \cdot \hat{\mathbf{c}}_I$$

$$\mathcal{C}_{ImKr}^p = - \int_{\Omega} h_{Kr}^p \left(\frac{\partial h_{Im}^e}{\partial x} (\hat{\mathbf{c}}_I \cdot \hat{\mathbf{x}}) + \frac{\partial h_{Im}^e}{\partial y} (\hat{\mathbf{c}}_I \cdot \hat{\mathbf{y}}) + \frac{\partial h_{Im}^e}{\partial z} (\hat{\mathbf{c}}_I \cdot \hat{\mathbf{z}}) \right)$$

$$\mathcal{F}_{Im}^c = - \int_{\Omega} (\bar{\mathbf{\Lambda}} \cdot \nabla \mathbf{u}_h^g) : (\nabla h_{Im}^e \otimes \hat{\mathbf{c}}_I) + \int_{\Omega} k_0^2 \mathbf{u}_h^g \cdot (h_{Im}^e \hat{\mathbf{c}}_I).$$

We need now to take care of (4.21.r), repeated below for convenience:

$$- \sum_{j=1}^N \sum_{n=1}^{N_e} \left(\int_{\Omega} q_h^* \nabla \cdot (h_{Jn}^e (\hat{\alpha}_{Jn} \hat{\mathbf{a}}_J + \hat{\beta}_{Jn} \hat{\mathbf{b}}_J + \hat{\gamma}_{Jn} \hat{\mathbf{c}}_J)) \right) =$$

$$\int_{\Omega} q_h^* \nabla \cdot \mathbf{u}_h^g, \quad \forall q_h \in \mathbb{P}^h(\Omega) \quad (4.25. a)$$

Let us concentrate on a single testing function defined by its double index L and s in (4.21.o), i.e., we make

$$q_h = h_{LS}^p \hat{p}_{LS}, \quad (4.25.b)$$

where $\hat{p}_{LS} \in \mathbb{C}$ is arbitrary. After the substitution of (4.25.b) into (4.25.a), we hope to arrive at an equation like

$$\sum_{J=1}^N \sum_{n=1}^{N_e} \mathcal{P}_{LSJn}^a \hat{\alpha}_{Jn} + \mathcal{P}_{LSJn}^b \hat{\beta}_{Jn} + \mathcal{P}_{LSJn}^c \hat{\gamma}_{Jn} = \mathcal{G}_{LS}^p \quad (4.25.c)$$

The terms $\mathcal{C}_{lmJn}^a$, $\mathcal{C}_{lmJn}^b$, $\mathcal{C}_{lmJn}^c$, and $\mathcal{C}_{lmKr}^p$ will be mapped to the global matrix according to the Table 4.6 below.

TABLE 4.6 – MAPPING TO THE GLOBAL MATRIX

<i>Term</i>	<i>Row</i>	<i>Column</i>
$\mathcal{P}_{LSJn}^a$	$I(\hat{p}_{LS})$	$I(\hat{\alpha}_{Jn})$
$\mathcal{P}_{LSJn}^b$	$I(\hat{p}_{LS})$	$I(\hat{\beta}_{Jn})$
$\mathcal{P}_{LSJn}^c$	$I(\hat{p}_{LS})$	$I(\hat{\gamma}_{Jn})$

The index functions are given by (4.21.e) – (4.21.h). The term $\mathcal{G}_{LS}^p$ will be mapped to the position $I(\hat{p}_{LS})$ in the right-side global vector. With the help of (4.22.h) and (4.22.i), and the rules of tensor algebra we have been employing thus far, we arrive at the specific forms for these terms:

$$\mathcal{P}_{LSJn}^a = - \int_{\Omega} h_{LS}^p \left(\frac{\partial h_{Jn}^e}{\partial x} (\hat{\mathbf{a}}_J \cdot \hat{\mathbf{x}}) + \frac{\partial h_{Jn}^e}{\partial y} (\hat{\mathbf{a}}_J \cdot \hat{\mathbf{y}}) + \frac{\partial h_{Jn}^e}{\partial z} (\hat{\mathbf{a}}_J \cdot \hat{\mathbf{z}}) \right) \quad (4.25.d)$$

$$\mathcal{P}_{LSJn}^b = - \int_{\Omega} h_{LS}^p \left(\frac{\partial h_{Jn}^e}{\partial x} (\hat{\mathbf{b}}_J \cdot \hat{\mathbf{x}}) + \frac{\partial h_{Jn}^e}{\partial y} (\hat{\mathbf{b}}_J \cdot \hat{\mathbf{y}}) + \frac{\partial h_{Jn}^e}{\partial z} (\hat{\mathbf{b}}_J \cdot \hat{\mathbf{z}}) \right)$$

$$\mathcal{P}_{LSJn}^c = - \int_{\Omega} h_{LS}^p \left(\frac{\partial h_{Jn}^e}{\partial x} (\hat{\mathbf{c}}_J \cdot \hat{\mathbf{x}}) + \frac{\partial h_{Jn}^e}{\partial y} (\hat{\mathbf{c}}_J \cdot \hat{\mathbf{y}}) + \frac{\partial h_{Jn}^e}{\partial z} (\hat{\mathbf{c}}_J \cdot \hat{\mathbf{z}}) \right)$$

$$\mathcal{G}_{LS}^p = \int_{\Omega} h_{LS}^p \nabla \cdot \mathbf{u}_h^g$$

4.3.5 Final comments

In what regards the assembly process, the work is essentially done. From (4.22.g), (4.22.k), (4.22.s), (4.23.d), (4.24.d) and (4.25.d), we can construct our linear system (4.14.z). In the next chapter, we will be concerned with some features in the solution of this linear system, and also with the application of our meshfree method to problems arising in electromagnetic wave scattering.

Chapter 5

Experimental studies

The objective of this chapter is to assess some features concerning the numerical implementation of the method described in the last chapter.

In the first section, we shall take a look at the numerical integration of the terms which will ultimately figure as the entries in the global matrix. Since the numerical integration is a delicate issue in the meshfree methods, we present a recipe to ‘alleviate’ its cost.

The second section deals with the inf-sup condition and the problem of identifying compatible pairs of spaces.

The third section is very brief, and discusses the preconditioning techniques we employed to solve the global linear system.

Finally, the fourth section brings lots of examples of our meshfree method in the solution of wave scattering problems. We show that it works pretty well in two and three-dimensional cases.

5.1 Numerical integration

5.1.1 Basic integrals

After we get the final form of the entries in the matrix and in the vector which will form the global linear system in (4.22.g), (4.22.k), (4.22.s), (4.23.d), (4.24.d) and (4.25.d), we can begin to make assumptions in order to simplify the process of actually computing them.

The most patent of these assumptions regards the components of the PML tensor. In terms like $\mathcal{A}_{lmjn}^a$ in (4.22.g), whenever we get integrands involving the PML tensor, as

$$\int_{\Omega} \left(\left(\Lambda_x \frac{\partial h_{jn}^e}{\partial x} \frac{\partial h_{lm}^e}{\partial x} + \Lambda_y \frac{\partial h_{jn}^e}{\partial y} \frac{\partial h_{lm}^e}{\partial y} + \Lambda_z \frac{\partial h_{jn}^e}{\partial z} \frac{\partial h_{lm}^e}{\partial z} \right) \right) \quad (5.1)$$

we assume that they do not vary within the patch. In other words, instead of considering $\Lambda_x(\mathbf{x})$, where $\mathbf{x}$ varies over the patch Ω_I corresponding to the testing function h_{lm}^e , we shall consider Λ_x calculated at the nodal location $\mathbf{x}_I$ (which happens to be at the center of Ω_I). The same is also valid for Λ_y and Λ_z . In this way, the integral in (5.1) becomes:

$$\int_{\Omega} \left(\left(\Lambda_x(\mathbf{x}_I) \frac{\partial h_{jn}^e}{\partial x} \frac{\partial h_{Im}^e}{\partial x} + \Lambda_y(\mathbf{x}_I) \frac{\partial h_{jn}^e}{\partial y} \frac{\partial h_{Im}^e}{\partial y} + \Lambda_z(\mathbf{x}_I) \frac{\partial h_{jn}^e}{\partial z} \frac{\partial h_{Im}^e}{\partial z} \right) \right) \quad (5.2)$$

In this approximation, the components of the PML tensor become constants within Ω_I . But we argue that this approximation gets better and better as the size of the patches becomes smaller.

After this approximation, an interesting feature can be observed. If we represent testing functions by the indices I and m (regardless of their being a field component or pseudopressure testing function), and expansion functions by J and n (also regardless of their being a field component or a pseudopressure expansion function), it can be noticed that all the integrals boil down to certain *basic integrals* involving the product between a pair of functions. These basic integrals are in the Table 5.1 below.

TABLE 5.1 – BASIC INTEGRALS

$\int_{\Omega} h_{jn}^e h_{Im}^e$
$\int_{\Omega} \frac{\partial h_{jn}^e}{\partial x} \frac{\partial h_{Im}^e}{\partial x}, \quad \int_{\Omega} \frac{\partial h_{jn}^e}{\partial y} \frac{\partial h_{Im}^e}{\partial y}, \quad \int_{\Omega} \frac{\partial h_{jn}^e}{\partial z} \frac{\partial h_{Im}^e}{\partial z}$
$\int_{\Omega} h_{jn}^p \frac{\partial h_{Im}^e}{\partial x}, \quad \int_{\Omega} h_{jn}^p \frac{\partial h_{Im}^e}{\partial y}, \quad \int_{\Omega} h_{jn}^p \frac{\partial h_{Im}^e}{\partial z}$
$\int_{\Omega} h_{Im}^p \frac{\partial h_{jn}^e}{\partial x}, \quad \int_{\Omega} h_{Im}^p \frac{\partial h_{jn}^e}{\partial y}, \quad \int_{\Omega} h_{Im}^p \frac{\partial h_{jn}^e}{\partial z}$

A proper inspection of the entries in in (4.22.g), (4.22.k), (4.22.s), (4.23.d), (4.24.d) and (4.25.d) reveals that they can be reduced to combinations of the basic integrals in Table 5.1. Therefore, any integration process must focus on the evaluation of the integrals above.

5.1.2 Acceleration technique

Because the Shepard PU functions in (4.17.j) are non-polynomial, it is likely that the numerical integration based on Gaussian quadrature will require many points in order to attain a precise result. This is a delicate feature which plagues some meshfree methods, and the design of efficient integration rules constitutes one of the frontiers in research [De and Bathe, 2001], [Babuska *et al.*, 2009], [Ham *et al.*, 2014].

However, if the situation is such that the nodal distribution is uniform and all patches are the same size, then the cost of the numerical integrations can be drastically

reduced, provided we add some restrictions on the form assumed by the elements of the local spaces V_I^p in (4.19.c) and V_I^e in (4.19.k). This set of restrictions is characterized in Chart 5.1 below.

Chart 5.1 – Elements in the local spaces

The local spaces are

$$V_I^p = \text{span}\{\ell_{I,1}^p, \ell_{I,2}^p, \dots, \ell_{I,N_p}^p\}, \quad 1 \leq I \leq N \quad (5.3.a)$$

$$V_I^e = \text{span}\{\ell_{I,1}^e, \ell_{I,2}^e, \dots, \ell_{I,N_e}^e\}, \quad 1 \leq I \leq N, \quad (5.3.b)$$

according to (4.19.c), (4.19.k), (4.21.a) and (4.21.c). We assume that any function $\ell_{I,m}^e(\mathbf{x})$ is of the form

$$\ell_{I,m}^e(\mathbf{x}) = f_m(x - x_I, y - y_I, z - z_I), \quad 1 \leq I \leq N, \quad 1 \leq m \leq N_e \quad (5.3.c)$$

i.e., these functions depend on the difference between the point $\mathbf{x} = [x, y, z]^T$ (at which the function is calculated) and the nodal point $\mathbf{x}_I = [x_I, y_I, z_I]^T$. *The functions f_m are the same for all patches I .*

It is true that the Shepard PU function $\varphi_I^0(\mathbf{x})$ in (4.17.j) also has this same form, i.e., it depends on the difference between $\mathbf{x}$ and $\mathbf{x}_I$. Consequently, the meshfree basis functions $h_{Im}^e(\mathbf{x})$ defined by

$$h_{Im}^e(\mathbf{x}) = \varphi_I^0(\mathbf{x}) \ell_{I,m}^e(\mathbf{x}) \quad (5.3.d)$$

as in (4.19.y) will depend just on the difference between $\mathbf{x}$ and $\mathbf{x}_I$. The same conclusions hold for the pseudopressure spaces, i.e., if we assume that

$$\ell_{I,m}^p(\mathbf{x}) = g_m(x - x_I, y - y_I, z - z_I), \quad 1 \leq I \leq N, \quad 1 \leq m \leq N_p, \quad (5.3.e)$$

where *the g_m are the same for all patches I* , then the meshfree basis functions $h_{Im}^p(\mathbf{x}) = \varphi_I^0(\mathbf{x}) \ell_{I,m}^p(\mathbf{x})$ will also depend just on the difference between $\mathbf{x}$ and $\mathbf{x}_I$.

Suppose a two-dimensional uniform nodal distribution as in Fig.5.1. (The reasoning can be automatically and effortlessly transferred to the three dimensions. But the procedure to be introduced in the next lines is easier to present in two dimensions.) Let all the patches (associated with the nodes) be of the same size, and assume the local spaces have the form stated in Chart 5.1.

Consider the nodes I, J, K and L in Fig. 5.1. For any m and n , it is not difficult to conclude that

$$\int_{\Omega} h_{Jn}^e h_{Im}^e = \int_{\Omega} h_{Ln}^e h_{Km}^e, \quad (5.4.a)$$

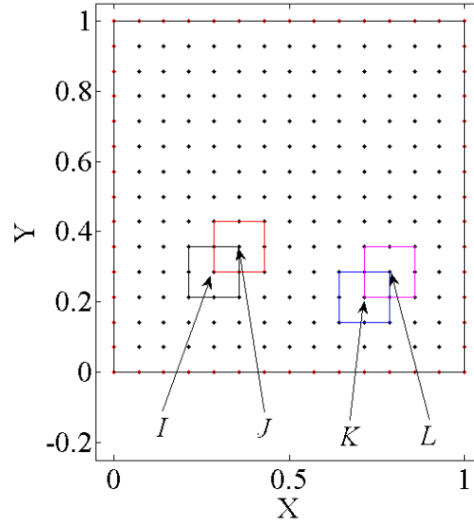


Fig. 5.1. A regular nodal distribution in two dimensions. All patches are the same size. The relative distances between nodes I and J and between K and L are the same.

because the relative position of nodes I and J is the same as that of nodes K and L . The same holds true for the other basic integrals in Table 5.1.

The consequence is that for any two pairs of nodes, if the relative positions of the nodes in each pair are the same, then the basic integrals evaluated for each pair will be the same. In other words: Let it be the pair formed by the nodes I and J and let it be another pair formed by the nodes K and L . If the relative position of nodes I and J and the relative position of nodes K and L are the same, then it follows that the basic integrals evaluated for the pair I and J will be the same as those evaluated for the pair K and L .

The conclusion is that in a regular arrangement of nodes, *the basic integrals need to be calculated only once*.

For example, suppose we are considering the interaction between nodes I and J in Fig. 5.1. By this we mean that we calculate all the basic integrals in Table 5.1. Later, when calculating the interaction between nodes K and L , these integrals do not need to be calculated again: Their values are available from the calculations regarding I and J .

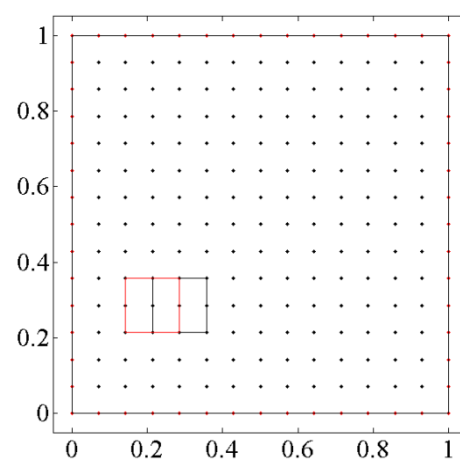
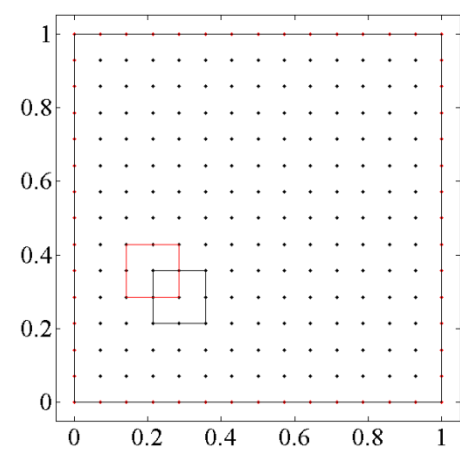
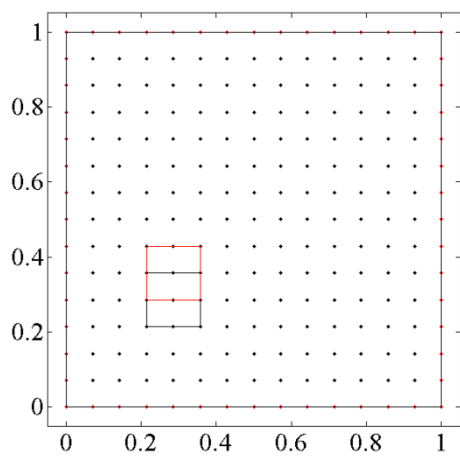
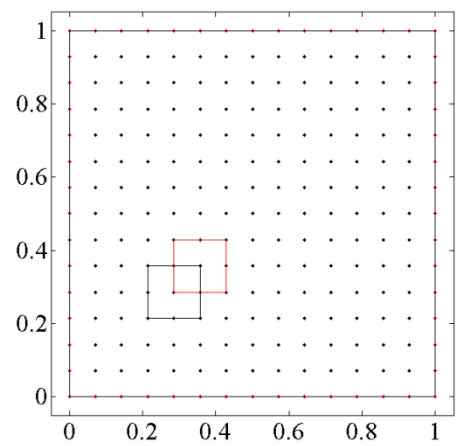
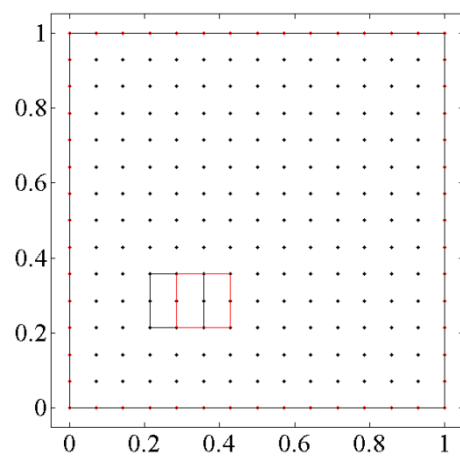
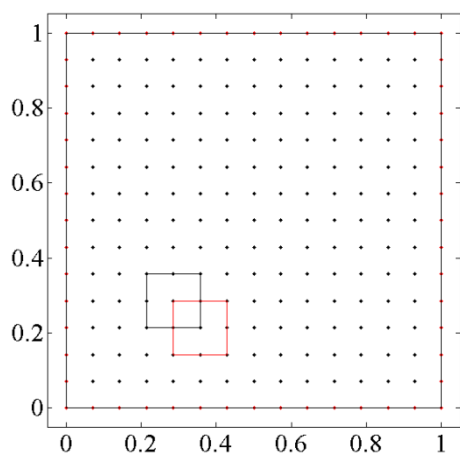
So the idea goes as follows:

First: Take a node I in the middle of a regular nodal arrangement.

Second: Determine all neighboring nodes which interact with I .

Third: Evaluate the interaction (basic integrals) between node I and each neighbor from the last step.

Fourth: Store this numerical information in suitable data structures.



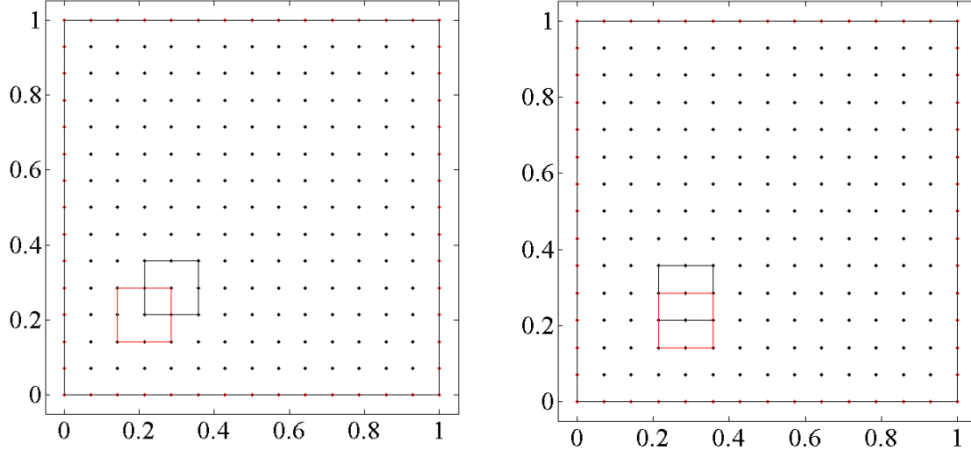


Fig. 5.2. The eight patches (in red) which are able to intersect a given patch (in black) in a given regular arrangement of nodes. Of course, the patch intersects itself, so in the end, for each patch I , there are 9 patches which intersect it. As J runs from 1 to 8, we get the eight figures above. Alternatively, one can say that each node I has 8 neighbors able to influence it (in addition to node I itself).

Fifth: Run through all the other nodes in the domain. For each node, determine all neighboring nodes able to interact with it.

Six: The interaction between this new node and its neighbors has already been calculated in the third step.

In this way, the only action that needs to be performed is a careful identification between nodes and its neighbors, and the subsequent mapping of the entries to the global matrix. When the regular arrangement of nodes is such that the size of the square patch is such that r_I is just the horizontal distance between two adjacent nodes as in Fig. 5.1, then each node I is influenced by itself and by the eight surrounding nodes, as in Fig. 5.2.

We shall not delve deeply into this subject, as the idea is sufficiently understandable, and also because the majority of work is done at the implementation level. So each one has a more or less clear idea on how to implement this ‘reuse approach’ according to the way he/she wants to develop his/her code.

But this ‘reuse approach’ can be employed in a (theoretically) infinite and regular nodal distribution. Actually, the nodal distributions are finite, which means that symmetry will be broken at the boundary nodes, i.e., these nodes do not have all the neighbor nodes that the nodes in the bulk of the domain have. Moreover, there will be situations in which the nodal distribution will not be regular (for example, when a scatterer with a complicated shape is considered).

The idea is to divide the nodal distribution into two parts: A *regular part* and a *non-regular part*. So if a total of N nodes is employed, then

$$N = N_r + N_{nr}, \quad (5.4. b)$$

where N_r is the number of nodes in the regular part of the distribution and N_{nr} the number of nodes in the non-regular part. We must now establish a criterion that allows us to say if a given node I belongs to the regular or to the non-regular part.

In the class of problems we are interested in, the outer boundary Γ_o will be a rectangle (remember, the explanation refers to the two-dimensional case). The nodal distribution can be set up in four steps as follows:

First: Begin by spreading nodes in a regular fashion throughout the rectangular region whose outer boundary is Γ_o , as in a grid.

Second: Adjust the size of the square patches so that for each patch I , r_I is just the horizontal distance between two adjacent nodes.

Third: Remove those nodes in the interior of the rectangular region that fall within the PEC scatterer.

Fourth: Spread nodes along the boundary of the PEC scatterer, i.e., along Γ_1 .

Fifth: Recalculate the size of the patches. They should not intersect Γ_1 , according to consideration 2 in Chart 4.2.

The third, fourth and fifth steps make the distribution ‘locally’ non-regular on and around the PEC surface Γ_1 . After the nodal distribution has been set up, we must loop through all nodes in order to find out if it falls within the regular or within the non-regular part. The criterion we established is in the form of an algorithm.

1. Take a node at $\mathbf{x}_I = [x_I, y_I]^T$, $1 \leq I \leq N$.

2. Consider the set of points which surround $\mathbf{x}_I$, i.e., consider the eight points

$$P_1 = [x_I - r_I, y_I - r_I]$$

$$P_2 = [x_I, y_I - r_I]$$

$$P_3 = [x_I + r_I, y_I - r_I]$$

$$P_4 = [x_I - r_I, y_I]$$

$$P_5 = [x_I + r_I, y_I]$$

$$P_6 = [x_I - r_I, y_I + r_I]$$

$$P_7 = [x_I, y_I + r_I]$$

$$P_8 = [x_I + r_I, y_I + r_I]$$

3. Is there a node at each one of the eight points from Step 2? If no, then node I belongs to the non-regular part of the nodal distribution. If yes, go to the next step.

4. Consider the patches associated with all the eight nodes located in $P_1 - P_8$. Are they the same size? If no, then node I belongs to the non-regular part of the nodal distribution. If yes, then node I belongs to the regular part.

The nodes which comprise the regular part of the nodal distribution can be treated in the same way as the nodes from the (theoretically) infinite regular nodal distribution discussed earlier. It means that we can take any one of them and calculate its interaction (i.e., the basic integrals) with all the eight neighbors. When considering any other node in the regular part, the interactions need not be calculated again: They are available from the previous calculation. The only work is to map the entries to the global matrix.

On the other hand, if a node I is in the non-regular part, the sizes of its associated patch and those of its neighbors will be different. In this case, the basic integrals must be calculated in the traditional way, i.e., there is no reuse procedure.

The extension of these ideas to three dimensions is straightforward. The difference is that there will be 26 nodes surrounding a given node $\mathbf{x}_I$, instead of just eight.

If the geometry of the computational domain Ω is conducive to a large number of nodes being able to be included in a regular distribution, then the gain in setting up the global matrix is enormous, particularly in three dimensions, where the numerical integrations are very expensive. Fortunately, this is the case, as for the category of problems in which we are interested, the domain is basically a parallelepiped with a hole within (the PEC scatterer). The nodal distribution will be regular in the bulk of the domain, and becomes non-regular only in the vicinity of the scatterer. A very attractive scenario, indeed.

5.1.3 Numerical quadrature

When it comes to the actual numerical integration of terms in Table 5.1, we employ the traditional Gaussian quadrature. The process will be illustrated for the first of them only; the reasoning can of course be extended to the others.

We want to compute the value of the integral

$$\mathbb{I} = \int_{\Omega} h_{jn}^e h_{Im}^e. \quad (5.5.a)$$

Since according to (4.17.g) the support of the test function h_{Im}^e is contained in the patch Ω_I , the integral above becomes

$$\mathbb{I} = \int_{\Omega_I} h_{jn}^e h_{Im}^e. \quad (5.5.b)$$

In two dimensions, the patch Ω_I is a square. Instead of applying the Gaussian quadrature to Ω_I (and therefore employ many integration points), we find it better to divide the square Ω_I into smaller squares, and then apply the Gaussian quadrature to each of these small squares (but this time with less integration points). In our experiments, we found that dividing Ω_I into 6×6 squares yields results with a good precision. In this way, the patch Ω_I is expressed as

$$\Omega_I = \bigcup_{k=1}^{36} \omega_k \quad (5.5. c)$$

i.e., as the union of the smaller squares ω_k . Of course, if $l \neq k$, then $\omega_l \cap \omega_k = \emptyset$, i.e., the smaller squares do not intersect with each other (except at their boundaries). The integral in (5.5.b) becomes

$$\mathbb{I} = \sum_{k=1}^{36} \int_{\omega_k} h_{jn}^e h_{lm}^e \quad (5.5. d)$$

We can now apply a simple 2-point quadrature rule in the x and y -directions of each of the integrals in (5.5.d). If we represent ω_k as a Cartesian product of intervals:

$$\omega_k = [a_k, b_k] \times [c_k, d_k], \quad (5.5. e)$$

then these ‘subintegrals’ can be computed as

$$\begin{aligned} \int_{\omega_k} h_{jn}^e h_{lm}^e &\cong \\ &\cong \frac{(b_k - a_k)}{2} \frac{(d_k - c_k)}{2} \sum_{i=1}^2 \sum_{j=1}^2 w_i w_j h_{jn}^e(x_i, y_j) h_{lm}^e(x_i, y_j), \end{aligned} \quad (5.5. f)$$

where the weights are given by $w_1 = 1$, $w_2 = 1$ and the coordinates x_i and y_j are given by

$$x_i = \frac{(b_k - a_k)}{2} \xi_i + \frac{(a_k + b_k)}{2} \quad (5.5. g)$$

$$y_j = \frac{(d_k - c_k)}{2} \xi_j + \frac{(c_k + d_k)}{2}. \quad (5.5. h)$$

The parameters ξ_i and ξ_j are given by

$$\xi_1 = -\sqrt{1/3} \quad (5.5. i)$$

$$\xi_2 = \sqrt{1/3}.$$

In three dimensions, the patch Ω_I is a cube, which is divided into $6 \times 6 \times 6$ little cubes. Therefore,

$$\Omega_I = \bigcup_{k=1}^{216} \omega_k \quad (5.5.j)$$

and consequently

$$\mathbb{I} = \int_{\Omega_I} h_{jn}^e h_{lm}^e = \sum_{k=1}^{216} \int_{\omega_k} h_{jn}^e h_{lm}^e. \quad (5.5.k)$$

Each little cube can be represented as a Cartesian product of intervals:

$$\omega_k = [a_k, b_k] \times [c_k, d_k] \times [e_k, f_k], \quad (5.5.l)$$

which allows the ‘subintegrals’ to be written as

$$\begin{aligned} \int_{\omega_k} h_{jn}^e h_{lm}^e &\cong \\ &\cong \frac{(b_k - a_k)}{2} \frac{(d_k - c_k)}{2} \frac{(f_k - e_k)}{2} \sum_{i=1}^2 \sum_{j=1}^2 \sum_{l=1}^2 w_i w_j w_l h_{jn}^e(x_i, y_j, z_l) h_{lm}^e(x_i, y_j, z_l) \quad . \end{aligned} \quad (5.5.m)$$

The weights are the same as those from the two-dimensional case, and the coordinates x_i and y_j are exactly those from (5.5.g) and (5.5.h). The coordinate z_l is given by

$$z_l = \frac{(f_k - e_k)}{2} \xi_l + \frac{(e_k + f_k)}{2} \quad (5.5.n)$$

The parameters ξ_l are those from (5.5.i).

5.2 The inf-sup stability test

When setting up the nodal distribution, during the first step outlined in Section 5.1.2 (which says that we begin with a regular distribution over the rectangle/parallelepiped whose surface is Γ_o), we can retrieve the value of the distance between two adjacent nodes and call it h . This h is sometimes called the *discretization length*, and intuitively, it gets smaller as more and more nodes are considered. This is the meaning of h referred to at the beginning of Section 4.1.1.

In this way, each nodal distribution has its associated discretization length h , and at the same time it serves as a basis for the finite-dimensional subspaces $\mathbb{V}_\tau^h(\Omega)$ and $\mathbb{P}^h(\Omega)$. So we can, in a sense, ‘identify’ a value of h and a pair of spaces $\mathbb{V}_\tau^h(\Omega)$ and $\mathbb{P}^h(\Omega)$. This is the reason for the superscript h in both of them.

According to the discussion in Section 4.3.2, the pair of finite-dimensional spaces $\mathbb{V}_\tau^h(\Omega)$ and $\mathbb{P}^h(\Omega)$ must obey the inf-sup condition

$$\exists \beta_b^h > 0 \quad s. t. \quad \inf_{q_h \in \mathbb{P}^h(\Omega) \setminus \{0\}} \sup_{v_h \in \mathbb{V}_\tau^h(\Omega) \setminus \{0\}} \frac{\left| - \int_\Omega q_h \nabla \cdot v_h \right|}{\|v_h\|_{H^1(\Omega)^3} \|q_h\|_{L^2(\Omega)}} \geq \beta_b^h. \quad (5.6.a)$$

The inf-sup condition β_b^h depends on h (i.e., on the finite-dimensional subspaces $\mathbb{V}_\tau^h(\Omega)$ and $\mathbb{P}^h(\Omega)$). The role of the inf-sup condition (5.6.a) was studied in Section 4.2.2: it is ultimately responsible for the uniqueness of the solution to the global linear system. But uniqueness of the numerical solution is related to the fact the global matrix is invertible. The question is that if the inf-sup is not obeyed, invertibility of the global matrix is put at risk. Since $\beta_b^h = 0$ implies that the discretized problem is not solvable, we must guard ourselves against this situation.

Suppose we constructed a pair of discrete spaces $\mathbb{V}_\tau^h(\Omega)$ and $\mathbb{P}^h(\Omega)$. If they pass the inf-sup condition, i.e., if we find a $\beta_b^h > 0$ such that (5.6.a) is satisfied, then it is fine, and the solution to the problem can be found. But suppose now that we want a more precise solution, and we construct a refined pair of spaces $\mathbb{V}_\tau^{h'}(\Omega)$ and $\mathbb{P}^{h'}(\Omega)$, based on a discretization such that $h' < h$. This new pair must be tested again to verify if they pass the inf-sup test, i.e., if we can find another $\beta_b^{h'} > 0$ satisfying (5.6.a). Let us say that this new pair of spaces again passes the inf-sup test, which allows us to find a more precise solution. Suppose now that we want an even more precise solution, so we construct another refined pair of spaces $\mathbb{V}_\tau^{h''}(\Omega)$ and $\mathbb{P}^{h''}(\Omega)$, based on a discretization such that $h'' < h'$. We must apply the test again and verify if we can find a $\beta_b^{h''} > 0$ which satisfies (5.6.a).

It is instructive to observe the behavior of these inf-sup constants as the h gets smaller, i.e., as the discretizations get more and more refined. Of course, *they should always stay away from zero*. Even if they do not assume the value zero, very small values for this constant may indicate that the global matrix is ‘getting close to a singular matrix’, and it is likely that numerical problems will occur. (Moreover, the estimate (4.6.h) in Theorem 4.1 says that the norm of the solution depends on a constant K_2 multiplying the norm of the functional g^* , which, according to the identification (4.5.h), is related to the lifting function $\mathbf{u}_h^g$ whose form we studied in (4.22.l) – (4.22.r). But the estimate (A2.62) in Appendix A.2 reveals that this constant K_2 is inversely proportional to the inf-sup constant. So if the inf-sup constant approaches zero as h gets smaller, it may happen that the solution becomes unbounded.)

The idea to inspect the values of the inf-sup constant as the discretization length h gets smaller is called the *inf-sup test* and it is due to K. J. Bathe [Bathe, 2001], [De and Bathe², 2001].

It is said that a family of pairs of finite-dimensional subspaces of $\mathbb{V}_\tau(\Omega)$ and $\mathbb{P}(\Omega)$ pass the test if the *stability* criterion is satisfied:

$$\exists \beta_b > 0 \text{ such that } \lim_{h \rightarrow 0} \beta_b^h = \beta_b \quad (5.6.b)$$

i.e., there should exist a positive constant β_b , independent of h , such that the inf-sup constants β_b^h of all finite-dimensional subspaces $\mathbb{V}_\tau^h(\Omega)$ and $\mathbb{P}^h(\Omega)$ converge to this β_b .

In practice, it takes a sequence of pairs of subspaces $\mathbb{V}_\tau^h(\Omega)$ and $\mathbb{P}^h(\Omega)$ such that $h \rightarrow 0$, finding their associated inf-sup values β_b^h , and then observing what happens to this sequence of values. If they approach zero, then these spaces fail the test. Ideally, they should converge to a positive value.

When constructing our meshfree subspaces for $\mathbb{V}_\tau(\Omega)$ and $\mathbb{P}(\Omega)$, we consider different choices for the local spaces V_I^p and V_I^e in (5.3.b) and (5.3.a), respectively. These local spaces will originate global spaces $\mathbb{V}_\tau^h(\Omega)$ and $\mathbb{P}^h(\Omega)$ with different characteristics, and we must find out if they pass the inf-sup test (5.6.b). In this way, we can identify which pair of meshfree spaces form compatible pairs, in the sense that they not only satisfy the inf-sup condition, but that they continue to satisfy it as the discretization length gets smaller.

Some observations are in order. Does it mean that, given a problem stated in any computational domain Ω , one needs to find the inf-sup values associated with a family of discretizations set up in Ω ? Ideally, yes. But in order to find the inf-sup values, one needs to solve an eigenvalue problem, as in (4.16.n). However, solving these eigenproblems may be a very expensive task, particularly when the number of DoF's involved in the problem becomes larger as $h \rightarrow 0$. What is generally done is to apply the inf-sup test to simple domains Ω , [De and Bathe², 2001], and extend the conclusions to larger/more complicated domains. (Much in the same way as in the experimental study of convergence rates of a given meshfree/finite element space: One usually chooses a simple domain, find the convergence rates and then extends the conclusion to other domains.)

In this work, the terms in the local basis are monomials. For two-dimensional problems, we inspect meshfree spaces whose local bases are given by

$$Z^0 = \text{span}\{1\} \quad (5.6.c)$$

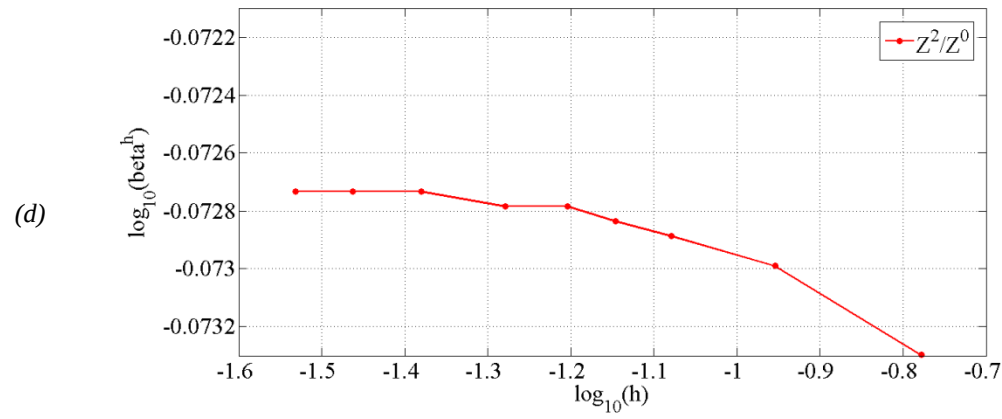
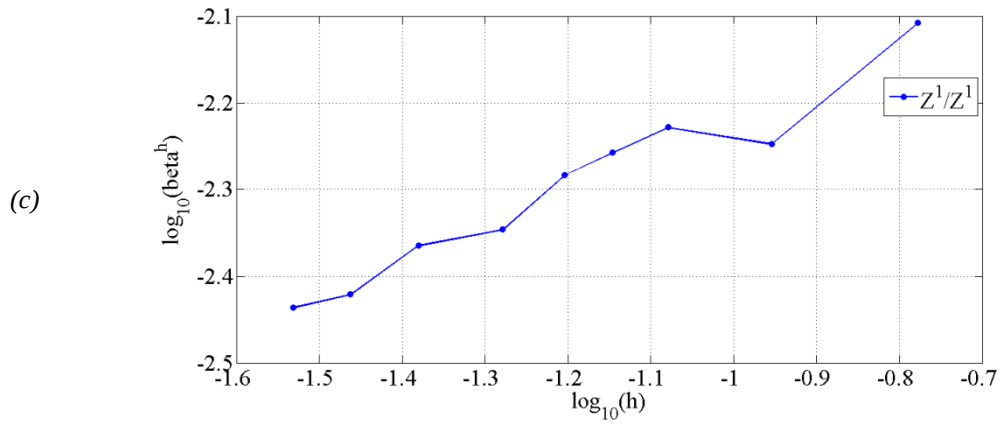
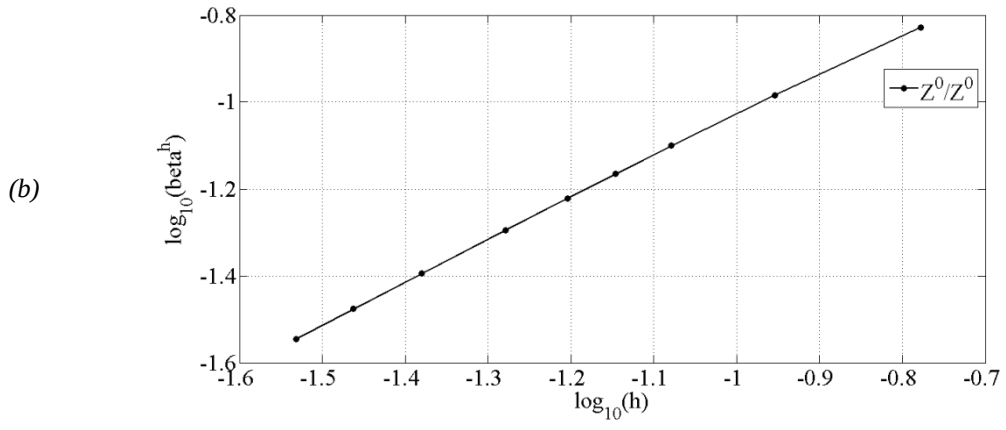
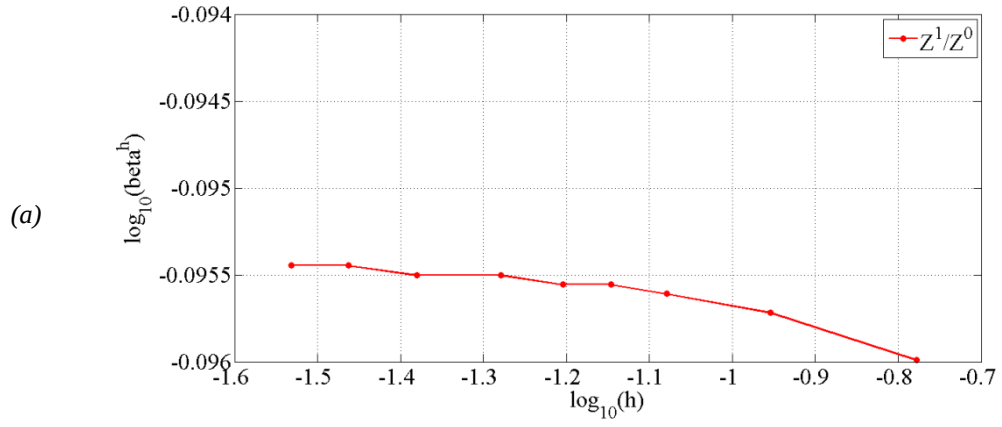
$$Z^1 = \text{span}\{1, X_I, Y_I\} \quad (5.6.d)$$

$$Z^2 = \text{span}\{1, X_I, Y_I, X_I^2, X_I Y_I, Y_I^2\} \quad (5.6.e)$$

where $1 \leq I \leq N$ and

$$X_I = \frac{x - x_I}{r_I} \quad (5.6.f)$$

$$Y_I = \frac{y - y_I}{r_I}. \quad (5.6.g)$$



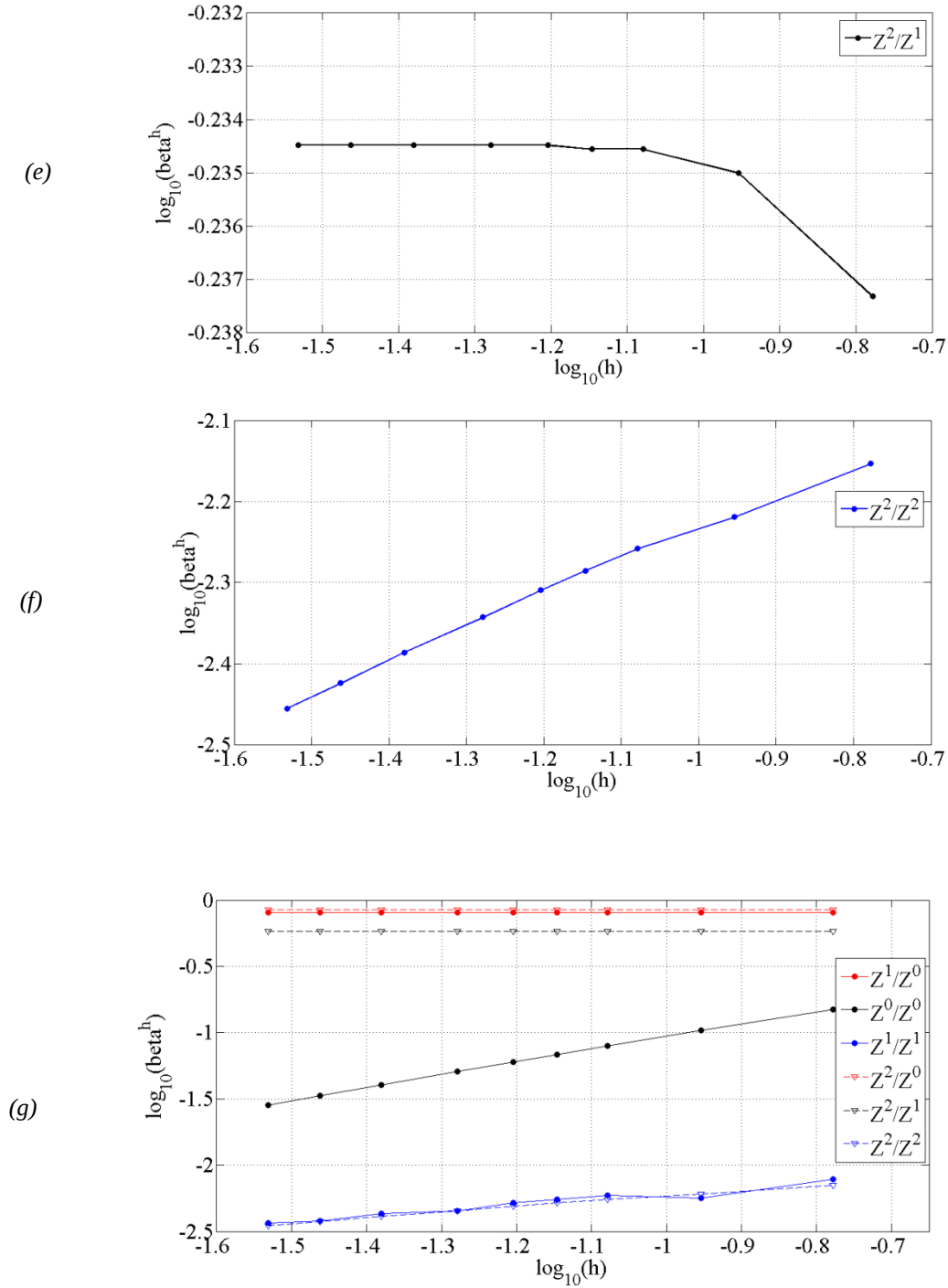


Fig. 5.3. Results for the inf-sup stability test in 2 dimensions. (a) The pair Z^1/Z^0 passes the test, as the inf-sup values are, for all practical purposes, constant (observe how the y-coordinates are almost constant). (b) The inf-sup values for the Z^0/Z^0 pair steadily decrease with h , and therefore fail the test. (c) The inf-sup values for the Z^1/Z^1 pair also decrease with h , but not in a steady way. But even so, they fail to converge to a positive value, and therefore do not pass the test. (d) The Z^2/Z^0 pair also passes the test, as the inf-sup values are almost constant (i.e., they stabilize at a positive value). (e) The same conclusion hold for the Z^2/Z^1 pair: It also passes the test. (f) The values for the Z^2/Z^2 pair decrease with h , and therefore fail the test. (g) When the results are plotted on the same graph, it becomes evident which pairs pass and which fail the test.

In other words, the local spaces V_I^e and V_I^p in (5.3.b) and (5.3.a) will be chosen among (5.6.c) – (5.6.e). These local spaces have the same form for all patches. For example we can take $V_I^p = Z^0$, $1 \leq I \leq N$, and $V_I^e = Z^1$, $1 \leq I \leq N$. This choice will produce global spaces V^p and V^e , that by their turn will be used in the construction of $\mathbb{P}^h(\Omega)$ and $\mathbb{V}_\tau^h(\Omega)$, according to (4.19.e), (4.19.o) and (4.19.p). After we get these finite-dimensional subspaces, the inf-sup stability test described earlier must be applied, in order to find out if they form a compatible pair.

We have tested a number of combinations of these local spaces and applied the inf-sup stability test. The domain Ω is the square $[0,1] \times [0,1]$. If we choose Z^1 for the V_I^e and Z^0 for the V_I^p , this combination will be referred to as Z^1/Z^0 . The same applies to the other choices. The result is in Fig. 5.3.

The analysis of Fig. 5.3 reveals that the pairs Z^1/Z^0 , Z^2/Z^0 and Z^2/Z^1 pass the test, since they converge to a value away from zero as the discretization h decreases. On the other hand, the pairs Z^0/Z^0 , Z^1/Z^1 and Z^2/Z^2 do not pass the test. The reason is that the associated inf-sup values steadily decrease with h , thus violating (5.6.b). The space Z^0/Z^0 is peculiar: In addition to the decreasing inf-sup values, we get zero eigenvalues when solving the eigenproblem (4.16.n), which indicate the presence of spurious modes.

In three dimensions, we inspect meshfree spaces whose local bases are given by

$$Z^0 = \text{span}\{1\} \quad (5.6.h)$$

$$Z^1 = \text{span}\{1, X_I, Y_I, Z_I\} \quad (5.6.i)$$

where X_I and Y_I are as in (5.6.f) and (5.6.g), respectively, and

$$Z_I = \frac{z - z_I}{r_I} \quad (5.6.j)$$

The procedure is analogous to that in the two-dimensional case, but the domain Ω is now the cube $[0,1] \times [0,1] \times [0,1]$, and the inf-sup stability test is applied to certain choices for the local spaces. The result is in Fig. 5.4.

According to Fig. 5.4, the pair Z^1/Z^0 is the only one which passes the test. The pairs Z^0/Z^0 and Z^1/Z^1 fail the test, as the inf-sup values also decrease with h . As it happens in the two-dimensional case, there are zero eigenvalues associated with the pair Z^0/Z^0 .

Now that we have identified which choices for the local spaces yield compatible pairs, i.e., pairs which satisfy the discrete inf-sup condition, they can be safely employed in the construction of our meshfree spaces $\mathbb{V}_\tau^h(\Omega)$ and $\mathbb{P}^h(\Omega)$. We can now move on and apply them to the solution of the scattering problems. Before we proceed, some clarification regarding the solution of the global linear system is in order.

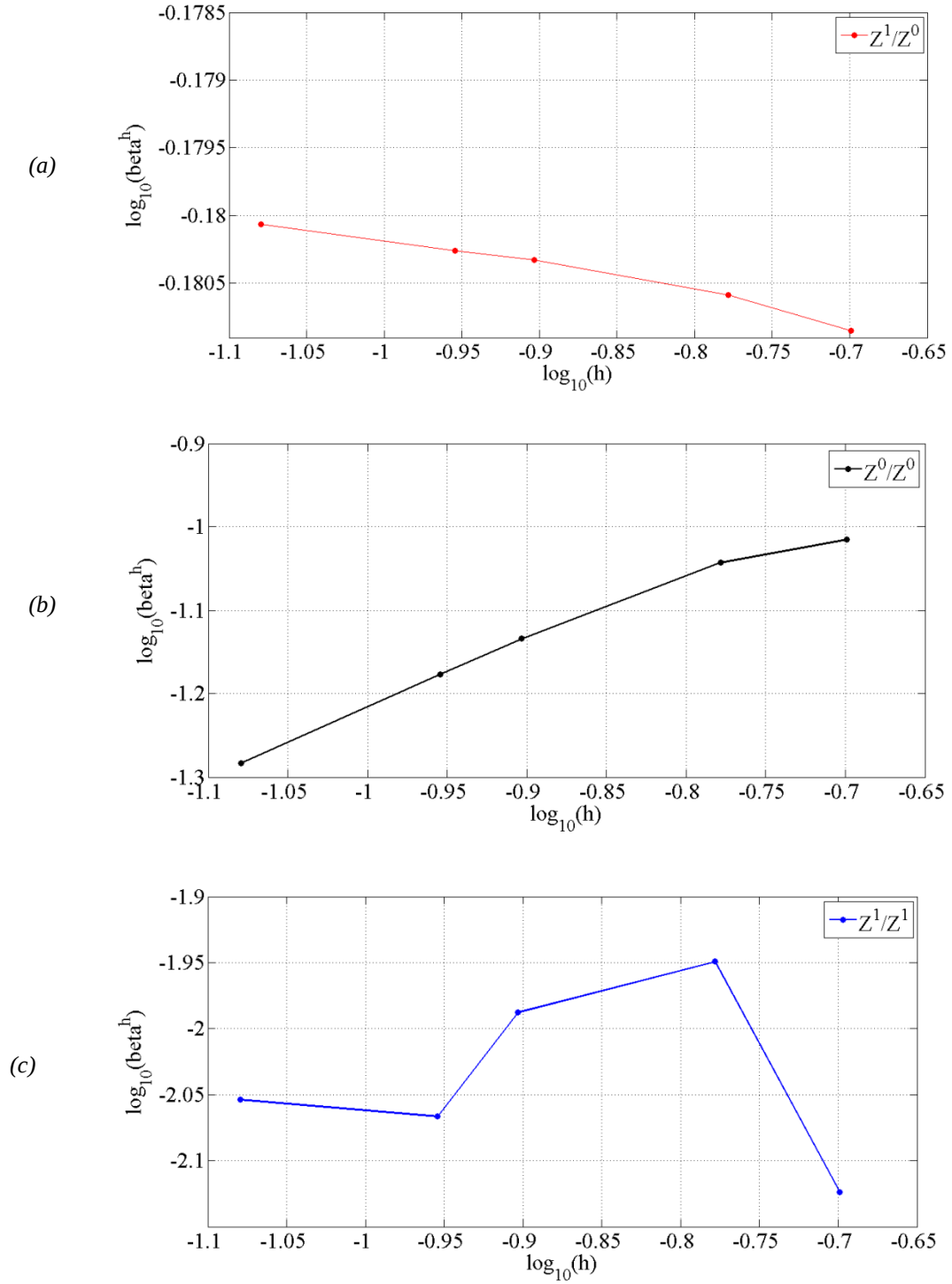


Fig. 5.4. Results for the inf-sup stability test in 3 dimensions. (a) The inf-sup values for the Z^1/Z^0 are slightly increasing, and converge to a positive value (observe how the y-coordinates are almost constant). (b) The inf-sup values for the pair Z^0/Z^0 steadily decrease with h , and therefore fail the test. (c) The values regarding the Z^1/Z^1 pair exhibit an erratic behavior, and fail to converge to some value. It cannot satisfy (5.6.b).

5.3 Preconditioning

According to (4.14.z), we are led to a (sparse) global linear system of the form

Find $(\bar{\mathbf{e}}, \bar{\mathbf{p}}) \in \mathbb{C}^Q \times \mathbb{C}^M$ such that

$$\begin{bmatrix} \bar{\mathbf{A}} & \bar{\mathbf{B}}^T \\ \bar{\mathbf{B}} & \mathbf{0} \end{bmatrix} \begin{bmatrix} \bar{\mathbf{e}} \\ \bar{\mathbf{p}} \end{bmatrix} = \begin{bmatrix} \bar{\mathbf{f}} \\ \bar{\mathbf{g}} \end{bmatrix} \quad (5.7.a)$$

When considering more refined discretizations, it is likely that the total number of DoF's will be considerably large, particularly in three-dimensional problems. By this we mean that, in three-dimensions, the total number of DoF's will, in all probability, be larger than 100 000. In this way, trying to solve the system (5.7.a) by a direct method will not be a feasible option.

The system in (5.7.a) will be solved by an iterative method. We found that the *generalized minimum residual* (GMRES) method suits our purposes [van der Vorst, 2009], [Saad, 2003]. However, as it is known, iterative methods for the solution of a given linear system may suffer from slow convergence, or even fail to converge at all. In other words, the iterative method needs *preconditioning*.

The system (5.7.a) can be written in the familiar form as

$$\mathbf{K}\mathbf{u} = \mathbf{F}, \quad (5.7.b)$$

where $\mathbf{K}$ is the associated sparse matrix, $\mathbf{u}$ is the vector of unknowns and $\mathbf{F}$ is a known vector. The GMRES algorithm, when applied directly to (5.7.b), may not work properly. The preconditioning is just a matrix $\mathbf{M}$ which operates as

$$\mathbf{M}^{-1}\mathbf{K}\mathbf{u} = \mathbf{M}^{-1}\mathbf{F}. \quad (5.7.c)$$

The solution of both linear systems (5.7.b) and (5.7.c) are the same. However, the GMRES (or any other iterative algorithm) should work better in (5.7.c) than in (5.7.b). In loose terms, the matrix $\mathbf{M}^{-1}\mathbf{K}$ has 'nicer' properties than the matrix $\mathbf{K}$, which allows the performance of the GMRES to improve significantly.

Trying to find suitable preconditioning matrices $\mathbf{M}$ is a very complicated problem, and it constitutes an area of research by its own [Saad, 2003]. It should satisfy some criteria, one of them is that the process of getting $\mathbf{M}$ should be more or less inexpensive.

The matrix in (5.7.a) has a saddle-point structure [Boffi *et al.*, 2013]. There is a class of preconditioners for saddle-point problems, documented in the literature [Benzi and Golub, 2004], [Benzi and Wathen, 2008], [Quarteroni, 2009]. Our choice for the preconditioning matrix $\mathbf{M}$ is

$$\mathbf{M} = \begin{bmatrix} \bar{\mathbf{A}} & \mathbf{0} \\ \mathbf{0} & \bar{\mathbf{B}}\bar{\mathbf{D}}^{-1}\bar{\mathbf{B}}^T \end{bmatrix}. \quad (5.7.d)$$

The matrix $\bar{\mathbf{D}}$ in (5.7.d) is a diagonal matrix, whose entries are those in the diagonal of $\bar{\mathbf{A}}$, i.e.,

$$[\bar{\mathbf{D}}]_{ij} = \begin{cases} [\bar{\mathbf{A}}]_{ij}, & i = j \\ 0, & i \neq j \end{cases} \quad (5.7.e)$$

The inverse matrix $\bar{\mathbf{D}}^{-1}$ is therefore very easy to compute:

$$[\bar{\mathbf{D}}^{-1}]_{ij} = \begin{cases} 1/[\bar{\mathbf{A}}]_{ij}, & i = j \\ 0, & i \neq j \end{cases} \quad (5.7.f)$$

In a sense, the computation of the preconditioning matrix $\mathbf{M}$ in (5.7.d) is not complicated, and we found that it works fine in conjunction with the GMRES.

It should be mentioned that actual research of finding the single most suitable preconditioning matrix $\mathbf{M}$ is beyond the scope of this thesis. Nevertheless, it constitutes an excellent proposal for a future work.

5.4 Case studies

In all examples to follow, both in two and three dimensions, we shall always employ the pair Z^1/Z^0 . The reason is that the pairs described by higher order terms produce more DoF's. The pair Z^1/Z^0 is the 'simplest' of those pairs which pass the inf-sup test, and it is worthwhile to dedicate some attention to evaluate its performance when applied to different problems.

5.4.1 Free-space: Error

In order to retrieve the discretization error, we consider a cubic region $(0,1) \times (0,1) \times (0,1)$ (in meters). We want to solve the problem

Find $(\mathbf{E}, p)$ such that

$$\nabla^2 \mathbf{E} + k_0^2 \mathbf{E} + \nabla p = \mathbf{0}, \quad \text{in } \Omega \quad (5.8.a)$$

$$\nabla \cdot \mathbf{E} = 0, \quad \text{in } \Omega \quad (5.8.b)$$

$$\mathbf{E} = e^{-jk_0 x} \hat{\mathbf{z}}, \quad \text{at } \Gamma \quad (5.8.c)$$

This problem represents a cubic region in free-space, in which a plane wave propagates. It does not represent a scattering problem, but it is useful as a means to extract convergence rates, since the analytical solution to this problem is just

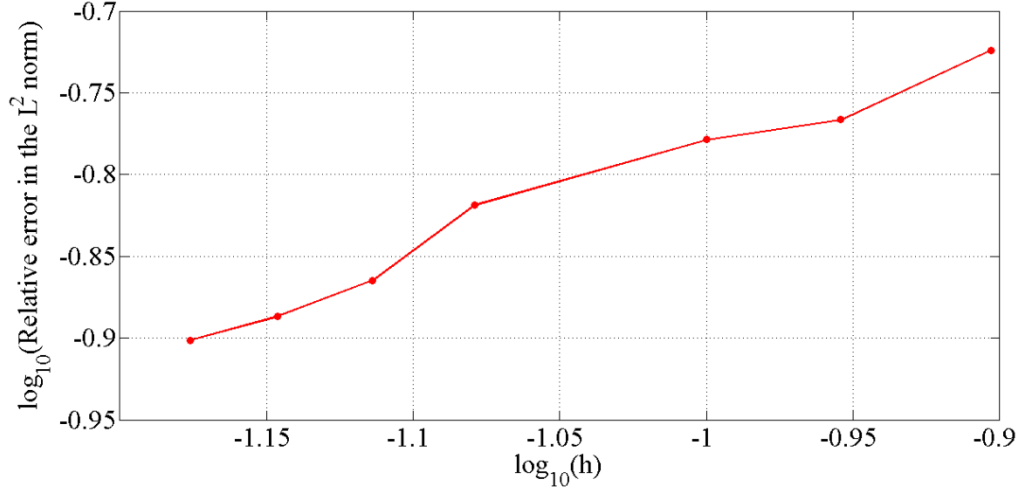


Fig. 5.5. The relative error in the free-space problem as a function of the discretization length h .

$$\mathbf{E}(\mathbf{x}) = e^{-jk_0 x} \hat{\mathbf{z}}, \quad \text{in } \Omega. \quad (5.8.d)$$

The reason is that there are neither sources nor scatterers to disturb the field. The variational formulation of this problem resembles that of the Navier-Stokes system. We can apply the trace operator γ_0^d from (2.58), since all components of the electric field are prescribed at the boundary (and not just the tangential components, as it happens for the scattering problems).

The sole purpose of this example is to measure the relative error resulting from the meshfree approximation, i.e., we evaluate

$$\epsilon(h) = \frac{\|\mathbf{E} - \mathbf{E}_h\|_{L^2(\Omega)^3}}{\|\mathbf{E}\|_{L^2(\Omega)^3}}, \quad (5.8.e)$$

where $\mathbf{E}$ is that from (5.8.d). Of course, the relative error ϵ is a function of the discretization length h . So we evaluate (5.8.e) for different pairs of spaces (for the components of the electric field and for the pseudopressure). The result is in Fig. 5.5.

Figure 5.5 reveals that the relative error decreases as h gets smaller. A linear regression applied to the curve in Fig. 5.5 reveals that the relation between ϵ and h is approximately given by the form (where C is a positive constant):

$$\epsilon(h) = Ch^{0.644}. \quad (5.8.f)$$

5.4.2 Scattering of a TE^z plane wave by a circular cylinder

The problem concerning the scattered field by a PEC circular cylinder has an analytical solution, given in terms of series of Hankel functions [Balanis, 1989]. Let it be a square region $\Omega = (-0.5, 0.5) \times (-0.5, 0.5)$ (in meters). In this region we make a circular hole whose radius is $a = 1/6$. This corresponds to the cross section of a PEC circular cylinder of the same radius.

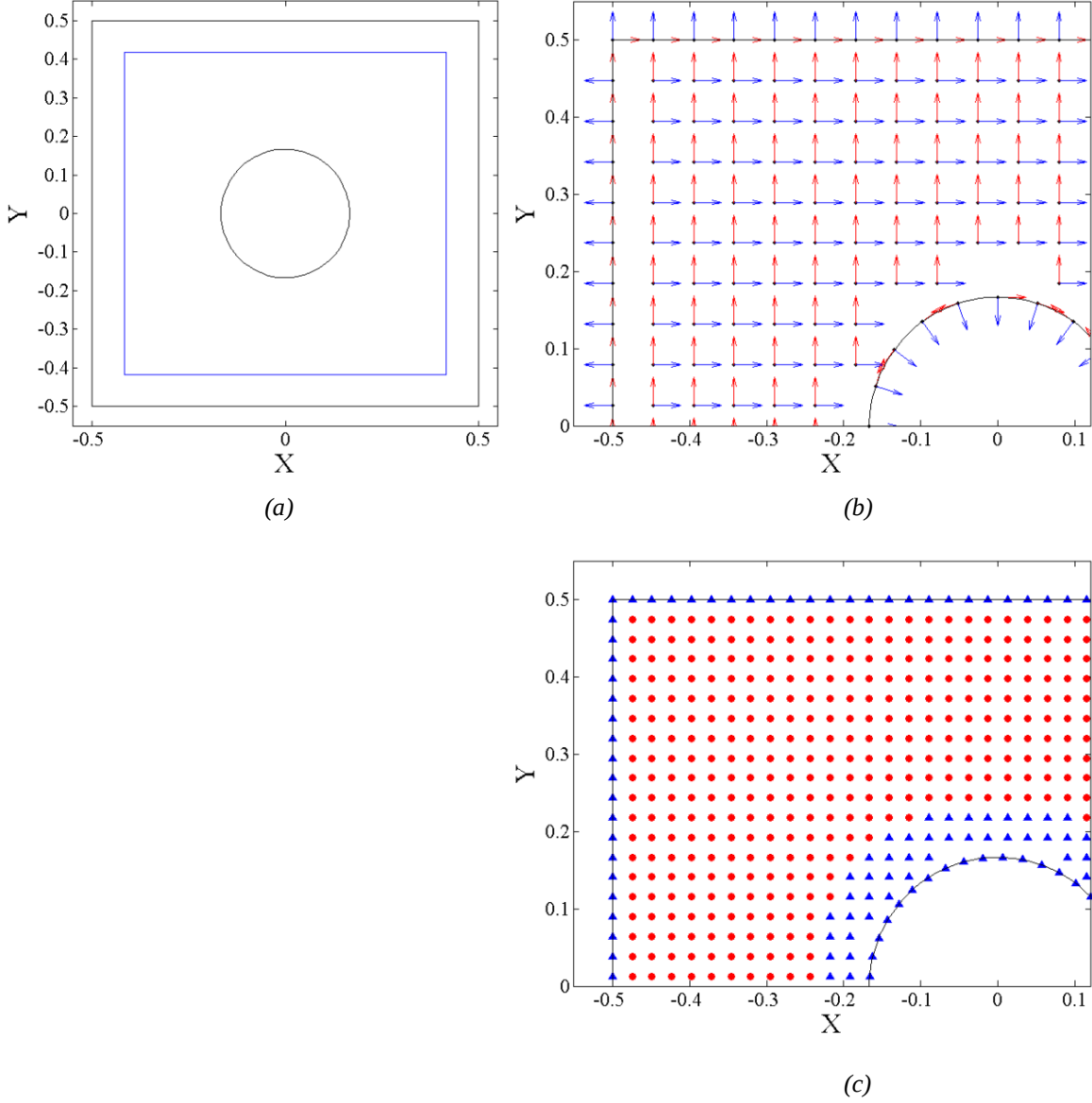


Fig. 5.6. (a) The computational domain, comprising the circular scatterer (the cylinder cross section) and the PML. (b) The elemental directions associated with each node. For the interior nodes, they are just the Cartesian directions $\hat{x}$ and $\hat{y}$. For the scatterer nodes, they happen to be the normal and tangential directions at the location of each node. (c) In this portion of the domain, we can see the nodes in the regular part of the distribution (represented by red circles) and the nodes in the non-regular part (blue triangles). The nodes in the regular part are in the bulk of the domain, whereas the nodes in the non-regular part happen to be on and around the boundaries.

We choose a wavenumber $k_0 = 24\pi$, which implies that the radius of the cylinder is such that $a = 2\lambda_0$. (λ_0 is just the free-space wavelength.)

The width of the PML is chosen to be $w_{PML} = 1/12$, or $w_{PML} = \lambda_0$. The incident field $\mathbf{E}^{inc}$ is given by

$$\mathbf{E}^{inc}(\mathbf{x}) = e^{-jk_0x} \hat{\mathbf{y}}, \quad \mathbf{x} \in \bar{\Omega}, \quad (5.9.a)$$

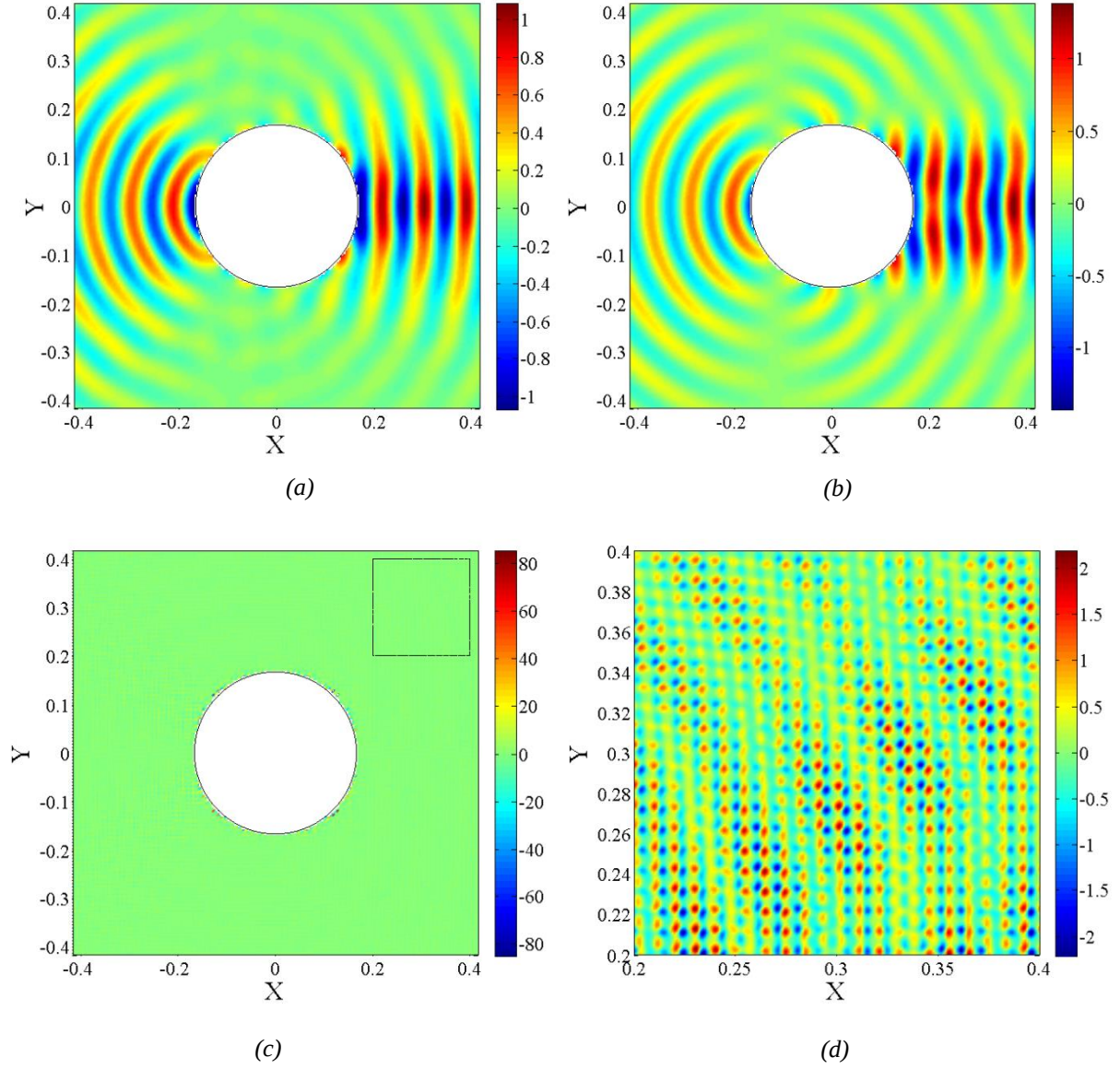


Fig. 5.7. The first two figures illustrate the y -component of the scattered electric field, in volts/meter. The region of the domain within the PML layer is not considered. (a) Numerical solution (real part). (b) Analytical solution (real part). (c) The numerical divergence. (d) The numerical divergence, calculated within the square in Fig. 5.7.c.

(in volts/meter) which allows the lifting function $\mathbf{u}_h^g$ to be easily calculated according to the procedure outlined in the Section 4.3.4.

Figure 5.6.a shows the whole computational domain, and Fig. 5.6.b shows a portion of the domain with some nodes and their corresponding elemental directions. Figure 5.6.c shows a portion of the nodal distribution, and illustrates which nodes fall within the regular and non-regular portions according to the discussion from Section 5.1.2.

The problem is discretized with 9192 nodes, originating a total of 61656 DoF's. The final linear system can be solved by a direct method. The results are in Fig. 5.7, which

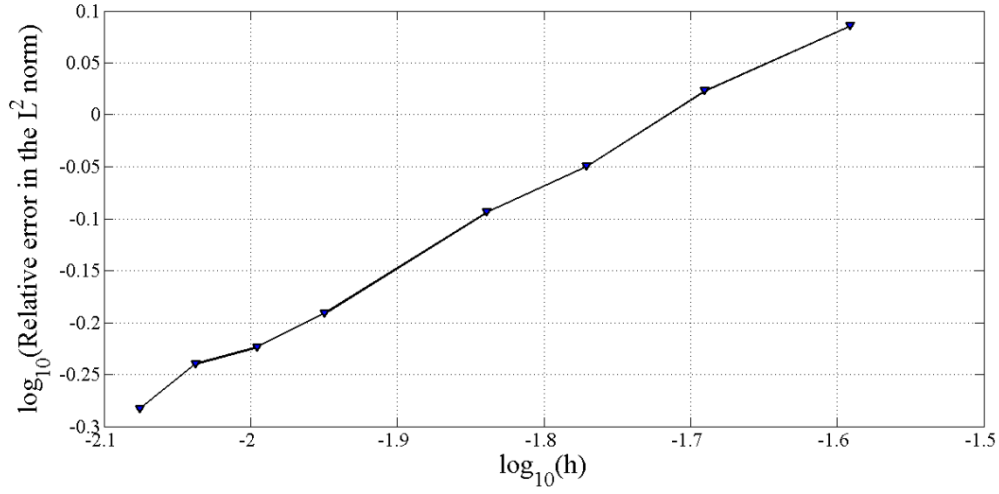


Fig. 5.7. (Cont.) (e) The relative error between the numerical and analytical solutions as a function of the discretization length h .

shows that the meshfree and the analytical solutions are in good agreement with each other. The divergence is imposed as zero in the weak sense, according to 2.156.b. Consequently, it means that the divergence is not zero pointwise (as revealed by Fig. 5.7.d), but that the integral of $\nabla \cdot \mathbf{E}^s$ multiplied by any test function q_h from $\mathbb{P}^h(\Omega)$ is zero.

We can measure the error between the numerical and analytical solutions in the portion of the computational domain Ω excluding the PML region (which we can denote by Ω_{PML}). If we express this subset of Ω as $\Omega \setminus \Omega_{PML}$, then we evaluate

$$\epsilon(h) = \frac{\|\mathbf{E} - \mathbf{E}_h\|_{L^2(\Omega \setminus \Omega_{PML})^3}}{\|\mathbf{E}\|_{L^2(\Omega \setminus \Omega_{PML})^3}}, \quad (5.9.b)$$

where $\|\cdot\|_{L^2(\Omega \setminus \Omega_{PML})^3}$ indicates that the integrations are carried out at $\Omega \setminus \Omega_{PML}$.

The result is in Fig. 5.7.e. A linear regression shows that the relation between ϵ and h is approximately:

$$\epsilon(h) = Ch^{0.7641}. \quad (5.9.c)$$

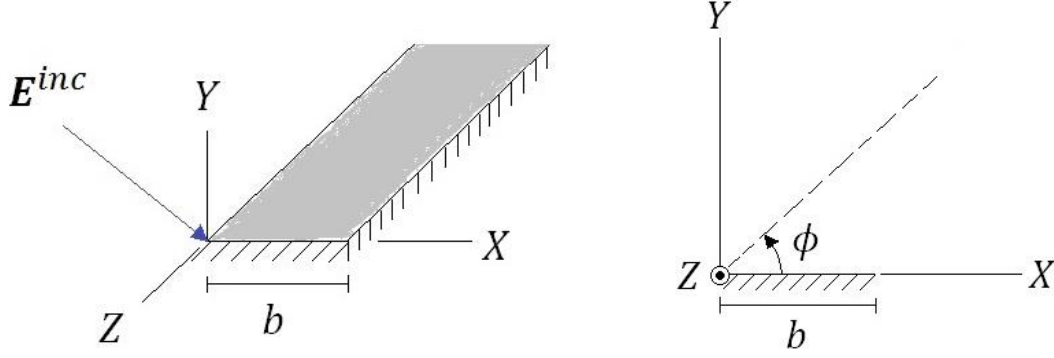


Fig. 5.8. *Left*: The PEC strip extends to infinity along the z -direction. *Right*: In the XY plane, we can set up the standard polar system of coordinates. In this way, the incidence and observation angles ϕ_i and ϕ_s are measured as indicated in the figure.

5.4.3 Scattering of a TE^z plane wave by a conducting strip

The problem concerning the scattering of a TE^z polarized plane wave by a conducting strip is examined next. The geometry of the problem is illustrated in Fig. 5.8.

The computational domain is a square region $\Omega = (-0.35, 0.65) \times (-0.5, 0.5)$ (in meters), in which we make a ‘hole’ of zero thickness and width b equal to 0.3. This ‘hole’ is indeed the cross-section of the strip, and occupies the interval $0 \leq x \leq 0.3$, $y = 0$.

The wavenumber is $k_0 = 40\pi$; in this way, $b = 6\lambda_0$. We choose the width of the PML layer to be $w_{PML} = 0.2$, which implies that $w_{PML} = 4\lambda_0$.

The incident field is a TE^z polarized plane; the associated magnetic field $\mathbf{H}^{inc}$ has a z -component given by

$$\mathbf{H}^{inc}(\mathbf{x}) = H_0 e^{-j\mathbf{k} \cdot \mathbf{x}} \hat{\mathbf{z}}, \quad \mathbf{x} \in \bar{\Omega} \quad (5.10.a)$$

in which H_0 is the amplitude of the incident field (in amperes/meter). The position vector $\mathbf{x}$ and the wavevector $\mathbf{k}$ are expressed as

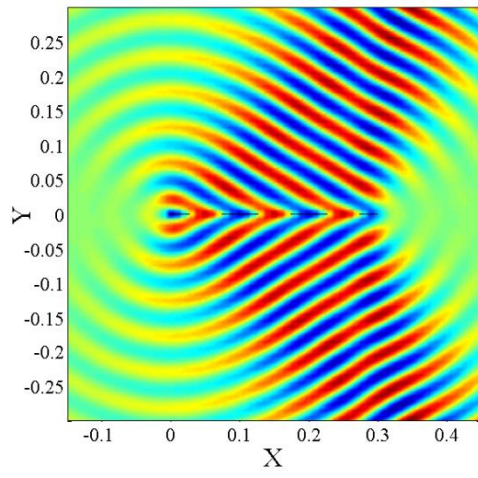
$$\mathbf{x} = [x, y]^T \quad (5.10.b)$$

$$\mathbf{k} = k_0 \hat{\mathbf{k}} = k_0 [k_x, k_y]^T, \quad (5.10.c)$$

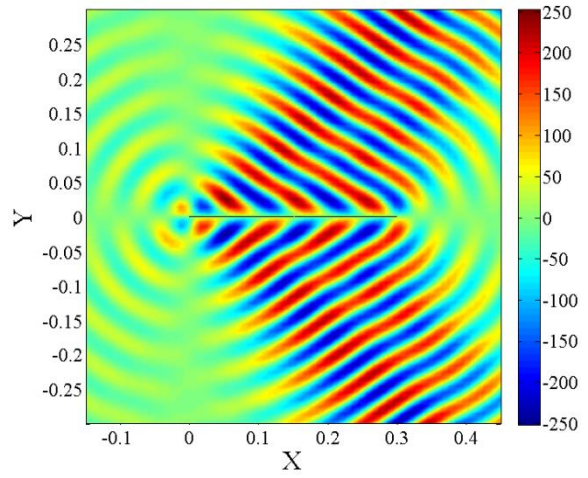
where $\hat{\mathbf{k}} = [k_x, k_y]^T$ is a unit vector pointing in the direction towards which the plane wave propagates. According to Fig. 5.8, it is given by

$$\hat{\mathbf{k}} = \begin{bmatrix} -\cos \phi_i \\ -\sin \phi_i \end{bmatrix}, \quad (5.10.d)$$

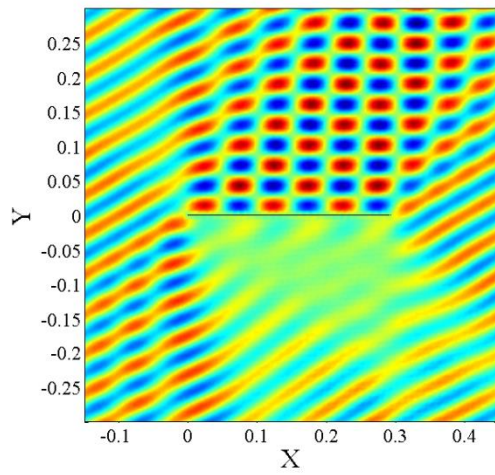
which allow us to ultimately rewrite (5.10.a) as



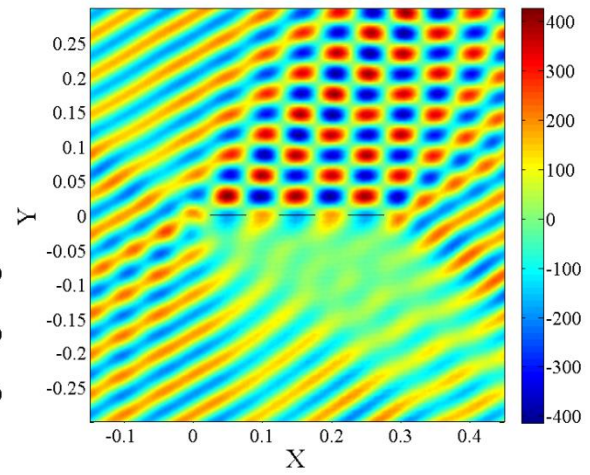
(a)



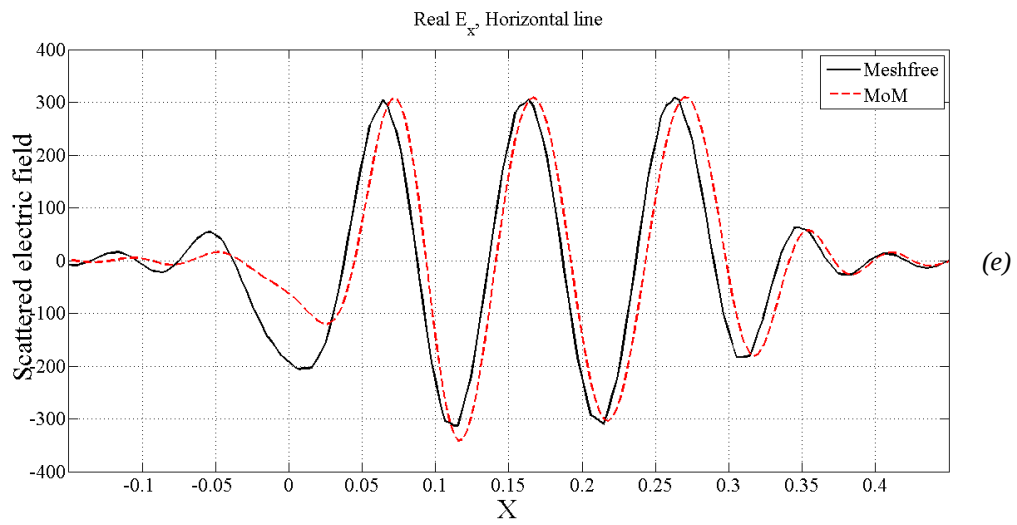
(b)



(c)



(d)



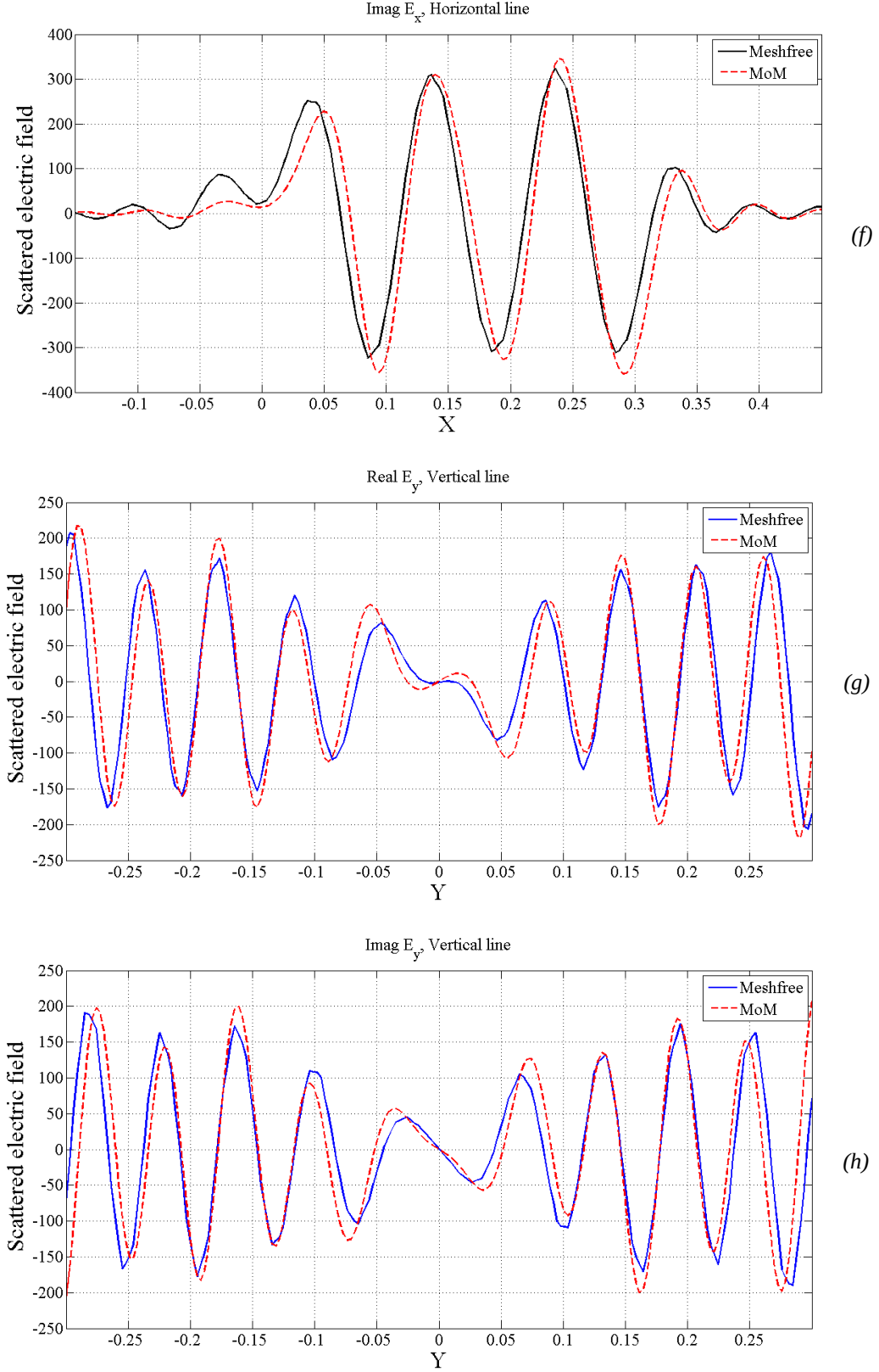


Fig. 5.9. The real part of scattered electric field $\mathbf{E}^s$, in volts/meter. (a) The x -component. (b) The y -component. The next two figures deal with the real part of the total field $\mathbf{E} = \mathbf{E}^s + \mathbf{E}^{inc}$. (c) The x -component. Observe how it is zero along the strip, in accordance with the boundary condition governing the tangential component of $\mathbf{E}$ at the surface of a PEC (i.e., $\hat{\mathbf{n}} \times \mathbf{E} = \mathbf{0}$). (d) The y -component. The last two figures also illustrate the shadow region behind the strip (a region not illuminated by the incident wave). Figures (e), (f), (g) and (h) bring a comparison between the meshfree and the MoM solutions.

$$\mathbf{H}^{inc}(\mathbf{x}) = H_0 e^{jk_0(x \cos \phi_i + y \sin \phi_i)} \hat{\mathbf{z}}. \quad (5.10.e)$$

We are interested in the incident electric field, which can be recovered from (5.10.e) via Ampère's law in free-space:

$$\nabla \times \mathbf{H}^{inc} = j\omega\epsilon_0 \mathbf{E}^{inc}. \quad (5.10.f)$$

The result is

$$\mathbf{E}^{inc}(\mathbf{x}) = \eta_0 H_0 (\sin \phi_i \hat{\mathbf{x}} - \cos \phi_i \hat{\mathbf{y}}) e^{jk_0(x \cos \phi_i + y \sin \phi_i)}, \quad \mathbf{x} \in \bar{\Omega}, \quad (5.10.g)$$

where $\eta_0 = \sqrt{\mu_0/\epsilon_0} \cong 377$ ohms is the *vacuum impedance*.

In this geometry, the PEC surface Γ_1 is just the interval $0 \leq x \leq 0.3, y = 0$. The angle of incidence ϕ_i is

$$\phi_i = \pi - \pi/3, \quad (5.10.h)$$

and $H_0 = 1$. Our discretization takes 10201 nodes, which yields 68917 DoF's. The resulting linear system is solved by a direct method, and the results are in Fig. 5.9.

In order to find out if the results are accurate or not, we compare the meshfree solutions with those provided by the method of moments (MoM). The current density on the surface of the strip is calculated via the two-dimensional electric field integral equation (EFIE), which is discretized with 250 piecewise constant basis functions and 250 Dirac delta weighting functions (point matching). After the current is found, the scattered field near the strip can be calculated by suitable radiation integrals [Balanis, 1989].

The meshfree and MoM solutions are compared along two lines in the near-field region. The first is a horizontal line defined by

$$L_1: -0.15 \leq x \leq 0.45, \quad y = 0.05, \quad (5.10.i)$$

Some results are in Fig. 5.9.e (real part of E_x^S) and in Fig. 5.9.f (imaginary part of E_x^S). The second line is vertical, and defined by

$$L_2: x = 0.35, \quad -0.3 \leq y \leq 0.3. \quad (5.10.j)$$

Figures 5.9.g and 5.9.h bring the real and imaginary parts of E_y^S , respectively. From the comparison between the meshfree and MoM solutions, it is clear that both methods provide similar results to the strip problem.

5.4.4 The spherical cavity

We now turn to three-dimensional problems. Let it be a spherical domain Ω , limited by a PEC surface Γ . The radius of the sphere is simply $a = 1$. Our goal is to find the eigenvalues and eigenfunctions associated with the original problem

Find $(\mathbf{E}, k_0^2)$ such that

$$\nabla \times \nabla \times \mathbf{E} - k_0^2 \mathbf{E} = \mathbf{0}, \quad \mathbf{x} \in \Omega \quad (5.11.a)$$

$$\nabla \cdot \mathbf{E} = 0, \quad \mathbf{x} \in \Omega \quad (5.11.b)$$

$$\hat{\mathbf{n}} \times \mathbf{E} = \mathbf{0}, \quad \mathbf{x} \in \Gamma. \quad (5.11.c)$$

According to the reasoning from Chapter 1, this problem becomes

Find $(\mathbf{E}, p, k_0^2)$ such that

$$\nabla^2 \mathbf{E} + k_0^2 \mathbf{E} + \nabla p = \mathbf{0}, \quad \text{in } \Omega \quad (5.12.a)$$

$$\nabla \cdot \mathbf{E} = 0, \quad \text{in } \Omega \quad (5.12.b)$$

$$\hat{\mathbf{n}} \times \mathbf{E} = \mathbf{0}, \quad \mathbf{x} \in \Gamma, \quad (5.12.c)$$

i.e., the double curl has been substituted by the vector Laplacian and the pseudopressure p has been included in order to couple equations (5.11.d) and (5.11.e). When it comes to the finite-dimensional subspaces, the right choices for $\mathbf{E}_h$ and p_h are $\mathbb{V}_\tau^h(\Omega)$ and $\mathbb{P}^h(\Omega)$, respectively.

These finite-dimensional subspaces are the same as those from the scattering problem; the objective of this example is to verify if the modeling of three-dimensional curved geometries via the elemental directions yield accurate results. In weak form, the system (5.12) becomes

Find $(\mathbf{E}_h, p_h, k_0^2) \in \mathbb{V}_\tau^h(\Omega) \times \mathbb{P}^h(\Omega) \times \mathbb{R}^+$ such that

$$\int_{\Omega} \nabla \mathbf{E}_h : \nabla \mathbf{v}_h^* - \int_{\Omega} k_0^2 \mathbf{E}_h \cdot \mathbf{v}_h^* - \int_{\Omega} p_h \nabla \cdot \mathbf{v}_h^* = 0, \quad \forall \mathbf{v}_h \in \mathbb{V}_\tau^h(\Omega) \quad (5.13.a)$$

$$- \int_{\Omega} q_h^* \nabla \cdot \mathbf{E}_h = 0, \quad \forall q_h \in \mathbb{P}^h(\Omega). \quad (5.13.b)$$

The system above is an *eigenvalue problem in mixed form*, since it seeks to approximate two unknowns at once, $\mathbf{E}_h$ and p_h . The associated eigenvalues are the k_0^2 . In order to put (5.13) into a standard form, it can be rewritten as

Find $(\mathbf{E}_h, p_h, k_0^2) \in \mathbb{V}_\tau^h(\Omega) \times \mathbb{P}^h(\Omega) \times \mathbb{R}^+$ such that

$$\int_{\Omega} \nabla \mathbf{E}_h : \nabla \mathbf{v}_h^* - \int_{\Omega} p_h \nabla \cdot \mathbf{v}_h^* = k_0^2 \int_{\Omega} \mathbf{E}_h \cdot \mathbf{v}_h^*, \quad \forall \mathbf{v}_h \in \mathbb{V}_\tau^h(\Omega) \quad (5.14.a)$$

$$- \int_{\Omega} q_h^* \nabla \cdot \mathbf{E}_h = 0, \quad \forall q_h \in \mathbb{P}^h(\Omega). \quad (5.14.b)$$

The theory behind eigenvalue problems in mixed form is beyond the scope of this thesis [Boffi, 2010], [Boffi *et al.*, 2013]. The only detail that is relevant to us here is that the system (5.14) is well-posed if it obeys the same the inf-sup condition as that in (5.6.a). Since the pair of spaces $\mathbb{V}_\tau^h(\Omega)$ and $\mathbb{P}^h(\Omega)$ constructed out of the Z^1/Z^0 pair passes the test, we are justified in making this choice.

Another important observation is that, as the complex-valued components of the PML tensor are absent in (5.14), the eigenfunctions $\mathbf{E}_h$ are going to be real. In this way, after the discretization process (which is carefully studied in Section 4.3), we get a linear system of the form:

Find $(\bar{\mathbf{e}}, \bar{\mathbf{p}}) \in \mathbb{R}^Q \times \mathbb{R}^M$ such that

$$\begin{bmatrix} \bar{\mathbf{A}} & \bar{\mathbf{B}}^T \\ \bar{\mathbf{B}} & \mathbf{0} \end{bmatrix} \begin{bmatrix} \bar{\mathbf{e}} \\ \bar{\mathbf{p}} \end{bmatrix} = \begin{bmatrix} \mathbf{0} \\ \mathbf{0} \end{bmatrix} \quad (5.15.a)$$

From the form assumed by the entries in the submatrix $\bar{\mathbf{A}}$ in (4.22.g), (4.23.d) and (4.24.d), it can be observed that it is constituted by two parts,

$$\bar{\mathbf{A}} = \bar{\mathbf{A}}_s - k_0^2 \bar{\mathbf{A}}_m, \quad (5.15.b)$$

where $\bar{\mathbf{A}}_s$ and $\bar{\mathbf{A}}_m$ are sometimes referred to as the *stiffness* and *mass* matrices, respectively. Since k_0^2 is an eigenvalue (and therefore unknown), the system (5.15.a) should be rewritten as

Find $(\bar{\mathbf{e}}, \bar{\mathbf{p}}, k_0^2) \in \mathbb{R}^Q \times \mathbb{R}^M \times \mathbb{R}^+$ such that

$$\begin{bmatrix} \bar{\mathbf{A}}_s & \bar{\mathbf{B}}^T \\ \bar{\mathbf{B}} & \mathbf{0} \end{bmatrix} \begin{bmatrix} \bar{\mathbf{e}} \\ \bar{\mathbf{p}} \end{bmatrix} = k_0^2 \begin{bmatrix} \bar{\mathbf{A}}_m & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} \begin{bmatrix} \bar{\mathbf{e}} \\ \bar{\mathbf{p}} \end{bmatrix} \quad (5.15.c)$$

which is nothing else than a generalized eigenvalue problem. After the vector of coefficients $\bar{\mathbf{e}}$ has been determined, the corresponding eigenfunctions $\mathbf{E}_h$ are found through (4.14.f), which, after it has been worked out, becomes (4.21.l).

Figure 5.10 shows some nodes in a portion of the spherical global boundary Γ , together with the elemental directions. The first eigenfunctions agree with the corresponding analytical solutions, as will be illustrated by Figs. 5.11, 5.12, 5.13, 5.14 and 5.15. These analytical solutions are expressed in spherical coordinates as triple products involving a certain class of spherical Bessel functions, also known as Schelkunoff functions (which

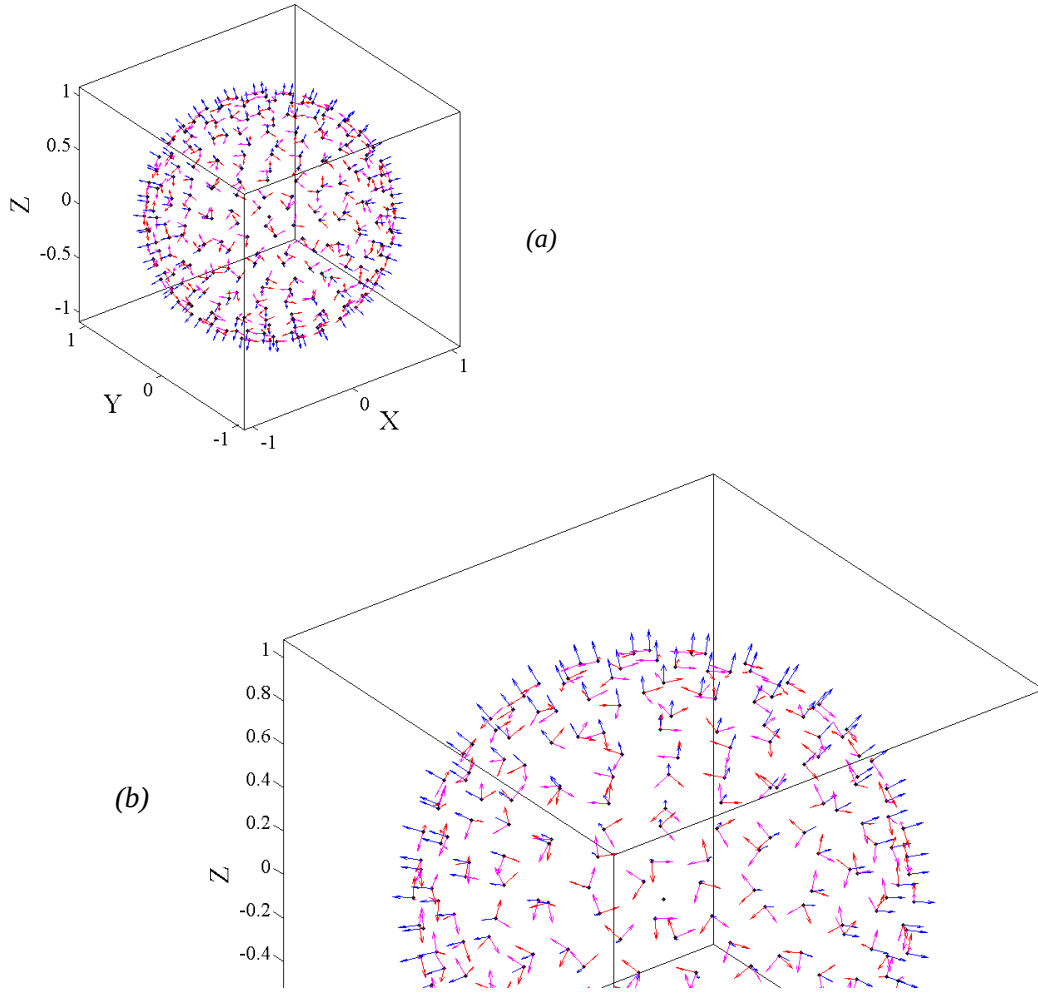


Fig. 5.10. (a) Nodes along the spherical surface, together with the elemental directions. (b) A zoom is applied to a portion of the surface in order to clarify the idea.

govern the dependence on the radius r), Legendre polynomials (which govern the dependence on the polar angle θ), and trigonometric terms (which govern the dependence on the azimuthal angle ϕ) [Balanis, 1989].

Each eigenvalue is determined by two indices: n and p (related to the p -th zero of the Schelkunoff function of order n for the TE^r modes, and to the p -th zero of the derivative of the Schelkunoff function of order n for the TM^r modes).

There are many modes associated to the same eigenvalue, known as the *degenerate modes*. Given an eigenvalue identified by n and p , the degenerate modes can be identified as follows: First, they are ascribed an index m such that $m = 0, 1, 2, \dots, n$. Second, if $m \neq 0$, then the mode displays either *even* symmetry or *odd* symmetry. The mode is said to be *even* if the dependence on the azimuthal angle ϕ is described by cosines (i.e., by terms such as $\cos m\phi$). It is said to be *odd* if the dependence is described by sines (i.e., by terms such as $\sin m\phi$).

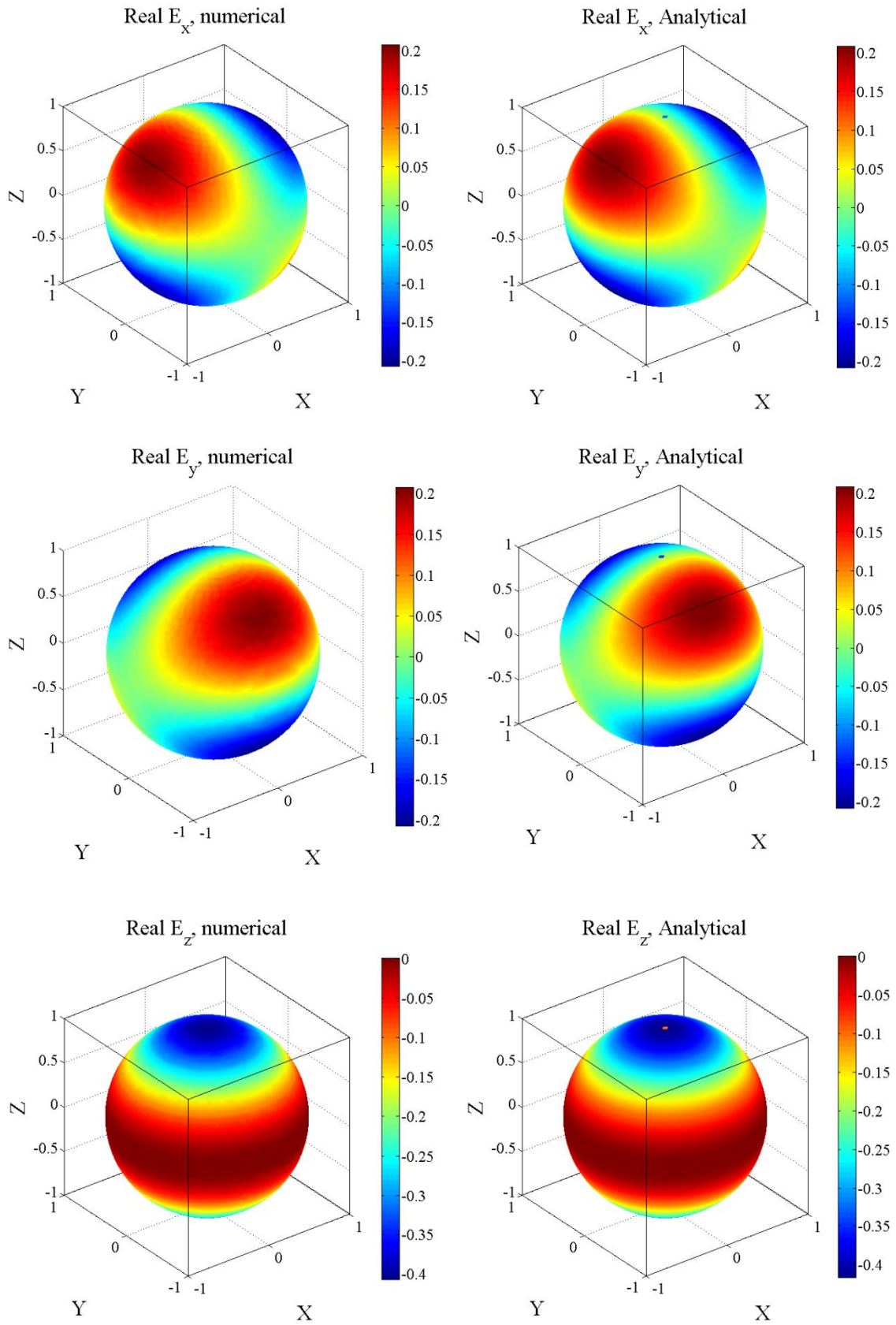


Fig. 5.11. The TM^r mode $\{1,1,0,-\}$.

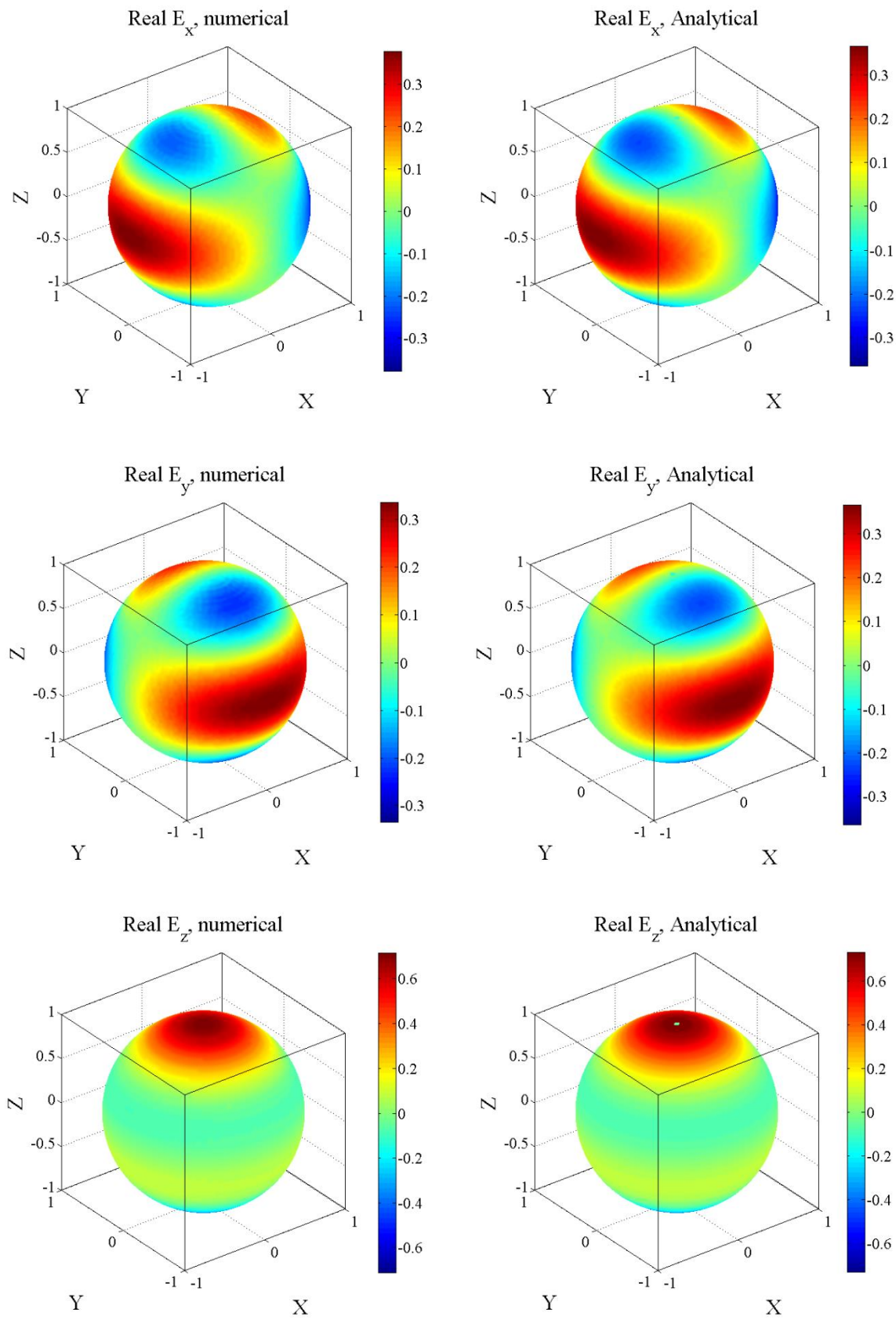


Fig. 5.12. The TM^r mode $\{2,1,0,-\}$.

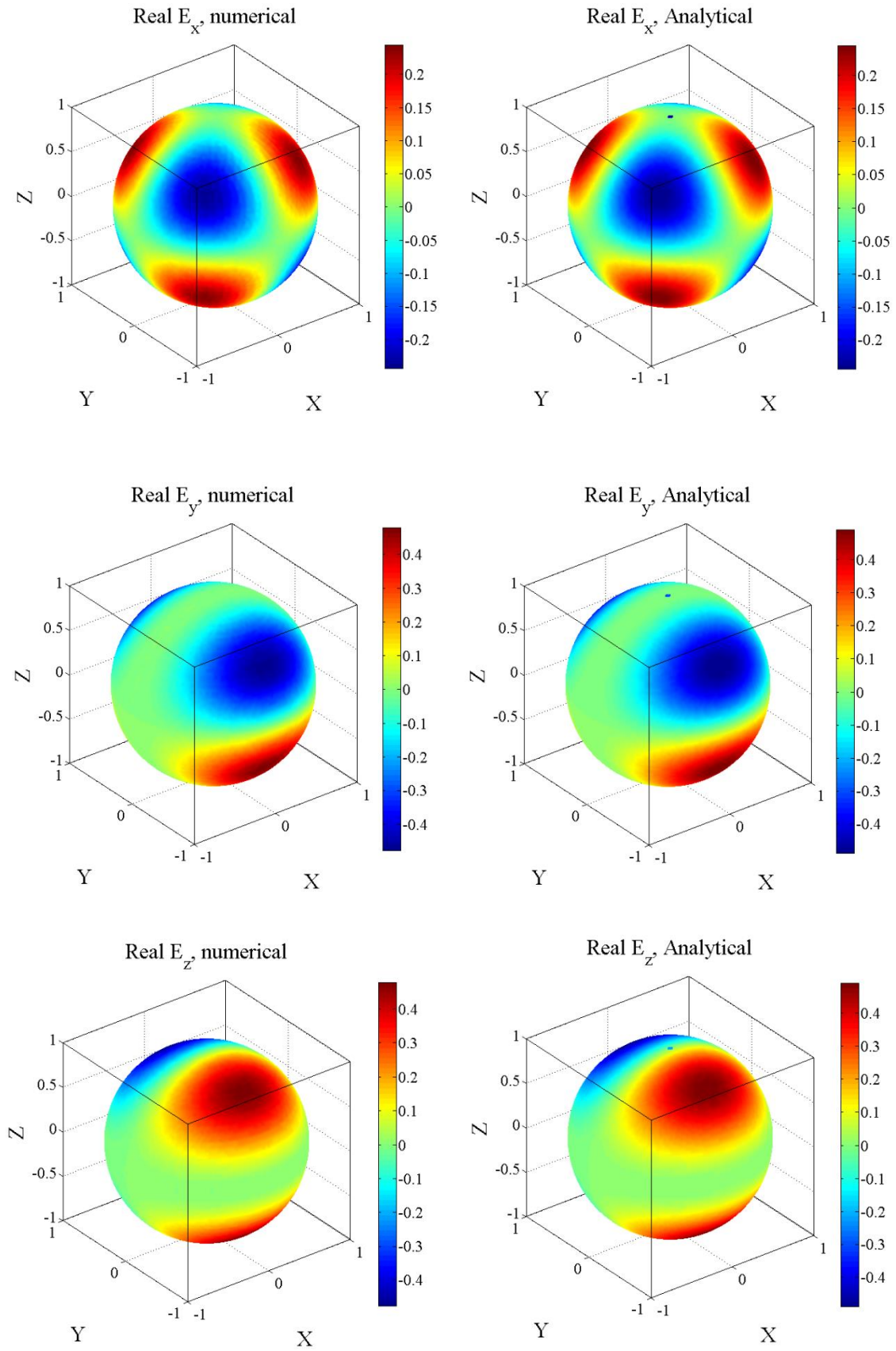


Fig. 5.13. The TM^r mode $\{2,1,1,odd\}$.

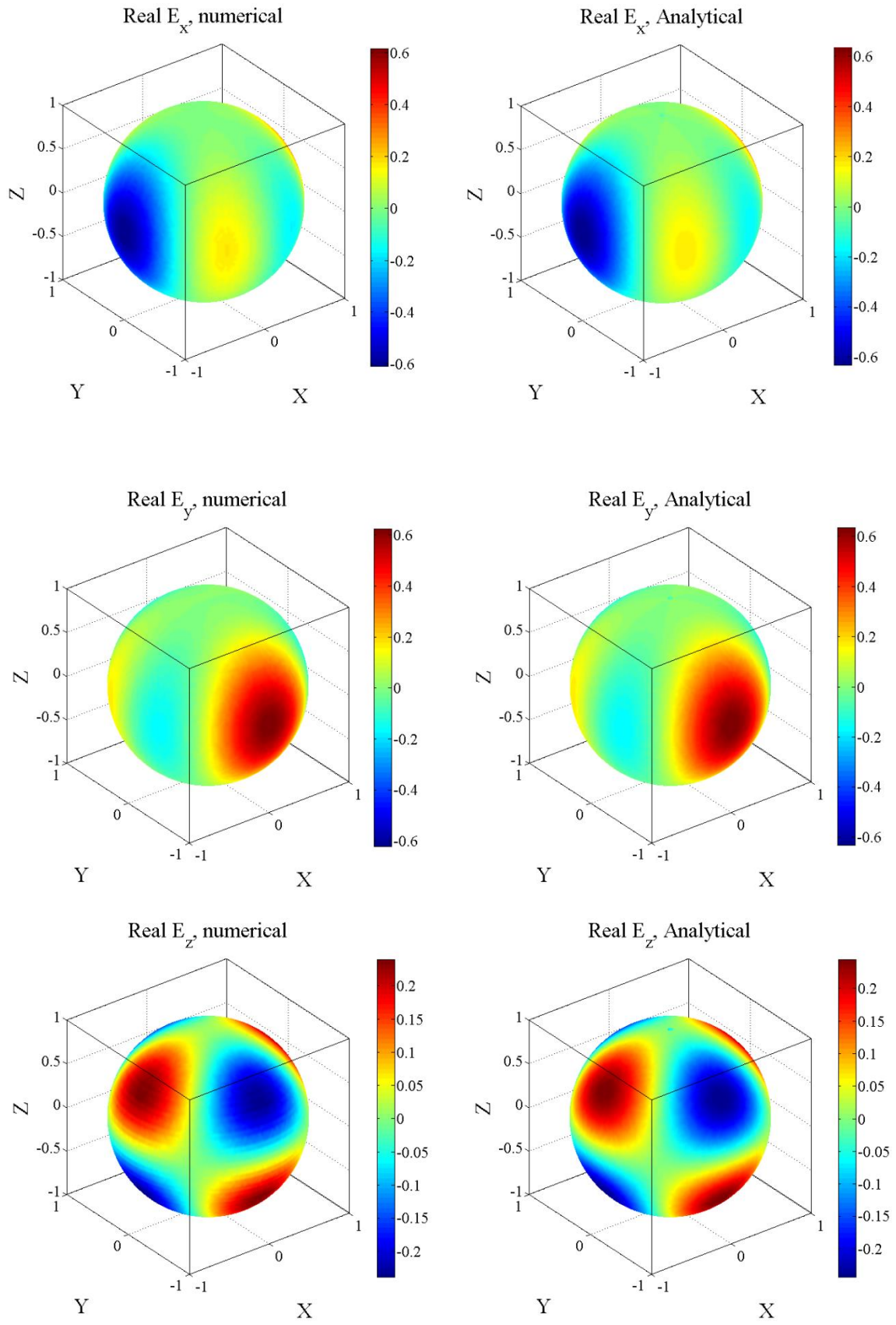


Fig. 5.14. The TM^r mode $\{2,1,2,even\}$.

In this way, a given mode is uniquely identified through a proper selection of 4 indices: n , p , m , and its symmetry (*even/odd*).

As stated earlier, our objective here is to find out if the meshfree spaces based on the elemental directions provide accurate solutions when applied to three-dimensional curvilinear geometries.

According to [Balanis, 1989] and [Harrington, 2001], if the eigenvalues are arranged in increasing order, the first two eigenvalues are associated with TM^r modes, whereas the third is related to TE^r modes. Since we are not interested in the higher-order modes, we concentrate just on the TM^r modes associated with the first two eigenvalues.

Let the four mode identifiers be assembled in a 4-tuple, as

$$\{n, p, m, s\}, \quad (5.15. d)$$

where s means ‘symmetry’. We study four modes; they are: $\{1,1,0,-\}$, $\{2,1,0,-\}$, $\{2,1,1,odd\}$, $\{2,1,2,even\}$. The field components are converted from the spherical to the Cartesian system, and the comparison between the numerical and analytical solutions is shown in Figs. 5.11, 5.12, 5.13 and 5.14, respectively. A total of 9273 nodes has been used in the discretization process, which leads to 104029 DoF’s.

The Figs. 5.11, 5.12, 5.13 and 5.14 display the field components on the surface of the sphere. It is true that the numerical and analytical solutions also agree at the interior volume of the sphere. In Fig. 5.15, we again consider the mode $\{2,1,0,-\}$, but now we display the solution along the YZ plane (i.e., we take the sphere and cut it open at the plane $x = 0$). At this plane, the mode $\{2,1,0,-\}$ has no x -component. So the computed y and z components are compared with their analytical counterparts.

When we compare the numerical and analytical solutions in Fig. 5.11, 5.12, 5.13, 5.14 and 5.15, it becomes evident that our meshfree spaces $\mathbb{V}_\tau^h(\Omega)$ based on elemental directions perform well when dealing with curved geometries. This is evidenced by Fig. 5.16, which measures the relative error between the numerical and analytical solutions corresponding to the first TM^r mode $\{1,1,0,-\}$ (that of Fig. 5.11):

$$\epsilon^{\{1,1,0,-\}}(h) = \frac{\|\mathbf{E}^{\{1,1,0,-\}} - \mathbf{E}_h^{\{1,1,0,-\}}\|_{L^2(\Omega)^3}}{\|\mathbf{E}^{\{1,1,0,-\}}\|_{L^2(\Omega)^3}}, \quad (5.15. e)$$

A linear regression applied to the curve in Fig. 5.16 shows that the relation between $\epsilon^{\{1,1,0,-\}}$ and h is approximately given by (where C is a positive constant):

$$\epsilon^{\{1,1,0,-\}}(h) = Ch^{1.138}, \quad (5.15. f)$$

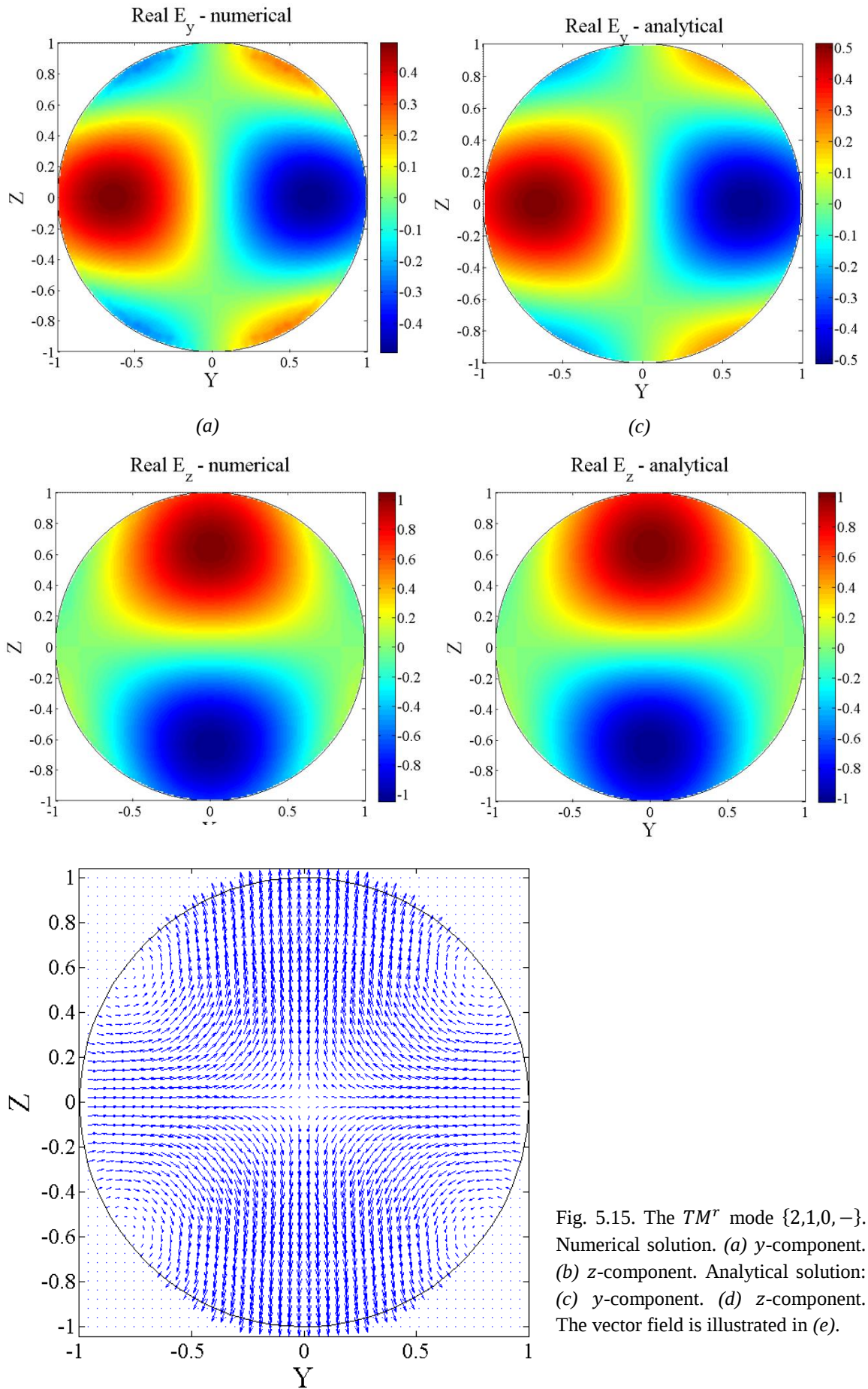


Fig. 5.15. The TM^r mode $\{2,1,0,-\}$. Numerical solution. (a) y -component. (b) z -component. Analytical solution: (c) y -component. (d) z -component. The vector field is illustrated in (e).

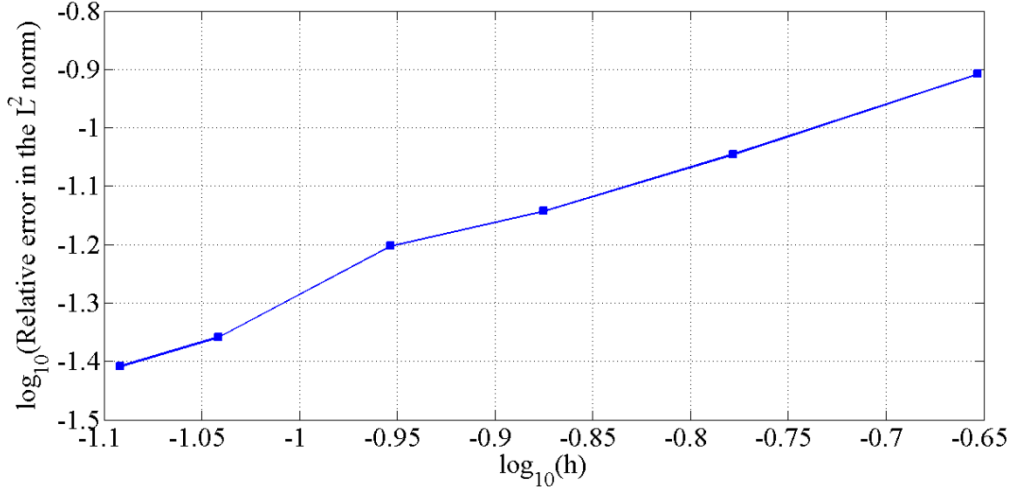


Fig. 5.16. The relative error for the TM^r mode $\{1,1,0,-\}$ as a function of the discretization length h .

5.4.5 Scattering by PEC plates

The great objective of this thesis is to develop a meshfree method able to calculate scattered fields by PEC targets in three dimensions. This area of study has a wide range of applications, particularly in the military (for example, in determining the radar cross sections of flying objects) [Kwon *et al.*, 2001]. However, realistic targets such as missiles and airplanes require a very precise description, which usually leads to problems with a huge number of DoF's. In these cases, it is likely that the resulting problem is solvable only with the help of a supercomputer.

At any rate, solving these large problems is not our goal. We are concerned here with providing a totally meshfree solution method able to deal with this category of problem; our purpose will be fulfilled if we show that we can solve 'smaller' problems in this same category. If the method proves successful, subsequent research can concentrate on the extension of the technique to larger problems.

We shall now study the three-dimensional scattering of plane waves by rectangular PEC plates. We think that this example is challenging enough to serve as a test to find out if the overall method we have been devising (which comprises the mixed formulation, the 'acoustic' PML, the reuse approach in the integration of the weak forms, the elemental directions and the preconditioning matrix) is able to solve this kind of problem. The geometry is illustrated in Fig. 5.17; the domain Ω is a box described by the intervals (in meters):

$$-0.5 \leq x \leq 0.5 \quad (5.16.a)$$

$$-0.5 \leq y \leq 0.5$$

$$-0.1\bar{6} \leq z \leq 0.1\bar{6}$$

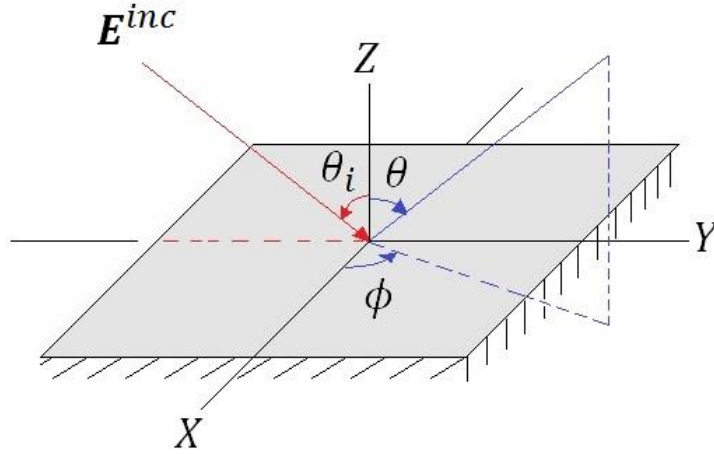


Fig. 5.17. We can set up the standard spherical system of coordinates. In this way, any direction can be identified by a pair of angles (θ, ϕ) . The incident field $\mathbf{E}^{inc}$ has the direction determined by θ_{inc} (as indicated in the figure) and by $\phi_{inc} = 3\pi/2$ (i.e., along the dashed red line). In what regards the dimensions, the plate is $b \times b$ (b in meters).

The PEC surface Γ_1 is just a zero-thickness square placed at the center of the domain:

$$-0.\bar{3} \leq x \leq 0.\bar{3} \quad (5.16.b)$$

$$-0.\bar{3} \leq y \leq 0.\bar{3}$$

$$z = 0$$

In (5.16.a), $0.1\bar{6} = 0.1666 \dots = 1/6$, and in (5.16.b), $0.\bar{3} = 0.333 \dots = 1/3$. The free-space wavenumber is given by $k_0 = 18\pi$, which implies that the width b of the plate is such that $b = 6\lambda_0$.

The width of the PML layer is $w_{PML} = 1/12$, or $w_{PML} = 0.75\lambda_0$.

In what regards the incident field, it is a plane wave whose wavevector is

$$\mathbf{k} = k_0 \hat{\mathbf{k}} = k_0 [k_x, k_y, k_z]^T, \quad (5.16.c)$$

where $\hat{\mathbf{k}} = [k_x, k_y, k_z]^T$ is a unit vector pointing in the direction towards which the plane wave propagates. According to Fig. 5.13,

$$\hat{\mathbf{k}} = -\hat{\mathbf{r}}, \quad (5.16.d)$$

i.e., $\hat{\mathbf{k}}$ is just the negative of the unit radial vector. It is known that the conversion from spherical to Cartesian coordinates is given by

$$\begin{bmatrix} k_x \\ k_y \\ k_z \end{bmatrix} = \begin{bmatrix} \sin \theta_i \cos \phi_i & \cos \theta_i \cos \phi_i & -\sin \phi_i \\ \sin \theta_i \sin \phi_i & \cos \theta_i \sin \phi_i & \cos \phi_i \\ \cos \theta_i & -\sin \theta_i & 0 \end{bmatrix} \begin{bmatrix} k_r \\ k_\theta \\ k_\phi \end{bmatrix}. \quad (5.16.e)$$

According to (5.16.d), $k_r = -1$, $k_\theta = 0$, and $k_\phi = 0$. In this way, the Cartesian components of $\hat{\mathbf{k}}$ become

$$\begin{bmatrix} k_x \\ k_y \\ k_z \end{bmatrix} = \begin{bmatrix} -\sin \theta_i \cos \phi_i \\ -\sin \theta_i \sin \phi_i \\ -\cos \theta_i \end{bmatrix} \quad (5.16.f)$$

The three Cartesian components of $\hat{\mathbf{k}}$ are completely determined by the pair of angles (θ_i, ϕ_i) .

In the sequel, we will consider two polarizations for the incident plane wave: The TE^x polarization, whose incident magnetic field is given by

$$\mathbf{H}^{inc}(\mathbf{x}) = H_0 e^{-j\mathbf{k} \cdot \mathbf{x}} \hat{\mathbf{x}}, \quad \mathbf{x} \in \bar{\Omega} \quad (5.16.g)$$

and the TM^x polarization, whose incident electric field is

$$\mathbf{E}^{inc}(\mathbf{x}) = E_0 e^{-j\mathbf{k} \cdot \mathbf{x}} \hat{\mathbf{x}}, \quad \mathbf{x} \in \bar{\Omega} \quad (5.16.h)$$

The position vector $\mathbf{x} = [x, y, z]^T$ and Ampère's law in free-space (5.10.f) allows us to determine the electric field associated to $\mathbf{H}^{inc}$ in (5.16.g):

$$\mathbf{E}^{inc}(\mathbf{x}) = \eta_0 H_0 (\cos \theta_i \hat{\mathbf{y}} - \sin \theta_i \sin \phi_i \hat{\mathbf{z}}) e^{-j\mathbf{k} \cdot \mathbf{x}}, \quad \mathbf{x} \in \bar{\Omega}. \quad (5.16.i)$$

So if we want to study the scattering of a TE^x wave, the incident field is given by (5.16.i). On the other hand, if the scattering of a TM^x wave is needed, then the incident field is that in (5.16.h).

The results for the TM^x and TE^x polarizations are in Figs. 5.18 and 5.19, respectively, where the Cartesian components of the scattered field are plotted on a surface surrounding the plate. The fields on this surface will later 'induce' equivalent currents, which by their turn will determine the far-field behavior. This will be duly explained in Section 5.4.6. The parameters of our simulations are in Table 5.2 below.

TABLE 5.2 –SIMULATION FACTS

Parameters	TM^x	TE^x
θ_i	$\pi/4$	$\pi/6$
ϕ_i	$3\pi/2$	$3\pi/2$
Field amplitude	$E_0 = 1$	$H_0 = 1$
Number of nodes	27 735	27 735
Number of DoF's	307 667	307 667
GMRES iterations	200	200
Relative residual	6.6×10^{-4}	1.66×10^{-4}

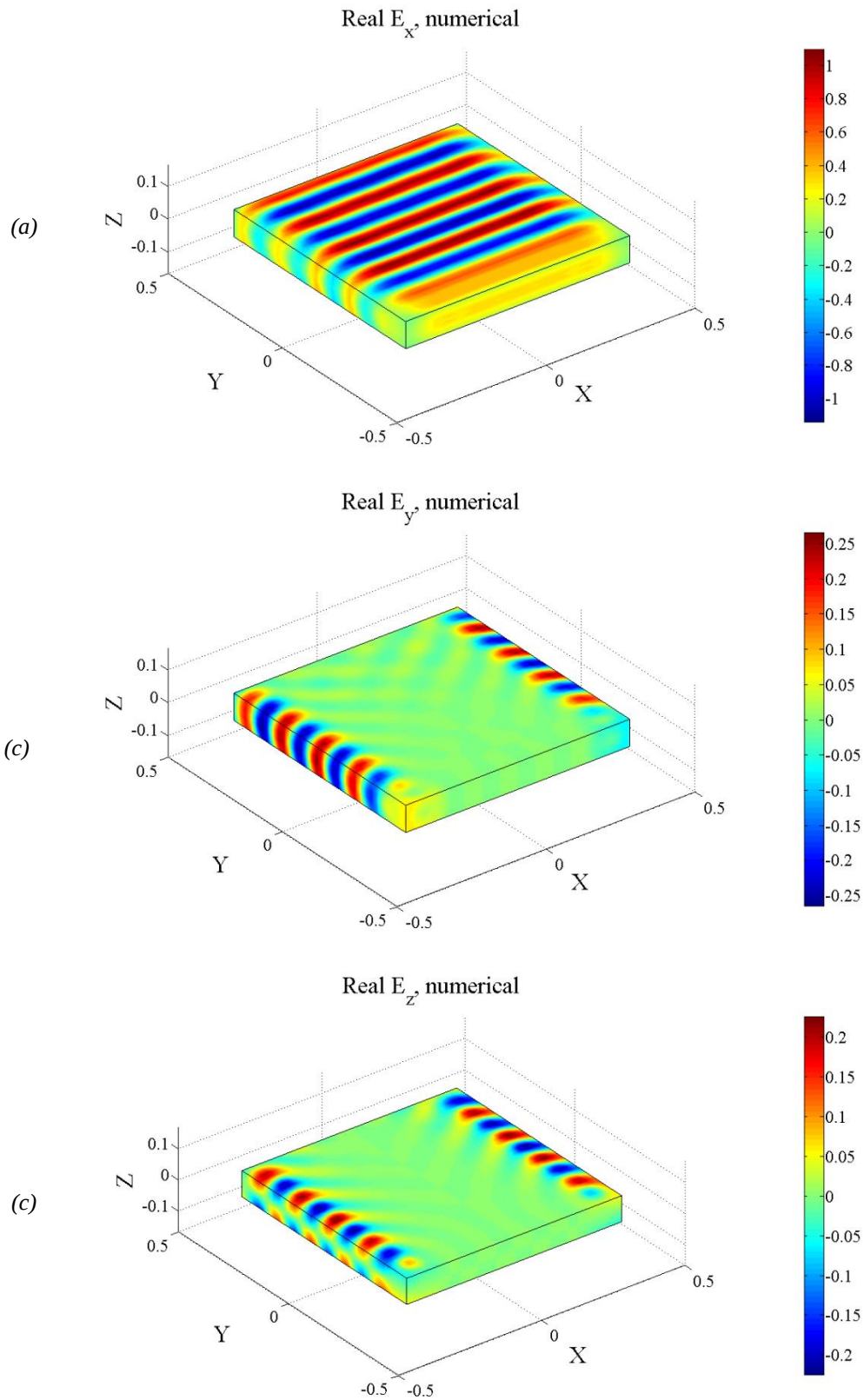


Fig. 5.18. Results for the scattering of a TM^x wave: The real part of the Cartesian components of the scattered electric field (in volts/meter) on a surface in free-space surrounding the scatterer (i.e., the PEC plate). (a) The x -component. (b) The y -component. (c) The z -component.

As it happened for the strip problem in Section 5.4.3, the plate problems also do not have analytical solutions. In order to discover if these results are meaningful or not, we need the notion of *radar cross section*, to be introduced next.

5.4.6 Radar cross sections

The true meaning of radar cross sections (RCS) is discussed in [Balanis, 1989]. For our purposes, we need only the mathematical definition. In three dimensions, given the observation angles (θ_s, ϕ_s) in spherical coordinates, the *radar cross-section* is defined by

$$\sigma_{3D}(\theta_s, \phi_s) := \lim_{r_s \rightarrow \infty} 4\pi r_s^2 \frac{|\mathbf{E}^s(r_s, \theta_s, \phi_s)|^2}{|\mathbf{E}^{inc}(r_s, \theta_s, \phi_s)|^2}, \quad (5.17.a)$$

where r_s is the observation radius r_s . (As $r_s \rightarrow \infty$, it is expected that r_s will somehow be cancelled at the right side of (5.17.a), so that σ_{3D} will ultimately depend just on the angles θ_s and ϕ_s .)

In two dimensions, the radar cross section is sometimes termed the *scattering width* (SW) [Balanis, 1989], [Peterson *et al.*, 1998]. Given the observation angle ϕ_s in polar coordinates, it is defined by

$$\sigma_{2D}(\phi_s) := \lim_{\rho_s \rightarrow \infty} 2\pi \rho_s \frac{|\mathbf{E}^s(\rho_s, \phi_s)|^2}{|\mathbf{E}^{inc}(\rho_s, \phi_s)|^2}. \quad (5.17.b)$$

The observation radius ρ_s is also expected to be cancelled at the right side of (5.17.b).

The unit of the RCS is just the unit for the area. It implies that in SI it is measured in square meters m^2 . It is usual to calculate the *normalized radar cross section*

$$\sigma_{3D}^n(\theta_s, \phi_s) = \frac{\sigma_{3D}(\theta_s, \phi_s)}{\lambda_0^2}, \quad (5.17.c)$$

i.e., the RCS (5.17.a) is divided by the free-space wavelength squared. In this way, σ_{3D}^n is dimensionless, which allows the magnitude of this quantity to be expressed in decibels:

$$\sigma_{3D}^n(\theta_s, \phi_s)|_{dB} = 10 \log_{10}(\sigma_{3D}^n(\theta_s, \phi_s)). \quad (5.17.d)$$

Analogously, the unit of the SW is just the unit for the length, which happens to be the meter in the SI. It is also usual to calculate the *normalized scattering width*

$$\sigma_{2D}^n(\phi_s) = \frac{\sigma_{2D}(\phi_s)}{\lambda_0}, \quad (5.17.e)$$

which is a dimensionless quantity. When expressed in decibels, it becomes

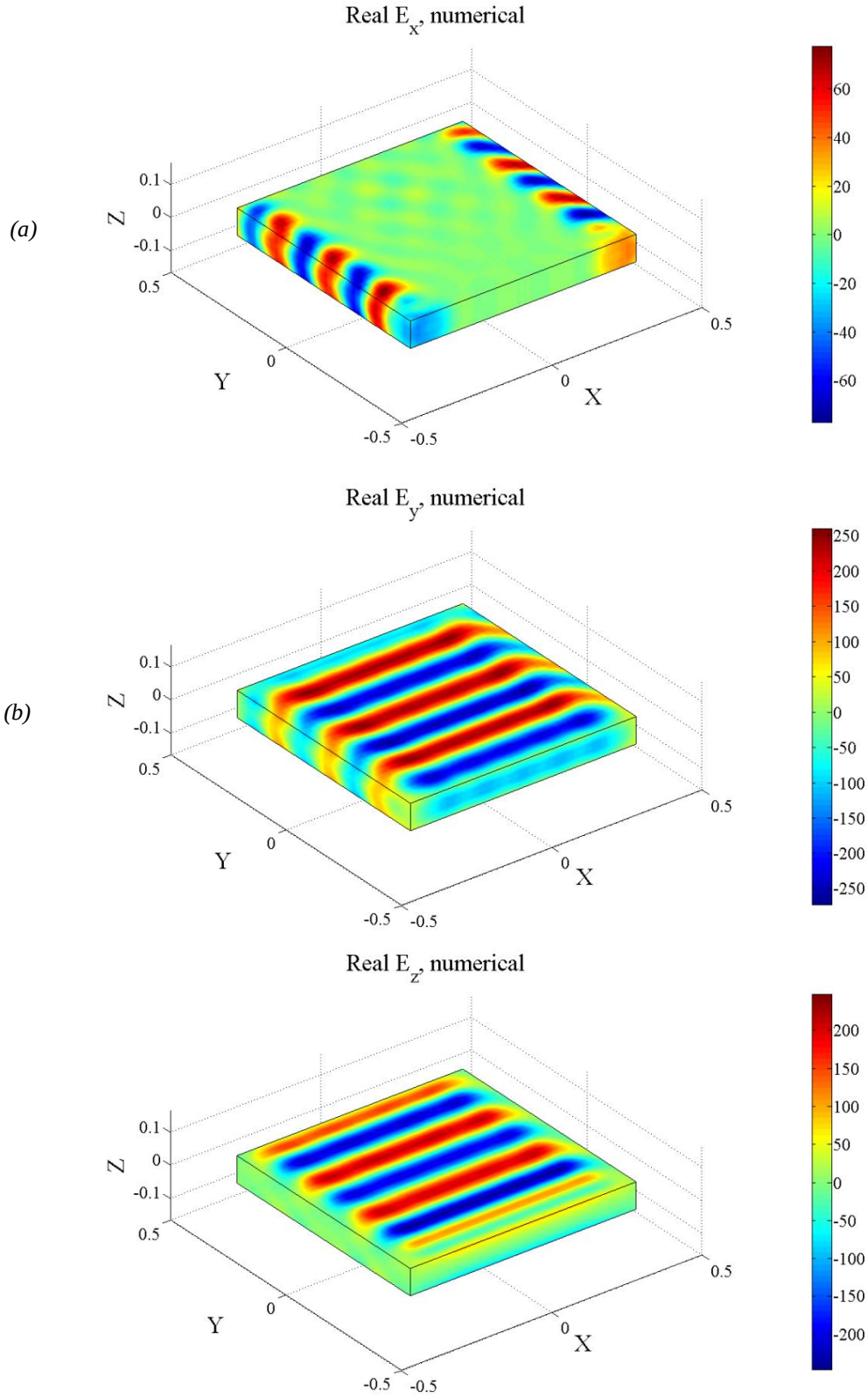


Fig. 5.19. Results for the scattering of a TE^x wave: The real part of the Cartesian components of the scattered electric field (in volts/meter) on a surface in free-space surrounding the scatterer (i.e., the PEC plate). (a) The x -component. (b) The y -component. (c) The z -component.

$$\sigma_{2D}^n(\phi_s)|_{dB} = 10 \log_{10}(\sigma_{2D}^n(\phi_s)). \quad (5.17.f)$$

The procedure for calculating either the RCS or the SW is extensively discussed in the literature. It relies basically on two results from electromagnetic theory: The surface equivalence principle and the far-field approximation [Peterson *et al.*, 1998], [Balanis, 1989]. We shall briefly outline the main steps.

5.4.6.1 Three dimensions

We first place an imaginary closed surface Σ surrounding the PEC scatterer. (Σ is sometimes termed the Huygens surface.) By this we mean that the scatterer surface Γ_1 is contained in the volume encircled by Σ . In this work, Σ is a box whose ‘size’ is larger than Γ_1 and smaller than Γ_o .

After we solve the scattering problem via our meshfree method, the scattered electric field $\mathbf{E}_h^s$ can be found at any point from the surface Σ . From the derivatives of $\mathbf{E}_h^s$, the components of the scattered magnetic field $\mathbf{H}_h^s$ can be calculated via Faraday’s law in free-space:

$$\nabla \times \mathbf{E}_h^s = -j\omega\mu_0\mathbf{H}_h^s. \quad (5.18.a)$$

The surface equivalence principle says that the scattered field at a point away from the scatterer can be determined by ‘equivalent currents’ defined over a closed surface around the scatterer, such as Σ . In a version of this principle called Love’s surface equivalence principle [Balanis, 1989], we consider the fields to be zero within the volume encircled by Σ . The standard boundary conditions tell us that there are equivalent currents flowing over Σ , given by

$$\mathbf{J}_{eq} = \hat{\mathbf{n}} \times \mathbf{H}_h^s \quad (5.18.b)$$

$$\mathbf{M}_{eq} = -\hat{\mathbf{n}} \times \mathbf{E}_h^s, \quad (5.18.c)$$

where $\mathbf{J}_{eq}$ is the electric current density and $\mathbf{J}_{eq}$ is the magnetic current density.

Let an observation point $\mathbf{x}_s \in \mathbb{R}^3$ be represented by its spherical coordinates (r_s, θ_s, ϕ_s) . Analogously, let any ‘source point’ $\mathbf{x}' \in \Sigma$ also be represented by its spherical coordinates (r', θ', ϕ') . The currents in (5.18.b) and (5.18.c) give rise to the magnetic and electric vector potentials $\mathbf{A}$ and $\mathbf{F}$:

$$\mathbf{A}(\mathbf{x}_s) = \frac{\mu_0}{4\pi} \oint_{\Sigma} \mathbf{J}_{eq}(\mathbf{x}') \frac{e^{-jk_0 R}}{R} \quad (5.18.d)$$

$$\mathbf{F}(\mathbf{x}_s) = \frac{\varepsilon_0}{4\pi} \oint_{\Sigma} \mathbf{M}_{eq}(\mathbf{x}') \frac{e^{-jk_0 R}}{R}, \quad (5.18.e)$$

where $R = |\mathbf{x}_s - \mathbf{x}'|$. If the observation point $\mathbf{x}_s$ is very far from the surface Γ_1 , then the *far-field approximation* can be employed. It says that, provided that $k_0 R \gg 1$, or more specifically, that [Balanis, 1989]

$$r_s \geq \frac{2D^2}{\lambda_0}, \quad (5.18.f)$$

where D is the diameter of the scatterer, then the following approximation is valid:

$$\frac{1}{R} \approx \frac{1}{r_s} \quad (5.18.g)$$

$$e^{-jk_0 R} \approx e^{-jk_0(r_s - r' \cos \psi)}. \quad (5.18.h)$$

The quantity ψ is the angle between the vectors $\mathbf{x}_s$ and $\mathbf{x}'$. In what regards the RCS, it is obvious that (5.18.f) holds, since $r_s \rightarrow \infty$ according to (5.17.a). So we are entitled to employ (5.18.g) and (5.18.h) in (5.18.d) and (5.18.f). We get:

$$\mathbf{A}(\mathbf{x}_s) = \frac{\mu_0}{4\pi} \frac{e^{-jk_0 r_s}}{r_s} \oint_{\Sigma} \mathbf{J}_{eq}(\mathbf{x}') e^{jk_0 r' \cos \psi} \quad (5.18.i)$$

$$\mathbf{F}(\mathbf{x}_s) = \frac{\varepsilon_0}{4\pi} \frac{e^{-jk_0 r_s}}{r_s} \oint_{\Sigma} \mathbf{M}_{eq}(\mathbf{x}') e^{jk_0 r' \cos \psi} \quad (5.18.j)$$

The scattered fields produced by the vector potentials $\mathbf{A}$ and $\mathbf{F}$ are given by

$$\mathbf{E}^s = -j\omega \left(\mathbf{A} + \frac{1}{k_0^2} \nabla \nabla \cdot \mathbf{A} \right) - \frac{1}{\varepsilon_0} \nabla \times \mathbf{F} \quad (5.18.k)$$

$$\mathbf{H}^s = -j\omega \left(\mathbf{F} + \frac{1}{k_0^2} \nabla \nabla \cdot \mathbf{F} \right) + \frac{1}{\mu_0} \nabla \times \mathbf{A} \quad (5.18.l)$$

It should be noticed that the scattered field $\mathbf{E}^s$ in (5.18.k) is not the finite-dimensional scattered electric field $\mathbf{E}_h^s$. The field $\mathbf{E}^s$ will be determined at positions very far from the scatterer, whereas $\mathbf{E}_h^s$ exists only near the scatterer. The near-field $\mathbf{E}_h^s$ ‘produces’ the equivalent currents in (5.18.b) and (5.18.c), which by their turn produce the field $\mathbf{E}^s$. In a sense, $\mathbf{E}^s$ is related to $\mathbf{E}_h^s$. This procedure is necessary because the nodal cloud cannot be extended to far distances (otherwise the total number of DoF’s in the problem would blow up).

When the operator $\nabla \nabla \cdot$ is applied to $\mathbf{A}$ and $\mathbf{F}$ in (5.18.k) and (5.18.l), one discovers that it gives rise to higher-order terms proportional to $1/r_s^2$, $1/r_s^3$, etc., and therefore can be neglected as far as far-field calculations are concerned. So the second term from (5.18.k) and (5.18.l) is discarded:

$$\mathbf{E}^s = -j\omega \mathbf{A} - \frac{1}{\varepsilon_0} \nabla \times \mathbf{F} \quad (5.18.m)$$

$$\mathbf{H}^s = -j\omega\mathbf{F} + \frac{1}{\mu_0}\nabla \times \mathbf{A} \quad (5.18.n)$$

There are some observations [Balanis, 1989] that can make the reasoning easier.

First: The electric field $\mathbf{E}^s$ is produced by a contribution from $\mathbf{A}$ and a contribution from $\mathbf{F}$:

$$\mathbf{E}^s = \mathbf{E}_A + \mathbf{E}_F, \quad (5.18.o)$$

$$\mathbf{E}_A = -j\omega\mathbf{A} \quad (5.18.p)$$

$$\mathbf{E}_F = -\frac{1}{\varepsilon_0}\nabla \times \mathbf{F}. \quad (5.18.q)$$

In the same way, the magnetic field $\mathbf{H}^s$ is made up from two contributions:

$$\mathbf{H}^s = \mathbf{H}_A + \mathbf{H}_F, \quad (5.18.r)$$

$$\mathbf{H}_A = \frac{1}{\mu_0}\nabla \times \mathbf{A} \quad (5.18.s)$$

$$\mathbf{H}_F = -j\omega\mathbf{F}. \quad (5.18.t)$$

Second: The radiated fields $\mathbf{E}_A$, $\mathbf{H}_A$, $\mathbf{E}_F$ and $\mathbf{H}_F$ have no radial components. In particular, (5.18.p) becomes:

$$(E_A)_r = 0 \quad (5.18.u)$$

$$(E_A)_\theta = -j\omega A_\theta \quad (5.18.v)$$

$$(E_A)_\phi = -j\omega A_\phi, \quad (5.18.w)$$

and (5.18.t) becomes:

$$(H_F)_r = 0 \quad (5.19.a)$$

$$(H_F)_\theta = -j\omega F_\theta \quad (5.19.b)$$

$$(H_F)_\phi = -j\omega F_\phi. \quad (5.19.c)$$

The (r, θ, ϕ) in the last six expressions actually refer to the observation point (r_s, θ_s, ϕ_s) .

Third: The fields $(\mathbf{E}_A, \mathbf{H}_A)$ and $(\mathbf{E}_F, \mathbf{H}_F)$ are TEM^r , which means that

$$\mathbf{E}_F = -\eta_0 \hat{\mathbf{r}}_s \times \mathbf{H}_F \quad (5.19.d)$$

$$\hat{\mathbf{r}}_s = \frac{\mathbf{x}_s}{\|\mathbf{x}_s\|} \quad (5.19.e)$$

The advantage of (5.19.d) over (5.18.q) is that the curl does not need to be calculated. From (5.19.a) – (5.19.c) and (5.19.d) we get

$$(E_F)_r = 0 \quad (5.19.f)$$

$$(E_F)_\theta = -j\omega\eta_0 F_\phi \quad (5.19.g)$$

$$(E_F)_\phi = j\omega\eta_0 F_\theta. \quad (5.19.h)$$

The combination of (5.18.u) – (5.18.w) and (5.19.f) – (5.19.h) allows us to write the components of $\mathbf{E}^s$ in (5.18.o) as

$$(E^s)_r = 0 \quad (5.19.i)$$

$$(E^s)_\theta = -j\omega(A_\theta + \eta_0 F_\phi) \quad (5.19.j)$$

$$(E^s)_\phi = -j\omega(A_\phi - \eta_0 F_\theta). \quad (5.19.k)$$

We can now get back to (5.18.i) and (5.18.j). If we introduce the terms

$$\mathbf{N}(\mathbf{x}_s) = \oint_{\Sigma} \mathbf{J}_{eq}(\mathbf{x}') e^{jk_0 r' \cos \psi} \quad (5.20.a)$$

$$\mathbf{L}(\mathbf{x}_s) = \oint_{\Sigma} \mathbf{M}_{eq}(\mathbf{x}') e^{jk_0 r' \cos \psi} \quad (5.20.b)$$

then the radiation integrals in (5.18.i) and (5.18.j) become

$$\mathbf{A}(\mathbf{x}_s) = \frac{\mu_0}{4\pi} \frac{e^{-jk_0 r_s}}{r_s} \mathbf{N}(\mathbf{x}_s) \quad (5.20.c)$$

$$\mathbf{F}(\mathbf{x}_s) = \frac{\varepsilon_0}{4\pi} \frac{e^{-jk_0 r_s}}{r_s} \mathbf{L}(\mathbf{x}_s) \quad (5.20.d)$$

We can combine (5.19.j), (5.19.k), (5.20.c) and (5.10.d) in order to discover that

$$(E^s)_\theta = -\frac{jk_0 e^{-jk_0 r_s}}{4\pi r_s} (\eta_0 N_\theta + L_\phi) \quad (5.20.e)$$

$$(E^s)_\phi = +\frac{jk_0 e^{-jk_0 r_s}}{4\pi r_s} (L_\theta - \eta_0 N_\phi) \quad (5.20.f)$$

Inspection of (5.20.e) and (5.20.f) reveals that we need to calculate the θ - and ϕ -spherical components of $\mathbf{N}$ and $\mathbf{L}$. [Remember, they refer to the observation point (r_s, θ_s, ϕ_s)]. However, the equivalent currents $\mathbf{J}_{eq}$ and $\mathbf{M}_{eq}$ are expressed in Cartesian coordinates. So we need a conversion between these two coordinate systems. If we represent $\mathbf{J}_{eq} = [J_x, J_y, J_z]^T$ and $\mathbf{M}_{eq} = [M_x, M_y, M_z]^T$, then the corresponding spherical components can be found through

$$\begin{bmatrix} J_r \\ J_\theta \\ J_\phi \end{bmatrix} = \begin{bmatrix} \sin \theta_s \cos \phi_s & \sin \theta_s \sin \phi_s & \cos \theta_s \\ \cos \theta_s \cos \phi_s & \cos \theta_s \sin \phi_s & -\sin \theta_s \\ -\sin \phi_s & \cos \phi_s & 0 \end{bmatrix} \begin{bmatrix} J_x \\ J_y \\ J_z \end{bmatrix}. \quad (5.20.g)$$

$$\begin{bmatrix} M_r \\ M_\theta \\ M_\phi \end{bmatrix} = \begin{bmatrix} \sin \theta_s \cos \phi_s & \sin \theta_s \sin \phi_s & \cos \theta_s \\ \cos \theta_s \cos \phi_s & \cos \theta_s \sin \phi_s & -\sin \theta_s \\ -\sin \phi_s & \cos \phi_s & 0 \end{bmatrix} \begin{bmatrix} M_x \\ M_y \\ M_z \end{bmatrix}. \quad (5.20.h)$$

With the help of (5.20.g) and (5.20.h), the components of the vectors in (5.20.a) and (5.20.b) become

$$N_\theta = \oint_{\Sigma} (J_x \cos \theta_s \cos \phi_s + J_y \cos \theta_s \sin \phi_s - J_z \sin \theta_s) e^{jk_0 r' \cos \psi} \quad (5.20.i)$$

$$L_\theta = \oint_{\Sigma} (M_x \cos \theta_s \cos \phi_s + M_y \cos \theta_s \sin \phi_s - M_z \sin \theta_s) e^{jk_0 r' \cos \psi} \quad (5.20.j)$$

$$N_\phi = \oint_{\Sigma} (-J_x \sin \phi_s + J_y \cos \phi_s) e^{jk_0 r' \cos \psi} \quad (5.20.k)$$

$$L_\phi = \oint_{\Sigma} (-M_x \sin \phi_s + M_y \cos \phi_s) e^{jk_0 r' \cos \psi} \quad (5.20.l)$$

The phase factor $jk_0 r' \cos \psi$ also deserves attention. Because ψ is the angle between the vectors $\mathbf{x}_s$ and $\mathbf{x}'$, the definition of dot product between two vectors gives us

$$\mathbf{x}_s \cdot \mathbf{x}' = \|\mathbf{x}_s\| \|\mathbf{x}'\| \cos \psi. \quad (5.20.m)$$

Since $\|\mathbf{x}_s\| = r_s$ and $\|\mathbf{x}'\| = r'$, we get

$$r' \cos \psi = \hat{\mathbf{r}}_s \cdot \mathbf{x}', \quad (5.20.n)$$

where $\hat{\mathbf{r}}_s$ has been defined at (5.19.e). In Cartesian coordinates, any source point $\mathbf{x}' \in \Sigma$ can be represented by

$$\mathbf{x}' = x' \hat{\mathbf{x}} + y' \hat{\mathbf{y}} + z' \hat{\mathbf{z}} \quad (5.20.o)$$

Moreover, as the spherical-to-Cartesian conversion is given by

$$\begin{bmatrix} (\hat{\mathbf{r}}_s)_x \\ (\hat{\mathbf{r}}_s)_y \\ (\hat{\mathbf{r}}_s)_z \end{bmatrix} = \begin{bmatrix} \sin \theta_s \cos \phi_s & \cos \theta_s \cos \phi_s & -\sin \phi_s \\ \sin \theta_s \sin \phi_s & \cos \theta_s \sin \phi_s & \cos \phi_s \\ \cos \theta_s & -\sin \theta_s & 0 \end{bmatrix} \begin{bmatrix} (\hat{\mathbf{r}}_s)_r \\ (\hat{\mathbf{r}}_s)_\theta \\ (\hat{\mathbf{r}}_s)_\phi \end{bmatrix}, \quad (5.20.p)$$

and as obviously $(\hat{\mathbf{r}}_s)_r = 1$, $(\hat{\mathbf{r}}_s)_\theta = 0$ and $(\hat{\mathbf{r}}_s)_\phi = 0$,

$$\hat{\mathbf{r}}_s = \sin \theta_s \cos \phi_s \hat{\mathbf{x}} + \sin \theta_s \sin \phi_s \hat{\mathbf{y}} + \cos \theta_s \hat{\mathbf{z}}. \quad (5.20.q)$$

In this way, from (5.20.n), (5.20.o) and (5.20.q),

$$e^{jk_0 r' \cos \psi} = e^{jk_0 (x' \sin \theta_s \cos \phi_s + y' \sin \theta_s \sin \phi_s + z' \cos \theta_s)} \quad (5.20.r)$$

Now that we are able to calculate (5.20.i) – (5.20.l), the θ - and ϕ -spherical components of the scattered electric field in (5.20.e) and (5.20.f) can be determined. The square of the modulus of these complex-valued quantities is

$$|(E^s)_\theta|^2 = \frac{k_0^2}{(4\pi r_s)^2} |\eta_0 N_\theta + L_\phi|^2 \quad (5.21.a)$$

$$|(E^s)_\phi|^2 = \frac{k_0^2}{(4\pi r_s)^2} |L_\theta - \eta_0 N_\phi|^2 \quad (5.21.b)$$

Because E^s has no radial component,

$$|E^s(r_s, \theta_s, \phi_s)|^2 = \frac{k_0^2}{(4\pi r_s)^2} (|\eta_0 N_\theta + L_\phi|^2 + |L_\theta - \eta_0 N_\phi|^2) \quad (5.21.c)$$

The incident fields in (5.16.h) and (5.16.i) are plane waves, whose amplitude does not depend on the radial distance r_s . Actually, the squared amplitude of these plane waves is constant throughout the space $\mathbb{R}^3$:

$$|E^{inc}(r_s, \theta_s, \phi_s)|^2 = \Psi, \quad (5.21.d)$$

where

$$\Psi = |E_0|^2, \quad TM^x \quad (5.21.e)$$

$$\Psi = |\eta_0 H_0|^2 (1 - (\sin \theta_i)^2 (\cos \phi_i)^2), \quad TE^x \quad (5.21.f)$$

With the information provided by (5.21.c) and (5.21.d), the RCS in (5.17.a) becomes

$$\sigma_{3D}(\theta_s, \phi_s) = \lim_{r_s \rightarrow \infty} 4\pi r_s^2 \frac{k_0^2}{(4\pi r_s)^2 \Psi} (|\eta_0 N_\theta + L_\phi|^2 + |L_\theta - \eta_0 N_\phi|^2), \quad (5.21.g)$$

The r_s^2 term gets cancelled. Furthermore, none of the integrals in (5.20.i) – (5.20.l) depends on the distance r_s . In this way, the right side of (5.21.g) does not depend on r_s , and we are safe to pass to the limit. Finally, we get the expression for the RCS:

$$\sigma_{3D}(\theta_s, \phi_s) = \frac{k_0^2}{4\pi \Psi} (|\eta_0 N_\theta + L_\phi|^2 + |L_\theta - \eta_0 N_\phi|^2). \quad (5.21.h)$$

The procedure for calculating the RCS can be summarized in the Chart 5.2 below.

Chart 5.2 – Calculating the RCS

Step 1. Set up an imaginary closed surface Σ around the scatterer.

Step 2. Calculate the equivalent currents $\mathbf{J}_{eq}$ and $\mathbf{M}_{eq}$ on Σ , according to (5.18.b) and (5.18.c).

Step 3. Choose an observation point (at infinity) characterized by the angles (θ_s, ϕ_s) .

Step 4. Calculate the phase term (5.20.r).

Step 5. Evaluate the integrals N_θ , N_ϕ , L_θ and L_ϕ in (5.20.i) – (5.20.l).

Step 6. Calculate the RCS in (5.21.h).

Step 7. Choose another observation point (θ_s, ϕ_s) and go back to Step 4.

5.4.6.2 Two dimensions

The process for getting the scattering width SW in (5.17.b) is derived from that of the RCS. We must refer back to the geometry in Fig. 5.8. Since the strip extends to infinity along the z -direction, our three-dimensional imaginary surface Σ is not closed. It is set up as follows. Let us place an imaginary closed curve σ around the strip cross-section in the XY plane. Then we make

$$\Sigma = \sigma \times (-\infty, +\infty) \quad (5.22.a)$$

The expressions for the wave potentials in (5.18.d) and (5.18.e) give the values of $\mathbf{A}$ and $\mathbf{F}$ at observation points $\mathbf{x}_s \in \mathbb{R}^3$. In cylindrical coordinates, the observation point $\mathbf{x}_s$ can be represented as $[\rho_s, \phi_s, z_s]^T$, and as $[x_s, y_s, z_s]^T$ in Cartesian coordinates. Analogously, a source point $\mathbf{x}' \in \Sigma$ has the cylindrical and Cartesian representation as $[\rho', \phi', z']^T$ and $[x', y', z']^T$, respectively. We can write

$$R = |\mathbf{x}_s - \mathbf{x}'| = \sqrt{(x_s - x')^2 + (y_s - y')^2 + (z_s - z')^2} \quad (5.22.b)$$

$$= \sqrt{|\boldsymbol{\rho}_s - \boldsymbol{\rho}'|^2 + (z_s - z')^2}, \quad (5.22.c)$$

where

$$\boldsymbol{\rho}_s = x_s \hat{\mathbf{x}} + y_s \hat{\mathbf{y}} \quad (5.22.d)$$

$$\boldsymbol{\rho}' = x' \hat{\mathbf{x}} + y' \hat{\mathbf{y}}. \quad (5.22.e)$$

The potentials $\mathbf{A}$ and $\mathbf{F}$ become

$$\mathbf{A}(\mathbf{x}_s) = \frac{\mu_0}{4\pi} \oint_{\Sigma} \mathbf{J}_{eq}(\mathbf{x}') \frac{e^{-jk_0 \sqrt{|\boldsymbol{\rho}_s - \boldsymbol{\rho}'|^2 + (z_s - z')^2}}}{\sqrt{|\boldsymbol{\rho}_s - \boldsymbol{\rho}'|^2 + (z_s - z')^2}} d\Sigma \quad (5.22.f)$$

$$\mathbf{F}(\mathbf{x}_s) = \frac{\varepsilon_0}{4\pi} \oint_{\Sigma} \mathbf{M}_{eq}(\mathbf{x}') \frac{e^{-jk_0 \sqrt{|\boldsymbol{\rho}_s - \boldsymbol{\rho}'|^2 + (z_s - z')^2}}}{\sqrt{|\boldsymbol{\rho}_s - \boldsymbol{\rho}'|^2 + (z_s - z')^2}} d\Sigma. \quad (5.22.g)$$

Since the geometry of this problem is invariant along z , the equivalent currents $\mathbf{J}_{eq}$ and $\mathbf{M}_{eq}$ do not depend on z , i.e., $\mathbf{J}_{eq}(\mathbf{x}') = \mathbf{J}_{eq}(\boldsymbol{\rho}')$ and $\mathbf{M}_{eq}(\mathbf{x}') = \mathbf{M}_{eq}(\boldsymbol{\rho}')$. Moreover, the differential element $d\Sigma$ is equal to $d\sigma dz'$, where $d\sigma$ a differential length along the curve σ . Then,

$$\mathbf{A}(\mathbf{x}_s) = \frac{\mu_0}{4\pi} \oint_{\sigma} \mathbf{J}_{eq}(\boldsymbol{\rho}') \left(\int_{-\infty}^{+\infty} \frac{e^{-jk_0\sqrt{|\boldsymbol{\rho}_s - \boldsymbol{\rho}'|^2 + (z_s - z')^2}}}{\sqrt{|\boldsymbol{\rho}_s - \boldsymbol{\rho}'|^2 + (z_s - z')^2}} dz' \right) d\sigma \quad (5.22.h)$$

$$\mathbf{F}(\mathbf{x}_s) = \frac{\varepsilon_0}{4\pi} \oint_{\sigma} \mathbf{M}_{eq}(\boldsymbol{\rho}') \left(\int_{-\infty}^{+\infty} \frac{e^{-jk_0\sqrt{|\boldsymbol{\rho}_s - \boldsymbol{\rho}'|^2 + (z_s - z')^2}}}{\sqrt{|\boldsymbol{\rho}_s - \boldsymbol{\rho}'|^2 + (z_s - z')^2}} dz' \right) d\sigma \quad (5.22.i)$$

It is known that [Balanis, 1989]:

$$\int_{-\infty}^{+\infty} \frac{e^{-jk_0\sqrt{|\boldsymbol{\rho}_s - \boldsymbol{\rho}'|^2 + (z_s - z')^2}}}{\sqrt{|\boldsymbol{\rho}_s - \boldsymbol{\rho}'|^2 + (z_s - z')^2}} dz' = -j\pi H_0^{(2)}(k_0|\boldsymbol{\rho}_s - \boldsymbol{\rho}'|), \quad (5.22.j)$$

where $H_0^{(2)}(\cdot)$ is the Hankel function of the second type. In this way,

$$\mathbf{A}(\mathbf{x}_s) = -\frac{j\mu_0}{4} \oint_{\sigma} \mathbf{J}_{eq}(\boldsymbol{\rho}') H_0^{(2)}(k_0|\boldsymbol{\rho}_s - \boldsymbol{\rho}'|) d\sigma \quad (5.22.k)$$

$$\mathbf{F}(\mathbf{x}_s) = -\frac{j\varepsilon_0}{4} \oint_{\sigma} \mathbf{M}_{eq}(\boldsymbol{\rho}') H_0^{(2)}(k_0|\boldsymbol{\rho}_s - \boldsymbol{\rho}'|) d\sigma. \quad (5.22.l)$$

For very large arguments, it is known that the Hankel functions satisfy [Balanis, 1989]:

$$H_n^{(2)}(t) \approx \sqrt{\frac{2}{\pi t}} e^{-j\left(t - \frac{n\pi}{2} - \frac{\pi}{4}\right)}, \quad \text{as } t \rightarrow \infty. \quad (5.22.m)$$

Since $e^{j\pi/4} = \sqrt{j}$, when we take $n = 0$ in (5.22.m), we get

$$\mathbf{A}(\mathbf{x}_s) = \mu_0 \sqrt{\frac{j}{8\pi k_0}} \oint_{\sigma} \mathbf{J}_{eq}(\boldsymbol{\rho}') \frac{e^{-jk_0|\boldsymbol{\rho}_s - \boldsymbol{\rho}'|}}{\sqrt{|\boldsymbol{\rho}_s - \boldsymbol{\rho}'|}} d\sigma \quad (5.22.n)$$

$$\mathbf{F}(\mathbf{x}_s) = \varepsilon_0 \sqrt{\frac{j}{8\pi k_0}} \oint_{\sigma} \mathbf{M}_{eq}(\boldsymbol{\rho}') \frac{e^{-jk_0|\boldsymbol{\rho}_s - \boldsymbol{\rho}'|}}{\sqrt{|\boldsymbol{\rho}_s - \boldsymbol{\rho}'|}} d\sigma \quad (5.22.o)$$

When the observation point is very far from the scatterer, i.e., when ρ_s is large, it holds the approximation

$$\frac{1}{|\boldsymbol{\rho}_s - \boldsymbol{\rho}'|} \approx \frac{1}{\rho_s} \quad (5.22.p)$$

$$e^{-jk_0|\boldsymbol{\rho}_s - \boldsymbol{\rho}'|} \approx e^{-jk_0(\rho_s - \rho' \cos(\phi_s - \phi'))} \quad (5.22.q)$$

Substitution of (5.22.p) and (5.22.q) into (5.22.n) and (5.22.o) leads us to

$$\mathbf{A}(\mathbf{x}_s) = \mu_0 \sqrt{\frac{j}{8\pi k_0}} \frac{e^{-jk_0 \rho_s}}{\sqrt{\rho_s}} \oint_{\sigma} \mathbf{J}_{eq}(\boldsymbol{\rho}') e^{jk_0 \cos(\phi_s - \phi')} d\sigma \quad (5.22.r)$$

$$\mathbf{F}(\mathbf{x}_s) = \varepsilon_0 \sqrt{\frac{j}{8\pi k_0}} \frac{e^{-jk_0 \rho_s}}{\sqrt{\rho_s}} \oint_{\sigma} \mathbf{M}_{eq}(\boldsymbol{\rho}') e^{jk_0 \cos(\phi_s - \phi')} d\sigma. \quad (5.22.s)$$

Now it can be verified that no term in the right side of the integrals above depend on the z_s coordinate. We can rewrite (5.22.r) and (5.22.s) as

$$\mathbf{A}(\boldsymbol{\rho}_s) = \mu_0 \sqrt{\frac{j}{8\pi k_0}} \frac{e^{-jk_0 \rho_s}}{\sqrt{\rho_s}} \mathbf{N}(\boldsymbol{\rho}_s) \quad (5.22.t)$$

$$\mathbf{F}(\boldsymbol{\rho}_s) = \varepsilon_0 \sqrt{\frac{j}{8\pi k_0}} \frac{e^{-jk_0 \rho_s}}{\sqrt{\rho_s}} \mathbf{L}(\boldsymbol{\rho}_s). \quad (5.22.u)$$

where

$$\mathbf{N}(\boldsymbol{\rho}_s) = \oint_{\sigma} \mathbf{J}_{eq}(\boldsymbol{\rho}') e^{jk_0 \cos(\phi_s - \phi')} d\sigma \quad (5.22.v)$$

$$\mathbf{L}(\boldsymbol{\rho}_s) = \oint_{\sigma} \mathbf{M}_{eq}(\boldsymbol{\rho}') e^{jk_0 \cos(\phi_s - \phi')} d\sigma \quad (5.22.w)$$

The equivalent currents $\mathbf{J}_{eq}$ and $\mathbf{M}_{eq}$ depend on the source points $\boldsymbol{\rho}' \in \sigma$. These currents in principle have three components, according to (5.18.c) and (5.18.d). Their conversion into spherical coordinates is given by (5.20.g) and (5.20.h). This allows us to calculate N_{θ} , N_{ϕ} , L_{θ} , and L_{ϕ} :

$$N_{\theta} = \oint_{\sigma} (J_x \cos \theta_s \cos \phi_s + J_y \cos \theta_s \sin \phi_s - J_z \sin \theta_s) e^{jk_0 \cos(\phi_s - \phi')} d\sigma \quad (5.23.a)$$

$$L_{\theta} = \oint_{\sigma} (M_x \cos \theta_s \cos \phi_s + M_y \cos \theta_s \sin \phi_s - M_z \sin \theta_s) e^{jk_0 \cos(\phi_s - \phi')} d\sigma \quad (5.23.b)$$

$$N_{\phi} = \oint_{\sigma} (-J_x \sin \phi_s + J_y \cos \phi_s) e^{jk_0 \cos(\phi_s - \phi')} d\sigma \quad (5.23.c)$$

$$L_{\phi} = \oint_{\sigma} (-M_x \sin \phi_s + M_y \cos \phi_s) e^{jk_0 \cos(\phi_s - \phi')} d\sigma \quad (5.23.d)$$

In what regards the spherical components of the scattered electric field, the same expressions as those of (5.19.j) and (5.19.k) apply here:

$$(E^s)_\theta = -j\omega(A_\theta + \eta_0 F_\phi) \quad (5.23.e)$$

$$(E^s)_\phi = -j\omega(A_\phi - \eta_0 F_\theta). \quad (5.23.f)$$

With the help of (5.22.t) and (5.22.u), the two expressions above become

$$(E^s)_\theta = -j\omega\varepsilon_0\eta_0 \sqrt{\frac{j}{8\pi k_0}} \frac{e^{-jk_0\rho_s}}{\sqrt{\rho_s}} (\eta_0 N_\theta + L_\phi) \quad (5.23.g)$$

$$(E^s)_\phi = -j\omega\varepsilon_0\eta_0 \sqrt{\frac{j}{8\pi k_0}} \frac{e^{-jk_0\rho_s}}{\sqrt{\rho_s}} (\eta_0 N_\phi - L_\theta) \quad (5.23.h)$$

Since $\mathbf{E}^s$ has no radial component, and because no quantity depends on the z_s coordinate,

$$|\mathbf{E}^s(\rho_s)|^2 = \frac{k_0}{8\pi} \frac{1}{\rho_s} \left(|\eta_0 N_\theta + L_\phi|^2 + |L_\theta - \eta_0 N_\phi|^2 \right) \quad (5.23.i)$$

According to (5.10.g), the square of the modulus of the incident field $\mathbf{E}^{inc}$ is simply

$$|\mathbf{E}^{inc}(\rho_s)|^2 = |\eta_0 H_0|^2 \quad (5.23.j)$$

From the definition of scattering width SW in (5.17.b), and (5.23.i), (5.23.j),

$$\sigma_{2D}(\phi_s) = \lim_{\rho_s \rightarrow \infty} 2\pi\rho_s \frac{k_0}{8\pi} \frac{1}{\rho_s} \frac{1}{|\eta_0 H_0|^2} \left(|\eta_0 N_\theta + L_\phi|^2 + |L_\theta - \eta_0 N_\phi|^2 \right) \quad (5.23.k)$$

The distance ρ_s gets cancelled in the right side, which allows us to pass to the limit as $\rho_s \rightarrow \infty$. Since no term in the right side of (5.23.k) depends on ρ_s , the SW depends just on the observation angle ϕ_s . We finally get

$$\sigma_{2D}(\phi_s) = \frac{k_0}{4|\eta_0 H_0|^2} \left(|\eta_0 N_\theta + L_\phi|^2 + |L_\theta - \eta_0 N_\phi|^2 \right) \quad (5.23.l)$$

The procedure for calculating the SW can be summarized in the Chart 5.3 below.

Chart 5.3 – Calculating the SW

Step 1. Set up an imaginary closed curve σ around the scatterer.

Step 2. Calculate the equivalent currents $\mathbf{J}_{eq}$ and $\mathbf{M}_{eq}$ on σ , according to (5.18.b) and (5.18.c).

Step 3. Choose an observation point (at infinity) characterized by the angle ϕ_s .

Step 4. Evaluate the integrals N_θ , N_ϕ , L_θ and L_ϕ in (5.23.a) – (5.23.d).

Step 5. Calculate the SW in (5.23.l).

Step 6. Choose another observation point ϕ_s .and go back to Step 3.

5.4.6.3 Physical Optics

In order to find out if the solutions to the scattering problems in Sections 5.4.3 and 5.4.5 are reliable, we need to compare the results with some standard. These problems lack analytical solutions, and so we need another standard to compare with.

In this work, we are going to compare the results provided by our meshfree method with those from the *physical optics approximation* (PO).

The problem regarding the scattering of waves by PEC obstacles is of much practical concern, and there are alternate methods by which they can be formulated. The physical optics is one of them.

When trying to find the scattered fields, one needs the current distributions on the surface of the PEC obstacle. If the current is known, then the vector potentials, and consequently the scattered fields, can be found via radiation integrals such as (5.18.d). However, if the obstacle is not an infinite and flat PEC surface, then the current density is generally unknown. For more general geometries, and when the only available information besides the geometry of the target is the incident field, one can find suitable *approximations* for the current densities. Once these are found, the scattered fields are calculated through (5.18.d).

In the physical optics approximation, given the geometry of the conductor (with the normal $\hat{\mathbf{n}}$ defined almost everywhere on its surface) and the incident field $(\mathbf{E}^{inc}, \mathbf{H}^{inc})$, the current density at the surface of the PEC obstacle is taken as

$$\mathbf{J}_{PO} = 2\hat{\mathbf{n}} \times \mathbf{H}^{inc} . \quad (5.24)$$

The approximation provided by (5.24) is meaningful, provided the scatterer is electrically large.

In what regards the physical optics approximation, this is all we need to know in this work. More details and an extensive explanation can be found in [Balanis, 1989].

For the problems discussed in this chapter (the scattering of plane waves by conducting strips and plates), the physical optics approximation provides closed results for the radar cross sections. In a sense, the RCS calculated by PO and those resulting from the ‘full theory’ agree with each other near the specular direction. (By specular direction it is meant the direction along which the incident wave is reflected by the conducting surface.) The predictions of the PO become less accurate away from the specular directions. One of the reasons is that, since the PO employs the approximation (5.24), which is valid only when the flat conductor is infinite, when in reality it is not, the PO fails to take the edge diffraction effects into account. But the results from the PO

are accurate near the specular directions, and as such provide a standard against which we can compare the results of our meshfree calculations.

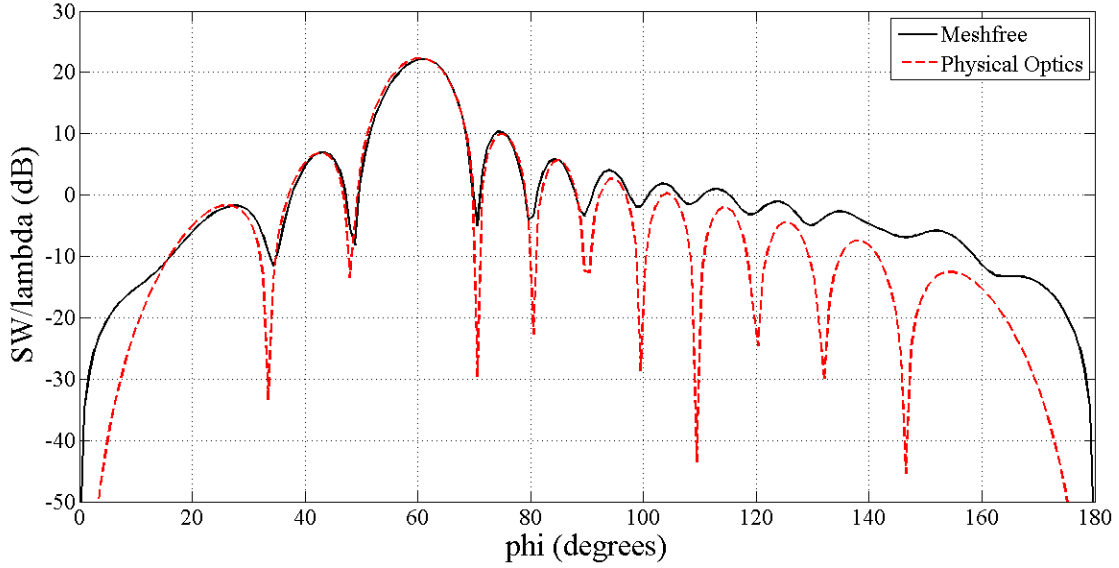
Although the predictions of the PO do not match exactly those from the ‘full theory’ as described above, they can nonetheless be used as standards against which results provided by another numerical method can be compared [Heldring *et al.*, 2002].

The results concerning the PO approximations for the geometries in the problems that interest us are taken from [Balanis, 1989] and are summarized in Table 5.3 below.

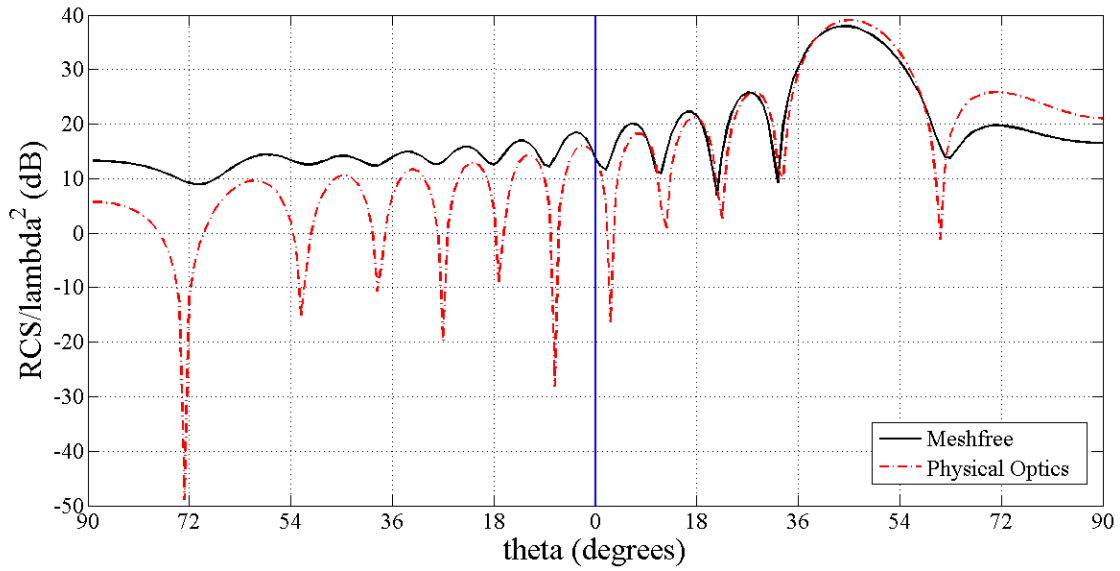
TABLE 5.3 – PHYSICAL OPTICS APPROXIMATION

Problem	PO Expression
Scattering of a TE^z plane wave by a PEC strip (Section 5.4.3)	$\sigma_{2D}(\phi_s) = \frac{2\pi b^2}{\lambda_0} \left(\sin \phi_s \frac{\sin X}{X} \right)^2 \quad (5.25.a)$ $X = \frac{k_0 b}{2} (\cos \phi_s + \cos \phi_i) \quad (5.25.b)$
Scattering of a TM^x plane wave by a PEC plate (Section 5.4.5)	$\sigma_{3D}(\theta_s, \phi_s) = 4\pi \left(\frac{b^2}{\lambda_0} \right)^2 Z \left(\frac{\sin X}{X} \right)^2 \left(\frac{\sin Y}{Y} \right)^2 \quad (5.26.a)$ $Z = \cos^2 \theta_i (\cos^2 \theta_s \cos^2 \phi_s + \sin^2 \phi_s) \quad (5.26.b)$ $X = \frac{k_0 b}{2} \sin \theta_s \cos \phi_s \quad (5.26.c)$ $Y = \frac{k_0 b}{2} (\sin \theta_s \sin \phi_s - \sin \theta_i) \quad (5.26.d)$
Scattering of a TE^x plane wave by a PEC plate (Section 5.4.5)	$\sigma_{3D}(\theta_s, \phi_s) = 4\pi \left(\frac{b^2}{\lambda_0} \right)^2 Z \left(\frac{\sin X}{X} \right)^2 \left(\frac{\sin Y}{Y} \right)^2 \quad (5.27.a)$ $Z = \cos^2 \theta_s \cos^2 \phi_s + \cos^2 \phi_s \quad (5.27.b)$ $X = \frac{k_0 b}{2} \sin \theta_s \cos \phi_s \quad (5.27.c)$ $Y = \frac{k_0 b}{2} (\sin \theta_s \sin \phi_s - \sin \theta_i) \quad (5.27.d)$

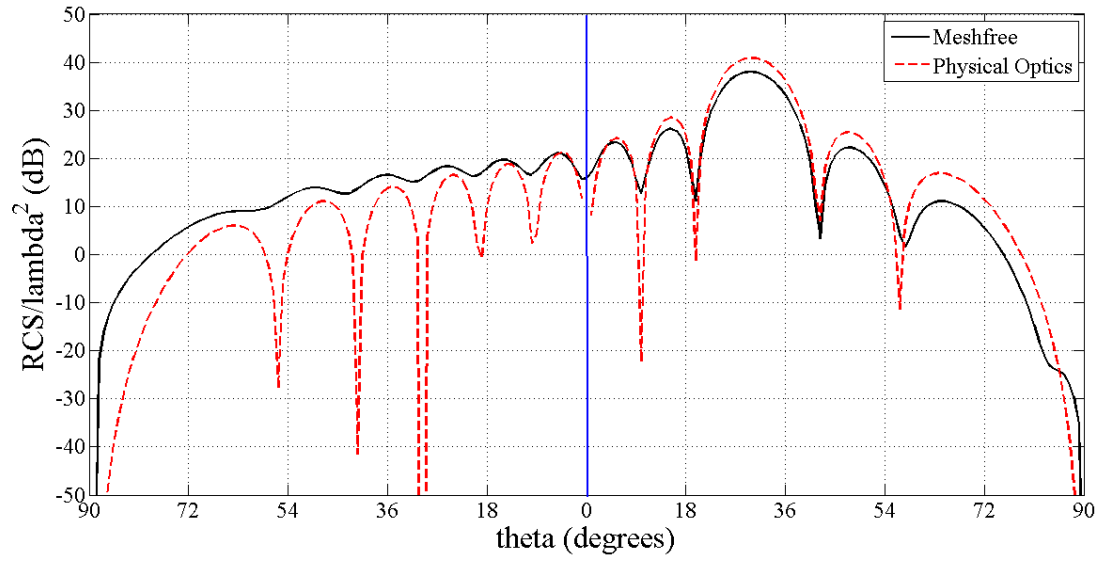
The RCS (and the SW) for each case has been calculated twice: First, we calculated the near-fields through our meshfree methods and from them we obtained the RCS (and the SW) via the procedure outlined in Charts 5.2 and 5.3. Second, the PO approximation to the RCS (and the SW) has been calculated from the expressions in Table 5.3. The results are in Fig. 5.20.



(a)



(b)



(c)

Fig. 5.20. Radar cross sections. (a) The normalized scattering width (SW) in decibels, according to (5.17.f) for the bi-dimensional strip problem. The observation angle ϕ_s is like that indicated in Fig. 5.8.b. (b) The normalized radar cross section (RCS) in decibels, according to (5.17.d) for the three-dimensional scattering of a TM^x wave by a PEC plate. (c) Normalized RCS in decibels for the scattering of a TE^x wave by a PEC plate. In the last two figures, the RCS is calculated in two regions, and the results are separated from each other by a blue line at the center of the graph. In region 1 (at the right of the blue line), $\phi_s = 90^\circ$ and $0 \leq \theta_s \leq 90^\circ$. In region 2 (at the left), $\phi_s = 270^\circ$ and $0 \leq \theta_s \leq 90^\circ$.

The results provided by the ‘full theory’ and the PO approximation agree with each other in the vicinity of the specular directions. When we consider directions away from the specular directions, there is still some concordance between the curves, particularly in what concerns the relative positions of the maxima and minima. The overall behavior of the two curves, in a sense, corresponds to what has been predicted earlier.

Chapter 6

Conclusions

6.1 Concluding remarks

We have finally arrived at the end of this thesis. Our primary objective was to find a nodal meshfree method aimed at solving vector problems in electromagnetism subject to the divergence-free constraint. The method of finite spheres, which relies on the partition of unity paradigm, provided a solid basis for the construction of our method.

As the work evolved, we felt that the task would be more mathematically demanding than we initially thought. We had to choose carefully which function spaces would be used and the right formulation to be employed. But that is not all: After the formulation had been established, it was necessary to show that it is consistent, or, as it is said, that it is well-posed. At precisely this point we realized that, if we were to actually provide a good formulation, it should be justified. And to justify it, we had to resort to concepts available only at a somehow higher mathematical level.

In a sense, we had to construct a ‘theory’ to justify our formulation. Fortunately, it was not necessary to begin from the scratch: We took the theory already developed for the Navier-Stokes system and adapted it to the wave scattering system, which is what ultimately interests us here.

Now that the work is complete, it can be observed that the theoretical aspects fit our meshfree method, and vice-versa. This is not coincidence: It was planned to be so. Moreover, the forms assumed by the theory and by the method were not conceived at once. We began with an aspect of the theory, and found that it needed some adjustments to fit the numerical method. In the same way, some aspects of the method had to be modified in order to accommodate the theoretical requirements. It took some time to figure out all the adjustments that had to be made so that the theory and the method could match each other.

The examples show that the method works well when applied to problems concerning the electromagnetic wave scattering by conducting objects in three-dimensions. The application of our method to the scattering by metallic plates can be viewed as a template: Any problem in this category can be solved by exactly the same way described in the thesis. Of course, more complicated targets will demand more computational power. But even more important is the fact that we have found a way to do it, i.e., we have now a recipe about how to solve such problems.

So that is it. When we trace a line going from our earlier works to the point where we are now, we are able to conclude that a formidable progress has been made. Much has been learned along the way, and we are grateful for all the knowledge gathered during these Ph.D. years.

6.2 Future work

Of course, there are some points raised during the development of this thesis that have not been addressed. We identified at least three of them, which are worth considering in future works.

6.2.1 The tangential trace operator

According to Section 2.1.4, there is a trace operator

$$\gamma_0^d: H^1(\Omega)^3 \rightarrow H^{1/2}(\Gamma)^3 \quad (6.1)$$

which is used in connection with the non-homogeneous Dirichlet boundary conditions in the Navier-Stokes system. It says that, when we know *all the three components* of the velocity field $\mathbf{u}$ at the boundary Γ , i.e., when we know that $\mathbf{u} = \mathbf{g}$ at Γ , if $\mathbf{g}$ is an element of $H^{1/2}(\Gamma)^3$, then one can find a function $\mathbf{u}^g \in H^1(\Omega)^3$ such that $\gamma_0^d \mathbf{u}^g = \mathbf{g}$. Since the velocity field $\mathbf{u}$ and $\mathbf{u}^g$ are in $H^1(\Omega)^3$, we can form the decomposition (2.78),

$$\mathbf{u} = \mathbf{u}^0 + \mathbf{u}^g \quad (6.2)$$

thus allowing the problem to be formulated in terms of $\mathbf{u}^0$, which obeys homogeneous Dirichlet conditions on all its components (i.e., all components of $\mathbf{u}^0$ are zero at Γ).

Analogously, in Section 2.2.3.4, there is a trace operator

$$\gamma_t: H(\mathbf{curl}; \Omega) \rightarrow Y(\Gamma), \quad (6.3)$$

which is used in connection with the non-homogeneous Dirichlet boundary conditions in the traditional formulation for the scattering system. When we know the tangential component of the scattered electric field $\mathbf{E}^s$ at Γ , i.e., when we know that $\hat{\mathbf{n}} \times \mathbf{E}^s = \mathbf{g}$ at Γ , if $\mathbf{g}$ is an element of $Y(\Gamma)$, then we can find a function $\mathbf{u}^g \in H(\mathbf{curl}; \Omega)$ such that $\gamma_t \mathbf{u}^g = \mathbf{g}$. Since both $\mathbf{E}^s$ and $\mathbf{u}^g$ are in $H(\mathbf{curl}; \Omega)$, we can form the decomposition

$$\mathbf{E}^s = \mathbf{E}^0 + \mathbf{u}^g, \quad (6.4)$$

thus allowing the problem to be formulated in terms of $\mathbf{E}^0$, which is such that $\gamma_t \mathbf{E}^0 = \hat{\mathbf{n}} \times \mathbf{E}^s = \mathbf{0}$ at Γ .

Both trace operators from (6.1) and (6.3) are backed by well-established theories. But as discussed in Section 2.2.3.4, what we really want is a characterization of the ‘inverse’ of the tangential trace operator

$$\gamma_t: H^1(\Omega)^3 \rightarrow Y(\Gamma). \quad (6.5)$$

The operator in (6.5) operates in the same way as the operator in (6.3). Since $H^1(\Omega)^3 \subset H(\mathbf{curl}; \Omega)$, it is just a restriction of the operator in (6.3) to those functions from the subspace $H^1(\Omega)^3$. We are interested in the opposite question: Given an element $\mathbf{g}$ from $Y(\Gamma)$, can we find an element in the more regular space $H^1(\Omega)^3$ whose image under γ_t is $\mathbf{g}$? According to the reasoning from Section 2.2.3.4, it is not unreasonable to expect that. Moreover, as it was shown, for any $\mathbf{g} \in Y(\Gamma)$ we can find a function in $H^1(\Omega)^3$ whose tangential trace is arbitrarily close to $\mathbf{g}$.

This does not satisfy us: We want a formal proof concerning the existence of the trace operator in (6.5). Maybe such trace operator exists from $H^1(\Omega)^3$ into a subspace of $Y(\Gamma)$. But which subspace? Moreover, maybe there are classes of domains Ω for which the operator in (6.5) is well-defined. But which classes?

In order to find an answer to these questions, we need to delve deeper into the theory of traces in Sobolev spaces.

6.2.2 Complex eigenvalues

In Section 3.3.6.7, we argued that the eigenproblem in (3.77.d)

Find $\mathbf{w}_\sigma \in \mathcal{X}^0$ such that

$$\int_{\Omega} (\bar{\mathbf{A}} \cdot \nabla \mathbf{w}_\sigma) : \nabla \mathbf{v}^* = \sigma \int_{\Omega} \mathbf{w}_\sigma \cdot \mathbf{v}^*, \quad \forall \mathbf{v} \in \mathcal{X}^0. \quad (6.6)$$

is likely to admit complex eigenvalues σ , due to the complex-valued components of the PML tensor $\bar{\mathbf{A}}$ at the left side of (6.6). We want a formal proof of this fact. If we find it, then it follows that the free-space wavenumber k_0^2 will never be an eigenvalue of (6.6), since it is a real number.

We believe that the answer will ultimately be provided by some argument from spectral theory.

6.2.3 Preconditioning

In section 5.3, we presented some discussion about the role of preconditioning matrices in the solution of large linear systems. We managed to find a cheap preconditioner, the matrix $\mathbf{M}$ given by (5.7.d):

$$\mathbf{M} = \begin{bmatrix} \bar{\mathbf{A}} & \mathbf{0} \\ \mathbf{0} & \bar{\mathbf{B}} \bar{\mathbf{D}}^{-1} \bar{\mathbf{B}}^T \end{bmatrix}. \quad (6.7)$$

The experimental results from Section 5.4 show that the GMRES together with the preconditioner in (6.7) was able to converge and deliver the right results in a reasonable number of iterations. One may ask: Is there another preconditioner that, when employed

in conjunction with the GMRES, is able to provide the correct answer but in a significantly smaller number of iterations? Suppose such a preconditioner exists. Is it as easy to construct as that in (6.7)?

Insight into these questions can be gained if more investigations are made in what concerns the use of preconditioners in the solution of sparse linear systems.

Appendix 1

Theorem 3.8

Theorem 3.8 concerns the well-posedness of non-coercive problems. It is restated below for convenience.

Theorem 3.8: Non-coercive problems – Suppose the following hypotheses are true:

- (i) V and H are two Hilbert spaces satisfying the requirements of Theorem 3.7, i.e., $V \hookrightarrow H$.
- (ii) The map $I_{V \rightarrow H}$ is compact, i.e., $I_{V \rightarrow H} \in \mathcal{K}(V, H)$.
- (iii) $a(\cdot, \cdot): V \times V \rightarrow \mathbb{C}$ is a continuous sesquilinear form.
- (iv) The sesquilinear form from item (iii) satisfies the property: There exist constants $\eta > 0$ and $\kappa_0 \geq 0$ such that

$$\operatorname{Re}\{a(u, u)\} + \kappa_0 \|I_{V \rightarrow H}(u)\|_H^2 \geq \eta \|u\|_V^2. \quad \forall u \in V \quad (3.65)$$

It can be concluded that if the solution to the homogeneous (zero-data) problem

Find $u \in V$ such that

$$a(u, v) = 0, \quad \forall v \in V \quad (3.66)$$

is the zero element $u = 0_V$, then it is true that:

(a) The solution to the general problem

Find $u \in V$ such that

$$a(u, v) = \langle F, v \rangle_{V^*, V}, \quad \forall v \in V \quad (3.67)$$

exists and is unique for every functional $F \in V^*$.

(b) The solution u from (a) depends continuously on the data, i.e., there exists a positive constant C_{FA} such that

$$\|I_{V \rightarrow H}(u)\|_H \leq C_{FA} \|F\|_{V^*} \quad (3.68)$$

Proof: The first part is devoted to proving existence and uniqueness of the solution. The second part deals with the boundedness (continuity) of the solution.

Part I: Existence and Uniqueness

Let it be the original problem (3.64):

$$\begin{aligned} & \text{Find } u \in V \text{ such that} \\ & a(u, v) = \langle F, v \rangle_{V^*, V}, \quad \forall v \in V \end{aligned} \quad (A1.1)$$

Of course, we can add the same quantity to both sides in (A1.1) and get the equivalent problem:

$$\begin{aligned} & \text{Find } u \in V \text{ such that} \\ & a(u, v) + \lambda_0(I_{V \rightarrow H}(u), I_{V \rightarrow H}(v))_H = \langle F, v \rangle_{V^*, V} + \lambda_0(I_{V \rightarrow H}(u), I_{V \rightarrow H}(v))_H \quad \forall v \in V. \end{aligned} \quad (A1.2)$$

Since u and v are in V , they are transferred to H via $I_{V \rightarrow H}$, and an inner product of their images in H is formed and added to both sides in (A1.2). Applying Theorem 3.7 in the inner product at the right side in (A1.2) [by making $w = I_{V \rightarrow H}(u)$] we get

$$\begin{aligned} & \text{Find } u \in V \text{ such that} \\ & a(u, v) + \lambda_0(I_{V \rightarrow H}(u), I_{V \rightarrow H}(v))_H = \langle F, v \rangle_{V^*, V} + \lambda_0 \langle T \circ I_{V \rightarrow H}(u), v \rangle_{V^*, V}, \quad \forall v \in V \end{aligned} \quad (A1.3)$$

Moreover, according to Theorem 3.7, the operator T in (A1.3) is an element from $\mathcal{L}(H, V^*)$, i.e.,

$$T \in \mathcal{L}(H, V^*) \quad (A1.4)$$

Hypothesis (ii) says that $I_{V \rightarrow H} \in \mathcal{K}(V, H)$. This hypothesis together with (A1.4) above and Theorem 3.5 imply that

$$J_{V \rightarrow V^*} := T \circ I_{V \rightarrow H} \in \mathcal{K}(V, V^*), \quad (A1.5)$$

i.e., the map $J_{V \rightarrow V^*}$ is compact. So (A1.3), it gets simplified to

$$\begin{aligned} & \text{Find } u \in V \text{ such that} \\ & a(u, v) + \lambda_0(I_{V \rightarrow H}(u), I_{V \rightarrow H}(v))_H = \langle F, v \rangle_{V^*, V} + \lambda_0 \langle J_{V \rightarrow V^*}(u), v \rangle_{V^*, V}, \quad \forall v \in V \end{aligned} \quad (A1.6)$$

and consequently to

$$\begin{aligned} & \text{Find } u \in V \text{ such that} \\ & a(u, v) + \lambda_0(I_{V \rightarrow H}(u), I_{V \rightarrow H}(v))_H = \langle F + \lambda_0 J_{V \rightarrow V^*}(u), v \rangle_{V^*, V}, \quad \forall v \in V \end{aligned} \quad (A1.7)$$

In the left side of (A1.7), if we fix u , the map

$$a(u, \cdot) + \lambda_0(I_{V \rightarrow H}(u), I_{V \rightarrow H}(\cdot))_H : V \rightarrow \mathbb{C} \quad (A1.8)$$

is linear and continuous. Linearity is obvious. In order to see that it is continuous, the triangle inequality says that for any $v \in V$,

$$\left| a(u, v) + \lambda_0(I_{V \rightarrow H}(u), I_{V \rightarrow H}(v))_H \right| \leq |a(u, v)| + \left| \lambda_0(I_{V \rightarrow H}(u), I_{V \rightarrow H}(v))_H \right| \quad (A1.9)$$

$$\leq \alpha_a \|u\|_V \|v\|_V + \lambda_0 \|I_{V \rightarrow H}(u)\|_H \|I_{V \rightarrow H}(v)\|_H \quad (A1.10)$$

$$\leq \alpha_a \|u\|_V \|v\|_V + \lambda_0 C_e^2 \|u\|_V \|v\|_V \quad (A1.11)$$

$$\leq (\alpha_a \|u\|_V + \lambda_0 C_e^2 \|u\|_V) \|v\|_V \quad (A1.12)$$

In (A1.10), the definition of continuous sesquilinear forms (3.3) have been used, and also the Cauchy-Schwarz inequality regarding the inner product in H . In (A1.11), C_e is just the embedding constant from (3.60). Inequality (A1.12) allows us to conclude that

$$\frac{\left| a(u, v) + \lambda_0(I_{V \rightarrow H}(u), I_{V \rightarrow H}(v))_H \right|}{\|v\|_V} \leq \alpha_a \|u\|_V + \lambda_0 C_e^2 \|u\|_V, \quad \forall v \in V \setminus \{0\} \quad (A1.13)$$

and consequently that

$$\sup_{v \in V \setminus \{0\}} \frac{\left| a(u, v) + \lambda_0(I_{V \rightarrow H}(u), I_{V \rightarrow H}(v))_H \right|}{\|v\|_V} \leq (\alpha_a + \lambda_0 C_e^2) \|u\|_V \quad (A1.14)$$

But the left side in (3.66.n) is just the definition of the norm in V^* . Then,

$$\left\| a(u, \cdot) + \lambda_0(I_{V \rightarrow H}(u), I_{V \rightarrow H}(\cdot))_H \right\|_{V^*} \leq (\alpha_a + \lambda_0 C_e^2) \|u\|_V, \quad (A1.15)$$

and continuity has been proved. Let us call the map (A1.8) by Lu , since u has been fixed. Then, $Lu \in V^*$, defined as:

$$\langle Lu, v \rangle_{V^*, V} = a(u, v) + \lambda_0(I_{V \rightarrow H}(u), I_{V \rightarrow H}(v))_H, \quad \forall v \in V \quad (A1.16)$$

Expression (A1.15) then means that

$$\|Lu\|_{V^*} \leq (\alpha_a + \lambda_0 C_e^2) \|u\|_V. \quad (A1.17)$$

We now investigate how Lu depends on u , which had been fixed. In a sense, there is an operator L which maps $u \in V$ to $Lu \in V^*$. From (A1.16), it is clearly linear in u , i.e.,

$$L(\alpha_1 u_1 + \alpha_2 u_2) = \alpha_1 Lu_1 + \alpha_2 Lu_2, \quad \text{in } V^*, \quad (A1.18)$$

where α_1 and α_2 are arbitrary complex numbers. The operator L is also bounded, as

$$\|L\|_{\mathcal{L}(V, V^*)} := \sup_{u \in V \setminus \{0\}} \frac{\|Lu\|_{V^*}}{\|u\|_V} \leq \alpha_a + \lambda_0 C_e^2, \quad (A1.19)$$

with the help of (A1.17). Since L is bounded and linear, then $L \in \mathcal{L}(V, V^*)$.

Employing (A1.16), problem (A1.7) assumes the form

$$\begin{aligned} & \text{Find } u \in V \text{ such that} \\ & \langle Lu, v \rangle_{V^*, V} = \langle F + \lambda_0 J_{V \rightarrow V^*}(u), v \rangle_{V^*, V}, \quad \forall v \in V \end{aligned} \quad (A1.20)$$

In operator form, (A1.20) becomes

$$\begin{aligned} & \text{Find } u \in V \text{ such that} \\ & Lu = F + \lambda_0 J_{V \rightarrow V^*}(u), \quad \text{in } V^* \end{aligned} \quad (A1.21)$$

In order to solve (A1.21), the operator $L \in \mathcal{L}(V, V^*)$ must admit an inverse, i.e., L must be one-to-one. We claim that L is one-to-one. To show that L is one-to-one is the same as to show that

$$\text{Ker } L = \{0_V\}, \quad (A1.22)$$

i.e., that the kernel of L is just the zero element from V . So let us analyze the kernel

$$\text{Ker } L = \{w \in V \mid Lw = 0_{V^*} \text{ in } V^*\}. \quad (A1.23)$$

Suppose $w \in \text{Ker } L$. Then $Lw = 0_{V^*}$ in V^* and consequently, with the help of (A1.16),

$$\langle Lw, v \rangle_{V^*, V} = a(w, v) + \lambda_0 (I_{V \rightarrow H}(w), I_{V \rightarrow H}(v))_H = \langle 0_{V^*}, v \rangle_{V^*, V} = 0, \quad \forall v \in V \quad (A1.24)$$

Since $w \in V$, take $v = w$ in (A1.24). One finds that

$$a(w, w) + \lambda_0 (I_{V \rightarrow H}(w), I_{V \rightarrow H}(w))_H = 0, \quad \forall w \in V \quad (A1.25)$$

Expression (A1.25) means that both real and imaginary parts of the left side are equal to zero. It is given that λ_0 is a positive real number; moreover, $(I_{V \rightarrow H}(w), I_{V \rightarrow H}(w))_H$ is also a positive real number, since it is the inner product between the same quantities. So the real part of the left side in (A1.25) becomes

$$\text{Re}\{a(w, w)\} + \lambda_0 \|I_{V \rightarrow H}(w)\|_H^2 = 0. \quad (A1.26)$$

But if we take the hypothesis (iv) from Theorem 3.8 into consideration, we form the expression

$$\beta \|w\|_V^2 \leq \text{Re}\{a(w, w)\} + \lambda_0 \|I_{V \rightarrow H}(w)\|_H^2 = 0, \quad (A1.27)$$

which implies that $\beta \|w\|_V^2 \leq 0$ and consequently $\|w\|_V^2 = 0$, since β is a positive real number and the norm squared can never be smaller than zero. Of course, $\|w\|_V^2 = 0$ implies that $\|w\|_V = 0$, and from this we conclude that $w = 0_V$, by one of the norm axioms. We have just proved that, if $w \in \text{Ker } L$, then $w = 0_V$, which is the same as saying that $\text{Ker } L = \{0_V\}$. So (A1.22) has been established as a truth, and consequently, the inverse operator L^{-1} exists.

We may inquire more about the inverse operator L^{-1} . We may ask: Is it linear and continuous (bounded)? In other words, is it true that $L^{-1} \in \mathcal{L}(V^*, V)$? Yes, it is true. To see that it is linear, we recall (A1.16) and notice that, given an arbitrary functional $g' \in V^*$, the action of the inverse operator is characterized by

$$w = L^{-1}g' \Leftrightarrow a(w, v) + \lambda_0(I_{V \rightarrow H}(w), I_{V \rightarrow H}(v))_H = \langle g', v \rangle_{V^*, V}, \quad \forall v \in V \quad (\text{A1.28})$$

So let us consider a functional $g'_1 \in V^*$. Then consider the problem of finding a w_1 such that

$$a(w_1, v) + \lambda_0(I_{V \rightarrow H}(w_1), I_{V \rightarrow H}(v))_H = \langle g'_1, v \rangle_{V^*, V}, \quad \forall v \in V \quad (\text{A1.29})$$

which according to (A1.28) is equivalent to $w_1 = L^{-1}g'_1$. Multiply (A1.29) by an arbitrary $\alpha_1 \in \mathbb{C}$ and get

$$a(\alpha_1 w_1, v) + \lambda_0(I_{V \rightarrow H}(\alpha_1 w_1), I_{V \rightarrow H}(v))_H = \langle \alpha_1 g'_1, v \rangle_{V^*, V}, \quad \forall v \in V, \quad (\text{A1.30})$$

since the sesquilinear form, the inner product, the embedding map and the duality pairing are all linear. According to (A1.28), this is equivalent to $\alpha_1 w_1 = L^{-1}\alpha_1 g'_1$. Take now another functional functional $g'_2 \in V^*$ and find a solution w_2 to the problem

$$a(w_2, v) + \lambda_0(I_{V \rightarrow H}(w_2), I_{V \rightarrow H}(v))_H = \langle g'_2, v \rangle_{V^*, V}, \quad \forall v \in V, \quad (\text{A1.31})$$

which is equivalent to $w_2 = L^{-1}g'_2$. Multiply (A1.31) by an arbitrary $\alpha_2 \in \mathbb{C}$ and get

$$a(\alpha_2 w_2, v) + \lambda_0(I_{V \rightarrow H}(\alpha_2 w_2), I_{V \rightarrow H}(v))_H = \langle \alpha_2 g'_2, v \rangle_{V^*, V}, \quad \forall v \in V, \quad (\text{A1.32})$$

which is equivalent to $\alpha_2 w_2 = L^{-1}\alpha_2 g'_2$. We now sum (A1.30) and (A1.32) and arrive at

$$\begin{aligned} a(\alpha_1 w_1 + \alpha_2 w_2, v) + \lambda_0(I_{V \rightarrow H}(\alpha_1 w_1 + \alpha_2 w_2), I_{V \rightarrow H}(v))_H = \\ \langle \alpha_1 g'_1 + \alpha_2 g'_2, v \rangle_{V^*, V}, \quad \forall v \in V \end{aligned} \quad (\text{A1.33})$$

which is equivalent to $\alpha_1 w_1 + \alpha_2 w_2 = L^{-1}(\alpha_1 g'_1 + \alpha_2 g'_2)$. But $w_1 = L^{-1}g'_1$ and $w_2 = L^{-1}g'_2$, so we finally get that

$$L^{-1}(\alpha_1 g'_1 + \alpha_2 g'_2) = \alpha_1 L^{-1}g'_1 + \alpha_2 L^{-1}g'_2 \quad (\text{A1.34})$$

Linearity of L^{-1} has been established. In order to find out if L^{-1} is continuous, we refer back to (A1.28) and begin by observing that

$$\left| a(w, v) + \lambda_0(I_{V \rightarrow H}(w), I_{V \rightarrow H}(v))_H \right| = |\langle g', v \rangle_{V^*, V}|, \quad \forall v \in V \quad (\text{A1.35})$$

By making $v = w$, it becomes

$$|a(w, w) + \lambda_0\|I_{V \rightarrow H}(w)\|_H^2| = |\langle g', w \rangle_{V^*, V}| \quad (\text{A1.36})$$

Since g' is an element from V^* , it is bounded, i.e., $\|g'\|_{V^*}$ is finite, and moreover,

$$|\langle g', w \rangle_{V^*, V}| \leq \|g'\|_{V^*} \|w\|_V \quad (A1.37)$$

Also, the real part of a complex number is smaller than or equal to its modulus, so we get

$$\operatorname{Re}\{a(w, w)\} + \lambda_0 \|I_{V \rightarrow H}(w)\|_H^2 \leq |a(w, w) + \lambda_0 \|I_{V \rightarrow H}(w)\|_H^2| \quad (A1.38)$$

From (A1.38), (A1.36) and (A1.37) we conclude that

$$\operatorname{Re}\{a(w, w)\} + \lambda_0 \|I_{V \rightarrow H}(w)\|_H^2 \leq \|g'\|_{V^*} \|w\|_V \quad (A1.39)$$

Hypothesis (iv) from Theorem 3.8 then reveals that

$$\beta \|w\|_V^2 \leq \|g'\|_{V^*} \|w\|_V \quad (A1.40)$$

or

$$\|w\|_V \leq \frac{1}{\beta} \|g'\|_{V^*}, \quad (A1.41)$$

According to (A1.28), $w = L^{-1}g'$, and g' is an arbitrary element from V^* . So it is true that

$$\|L^{-1}g'\|_V \leq \frac{1}{\beta} \|g'\|_{V^*}, \quad \forall g' \in V^*, \quad (A1.42)$$

which allows us to conclude that

$$\|L^{-1}\|_{\mathcal{L}(V^*, V)} := \sup_{g' \in V^* \setminus \{0\}} \frac{\|L^{-1}g'\|_V}{\|g'\|_{V^*}} \leq \frac{1}{\beta} < \infty, \quad (A1.43)$$

as $\beta > 0$. In this way, the continuity of L^{-1} has been established. Since L^{-1} is linear and continuous, $L^{-1} \in \mathcal{L}(V^*, V)$.

We now apply L^{-1} to (A1.21) and get the equivalent problem

$$\begin{aligned} & \text{Find } u \in V \text{ such that} \\ & u = L^{-1}F + \lambda_0 L^{-1} \circ J_{V \rightarrow V^*}(u), \quad \text{in } V \end{aligned} \quad (A1.44)$$

Problem (A1.44) can be rewritten as

$$\begin{aligned} & \text{Find } u \in V \text{ such that} \\ & (I_V - \lambda_0 L^{-1} \circ J_{V \rightarrow V^*})u = L^{-1}F, \quad \text{in } V \end{aligned} \quad (A1.45)$$

Since $J_{V \rightarrow V^*} \in \mathcal{K}(V, V^*)$, from (A1.5) and $L^{-1} \in \mathcal{L}(V^*, V)$, then $L^{-1} \circ J_{V \rightarrow V^*} \in \mathcal{K}(V, V)$, according to Theorem 3.5.

We are at a position to apply Theorem 3.6. When applied to the compact operator $\lambda_0 L^{-1} \circ J_{V \rightarrow V^*}$, it says that

$$\text{Ker}(I_V - \lambda_0 L^{-1} \circ J_{V \rightarrow V^*}) = \{0_V\} \Leftrightarrow R(I_V - \lambda_0 L^{-1} \circ J_{V \rightarrow V^*}) = V \quad (\text{A1.46})$$

We are particularly interested in the implication $\Rightarrow$, which says that

$$\text{If } \text{Ker}(I_V - \lambda_0 L^{-1} \circ J_{V \rightarrow V^*}) = \{0_V\} \text{ then } R(I_V - \lambda_0 L^{-1} \circ J_{V \rightarrow V^*}) = V \quad (\text{A1.47})$$

Expression above means that if the operator $I_V - \lambda_0 L^{-1} \circ J_{V \rightarrow V^*}$ is injective (one-to-one), then its range is the whole of V , i.e., the aforementioned operator is also surjective. Therefore injectivity implies surjectivity, or in other words, *uniqueness implies existence*. Let us characterize $\text{Ker}(I_V - \lambda_0 L^{-1} \circ J_{V \rightarrow V^*})$. From the definition of kernel:

$$\text{Ker}(I_V - \lambda_0 L^{-1} \circ J_{V \rightarrow V^*}) = \{u \in V \mid (I_V - \lambda_0 L^{-1} \circ J_{V \rightarrow V^*})u = 0\} \quad (\text{A1.48})$$

$$= \{u \in V \mid u - \lambda_0 L^{-1} \circ J_{V \rightarrow V^*} u = 0\} \quad (\text{A1.49})$$

$$= \{u \in V \mid u = \lambda_0 L^{-1} \circ J_{V \rightarrow V^*} u\} \quad (\text{A1.50})$$

Operating with $L \in \mathcal{L}(V, V^*)$ on both sides of (A1.50),

$$= \{u \in V \mid Lu = \lambda_0 J_{V \rightarrow V^*} u, \text{ in } V^*\} \quad (\text{A1.51})$$

$$= \{u \in V \mid \langle Lu, v \rangle_{V^*, V} = \lambda_0 \langle J_{V \rightarrow V^*} u, v \rangle_{V^*, V}, \forall v \in V\} \quad (\text{A1.52})$$

From the definition of $\langle Lu, v \rangle_{V^*, V}$ in (A1.16), from Theorem 3.7 [by making $w = I_{V \rightarrow H}(u)$], and from $J_{V \rightarrow V^*}$ in (A1.5) we arrive at

$$= \left\{ u \in V \mid a(u, v) + \lambda_0 (I_{V \rightarrow H}(u), I_{V \rightarrow H}(v))_H = \lambda_0 (I_{V \rightarrow H}(u), I_{V \rightarrow H}(v))_H, \forall v \in V \right\} \quad (\text{A1.53})$$

Finally,

$$= \{u \in V \mid a(u, v) = 0, \forall v \in V\} \quad (\text{A1.54})$$

Let us now characterize $R(I_V - \lambda_0 L^{-1} \circ J_{V \rightarrow V^*})$ in (A1.47). When we say that $R(I_V - \lambda_0 L^{-1} \circ J_{V \rightarrow V^*}) = V$, it means that

$$\forall z \in V \exists u \in V (I_V - \lambda_0 L^{-1} \circ J_{V \rightarrow V^*})u = z \quad (\text{A1.55})$$

$$\forall z \in V \exists u \in V u - \lambda_0 L^{-1} \circ J_{V \rightarrow V^*} u = z \quad (\text{A1.56})$$

Operating with $L \in \mathcal{L}(V, V^*)$ on both sides of (A1.56),

$$\forall z \in V \exists u \in V Lu - \lambda_0 J_{V \rightarrow V^*} u = Lz, \text{ in } V^*. \quad (\text{A1.57})$$

$$\forall z \in V \exists u \in V \langle Lu, v \rangle_{V^*, V} - \langle J_{V \rightarrow V^*} u, v \rangle_{V^*, V} = \langle Lz, v \rangle_{V^*, V}, \forall v \in V \quad (\text{A1.58})$$

From the definition of $\langle Lu, v \rangle_{V^*, V}$ in (A1.16), from Theorem 3.7 [by making $w = I_{V \rightarrow H}(u)$], and from $J_{V \rightarrow V^*}$ in (A1.5) we arrive at

$$\forall z \in V \quad \exists u \in V \quad a(u, v) = \langle Lz, v \rangle_{V^*, V}, \quad \forall v \in V. \quad (\text{A1.59})$$

Let $F \in V^*$ be arbitrary. Then $L^{-1}F \in V$. If we make $z = L^{-1}F$ in (A1.59), we see that

$$\exists u \in V \quad a(u, v) = \langle L \circ L^{-1}F, v \rangle_{V^*, V}, \quad \forall v \in V. \quad (\text{A1.60})$$

Since $L \circ L^{-1}$ is just the identity operator on V^* and $F \in V^*$ is arbitrary, we get

$$\forall F \in V^* \quad \exists u \in V \quad a(u, v) = \langle F, v \rangle_{V^*, V}, \quad \forall v \in V. \quad (\text{A1.61})$$

This is the conclusion we get from the fact that $R(I_V - \lambda_0 L^{-1} \circ J_{V \rightarrow V^*}) = V$.

The main result we proved was (A1.47). We have shown that the kernel which appears at the left side from (A1.47) is given by (A1.54), whereas the conclusion $R(I_V - \lambda_0 L^{-1} \circ J_{V \rightarrow V^*}) = V$ implies (A1.61). Graphically,

$$\begin{array}{ccc} \underbrace{\text{Ker}(I_V - \lambda_0 L^{-1} \circ J_{V \rightarrow V^*}) = \{0_V\}} & \Rightarrow & \underbrace{R(I_V - \lambda_0 L^{-1} \circ J_{V \rightarrow V^*}) = V} \\ \parallel & & \Downarrow \\ \{u \in V \mid a(u, v) = 0, \forall v \in V\} & & (\text{A1.61}) \end{array} \quad (\text{A1.62})$$

Saying that $\{u \in V \mid a(u, v) = 0, \forall v \in V\}$ is equal to $\{0_V\}$ is just to state that the solution to the homogeneous (zero-data) problem

$$\begin{array}{l} \text{Find } u \in V \text{ such that} \\ a(u, v) = 0, \quad \forall v \in V \end{array} \quad (\text{A1.63})$$

is the zero element $u = 0_V$. And (A1.61) is equivalent to saying that the solution to the general problem

$$\begin{array}{l} \text{Find } u \in V \text{ such that} \\ a(u, v) = \langle F, v \rangle_{V^*, V}, \quad \forall v \in V \end{array} \quad (\text{A1.64})$$

exists for every functional $F \in V^*$.

Therefore, if we prove that the solution to the homogeneous problem (A1.63) is 0_V (which is the same as proving the uniqueness of an eventual solution), it automatically follows that the solution to the general problem (A1.64) does indeed exist for any ‘source’ F . Again, injectivity implies surjectivity. If we prove injectivity, then we get injectivity plus surjectivity, which is a very positive scenario.

And thus, conclusion (a) from Theorem 3.8 is proved.

Part II: Boundedness (Continuity)

We now suppose that the solution to the homogeneous problem (3.63) is the zero element $u = 0_V$, so that we know that the solution to the general problem (3.64) exists and is unique for any $F \in V^*$.

The claim that such a solution depends continuously on the data $F \in V^*$ means that

$$\exists C_{FA} > 0 \quad \forall F \in V^* \quad \|I_{V \rightarrow H}(u)\|_H \leq C_{FA} \|F\|_{V^*}, \quad (A1.65)$$

where u is the solution to

$$\begin{aligned} & \text{Find } u \in V \text{ such that} \\ & a(u, v) = \langle F, v \rangle_{V^*, V}, \quad \forall v \in V \end{aligned} \quad (A1.66)$$

In (A1.65), C_{FA} is a positive constant independent of u . Suppose (A1.65) is not true. Then the negation of (A1.65) is

$$\forall C_{FA} > 0 \quad \exists F \in V^* \quad \|I_{V \rightarrow H}(u)\|_H > C_{FA} \|F\|_{V^*}. \quad (A1.67)$$

It means that for any choice of a positive C_{FA} , there is a corresponding $F \in V^*$ such that $\|I_{V \rightarrow H}(u)\|_H > C_{FA} \|F\|_{V^*}$, where u is the solution to (A1.66). Now take successively $C_{FA} = 1, 2, 3, \dots, n, \dots$, i.e., we consider each natural number as a choice for C_{FA} . We deduce the existence of a sequence of functionals $\{F_n\}_{n=1}^\infty \subset V^*$ such that for each F_n ,

$$\|I_{V \rightarrow H}(u_n)\|_H > n \|F_n\|_{V^*} \quad (A1.68)$$

where u_n is the solution to

$$\begin{aligned} & \text{Find } u_n \in V \text{ such that} \\ & a(u_n, v) = \langle F_n, v \rangle_{V^*, V}, \quad \forall v \in V. \end{aligned} \quad (A1.69)$$

So we have got first a sequence of functionals $\{F_n\}_{n=1}^\infty \subset V^*$, which produces a sequence $\{u_n\}_{n=1}^\infty \subset V$, which finally produces another sequence $\{I_{V \rightarrow H}(u_n)\}_{n=1}^\infty \subset H$, whose elements are related to the elements of the original sequence of functionals through (A1.68).

If any member F_n is multiplied by a scalar γ , then u_n is also multiplied by γ , since (A1.69) is a linear problem. As the embedding map $I_{V \rightarrow H}$ is also linear, $I_{V \rightarrow H}(\gamma u_n)$ becomes $\gamma I_{V \rightarrow H}(u_n)$.

For each $n \in \mathbb{N}$, we multiply F_n by the inverse of $\|I_{V \rightarrow H}(u_n)\|_H$, i.e., we form a new sequence of functionals $\{G_n\}_{n=1}^\infty \subset V^*$, where

$$G_n = \frac{1}{\|I_{V \rightarrow H}(u_n)\|_H} F_n. \quad (A1.70)$$

The sequence of functionals $\{G_n\}_{n=1}^\infty$ obviously produces a new sequence $\{w_n\}_{n=1}^\infty \subset V$, where w_n is the solution to

Find $w_n \in V$ such that

$$a(w_n, v) = \langle G_n, v \rangle_{V^*, V}, \quad \forall v \in V. \quad (A1.71)$$

Of course, w_n is just u_n divided by $\|I_{V \rightarrow H}(u_n)\|_H$. This sequence $\{w_n\}_{n=1}^\infty$ finally produces a sequence $\{I_{V \rightarrow H}(w_n)\}_{n=1}^\infty \subset H$, for which

$$I_{V \rightarrow H}(w_n) = I_{V \rightarrow H} \left(\frac{u_n}{\|I_{V \rightarrow H}(u_n)\|_H} \right) = \frac{1}{\|I_{V \rightarrow H}(u_n)\|_H} I_{V \rightarrow H}(u_n) \quad (A1.72)$$

Then, for each n ,

$$\|I_{V \rightarrow H}(w_n)\|_H = \frac{1}{\|I_{V \rightarrow H}(u_n)\|_H} \|I_{V \rightarrow H}(u_n)\|_H > \frac{1}{\|I_{V \rightarrow H}(u_n)\|_H} n \|F_n\|_{V^*}, \quad (A1.73)$$

where the inequality came from (A1.68). Expression above gets simplified to

$$\|I_{V \rightarrow H}(w_n)\|_H = 1 > \frac{1}{\|I_{V \rightarrow H}(u_n)\|_H} n \|F_n\|_{V^*}. \quad (A1.74)$$

But

$$\frac{1}{\|I_{V \rightarrow H}(u_n)\|_H} \|F_n\|_{V^*} = \left\| \frac{1}{\|I_{V \rightarrow H}(u_n)\|_H} \right\|_{V^*} = \|G_n\|_{V^*}, \quad (A1.75)$$

which allows (A1.74) to become

$$\|I_{V \rightarrow H}(w_n)\|_H = 1 > n \|G_n\|_{V^*}, \quad \forall n \in \mathbb{N} \quad (A1.76)$$

The great conclusion thus far amounts to this: We have got a new sequence of functionals $\{G_n\}_{n=1}^\infty \subset V^*$, which produces a sequence $\{w_n\}_{n=1}^\infty \subset V$ through (A1.71). Moreover, this sequence in V produces a sequence $\{I_{V \rightarrow H}(w_n)\}_{n=1}^\infty \subset H$ such that *each of its terms has unit norm*, according to (A1.76).

Expression (A1.76) reveals a striking fact:

$$\|G_n\|_{V^*} < \frac{1}{n}, \quad \forall n \in \mathbb{N}, \quad (A1.77)$$

which implies that

$$\lim_{n \rightarrow \infty} \|G_n\|_{V^*} = 0. \quad (A1.78)$$

Now we claim that the sequence $\{w_n\}_{n=1}^\infty \subset V$ is bounded. We have, for each $n \in \mathbb{N}$:

$$a(w_n, v) = \langle G_n, v \rangle_{V^*, V}, \quad \forall v \in V. \quad (A1.79)$$

When we add the same quantity to both sides, it becomes:

$$(A1.80)$$

$$a(w_n, v) + \lambda_0(I_{V \rightarrow H}(w_n), I_{V \rightarrow H}(v))_H = \langle G_n, v \rangle_{V^*, V} + \lambda_0(I_{V \rightarrow H}(w_n), I_{V \rightarrow H}(v))_H \quad \forall v \in V$$

Now make $v = w_n$ and take the absolute value of each side:

$$|a(w_n, w_n) + \lambda_0\|I_{V \rightarrow H}(w_n)\|_H^2| = |\langle G_n, w_n \rangle_{V^*, V} + \lambda_0\|I_{V \rightarrow H}(w_n)\|_H^2| \quad (A1.81)$$

Since the real part of a complex number is smaller than or equal to its modulus, we get

$$\operatorname{Re}\{a(w_n, w_n)\} + \lambda_0\|I_{V \rightarrow H}(w_n)\|_H^2 \leq |\langle G_n, w_n \rangle_{V^*, V} + \lambda_0\|I_{V \rightarrow H}(w_n)\|_H^2| \quad (A1.82)$$

Hypothesis (iv) in Theorem 3.8 allows us to write

$$\beta\|w_n\|_V^2 \leq |\langle G_n, w_n \rangle_{V^*, V} + \lambda_0\|I_{V \rightarrow H}(w_n)\|_H^2| \leq \|G_n\|_{V^*}\|w_n\|_V + \lambda_0, \quad (A1.83)$$

where the triangle inequality and the fact that $\|I_{V \rightarrow H}(w_n)\|_H = 1$ in (A1.76) have been used. We get

$$\beta\|w_n\|_V^2 \leq \|G_n\|_{V^*}\|w_n\|_V + \lambda_0. \quad (A1.84)$$

which is rewritten as

$$\beta\|w_n\|_V^2 - \|G_n\|_{V^*}\|w_n\|_V - \lambda_0 \leq 0, \quad (A1.85)$$

which is a standard quadratic inequality, whose solution is

$$\frac{\|G_n\|_{V^*} - \sqrt{\|G_n\|_{V^*}^2 + 4\beta\lambda_0}}{2\beta} \leq \|w_n\|_V \leq \frac{\|G_n\|_{V^*} + \sqrt{\|G_n\|_{V^*}^2 + 4\beta\lambda_0}}{2\beta} \quad (A1.86)$$

(Remember that it is always true that $\|w_n\|_V \geq 0$.) By concentrating on the right side of (A1.86) and observing that $\|G_n\|_{V^*}^2 + 4\beta\lambda_0 \leq (\|G_n\|_{V^*} + 2\sqrt{\beta\lambda_0})^2$, we get

$$\|w_n\|_V \leq \frac{\|G_n\|_{V^*} + \sqrt{\beta\lambda_0}}{\beta}, \quad \forall n \in \mathbb{N} \quad (A1.87)$$

Since all G_n are in V^* , they are *bounded* linear functionals, and therefore are finite. Moreover, according to (A1.78), the sequence $\{\|G_n\|_{V^*}\}_{n=1}^\infty \subset \mathbb{R}^+$ is convergent. It is a known fact that convergent sequences are bounded [Kreyszig, 1989], so there is a constant M such that $\|G_n\|_{V^*} \leq M$, for any n [according to the definition (3.57)]. We can go further and see that this constant is 1, from (A1.77). So

$$\|w_n\|_V \leq \frac{1 + \sqrt{\beta\lambda_0}}{\beta}, \quad \forall n \in \mathbb{N} \quad (A1.88)$$

which is the same as saying that the sequence $\{w_n\}_{n=1}^\infty$ is bounded.

In order to proceed, we need two theorems concerning compact operators in Hilbert spaces [Salsa, 2008]. For the notion of weak convergence, see [Brezis, 2010].

Theorem A1.1: Convergent subsequences – Let $\mathcal{H}$ be a Hilbert space. If a sequence $\{z_n\}_{n=1}^{\infty} \subset \mathcal{H}$ is bounded, then $\{z_n\}_{n=1}^{\infty}$ admits a subsequence $\{z_{n_j}\}_{j=1}^{\infty} \subset \{z_n\}_{n=1}^{\infty}$ which converges weakly to an element $z^0 \in \mathcal{H}$, i.e.,

$$z_{n_j} \rightharpoonup z^0 \text{ in } \mathcal{H}. \quad (\text{A1.89})$$

A nice property of compact operators is that they convert weakly convergent sequences into strongly convergent sequences. This is stated in the next theorem.

Theorem A1.2: From weak to strong – Suppose $\mathcal{H}_1$ and $\mathcal{H}_2$ are two Hilbert spaces, and let $T \in \mathcal{L}(\mathcal{H}_1, \mathcal{H}_2)$. Let $\{z_n\}_{n=1}^{\infty}$ be an arbitrary sequence in $\mathcal{H}_1$. Then

$$T \in \mathcal{K}(\mathcal{H}_1, \mathcal{H}_2) \Leftrightarrow ((z_n \rightharpoonup z^0 \text{ in } \mathcal{H}_1) \text{ implies } (Tz_n \rightarrow Tz^0 \text{ in } \mathcal{H}_2)) \quad (\text{A1.90})$$

Applying Theorem A1.1 to the bounded sequence $\{w_n\}_{n=1}^{\infty}$ in V allows us to conclude that there is a subsequence $\{w_{n_j}\}_{j=1}^{\infty} \subset \{w_n\}_{n=1}^{\infty}$ such that

$$w_{n_j} \rightharpoonup w^0 \text{ in } V. \quad (\text{A1.91})$$

Hypothesis (ii) from Theorem 3.8 gives us that the embedding map $I_{V \rightarrow H}$ is compact, i.e., $I_{V \rightarrow H} \in \mathcal{K}(V, H)$. Theorem A1.2 therefore says that

$$I_{V \rightarrow H} w_{n_j} \rightarrow I_{V \rightarrow H} w^0 \text{ in } H. \quad (\text{A1.92})$$

The question is that, as we let $j \rightarrow \infty$ (and consequently $n_j \rightarrow \infty$), problem (A1.71) becomes

$$\begin{aligned} & \text{Find } w_{n_j} \in V \text{ such that} \\ & a(w_{n_j}, v) = \langle G_{n_j}, v \rangle_{V^*, V}, \quad \forall v \in V. \end{aligned} \quad (\text{A1.93})$$

Since the sesquilinear form a is continuous, for any $v \in V$, it induces a functional $A^T v \in V^*$ whose action on w_{n_j} is given by

$$a(w_{n_j}, v) =: \langle w_{n_j}, A^T v \rangle_{V, V^*}, \quad (\text{A1.94})$$

which allows us to write (A1.93) as

$$\begin{aligned} & \text{Find } w_{n_j} \in V \text{ such that} \\ & \langle w_{n_j}, A^T v \rangle_{V, V^*} = \langle G_{n_j}, v \rangle_{V^*, V}, \quad \forall v \in V. \end{aligned} \quad (\text{A1.95})$$

Since $w_{n_j} \rightharpoonup w^0$, according to (A1.91), and G_n converges (strongly) to 0_{V^*} , problem above becomes

$$\begin{aligned} & \text{Find } w^0 \in V \text{ such that} \\ & \langle w^0, A^T v \rangle_{V, V^*} = 0, \quad \forall v \in V. \end{aligned} \quad (A1.96)$$

Or, from (A1.94),

$$\begin{aligned} & \text{Find } w^0 \in V \text{ such that} \\ & a(w^0, v) = 0, \quad \forall v \in V. \end{aligned} \quad (A1.97)$$

But we supposed from the outset that the solution to the homogeneous problem (3.63) is 0_V , so we conclude that $w^0 = 0_V$.

From (A1.92) we get that

$$I_{V \rightarrow H} w_{n_j} \rightarrow I_{V \rightarrow H} w^0 = I_{V \rightarrow H} 0_V = 0_H \quad \text{in } H, \quad (A1.98)$$

i.e., the subsequence $\{I_{V \rightarrow H} w_{n_j}\}_{j=1}^{\infty}$ converges to 0_H .

On the other hand, (A1.76) says that

$$\|I_{V \rightarrow H}(w_n)\|_H = 1, \quad \forall n \in \mathbb{N}, \quad (A1.99)$$

i.e., all elements from the sequence $\{I_{V \rightarrow H}(w_n)\}_{n=1}^{\infty}$ have unit norm. Therefore, all elements from the subsequence $\{I_{V \rightarrow H} w_{n_j}\}_{j=1}^{\infty}$ also have unit norm.

We say that $I_{V \rightarrow H} w_{n_j} \rightarrow 0_H$ in H if $\|I_{V \rightarrow H} w_{n_j} - 0_H\|_H \rightarrow 0$, which implies that $\|I_{V \rightarrow H} w_{n_j}\|_H \rightarrow 0$. But from (A1.99), that does not happen, so the subsequence $\{I_{V \rightarrow H} w_{n_j}\}_{j=1}^{\infty}$ does not converge to zero. We have just arrived at a contradiction, so (A1.67) is false, and consequently, (A1.65) is true.

The solution u to (A1.66) does depend continuously on the data, with respect to the norm in the Hilbert space H .

Appendix 2

Theorem 3.9

In the mixed formulation resulting from our electromagnetic wave scattering problem, the sesquilinear form a is *not coercive*, contrary to what happens in a large number of problems from mechanics. If we are to propose a meshfree method based on this formulation, we are compelled to show first that it is indeed well-posed. We therefore construct an adaptation of the theory of mixed formulations in which the Fredholm Alternative is taken into account. The result is stated in the theorem below.

Theorem 3.9: Well-posedness of mixed formulations, non-coercive case – Let $\mathcal{X}$ and $\mathcal{Y}$ be two Hilbert spaces, and let $a: \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{C}$ and $b: \mathcal{X} \times \mathcal{Y} \rightarrow \mathbb{C}$ be two continuous sesquilinear forms, i.e., there are positive constants α_a and α_b such that:

(i) a is continuous, i.e.,

$$|a(x, v)| \leq \alpha_a \|x\|_{\mathcal{X}} \|v\|_{\mathcal{X}}, \quad \forall x, v \in \mathcal{X} \quad (3.69.a)$$

(ii) b is continuous, i.e.,

$$|b(x, y)| \leq \alpha_b \|x\|_{\mathcal{X}} \|y\|_{\mathcal{Y}}, \quad \forall x \in \mathcal{X} \quad \forall y \in \mathcal{Y} \quad (3.69.b)$$

Let $\mathcal{X}^0$ be the kernel of the sesquilinear form b i.e.,

$$\mathcal{X}^0 = \text{Ker } b = \{x \in \mathcal{X} \mid b(x, y) = 0, \quad \forall y \in \mathcal{Y}\}. \quad (3.69.c)$$

Consider a third Hilbert space H such that $\mathcal{X}^0$ and H satisfy the requirements of Theorem 3.7, i.e.,

(iii) $\mathcal{X}^0$ is continuously embedded into H , i.e., $\mathcal{X}^0 \hookrightarrow H$.

Moreover, it holds that:

(iv) The map $I_{\mathcal{X}^0 \rightarrow H}$ is compact, i.e., $I_{\mathcal{X}^0 \rightarrow H} \in \mathcal{K}(\mathcal{X}^0, H)$.

(v) The sesquilinear form a satisfies the following property on the kernel $\mathcal{X}^0$: There exist constants $\eta > 0$ and $\kappa_0 \geq 0$ such that

$$\text{Re}\{a(u, u)\} + \kappa_0 \|I_{\mathcal{X}^0 \rightarrow H}(u)\|_H^2 \geq \eta \|u\|_{\mathcal{X}}^2, \quad \forall u \in \mathcal{X}^0. \quad (3.69.d)$$

(vi) The sesquilinear form b satisfies the inf-sup condition, i.e., there is a positive constant $\beta_b > 0$ such that

$$\inf_{y \in \mathcal{Y} \setminus \{0\}} \sup_{x \in \mathcal{X} \setminus \{0\}} \frac{|b(x, y)|}{\|x\|_{\mathcal{X}} \|y\|_{\mathcal{Y}}} \geq \beta_b. \quad (3.69.e)$$

(vii) The solution to the homogeneous (zero-data) problem at the kernel $\mathcal{X}^0$

Find $w \in \mathcal{X}^0$ such that

$$a(w, v) = 0, \quad \forall w \in \mathcal{X}^0 \quad (3.69.f)$$

is the zero element $w = 0$. Furthermore, let us assume that:

(viii) The original space $\mathcal{X}$ is also continuously embedded H , i.e., $\mathcal{X} \hookrightarrow H$.

(ix) The spaces $\mathcal{X}$ and $\mathcal{X}^0$ are subspaces of H , i.e., $\mathcal{X} \subset H$ and $\mathcal{X}^0 \subset H$ (which implies that $I_{\mathcal{X} \rightarrow H}$ and $I_{\mathcal{X}^0 \rightarrow H}$ are inclusion maps).

Then it can be concluded that for each $f^* \in \mathcal{X}^*$ and $g^* \in \mathcal{Y}^*$, there is a unique solution to the mixed problem

Find $(u, p) \in \mathcal{X} \times \mathcal{Y}$ such that

$$a(u, x) + b(x, p) = \langle f^*, x \rangle_{\mathcal{X}^*, \mathcal{X}} \quad \forall x \in \mathcal{X} \quad (3.69.g)$$

$$b(u, y) = \langle g^*, y \rangle_{\mathcal{Y}^*, \mathcal{Y}} \quad \forall y \in \mathcal{Y}$$

It also follows that the solution u depends continuously on the data f^* and g^* in the H norm, i.e., there are positive constants K_1 and K_2 such that

$$\|u\|_H \leq K_1 \|f^*\|_{\mathcal{X}^*} + K_2 \|g^*\|_{\mathcal{Y}^*} \quad (3.69.h)$$

Proof: Consider problem (3.69.g), for which the sesquilinear forms a and b obey requirements (i) and (ii), respectively, and let $f^* \in \mathcal{X}^*$ and $g^* \in \mathcal{Y}^*$ be arbitrary functionals.

Part I – Existence

The inf-sup condition from requirement (vi) holds; and we know from conclusion (iii) in Theorem 3.3 that such a condition is equivalent to the fact that operator $B: \mathcal{X} \rightarrow \mathcal{Y}^*$ is surjective. If we write (3.69.g) in the operator form (3.32), then we see that $Bu = g^*$. But since the operator B is surjective, there exists an element u_g from $\mathcal{X}$ such that $Bu_g = g^*$.

We also know from conclusion (v) in Theorem 3.3 that the inf-sup condition (3.69.e) is equivalent to

$$\sup_{y \in \mathcal{Y} \setminus \{0\}} \frac{\langle Bx, y \rangle_{\mathcal{Y}^*, \mathcal{Y}}}{\|y\|_{\mathcal{Y}}} \geq \beta_b \|x\|_{\mathcal{X}}, \quad \forall x \in (\mathcal{X}^0)^\perp, \quad (A2.1)$$

In the most general case, the functional g^* can be any element from $\mathcal{Y}^*$. (We assume that it is different from zero; otherwise, we can jump to (A2.8) and make $u_g = 0$ there.) Since $Bu_g = g^*$, it means that

$$b(u_g, y) = \langle g^*, y \rangle_{\mathcal{Y}^*, \mathcal{Y}} \quad \forall y \in \mathcal{Y}, \quad (A2.2)$$

and consequently $u_g \notin \mathcal{X}^0 = \text{Ker } b$, defined in (3.69.c). The Hilbert space $\mathcal{X}$ can be decomposed as $\mathcal{X} = \mathcal{X}^0 \oplus (\mathcal{X}^0)^\perp$, because $\mathcal{X}^0$ is a null-space and null-spaces are closed [Kreyszig]. Since $u_g \notin \mathcal{X}^0$, then $u_g \in (\mathcal{X}^0)^\perp$. We now make $x = u_g$ in (A2.1) and observe that

$$\sup_{y \in \mathcal{Y} \setminus \{0\}} \frac{\langle g^*, y \rangle_{\mathcal{Y}^*, \mathcal{Y}}}{\|y\|_{\mathcal{Y}}} = \sup_{y \in \mathcal{Y} \setminus \{0\}} \frac{\langle Bu_g, y \rangle_{\mathcal{Y}^*, \mathcal{Y}}}{\|y\|_{\mathcal{Y}}} \geq \beta_b \|u_g\|_{\mathcal{X}}, \quad (A2.3)$$

The leftmost supremum in (A2.3) is just the norm of the functional g^* , so we conclude that

$$\|u_g\|_{\mathcal{X}} \leq \frac{1}{\beta_b} \|g^*\|_{\mathcal{Y}^*}. \quad (A2.4)$$

We now write the original solution u as

$$u = u^s + u_g. \quad (A2.5)$$

When we substitute (A2.5) into the original system (3.69.g), we find that u^s is the true unknown:

Find $(u^s, p) \in \mathcal{X} \times \mathcal{Y}$ such that

$$a(u^s, x) + b(x, p) = \langle f^*, x \rangle_{\mathcal{X}^*, \mathcal{X}} - a(u_g, x), \quad \forall x \in \mathcal{X} \quad (A2.6)$$

$$b(u^s, y) = 0, \quad \forall y \in \mathcal{Y}$$

Since the sesquilinear form a is continuous, it is not difficult to see that $a(u_g, \cdot)$ defines a bounded and linear functional on $\mathcal{X}$, i.e., $a(u_g, \cdot) \in \mathcal{X}^*$. We may write this as

$$a(u_g, x) =: \langle Au_g, x \rangle_{\mathcal{X}^*, \mathcal{X}}, \quad \forall x \in \mathcal{X}. \quad (A2.7)$$

Consequently, $Au_g \in \mathcal{X}^*$ and (A2.6) assumes the form

Find $(u^s, p) \in \mathcal{X} \times \mathcal{Y}$ such that

$$a(u^s, x) + b(x, p) = \langle f^* - Au_g, x \rangle_{\mathcal{X}^*, \mathcal{X}}, \quad \forall x \in \mathcal{X} \quad (A2.8)$$

$$b(u^s, y) = 0, \quad \forall y \in \mathcal{Y}$$

From the second equation in (A2.8), we learn that $u^s \in \mathcal{X}^0$. Since $\mathcal{X}^0 \subset \mathcal{X}$, the first equation in (A2.8) is of course valid when the test functions are taken from $\mathcal{X}^0$. In other words, we can restrict the problem (A2.8) to $\mathcal{X}^0$ and get

Find $u^s \in \mathcal{X}^0$ such that

$$a(u^s, x) + b(x, p) = \langle f^* - Au_g, x \rangle_{\mathcal{X}^*, \mathcal{X}}, \quad \forall x \in \mathcal{X}^0. \quad (A2.9)$$

Since $x \in \mathcal{X}^0 = \text{Ker } b$, according to (3.69.c), $b(x, y) = 0$ for any $y \in \mathcal{Y}$. But $p \in \mathcal{Y}$, and then our problem becomes

$$\begin{aligned} & \text{Find } u^s \in \mathcal{X}^0 \text{ such that} \\ & a(u^s, x) = \langle f^* - Au_g, x \rangle_{\mathcal{X}^*, \mathcal{X}}, \quad \forall x \in \mathcal{X}^0. \end{aligned} \quad (\text{A2.10})$$

There is just a small technicality: Since $\mathcal{X}^0 \subset \mathcal{X}$, then it is obvious that $\mathcal{X}^* \subset (\mathcal{X}^0)^*$, i.e., bounded linear functionals acting on the whole space $\mathcal{X}$, when restricted in their action to the subspace $\mathcal{X}^0$, also define functionals on $\mathcal{X}^0$. So there is no harm in writing (A2.10) in a slightly modified form:

$$\begin{aligned} & \text{Find } u^s \in \mathcal{X}^0 \text{ such that} \\ & a(u^s, x) = \langle f^* - Au_g, x \rangle_{(\mathcal{X}^0)^*, \mathcal{X}^0}, \quad \forall x \in \mathcal{X}^0, \end{aligned} \quad (\text{A2.11})$$

where

$$\langle f^* - Au_g, x \rangle_{(\mathcal{X}^0)^*, \mathcal{X}^0} := \langle f^* - Au_g, x \rangle_{\mathcal{X}^*, \mathcal{X}}, \quad \forall x \in \mathcal{X}^0. \quad (\text{A2.12})$$

This is the point at which hypotheses (iii), (iv) and (v) from our Theorem 3.9 play their role in the solvability of problem (A2.11). We assumed in hypothesis (vii) that the solution to the homogeneous (zero-data) problem

$$\begin{aligned} & \text{Find } w \in \mathcal{X}^0 \text{ such that} \\ & a(w, x) = 0, \quad \forall x \in \mathcal{X}^0 \end{aligned} \quad (\text{A2.13})$$

is the zero element $0_{\mathcal{X}^0} = 0_{\mathcal{X}}$. Then it follows, via Theorem 3.8, that the solution u^s to (A2.11) exists and is unique. Moreover, it holds the estimate

$$\|I_{\mathcal{X}^0 \rightarrow H}(u^s)\|_H \leq C_{FA} \|f^* - Au_g\|_{(\mathcal{X}^0)^*}, \quad (\text{A2.14})$$

i.e., the element u^s measured in the norm of the Hilbert space H depends on the functionals f^* and Au_g . Moreover,

$$\|f^* - Au_g\|_{(\mathcal{X}^0)^*} = \sup_{x \in \mathcal{X}^0 \setminus \{0\}} \frac{\langle f^* - Au_g, x \rangle_{(\mathcal{X}^0)^*, \mathcal{X}^0}}{\|x\|_{\mathcal{X}^0}} \quad (\text{A2.15})$$

$$= \sup_{x \in \mathcal{X}^0 \setminus \{0\}} \frac{\langle f^* - Au_g, x \rangle_{\mathcal{X}^*, \mathcal{X}}}{\|x\|_{\mathcal{X}}} \quad (\text{A2.16})$$

$$\leq \sup_{x \in \mathcal{X} \setminus \{0\}} \frac{\langle f^* - Au_g, x \rangle_{\mathcal{X}^*, \mathcal{X}}}{\|x\|_{\mathcal{X}}} \quad (\text{A2.17})$$

$$= \|f^* - Au_g\|_{\mathcal{X}^*} \quad (\text{A2.18})$$

The move from (A2.15) to (A2.16) is justified by (A2.12) and the due to the fact that $\|\cdot\|_{\mathcal{X}^0} = \|\cdot\|_{\mathcal{X}}$, since $\mathcal{X}^0 \subset \mathcal{X}$. Since the supremum over a subspace is smaller than or equal to the supremum over the whole space, (A2.17) follows from (A2.16). Finally, (A2.18) is just the ordinary definition of the norm of a functional on $\mathcal{X}$. The estimate in (A2.14) is modified into

$$\|I_{\mathcal{X}^0 \rightarrow H}(u^s)\|_H \leq C_{FA} \|f^* - Au_g\|_{\mathcal{X}^*}. \quad (\text{A2.19})$$

We now get back to (A2.10); with the help of (A2.7) and (A2.5), it becomes

$$a(u, x) - \langle f^*, x \rangle_{\mathcal{X}^*, \mathcal{X}} = 0, \quad \forall x \in \mathcal{X}^0. \quad (\text{A2.20})$$

(Notice that we no longer use the phrase ‘Find $u^s \in \mathcal{X}^0$ such that’ because the existence of both u^s and u_g have already been established.) Thanks to the continuity of the sesquilinear form a , it is not difficult to see that $a(u, \cdot) - f^*$ defines a linear and bounded functional on $\mathcal{X}^0$, i.e., $a(u, \cdot) - f^* \in (\mathcal{X}^0)^*$. Make

$$a(u, \cdot) - f^* = F^* \text{ in } (\mathcal{X}^0)^*, \quad (\text{A2.21})$$

i.e.,

$$a(u, x) - \langle f^*, x \rangle_{\mathcal{X}^*, \mathcal{X}} = \langle F^*, x \rangle_{\mathcal{X}^*, \mathcal{X}}, \quad \forall x \in \mathcal{X}^0 \quad (\text{A2.22})$$

Since according to (A2.20) the action of this functional is zero on all elements from $\mathcal{X}^0$, it follows that

$$F^* \in \mathcal{X}_\#^0, \quad (\text{A2.23})$$

i.e., this functional belongs to the annihilator of $\mathcal{X}^0$. From (3.69.c), we get that

$$\mathcal{X}^0 = \text{Ker } b = \{x \in \mathcal{X} \mid b(x, y) = 0, \quad \forall y \in \mathcal{Y}\} \quad (\text{A2.24})$$

$$= \{x \in \mathcal{X} \mid \langle Bx, y \rangle_{\mathcal{Y}^*, \mathcal{Y}} = 0, \quad \forall y \in \mathcal{Y}\} \quad (\text{A2.25})$$

$$= \{x \in \mathcal{X} \mid Bx = 0_{\mathcal{Y}^*}\} \quad (\text{A2.26})$$

$$= \text{Ker } B \quad (\text{A2.27})$$

Since $\mathcal{X}^0 = \text{Ker } B$, from (A2.23) we learn that

$$F^* \in (\text{Ker } B)_\# \quad (\text{A2.28})$$

We know from (3.26) that $B: \mathcal{X} \rightarrow \mathcal{Y}^*$. The space $\mathcal{X}$ is a Hilbert space, and therefore a Banach space. Since $\mathcal{Y}$ is a Hilbert space, its dual $\mathcal{Y}^*$ is also a Hilbert space [Kreyszig, 1989]. Consequently $\mathcal{Y}^*$ is a Banach space. Moreover, due to the continuity of the sesquilinear form b , it is not difficult to see that $B \in \mathcal{L}(\mathcal{X}, \mathcal{Y}^*)$.

We may then apply Theorem 3.10 to the operator B and conclude that

$$R(B^T) = (Ker B)_\# . \quad (A2.29)$$

From (A2.28) and (A2.29), we observe that

$$F^* \in R(B^T). \quad (A2.30)$$

Also, from (3.28) we learn that $B^T: \mathcal{Y} \rightarrow \mathcal{X}^*$ is a linear transformation (as b is a sesquilinear form). It is known that the range of linear transformations is a linear space in itself [Kreyszig, 1989]. So if (A2.30) is true, than

$$-F^* \in R(B^T). \quad (A2.31)$$

The meaning of (A2.31) is twofold. First, the functional $F^* \in (\mathcal{X}^0)^*$ in (A2.21) is also in $\mathcal{X}^*$. Since $\mathcal{X}^0 \subset \mathcal{X}$, then $\mathcal{X}^* \subset (\mathcal{X}^0)^*$. We initially took F^* to be in $(\mathcal{X}^0)^*$, and discovered that it actually belongs to the subspace $\mathcal{X}^*$. So we have refined our knowledge about F^* . Second, there exists an element $p \in \mathcal{Y}$ such that

$$B^T p = -F^* \text{ in } \mathcal{X}^* . \quad (A2.32)$$

Expression (A2.32), when worked out with the help of (A2.21), reveals that

$$a(u, \cdot) + B^T p = f^* \text{ in } \mathcal{X}^* , \quad (A2.33)$$

or

$$a(u, x) + \langle B^T p, x \rangle_{\mathcal{X}^*, \mathcal{X}} = \langle f^*, x \rangle_{\mathcal{X}^*, \mathcal{X}} , \quad \forall x \in \mathcal{X} \quad (A2.34)$$

Finally, (A2.34) implies that

$$a(u, x) + b(x, p) = \langle f^*, x \rangle_{\mathcal{X}^*, \mathcal{X}} , \quad \forall x \in \mathcal{X} , \quad (A2.35)$$

which is nothing else than the first equation from the original system (3.69.g). Expression above says that the solution (u, p) to our problem *exists*. This follows from (A2.32), which establishes the existence for p , and from the existence of u^s and u_g . According to (A2.5), if u^s and u_g exist, then obviously $u = u^s + u_g$ also exists.

Part II: Uniqueness

Now that we know the solution (u, p) exists, we need to show that it is unique. We say that $(u, p) \in \mathcal{X} \times \mathcal{Y}$ is the solution to the original problem (3.69.g) if

$$a(u, x) + b(x, p) = \langle f^*, x \rangle_{\mathcal{X}^*, \mathcal{X}} \quad \forall x \in \mathcal{X} \quad (A2.36)$$

$$b(u, y) = \langle g^*, y \rangle_{\mathcal{Y}^*, \mathcal{Y}} \quad \forall y \in \mathcal{Y}$$

Suppose $(u_2, p_2) \in \mathcal{X} \times \mathcal{Y}$ is another solution to problem (3.69.g). Then

$$a(u_2, x) + b(x, p_2) = \langle f^*, x \rangle_{\mathcal{X}^*, \mathcal{X}} \quad \forall x \in \mathcal{X} \quad (A2.37)$$

$$b(u_2, y) = \langle g^*, y \rangle_{\mathcal{Y}^*, \mathcal{Y}} \quad \forall y \in \mathcal{Y}$$

If we subtract the first equation in (A2.37) from the first in (A2.36) (and likewise for the second equations), we get

$$a(u - u_2, x) + b(x, p - p_2) = 0 \quad \forall x \in \mathcal{X} \quad (\text{A2.38})$$

$$b(u - u_2, y) = 0 \quad \forall y \in \mathcal{Y}$$

From the second equation in (A2.38), we observe that $u - u_2 \in \text{Ker } b = \mathcal{X}^0$. Since $\mathcal{X}^0 \subset \mathcal{X}$, it follows that

$$a(u - u_2, x) + b(x, p - p_2) = 0 \quad \forall x \in \mathcal{X}^0 \quad (\text{A2.39})$$

But since $x \in \mathcal{X}^0$ in (A2.39), then $b(x, p - p_2) = 0$. Consequently,

$$a(u - u_2, x) = 0 \quad \forall x \in \mathcal{X}^0 \quad (\text{A2.40})$$

Expression (A2.40) is just the homogeneous problem at the kernel; according to hypothesis (vii), its solution is zero. We conclude that $u - u_2 = 0_{\mathcal{X}^0} = 0_x$ and then,

$$u_2 = u. \quad (\text{A2.41})$$

Since $u - u_2 = 0_x$, the first equation in (A2.38) gives

$$b(x, p - p_2) = 0, \quad \forall x \in \mathcal{X} \quad (\text{A2.42})$$

which is the same as

$$\langle x, B^T(p - p_2) \rangle_{\mathcal{X}, \mathcal{X}^*} = 0, \quad \forall x \in \mathcal{X}, \quad (\text{A2.43})$$

according to the definition of the operator B^T in (3.29). Expression (A2.43) implies that

$$B^T(p - p_2) = 0_{\mathcal{X}^*}, \quad (\text{A2.44})$$

where $0_{\mathcal{X}^*}$ is the zero functional (the zero element) from the dual space $\mathcal{X}^*$. It is assumed in hypothesis (vi) that the inf-sup condition (3.69.e) holds true. According to conclusion (ii) from Theorem 3.3, the inf-sup condition is equivalent to the fact that B^T is injective (i.e., B^T is one-to-one), which means that $\text{Ker } B^T = \{0_y\}$. But (A2.44) says that $(p - p_2) \in \text{Ker } B^T$. Consequently, $p - p_2 = 0_y$, or

$$p_2 = p \quad (\text{A2.45})$$

We have just showed that, if (u_2, p_2) is any eventual solution to the original problem (A2.6), then it is equal to (u, p) , whose existence has been proved in Part I. Therefore, the solution (u, p) is unique.

Part III: Boundedness

According to (A2.5), $u = u^s + u_g$. The estimates on these two parts are

$$\|u_g\|_{\mathcal{X}} \leq \frac{1}{\beta_b} \|g^*\|_{\mathcal{Y}^*}, \quad (\text{A2.46})$$

proven in (A2.4), and

$$\|I_{\mathcal{X}^0 \rightarrow H}(u^s)\|_H \leq C_{FA} \|f^* - Au_g\|_{\mathcal{X}^*} \quad (\text{A2.47})$$

established in (A2.19). We can work (A2.47) out and observe that

$$\|I_{\mathcal{X}^0 \rightarrow H}(u^s)\|_H \leq C_{FA} \|f^* - Au_g\|_{\mathcal{X}^*} \quad (\text{A2.48})$$

$$\leq C_{FA} (\|f^*\|_{\mathcal{X}^*} + \|Au_g\|_{\mathcal{X}^*}) \quad (\text{A2.49})$$

$$\leq C_{FA} (\|f^*\|_{\mathcal{X}^*} + \alpha_a \|u_g\|_{\mathcal{X}}) \quad (\text{A2.50})$$

$$\leq C_{FA} \left(\|f^*\|_{\mathcal{X}^*} + \frac{\alpha_a}{\beta_b} \|g^*\|_{\mathcal{Y}^*} \right) \quad (\text{A2.51})$$

In (A2.49) the triangle inequality has been employed. The move from (A2.49) to (A2.50) is justified by the fact that the operator A is induced by the continuous sesquilinear form a , according to (A2.7). From this point to (A2.51), it suffices to consider (A2.46).

The Fredholm Alternative gives us estimates concerning the third ‘auxiliary’ Hilbert space H . We have got a funny fact in which u^s , the portion of the solution which lies at the kernel $\mathcal{X}^0$, is measured in the norm of H , whereas u_g is measured in the norm of $\mathcal{X}$.

If we assume further that the original Hilbert space $\mathcal{X}$ is also embedded in H , i.e., if $\mathcal{X} \hookrightarrow H$ [hypothesis (viii)], then there is a continuous map $I_{\mathcal{X} \rightarrow H}: \mathcal{X} \rightarrow H$, i.e.,

$$\|I_{\mathcal{X} \rightarrow H}(w)\|_H \leq C'_e \|w\|_{\mathcal{X}}, \quad \forall w \in \mathcal{X}, \quad (\text{A2.52})$$

where C'_e is a constant independent of w . In principle, the constants from the embeddings $\mathcal{X} \hookrightarrow H$ and $\mathcal{X}^0 \hookrightarrow H$ may be different from each other. From (A2.46) and (A2.52), in which we make $w = u_g$, we get

$$\|I_{\mathcal{X} \rightarrow H}(u_g)\|_H \leq \frac{C'_e}{\beta_b} \|g^*\|_{\mathcal{Y}^*} \quad (\text{A2.53})$$

If $\mathcal{X} \subset H$ and $\mathcal{X}^0 \subset H$, which implies that $I_{\mathcal{X} \rightarrow H}$ and $I_{\mathcal{X}^0 \rightarrow H}$ are inclusion maps [hypothesis (ix)], then

$$I_{\mathcal{X} \rightarrow H}(u_g) = u_g \quad (\text{A2.54})$$

$$I_{\mathcal{X}^0 \rightarrow H}(u^s) = u^s \quad (\text{A2.55})$$

Estimates (A2.51) and (A2.53) therefore simplify to

$$\|u^s\|_H \leq C_{FA} \left(\|f^*\|_{\mathcal{X}^*} + \frac{\alpha_a}{\beta_b} \|g^*\|_{\mathcal{Y}^*} \right) \quad (A2.56)$$

$$\|u_g\|_H \leq \frac{C'_e}{\beta_b} \|g^*\|_{\mathcal{Y}^*} \quad (A2.57)$$

Since it is true that $u \in \mathcal{X}$, we can form the chain of results:

$$\|I_{\mathcal{X} \rightarrow H}(u)\|_H = \|u\|_H \quad (A2.58)$$

$$= \|u^s + u_g\|_H \quad (A2.59)$$

$$\leq \|u^s\|_H + \|u_g\|_H \quad (A2.60)$$

$$\leq C_{FA} \|f^*\|_{\mathcal{X}^*} + \frac{(C_{FA}\alpha_a + C'_e)}{\beta_b} \|g^*\|_{\mathcal{Y}^*} \quad (A2.61)$$

The equality in (A2.58) is justified by the fact that $I_{\mathcal{X} \rightarrow H}$ is an inclusion map. In (A2.60), the usual triangle inequality has been employed. At last, (A2.61) follows from (A2.56) and (A2.57). Therefore our estimate on the solution u is given by

$$\|u\|_H \leq C_{FA} \|f^*\|_{\mathcal{X}^*} + \frac{(C_{FA}\alpha_a + C'_e)}{\beta_b} \|g^*\|_{\mathcal{Y}^*} \quad (A2.62)$$

In order to find an estimate for p , we recall that the inf-sup condition (3.69.e) – which is assumed to hold – is equivalent to

$$\sup_{x \in \mathcal{X} \setminus \{0\}} \frac{\langle x, B^T y \rangle_{\mathcal{X}, \mathcal{X}^*}}{\|x\|_{\mathcal{X}}} \geq \beta_b \|y\|_{\mathcal{Y}}, \quad \forall y \in \mathcal{Y} \quad (A2.63)$$

according to conclusion (iv) in Theorem 3.3. We take $y = p$ in (A2.63) and the definition of the norm of a functional to conclude that

$$\beta_b \|p\|_{\mathcal{Y}} \leq \|B^T p\|_{\mathcal{X}^*}. \quad (A2.64)$$

From (A2.34), we see that

$$a(u, x) + \langle B^T p, x \rangle_{\mathcal{X}^*, \mathcal{X}} = \langle f^*, x \rangle_{\mathcal{X}^*, \mathcal{X}}, \quad \forall x \in \mathcal{X}, \quad (A2.65)$$

from which it follows that for any $x \in \mathcal{X}$,

$$|\langle B^T p, x \rangle_{\mathcal{X}^*, \mathcal{X}}| \leq |\langle f^*, x \rangle_{\mathcal{X}^*, \mathcal{X}} - a(u, x)| \quad (A2.66)$$

$$\leq |\langle f^*, x \rangle_{\mathcal{X}^*, \mathcal{X}}| + |a(u, x)| \quad (A2.67)$$

$$\leq (\|f^*\|_{\mathcal{X}^*} + \alpha_a \|u\|_{\mathcal{X}}) \|x\|_{\mathcal{X}} \quad (A2.68)$$

Consequently,

$$\frac{|\langle B^T p, x \rangle_{\mathcal{X}^*, \mathcal{X}}|}{\|x\|_{\mathcal{X}}} \leq \|f^*\|_{\mathcal{X}^*} + \alpha_a \|u\|_{\mathcal{X}}, \quad \forall x \in \mathcal{X} \setminus \{0\} \quad (\text{A2.69})$$

Moreover,

$$\|B^T p\|_{\mathcal{X}^*} := \sup_{x \in \mathcal{X} \setminus \{0\}} \frac{|\langle B^T p, x \rangle_{\mathcal{X}^*, \mathcal{X}}|}{\|x\|_{\mathcal{X}}} \leq \|f^*\|_{\mathcal{X}^*} + \alpha_a \|u\|_{\mathcal{X}}. \quad (\text{A2.70})$$

From (A2.64) and (A2.70), we conclude that

$$\|p\|_y \leq \frac{\|f^*\|_{\mathcal{X}^*} + \alpha_a \|u\|_{\mathcal{X}}}{\beta_b}. \quad (\text{A2.71})$$

■

Unfortunately, we are not able to provide an estimate for p based on the data $\|f^*\|_{\mathcal{X}^*}$ and $\|g^*\|_{y^*}$ only, as it is done for u in (A2.62). The question is that (A2.71) depends on u measured in the norm of $\mathcal{X}$, whereas in (A2.62) u is measured in the norm of H .

This little issue is due to the fact that the Fredholm Alternative provides estimates for the norm of the solution with respect to the auxiliary Hilbert space H , and not with respect to the original space $\mathcal{X}$.

Appendix 3

List of Symbols

This is a non-exhaustive list concerning the symbols standing for some of the mathematical objects which appear in this work. Each entry is described by three fields: First: The symbol of the object. Second: A brief description of it. Third: The first page in which the symbol appears.

$\mathcal{B}(\mathbf{x}, t)$	Time-dependent magnetic flux density	4
$C^m(\Omega)$	Space of m -times continuously differentiable functions	35
$C_0^\infty(\Omega)$	Space of compactly supported and infinitely differentiable functions	25
$C^\infty(\bar{\Omega})$	Space of m -times uniformly and continuously differentiable functions	35
$\mathcal{D}(\mathbf{x}, t)$	Time-dependent electric flux density	4
$\bar{\bar{\mathbf{D}}}$	Strain rate tensor	15
$\mathcal{D}_\tau(\Omega)$	A subspace of $C^\infty(\bar{\Omega})^3$	53
$\mathbf{E}(\mathbf{x})$	Time-harmonic electric field	6
$\mathbf{E}^s(\mathbf{x})$	Time-harmonic scattered electric field	10
$\mathbf{E}^{inc}(\mathbf{x})$	Time-harmonic incident electric field	10
$\mathcal{E}(\mathbf{x}, t)$	Time-dependent electric field	4
$\mathbb{E}^h(\Omega)$	Finite-dimensional subspace of $H^1(\Omega)^3$	118
$\mathbb{V}_\tau^h(\Omega)$	A subspace of $\mathbb{E}^h(\Omega)$	119
$\mathcal{H}(\mathbf{x}, t)$	Time-dependent magnetic field	4
$H^1(\Omega)$	Sobolev space $W^{1,2}(\Omega)$	29
$H_0^1(\Omega)$	A subspace of $H^1(\Omega)$	35
$H^1(\Omega)^3$	‘Three-dimensional’ $H^1(\Omega)$ space	32
$H^{1/2}(\Gamma)$	Range of the trace operator γ_0	36
$H^{-1/2}(\Gamma)$	Dual space of $H^{1/2}(\Gamma)$	60
$H(\mathbf{curl}; \Omega)$	A particular Sobolev space	58
$H_0(\mathbf{curl}; \Omega)$	Subspace of $H(\mathbf{curl}; \Omega)$	59
$\bar{\bar{\mathbf{I}}}$	Identity tensor	13

$\mathbf{J}(\mathbf{x}, t)$	Time-dependent electric current density	4
$\bar{\bar{\mathbf{J}}}$	Viscous stress tensor	15
$\mathbf{J}_S(\mathbf{x})$	Time-harmonic source current density	6
$\mathbf{J}_S(\mathbf{x}, t)$	Time-dependent source current density	5
Ker	Kernel, or null space of an operator	36
$\mathcal{K}(\mathcal{X}, \mathcal{Y})$	Space of compact operators	87
$\mathcal{L}(\mathcal{X}, \mathcal{Y})$	Space of bounded and linear operators	87
$L^p(\Omega)$	Lebesgue space, index p	26
$L^1_{loc}(\Omega)$	Space of locally summable functions	27
$L^2(\Omega)^3$	‘Three-dimensional’ $L^2(\Omega)$ space	31
$L^2_0(\Omega)$	Zero-average $L^2(\Omega)$ space	41
$\mathbb{P}^h(\Omega)$	Finite-dimensional subspace of $L^p(\Omega)$	118
V_I	Local space associated with patch I (unspecified)	137
V_I^e	Local space associated with patch I (electric field)	147
V_I^p	Local space associated with patch I (pseudopressure)	147
$\mathbb{V}_\tau(\Omega)$	A subspace of $H^1(\Omega)^3$	64
$Y(\Gamma)$	Range of the tangential trace operator $\boldsymbol{\gamma}_t$	60
$\boldsymbol{\rho}(\mathbf{x}, t)$	Time-dependent electric charge density	4
Tr	Trace of a tensor	18
$\hat{\mathbf{a}}_I, \hat{\mathbf{b}}_I, \hat{\mathbf{c}}_I$	Elemental directions associated with node I	148
f	Frequency	6
$\mathbf{f}$	Mass density of forces	15
$\mathbf{g}$	Non-homogeneous Dirichlet boundary condition	37
$h_{Im}^e(\mathbf{x})$	Two-index basis function for the electric field	149
$h_{Im}^p(\mathbf{x})$	Two-index basis function for the pseudopressure	145
k_0	Free-space wavenumber	7
$\ell_{I,m}$	m -th function in the local basis for the patch I	137
$\hat{\mathbf{n}}$	Outward-pointing unit normal vector	5
p	Pseudopressure (Lagrange multiplier)	13
supp	Support of a function	25

$\mathbf{u}$	Velocity field	15
Γ	Boundary of the computational domain Ω	5
Γ_o	Outer boundary of the computational domain	8
Γ_1	Boundary of the PEC scatterer	8
$\Lambda_x, \Lambda_y, \Lambda_z$	Components of the PML tensor $\bar{\bar{\Lambda}}$	51
$\bar{\bar{\Lambda}}$	PML tensor	51
Ω	Computational domain	5
Ω_I	A patch associated with node I	136
Ω_o	The interior of Γ_o	8
$\bar{\Omega}$	Closure of the computational domain Ω	10
ε_0	Free-space electric permittivity	5
ε_r	Relative electric permittivity	5
γ_0	Trace operator	35
γ_0^d	‘Multidimensional’ trace operator	37
γ_t	Tangential trace operator	60
$\varphi_I^0(\mathbf{x})$	PU function associated with patch I	138
η_0	Vacuum impedance	185
λ	Lamé coefficient	19
λ_0	Free-space wavelength	7
μ	Lamé coefficient, dynamic viscosity	19
μ_0	Free-space magnetic permeability	4
ν	Kinematic viscosity	22
ω	Angular frequency	6
μ_r	Relative magnetic permeability	4
$\rho(\mathbf{x})$	Time-harmonic electric charge density	6
ρ	Mass density	15
σ	Electric conductivity	5
$\bar{\bar{\sigma}}$	Cauchy stress tensor	15
$\partial\Omega$	Boundary of the computational domain Ω	5
$\nabla \times$	The curl operator	4

$\nabla \cdot$	The divergence operator	4
∇^2	The Laplacian operator	11
$\otimes$	Tensor product operator	15

Bibliography

[Abraham *et al.*, 1988] R. Abraham, J. E. Marsden and T. Ratiu, *Manifolds, Tensor Analysis and Applications*, 2nd Edition, Applied Mathematical Sciences Series, Book 75, Springer, 1988.

[Atluri and Shen, 2002] S. Atluri and S. Shen, “The meshless local Petrov-Galerkin method: A simple and less-costly alternative to the finite-element and boundary element methods”, *CMES: Computer Modeling in Engineering & Sciences*, vol.3, no.1, pp.11-51, 2002.

[Babuska and Melenk, 1997] I. Babuska and J. M. Melenk, “The partition of unity method”, *International Journal for Numerical Methods in Engineering*, vol.40, pp.727-758.

[Babuska *et al.*, 2009] I. Babuska, U. Banerjee, J. E. Osborn and Q. Zhang, “Effect of numerical integration on meshless methods”, *Computer Methods in Applied Mechanics and Engineering*, vol.198, pp.2886-2897, 2009.

[Balanis, 1989] C. Balanis, *Advanced Engineering Electromagnetics*, John Wiley & Sons, 1989.

[Bathe, 1996] K.-J. Bathe, *Finite Element Procedures*, Prentice Hall, 1996.

[Bathe, 2001] K.-J. Bathe, “The inf-sup condition and its evaluation for mixed finite elements”, *Computers and Structures*, vol.79, pp.243-252, 2001.

[Belytschko *et al.*, 1994] T. Belytschko, Y. Y. Lu and L. Gu, “Element-free Galerkin methods”, *International Methods for Numerical Methods in Engineering*, vol.37, Issue 2, pp.229-256, 1994.

[Benzi and Golub, 2004] M. Benzi and G. H. Golub, “A Preconditioner for Generalized Saddle Point Problems”, *SIAM Journal on Matrix Analysis and Applications*, vol.26, pp.20-41, 2004.

[Benzi and Wathen, 2008] M. Benzi and A. J. Wathen, “Some Preconditioning Techniques for Saddle Point Problems”, *Model Order Reduction: Theory, Research Aspects and Applications*, Springer Mathematics in Industry Series, Book 13, pp.195-211, 2008.

[Bermúdez *et al.*, 2004] A. Bermúdez, L. Hervella-Nieto, A. Prieto and R. Rodríguez, “An exact bounded PML for the Helmholtz equation”, *Comptes Rendus de l’Académie des Sciences de Paris, Série I, Mathématiques*, 339, pp.803-808, 2004.

[Bermúdez *et al.*, 2007] A. Bermúdez, L. Hervella-Nieto, A. Prieto and R. Rodríguez, “An optimal perfectly matched layer with unbounded absorbing function for time-harmonic acoustic scattering problems”, *Journal of Computational Physics*, vol. 223, pp.469-488, 2007.

[Bermúdez *et al.*, 2010] A. Bermúdez, L. Hervella-Nieto, A. Prieto and R. Rodríguez, “Perfectly Matched Layers for Time-Harmonic Second Order Elliptic Problems”, *Archives of Computational Methods in Engineering*, vol.17, pp.77-107, 2010.

[Boffi, 2010] D. Boffi, “Finite Element Approximation of Eigenvalue Problems”, *Acta Numerica*, vol.19, pp.1-120, 2010.

[Boffi *et al.*, 2013] D. Boffi, F. Brezzi and M. Fortin, *Mixed Finite Element Methods and Applications*, Springer Series in Computational Mathematics, Book 44, 2013.

[Bogovskii, 1980] M. E. Bogovskii, “Solution of some vector analysis problems connected with operators div and grad”, *Trudy Seminar S. Sobolev*, No.1, vol.49, pp.5-40, Akademia Nauk SSSR, Sibirskoe Otdelnie Matematiki, Novosibirsk, Russia, 1980.

[Bossavit, 1997] A. Bossavit, *Computational Electromagnetism: Variational Formulations, Complementarity, Edge Elements*, Academic Press, 1997.

[Bottauscio *et al.*, 2006] O. Bottauscio, M. Chiampi and A. Manzini, “Element-free Galerkin method in eddy-current problems with ferromagnetic media”, *IEEE Transactions on Magnetics*, vol.42, no.5, pp.1577-1584, 2006.

[Boyer and Fabrie, 2012] F. Boyer and P. Fabrie, *Mathematical Tools for the Study of the Incompressible Navier-Stokes Equations and Related Models*, Applied Mathematical Sciences Series, Book 183, Springer, 2012.

[Brezis, 2010] H. Brezis, *Functional Analysis, Sobolev Spaces and Partial Differential Equations*, Springer Universitext, 2010.

[Brezzi and Bathe, 1990] F. Brezzi and K.-J. Bathe, “A Discourse on the Stability Conditions for Mixed Finite Element Formulations”, *Computer Methods in Applied Mechanics and Engineering*, vol.82, pp.27-57, 1990.

[Brezzi and Fortin, 1991] F. Brezzi and M. Fortin, *Mixed and Hybrid Finite Elements*, Springer Series in Computational Mathematics, Book 15, 1991.

[Böhmer, 2010] K. Böhmer, *Numerical Methods for Nonlinear Elliptic Partial Differential Equations: A Synopsis*, Numerical Mathematics and Scientific Computation Series, Oxford University Press, 2010.

[Chapelle and Bathe, 2011] D. Chapelle and K.-J. Bathe, *The Finite Element Analysis of Shells: Fundamentals*, 2nd Edition, Springer Computational Solid and Fluid Mechanics Series, 2011.

- [Cheney, 2001] W. Cheney, *Analysis for Applied Mathematics*, Springer Graduate Texts in Mathematics, Book 208, 2001.
- [Cingoski *et al.*, 1998] V. Cingoski, N. Miyamoto and H. Yamashita, “Element-Free Galerkin Method for Electromagnetic Field Computations”, *IEEE Transactions on Magnetism*, vol. 34, no.5, pp.3236-3239, 1998.
- [Conway, 1994] J. B. Conway, *A Course in Functional Analysis*, 2nd Edition, Springer Graduate Texts in Mathematics, Book 96, 1994.
- [Crossley, 2005] M. Crossley, *Essential Topology*, Springer, 2005.
- [De and Bathe, 2000] S. De and K. J. Bathe, “The Method of Finite Spheres”, *Computational Mechanics*, vol.25, pp.329-345, 2000.
- [De and Bathe, 2001] S. De and K. J. Bathe, “The Method of Finite Spheres with Improved Numerical Integration”, *Computers & Structures*, vol.79, pp.2183-2196, 2001.
- [De and Bathe², 2001] S. De and K. J. Bathe, “Displacement/Pressure Mixed Interpolation in the Method of Finite Spheres”, *International Journal for Numerical Methods in Engineering*, vol. 51, pp.275-292, 2001.
- [De *et al.*, 2003] S. De, J. W. Hong and K. J. Bathe, “On the Method of Finite Spheres in Applications: Towards the Use with ADINA and in a Surgical Simulator”, *Computational Mechanics*, vol.31, pp.27-37, 2003.
- [Dehghan and Mirzaei, 2008] M. Dehghan and D. Mirzaei, “The Meshless Local Petrov-Galerkin (MLPG) method for the generalized two-dimensional nonlinear Schrödinger equation”, *Engineering Analysis with Boundary Elements*, vol.32, pp.747-756, 2008.
- [Duarte and Oden, 1996] C. A. Duarte and J. T. Oden, “H-p clouds – an h-p meshless method”, *Numerical Methods for Partial Differential Equations*, vol.12, pp.673-705.
- [Embar *et al.*, 2010] A. Embar, J. Dolbow and I. Harari, “Imposing Dirichlet boundary conditions with Nitsche’s method and spline-based finite elements”, *International Journal for Numerical Methods in Engineering*, vol.83, pp.877-898, 2010.
- [Ern and Guermond, 2004] A. Ern and J.-L. Guermond, *Theory and Practice of Finite Elements*, Applied Mathematical Sciences Series, Book 159, Springer, 2004.
- [Evans, 2010] L. Evans, *Partial Differential Equations*, 2nd Edition, Graduate Studies in Mathematics, Vol. 19, American Mathematical Society, 2010.
- [Galdi, 2011] G. P. Galdi, *An Introduction to the Mathematical Theory of the Navier-Stokes Equations: Steady-State Problems*, 2nd Edition, Springer Monographs in Mathematics, Book 168, 2011.

[Gerbeau *et al.*, 2006] J.-F. Gerbeau, C. Le Bris and T. Lelièvre, *Mathematical Methods for the Magnetohydrodynamics of Liquid Metals*, Numerical Mathematics and Scientific Computation Series, Oxford University Press, 2006.

[Gingold and Monaghan, 1977] R. A. Gingold and J. J. Monaghan, “Smoothed particle hydrodynamics: Theory and application to non-spherical stars”, *Monthly Notices of the Royal Astronomical Society*, vol.181, pp.375-379, 1977.

[Girault and Raviart, 1986] V. Girault and P.-A. Raviart, *Finite Element Methods for Navier-Stokes Equations: Theory and Algorithms*, Springer Series in Computational Mathematics, Book 5, 1986.

[Glowinski *et al.*, 2003] R. Glowinski, P. G. Ciarlet and J. L. Lions, *Handbook of Numerical Analysis – Vol. IX: Numerical Methods for Fluids*, Elsevier 2003.

[Gross and Reusken, 2011] S. Gross and A. Reusken, *Numerical Methods for Two-phase Incompressible Flows*, Springer Series in Computational Mathematics, Book 40, 2011.

[Ham and Bathe, 2012] S. Ham and K.-J. Bathe, “A finite element method enriched for wave propagation problems”, *Computers and Structures*, vol. 94, pp.1-12, 2012.

[Ham *et al.*, 2014] S. Ham, B. Lai and K.-J. Bathe, “The method of finite spheres for wave propagation problems”, *Computers and Structures*, vol. 142, pp.1-14, 2014.

[Hanson and Yakovlev, 2002] G. Hanson and A. Yakovlev, *Operator Theory for Electromagnetics: An Introduction*, Springer, 2002.

[Harrington, 2001] R. F. Harrington, *Time-Harmonic Electromagnetic Fields*, 2nd Edition, Wiley-IEEE Press, 2001.

[Heldring *et al.*, 2002] A. Heldring, J. M. Rius and L. Ligthart, “New Block ILU Preconditioner Scheme for Numerical Analysis of Very Large Electromagnetic Problems”, *IEEE Transactions on Magnetics*, vol.38, pp.337-340.

[Houston *et al.*, 2005] P. Houston, I. Perugia, A. Schneebeli and D. Schötzau, “Mixed discontinuous Galerkin approximation of the Maxwell operator: The indefinite case”, *ESAIM: Mathematical Modeling and Numerical Analysis*, vol.39, no.4, pp.727-753, 2005.

[Ihlenburg, 1998] F. Ihlenburg, *Finite Element Analysis of Acoustic Scattering*, Applied Mathematical Sciences Series, Book 132, Springer, 1998.

[Irgens, 2008] F. Irgens, *Continuum Mechanics*, Springer, 2008.

[Kreyszig, 1989] E. Kreyszig, *Introductory Functional Analysis with Applications*, Wiley Classics Library, Wiley, 1989.

- [Kwon *et al.*, 2001] D.-H. Kwon, R. J. Burkholder and P. H. Pathak, “Efficient Method of Moments Formulation for Large PEC Scattering Problems Using Asymptotic Phasefront Extraction (APE)”, *IEEE Transactions on Antennas and Propagation*, vol.49, pp.583-591, 2001.
- [Leoni, 2009] G. Leoni, *A First Course in Sobolev Spaces*, Graduate Studies in Mathematics, Vol. 105, American Mathematical Society, 2009.
- [Li and Liu, 2007] S. Li and W. K. Liu, *Meshfree Particle Methods*, Springer, 2007.
- [Li *et al.*, 2003] Q. Li, S. Shen, Z. Han and S. Atluri, “Application of meshless Petrov-Galerkin (MLPG) to problems with singularities and material discontinuities in 3-D elasticity”, *CMES: Computer Modeling in Engineering & Sciences*, vol.4, no.5, pp.571-585, 2003.
- [Lima and Mesquita, 2013] N. Z. Lima and R. C. Mesquita, “Point interpolation methods based on weakened-weak formulations”, *Journal of Microwaves, Optoelectronics and Electromagnetic Applications*, vol.12, no.2, pp.506-523, 2013.
- [Liu, 2010] G. R. Liu, *Mesh Free Methods: Moving Beyond the Finite Element Method*, 2nd Edition, CRC Press, 2010.
- [Liu and Liu, 2003] G. R. Liu and M. B. Liu, *Smoothed Particle Hydrodynamics: A Meshfree Particle Method*, World Scientific Publishing Company, 2003.
- [Liu and Liu, 2010] M. B. Liu and G. R. Liu, “Smoothed Particle Hydrodynamics (SPH): an Overview and Recent Developments”, *Archives of Computational Methods in Engineering*, vol.17, pp.25-76, 2010.
- [Lu and Shanker, 2007] C. Lu and B. Shanker, “Generalized finite element method for vector electromagnetic problems”, *IEEE Transactions on Antennas and Propagation*, vol.55, no.5, pp.1369-1381, 2007.
- [Lynch and Paulsen, 1991] D. R. Lynch and K. D. Paulsen, “Origin of Vector Parasites in Numerical Maxwell Solutions” *IEEE Transactions on Microwave Theory and Techniques*, vol. 39, no.3, pp.383-394, 1991.
- [Manzin and Bottauscio, 2008] A. Manzin and O. Bottauscio, “Element-free Galerkin method for the analysis of electromagnetic-wave scattering”, *IEEE Transaction on Magnetics*, vol.44, no.6, pp.1366-1369, 2008.
- [Maréchal, 1998] Y. Maréchal, “Some Meshless Methods for Electromagnetic Field Computations”, *IEEE Transactions on Magnetics*, vol.34, no.5, pp.3351-3354, 1998.
- [Melenk and Babuska, 1996] J. M. Melenk and I. Babuska, “The partition of unity finite element method: basic theory and applications”, *Computer Methods in Applied Mechanics and Engineering*, vol.139, pp.289-314.

- [Moiola and Spence, 2014] A. Moiola and E. A. Spence, “Is the Helmholtz equation really sign-indefinite?”, *SIAM Review*, vol.56, No.2, pp.274-312, 2014.
- [Monk, 2003] P. Monk, *Finite Element Methods for Maxwell’s Equations*, Numerical Mathematics and Scientific Computation Series, Oxford University Press, 2003.
- [Munkres, 2000] J. Munkres, *Topology*, 2nd Edition, Pearson-Prentice Hall, 2000.
- [Necas, 1962] J. Necas, “Sur une méthode pour résoudre les équations aux dérivées partielles de type elliptique, voisine de la variationnelle”, *Annali della Scuola Normale Superiore di Pisa – Classe di Scienze*, vol.16, pp.305-326, 1962.
- [Nguyen *et al.*, 2011] N. C. Nguyen, J. Peraire and B. Cockburn, “Hybridizable discontinuous Galerkin methods for the time-harmonic Maxwell’s equations”, *Journal of Computational Physics*, vol.230, pp.7151-7175, 2011.
- [Nicomedes *et al.*, 2011] W. Nicomedes, R. Mesquita and F. Moreira, “A meshless local Petrov-Galerkin method for three dimensional scalar problems”, *IEEE Transactions on Magnetics*, vol.47, no.5, pp.1214-1217, 2011.
- [Nicomedes *et al.*, 2012] W. Nicomedes, R. Mesquita and F. Moreira, “The Meshless Local Petrov-Galerkin Method in Two-Dimensional Electromagnetic Wave Analysis”, *IEEE Transactions on Antennas and Propagation*, vol.60, no.4, pp.1957-1968, 2012.
- [Nicomedes *et al.*², 2012] W. Nicomedes, R. Mesquita and F. Moreira, “Calculating the band structure of photonic crystals through the meshless local Petrov-Galerkin (MLPG) method and periodic shape functions”, *IEEE Transactions on Magnetics*, vol.48, no.2, pp.551-554, 2012.
- [Parreira *et al.*, 2006] G. Parreira, A. Fonseca and R. Mesquita, “The element-free Galerkin method in three-dimensional electromagnetic problems”, *IEEE Transactions on Magnetics*, vol. 42, no.4, pp.711-714, 2006.
- [Perugia *et al.*, 2002] I. Perugia, D. Schötzau and P. Monk, “Stabilized interior penalty method for the time-harmonic Maxwell equations”, *Computer Methods in Applied Mechanics and Engineering*, vol.191, pp.4675-4697, 2002.
- [Peterson *et al.*, 1998] A. F. Peterson, S. L. Ray and R. Mittra, *Computational Methods for Electromagnetics*, IEEE Press Series on electromagnetic waves, IEEE Press, 1998.
- [Pimenta *et al.*, 2013] L. Pimenta, G. Pereira, N. Michael, R. Mesquita, L. Chaimowicz and V. Kumay, “Swarm Coordination Based on Smoothed Particle Hydrodynamics Technique”, *IEEE Transactions on Robotics*, vol.29, pp.383-399, 2013.
- [Proekt and Tsukerman, 2002] L. Proekt and I. Tsukerman, “Method of overlapping patches for electromagnetic computation”, *IEEE Transactions on Magnetics*, vol.38, no.2, pp.741-744, 2002.
- [Quarteroni, 2009] A. Quarteroni, *Numerical Models for Differential Problems*, MS&A: Modeling, Simulation & Applications, Volume 2, Springer, 2009.

- [Quarteroni and Valli, 1994] A. Quarteroni and A. Valli, *Numerical Approximation of Partial Differential Equations*, Springer Series in Computational Mathematics, Book 23, 1994.
- [Roberts and Thomas, 1991] J. E. Roberts and J.-M. Thomas, “Mixed and Hybrid Methods”, *Handbook of Numerical Analysis – Vol. II: Finite Element Methods*, Elsevier, 1991.
- [Rynne and Youngson, 2007] B. Rynne and M. A. Youngson, *Linear Functional Analysis*, 2nd Edition, Springer Undergraduate Mathematics Series, 2007.
- [Saad, 2003] Y. Saad, *Iterative Methods for Sparse Linear Systems*, 2nd Edition, Society for Industrial and Applied Mathematics – SIAM, 2003.
- [Sacks *et al.*, 1995] Z. S. Sacks, D. M. Kingsland, R. Lee and J. F. Lee, “A Perfectly Matched Anisotropic Absorber for Use as an Absorbing Boundary Condition”, *IEEE Transactions on Antennas and Propagation*, vol.43, no.12, pp.1460-1463, 1995.
- [Salsa, 2008] S. Salsa, *Partial Differential Equations in Action: From Modelling to Theory*, Springer Universitext, 2008.
- [Searcóid, 2007] M. Searcóid, *Metric Spaces*, Springer, 2007.
- [Soares *et al.*, 2014] R. Soares, F. Moreira, R. Mesquita, D. Lowther and N. Z. Lima, “A Modified Meshless Local Petrov-Galerkin Applied to Electromagnetic Axisymmetric Problems”, *IEEE Transactions on Magnetics*, v.50, pp.513-516, 2014.
- [Soares Jr., 2009] D. Soares Jr., “Numerical modeling of electromagnetic wave propagation by meshless local Petrov-Galerkin formulations”, *CMES: Computer Modeling in Engineering & Sciences*, vol.50, no.2, pp.97-114, 2009.
- [Strouboulis *et al.*, 2001] T. Strouboulis, K. Copps and I. Babuska, “The generalized finite element method”, *Computer Methods in Applied Mechanics and Engineering*, vol.190, pp.4081-4193, 2001.
- [Tao, 2011] T. Tao, *An Introduction to Measure Theory*, Graduate Studies in Mathematics, Vol. 126, American Mathematical Society, 2011.
- [van Bladel, 2007] J. van Bladel, *Electromagnetic Fields*, 2nd Edition, Wiley-IEEE Press, 2007.
- [van der Vorst, 2009] H. A. van der Vorst, *Iterative Krylov Methods for Large Linear Systems*, Cambridge Monographs on Applied and Computational Mathematics, Book 13, 2009.
- [Vavourakis, 2009] V. Vavourakis, “A meshless local boundary integral equation method for two-dimensional steady elliptic problems”, *Computational Mechanics*, vol.44, pp.777-790, 2009.
- [Yu and Chen, 2009] Y. Yu and Z. Chen, “A 3-D radial point interpolation method for meshless time-domain modeling”, *IEEE Transactions on Microwave Theory and Techniques*, vol.57, no.8, pp.2015-2020, 2009.

[Yu and Chen, 2010] Y. Yu and Z. Chen, “Towards the development of an unconditionally stable time-domain meshless method”, *IEEE Transactions on Microwave Theory and Techniques*, vol.58, no.3, pp.578-586, 2010.

[Zeytounian, 2012] R. Zh. Zeytounian, *Navier-Stokes-Fourier Equations: A Rational Asymptotic Modelling Point of View*, Springer, 2012.

[Zhang and Batra, 2009] G. M. Zhang and R. C. Batra, “Symmetric smoothed particle hydrodynamics (SSPH) method and its application to elastic problems”, *Computational Mechanics*, vol.43, pp.321-340, 2009.