Christian Constanda
Editor

Computational and Analytic Methods in Science and Engineering

Birkhäuser

Computational and Analytic Methods in Science and Engineering

Christian Constanda
Editor

Computational and Analytic Methods in Science and Engineering

Birkhäuser

Editor
Christian Constanda
The Charles W. Oliphant Professor of Mathematics
The University of Tulsa
Tulsa, OK, USA

ISBN 978-3-030-48185-8 ISBN 978-3-030-48186-5 (eBook)
https://doi.org/10.1007/978-3-030-48186-5

Mathematics Subject Classification: 00B15, 74G10, 93E25

This book is published under the imprint Birkhäuser, www.birkhauser-science.com, by the registered company Springer Nature Switzerland AG.
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

Preface

The international conferences on Computational and Mathematical Methods in Science and Engineering (CMMSE) are annual events where professionals of a variety of denominations who use analytic and numerical methods of investigation communicate the most recent results of their research.

The latest edition of this well-established series of meetings took place in the resort of Costa Ballena, Rota, Cadiz, Spain, June 30–July 6, 2019, and included a special session on the applications of integral methods to scientific developments in a variety of fields, such as pure analysis, numerical techniques, mathematical biology, petroleum engineering, and continuum mechanics. The chapters in this volume, arranged alphabetically by first author's name, represent a collection of selected, peer-reviewed articles presented in that session.

On behalf of the participants, I wish to express my appreciation to the organizing committee—in particular, to its chairman, Jesús Vigo Aguiar—for underwriting the success of the conference by providing an environment conducive to the forging of good interpersonal relationships and the creation of synergies that will lead to further advancements in the construction and use of an essential class of techniques for the qualitative and quantitative study of mathematical models.

Finally, I would also like to thank the reviewers for their thorough and timely responses and Christopher Tominich and his team at Birkhäuser–New York for their courteous and professional handling of the publication process.

Tulsa, OK, USA
March 2020

Christian Constanda

Contents

1 New Numerical Results for the Optimization of Neumann Eigenvalues 1
Daniel Abele and Andreas Kleefeld

2 Transient Convection-Diffusion-Reaction Problems with Variable Velocity Field by Means of DRBEM with Different Radial Basis Functions 21
Salam Adel Al-Bayati and Luiz C. Wrobel

3 On a Parametric Representation of the Angular Neutron Flux in the Energy Range from 1 eV to 10 MeV 45
Luiz F. F. Chaves Barcellos, Bardo E. J. Bodmann, and Marco T. Vilhena

4 A Boundary Integral Equation Formulation for Advection–Diffusion–Reaction Problems with Point Sources 61
Luiz F. Bez, Rogério J. Marczak, Bardo E. J. Bodmann, and Marco T. Vilhena

5 Displacement Boundary Value Problem for a Thin Plate in an Unbounded Domain 75
Christian Constanda and Dale Doty

6 A Dirichlet Spectral Problem in Domains Surrounded by Thin Stiff and Heavy Bands 101
Delfina Gómez, Sergey A. Nazarov, and Maria–Eugenia Pérez-Martínez

7 Spectral Homogenization Problems in Linear Elasticity with Large Reaction Terms Concentrated in Small Regions of the Boundary 127
Delfina Gómez, Sergey A. Nazarov, and Maria-Eugenia Pérez-Martínez

8 The Mathematical Modelling of the Motion of Biological Cells in Response to Chemical Signals 151
Paul J. Harris

9 Numerical Calculation of Interior Transmission Eigenvalues with Mixed Boundary Conditions 173
Andreas Kleefeld and Jijun Liu

10 An Inequality for Hölder Continuous Functions Generalizing a Result of Carlo Miranda 197
Massimo Lanza de Cristoforis

11 Two-Phase Three-Component Flow in Porous Media: Mathematical Modeling of Dispersion-Free Pressure Behavior 223
Luara K. S. Sousa, Luana C. M. Cantagesso, Adolfo P. Pires, and Alvaro M. M. Peres

12 Error Analysis and the Role of Permutation in Dynamic Iteration Schemes 239
Barbara Zubik-Kowal

Correction to: An Inequality for Hölder Continuous Functions Generalizing a Result of Carlo Miranda C1

Index 257

Contributors

Daniel Abele Forschungszentrum Jülich, Jülich Supercomputing Centre, Jülich, Germany

Salam Adel Al-Bayati College of Sciences, Department of Mathematics and Computer Applications, AL-Nahrain University, Baghdad, Iraq

Luiz F. F. Chaves Barcellos Federal University of Rio Grande do Sul, Porto Alegre, RS, Brazil

Luiz F. Bez Federal University of Rio Grande do Sul, Porto Alegre, RS, Brazil

Bardo E. J. Bodmann Federal University of Rio Grande do Sul, Porto Alegre, RS, Brazil

Luana C. M. Cantagesso Universidade Estadual do Norte Fluminense, Macaé, RJ, Brazil

Christian Constanda The University of Tulsa, Tulsa, OK, USA

Dale Doty The University of Tulsa, Tulsa, OK, USA

Delfina Gómez Universidad de Cantabria, Santander, Spain

Paul J. Harris The University of Brighton, Brighton, UK

Andreas Kleefeld Forschungszentrum Jülich GmbH, Jülich Supercomputing Centre, Jülich, Germany

Massimo Lanza de Cristoforis Dipartimento di Matematica 'Tullio Levi-Civita', Università degli Studi di Padova, Padova, Italy

Jijun Liu School of Mathematics/Seu-Yau Center, Southeast University, Nanjing, China

Rogério J. Marczak Federal University of Rio Grande do Sul, Porto Alegre, RS, Brazil

Sergey A. Nazarov Saint-Petersburg State University, St. Petersburg, Russia
Institute of Problems of Mechanical Engineering RAS, St. Petersburg, Russia

Alvaro M. M. Peres Universidade Estadual do Norte Fluminense, Macaé, RJ, Brazil

Maria-Eugenia Pérez-Martínez Universidad de Cantabria, Santander, Spain

Adolfo P. Pires Universidade Estadual do Norte Fluminense, Macaé, RJ, Brazil

Luara K. S. Sousa Universidade Estadual do Norte Fluminense, Macaé, RJ, Brazil

Marco T. Vilhena Federal University of Rio Grande do Sul, Porto Alegre, RS, Brazil

Luiz C. Wrobel Institute of Materials and Manufacturing, Brunel University London, Uxbridge, UK
Department of Civil and Environmental Engineering, Pontifical Catholic University of Rio de Janeiro (PUC-Rio), Brazil

Barbara Zubik-Kowal Department of Mathematics, Boise State University, Boise, ID, USA

Chapter 1
New Numerical Results for the Optimization of Neumann Eigenvalues

Daniel Abele and Andreas Kleefeld

1.1 Introduction

We will discuss the optimization of interior Neumann eigenvalues with respect to the shape of the domain. To state the problem precisely, consider an open, possibly disconnected set $\Omega \in \mathbb{R}^2$ with smooth boundary $\partial\Omega$. The normal onto the boundary at point $x \in \partial\Omega$ directed into the exterior is $\nu := \nu(x)$. Interior Neumann eigenvalues are values $\lambda = \kappa^2 \in \mathbb{R}$ for which the boundary value problem (BVP)

$$\Delta u + \kappa^2 u = 0 \quad \text{in } \Omega \tag{1.1a}$$

$$\frac{\partial u}{\partial \nu} = 0 \quad \text{on } \partial\Omega \tag{1.1b}$$

has non-trivial solutions. Precisely, Eq. (1.1a) is the Helmholtz equation with wavenumber κ in the interior of Ω and Eq. (1.1b) is the homogeneous Neumann boundary condition. The problem arises in the study of acoustic scattering [CoKr13]. It is well-known that the eigenvalues are discrete, real, and nonnegative:

$$0 = \lambda_0 \leq \lambda_1 \leq \lambda_2 \leq \dots .$$

The eigenvalues depend on the domain. The optimization problem for the k-th eigenvalue is

$$\max_{\Omega} \{\lambda_k(\Omega)\}$$

$$\text{s.t. } |\Omega| = 1$$

D. Abele · A. Kleefeld (✉)
Forschungszentrum Jülich, Jülich Supercomputing Centre, Jülich, Germany
e-mail: d.abele@fz-juelich.de; a.kleefeld@fz-juelich.de

C. Constanda (ed.), *Computational and Analytic Methods in Science and Engineering*, https://doi.org/10.1007/978-3-030-48186-5_1

Table 1.1 Recent results for the numerical optimization of interior Neumann eigenvalue λ_k

k	[AnFr12]	[AnOu17]	[Kl19]
1	–	10.66(2)	–
2	–	21.28(4)	–
3	32.79 (3)	32.90(3)	32.9018 (3)
4	43.43 (5)	43.86(3)	43.8694 (3)
5	54.08 (7)	55.17(3)	–
6	67.04 (4)	67.33(4)	–
7	77.68 (6)	77.99(6)	–
8	89.22 (4)	89.38(4)	–
9	101.73 (4)	101.83(4)	–
10	113.86 (5)	114.16(5)	–

The value in parentheses is the multiplicity

with $|\Omega|$ denoting the area of the domain. The area must be constrained as the eigenvalues are inversely proportional to the area. So the goal is to find the shape of the domain that maximizes λ_k for $k > 0$ among all domains of constant area. The eigenvalue λ_0, for which (1.1) has only constant solutions, is ignored here as it is always zero.

There has been some theoretical and numerical work in this area. Szegö and a little later Weinberger have shown that the first eigenvalue is maximized by a disk [Sz54, We56]. The second eigenvalue is maximized by the union of two disjoint disks of the same size [GiNaPo09]. It is so far unknown if maximizers for higher eigenvalues exist. However, it has been shown that disjoint unions of disks do not maximize all eigenvalues [PoRo10], so there is room for exploration. Recent numerical results suggest that maximizers for the first ten eigenvalues exist and follow a certain system [AnOu17, AnFr12]. That system has been exploited to get more precise results for some eigenvalues [Kl19]. Those numerical results are summarized in Table 1.1.

1.1.1 Contribution

This work expands on the idea of [Kl19]. We show that the parametrization of shapes presented there is not very successful beyond the fourth eigenvalue. By introducing additional parameters we managed to get improved optimization results for some eigenvalues, while still using fewer parameters than a general Fourier series approach. As the performance of the eigenvalue solver directly affects the achievable precision, we discuss the employed methods and the implementation in greater detail and explain some adaptations that make optimization feasible on a larger scale than before: more eigenvalues, more degrees of freedom and greater precision. In particular, we developed a strongly scaling parallelization scheme.

1.1.2 Outline

In Sect. 1.2 we present the method of computing the eigenvalues and its implementation. The numerical methods—boundary element method to discretize the BVP and the contour integral method of Beyn to solve the nonlinear eigenvalue problem—are discussed in detail in Sects. 1.2.1 and 1.2.2. That discussion motivates the parallelization scheme that is explained in Sect. 1.2.3. Section 1.3 is dedicated to the actual optimization. After a quick summary of the shape parametrization of [Kl19] we present the disappointing results of using that parametrization in the maximization of further eigenvalues. We then extend the parameter space and show the much improved results. Finally we will give our conclusion and a small outlook in Sect. 1.4.

1.2 Computation of Eigenvalues

The process to compute eigenvalues is the same as in [Kl19] with the exception of parallelization and some other important modifications that have major implications on the required computational effort. First, the BVP (1.1) is discretized using the boundary element method. The resulting homogeneous linear system is a nonlinear eigenvalue problem that is solved with the method of Beyn. To motivate the parallelization, we will give a quick summary of these methods while the modifications are highlighted and discussed in detail.

1.2.1 The Boundary Element Method

The theory of this method is covered in [CoKr83]. Using a single layer potential ansatz, the BVP (1.1) is first converted into the integral equation of the second kind

$$\frac{1}{2}\psi(x) + \int_{\partial\Omega} \frac{\partial}{\partial \nu(x)} \Phi_\kappa(x, y)\psi(y)\,\mathrm{d}s(y) = 0\,, \quad x \in \partial\Omega \tag{1.2}$$

whose solution $\psi \in C(\partial\Omega)$ is the density of the solution of the BVP. The kernel

$$\Phi_\kappa(x, y) := \frac{\mathrm{i}}{4} H_0^{(1)}(\kappa \|x - y\|)$$

with $H_0^{(1)}$ the Hankel function of the first kind of order zero is the fundamental solution of the PDE (1.1a). We choose n points $x_i, i = 1, \ldots, n$ on the boundary that form $n/2$ boundary elements with two endpoints and one midpoint. The integral equation (1.2) is discretized using piecewise quadratic interpolation of ψ and the

boundary. Collocation results in a homogeneous linear system

$$\underbrace{\left(\frac{1}{2}\mathbf{I}+\mathbf{M}(\kappa)\right)}_{\mathbf{T}(\kappa)}\vec{\psi}=0 \tag{1.3}$$

with system matrix $\mathbf{T}(\kappa)$ and identity matrix $\mathbf{I}\in\mathbb{C}^{n\times n}$. The entries of matrix $\mathbf{M}(\kappa)\in\mathbb{C}^{n\times n}$ are integrals over the quadratic boundary elements of the form

$$\int_0^1 \frac{\partial}{\partial\nu_{i,k}}\Phi_\kappa(x_i,g_k(t))L_j(t)\left\|g_k'(t)\right\|\,\mathrm{d}t \tag{1.4}$$

with x_i the collocation point, $k=1,\ldots,n/2$ the index of the boundary element, $g_k:[0,1]\ni t\mapsto L_1(t)x_{2k-1}+L_2(t)x_{2k}+L_3(t)x_{2k+1}$ the quadratic interpolation polynomial of the k-th boundary element, $\nu_{i,k}$ some approximation of the normal onto the boundary at x_i that may depend on the boundary element (see below), and L_j, $j=1,2,3$ the j-th quadratic Lagrange basis polynomial. The integral describes the influence of the j-th point of boundary element k on collocation point x_i, thus index i is the row index of the matrix entry and indices j and k depend on the column index. Elements in odd indexed rows correspond to the endpoints of boundary elements $(x_1,x_3,\ldots)$ and as such are the sum of two such integrals with different j,k as they belong to two different boundary elements. The kernel

$$\frac{\partial}{\partial\nu_{i,k}}\Phi_\kappa(x,y)=-\frac{\mathrm{i}\kappa}{4}H_1^{(1)}(\kappa\|x-y\|)\frac{1}{\|x-y\|}\left\langle x-y,\nu_{i,k}\right\rangle \tag{1.5}$$

has a singularity at $x=y$ the type of which depends on the choice of $\nu_{i,k}$.

Handling of the Singular Kernel

The singularity of the kernel (1.5) must be handled correctly when evaluating the integrals (1.4) numerically. Note first that there is only a singularity in integral (1.4) if collocation point x_i is part of boundary element k, i.e. $g_k(t_0)=x_i$ for some $t_0\in[0,1]$. Otherwise the integrand is continuous and does not present any specific challenge to numerical quadrature. Let us now assume that there is a singularity in the k-th boundary element. We will examine the limit of the kernel by isolating the singular part. First the Hankel function is replaced by $H_1^{(1)}(z)=J_1(z)+\mathrm{i}Y_1(z)$ where J_1 and Y_1 are the Bessel functions of the first and second kind of order one.

As $z \to 0$, $J_1(z)$ tends to zero, so we have

$$\lim_{y\to x}\left(-\frac{\mathrm{i}\kappa}{4}H_1^{(1)}(\kappa\,\|x-y\|)\left\langle\frac{x-y}{\|x-y\|},\nu_{i,k}\right\rangle\right)$$
$$=\lim_{y\to x}\left(\frac{\kappa}{4}Y_1(\kappa\,\|x-y\|)\left\langle\frac{x-y}{\|x-y\|},\nu_{i,k}\right\rangle\right). \tag{1.6}$$

We replace Y_1 with its power series expansion

$$Y_1(z) = -\frac{2}{\pi z}+\frac{2}{\pi}\ln(\frac{1}{2}z)J_1(z)$$
$$+\frac{1}{\pi}\sum_{k=0}^{\infty}\frac{(-1)^k}{k!(k+1)!}(\psi_0(k+1)+\psi_0(k+2))\left(\frac{1}{2}z\right)^{2k+1}$$

with ψ_0 the digamma function. Of this expansion only the first summand is infinite. The remaining summands again tend to zero as $z \to 0$. This reduces (1.6) to

$$\lim_{y\to x}\left(-\frac{1}{2\pi}\left\langle\frac{x-y}{\|x-y\|^2},\nu_{i,k}\right\rangle\right). \tag{1.7}$$

The limit does not exist for any general vectors x, y, and $\nu_{i,k}$, so further constraints are applied. The points x and y lie on the graph of g_k with $x = g(t_0)$. By Taylor's theorem there exists a function $h : \mathbb{R} \to \mathbb{R}^2$ with $\lim_{t\to t_0} h(t) = 0$ such that

$$g_k(t) = g_k(t_0)+g_k'(t_0)(t-t_0)+\frac{1}{2}g_k''(t_0)(t-t_0)^2+h(t)(t-t_0)^2.$$

Inserting this into (1.7), the remainder term vanishes as it tends to zero. Let $\nu_{i,k}$ be the normal vector onto g_k at t_0. Then, the linear term vanishes as $g_k'(t_0)$ is the tangent onto g_k at t_0 and $\langle g_k'(t_0),\nu_{i,k}\rangle = 0$. The constant terms cancel and only the quadratic term remains. Hence, we are left with

$$\lim_{t\to t_0}\left(\frac{1}{4\pi}\left\langle\frac{g_k''(t_0)(t-t_0)^2}{\|g_k(t_0)-g_k(t)\|^2},\nu_{i,k}\right\rangle\right).$$

The difference quotient tends to the derivative and we find the result

$$\frac{1}{4\pi}\left\langle\frac{g_k''(t_0)}{\|g_k'(t_0)\|^2},\nu_{i,k}\right\rangle.$$

So under the assumptions from the beginning that x_i is part of element k, the singularity in the kernel (1.5) is removable and the integrals (1.4) are proper if $\langle g_k'(t_0),\nu_{i,k}\rangle = 0$. Note that the existence of the integral in (1.2) can be shown

with the same argument. In [Kl19], the normal onto the exact boundary, which is generally not normal to g_k, was used in the numeric computation. This results in infinite singularities that are difficult to handle. Here, we will use normals onto the interpolation polynomials. We denote the normal onto g_k at t_0 as $g_k^{\perp}(t_0)$. For even indices i (which correspond to the midpoint of element $i/2$), we always use $g_{i/2}^{\perp}(1/2)$. For odd indices i, the point x_i is part of two boundary elements and we have $g_{(i+1)/2-1}(1) = g_{(i+1)/2}(0) = x_i$. We have a choice of two different normals at x_i. When we integrate over element $(i+1)/2-1$, we need to choose $g_{(i+1)/2-1}^{\perp}(1)$ and analogously for element $(i+1)/2$. Otherwise the choice is arbitrary, so we alternate to avoid introducing bias. In summary

$$\nu_{i,k} = \begin{cases} g_{i/2}^{\perp}(\frac{1}{2}) & \text{if } i \bmod 2 = 0 \\ g_{(i+1)/2-1}^{\perp}(1) & \text{if } i \bmod 2 = 1 \text{ and } k \bmod 2 = (\frac{i+1}{2} - 1) \bmod 2 \\ g_{(i+1)/2}^{\perp}(0) & \text{if } i \bmod 2 = 1 \text{ and } k \bmod 2 = \frac{i+1}{2} \bmod 2 \end{cases}.$$

Now the integrals can be evaluated without any special handling of the singularity. This makes the quadrature less expensive. We have used a routine that uses the 15 point Gauss–Kronrod rule to solve the integrals to within a relative and absolute tolerance of 10^{-10}.

Exploiting Symmetries of the Domain

The shapes considered in this work are all symmetric to some degree. This fact can be exploited to reduce the required work. In the integrand of the integrals (1.4) with kernel (1.5), the collocation point x_i, normal $\nu_{i,k}$, and integration point $y = g_k(t)$ exclusively exist in norms and scalar products. Those are invariant under rotation:

$$\|\mathbf{R}v\| = \|v\|$$
$$\langle \mathbf{R}v, \mathbf{R}w\rangle = \langle v, w\rangle$$

for all $v, w \in \mathbb{R}^2$ and all rotation transformations $\mathbf{R}$. Under reflection, scalar products switch signs but this is compensated by the switching of the integration bounds, so the resulting integral is again invariant. Thus, if x_a is the image of x_b and x_c is the image of x_d under reflection or rotation, then $\mathbf{M}(\kappa)_{ac} = \mathbf{M}(\kappa)_{bd}$. For this to work, n must be divisible by two times the degree of symmetry and the boundary elements must have the same symmetries as the shape itself (Fig. 1.1).

As an example for what effect this has on matrix $\mathbf{M}(\kappa)$ we will discuss the suspected shape maximizer of λ_3, which has degree of symmetry six (three rotations times two for reflection symmetry, Fig. 1.1). For simplicity, we assume that the first collocation point x_1 lies on a symmetry axis. Threefold rotational symmetry and the corresponding shifting of rows and columns by $n/3$ lead to the matrix having 3×3

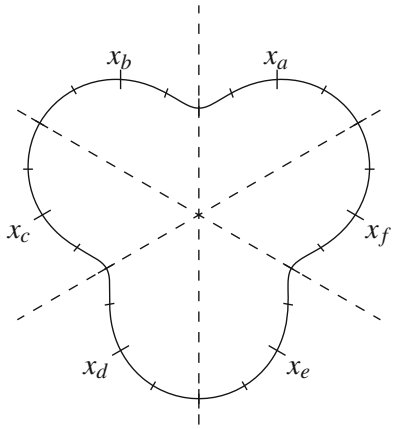


Fig. 1.1 Suspected shape maximizer of λ_3, discretized using $n = 24$ points that form 12 boundary elements. Long ticks mark the endpoints, short ticks the midpoints. The shape has degree of symmetry six, three rotations times two for reflection symmetry. The discretization parameter n is divisible by 12, two times the degree of symmetry. The boundary elements have the same symmetry as the shape itself. The points x_a, x_b, x_c, x_d, x_e, and x_f are images of each other

block structure

$$\mathbf{M}(\kappa) = \begin{pmatrix} \mathbf{A} & \mathbf{B} & \mathbf{C} \\ \mathbf{C} & \mathbf{A} & \mathbf{B} \\ \mathbf{B} & \mathbf{C} & \mathbf{A} \end{pmatrix}$$

with some blocks $\mathbf{A}, \mathbf{B}, \mathbf{C} \in \mathbb{C}^{n/3 \times n/3}$ that each have to be evaluated only once, reducing the required work to a third. Reflection symmetry and the corresponding reflection of both rows and columns of the matrix lead to the matrix being centrally symmetric with respect to the entry at row and column $n/2 + 1$, e.g. for $n = 4$ the matrix

$$\begin{pmatrix} a & e & b & e \\ f & g & h & i \\ c & j & d & j \\ f & i & h & g \end{pmatrix}$$

is point symmetric with respect to entry (3, 3) using periodic indices. Note that this example is just for illustration as $n = 4$ would not be valid for degree of symmetry six due to the restrictions mentioned above. Each value repeats twice except where row index and column index both correspond to fixed points of the reflection, i.e. to points that lie on the symmetry axis. As the boundary elements are symmetric and x_1 lies on the axis, so does $x_{n/2+1}$. In our example, values of entries (1, 1), $(1, n/2 + 1)$, $(n/2 + 1, 1)$, and $(n/2 + 1, n/2 + 1)$ (or a, b, c, and d) do not repeat.

Rotation and reflection symmetry combine so that all values repeat six times, except for the fixed points of reflection which only repeat three times.

Rotation symmetry of degree r reduces the number of matrix entries that must be evaluated to n^2/r. Reflection symmetry further reduces the amount to $n^2/2r+4$. Except for infinitely symmetric shapes like a circle we usually have $n \gg r$, so the complexity of the method with respect to the number of collocation points does not change, but the total amount of work is reduced by a factor approaching $2r$.

1.2.2 Beyn's Contour Integral Method

The eigenvalues of the nonlinear eigenvalue problem (1.3) are computed using the contour integral method W.-J. Beyn presented in [Be12]. For simplicity, we assume here that all eigenvalues are simple, but the method works identically for multiple eigenvalues. Given an operator $\mathbf{T} : \Gamma \to \mathbb{C}^{n\times n}$ that is holomorphic on a domain $\Gamma \subset \mathbb{C}$ and a closed contour $C \subset \Gamma$ with its interior $\mathrm{int}(C) \subset \Gamma$, the method computes all eigenvalues of $\mathbf{T}$ in the interior $\mathrm{int}(C)$.

Let $\kappa_i, i = 1,\ldots,k$ be the eigenvalues of $\mathbf{T}$ in the interior $\mathrm{int}(C)$ and v_i and w_i the corresponding left and right eigenvectors that are normalized so that $w_i^{\mathrm{H}}\mathbf{T}'(\kappa_i)v_i = 1$. Then the equation

$$\frac{1}{2\pi\mathrm{i}}\int_C f(\kappa)\mathbf{T}(\kappa)^{-1}\,\mathrm{d}\kappa = \sum_{i=1}^{k} f(\kappa_i)v_i w_i^{\mathrm{H}} \tag{1.8}$$

holds for all holomorphic functions $f : \Gamma \to \mathbb{C}$ [Be12, Theorem 2.9]. We additionally assume that $k < n$, which is sufficient for our purposes. Beyn describes an extension to the method for $k \geq n$. Applying (1.8) to the functions $f_0(\kappa) = 1$ and $f_1(\kappa) = \kappa$ and multiplying with a random matrix $\mathbf{Z} \in \mathbb{C}^{n\times m}$ from the right yields two equations

$$\mathbf{A}_0 = \int_C \mathbf{T}(\kappa)^{-1}\mathbf{Z}\,\mathrm{d}\kappa = \mathbf{V}\mathbf{W}^{\mathrm{H}}\mathbf{Z} \tag{1.9a}$$

$$\mathbf{A}_1 = \int_C \kappa\mathbf{T}(\kappa)^{-1}\mathbf{Z}\,\mathrm{d}\kappa = \mathbf{V}\mathbf{K}\mathbf{W}^{\mathrm{H}}\mathbf{Z} \tag{1.9b}$$

with $\mathbf{V} = (v_1,\ldots,v_m)$, $\mathbf{W} = (w_1,\ldots,w_m)$, and $\mathbf{K} = \mathrm{diag}(\kappa_1,\ldots,\kappa_m)$. The dimension m is an initial guess for k with $k \leq m \leq n$. Therefore, the matrix $\mathbf{Z}$ reduces the dimensions of $\mathbf{A}_0$ and $\mathbf{A}_1$ without reducing the rank k. Singular value decomposition (SVD) of $\mathbf{A}_0$ in reduced form yields

$$\mathbf{A}_0 = \mathbf{V}_0\mathbf{S}_0\mathbf{W}_0^{\mathrm{H}}$$

with matrices $\mathbf{V}_0 \in \mathbb{C}^{n\times k}$ and $\mathbf{W}_0 \in \mathbb{C}^{m\times k}$ and the diagonal matrix of singular values $\mathbf{S}_0 = \operatorname{diag}(\sigma_1, \ldots, \sigma_k)$. With this, the correctness of the initial guess m can be confirmed by comparing it with the actual computed rank. In case the check is failed, the method is started again with a higher guess m. Finally a matrix

$$\mathbf{B} = \mathbf{V}_0^{\mathrm{H}} \mathbf{A}_1 \mathbf{W}_0 \mathbf{S}_0^{-1} = \mathbf{Q}\mathbf{K}\mathbf{Q}^{-1}$$

is computed. The matrix is diagonalizable by construction with $\mathbf{Q} = \mathbf{V}_0^{\mathrm{H}}\mathbf{V}$ and has eigenvalues $\kappa_i, i = 1, \ldots, k$. Thus, the nonlinear eigenvalue problem is converted into a linear eigenvalue problem. If the eigenvalues are not simple, the method works identically but $\mathbf{K}$ will have Jordan normal form. The structure of multiplicity is preserved. We have enough knowledge about the location of eigenvalues to make the choice of C trivial. All that is left is the discretization of the contour integrals (1.9). Let the contour be described by a smooth mapping $h : [0, 2\pi] \to C$ with $h(0) = h(2\pi)$, e.g. the simplest contour, a circle with center μ and radius r, is described by $h(t) = \mu + r\mathrm{e}^{\mathrm{i}t}$. The interval $[0, 2\pi]$ is partitioned by the equally spaced nodes $t_j = 2\pi j/N$, $j = 1, \ldots, N$ with N a chosen discretization parameter. The approximations

$$\mathbf{A}_0 \approx \frac{1}{N\mathrm{i}} \sum_{j=1}^{N} \mathbf{T}(h(t_j))^{-1}\mathbf{Z}h'(t_j) \tag{1.10a}$$

$$\mathbf{A}_1 \approx \frac{1}{N\mathrm{i}} \sum_{j=1}^{N} h(t_j)\mathbf{T}(h(t_j))^{-1}\mathbf{Z}h'(t_j) \tag{1.10b}$$

are obtained by transforming the integrals onto the partitioned interval and applying the trapezoidal quadrature rule. Beyn shows that the error in the eigenvalues decays exponentially with N [Be12, Corollary 4.8].

The operator $\mathbf{T}$ must be evaluated N times. This is by far the most expensive part of the algorithm. As k is usually much smaller than n, the introduction of the random matrix $\mathbf{Z}$ makes the matrices small enough so that the effort required for linear algebra operations is small. While the solving of linear systems $\mathbf{T}(h(t_j))^{-1}\mathbf{Z}$ is still noticeable, the other operations (SVD, solving the linear eigenvalue problem) are completely negligible.

1.2.3 Parallelization

For optimization with many iterations, the eigenvalue solver is required to be fast. The evaluation of both the matrix $\mathbf{M}(\kappa)$ and the contour integrals in Beyn's method are well suited for parallelization. This covers almost the entire computation. In our implementation for $n = 1152$ and $N = 48$, almost 100% of the time is spent

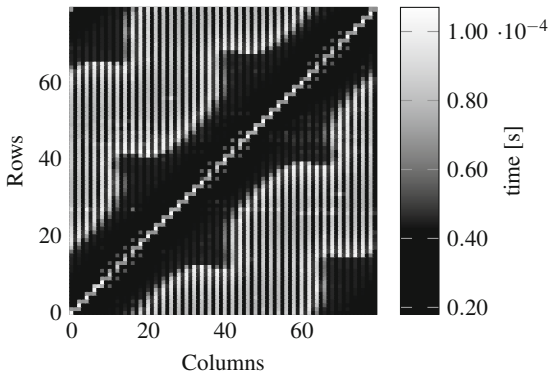


Fig. 1.2 Time required to evaluate each entry of $\mathbf{M}(\kappa)$ for a domain with threefold rotational symmetry. Vertical stripes correspond to alternating end- and midpoints of boundary elements. The singularity of the kernel of the integral equation causes a band around the diagonal whose boundary follows the shape of the domain

on the evaluation of the contour integrals (1.10), of which the evaluation of $\mathbf{M}(\kappa)$ requires 98.8% and solving the linear systems $\mathbf{T}(\kappa)^{-1}\mathbf{Z}$ requires 1.2%. Everything else is completely negligible. While the exact gains that can be expected depend on the implementation and the system, the principles outlined here are universal. Most relevant for performance are the routines that evaluate the Hankel function and perform quadrature. Our program is implemented in C, using GNU Scientific Library (GSL) [Ga09] as a general framework and for quadrature specifically. For the Hankel function we use the FORTRAN routine provided by Amos [Am86].

Each entry of the $n \times n$ matrix can be evaluated independently without any synchronization or communication. Row cyclical distribution is a simple and effective way to balance the workload (Fig. 1.2). Columns that correspond to an endpoint of a boundary element require the evaluation of two integrals, whereas columns that correspond to a midpoint require only one. Additionally, the time necessary to evaluate a single integral depends strongly on the distance from the singularity of the kernel, i.e. the diagonal. Row distribution removes both these imbalances. For shapes less regular than a disk and collocation points that are not perfectly evenly spaced, the number of collocation points that are close to the singularity varies smoothly between rows. So the rows need to be distributed cyclically. Without any communication or synchronization and with 98.8% of the computation parallelized, strong scaling is expected for this strategy. The implementation, e.g. using OpenMP, is trivial.

Each of the N summands of the trapezoidal rule that approximates the contour integrals in Beyn's method can also be evaluated independently without any communication except for one sum reduce operation at the end. In regards to workload balancing, the time to evaluate $\mathbf{T}(\kappa)$ in our implementation depends mainly on the sign of the imaginary part of wave number κ (Fig. 1.3). The integrals involving the Hankel function are more expensive for $\mathrm{Im}(\kappa) < 0$. The eigenvalues are real, so a contour that is centered on the real axis is used. Cyclical distribution of summands is generally solid, although not very flexible regarding the number of tasks. Some tasks may end up with fewer summands from one half space than the other. The parallelized part is slightly larger than for the first strategy as basically

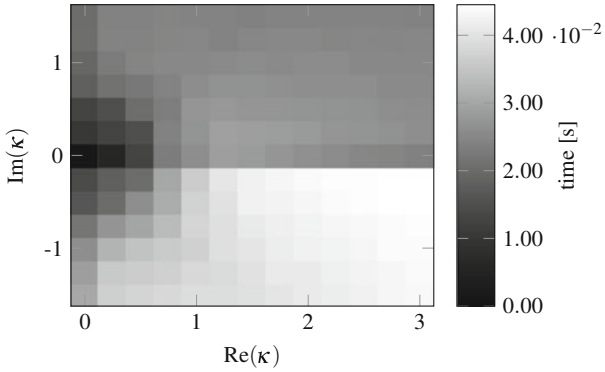


Fig. 1.3 Time required to evaluate $\mathbf{T}(\kappa)$ for different wave numbers κ. The evaluation of integrals involving the Hankel function $H_1^{(1)}$ is significantly more expensive for arguments with negative imaginary part. The singularity at the origin is of no concern as we are not interested in zero eigenvalues

100% of the computation is covered. This improves scaling compared to the first strategy. However, the parameter N is typically not very high (<50) so the degree of parallelism is severely constrained. It is advisable to employ both strategies, e.g. in a hybrid (mixed shared memory and distributed memory) application. Our implementation achieves a speedup of ~520 with 576 physical/1152 virtual cores on the JURECA cluster [Ju18] which likely can be further improved. The time to evaluate the eigenvalues is reduced to below one second, which allows large scale optimization.

1.3 Optimization of Eigenvalues

Recall that the constrained optimization problem we are trying to solve is

$$\begin{aligned} &\max_{\Omega}\{\lambda_k(\Omega)\} \\ &\text{s.t. } |\Omega| = 1 \end{aligned} \tag{1.11}$$

for some fixed $k \in \mathbb{N}$ with $|\Omega|$ denoting the area of the domain Ω. By applying the known relations

$$\lambda_k(a\Omega) = a^{-2}\lambda_k(\Omega)$$
$$|a\Omega| = a^2\,|\Omega|$$

with $a\Omega$ denoting the homothety of Ω by the factor a, we can convert (1.11) into the equivalent unconstrained problem

$$\max_{\Omega}\{\lambda_k\,|\Omega|\}\,.$$

In the numeric treatment, it is sufficient to consider connected domains. The spectrum of a disconnected domain $\Omega_1 \cup \Omega_2$ with $\Omega_1 \cap \Omega_2 = \emptyset$ is the ordered union of the spectrums of the component domains Ω_1 and Ω_2. If the maximums $\lambda_i^* = \max_\Omega\{\lambda_i\}$ and corresponding maximizers Ω_i^* (connected or disconnected) for $i = 1, \ldots, k-1$ are known, then the maximum of λ_k over disconnected domains is

$$\lambda_k^* = \max_{1\le i\le \frac{n}{2}}\{\lambda_i^* + \lambda_{k-i}^*\} \tag{1.12}$$

and the corresponding disconnected maximizer is

$$\Omega_k^* = \left(\sqrt{\frac{\lambda_j^*}{\lambda_j^* + \lambda_{k-j}^*}}\Omega_j^*\right) \cup \left(\sqrt{\frac{\lambda_{k-j}^*}{\lambda_j^* + \lambda_{k-j}^*}}\Omega_{k-j}^*\right)$$

where j is the integer that maximizes (1.12). If the maximum over all domains is greater than (1.12), then the maximizer must be connected [PoRo10].

In [Kl19], Kleefeld introduced equipotentials to parametrize the domain. They are described by the implicit function

$$\sum_{i=1}^{m}\frac{1}{\|x - p_i\|^{2\alpha}} = c \tag{1.13}$$

with m fixed base points p_i and free parameters c and α. To match the shapes reported in [AnOu17, Fig. 2], the base points are chosen so they form equilateral triangles of side length $\sqrt{3}/2$, three points on one triangle to maximize λ_3 and four points on two triangles to maximize λ_4 (see first row of Fig. 1.4). Points on the boundary required by the boundary element method are generated by transforming the equation into polar coordinates and using a root finding algorithm to compute the radiuses r_i for evenly spaced angles ϕ_i. The points (r_i, ϕ_i) are then transformed back into Cartesian coordinates. The area of the domain is computed to high accuracy by approximating the domain as a polygon with $100 \cdot n$ sides. With this method, Kleefeld improved on the values found by Antunes and Oudet with the maximum of λ_3 to 32.9018 over 32.90 and the maximum of λ_4 to 43.8694 over 43.86.

With the improvements for the method of computation presented above, we can now try the scheme on the higher eigenvalues. Unless otherwise noted, we have used discretization parameters $n = 1152$ and $N = 48$. Convergence experiments suggest this is generally enough for six significant digits in the eigenvalues. To be safe, we check the results of optimization with finer discretization. As the

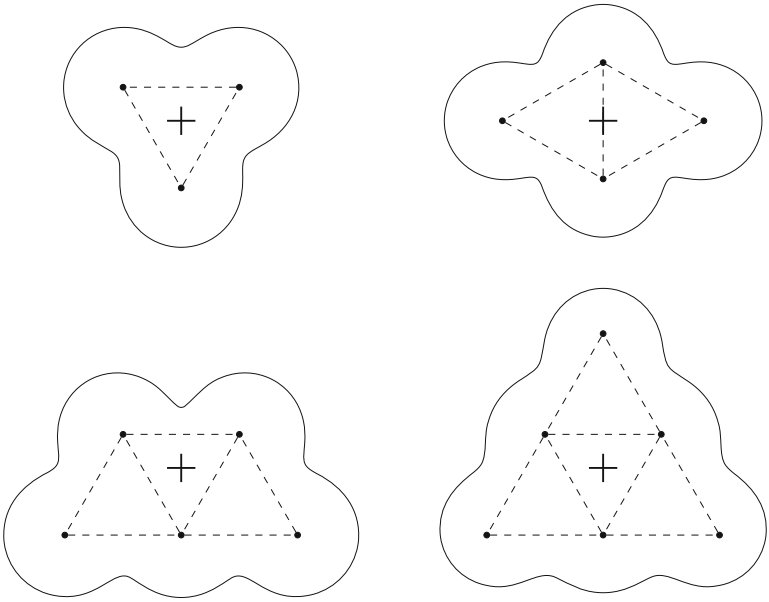


Fig. 1.4 Arrangement of base points for equipotentials to maximize λ_3, λ_4, λ_5, and λ_6. The points form a regular triangular grid. A cross marks the origin and rotation center

gradient of the objective function is not trivially computable, we use a routine provided by GSL that implements the Nelder–Mead simplex method. To avoid local maximums, we try different starting values. For two parameters, it is possible to exhaustively probe the parameter space to find good starting values. The extended, higher dimensional parameter space that is presented later is randomly probed instead. Simplex algorithms are known to terminate prematurely even without the presence of local minimums. So after the optimization routine terminates, it is restarted with the step size reset to its initial value. If the previous result is indeed the maximum (or close enough), the restarted optimization routine quickly terminates again. Eigenvalues are truncated to six significant digits. Shape parameters are given with more significant digits so that the results can be reliably reproduced. The precise error propagation is unknown.

We use the results of Antunes and Oudet [AnOu17] as references for comparison. But there is some uncertainty regarding the precision of those values. The first eigenvalue λ_1 has been proven to be maximized by a disk. The spectrum for the disk can be stated analytically. It is composed of values $\pi j'^2_{pq}$ where j'_{pq} is the q-th positive zero of J'_p, the first derivative of the Bessel function of the first kind of order p. These values can be computed very accurately with root finding algorithms. The first eigenvalue is approximately 10.649866. The second eigenvalue λ_2 is maximized by the union of two disks of the same size. Following the rules for spectrum of disconnected shapes outlined above, the maximum is precisely $2\lambda_1$ or approximately 21.299733. However, the values given in [AnOu17] are 10.66 for λ_1

and 21.28 for λ_2. This discrepancy is not discussed in their paper. While a value lower than the analytical value could be simply caused by incomplete optimization, a higher value calls into question the precision of the eigenvalue solver. This has direct implications for λ_7 as well, for which the maximizer found by Antunes and Oudet is a disjoint union of the maximizers of λ_1 and λ_6. The maximum value for λ_7 therefore should be the sum of maximums for λ_1 and λ_6. It appears like the authors used their own inaccurate value for λ_1, so λ_7 is inaccurate as well. Of course all their results could be affected by an inaccurate solver, so any comparison can only be tentative.

The base points for the maximizers of $\lambda_k, k = 5, 6, 8, 9, 10$ are k points on a regular equilateral triangular grid as they were for $k = 3, 4$ (see Fig. 1.4). For λ_7, Antunes and Oudet did not find a connected shape maximizer. The maximizer can then be constructed from the maximizers for $k < 7$, so it is improved automatically with those. The positive results from [Kl19] have been confirmed with 32.9018, 32.9018, 32.9018 for λ_3 and 43.8693, 43.8693, 43.8693 for λ_4. The parameters differ slightly from [Kl19] with $c = 1.687730810, \alpha = 2.019822714$ and $c = 2.084610015, \alpha = 2.541256146$, respectively. This might be explained by the finer discretization used. Unfortunately, equipotentials work less and less well for higher k and less symmetric shapes. For λ_5, which in [AnOu17] has multiplicity three, we have 54.5401, 54.5401, 56.0889 for $c = 2.380671137$ and $\alpha = 3.914738607$. This is significantly lower than the reference value 55.17 and the multiplicity is not reproduced. For λ_6 (multiplicity four), where the suspected shape is more regular, we get 67.0440, 67.0440, 67.0440, 67.0440 with $c = 2.849410261, \alpha = 0.660868556$, which is a bit closer to the reference value of 67.33 but still some distance away. For λ_{10} (multiplicity five), we even get 109.988 109.988 109.988 118.955 118.955 instead of the previous value 114.16 with $c = 1.567009307, \alpha = 5.196376634$. The eigenvalues for $k = 7, 8, 9$ that have been skipped have not been tried as there was no reason to believe they would fare better.

The results strongly suggest that equipotentials as they are in (1.13) are not general representations of the shape maximizers. They have shown potential but need refinement. So far, the base points of the equipotentials have been arranged on a completely regular triangular grid and all base points are weighted equally. As there is no particular reason for this regularity other than visual intuition, breaking it might prove beneficial. So the base points will be allowed to deviate slightly from their regular position. The weight for the base points in the sum of potentials will be allowed to deviate from the regular weight of one. The imagined balls around base points expand as their weight increases. In general, the boundary of the shape moves away from such points. This final equation reads

$$\sum_{i=1}^{m} \frac{1+\hat{\delta}_i}{\left\|x-(p_i+\hat{\epsilon}_i)\right\|^{2\alpha}} = c$$

where $\hat{\epsilon}_i \in \mathbb{R}^2$ is the irregularity of position and $\hat{\delta}_i \in \mathbb{R}$ is the irregularity of weight of base point i.

Some of the new parameters are fixed so that the shapes are unique and there are no dependencies between parameters. The eigenvalues do not depend on the absolute position of the domain in space, only on the relative position of its base points, so at least one base point should remain fixed during optimization. One weight should remain fixed to avoid a dependency between the weights and parameter c. It is always possible to normalize one weight to a value of one without changing the shape by dividing all weights and c by that weight. The degrees of freedom are further reduced by a requirement that no rotation or reflection symmetries of the regular base points are broken. It must be said that at this point the conjecture that the symmetries are meaningful is unproven. However, based on the results of Antunes and Oudet, the conjecture seems reasonable and it keeps the number of parameters low. So most of the parameters $\hat{\delta}_i$ and $\hat{\epsilon}_i$ will be fixed to zero. For example, the shape for λ_3 will have no additional free parameters. One point must be fixed for uniqueness, the others to preserve rotation symmetry. The shape for λ_4 consists of two pairs of points that are images of each other. One pair must be fixed. The other is free but can only move along the symmetry axis. So there is one free coordinate and one free weight. The free, non-zero parameters will be denoted as $\epsilon_i, i = 1, \ldots, f_\epsilon$ and $\delta_i, i = 1, \ldots, f_\delta$ with f_δ and f_ϵ the number of degrees of freedom (Table 1.2). The free parameters can be assigned to base points almost arbitrarily as long as symmetry is conserved. We included our chosen assignment in the tables of results (Tables 1.3 and 1.4). Due to rotation symmetry, some irregularities of position ϵ_i are not axis aligned but point toward the rotation center. For convenience of implementation we avoided irregularities of position that point away from the domain so that a positive first optimization step does not tear the shape apart.

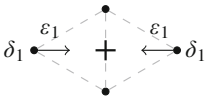
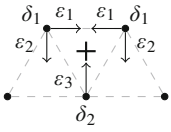
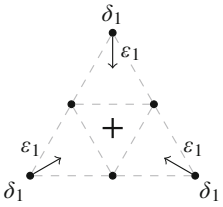
The introduction of ϵ and δ drastically improves the results over just two parameters. Figure 1.5 shows the shapes and optimized eigenvalues. Tables 1.3 and 1.4 show the full numerical results including parameters. The maximum for λ_3 remains unchanged as it did not gain any additional free parameters. The value of λ_4 got another small boost to 43.8700. For $k = 4, 6, 8, 10$, we achieved

Table 1.2 Degree of freedom of positions f_ϵ (each coordinate is counted separately) and weights f_δ and total degree of freedom f (including c, α) of the equipotential that is used to maximize λ_k after symmetry and uniqueness of the shape and independence of parameters is handled

k	f_ϵ	f_δ	f
3	0	0	2
4	1	1	4
5	3	2	7
6	1	1	4
8	6	4	12
9	7	4	13
10	1	2	5

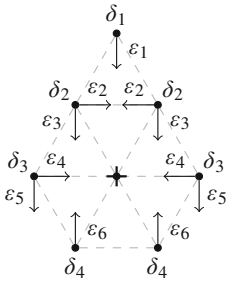
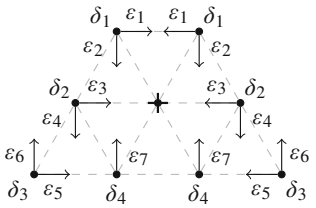
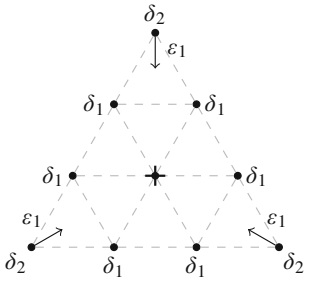
Degree of freedom is generally greater for shapes with fewer symmetries or more base points

Table 1.3 Optimization results for interior Neumann eigenvalues $\lambda_k, k = 3, \ldots, 7$ using extended equipotentials

k	Reference	Maximum	Parameters	Base points and irregularities
3	32.90 (3)	32.9018 32.9018 32.9018	$c = 1.687730810$ $\alpha = 2.019822714$	
4	43.86 (3)	43.8700 43.8700 43.8700	$c = 1.942568636$ $\alpha = 2.751523202$ $\epsilon_1 = -1.314531646 \cdot 10^{-2}$ $\delta_1 = -4.623467053 \cdot 10^{-2}$	
5	55.17 (3)	55.1498 55.1498 55.1498	$c = 1.548694899$ $\alpha = 2.231247849$ $\epsilon_1 = -8.845230330 \cdot 10^{-2}$ $\epsilon_2 = -4.509337199 \cdot 10^{-2}$ $\epsilon_3 = -4.354727490 \cdot 10^{-2}$ $\delta_1 = -1.979312992 \cdot 10^{-1}$ $\delta_2 = -1.671890335 \cdot 10^{-1}$	
6	67.33 (4)	67.3364 67.3364 67.3364 67.3364	$c = 2.027170345$ $\alpha = 1.706097040$ $\epsilon_1 = 1.577407017 \cdot 10^{-1}$ $\delta_1 = 6.001214705 \cdot 10^{-3}$	
7	77.99 (6)	77.9862 77.9862 77.9862 77.9862 77.9862 77.9862	–	–

The second column gives the reference value of [AnOu17] with multiplicity in parentheses. The third column contains the maximal eigenvalue that was found by us and as many of the following eigenvalues as the multiplicity requires. The third column contains the equipotential parameters of the shape maximizer. The figures in the fourth column show the base points of the equipotential and the assignment of free irregularity parameters ϵ and δ. The shape maximizer for λ_7 in [AnOu17] is a disconnected shape that is a union of the shape for λ_6 with a disk and we did not run numerical optimization on it. The values are the sum of the analytical maximum for λ_1 and our new maximum for λ_6

Table 1.4 Optimization results for interior Neumann eigenvalues λ_k, $k = 8, 9, 10$ using extended equipotentials

k	Reference	Maximum	Parameters	Base points and irregularities
8	89.38 (4)	89.8620 89.8620 89.8620 89.8621	$c = 1.942964474$ $\alpha = 1.810828390$ $\epsilon_1 = 1.219776174 \cdot 10^{-1}$ $\epsilon_2 = -9.776658965 \cdot 10^{-2}$ $\epsilon_3 = -4.652290511 \cdot 10^{-2}$ $\epsilon_4 = -6.000769737 \cdot 10^{-2}$ $\epsilon_5 = -7.584457864 \cdot 10^{-2}$ $\epsilon_6 = -2.247915505 \cdot 10^{-1}$ $\delta_1 = 5.396514489 \cdot 10^{-1}$ $\delta_2 = 2.082674393 \cdot 10^{-1}$ $\delta_3 = 1.353703658 \cdot 10^{-1}$ $\delta_4 = 9.643159176 \cdot 10^{-2}$	
9	101.83 (4)	101.752 101.752 101.752 101.752	$c = 1.506287804$ $\alpha = 1.928595020$ $\epsilon_1 = -2.021261311 \cdot 10^{-1}$ $\epsilon_2 = -1.184995442 \cdot 10^{-1}$ $\epsilon_3 = -1.272843752 \cdot 10^{-1}$ $\epsilon_4 = -1.075608953 \cdot 10^{-1}$ $\epsilon_5 = -3.596435931 \cdot 10^{-2}$ $\epsilon_6 = 7.343083116 \cdot 10^{-3}$ $\epsilon_7 = 8.586462162 \cdot 10^{-2}$ $\delta_1 = -3.306712889 \cdot 10^{-2}$ $\delta_2 = 5.598216794 \cdot 10^{-1}$ $\delta_3 = 7.664620451 \cdot 10^{-3}$ $\delta_4 = 1.043695363 \cdot 10^{0}$	
10	114.16 (5)	114.187 114.187 114.187 114.187 114.187	$c = 0.899837214$ $\alpha = 2.708323325$ $\epsilon_1 = -4.458971106 \cdot 10^{-2}$ $\delta_1 = 1.150148658 \cdot 10^{0}$ $\delta_2 = -2.824155602 \cdot 10^{-1}$	

The second column gives the reference value of [AnOu17] with multiplicity in parentheses. The third column contains the maximal eigenvalue that was found by us and as many of the following eigenvalues as the multiplicity requires. The third column contains the equipotential parameters of the shape maximizer. The figures in the fourth column show the base points of the equipotential and the assignment of free irregularity parameters ϵ and δ

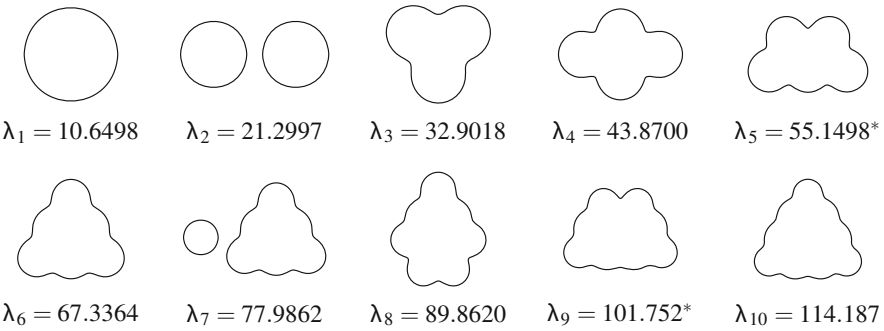


Fig. 1.5 Shape maximizers for interior Neumann eigenvalues $\lambda_k, k = 1, \ldots, 10$. An asterisk marks values where the reference value by [AnOu17] has not been matched or exceeded. The first two eigenvalues, which are proven theoretically, are included for completeness. For $k = 3, 4, 5, 6, 8, 9, 10$ the maximizers were found by optimizing the (extended) parameters of equipotentials. The maximizer for the seventh eigenvalue is a union of the scaled maximizers for the first and sixth eigenvalues and was not optimized on its own. All shapes have been scaled so they have the same area

higher maximums than the reference value, sometimes by a small amount like $\lambda_6 = 67.3364$ over 67.33 that could be interpreted as just an increase in precision, sometimes significantly so with $\lambda_8 = 89.8620$ over 89.38 and $\lambda_{10} = 114.187$ over 114.16. Using the improved value for λ_6 and the precise value for λ_1 (see above), λ_7 can also be considered improved even though the reference value is higher. The result for λ_5 is now much closer than it was using just two parameters but is still short of the reference value by about $2 \cdot 10^{-2}$. The difference is small enough that it may still be caused by an inaccurate reference value. For λ_9, which is also too small, the distance to the reference value is almost certainly too big to be explained in that way. Maybe coincidentally, similar trapezoid shapes are used in both cases (λ_9 and λ_5) where the reference value has not been matched or exceeded.

The multiplicities given in [AnOu17] have not been precisely reproduced in all cases. For example there is a small gap in between the values for λ_8. For almost all shapes, the eigenvalues have multiplicities one or two. With changing shapes, some of those groups of identical eigenvalues increase, others decrease. In most cases, both Antunes and Oudet and us have found the optimum where two groups merge, producing multiplicities of three or four. Note that for unions of disjoint shapes, higher multiplicities are expected, as the multiplicities of the component shapes accumulate. Connected shapes where more than two groups merge may not exist and the values for λ_8 may simply be a near miss, where three groups almost merge. On the other hand it is possible that we are simply not able to represent such shapes with the chosen parametrization or that the optimization routine missed them.

1.4 Conclusion

We have presented a way to efficiently and precisely compute interior Neumann eigenvalues for two dimensional domains. Along the way, we highlighted a few techniques to reduce the time to solution. The strongly scaling parallelization in particular allowed us to use the implemented solvers in the optimization of the eigenvalues with respect to the shape of the domain. We refined the parametrization of the shapes developed by previous research and found improved maximums for most of the first ten eigenvalues.

The very specialized parametrization we presented requires fewer parameters than more general approaches like Fourier series. This makes numerical optimization much cheaper. But the new parametrization is unfortunately far from compact, especially for higher eigenvalues. It is therefore unlikely to be helpful in any theoretical proof of shape maximizers. Numerical optimization was also not equally successful in all cases. The general idea seems promising, but further adaptations will be necessary. Ultimately, an entirely new idea might be called for.

It should prove insightful to study even higher eigenvalues than in this work. Both the method of solution and the parametrization can also be extended without great modifications to three dimensions. Similar results as for the interior Neumann problem also exist for Dirichlet boundary conditions. So research similar to the one in this work is possible in that area. We have concentrated on acoustic scattering. One could also study electromagnetic or elastic problems.

The source code for the program is available at the URL below. We invite researchers to check the program, extend it, or use it in their own research.

```
https://gitlab.version.fz-juelich.de/abele2/shapeopt
```

Acknowledgments The authors gratefully acknowledge the computing time provided on the supercomputer JURECA at Jülich Supercomputing Centre (JSC).

References

[Am86] Amos, D.E.: Algorithm 644: a portable package for Bessel functions of a complex argument and nonnegative order. ACM Trans. Math. Softw. **12**(3), 265–273 (1986)

[AnFr12] Antunes, P.R.S., Freitas, P.: Numerical optimization of low eigenvalues of the Dirichlet and Neumann Laplacians. J. Optim. Theory Appl. **154**, 235–257 (2012)

[AnOu17] Antunes, P.R.S., Oudet, E.: Numerical results for extremal problem for eigenvalues of the Laplacian. In: Henrot, A. (ed.) Shape Optimization and Spectral Theory, pp. 398–412. De Gruyter, Warzow/Berlin (2017)

[Be12] Beyn, W.-J.: An integral method for solving nonlinear eigenvalue problems. Linear Algebra Appl. **436**, 3839–3863 (2012)

[CoKr83] Colton, D., Kress, R.: Integral Equation Methods in Scattering Theory (Wiley, New York, 1983)

[CoKr13] Colton, D., Kress, R.: Inverse Acoustic and Electromagnetic Scattering Theory. Springer, New York (2013)

[Ga09] Galassi, M., Davies, J., Theiler, J., Gough, B., Jungman, G., Alken, P., Booth, M., Rossi, F.: GNU Scientific Library Reference Manual, 3rd edn. Network Theory Ltd., Bristol (2009)

[GiNaPo09] Girouard, A., Nadirashvili, N., Polterovich, I.: Maximization of the second positive Neumann eigenvalue for planar domains. J. Differ. Geom. **83**, 637–662 (2009)

[Ju18] Jülich Supercomputing Centre: JURECA: Modular supercomputer at Jülich Supercomputing Centre. J. Large-Scale Res. Facil. **4**, A132 (2018)

[Kl19] Kleefeld, A.: Shape optimization for interior Neumann and transmission eigenvalues. In: Constanda, C., Harris, P. (eds.) Integral Methods in Science and Engineering, pp. 185–196. Birkhäuser, Cham (2019)

[PoRo10] Poliquin, G., Roy-Fortin, G.: Wolf-Keller theorem for Neumann eigenvalues. Ann. Sci. Math. Québec **36**, 169–178 (2012)

[Sz54] Szegö, G.: Inequalities for certain eigenvalues of a membrane of given area. Arch. Ration. Mech. Anal. **3**, 343–356 (1954)

[We56] Weinberger, H.F.: An isoperimetric inequality for the N-dimensional free membrane problem. Arch. Ration. Mech. Anal. **5**, 633–636 (1956)

Chapter 2
Transient Convection-Diffusion-Reaction Problems with Variable Velocity Field by Means of DRBEM with Different Radial Basis Functions

Salam Adel Al-Bayati and Luiz C. Wrobel

2.1 Introduction

The solution of convection-diffusion-reaction problems is a difficult task for all numerical methods because of the nature of the governing equation, which includes first-order and second-order partial derivatives in space [PaEtAl92, AlWr17, Wr02, Al02, BrEtAl12, AlWr18a, AlWr18b, AlWr19]. The convection-diffusion equation is the basis of many physical and chemical phenomena, and its use has also spread in economics, financial forecasting and other fields [Mo96]. The DRBEM, initially applied to transient heat conduction problems by Wrobel et al. [WoEtAl86], interprets the time derivative in the diffusion equation as a body force and employs the fundamental solution to the corresponding steady-state equation to generate a boundary integral equation. When the steady-state fundamental solution is used in the DRBEM to approximate transient problems, other techniques should be employed to approximate the solution's functional dependence on the temporal variables. Aral and Tang [ArTa89] used the fundamental solution of the Laplace equation, but made use of a secondary reduction process, called SR-BEM, to arrive at a boundary-only formulation. They presented the results of transient convection-diffusion problems with or without first-order chemical reaction for low to moderate

S. A. Al-Bayati (✉)
College of Sciences, Department of Mathematics and Computer Applications, AL-Nahrain University, Baghdad, Iraq
e-mail: salam_ahmed@sc.nahrainuniv.edu.iq

L. C. Wrobel
Institute of Materials and Manufacturing, Brunel University London, Uxbridge, UB8 3PH, UK

Department of Civil and Environmental Engineering, Pontifical Catholic University of Rio de Janeiro (PUC-Rio), Brazil
e-mail: Luiz.Wrobel@brunel.ac.uk

C. Constanda (ed.), *Computational and Analytic Methods in Science and Engineering*, https://doi.org/10.1007/978-3-030-48186-5_2

Péclet numbers. Martin [Ma05] proposed a Schwartz waveform relaxation algorithm for the unsteady diffusive-convective equation, which uses domain decomposition methods and applies an iterative algorithm directly to the time-dependent problem. Partridge and Sensale [PaSe00] have used the method of fundamental solution with dual reciprocity and subdomain approach to solve convection-diffusion problems. The time integration scheme is the FDM with a relaxation procedure, which is iterative in nature and needs a carefully selected time increment. Regarding the DRBEM formulation presented in this work, a backward finite difference scheme is adopted, Smith [Sm85].

In this article, the DRBEM is also employed to discretise the spatial partial derivatives in the two-dimensional diffusive-convective-reactive type problem. Thus, the problem is ultimately described in terms of boundary values only, consequently reducing its dimensionality by one [WrDe91]. We use the fundamental solution to the steady-state convection-diffusion-reaction equation and transform the domain integral arising from the time derivative term using a set of coordinate functions and particular solutions which satisfy the associated non-homogeneous steady-state convection-diffusion-reaction problem. Further, only a simple set of cubic radial basis functions has been previously used in this formulation. We consider two other sets of coordinate functions, non-augmented thin-plate spline (TPS) and multiquadric (MQ) radial basis functions, and analyse their performance in conjunction with the order of time integration algorithms for convection-diffusion-reaction problems. This work also focuses on the search for the optimal shape parameter when utilising the multiquadric radial basis function (MQ-RBF). This is due to the lack of information on choosing the best shape parameter, forcing the user having to make an 'ad-hoc' decision. Recent numerical experiments available in the literature, nevertheless, showed that the MQ-RBF has shown great potential when dealing with complicated PDEs in two dimensions if an adequate shape value is provided.

A brief outline of the rest of this paper is as follows. Section 2.2 reviews the mathematical representation of convection-diffusion- reaction problems. Section 2.3 derives the boundary element formulation of the governing equation using the steady-state fundamental solution of the corresponding equation. In Sects. 2.4 and 2.5, the DRBEM formulation and its discretisation are developed for the 2D transient convection-diffusion-reaction problem. A two-level time marching procedure for the proposed model is implemented in Sect. 2.6. Section 2.7 gives the description of the coordinate functions and the choice of the three radial basis functions. Section 2.8 compares and investigates the solution profiles for the present numerical experiments with the analytical solution of the tested cases. Computational aspects are included to demonstrate the performance of the approach in Sect. 2.9. Finally, some conclusions and remarks are provided in the last section.

2.2 Governing Equation

The two-dimensional transient convection-diffusion-reaction problem over a domain Ω in $\mathbb{R}^2$ bounded by a boundary Γ, for isotropic materials, is governed by the following PDE:

$$D\nabla^2\phi\,(x,y,t) - v_x\,(x,y)\,\frac{\partial\phi\,(x,y,t)}{\partial x} - v_y\,(x,y)\,\frac{\partial\phi\,(x,y,t)}{\partial y} - k\,\phi\,(x,y,t) = \frac{\partial\phi\,(x,y,t)}{\partial t}, \quad (x,y)\in\Omega, \quad t>0. \tag{2.1}$$

In Eq. (2.1), ϕ represents the concentration of a substance, treated as a function of space and time. The velocity components v_x and v_y along the x and y directions are assumed to vary in space. Besides, D is the diffusivity coefficient and k represents the first-order reaction constant or adsorption coefficient. The boundary conditions are

$$\phi = \bar{\phi} \quad \text{over} \quad \Gamma_D$$

$$q = \frac{\partial\phi}{\partial n} = \bar{q} \quad \text{over} \quad \Gamma_N,$$

where Γ_D and Γ_N are the Dirichlet and Neumann parts of the boundary with $\Gamma = \Gamma_D \cup \Gamma_N$, and $\Gamma_D \cap \Gamma_N = \emptyset$ (see Fig. 2.1). The initial condition over the domain Ω is

$$\phi\,(x,y,t=0) = \phi_0\,(x,y)\,, \quad (x,y)\in\Omega.$$

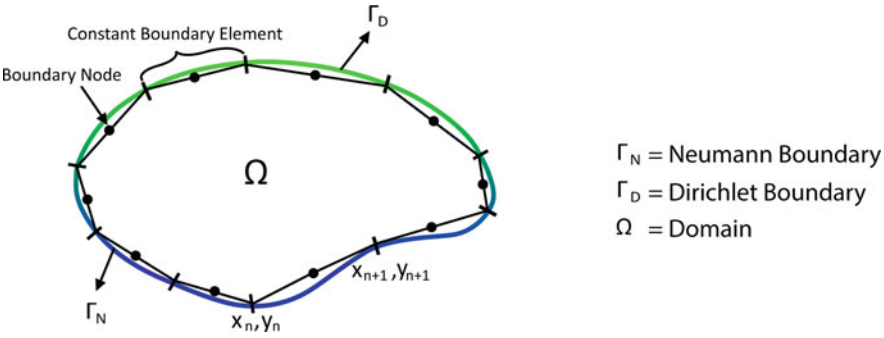


Fig. 2.1 Definition of domain, boundary and constant elements

The parameter that describes the relative influence of the convective and diffusive components is called the Péclet number, Pé $= |v|\, L/D$, where $v = \left(v_x^2 + v_y^2\right)^{1/2}$ is the velocity and L is a characteristic length of the domain. For small values of Pé, Eq. (2.1) behaves as a parabolic differential equation, while for large values the equation becomes more like hyperbolic. These changes in the structure of the PDE according to the values of the Péclet number have significant effects on its numerical solution.

2.3 BEM Formulation of Transient Convection-Diffusion-Reaction Problems

Let us consider a region $\Omega \subset \mathbb{R}^2$ bounded by a piecewise smooth boundary Γ. The transport of ϕ in the presence of a reaction term is governed by the two-dimensional transient convection-diffusion-reaction Eq. (2.1). The variable ϕ can be interpreted as temperature for heat transfer problems, concentration for dispersion problems, etc., and will be herein referred to as a potential. For the sake of obtaining an integral equation equivalent to the above PDE, a fundamental solution of Eq. (2.1) is necessary. However, fundamental solutions are only available for the case of constant velocity fields. At this stage, the variable velocity components $v_x = v_x(x, y)$ and $v_y = v_y(x, y)$ are decomposed into average (constant) terms $\bar{v}_x$ and $\bar{v}_y$, and perturbations $P_x = P_x\,(x, y)$ and $P_y = P_y\,(x, y)$, such that

$$v_x\,(x, y) = \bar{v}_x + P_x\,(x, y) \qquad v_y\,(x, y) = \bar{v}_y + P_y\,(x, y)\,.$$

Now, we can re-write Eq. (2.1) to take the form

$$D\nabla^2\phi\,(x, y, t) - \bar{v}_x \frac{\partial \phi\,(x, y, t)}{\partial x} - \bar{v}_y \frac{\partial \phi\,(x, y, t)}{\partial y} - k\,\phi\,(x, y, t) = \frac{\partial \phi\,(x, y, t)}{\partial t} + P_x \frac{\partial \phi\,(x, y, t)}{\partial x} + P_y \frac{\partial \phi\,(x, y, t)}{\partial y}. \tag{2.2}$$

Next, one can transform the differential equation (2.2) into an equivalent integral equation as follows [WrDe91]:

$$\phi\,(\xi) - D \int_\Gamma \phi^* \frac{\partial \phi}{\partial n}\, d\Gamma + D \int_\Gamma \phi\, \frac{\partial \phi^*}{\partial n}\, d\Gamma + \int_\Gamma \phi\, \phi^*\, \bar{v}_n\, d\Gamma = -\int_\Omega \left[\frac{\partial \phi}{\partial t} + \left(P_x \frac{\partial \phi}{\partial x} + P_y \frac{\partial \phi}{\partial y}\right)\right] \phi^*\, d\Omega,\ \ \xi \in \Omega, \tag{2.3}$$

where $\bar{v}_n = \mathbf{v}.\mathbf{n}$, $\mathbf{n}$ is the unit outward normal vector and the dot stands for scalar product and $v = \left(v_x, v_y\right)$. In the above equation, ϕ^* is the fundamental solution of the steady-state convection-diffusion-reaction equation with constant coefficients. For two-dimensional problems, ϕ^* is given by

$$\phi^*(\xi, \chi) = \frac{1}{2\pi D} e^{-\left(\frac{\bar{v}.\mathbf{r}}{2D}\right)} K_0(\mu \mathbf{r}), \quad \mathbf{r} = |\xi - \chi|,$$

where

$$\mu = \left[\left(\frac{|\bar{v}|}{2D}\right)^2 + \frac{k}{D}\right]^{\frac{1}{2}}, \quad \bar{v} = \left(\bar{v}_x, \bar{v}_y\right),$$

in which ξ and χ are the source and field points, respectively, and r is the modulus of $\mathbf{r}$, the distance vector between the source and field points. The derivative of the fundamental solution with respect to the outward normal is given by

$$\frac{\partial \phi^*}{\partial n} = \frac{1}{2\pi D} e^{-\left(\frac{\bar{v}.\mathbf{r}}{2D}\right)} \left[-\mu K_1(\mu \mathbf{r}) \frac{\partial r}{\partial n} - \frac{\bar{v}_n}{2D} K_0(\mu \mathbf{r})\right].$$

In the above, K_0 and K_1 are Bessel functions of second kind, of orders zero and one, respectively. The exponential term is responsible for the inclusion of the correct amount of 'upwind' into the formulation [RaŠk13]. Equation (2.3) is valid for source points ξ inside the domain Ω. A similar expression can be obtained, by implementing Green's second identity and a limit analysis, for source points ξ on the boundary Γ, in the form

$$c(\xi)\phi(\xi) - D\int_\Gamma \phi^* \frac{\partial \phi}{\partial n} d\Gamma + D\int_\Gamma \phi \frac{\partial \phi^*}{\partial n} d\Gamma + \int_\Gamma \phi \phi^* \bar{v}_n d\Gamma$$

$$= -\int_\Omega \frac{\partial \phi}{\partial t} \phi^* d\Omega - \int_\Omega \left(P_x \frac{\partial \phi}{\partial x} + P_y \frac{\partial \phi}{\partial y}\right) \phi^* d\Omega, \quad \xi \in \Gamma, \tag{2.4}$$

in which $c(\xi)$ is a function of the internal angle the boundary Γ makes at point ξ.

2.4 Standard Approach: DRBEM

In this section, we will discuss the transformation of the domain integral in Eqs. (2.3) and (2.4), and the DRBEM will be implemented to approximate the two domain integrals appearing in this formulation, first the domain integral of the time

derivative, and second the domain integral related to the velocity perturbation parts. Now, we start by expanding the time derivative $\frac{\partial \phi}{\partial t}$ in the form

$$\frac{\partial \phi\left(x, y, t\right)}{\partial t} = \sum_{i=1}^{n} f_i\left(x, y\right) \alpha_{1,i}\left(t\right). \tag{2.5}$$

The above series involves a set of known coordinate functions f_i and a set of unknown time-dependent coefficients $\alpha_{1,i}$. With this approximation, the first domain integral in Eq. (2.4) becomes

$$\int_{\Omega} \frac{\partial \phi}{\partial t} \phi^{*} d\Omega = \sum_{i=1}^{n} \alpha_{1,i} \int_{\Omega} f_i \phi^{*} d\Omega. \tag{2.6}$$

The next step is to consider that, for each function f_i, there exists a related function ψ_i which is a particular solution of the equation:

$$D\nabla^2 \psi - \bar{v}_x \frac{\partial \psi}{\partial y} - \bar{v}_y \frac{\partial \psi}{\partial x} - k\psi = f. \tag{2.7}$$

Now, the domain integral (2.6) can be recast in the form:

$$\int_{\Omega} \frac{\partial \phi}{\partial t} \phi^{*} d\Omega = \sum_{k=1}^{M} \alpha_{1,k} \int_{\Omega} \left(D\nabla^2 \psi_k - \bar{v}_x \frac{\partial \phi_k}{\partial y} - \bar{v}_y \frac{\partial \phi_k}{\partial x} - k\psi_k \right) \phi^{*} d\Omega. \tag{2.8}$$

Substituting expression (2.8) into (2.4), applying integration by parts to the right side of the resulting equation and doing some simplifications, one finally arrives at a boundary integral equation of the form

$$c\left(\xi\right)\phi\left(\xi\right) - D\int_{\Gamma} \phi^{*} \frac{\partial \phi}{\partial n} d\Gamma + D\int_{\Gamma} \phi \frac{\partial \phi^{*}}{\partial n} d\Gamma + \int_{\Gamma} \phi\, \phi^{*}\, \bar{v}_n\, d\Gamma$$

$$= \sum_{k=1}^{M} \alpha_{1,k} \left[c\left(\xi\right)\psi_k\left(\xi\right) - D\int_{\Gamma} \phi^{*} \frac{\partial \psi_k}{\partial n} d\Gamma + D\int_{\Gamma} \left[\left(\frac{\partial \phi^{*}}{\partial n} + \frac{\bar{v}_n}{D} \phi^{*} \right) \psi_k \right] d\Gamma \right]$$

$$- \int_{\Omega} \left(P_x \frac{\partial \phi_k}{\partial x} + P_y \frac{\partial \phi_k}{\partial y} \right) \phi^{*}\, d\Omega, \quad \xi \in \Gamma. \tag{2.9}$$

2.5 Discretization

To discretise the spatial domain, boundary elements were employed. The integrals over the boundary are approximated by a summation of integrals over individual boundary elements. For the numerical solution of the problem, Eq. (2.9) is discretized in the form

$$c_i\,\phi_i - D\sum_{j=1}^{N}\int_{\Gamma_j}\phi^*\,\frac{\partial\phi}{\partial n}\,d\Gamma + D\sum_{j=1}^{N}\int_{\Gamma_j}\left(\frac{\partial\phi^*}{\partial n}+\frac{\bar{v}_n}{D}\,\phi^*\right)\phi\,d\Gamma$$

$$=\sum_{k=1}^{M}\alpha_{1,k}\left[c_i\,\psi_{ik}\,(\xi) - D\sum_{j=1}^{N}\int_{\Gamma_j}\phi^*\frac{\partial\psi_k}{\partial n}\,d\Gamma + D\sum_{j=1}^{N}\int_{\Gamma_j}\left[\left(\frac{\partial\phi^*}{\partial n}+\frac{\bar{v}_n}{D}\,\phi^*\right)\psi_k\,d\Gamma\right]\right]$$

$$-\int_{\Omega}\left(P_x\,\frac{\partial\phi_k}{\partial x} + P_y\,\frac{\partial\phi_k}{\partial y}\right)\phi^*\,d\Omega, \tag{2.10}$$

where the index i means the values at the source point ξ and N elements have been employed. The domain integral on the right-hand side prevents us from obtaining a boundary-only equation.

Now, in order to obtain a boundary integral which is equivalent to the domain integral in expressions (2.9) and (2.10), a dual reciprocity approximation is again implemented [AlWr17]. Applying this to the domain integral of Eq. (2.10), the expression will be expanded in the form

$$P_x\,(x,y)\,\frac{\partial\phi}{\partial x} + P_y\,(x,y)\,\frac{\partial\phi}{\partial y} = \sum_{k=1}^{N}\alpha_{2,k}\,(t)\,f_k. \tag{2.11}$$

Expression (2.11) contains two diagonal matrices $P_x = \left(P_x\,(x_i,y_i)\,\delta_{i,j}\right)_{i,j=\overline{1,M}}$ and $P_y = \left(P_y\,(x_i,y_i)\,\delta_{i,j}\right)_{i,j=\overline{1,M}}$ while

$$\frac{\partial\phi}{\partial x} = \left(\frac{\partial\phi\,(x_i,y_i)}{\partial x}\right)^T_{i=\overline{1,M}}, \quad \frac{\partial\phi}{\partial y} = \left(\frac{\partial\phi\,(x_i,y_i)}{\partial x}\right)^T_{i=\overline{1,M}}$$

are column vectors and $\delta_{i,j}$ is the Kronecker delta symbol. Integrating Eq. (2.11) we obtain

$$\int_{\Omega}\left(P_x\frac{\partial\phi}{\partial x}+P_y\frac{\partial\phi}{\partial y}\right)\phi^*d\Omega = \sum_{k=1}^{M}\alpha_{2,k}\,(t)\int_{\Omega}f_k\,\phi^*d\Omega. \tag{2.12}$$

Now, substituting Eq. (2.12) into (2.10), we obtain

$$c_i\,\phi_i - D\sum_{j=1}^{N}\int_{\Gamma_j}\phi^*\,\frac{\partial\phi}{\partial n}\,d\Gamma + D\sum_{j=1}^{N}\int_{\Gamma_j}\left(\frac{\partial\phi^*}{\partial n}+\frac{\bar{v}_n}{D}\,\phi^*\right)\phi\,d\Gamma$$

$$= \sum_{k=1}^{M}\alpha_{1,k}\left[c_i\,\psi_{ik}\,(\xi) - D\sum_{j=1}^{N}\int_{\Gamma_j}\phi^*\frac{\partial\psi_k}{\partial n}\,d\Gamma + D\sum_{j=1}^{N}\int_{\Gamma_j}\left[\left(\frac{\partial\phi^*}{\partial n}+\frac{\bar{v}_n}{D}\,\phi^*\right)\psi_k\,d\Gamma\right]\right]$$

$$-\sum_{j=1}^{M}\alpha_{2,j}\,(t)\int_{\Omega} f_k\,\phi^*\,d\Omega.$$

The next step is to consider that, for each function f_k, there exists a related function ψ_k which represents the particular solution as in Eq. (2.7). We get

$$\int_{\Omega}\left(P_x\frac{\partial\phi}{\partial x}+P_y\frac{\partial\phi}{\partial y}\right)\phi^*\,d\Omega = \sum_{k=1}^{M}\alpha_{2,k}\int_{\Omega}\left(D\nabla^2\psi_k - \bar{v}_x\,\frac{\partial\phi_k}{\partial y} - \bar{v}_y\,\frac{\partial\phi_k}{\partial x} - k\psi_k\right)\phi^*\,d\Omega. \tag{2.13}$$

Substituting Eq. (2.13) into expression (2.9), and applying integration by parts to the domain integral of the resulting equation, one finally arrives at a boundary integral equation of the form

$$c_i\,\phi_i - D\sum_{j=1}^{N}\int_{\Gamma_j}\phi^*\,\frac{\partial\phi}{\partial n}\,d\Gamma + D\sum_{j=1}^{N}\int_{\Gamma_j}\left(\frac{\partial\phi^*}{\partial n}+\frac{\bar{v}_n}{D}\,\phi^*\right)\phi\,d\Gamma$$

$$= \sum_{k=1}^{M}\alpha_{1,k}\left[c_i\,\psi_{ik}\,(\xi) - D\sum_{j=1}^{N}\int_{\Gamma_j}\phi^*\frac{\partial\psi_k}{\partial n}\,d\Gamma + D\sum_{j=1}^{N}\int_{\Gamma_j}\left[\left(\frac{\partial\phi^*}{\partial n}+\frac{\bar{v}_n}{D}\,\phi^*\right)\psi_k\,d\Gamma\right]\right]$$

$$-\sum_{k=1}^{N}\alpha_{2,k}\left[c_i\,\psi_{ik}\,(\xi) - D\sum_{j=1}^{N}\int_{\Gamma_j}\phi^*\,\frac{\partial\psi_k}{\partial n}\,d\Gamma + \sum_{j=1}^{N}\int_{\Gamma_j}\left(\frac{\partial\phi^*}{\partial n}+\frac{\bar{v}_n}{D}\phi^*\right)\psi_k\,d\Gamma\right]. \tag{2.14}$$

Applying Eq. (2.14) to all boundary nodes using a collocation technique, taking into account the previous functions, results in the following system of algebraic equations:

$$H\phi - Gq = (H\psi - G\eta)\,\alpha_1\,(t) + (H\psi - G\eta)\,\alpha_2\,(t)\,. \tag{2.15}$$

In the above system, the same matrices H and G are used on both sides. Matrices ψ and η are also geometry-dependent square matrices (assuming, for simplicity, that the number of terms in Eq. (2.5) is equal to the number of boundary nodes), and ϕ, q, and α are vectors of nodal values. The next step in the formulation is to find an expression for the unknown vectors α. By applying expressions (2.5) and (2.11) to all boundary nodes and inverting, one arrives at:

$$\alpha_1 = F^{-1} \frac{\partial \phi}{\partial t},$$

and

$$\alpha_2 = F^{-1} \left(P_x (x, y) \frac{\partial \phi}{\partial x} + P_y (x, y) \frac{\partial \phi}{\partial y} \right),$$

which, substituted into (2.15) results in:

$$H\phi - Gq = (H\psi - G\eta) F^{-1} \frac{\partial \phi}{\partial t} + (H\psi - G\eta) F^{-1} \left(P_x \frac{\partial \phi}{\partial x} + P_y \frac{\partial \phi}{\partial y} \right).$$

Calling:

$$C = (H\psi - G\eta) F^{-1},$$

gives

$$H\phi - Gq = C \frac{\partial \phi}{\partial t} + C \left(P_x \frac{\partial \phi}{\partial x} + P_y \frac{\partial \phi}{\partial y} \right). \tag{2.16}$$

Next, we shall explain how to deal with the convective terms in Eq. (2.16).

2.6 Handling Convective Terms

In the present section, emphasis will be placed on the treatment of the convective terms. A mechanism must be established to relate the nodal values of ϕ to the nodal values of its derivatives.

Let us assume that the function ϕ can be represented by

$$\phi (x, y) = \sum_{k=1}^{M} \gamma_k (x, y) \beta_k, \tag{2.17}$$

where $\gamma_k (x, y)$ are known functions and β_k are constants. The upper bound M stands for the total number of terms in the series, i.e. boundary and internal points.

Now, by differentiating it with respect to x and y produces

$$\frac{\partial \phi}{\partial x} = \sum_{k=1}^{M} \frac{\partial \gamma_k}{\partial x} \beta_k \quad \text{and} \quad \frac{\partial \phi}{\partial y} = \sum_{k=1}^{M} \frac{\partial \gamma_k}{\partial y} \beta_k. \tag{2.18}$$

Applying Eq. (2.17) at all M nodes, a set of equations is produced that can be represented in matrix form by

$$\phi = \gamma \ \beta$$

with corresponding matrix equations for expressions (2.18) given as

$$\frac{\partial \phi}{\partial x} = \frac{\partial \gamma}{\partial x} \gamma^{-1} \phi \quad \text{and} \quad \frac{\partial \phi}{\partial y} = \frac{\partial \gamma}{\partial y} \gamma^{-1} \phi. \tag{2.19}$$

Therefore, substituting Eq. (2.19) into Eq. (2.16), the new expression will be

$$(H - P)\, \phi - Gq = C \frac{\partial \phi}{\partial t}, \tag{2.20}$$

where

$$P = C \left(P_x \frac{\partial \gamma}{\partial x} + P_y \frac{\partial \gamma}{\partial y} \right) \gamma^{-1}.$$

The coefficients of the diagonal perturbation matrix P are all geometry-dependent only. The differential algebraic system (2.20) has a form similar to the one obtained using the finite element method (FEM) and hence, can be solved by any standard time integration algorithm by incorporating suitable modifications to account for its *mixed nature*. It should be stressed that the coefficients of matrices H, G and C all depend on geometry only, thus they can be computed once and stored.

2.7 Time Marching Solution Scheme

This section will show how to handle the linear algebraic system (2.20) adopting time marching schemes [PaEtAl92, CaEtAl10, DiKa04]. A finite difference approximation for the time derivative term is given by

$$\frac{\partial \phi}{\partial t} = \frac{\phi^{i+1} - \phi^{i}}{\Delta t}.$$

Let us assume a linear variation of ϕ and q according to

$$\phi(t) = \left(1 - \theta_\phi\right)\phi^i + \theta_\phi \phi^{i+1},$$

$$q(t) = \left(1 - \theta_q\right) q^i + \theta_q q^{i+1},$$

where θ_ϕ and θ_q are parameters which position the values of ϕ and q between time levels m and $m+1$, and take values in the interval $0 \leq \theta_\phi, \theta_q \leq 1$. Next, employing a general two-level time integration scheme for solution of Eq. (2.20), the following discrete form is obtained:

$$\left[\frac{1}{\Delta t} C + \theta_\phi \{H - P\}\right] \phi^{m+1} - \theta_q G q^{m+1} = \left[\frac{1}{\Delta t} C - \left\{\left(1 - \theta_\phi\right)(H - P)\right\}\right] \phi^m + \left(1 - \theta_q\right) G q^m, \tag{2.21}$$

where ϕ^{m+1} and q^{m+1} represent the potential and flux at the $(m+1)$ th time step, Δt is the time step, ϕ^m and q^m are the potential and flux at the mth time step. Several tests were done here to choose the best values for θ and we decided to select the backward-difference scheme $\theta_\phi = 1$ and $\theta_q = 1$. In the time marching computation, the unknown quantities ϕ^m are updated at each time step by the new values obtained after solving Eq. (2.21). At the first time step, the concentration ϕ and heat flux q at all boundary and internal points are specified with initial values. The computation ends when all time steps are fulfilled [Wr02] or a steady state is reached. The right side of Eq. (2.21) is known at all times. Upon introducing the boundary conditions at time $(m+1)\,\Delta t$, the left side of the equation can be rearranged and the resulting system solved by using standard direct procedures such as Least Squares, Gauss elimination and LU decomposition. More details of the element properties, interpolation functions, time integration and equation system formulation used in this paper are described in Brebbia et al. [BrEtAl12].

2.8 The Choice of Radial Basis Functions

In recent years, the theory of radial basis functions (RBFs) has undergone intensive research and enjoyed considerable success as a technique for interpolating multivariable data and functions. A radial basis function, $\Psi\left(x - x_j\right) = \psi\left(\left\|x - x_j\right\|\right)$, depends upon the separation distances of a subset of data centres, $\left(x_j\right)_{j=\overline{1,N}}$. The distance, $\left\|x - x_j\right\|$, is usually taken to be the Euclidean metric, although other metrics are possible (for more details see Golberg and Chen [GoCh94]). The type of RBF used in the interpolation of the unknown variables normally plays an important role in determining the accuracy of the DRM [OoPo13]. Partridge et al.

[PaEtAl92] have shown that a variety of functions can in principle be used as global interpolation functions f_k. The approach used by Wrobel and DeFigueiredo [De90] was based on practical experience rather than formal mathematical analyses and motivated by a previous successful experience with axisymmetric diffusion problems in which a similar approach was used [Te87]. In the present work, decision has been made to follow [De90] by starting with a simple form of the particular solution ψ and finding the related expression for function f by substitution directly into Eq. (2.7). The resulting expressions are

$$\psi = r^3,$$
$$\eta = 3\ r\ \left[(x - x_k)\ n_x\ + (y - y_k)\ n_y\ \right],$$
$$f = 9\,D\,r\ -\ 3\,r\ \left[(x - x_k)\ v_x\ + (y - y_k)\ v_y\ \right] - k\,r^3,$$

in which (x_k, y_k) and (x, y) are the coordinates of the kth boundary or internal point and a general point, respectively. It is important to notice that the set of functions f produced depend not only on the distance r but also on the diffusivity D, velocity components v_x and v_y as well as the reaction rate k, therefore, it will behave differently when diffusion or convection is the dominating process. The most popular RBFs are labelled as: $r^{2m-2} \log r$ (generalised thin-plate spline), $\left(r^2 + c^2\right)^{m/2}$ (generalised multiquadric) and $e^{-\beta r}$ (Gaussian) where m is an integer number and $r = \left\| x - x_j \right\|$. Duchon [Du77] derived the thin-plate splines (TPS) as an optimum solution to the interpolation problem in a certain Hilbert space via the construction of a reproducing kernel. It is interesting to observe that Duchon's thin-plate splines function with $m = 2$ corresponds to the fundamental solution commonly used in the BEM technique to solve biharmonic problems.

Another popular RBF for the DRM is the multiquadric (MQ). However, despite MQ's excellent performance, it contains a free parameter, c, often referred to as the shape parameter that describes the relative 'flatness' of the RBFs about their centres. When c is small the resulting interpolating surface is pulled tightly to the data points, forming a cone-like basis functions. As c increases, the peak of the cone gradually flattens. The Hardy multiquadric functions with values of $m = 1$ and $c = 0$ are often referred to as conical functions and, with $m = 3$ and $c = 0$, as Duchon cubic. Even though TPS have been considered optimal in interpolating multivariate functions, they do only converge linearly, Powell [Po94]. On the other hand, the multiquadric (MQ) functions converge exponentially as shown by Madych and Nelson [MaNe90]. The tuning of the free parameter c can dramatically affect the quality of the solution obtained. Increasing the value of c will lead to a flatter RBF. This will, in general, improve the rate of convergence at the expense of increased numerical ill-conditioning of the resulting linear system [MaNe90]. Much effort has been made to search for the ideal shape parameter c when utilising the MQ-RBF. This is due to the lack of information on choosing the best shape parameter available in the literature, forcing the user having to make an 'ad-hoc' decision. It should also be noted that, following the procedures discussed in this section, the

Table 2.1 Radial basis functions

Name	Function
Multiquadric MQ	$\left(r^2 + c^2\right)^{1/2}$
Thin-Plate Spline TPS	$r^2 \log r$
Cubic RBF	r^3

MQ-RBF is used to approximate the function ψ, not the function f. After a process of investigation, the authors found the optimal value of the non-dimensional shape parameter for the current problems to be $c = 75$, and this value is used for all simulations.

The RBFs presented in Table 2.1 have been examined in this paper. Thin-plate splines and the multiquadric are conditionally positive definite functions (for more details see [OrEtAl11]).

2.9 Numerical Results and Discussions

The present section is concerned with the numerical application of the DRBEM for the solution of two-dimensional transient convection-diffusion-reaction problems with variable velocity. We shall examine some test examples to assess the robustness and accuracy of this new proposed formulations. For the validation and the performance of the proposed procedure, two benchmark problems with known analytical solution are considered.

2.9.1 Transient Convection-Diffusion-Reaction over a Square Channel with Time-Dependent Dirichlet Boundary Conditions and Tangential Velocity Field

In the first example, the domain is considered to be a unit square. We focus on solutions predicted all over the domain by using 19 internal nodes and fixed values of $D = 1\ \mathrm{m}^2/\mathrm{s}$, $k = 0$, and variable velocity $v_x\,(x)$ and $v_y\,(y)$ (m/s) as follows:

$$v_x\,(x) = \tan(x)\,,$$
$$v_y\,(y) = \tan(y)\,.$$

The test results are obtained for Eq. (2.1) with the following initial and boundary conditions. The initial condition is chosen as the analytical value of Eq. (2.22) for $t = 0$:

$$\phi\,(x, y, t = 0) = \sin(x) + \sin(y)\,.$$

Boundary conditions are chosen as:

$$\phi\,(x=0,\,y,\,t)=(\sin\,(y))\,e^{-2t},\ \phi\,(x=1,\,y,\,t)=(\sin\,(1)+\sin\,(y))\,e^{-2t}$$

$$\phi\,(x,\,y=0,\,t)=(\sin\,(x))\,e^{-2t},\ \phi\,(x,\,y=1,\,t)=(\sin\,(x)+\sin\,(1))\,e^{-2t}.$$

The analytic solution for the present case can be obtained from the following expression:

$$\phi\,(x,\,y,\,t)=(\sin\,(x)+\sin\,(y))\,e^{-2t}. \tag{2.22}$$

Figure 2.2 shows the geometrical mesh of the BEM model over a square channel. The boundary is discretised into 50 equally spaced constant elements per side. The analytical and the numerical solutions of this problem are shown in Fig. 2.3 at several time levels utilising the MQ-RBF and implementing a fully implicit scheme when $\theta = 1$ and time step $\Delta t = 0.05$. The result is obtained for the time evolution of the concentration profile along the centre line of the domain. Comparison between the above analytical solution and the numerical results shows an excellent agreement.

Figures 2.4 and 2.5 consider the results using the thin-plate spline TPS-RBF and Cubic RBF also with time step $\Delta t = 0.05$ s. Similar results as for the MQ-RBF have been obtained in both cases. Figure 2.6 shows the time evolution of the concentration distribution in comparison with the analytical solution at the centre points of the computational domain, i.e. $x = y = 0.5$ using the backward-difference

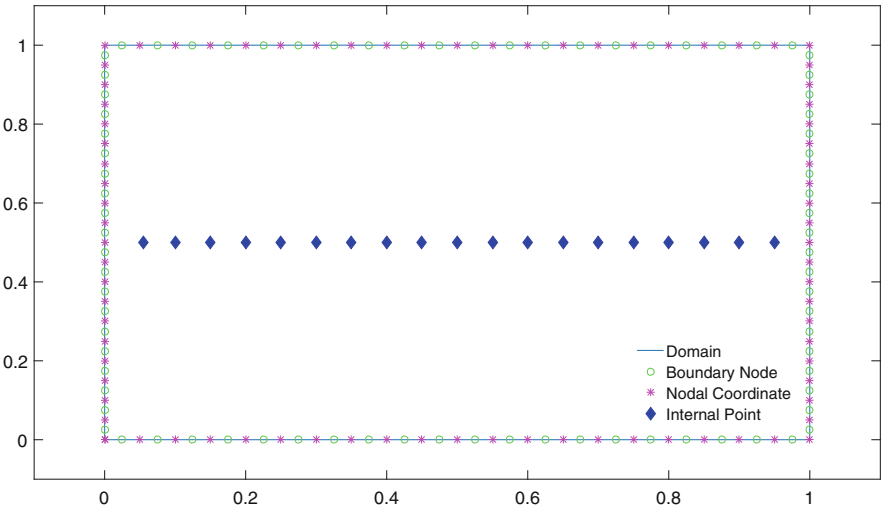


Fig. 2.2 Geometrical mesh of convection- diffusion problem with side length 1 m

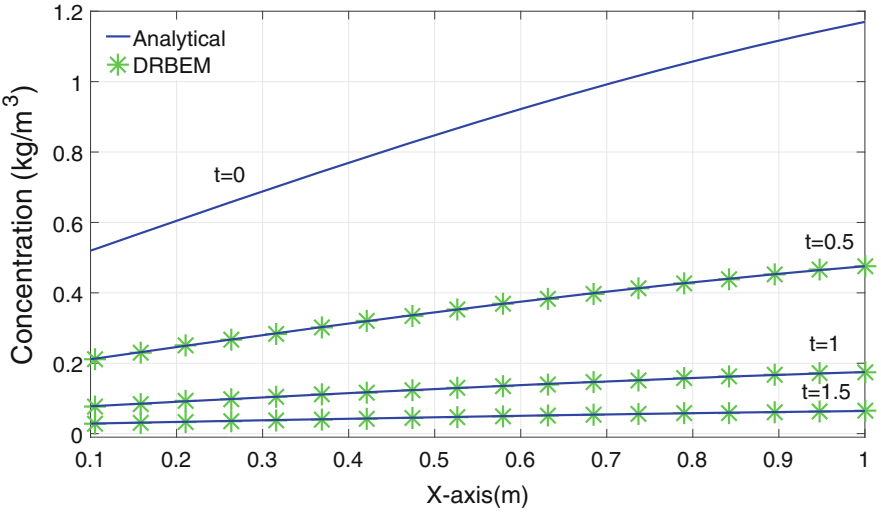


Fig. 2.3 Concentration profile for every 10 time steps using MQ-RBF: comparison between the analytical (solid line) and numerical (star points) solutions

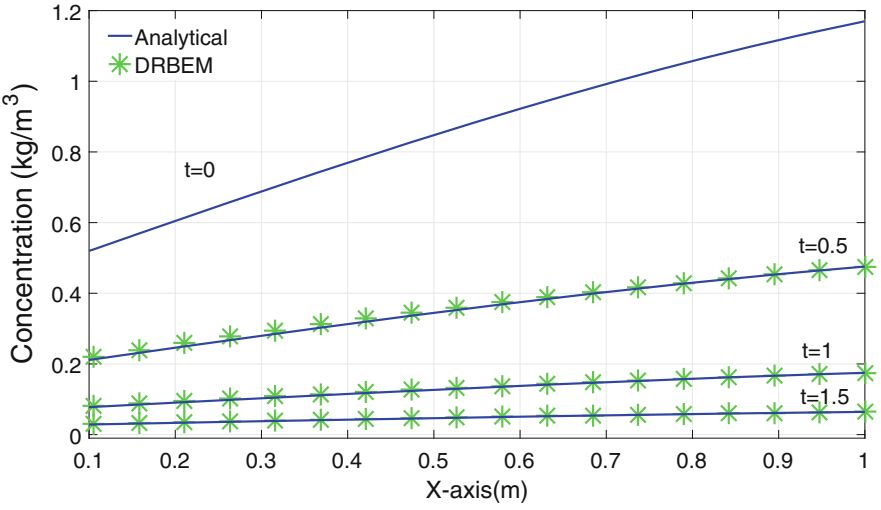


Fig. 2.4 Concentration profile for every 10 time steps using TPS-RBF: comparison between the analytical (solid line) and numerical (star points) solutions

procedure and TPS-RBF. Table 2.2 shows a comparison between the three different RBFs with time step value $\Delta t = 0.05$ s at time level $t = 0.5$. It can be seen that the results obtained by the multiquadric, thin-plate spline and cubic RBFs are reasonably similar. In order to estimate the simulation error, the root mean square norm is utilised as shown in Table 2.3. It is based on the difference between the

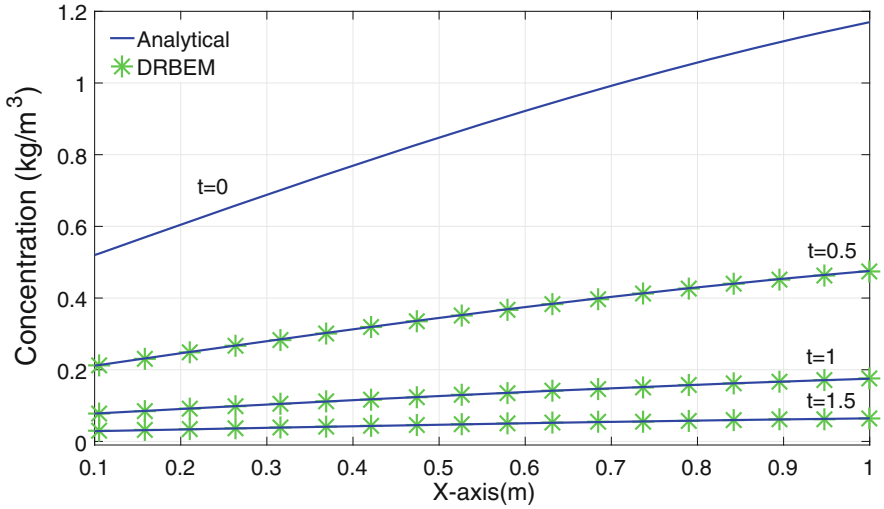


Fig. 2.5 Concentration profile for every 10 time steps using Cubic RBF: comparison between the analytical (solid line) and numerical (star points) solutions

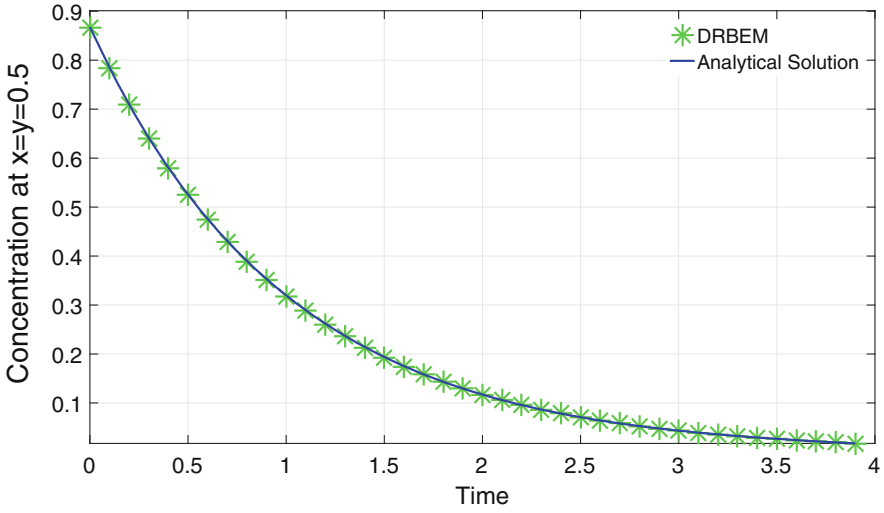


Fig. 2.6 Concentration distribution with time using TPS-RBF: comparison of analytical (solid line) and numerical solution (star points) for at $x = y = 0.5$

simulation results ϕ_{numer} and the analytical solution ϕ_{exact} as

$$RMS = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\frac{\left(\phi_{i,numer} - \phi_{i,exact}\right)^2}{\phi_{i,exact}^2}},$$

Table 2.2 Results for convection-diffusion-reaction at $t = 0.5$ for $\Delta t = 0.05$

x	Cubic	MQ	TPS	Analytical
0.055	0.1962	0.1939	0.2000	0.1965
0.15	0.2304	0.2285	0.2395	0.2313
0.25	0.2662	0.2680	0.2773	0.2673
0.35	0.3013	0.3020	0.2951	0.3025
0.50	0.3515	0.3490	0.3601	0.3527
0.60	0.3829	0.3847	0.3893	0.3840
0.75	0.4259	0.4258	0.4287	0.4271
0.85	0.4517	0.4509	0.4535	0.4527
0.95	0.4751	0.4759	0.4756	0.4756

Table 2.3 RMS error of DRBEM at $t = 2$ for decreasing Δt

$\theta = 1$, $f = r^2 \log(r)$, Problem 1			
	$\Delta t = 0.1$	$\Delta t = 0.05$	$\Delta t = 0.025$
RMS error in ϕ	0.0091	0.0067	0.0058

Table 2.4 Results for convection-diffusion-reaction problem using MQ-RBF with different values of the shape parameter c

x	c=100	c=75	c=50	c=25	c=5	Analytical
0.055	0.1970	0.1939	0.1861	0.1852	0.3752	0.1965
0.25	0.2678	0.2680	0.2468	0.2284	1.4938	0.2673
0.50	0.3524	0.3490	0.3185	0.3025	−3.5933	0.3527
0.75	0.4268	0.4258	0.3984	0.3909	−7.3263	0.4271
0.95	0.4759	0.4759	0.4708	0.4662	−1.6469	0.4756

where i denotes a nodal value, $\phi_{i,exact}$ is the analytical solution, $\phi_{i,numer}$ is the numerical solution and N is the total number of internal nodes. In Table 2.3 the error is seen to reduce as Δt decreases, as expected. Table 2.4 shows a comparison between five different values of the shape parameter c for MQ-RBF with time step value $\Delta t = 0.05$ s at time level $t = 0.5$. It is clear that the results obtained are reasonable and laying at same level of accuracy when the parameter $c = 75$ or 100. On the other hand, the results appear to lose their accuracy for smaller values of c.

From another point of view, taking a very high value of the shape parameter c creates collocation matrices which are poorly conditioned and require high-precision arithmetic to solve accurately. Using a relatively high non-dimensional shape parameter of 75, the collocation matrices are sufficiently well conditioned to be solved using quad-precision arithmetic (see [Ch12, StEtAl13, StPo15] for more details on the shape parameter c).

2.9.2 *Transient Convection-Diffusion-Reaction Problem over Rectangular Region with Mixed (Neumann–Dirichlet) Boundary Conditions*

As a final example, we investigate a convection-diffusion-reaction problem with linear reaction term. The velocity field is considered to be along the longitudinal direction and all the coefficients in the governing equation are constant. The numerical and analytical solutions are compared for different time steps Δt and reaction coefficient k. The geometry is considered to be [0.7 m × 1 m] as shown in Fig. 2.7. Potential values are imposed at the ends of the cross-section, i.e., at $x = 0$, $\phi = 300$ and at $x = 1$, $\phi = 10$. On the sides parallel to x, the lateral fluxes $q = 0$, the problem thus having mixed Neumann–Dirichlet boundary conditions:

$$\frac{\partial \phi}{\partial n}(x, 0, t) = \frac{\partial \phi}{\partial n}(x, 0.7, t) = 0, \ \ 0 \leq x \leq 1, \ \ t > 0,$$

$$\phi(0, y, t) = 300, \ \phi(1, y, t) = 10, \quad 0 \leq y \leq 0.7, \ \ t > 0$$

and the initial conditions are $\phi(x, y, 0) = 0$ at all points at $t = 0$. The values of the reaction parameter k are assumed to be $k = 1, 5, 10, 20$ and $40\,\mathrm{s}^{-1}$, $v_y = 0$ while v_x is considered to vary according to the formula:

$$v_x = kx + \log\left(\frac{10}{300}\right) x - \frac{k}{2}. \tag{2.23}$$

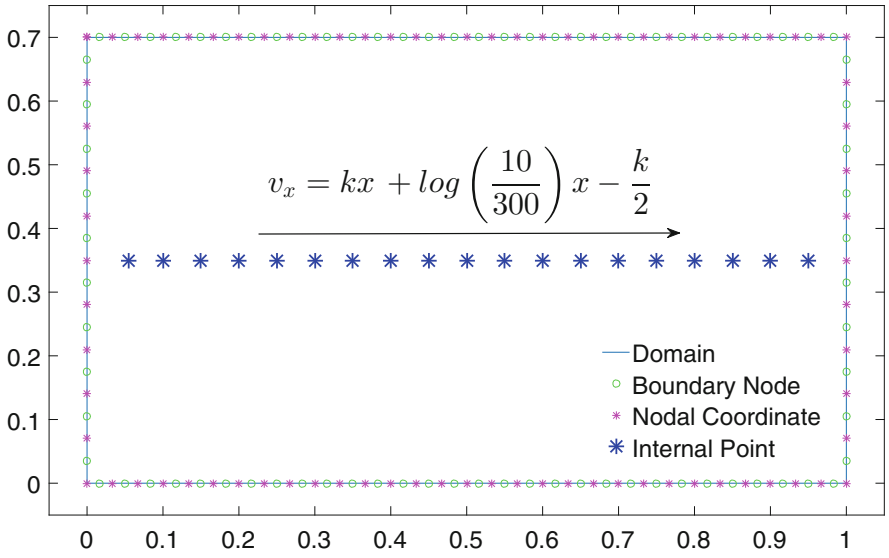


Fig. 2.7 Schematic representation of the rectangular channel model with side length 1 m

The steady-state solution is given in [PaSe00]

$$\phi(x, y) = 300 \exp\left[\frac{k}{2}x^2 + \log\left(\frac{10}{300}\right)x - \frac{k}{2}x\right].$$

In the numerical simulation with a fully implicit scheme, a diffusion coefficient $D = 1\,\mathrm{m}^2/\mathrm{s}$, a variable velocity v_x as described in Eq. (2.23), and a time step $\Delta t = 0.05\,\mathrm{s}$ were used with the results shown at $t = 2\,\mathrm{s}$, by which time the solution has converged to a steady state. Comparison between the above analytical solution and our numerical results are given in figures below, showing excellent agreement.

Case (i): $k = 1$ The first case is considered with the reaction value $k = 1$, which is analysed with the computational domain discretised into 80 constant elements and using 19 internal points. For the DRBEM model, only the TPS-RBF has been applied in all cases. Figure 2.8 shows the exact and numerical solutions, with 10 constant elements along the vertical sides and 30 along each horizontal side.

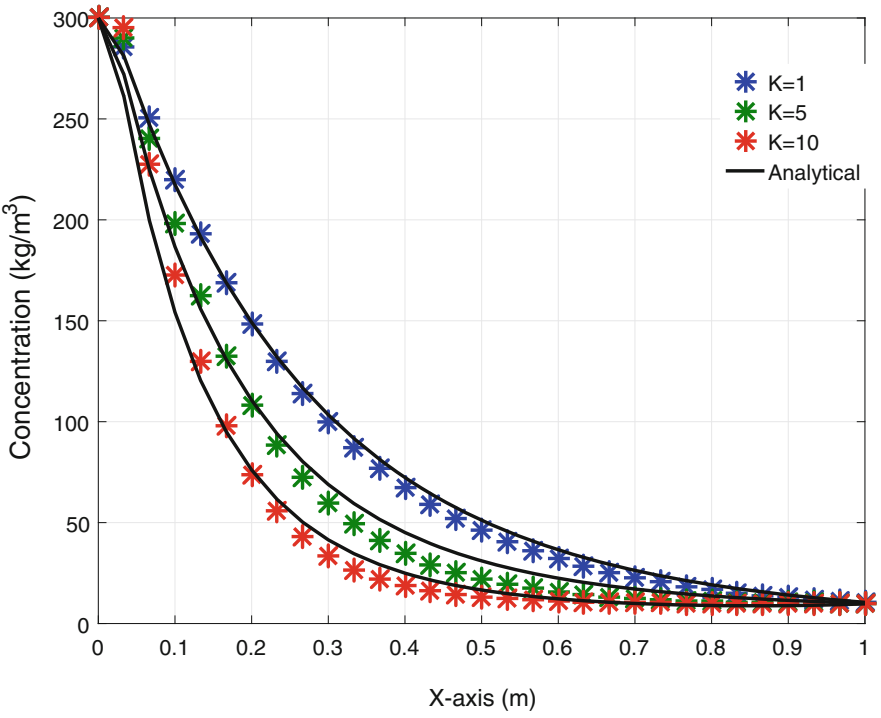


Fig. 2.8 Concentration profile ϕ distribution for bounded domain with different values of reaction k: Comparison between the analytical (solid line) and numerical (star points) solutions, for every 5 time steps, Problem 2

Case (ii): $k = 5$ In the second case, the contribution of the reactive term in Eq. (2.1) is increased to $k = 5$. In Fig. 2.8, results are compared for the same time-stepping scheme considered in the previous case. The results are still very reasonable for the discretisation employed, which is the same as for $k = 1$.

Case (iii): $k = 10$ For this case, the contribution of the reactive term in Eq. (2.1) is increased to $k = 10$. Figure 2.8 displays the results time for the same time-stepping scheme considered in the previous case.

Case (iv): $k = 20$ To see the effect of further increasing the value of k, the reaction coefficient is now $k = 20$. In this case, the maximum global Péclet number is equal to 10 (see Fig. 2.9).

Case (v): $k = 40$ The final test considers the reaction coefficient $k = 40$. A plot of the variation of the concentration ϕ along the x-axis is presented in Fig. 2.9. In this case, the maximum global Péclet number is 20. It is obvious that the agreement with the corresponding analytical solution is still very good.

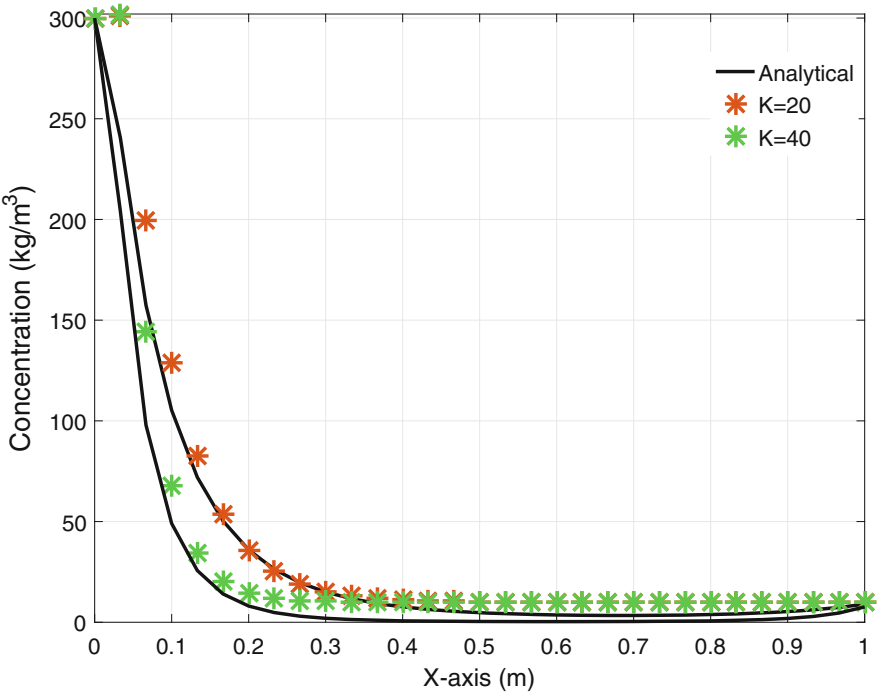


Fig. 2.9 Comparison between the analytical (solid line) and numerical (star points) solutions, for different values of the reaction k

2.10 Concluding Remarks

In this paper, we present a novel formulation of the DRBEM for solving two-dimensional transient convection-diffusion-reaction problems with spatial variable velocity field. This new formulation for this type of problems has been implemented to handle the time derivative part and the variable velocity field. The fundamental solution of the corresponding steady-state equation with constant coefficients has been utilised. The DRBEM is used to transform the domain integrals appearing in the BEM formulations into equivalent boundary integrals, thus retaining the boundary-only character of the standard BEM. Numerical applications for 2D time-dependent problems are demonstrated to show the validity of the proposed technique, and its accuracy was evaluated by applying it to two tests with different velocity fields. Moreover, numerical results show that the DRBEM does not present oscillations or damping of the wave front as may appear in other numerical techniques.

The results presented in Sect. 2.9 show the versatility of the method to solve time-dependent convection-diffusion-reaction problems involving variable velocity fields. We can note a distinct advantage of the present approach, which demonstrates very good accuracy even for high reaction values which increase the Péclet number for the cases studied. It is obvious that, as the velocity increases, the concentration distribution becomes steeper and more difficult to reproduce with numerical models. However, all BEM solutions are still in good agreement for moderate Péclet number (Pé $= 10$ and Pé $= 20$), but oscillations appear for high Péclet number; thus, more refined discretisations are required for these cases. We have made an extensive investigation for the last case studied by considering many different values of the reaction coefficient k. For all these various values of k the backward time-stepping scheme produces very good results in general. We have derived and implemented three RBFs and tested them with different types of problems.

Acknowledgments The first author gratefully acknowledges the Ministry of Higher Education and Scientific Research of Iraq (Al-Nahrain University) for the financial support and PhD scholarship.

References

[AlWr17] Al-Bayati, S.A., Wrobel, L.C.: DRBEM Formulation for Convection-Diffusion-Reaction Problems with Variable Velocity. In: Chappell, D.J. (ed.) Eleventh UK Conference on Boundary Integral Methods (UKBIM 11), pp. 5–14. Nottingham Trent University Press, Nottingham (2017)

[AlWr18a] Al-Bayati, S.A., Wrobel, L.C.: A novel dual reciprocity boundary element formulation for two-dimensional transient convection–diffusion–reaction problems with variable velocity. Eng. Anal. Bound. Elem. **94**, 60–68 (2018)

[AlWr18b] Al-Bayati, S.A., Wrobel, L.C.: The dual reciprocity boundary element formulation for convection-diffusion-reaction problems with variable velocity field using different radial basis functions. Int. J. Mech. Sci. **145**, 367–377 (2018)

[AlWr19] Al-Bayati, S.A., Wrobel, L.C.: Radial integration boundary element method for two-dimensional non-homogeneous convection–diffusion–reaction problems with variable source term. Eng. Anal. Bound. Elem. **101**, 89–101 (2019)

[Al02] Aliabadi, M.H.: The Boundary Element Method. Applications in Solids and Structures, vol 2. Wiley, Chichester (2002)

[ArTa89] Aral, M.M., Tang, Y.: A boundary-only procedure for transient transport problems with or without first-order chemical reaction. Appl. Math. Model. **13**(3), 130–137 (1989)

[De90] DeFigueiredo, D.B.: Boundary Element Analysis of Convection-Diffusion Problems, PhD Thesis, Wessex Institute of Technology, Southampton (1990)

[BrEtAl12] Brebbia, C.A., Telles, J.C.F., Wrobel, L.C.: Boundary Element Techniques: Theory and Applications in Engineering. Springer, Berlin (2012)

[CaEtAl10] Cao, L. Qin, Q., Zhao, N.: Application of DRM-Trefftz and DRM-MFS to transient heat conduction analysis. Recent Pat. Space Technol. **2**, 41–50 (2010)

[Ch12] Cheng, A.H.D.: Multiquadric and its shape parameter—A numerical investigation of error estimate, condition number, and round-off error by arbitrary precision computation. Eng. Anal. Bound. Elem. **36**(2), 220–239 (2012)

[DiKa04] Divo, E., Kassab, A.J.: Transient non-linear heat conduction solution by a dual reciprocity boundary element method with an effective posteriori error estimator. In:ASME 2004 International Mechanical Engineering Congress and Exposition, pp. 77–86 (2004)

[Du77] Duchon, J.: Splines minimizing rotation-invariant semi-norms in Sobolev spaces. In: Constructive Theory of Functions of Several Variables, pp. 85–100. Springer, Berlin (1977)

[GoCh94] Golberg, M.A., Chen, C.S.: The theory of radial basis functions applied to the BEM for inhomogeneous partial differential equations. Bound. Elem. Commun. **5**(2), 57–61 (1994)

[MaNe90] Madych, W.R., Nelson, S.A.: Multivariate interpolation and conditionally positive definite functions. II. Math. Comput. **54**(189), 211–230 (1990)

[Ma05] Martin, V.: An optimized Schwarz waveform relaxation method for the unsteady convection-diffusion equation in two dimensions. Appl. Numer. Math. **52**(4), 401–428 (2005)

[Mo96] Morton, K.W.: Numerical Solution of Convection-Diffusion Problems. Chapman & Hall, London (1996)

[OoPo13] Ooi, E.H., Popov, V.: Meshless solution of axisymmetric convection–diffusion equation: a comparison between two alternative RBIE implementations. Eng. Anal. Bound. Elem. **37**(4), 719–727 (2013)

[OrEtAl11] Orsini, P., Power, H., Lees, M.: The Hermite Radial basis function control volume method for multi-zones problems; A non-overlapping domain decomposition algorithm. Comput. Methods Appl. Mech. Eng. **200**(5), 477–493 (2011)

[PaSe00] Partridge, P.W., Sensale, B.: The method of fundamental solutions with dual reciprocity for diffusion and diffusion–convection using subdomains. Eng. Anal. Bound. Elem. **24** (9), 633–641 (2000)

[PaEtAl92] Partridge, P.W., Brebbia, C.A., Wrobel, L.C.: The dual reciprocity boundary element method. Southampton: Comp. Mech. Pub., Southampton (1992)

[Po94] Powell, M.J.D.: The uniform convergence of thin plate spline interpolation in two dimensions. Numer. Math. **68**(1), 107–128 (1994)

[RaŠk13] Ravnik, J., Škerget, L.: A gradient free integral equation for diffusion–convection equation with variable coefficient and velocity. Eng. Anal. Bound. Elem. **37**(4), 683–690 (2013)

[Sm85] Smith, G.D: Numerical Solution of Partial Differential Equations: Finite Difference Methods. Oxford University Press, Oxford (1985)

[StPo15] Stevens, D., Power, H.: The radial basis function finite collocation approach for capturing sharp fronts in time dependent advection problems. J. Comput. Phys. **298**, 423–445 (2015)

[StEtAl13] Stevens, D., Power, H., Meng, C.Y., Howard, D., Cliffe, K.A.: An alternative local collocation strategy for high-convergence meshless PDE solutions, using radial basis functions. J. Comput. Phys. **254**, 52–75 (2013)

[Te87] Telles, J.C.F.: A self-adaptive co-ordinate transformation for efficient numerical evaluation of general boundary element integrals. Int. J. Numer. Methods Eng. **24**(5), 959–973 (1987)

[Wr02] Wrobel, L.C.: The Boundary Element Method. Applications in Thermo-Fluids and Acoustics. Wiley, Chichester (2002)

[WrDe91] Wrobel, L.C., DeFigueiredo, D.B.: Numerical analysis of convection-diffusion problems using the boundary element method. Int. J. Numer. Methods Heat Fluid Flow **1**(1), 3–18 (1991)

[WoEtAl86] Wrobel, L.C., Brebbia, C.A., Nardini, D.: The dual reciprocity boundary element formulation for transient heat conduction. In: Finite Elements in Water Resources VI, pp. 801–811. Springer, Berlin (1986)

Chapter 3
On a Parametric Representation of the Angular Neutron Flux in the Energy Range from 1 eV to 10 MeV

Luiz F. F. Chaves Barcellos, Bardo E. J. Bodmann, and Marco T. Vilhena

3.1 Introduction

Neutron transport is relevant in a variety of applications, such as nuclear cancer therapy (for instance, Boron Neutron Capture Therapy), design and characterisation of new materials by the use of neutron scattering, energy production by nuclear reactions and many others. From a theoretical point of view, the neutron transport equation, i.e. the Boltzmann equation, and its solution is still a challenge [FeEtAl17, Ga08, La06]. Because of difficulties in determining the solution in full phase space with its seven variables (space, time, direction and energy) approaches in the literature resort to simpler models based on the diffusion equation, where the continuous energy dependence is approximated by multi-group models and the directional information integrated out so that the solution is the scalar neutron flux [OlEtAl17, OlEtAl19]. It is worth to mention also that works on the original transport equation for a stationary case make use of a discretisation of the spatial, the directional and the energy variable, known as the Lattice Boltzmann approach [BiPa12, ErHe13, WaEtAl17, WaEtAl18, WaEtAl19, YaEtAl17].

Methods that maintain the spatial variable as a continuous quantity but discretising the angular variable are based on the so-called S_N approximation [LaEtAl18, SeEtAl12, VaEtAl19]. There also exist approaches with continuous angular variables, the so-called P_N approximation [GhEtAl19], where the angular dependency is approximated by Legendre polynomials. None of these treats the transport problem in full three spatial dimensions, nevertheless, there do exist numerical transport codes that give up properties such as symmetries (geometrical as well

L. F. F. Chaves Barcellos · B. E. J. Bodmann (✉) · M. T. Vilhena
Federal University of Rio Grande do Sul, Porto Alegre, RS, Brazil
e-mail: bardo.bodmann@ufrgs.br

C. Constanda (ed.), *Computational and Analytic Methods in Science and Engineering*, https://doi.org/10.1007/978-3-030-48186-5_3

as dynamical) of the original problem due to the approximation of the differential operators in the transport equation.

The present work is an attempt to solve the transport problem using a physical Monte Carlo method. Physical means that all the interaction terms present in the Boltzmann equation such as scattering, capture and fission are considered in the simulation [BaEtAl17].

$$\begin{aligned}\frac{1}{v(E)}\frac{\partial}{\partial t}\Phi(\mathbf{r},\boldsymbol{\Omega},E,t)+\boldsymbol{\Omega}\cdot\nabla\Phi(\mathbf{r},\boldsymbol{\Omega},E,t)+\Sigma_t(\mathbf{r},E,t)\Phi(\mathbf{r},\boldsymbol{\Omega},E,t)\\ =\int_0^\infty dE'\int_{4\pi}d\boldsymbol{\Omega}'\left\{\Sigma_s(\mathbf{r},E'\to E,\boldsymbol{\Omega}'\to\boldsymbol{\Omega})\Phi(\mathbf{r},\boldsymbol{\Omega}',E',t)\right.\\ \left.+\frac{\nu(E)}{4\pi}\chi_f(E)\Sigma_f\Phi(\mathbf{r},\boldsymbol{\Omega}',E',t)\right\}+S(,\boldsymbol{\Omega},E,t).\end{aligned}\tag{3.1}$$

As a matter of fact, there do exist other simulators in the literature that solve the transport equation, but they differ in their adopted philosophy, many of them solve the problem by a mathematical Monte Carlo implementation, where the quantity of interest (scalar or angular flux among others) is determined as a result of the Monte Carlo simulation (see, for instance, [WeEtAl18]). Due to the fact that our method mimics neutron physics in the microscopic scale, the generated data set by the Monte Carlo simulation may be evaluated in a posterior analysis where the neutron density important for kinetics, the scalar neutron flux important for diffusion models or the angular neutron flux, solution of the transport equation may be obtained.

It is noteworthy that transport problems with dimension $D \geq 2$ need known angular fluxes at the boundary, so that a unique solution may be determined. These type of approaches usually use *adhoc* hypothesis for these angular fluxes, which can be replaced when determined by Monte Carlo simulations such as the one discussed in this contribution. Moreover, the resulting distributions of the simulation are parameterised so that the quantity of interest is represented as a formula and thus may be used directly in analytical approaches. Note that in the present approach, all the variables are continuous so that symmetries of the physical problem are preserved and no artefacts due to discretisations arise.

3.2 The Simulation

As a scenario we consider the following problem. The world of the simulation is defined by a domain in the form of a cube with edges of 100 cm length, where 72.0% of the volume is occupied by water H_2O and the remaining part is filled with uranium oxide with 0.895% U-235 enrichment and the medium is assumed to be in thermal equilibrium with $T = 568.9\ K$. The focus of this simulation is the spectral neutron distribution as well as the spectral neutron flux. To this end we simplify the data set by considering neutron tallies generated within a sphere with

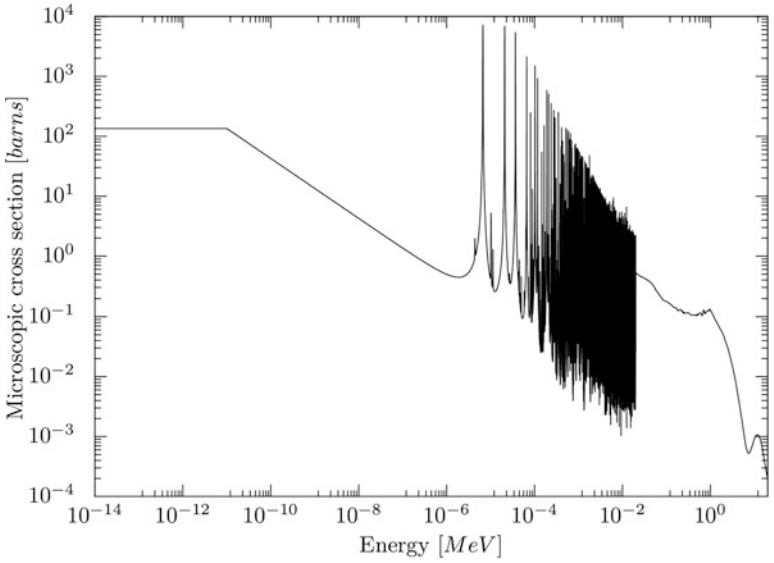


Fig. 3.1 Microscopic absorption cross section of Uranium-238

radius $R = 50$ cm in the centre of the cube and thus suppress influences of the cube boundary, which otherwise might spoil isotropy.

By virtue of the energy range stretching from 10^{-14} MeV to 10^{1} MeV the task of finding an acceptable parametrisation is significantly more difficult than determining the spatial and angular distributions. Due to the restriction of the volume where tallies are sampled from, the spatial distribution in the sphere is approximately homogeneous and the angular distribution compatible with isotropy. In the further focus is put on the energy dependence of the neutron flux in the aforementioned energy range.

As reported in previous works, the microscopic interaction cross sections are provided by a library of sectionally continuous functions [BaEtAl17, CaEtAl13]. An example is given in Fig. 3.1, where the parameterised microscopic absorption cross section for Uranium-238 is shown in the energy range between 10^{-14} MeV and 10^{1} MeV. The initial neutron number at simulation start was 10^{6} and the neutron energies of the initial neutron population follows the Watts distribution [Re08, St07] as shown in Fig. 3.2.

$$\chi(E) = 0.453 e^{-1.036\ MeV^{-1}E} \sinh \sqrt{2.29\ MeV^{-1}E}.$$

After a transient of approximately 10^{-5} s to 10^{-4} s the shape of the neutron distribution stabilises and remains unaltered so that a steady state condition prevails. The time evolution of the spectrum starting from a fission spectrum as initial condition is shown in Fig. 3.3, where for approximately 10^{-4} s the shape of the

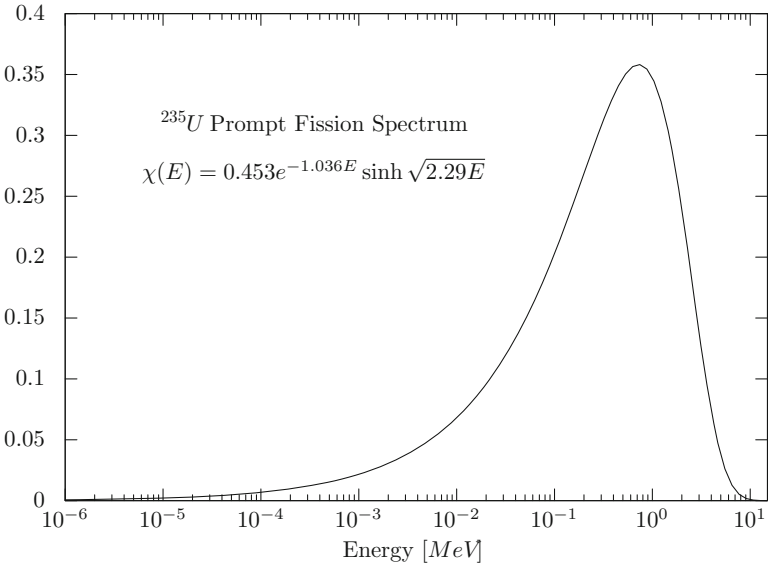


Fig. 3.2 Fission spectrum for Uranium-235

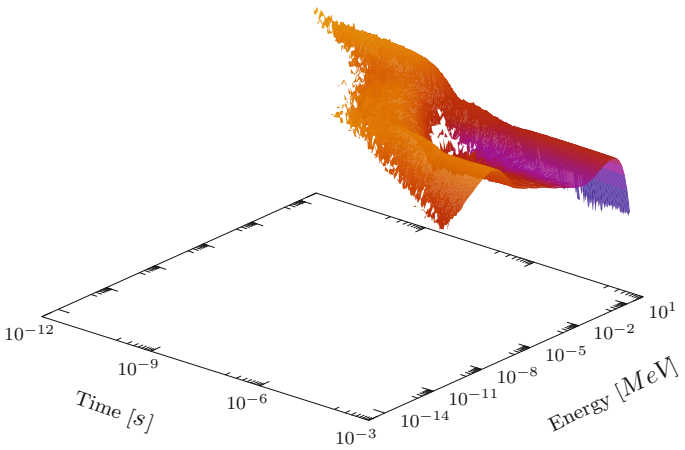


Fig. 3.3 Time evolution of the energy spectrum. The shown surface corresponds to the frequency per second and MeV

spectrum apparently does no longer change. In this figure the initial condition spreads over two orders in magnitude and the time and energy dependent surface represents the time evolution of the energy spectrum, which corresponds to the frequency per second and MeV for neutrons to appear in the intervals $(t, t + \Delta t] \otimes (E, E + \Delta E]$. For large enough times, which allow to assume a steady state regime, the shape of the neutron spectrum is the one shown in Fig. 3.4.

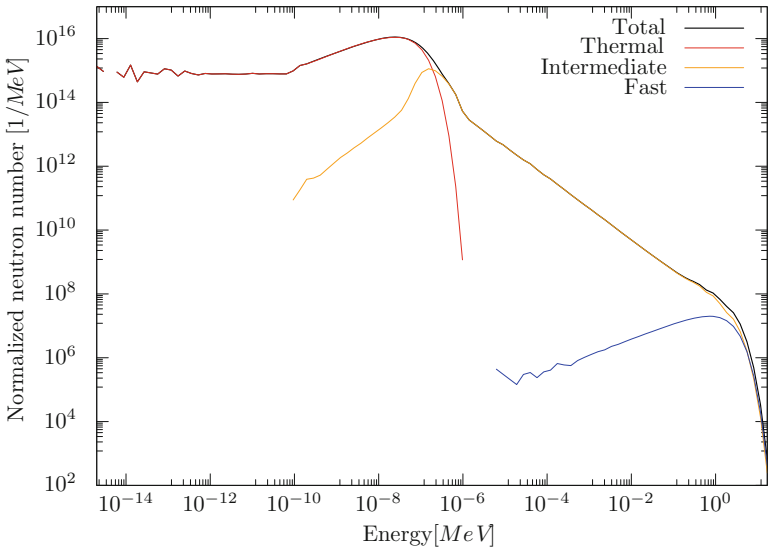


Fig. 3.4 Spectral neutron distribution for a steady state regime

This time scale for the transient phase of the simulation until an approximate steady state is attained is also in agreement with findings in the established literature for light water moderated reactors (see, for instance, [Re08, SpEtAl97, St07]). Two regions of the energy distribution are physically understood, the low energy region up to roughly 1 eV, which is dominated by a thermal equilibrium condition between the neutrons and the surrounding medium and the high energy end beyond 0.1 MeV where the fission distribution contributes to the observed shape. For the intermediate energy region, so far there does not exist an analytical representation, which is the principal issue of the present discussion.

For this reason, one of the tally information recorded in the created simulation data set is a tag which identifies the distribution a specific neutron belongs to, the thermal regime, the intermediate regime where down-scattering dominates and the fission region beyond approximately 0.1 MeV as shown in Fig. 3.4. It is noteworthy, that in the energy interval between 10^{-6} MeV and 10^{0} MeV the neutron spectrum follows with considerable accuracy a power law, which is expected from the dominating down-scattering interactions. Below roughly 1 eV up-scattering is to be taken into account which is also in agreement with the cut between thermal and fast neutrons in a two energy group approach where 1 eV is considered the transition energy [Re08, St07].

The evolution of the neutron population per Monte Carlo step for the total and the three distributions (thermal, intermediate and fission) is shown in Fig. 3.5. From the MC step 500 onward the ratios between two distributions remains constant until the end of the simulation. Also the variation in the neutron population is sufficiently

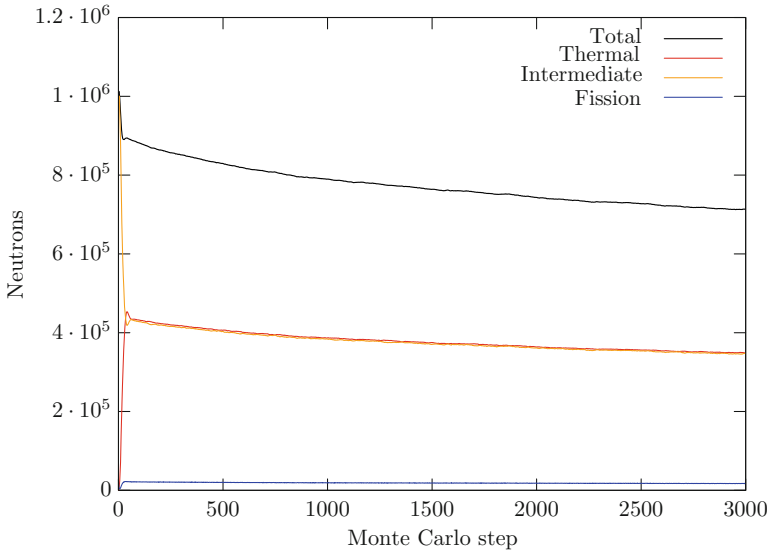


Fig. 3.5 Evolution of the total, the thermal, the intermediate and the fission neutron population per Monte Carlo step

small so that the steady state hypothesis is a valid approximation, which is further supported by the fact that the energy spectrum maintains its shape.

A further criterion to evaluate the regime of the simulation is criticality, or equivalently the effective multiplication factor k_{eff}. Upon using the neutron population balance definition for the multiplication factor, then the neutron life cycle has to be identified. The fact that the simulation method is a physical Monte Carlo implementation, allows for neutrons and their respective generation to be directly tagged. The effective multiplication factor as extracted from the simulation data for the sequence of generations is shown in Fig. 3.6, where up to the 50th generation an approximate steady state regime holds, while changes in the multiplication factor beyond generation 50 are due to the fact, that the simulation terminates and new neutrons are no longer created. The numerical values for the effective multiplication factor k_{eff} fluctuate around $\approx$0.995, which shows that the simulated system is close to critical, so that absorbed and leaked out neutrons are replaced by neutrons produced by fission. These considerations indicate that the configuration of the system is close to a true steady state and neutron data may be extracted for data evaluation between Monte Carlo steps 500 and 3000 (see Fig. 3.5). From one generated data set a variety of physical information may be obtained in a post-simulation analysis and it is also possible to extend the data set with additional simulations, a clear advantage of having chosen the physical Monte Carlo method.

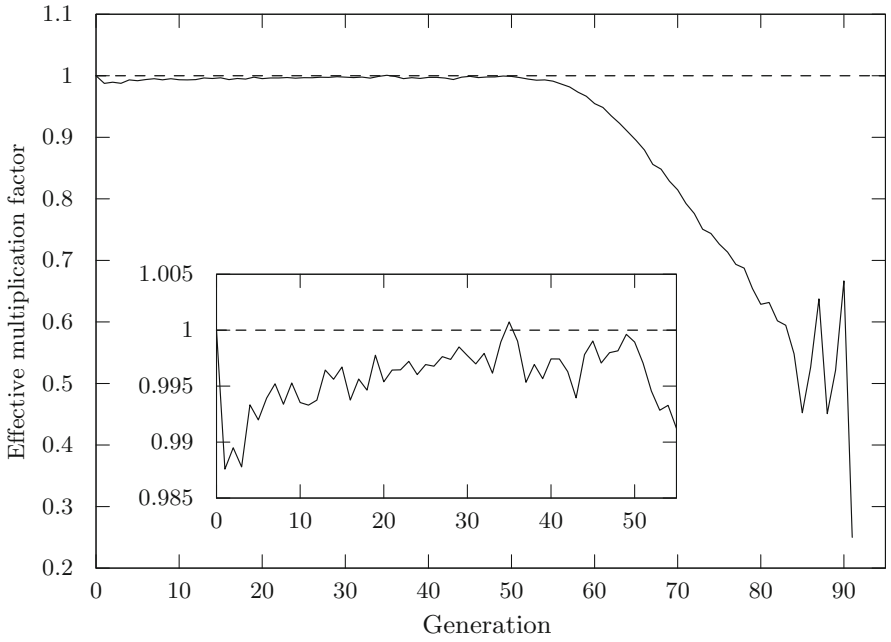


Fig. 3.6 The effective multiplication factor k_{eff} along the generations

3.3 Results

In the following we show some results that refer to analysing specific terms in the Boltzmann equation, where for the further discussion scattering contributions were considered especially interesting. Down-scattering of neutrons with initial energies E' to a specific final energy E is a substantial contribution in the integral term and moreover crucial for the shaping of neutron or flux distributions in the intermediate energy range (see Fig. 3.4). A second contribution from scattering is hidden in the total cross section Σ_t which also contains absorption terms, however, the former is more robust with respect to influences on the shape of the distributions since the initial energy is E and all possible final energies E' are integrated. Formally, the integral term plays the role of a source term increasing the angular neutron flux in the energy interval $[E, E + dE)$, while the scattering contribution to the total cross section has the effect of a drain reducing the angular flux.

Two scenarios were filtered from the simulation data set, scattering of neutrons with Uranium-238 and with Hydrogen-1. As a Uranium target is very much heavier than the neutron projectile, neutrons need an exorbitant number of collisions (more than 2000) in order to reduce their initially high energy from fission in the average $\sim$2 MeV to thermal energies below 1 eV. As a consequence one observes in Fig. 3.7 the sharp line from the initial to final energy correlation in this type of scattering

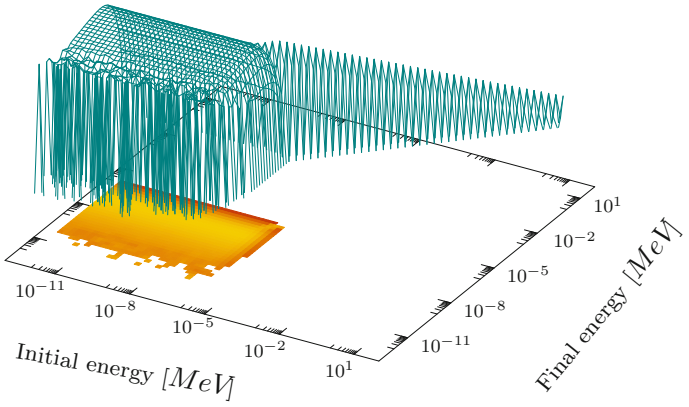


Fig. 3.7 Probability distribution dependence of initial and final energy for down-scattering of neutrons and the thermal regime for a target atom Uranium-238

reaction. At energies of approximately eV and below for both the initial and final energy, the neutrons enter in thermal equilibrium with the surrounding medium. As a consequence the spectral neutron distribution remains the same, since in the average neutrons gain energy in collisions as much as they lose energy. This property of the distribution is visible in the lower left part of Fig. 3.7, where for any initial energy the final energy follows approximately the shape of a Maxwell–Boltzmann distribution. The other extreme of a possible scattering scenario is the collision of neutrons with Hydrogen. Since neutrons and protons have a relative mass difference in the order of magnitude 10^{-3}, a collision has by far more possibilities for energy and momentum transfer than Uranium. In the average already 15 collisions are sufficient to bring neutrons down to thermal energies. Note, that for initial energies above eV up to MeV there is a considerable spread in the distribution for final energies after collision. The probability profile for down-scattering as well as the thermal region is shown in Fig. 3.8. The respective plots for the scattering reactions Figs. 3.7 and 3.8 may also be read in the reverse fashion. Considering one specific final energy one may analyse, what are possible initial energies. For Uranium-238, only in the thermal region a broad distribution relates the spectrum for initial energies to one final one. This property allows to simplify considerably the scattering integral for Uranium-238 since for higher energies than $\sim$1 eV there exists only a narrow peaked ridge, which makes it possible to approximate the integral by a one to one relation between the initial and final energies. This is different for Hydrogen-1, where the distribution for possible initial energies that lead to a specific final energy is broad and independent of the specific energy scale considered, which in our considerations spans from 10^{-13} MeV up to 10 MeV.

As already mentioned in the introduction, many approaches discretize the energy spectrum into a finite set of energy groups. Evidently, in such approaches average values for each energy group are used, but these work out fine when the fluxes vary only smoothly with energy. Though, in the energy regions, where the constituents

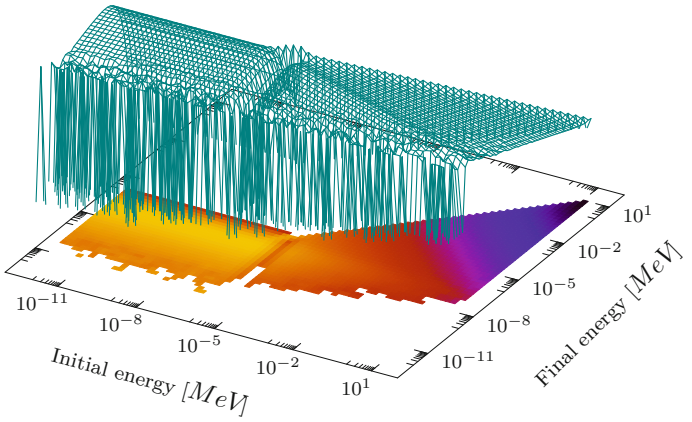


Fig. 3.8 Probability distribution dependence of initial and final energy for down-scattering of neutrons and the thermal regime for a target atom Hydrogen-1

of the medium have resonances, this property is hardly complied with. It is notable, that many transport codes which are used in nuclear engineering tasks work with energy groups so that the knowledge of the spectral angular flux could help to improve the determination of the averaged nuclear parameters in these type of approaches. Recalling, that the present Monte Carlo implementation mimics some part of real micro physics, one may use the tags and filter the energy dependence of the neutron flux, which may then be approximated by a parameterised formula. The present configuration was chosen such that angular and spatial contributions do not interfere in the energy distribution, for this reason only neutrons in the spherical centre of the whole domain were sampled, which left the spatial and solid angle distribution approximately homogeneous (not shown in this work). The comparison of the spectral neutron flux resulting from the Monte Carlo simulation together with its parametrisation is shown in Fig. 3.9. The parametric representation of the spectral neutron flux was obtained first dividing the whole energy interval in four sub-intervals (0 MeV, 10^{-10} MeV], (10^{-10} MeV, 10^{-6} MeV], (10^{-6} MeV, 0.01 MeV] and (0.01 MeV, ∞). In each interval for $\log(E)$ the results of the Monte Carlo simulation were fitted by four polynomial functions $f_i(E)$ ($i \in [1, 4]$ with degrees 1, 4, 1 and 6, respectively. The found coefficients are shown in Table 3.1. Since the simulation data span an energy interval over 14 orders in magnitude, it is convenient to represent the spectral neutron flux Φ in a double logarithmic scale and further as one formula instead of sectionally defined functions. To this end we define the energy window functions w_i for the four intervals and introduce the parameters a and b, which define the smoothness or abruptness of the transition between the intervals.

$$w_1(\log(E), a) = \frac{1}{2}\left(-\tanh(a\log(E/10^{-10})) + 1\right)$$

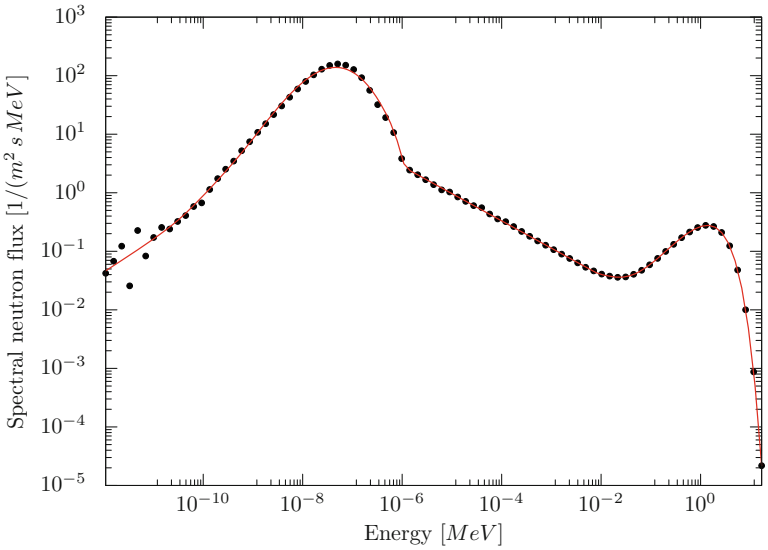


Fig. 3.9 The spectral neutron flux in the energy range from 10^{-13} MeV up to 10 MeV

Table 3.1 Fit coefficients of the spectral neutron flux in log–log scale and in the respective subdomains

	Function				
	f_1	f_2	f_3	f_4	
	Intervals in $[MeV]$				
	$(0\ ,\ 10^{-10}]$	$(10^{-10}\ ,\ 10^{-6}]$	$(10^{-6}\ ,\ 0.01]$	$(0.01\ ,\ \infty)$	
E^0	12.4706	−228.186	−5.48732	−1.32801	in $\log\left[MeV^{-1}m^{-2}s^{-1}\right]$
E^1	0.565077	−38.641	−0.477512	0.27156	in $\log\left[MeV^{-2}m^{-2}s^{-1}\right]$
E^2	–	−2.2306	–	−0.353689	in $\log\left[MeV^{-3}m^{-2}s^{-1}\right]$
E^3	–	−0.0513524	–	−0.130745	in $\log\left[MeV^{-4}m^{-2}s^{-1}\right]$
E^4	–	−0.000376707	–	−0.0363351	in $\log\left[MeV^{-5}m^{-2}s^{-1}\right]$
E^5	–	–	–	−0.00903293	in $\log\left[MeV^{-6}m^{-2}s^{-1}\right]$
E^6	–	–	–	−0.000867472	in $\log\left[MeV^{-7}m^{-2}s^{-1}\right]$

$$w_2(\log(E), a, b) = \frac{1}{2}\left(\tanh(a\log(E/10^{-10})) - \tanh(b\log(E/10^{-6}))\right)$$

$$w_3(\log(E), a, b) = \frac{1}{2}\left(\tanh(a\log(E/10^{-6})) - \tanh(b\log(E/0.01))\right)$$

$$w_4(\log(E), a) = \frac{1}{2}\left(\tanh(a\log(E/0.01)) + 1\right).$$

The spectral flux in a double linear scale is then given by

$$\Phi(E) = 10^{w_1(\log(E),1)\, f_1(log(E)) + w_2(\log(E),1,5)\, f_2(log(E)) + w_3(\log(E),5,100)\, f_3(log(E))} \times \\ \times 10^{w_4(\log(E),100)\, f_4(log(E))}$$

which may be directly used now for analytical approaches. In the further we analyse this parametrisation as a possible solution of the Boltzmann equation.

Due to the fact that all terms of the Boltzmann equation are taken care of in the physical Monte Carlo simulation one shall expect that apart from approximations that were made the parameterised spectral neutron flux interpolated from the result of the simulation data shall represent to a certain extent the solution of the Boltzmann equation. The terms that were neglected in Eq. (3.2) were the explicit time variation, the directional divergence and a possible external source. Then the removal term is represented by the total reaction rate (last term on the left-hand side of Eq. (3.2)), more explicitly the absorption rate and the scattering rate from an initial energy E to any other energy E' and the emission term is given by the fission rate times the neutron multiplicity and spectral weight together with the scattering rate from any initial energy E' to a specific final energy E (first and second term on the right-hand side of Eq. (3.2)).

Recalling that the simulation is not a perfect steady state and also isotropy is only approximate and on the other hand the evaluation of the scattering integral was done numerically using scattering probabilities (Uranium-235, Uranium-238, Oxygen-16 and Hydrogen-1) which for Uranium-238 and Hydrogen-1 are shown in Figs. 3.7 and 3.8, respectively, one cannot expect a perfect match between the left and right-hand side contributions presented in Fig. 3.10. This is especially a consequence of the huge energy interval that defines the limits on the integration (10^{-13} MeV and 10^{1} MeV). As was to be expected, emission and removal are in perfect coincidence at the high energy end where fission contributions dominate. In the intermediate range emission has larger values in comparison to removal. Nevertheless an encouraging detail is the identification of the resonance region of Uranium-238 between roughly 10^{1} eV and 10^{2} eV, which is an essential feature of the neutron life cycle. One of the possible discrepancies between the simulated emission and removal part of the balance equation is the spectral flux and its implication in the scattering integrals, since the latter was solved numerically and thus is subject to inaccuracies.

The emission curve shows considerable fluctuations in the thermal region which may be understood from the fact that the only way to get to these energies is from down-scattering which suffers from limitations by not sufficiently high sampling. On the contrary the removal curve is smooth because of the larger cross sections in this energy range and thus higher reaction rates and consequently better statistics. For lower energies the sampling of emission contributions is poorer manifest in larger fluctuations. According to statistical standards one has to accept that values from the removal curve are statistically compatible with the emission curve within

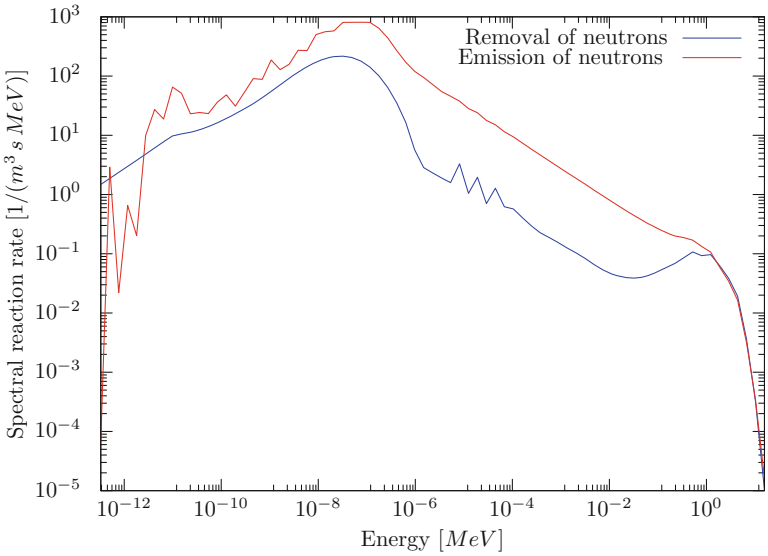


Fig. 3.10 Comparison of the removal and emission terms in the Boltzmann equation

the fluctuations in this energy range. A larger sampling will certainly improve the comparison in this range, which is a future work in progress.

In the intermediate energy range we face the problem of adequately approximating the scattering integral. Recalling, that the integral represents the balance for a specific final energy after scattering but summing up all possible initial energies, uncertainties in the initial energy contribution may influence significantly in the reaction rates and thus the balance of the left- and right-hand side in the Boltzmann equation, i.e. the comparison of the emission and removal contributions.

Another influence for inaccuracies in the obtained results was caused by the fact, that a steady state was only attained approximately. The adopted philosophy for the implementation of the physical Monte Carlo simulator has the disadvantage that it is difficult to adjust the system configuration such as to obtain a specific state, because this property appears as a result of the simulation as a consequence of a specific system configuration. In the presented simulation the evolution of the neutron population per Monte Carlo step showed a slightly sub-critical condition. As a consequence, the variation of the neutron population and the divergence of the fluxes change and may contribute to a discrepancy in the removal and emission terms comparison. Also the spatial distribution is not perfectly homogeneous and isotropic, which yields another contribution that may uneven the balance in the Boltzmann equation. Another aspect that shall be taken into account is that in order to adjust exact criticality one would need an iteration of Monte Carlo simulations, where from variations of the geometric and physical parameters the desired state could be matched for a specific case in consideration.

3.4 Conclusion

With the stage of the development of the present physical Monte Carlo simulator (built from scratch) and the generated results, we hopefully showed by our presented results, how an implementation following the paradigm of power computing may contribute to approach transport challenges such as the stationary Boltzmann equation. Although, the implementation relevant for the present work operates as a stand-alone program package, its C++ structure allows to integrate in a straight forward manner the elaborated and tested code directly in the GEANT simulation project. A return for our simulation intentions is that this insertion will open pathways to make use of the GEANT geometry package and its efficient resources to implement complex geometric domain setups, such as a reactor core with its fuels elements, control rods and structural details.

A few computational highlights emphasise the simulation's efficiency by the variety of results presented. The interaction probabilities such as the scattering probabilities exemplified in Fig. 3.7 and 3.8 result from number-crunching with standard computer architectures but attaining a simulation thread-time with 7.08 h × thread, a subsequent data assessment thread-time with 6.25 h × thread and moreover a reasonable memory administration with a maximum of 471 MB RAM per thread and at a cost of 452.7 GB stored simulation data for an ensemble of initially 10^6 neutrons with its subsequent neutrons generated from fission. The results, here for the spectral neutron flux, were parameterised which may be used in other approaches such as analytical ones or even for the more classical energy group approximations. In the latter, averages need the fluxes before-hand in order to compute reaction rates consistently, which are available from the simulations reported. Nevertheless, the results represent a scientific contribution on its own right as a novel approach to attack solving the Boltzmann transport equation by means of a physical Monte Carlo approach.

Evidently, there is need for further improvements by virtue of the discrepancy of the removal and emission terms when inserting the obtained spectral neutron flux in the Boltzmann equation. The only term in the equation, that was evaluated numerically was the scattering integral, which clearly needs a thorough revision. However, other approaches also fail to satisfy the Boltzmann equation by orders in magnitude comparable or even worse than our result. Very specific improvements necessary are in the energy ranges where the statistics is poorer but where cross sections are typically higher. Consequently, using more smooth distributions for scattering to evaluate the integrals in the Boltzmann equation will likely reduce a significant source for errors. Furthermore, simulations with importance sampling in the low statistics regions may improve the precision and smooth this part of the total distribution. Concluding, further simulations with better statistics and more accurate parametrisations certainly will open pathways for new optimised simulations and may also provide essential physical details for future (semi-)analytical models.

References

[BaEtAl17] Barcellos, L.F., Bodmann, B.E.J., Bogado Leite, S., Vilhena, M.T.: On a continuous energy Monte Carlo simulator for neutron transport: Optimisation with fission, intermediate and thermal distributions. In: Constanda, C., Dalla Riva, M., Lamberti, P.D., Musolino, P. (eds.) Integral Methods in Science and Engineering, vol. 2. Springer, Berlin (2017)

[BiPa12] Bindra, H., Patil, D.V.: Radiative or neutron transport modeling using a lattice Boltzmann equation framework. Phys. Rev. E **86**, 016706 (2012)

[CaEtAl13] De Camargo, D.Q., Bodmann, B.E.J., Vilhena, M.T., Bogado Leite, S.Q., Alvim, A.C.M.: A stochastic model for neutrons simulation considering the spectrum and nuclear properties with continuous dependence of energy. Prog. Nucl. Energy **69**, 59–63 (2013)

[ErHe13] Erasmus, B., Van Heerden, F.A.: The Lattice Boltzmann Method Applied to Neutron Transport. U.S. Department of Energy Office of Scientific and Technical Information, Oak Ridge (2013)

[FeEtAl17] Fernandes, J.C.L., Bodmann, B.E.J., Vilhena, M.T.: On multi-group neutron transport in planar one dimensional geometry: a solution for a localized pulsed source. Ann. Nucl. Energy **101**, 552–558 (2017)

[Ga08] Ganapol, B.D.: Analytical Benchmarks for Nuclear Engineering Applications Case Studies in Neutron Transport Theory. In: Organisation for Economic Co-Operation and Development OECD (2008)

[GhEtAl19] Ghazaie, S.H., Abbasi, M., Zolfaghari, A.: The multi-PN approximation to neutron transport equation. Prog. Nucl. Energy **110**, 64–74 (2019)

[LaEtAl18] Ladeia, C.A., Bodmann, B.E.J., Vilhena, M.T.: The radiative conductive transfer equation in cylinder geometry: Semi-analytical solution and a point analysis of convergence. J. Quant. Spectrosc. Radiat. Transf. **217**, 338–352 (2018)

[La06] Larsen, E.W.: An overview of neutron transport problems and simulation techniques. In: Graziani, F. (ed.) Computational Methods in Transport. Springer, Berlin (2006)

[OlEtAl17] Oliveira, F.R., Bodmann, B.E.J., Vilhena, M.T., Carvalho, F.: On an analytical formulation for the mono-energetic neutron space-kinetic equation in full cylinder symmetry. Ann. Nucl. Energy **99**, 253–257 (2017)

[OlEtAl19] Oliveira, F.R., Fernandes, J.C.L., Bodmann, B.E.J., Vilhena, M.T.: On an analytical solution for the two energy group neutron space-kinetic equation in heterogeneous cylindrical geometry. Ann. Nucl. Energy **133**, 216–220 (2019)

[Re08] Reuss, P.: Neutron Physics. EDP Sciences, Les Ulis (2008)

[SeEtAl12] Segatto, C.F., Vilhena, M.T., and Goncalez, T.T.: On the analytical solution of the neutron SN equation in a rectangle assuming an exponential exiting angular flux at boundary. Int. J. Nucl. Energy Sci. Technol. **7**, 45–56 (2012)

[SpEtAl97] Spriggs, G.D., Adams, K.J., Parsons, D.K.: On the definition of neutron lifetimes in multiplying and non-multiplying systems, LA-13260-MS (1997)

[St07] Stacey, W.M.: Nuclear Reactor Physics. WILEY-VCH Verlag GmbH & Co.KGaA, New York (2007)

[VaEtAl19] Valerio, F.L., Segatto, C.F., Vargas, R.M.F., Vilhena, M.T.: On the analytical representation for the SN radiative-conductive transfer solution in inhomogeneous plane parallel atmosphere with convergence analysis. J. Quant. Spectrosc. Radiat. Transf. **235**, 132–139 (2019)

[WaEtAl17] Wang, Y., Yan, L., Ma, Y.: Lattice Boltzmann solution of the transient Boltzmann transport equation in radiative and neutron transport. Phys. Rev. E **95**, 063313 (2017)

[WaEtAl18] Wang, Y., Xie, M., Ma, Y.: Neutron transport solution of lattice Boltzmann method and streaming-based block-structured adaptive mesh refinement. Ann. Nucl. Energy **118**, 249–259 (2018)

[WaEtAl19] Wang, Y., Ma, Y., Xie, M.: GPU accelerated lattice Boltzmann method in neutron kinetics problems. Ann. Nucl. Energy **129**, 350–365 (2019)

[WeEtAl18] Werner, C.J., Bull, J.S., Solomon, C.J., Brown, F.B., McKinney, G.W., Rising, M.E., Dixon, D.A., Martz, R.L., Hughes, H.G., Cox, L.J., Zukaitis, A.J., Armstrong, J.C., Forster, R.A., Casswell, L.: MCNP User's Manual Code Version 6.2, LA-UR-18-20808. Los Alamos National Laboratory, Los Alamos (2018)

[YaEtAl17] Yan, L., Wang, Y., Ma, Y., Li, W.: Finite volume lattice Boltzmann scheme for neutron/radiative transfer on unstructured mesh. Ann. Nucl. Energy **109**, 227–236 (2017)

Chapter 4
A Boundary Integral Equation Formulation for Advection–Diffusion–Reaction Problems with Point Sources

Luiz F. Bez, Rogério J. Marczak, Bardo E. J. Bodmann, and Marco T. Vilhena

4.1 Introduction

The advection–diffusion–reaction equation may be understood as an extension of the continuity equation and thus represents the simplest model to describe dispersion phenomena such as for particles, heat or more generally energy among others. Due to its simplicity, in risk and safety analysis this model is usually the first one to be employed to simulate and analyse consequences of events or even accidents in the chemical or the nuclear industry, for instance. So far, comparisons of results from the advection–diffusion–reaction model to experimental data, where they were available, proved the usefulness of this idealization of the real world. Nevertheless, there are some shortcomings, due to the difficulty with numerical representations of realistic landscapes, which in general demands power computing resources to produce acceptable solutions for the dispersion process. More specifically, problems of this type if solved via traditional numerical domain methods, usually require fine meshing, especially for problems with high velocity wind fields and large gradients, such as those from point sources, which demands significant computational investments for numerical simulations [MeEtAl17, Ro72]. Another alternative are analytical methods; however, so far they can be applied to problems with simple regular domains but are hardly employed for realistic domains like the ones with complex topography terrains [CuEtAl16]. One alternative is to employ analytical solutions in order to analyse a simplified version of the problem. This, however, is only practical for very specific circumstances and results in a loss in solution quality when applied to realistic situations. Hence, the present contribution is a discussion in

L. F. Bez · R. J. Marczak · B. E. J. Bodmann (✉) · M. T. Vilhena
Federal University of Rio Grande do Sul, Porto Alegre, RS, Brazil
e-mail: rato@mecanica.ufrgs.br; bardo.bodmann@ufrgs.br

C. Constanda (ed.), *Computational and Analytic Methods in Science and Engineering*, https://doi.org/10.1007/978-3-030-48186-5_4

this line, where the boundary element method is presented as an efficient procedure for computational simulations of the aforementioned issues.

Although, we consider a problem in a simple domain geometry, the forthcoming discussion is to shed light on some highlights of the boundary element method and its efficiency in numerically solving advection–diffusion–reaction type of problems. Moreover, from the Green Function Theory, the solution of an arbitrary source can readily be obtained from the knowledge of the solution for a point source. Ikeuchi and Onishi [IkOn83] showed that the boundary element method (BEM) can be used as an alternative for decreasing the problem size and maintaining solution quality. Also, the upwind effect of the mean velocity field is accounted for by the fundamental solution, allowing for the mesh which is being used to be coarser while the solution remains numerically efficient and stable even for very high local Peclet numbers [QiEtAl98]. In the further we present a BEM formulation for a boundary integral equation applied to solve the advection–diffusion–reaction problem for a point source in a rectangular domain. We also report on the numerical performance of the method and its sensibility to changes in the discretization, numerical integration quality and Peclet numbers, respectively. The present formulation has potential to be extended and applied to other scenarios.

4.2 Mathematical Formulation

We consider an advection–diffusion–reaction problem with constant coefficients and known source, as stated in Eq. (4.1).

$$\mathbf{v} \cdot \nabla\phi - d\nabla^2\phi + k\phi = f \tag{4.1}$$

Here ϕ is the concentration of a chemical substance, $\mathbf{v}$ is the advective velocity, d is the diffusivity, k is the reaction coefficient, and f is the source term.

Associated with this differential formulation we have a fundamental solution, which is of Green function type. In our case the fundamental solution ϕ_i^* is the solution to the problem Eq. (4.2), associated with a point source located at x_i, and with asymptotic behaviour $\phi_i^* \to 0$ for increasing distance to the source.

$$\mathbf{v} \cdot \boldsymbol{\nabla}\phi_i^* + d\,\boldsymbol{\nabla}^2\phi_i^* - k\,\phi_i^* = -\delta(\mathbf{x_i}) \tag{4.2}$$

The fundamental solution and its derivative, for a 2D problem, are shown in Eqs. (4.3) and (4.4).

$$\phi_i^*(\mathbf{x}) = \frac{1}{2\pi d}\exp\left(-\frac{\mathbf{v}\cdot\mathbf{r}}{2d}\right)K_0(\mu r) \tag{4.3}$$

$$\frac{\partial\phi_i^*(\mathbf{x})}{\partial\mathbf{n}} = -\frac{1}{2\pi d}\exp\left(-\frac{\mathbf{v}\cdot\mathbf{r}}{2d}\right)\left(\mu K_1(\mu r)\frac{\mathbf{r}\cdot\mathbf{n}}{r} + \frac{\mathbf{v}\cdot n}{2d}K_0(\mu r)\right) \tag{4.4}$$

Here $\mathbf{r} = \mathbf{x} - \mathbf{x_i}$ is the distance vector relative to the source location, K_v is the modified Bessel function of the second kind of order v, and μ is the coefficient defined in Eq. (4.5).

$$\mu = \sqrt{\left(\frac{v}{2d}\right)^2 + \frac{k}{d}} \tag{4.5}$$

We will solve numerically a type of an inverse problem, in a weighted residual sense, as stated in Eq. (4.6). Combining the latter with Eq. (4.2) and using the filtering property of the Dirac delta functional, we arrive at Eq. (4.7), the boundary integral equation (BIE).

$$\begin{aligned} -\int_\Omega \phi \left(\mathbf{v}\cdot\nabla\phi + d\,\nabla^2\phi - k\,\phi\right)\, d\Omega + \int_\Gamma \phi \left(d\,q_i^* + v_n\,\phi_i^*\right)\, d\Gamma \\ -\int_\Gamma q_n\,\phi_i^*\, d\Gamma = \int_\Omega f\,\phi_i^*\, d\Omega \end{aligned} \tag{4.6}$$

$$c_i\,\phi_i + \int_\Gamma \phi \left(d\,q_i^* + v_n\,\phi_i^*\right)\, d\Gamma - \int_\Gamma q_n\,\phi_i^*\, d\Gamma = \int_\Omega f\,\phi_i^*\, d\Omega \tag{4.7}$$

Here ϕ_i is the concentration at location $\mathbf{x_i}$, v_n, q_n, and q_i^* are defined in Eqs. (4.8) and $\mathbf{n}$ is a unit vector pointing outward and is normal to the boundary. The coefficient c_i depends on the location of the source and is given by Eq. (4.9).

$$\begin{aligned} v_n &= \mathbf{v}\cdot\mathbf{n} \\ q_n &= \nabla\phi\cdot\mathbf{n} \\ v_n &= \nabla\phi_i^*\cdot\mathbf{n} \end{aligned} \tag{4.8}$$

$$c_i = \begin{cases} 0 & \text{if } \mathbf{x_i} \notin \Omega \\ 1 & \text{if } \mathbf{x_i} \in \Omega \\ \frac{\theta_i}{2\pi} & \text{if } \mathbf{x_i} \in \Gamma \end{cases} \tag{4.9}$$

Let α be the internal angle of the boundary, then at a smooth point of the boundary, in a Lipschitz sense, the internal angle equals π, whereas in a corner point $\mathbf{x_i}$ the internal angle of the corner is θ_i. In a well posed boundary value problem one must either know ϕ or q_n in any portion of the boundary, with at least in one portion ϕ shall be known.

If the source point is located outside of the domain, all the integrals in Eq. (4.7) are regular. When the source point is on the boundary, the integrals containing ϕ_i^* have a weak integrable singularity, and the integrals containing q_i^* have a strong singularity and must be interpreted using the Cauchy principal value. In this

contribution we will work with point sources located at the nodes which are used to interpolate the boundary's geometry.

4.3 Numerical Implementation

This section presents the numerical implementation of the method, and is divided into the boundary discretization, the numerical integration, and source term treatment.

4.3.1 Discretization and Matrix System

The boundary is discretized using linear continuous elements everywhere, except for the points where the boundary has a corner or even a possible discontinuity as boundary condition. This is necessary since, in a corner, the outward normal vector does not have a well defined direction—its limits differ when we approach from either side of the corner. Likewise, q_n may differ at either side of the corner, as will become apparent in the benchmark study.

The solution for the corner treatment used in this study was to apply a discontinuous formulation for the corner node. This effectively duplicates the node and pushes it somewhere into the boundary elements forming the corner, by some parametric form. Using the boundary elements defined by N nodes we can write Eq. (4.7) in its discrete form, recalling that the shape functions ψ_j referring to each node j have compact support, while the fundamental solution ϕ_i^* for each source point i does not.

$$c_i\,\phi_i + \sum_{j=1}^{N} \phi_j \int_{\Gamma_j} \psi_j \left(d\,q_i^* + v_n\,\phi_i^*\right)\, d\Gamma_j - \sum_{j=1}^{N} q_{nj} \int_{\Gamma_j} \psi_j\,\phi_i^* d\Gamma_j = \int_{\Omega} f\,\phi_i^* d\Omega \tag{4.10}$$

Equations (4.11) and (4.12) define two matrices that will multiply the nodal values of ϕ and q_n, respectively, and Eq. (4.13) will define the source vector.

$$H_{ij} = \delta_{ij}\,c_i + \int_{\Gamma_j} \psi_j \left(d\,q_i^* + v_n\,\phi_i^*\right)\, d\Gamma_j \tag{4.11}$$

$$G_{ij} = \int_{\Gamma_j} \psi_j\,\phi_i^* d\Gamma_j \tag{4.12}$$

$$b_i = \int_{\Omega} f\,\phi_i^* d\Omega \tag{4.13}$$

Here δ_{ij} is the usual Kronecker delta. After having carried out all the integrations, the following system of equations arises:

$$H_{ij}\phi_j - G_{ij}q_{nj} = b_i \tag{4.14}$$

Combining Eqs. (4.10)–(4.13) by rearranging the columns of the matrices H_{ij} and G_{ij} in such a way that all the boundary unknowns are collected in a vector χ_j, we arrive at the final system of equations:

$$A_{ij}\chi_j = a_i \tag{4.15}$$

The system defined in Eq. (4.15) is then solved numerically. Matrix A_{ij} is a full and asymmetric matrix [BrEtAl84].

4.3.2 Numerical Integration

The integrations performed to assemble H_{ij} and G_{ij} can be either regular, weakly singular, or strongly singular. Each requires a different kind of numerical treatment. When the source point is on the element to be integrated, then the integral may be singular, otherwise it is regular.

Starting with the regular integrals, every integral, where the source point is located outside of the element to be integrated, is treated in the same way. This requires some care, since when the source point is close to the integrated domain—close but not in it—the integral is quasi-singular and needs special treatment to improve accuracy. Telles [Te87] showed that these problems arise when the source is too close to the domain to be integrated. By setting the discontinuity parameter α to be greater than 0.1 there is no need for special treatment in any boundary integral. When calculating concentration values in internal points, the point cannot be placed too close to the boundary; however, this critical distance value changes depending on the direction of the velocity vector. All regular integrals are evaluated with the Gauss–Legendre quadrature.

Qiu [QiEtAl98] showed the asymptotic behaviour of the singularities present in the 2D advection–diffusion–reaction problem. The G_{ij} matrix has weakly singular kernels to be calculated in its diagonal, and the H_{ij} matrix has a combination of strongly and weakly singular terms in its diagonal. Weakly singular kernels were integrated using the cubic coordinate transformation proposed by Telles [Te87].

Using the Gauss–Legendre quadrature and the Telles transformation, the entire G matrix can be calculated. However, the H matrix has terms that are strongly singular and cannot be handled by a coordinate transformation. In Eq. (4.17) the H matrix terms are to be integrated, nevertheless, the strongly and weakly singular parts may be treated separately.

$$H_{ii} = c_i + \int_{-1}^{1} J\,(a_i\xi + b_i)\,\exp\left(-\frac{\mathbf{v}\cdot\mathbf{r}}{2d}\right)\times \tag{4.16}$$
$$\times\left[-\frac{1}{2\pi}\left(\mu K_1(\mu r)\frac{\mathbf{r}\cdot\mathbf{n}}{r}+\frac{v_n}{2d}K_0(\mu r)\right)+\frac{v_n}{2\pi d}K_0(\mu r)\right]\,d\xi$$

In Eq. (4.17) we have, from left to right, the Jacobian of the transformation into normalized space, the shape function relating to the i-th node, the exponential term representing the upwind effect, the terms with K_1, which are strongly singular, and finally the terms with K_0 which are weakly singular.

In problems modelled with the Laplace fundamental solution and in advection–diffusion problems (without reaction) the diagonal term of the H matrix can be calculated indirectly, as proposed by Brebbia and Dominguez [BrDo92]. This technique calculates the diagonal term, including the free coefficient c_i, by combining the terms of the other columns in the matrix. In problems with a reaction contribution this technique cannot be applied.

In our case the geometrical interpolation is linear and as a consequence the radius $\mathbf{r}$ is always perpendicular to the boundary normal vector, making the term multiplying K_1 to be identically zero on the integration domain, thus leaving only terms with K_0 to be integrated. Those terms are weakly singular and are handled with the Telles coordinate transformation.

Figure 4.1 shows that the weakly singular kernels are regularized if the shape function in question tends to zero at the source point. This means that the only G-matrix terms that need the Telles transformation are the diagonal terms. This transformation can also be applied to the terms that are regularized by the shape functions.

Figure 4.2 shows the relative error of the weakly singular integrals, for an increasing number of quadrature points and using the Telles transformation. As shown in the figure, 12 points are needed to evaluate the kernels with a relative error of 10^{-6}, compared to a reference result obtained by the software Maple [MoEtAl11].

4.3.3 Analytic Treatment of the Source Term

In the cases studied in this contribution, we are considering the source to be the sum of M point sources. Modelling the point sources as Dirac delta distributions, and using their filtering property, we can evaluate the domain integral in Eq. (4.13) analytically. The source vector can then be evaluated as shown in Eq. (4.17).

$$b_i = \int_{\Omega}\sum_{k=1}^{M} S_k\,\delta_k\phi_i^*\,d\Omega = \sum_{k=1}^{M} S_k\phi_i^*(\mathbf{x_k}) \tag{4.17}$$

Here S_k is the source strength, and δ_k is a Dirac delta distribution at the point $\mathbf{x_k}$.

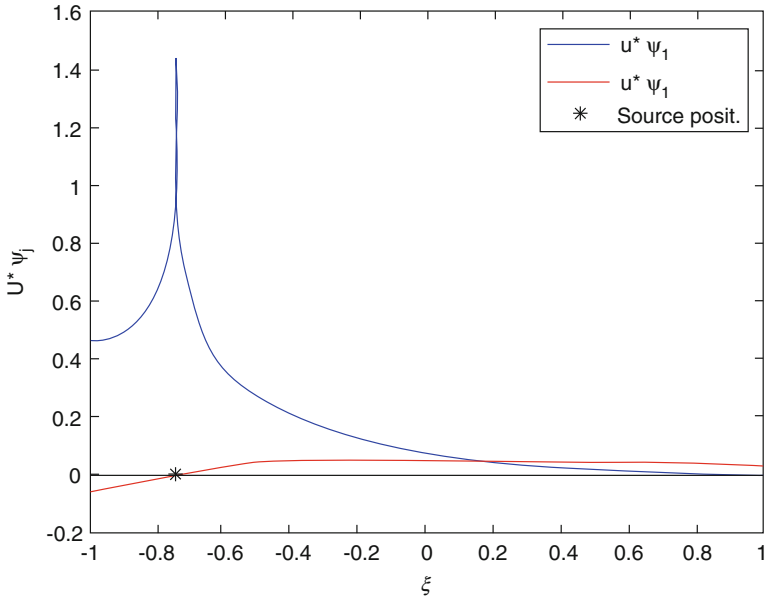


Fig. 4.1 Weakly singular kernel behaviour; when the shape functions tends to 1 at the source point the kernel is weakly singular, when the shape function tends to zero at the source point the kernel is regular

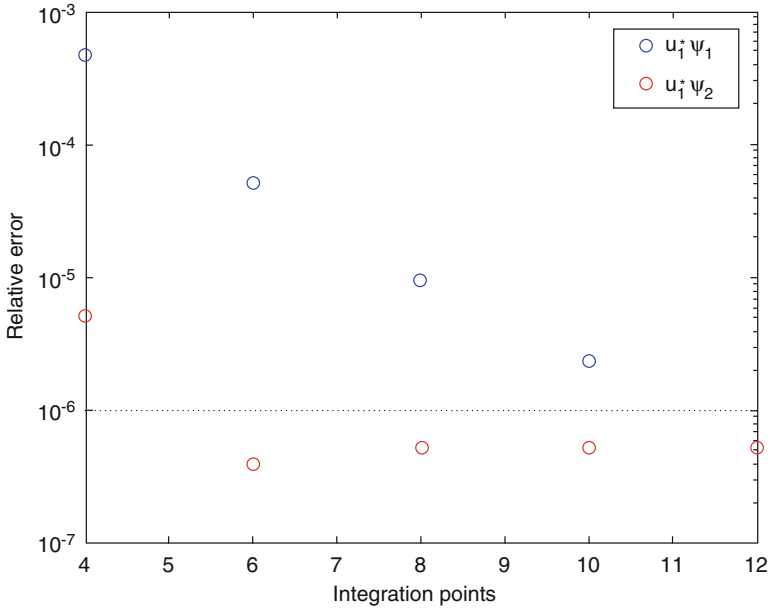


Fig. 4.2 Relative error of the weakly singular integrals, for an increasing number of quadrature points and using the Telles transformation

4.4 Benchmark Case

In this section we will test our implementation for a classical benchmark case. Figure 4.3 shows schematically a unity square domain, with a uniform velocity field in the positive x-direction. The left and right boundaries have primal boundary conditions, corresponding to concentrations of zero and one, respectively. The top and bottom boundaries have dual boundary conditions, with null flux and in this problem no reaction term was present.

This case has an analytical solution, given by Eq. (4.18). Note, that the higher the velocity, the higher the gradient presented by this solution, and traditionally the harder it is to solve this problem numerically.

$$\phi(x, y) = \frac{\exp\left(\frac{vx}{d}\right) - 1}{\exp\left(\frac{v}{d}\right) - 1} \tag{4.18}$$

This problem needs only boundary discretization and has no source term. Applying the methods described in Sect. 4.3 we will now analyse the errors presented by the method in this benchmark case. Figure 4.4 shows the boundary results, in a case with $v = 5$, $d = 1$, element size of 0.025 (resulting in 160 elements), and with 8 quadrature points for the integrals. In the following subsections, we will analyse the behaviour of the implemented method for a set of Peclet numbers, different mesh sizes, and different number of quadrature points.

4.4.1 Concentration Profiles for Various Peclet Numbers

Qiu [QiEtAl98] showed that the boundary element method for a problem with constant coefficients and using the advection–diffusion–reaction fundamental solution is stable for any Peclet number, independent of the used mesh size, given that the integrals are correctly evaluated. It is noteworthy, that for high Peclet numbers the evaluation of the fundamental solution may be problematic. When the scalar product of the velocity by the distance from the source point is large

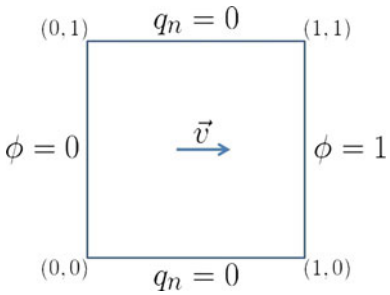


Fig. 4.3 Schematic description of the benchmark problem and its boundary conditions

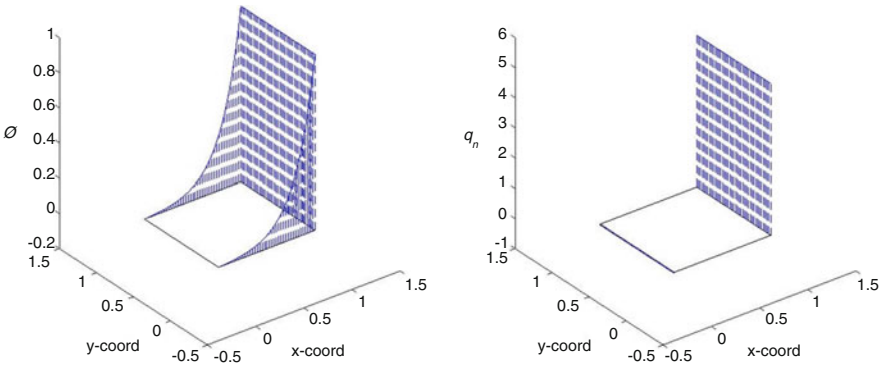


Fig. 4.4 Boundary results for the benchmark problem, with $v = 5$, $d = 1$, element size of 0.025, and 8 quadrature points for the integrals

and negative, the argument of the exponential becomes too large (depending on the floating point representation) and causes numerical overflow. At the same time, the implementation of the Bessel function of the second kind will cause underflow, while the product of the two should be a finite small number, but under these conditions it becomes erroneous or simply cannot be evaluated numerically. This shortcoming can be avoided by rewriting the fundamental solution as given in Eq. (4.19). The product of the Bessel function of the second kind and the exponential of the same argument in square brackets is called normalized Bessel function of the second kind, and its evaluation can be accomplished without numerical underflow or overflow. The same can be done for the derivatives of the fundamental solution.

$$\begin{aligned}\phi_i^*(\mathbf{x}) &= \frac{1}{2\pi d}\exp(-\frac{\mathbf{v}\cdot\mathbf{r}}{2d})K_0(\mu r)\\ &= \frac{1}{2\pi d}\exp(-\frac{\mathbf{v}\cdot\mathbf{r}}{2d}-\mu r)\left[\exp(\mu r)\,K_0(\mu r)\right]\end{aligned} \tag{4.19}$$

Figure 4.5 shows the results for concentration profiles at internal points of the domain ($y = 0.5$) of the benchmark problem. The problem is solved with 160 elements, 8 quadrature points for the integrals, $d = 1$, and with varying velocity. Along the x-direction a homogeneous mesh with element size 0.05 was used resulting in 20 elements along this direction. The Peclet numbers ran from 10^{-3} to 30 and the maximum relative error observed in these cases was of the order 10^{-4}.

4.4.2 Mesh Size Sensitivity

Table 4.1 shows the maximum relative error and the mean square error for boundary nodes and internal points (at $y = 0.5$) when varying the mesh size while keeping the

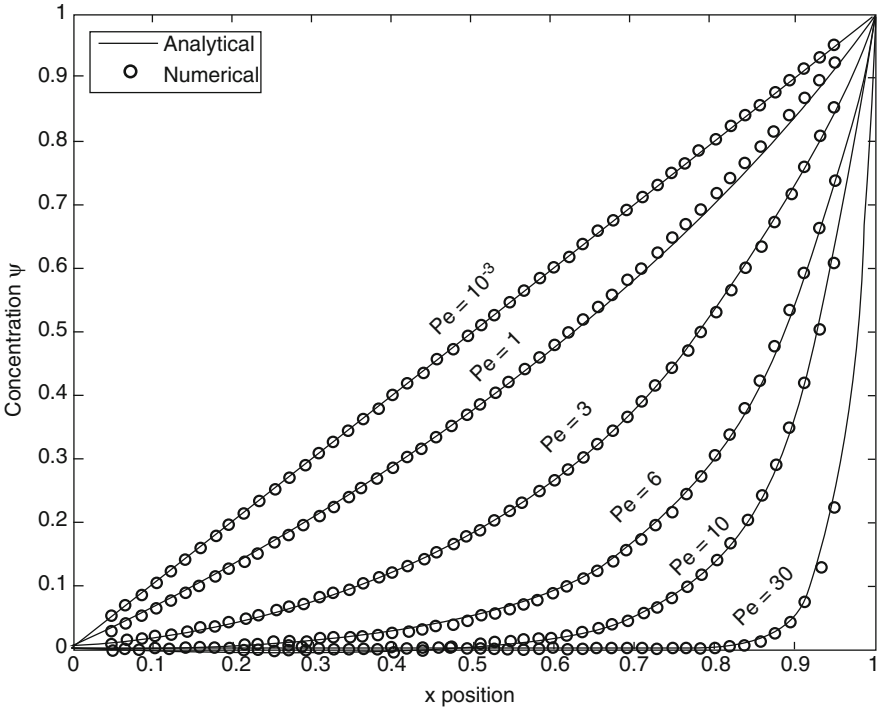


Fig. 4.5 Concentrations at internal points ($y = 0.5$), for various Peclet numbers

Table 4.1 Relative errors for varying element size on the benchmark problem with $v = 5$, $d = 1$, and 8 quadrature points

	Error at boundary nodes		Error at internal nodes	
Element size	$E_{\max}$	E_{sqrt}	$E_{\max}$	E_{sqrt}
0.100	6.38%	1.20%	0.55%	0.03%
0.050	3.17%	0.46%	0.13%	0.01%
0.025	1.57%	0.17%	0.03%	0.002%

velocity constant $v = 5$, the diffusivity $d = 1$, and the number of integration points set to 8. One observes that with relatively coarse meshes low numerical errors were attained. Results for internal points produced errors even smaller than the ones on the boundary nodes.

Table 4.2 Relative errors for varying number of quadrature points of the benchmark problem with $v = 5$, $d = 1$, and element size of 0.05

	Error at boundary nodes		Error at internal nodes	
Quadrature points	$E_{\max}$	E_{sqrt}	$E_{\max}$	E_{sqrt}
4	3.51%	0.51%	0.15%	0.01%
8	3.17%	0.46%	20.13%	0.01%
16	3.17%	0.46%	0.13%	0.01%

4.4.3 Sensitivity to the Quadrature Order

Table 4.2 shows absolute and relative errors for boundary and interior points (at $y = 0.5$) for three different numbers of quadrature points. A fixed mesh of element size 0.05 (20 elements for each size of the square) was used to solve a problem with velocity $v = 5$ and the diffusivity $d = 1$. Any difference in the numerical values in Table 4.2 is due to the quality of the numerical integration performed. As can be seen, no significant difference can be observed between routines using 8 or 16 quadrature points.

4.5 Applications with Point Sources

Once the solution of the advection–diffusion problem for one point source is obtained, one may extend the solution to in principle any source distribution by the use of the Green function theory. In the Fig. 4.6 the influence of the Peclet number, i.e. the advective to the diffusive transport rate, for point source responses is shown. The contour plots show in three different cases, with Peclet numbers of 2 (top), 20 (middle), and 200 (bottom) the increasing dominance of advection over diffusion. In the three problems the diffusivity was set to $d = 1$, a non-zero velocity was assumed for the x-direction only, no reactions were considered, and one point source with magnitude 10 was taken into account. For the boundary conditions null concentrations along the edge $x = 0$ and null diffusive fluxes on the remaining boundaries was understood.

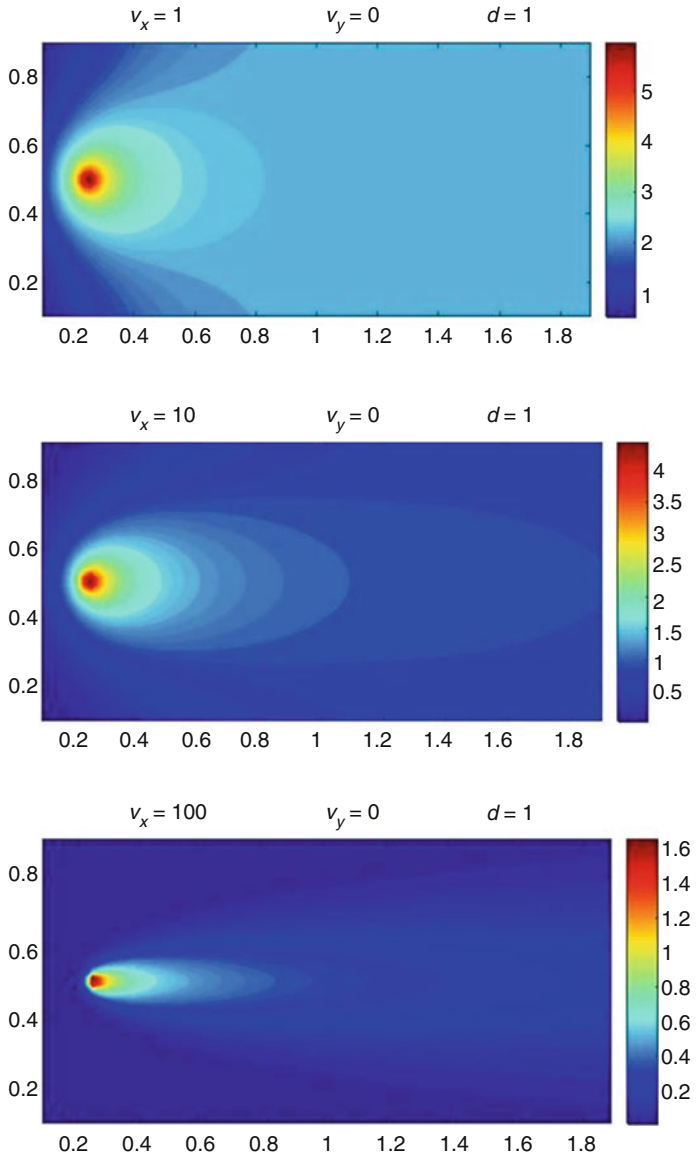


Fig. 4.6 Concentration fields for one point source problems with varying Peclet numbers $Pe = 2$ (top), 20 (middle), and 200 (bottom) are shown. The source magnitude was set to 10 for all cases

4.6 Conclusions

In this work we solved the advection–diffusion–reaction equation in a rectangular domain considering a point source. We also showed that the method is stable for a large range of Peclet numbers and maintains precision even with relatively coarse meshes and relatively few integration points. All singular integrals were regularized by a coordinate transformation [Te87], so that the standard Gauss–Legendre quadrature integration was efficient.

It is well known that the solution of an arbitrary domain subjected to a single known source point is a Green function of the problem. Therefore, the present formulation offers a numerical alternative to other methods provided the Green function method may be applied. The present results show stability, accuracy, and ability to capture the singularities around point sources, and can be used in principle for an arbitrary number of sources. The formulation presented, along with only the boundary discretization, has potential to be used in complex problems like those considering actual terrain topographies and can be extended to the three dimensional case without changing the presently prescribed procedure.

References

[BrDo92] Brebbia, C., Dominguez, J.: Boundary Elements, an Introductory Course. WIT Press, Boston (1992)

[BrEtAl84] Brebbia, C., Telles, J., Wrobel, L.: Boundary Element Techniques. Springer, Berlin (1984)

[CuEtAl16] Cunha, C., Carrer, J., Oliveira, M., Costa, V.: A study concerning the solution of advection-diffusion problems by the Boundary Element Method. Eng. Anal. Bound. Elem. **65**, 79–94 (2016)

[IkOn83] Ikeuchi, M., Onishi, K.: Boundary element solutions to steady convective diffusion equations. Appl. Math. Model. **7**(2), 115–118 (1983)

[MeEtAl17] Meneghetti, A., Bodmann, B.E.J., Vilhena, M.T.M.B.: A New Diffeomorph Conformal Methodology to Solve Flow Problems with Complex Boundaries by an Equivalent Plane Parallel Problem. In: Constanda, Ch., Dalla Riva, M., Lamberti, P.D., Musolino, P. (eds.) Integral Methods in Science and Engineering, vol 1, Theoretical Techniques, pp. 205–214. Birkhauser, New York (2017)

[MoEtAl11] Monagan, M.B., et al.: Maple 15 Programming Guide. Maplesoft, Waterloo (2011)

[QiEtAl98] Qiu, Z.H., Wrobel, L.C., Power, H.: Numerical solution of convection-diffusion problems at high Peclet number using boundary elements. Int. J. Numer. Methods Eng. **41**(5), 899–914 (1998)

[Ro72] Roache, P.: Computational Fluid Dynamics. Hermosa Publishers, Washington (1972)

[Te87] Telles, J.C.F.: A self-adaptive co-ordinate transformation for efficient numerical evaluation of general boundary element integrals. Int. J. Numer. Methods Eng. **24**(5), 959–973 (1987)

Chapter 5
Displacement Boundary Value Problem for a Thin Plate in an Unbounded Domain

Christian Constanda and Dale Doty

5.1 Introduction

Many times, solutions of boundary value problems for a mathematical model cannot be computed explicitly, but can be approximated to within acceptable tolerances by means of expansions in a complete set of functions in a Hilbert space such as L^2. This method acquires additional interest and usefulness when the choice of the functions in question is based on the structure of the layer potentials generated by the problem.

An expansion of this type may encounter difficulties in the case of an infinite domain, where the solution must fit a certain prescribed far-field pattern. Often, such a pattern requires the a priori knowledge of what is known as a rigid displacement in elasticity theory, which is not normally readily available.

In this chapter, we construct a generalized Fourier series method for the approximation of the solution to the Dirichlet problem associated with the bending of a thin elastic plate with transverse shear deformation that occupies a region $S^- \times [-h_0/2, h_0/2]$, $h_0 = \text{const}$, in $\mathbb{R}^3$, where S^- is the complement to $\mathbb{R}^2$ of a finite domain S^+ bounded by a simple, closed, C^2-curve ∂S. After the completion of the analytic part that describes and justifies the procedure, we illustrate the method by means of two numerical examples in the case of a homogeneous and isotropic material, with S^- representing the outside of a circle.

The Dirichlet, Neumann, and Robin problems in S^+ for this model have been investigated in [CoDo17a, CoDo17b, CoDo18, CoDo19a, CoDo19b, CoDo19c, CoDo20].

C. Constanda (✉) · D. Doty
The University of Tulsa, Tulsa, OK, USA
e-mail: christian-constanda@utulsa.edu; dale-doty@utulsa.edu

C. Constanda (ed.), *Computational and Analytic Methods in Science and Engineering*, https://doi.org/10.1007/978-3-030-48186-5_5

5.2 The Plate Equations

In what follows, $x(x_1, x_2)$ and $y(y_1, y_2)$ are generic points in $\mathbb{R}^2$, $|x - y|$ is the Cartesian distance between x and y, $C^{0,\alpha}(\partial S)$ and $C^{1,\alpha}(\partial S)$, $\alpha \in (0, 1)$, are the spaces of Hölder continuous functions and Hölder continuously differentiable functions on ∂S, respectively, $\|\cdot\|$ is the norm on $L^2(\partial S)$, $M^{(i)}$, $M_{(i)}$, and M^{T} are the columns, rows, and transpose of a matrix M, and I is the identity operator or matrix.

We denote by λ and μ the Lamé constants of the material, and by

$$\big(x_3 u_1(x_1, x_2), x_3 u_2(x_1, x_2), u_3(x_1, x_2)\big)^{\mathsf{T}}$$

the three-dimensional displacement of the points in the plate. Since our mathematical model is set up by means of averaging all the quantities involved in its description over the thickness h_0 of the plate, the unknown object in our analysis is the vector function

$$u = (u_1, u_2, u_3)^{\mathsf{T}}.$$

The equilibrium system of partial differential equations when the body forces and moments are negligible can be written as [Co16]

$$A(\partial_1, \partial_2)u(x) = 0, \tag{5.1}$$

where

$$A(\partial_1, \partial_2) = \begin{pmatrix} h^2\mu\Delta + h^2(\lambda+\mu)\partial_1^2 - \mu & h^2(\lambda+\mu)\partial_1\partial_2 & -\mu\partial_1 \\ h^2(\lambda+\mu)\partial_1\partial_2 & h^2\mu\Delta + h^2(\lambda+\mu)\partial_2^2 - \mu & -\mu\partial_2 \\ \mu\partial_1 & \mu\partial_2 & \mu\Delta \end{pmatrix},$$

$h = h_0/\sqrt{12}$, $\partial_\alpha = \partial/\partial x_\alpha$, $\alpha = 1, 2$, and $\Delta = \partial_1^2 + \partial_2^2$ is the two-dimensional Laplacian.

It is easily verified [Co16] that the columns $f^{(i)}$ of the matrix

$$F = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ -x_1 & -x_2 & 1 \end{pmatrix}, \quad x \in \mathbb{R}^2,$$

form a basis for the space $\mathscr{F}$ of rigid displacements, so an arbitrary rigid displacement can be written as

$$f = Fc,$$

where $c = (c_1, c_2, c_3)^{\mathsf{T}}$ is a constant vector. Also, the restrictions $f^{(i)}|_{\partial S}, i = 1, 2, 3$ form a basis for the null space of the boundary operator T.

The study of system (5.1) by means of the boundary integral equation method requires the knowledge of a 3×3 matrix $D(x, y)$ of fundamental solutions. Such a matrix, constructed in [Co16], has the symmetry

$$D(x, y) = D^{\mathsf{T}}(y, x) \tag{5.2}$$

and is accompanied by its associated matrix of singular solutions

$$P(x, y) = \big(T(\partial_y) D(y, x)\big)^{\mathsf{T}}, \tag{5.3}$$

where T is the boundary moment-force operator defined by

$$T(\partial_1, \partial_2) = \begin{pmatrix} h^2(\lambda + 2\mu) n_1 \partial_1 + h^2 \mu n_2 \partial_2 & h^2 \mu n_2 \partial_1 + h^2 \lambda n_1 \partial_2 & 0 \\ h^2 \lambda n_2 \partial_1 + h^2 \mu n_1 \partial_2 & h^2 \mu n_1 \partial_1 + h^2(\lambda + 2\mu) n_2 \partial_2 & 0 \\ \mu n_1 & \mu n_2 & \mu n_\alpha \partial_\alpha \end{pmatrix},$$

$n = (n_1, n_2)^{\mathsf{T}}$ is the unit of the outward normal to ∂S, and $n_\alpha \partial_\alpha = n_1 \partial_1 + n_2 \partial_2$.

5.3 Dirichlet Boundary Value Problem

Adopting the notation $x_p(r, \theta)$ for the point x in polar coordinates, we consider the space $\mathscr{A}$ of vector functions in S^- that, as $r \to \infty$, admits an asymptotic expansion of the form

$$\begin{aligned} u_1(r, \theta) &= r^{-1}\big[m_0 \sin\theta + 2m_1 \cos\theta - m_0 \sin(3\theta) + (m_2 - m_1)\cos(3\theta)\big] \\ &\quad + r^{-2}\big[(2m_3 + m_4)\sin(2\theta) + m_5 \cos(2\theta) - 2m_3 \sin(4\theta) + 2m_6 \cos(4\theta)\big] \\ &\quad + r^{-3}\big[2m_7 \sin(3\theta) + 2m_8 \cos(3\theta) + 3(m_9 - m_7)\sin(5\theta) \\ &\qquad\qquad + 3(m_{10} - m_8)\cos(5\theta)\big] + O(r^{-4}), \\ u_2(r, \theta) &= r^{-1}\big[2m_2 \sin\theta + m_0 \cos\theta + (m_2 - m_1)\sin(3\theta) + m_0 \cos(3\theta)\big] \\ &\quad + r^{-2}\big[(2m_6 + m_5)\sin(2\theta) - m_4 \cos(2\theta) + 2m_6 \sin(4\theta) + 2m_3 \cos(4\theta)\big] \\ &\quad + r^{-3}\big[2m_{10} \sin(3\theta) - 2m_9 \cos(3\theta) + 3(m_{10} - m_8)\sin(5\theta) \\ &\qquad\qquad + 3(m_7 - m_9)\cos(5\theta)\big] + O(r^{-4}), \end{aligned}$$

$$\begin{aligned} u_3(r,\theta) = {} & -(m_1+m_2)\ln r - \big[m_1+m_2+m_0\sin(2\theta)+(m_1-m_2)\cos(2\theta)\big] \\ & + r^{-1}\big[(m_3+m_4)\sin\theta + (m_5+m_6)\cos\theta - m_3\sin(3\theta)+m_6\cos(3\theta)\big] \\ & + r^{-2}\big[m_{11}\sin(2\theta)+m_{12}\cos(2\theta)+(m_9-m_7)\sin(4\theta) \\ & \qquad + (m_{10}-m_8)\cos(4\theta)\big] + O(r^{-3}), \end{aligned} \tag{5.4}$$

where $m_1, \dots, m_{12}$ are arbitrary real constants. We also consider the set

$$\mathscr{A}^* = \mathscr{A} \oplus \mathscr{F}.$$

In what follows, we assume that the origin lies in S^+. Direct calculation shows that, for y fixed, matrix $D(x_p, y)$ splits into the sum

$$D(x_p, y) = D^{\mathscr{A}}(x_p, y) + D^{\infty}(x_p, y); \tag{5.5}$$

here, $D^{\mathscr{A}}$ is a matrix whose columns $(D^{\mathscr{A}})^{(i)}$ are vector functions of class $\mathscr{A}$, and the entries of the residual D^{∞} are

$$D_{11}^{\infty}(x,y) = -a\mu^2\left\{\frac{x_1^2-x_2^2}{x_1^2+x_2^2} + 2 + \ln(x_1^2+x_2^2)\right\},$$

$$D_{21}^{\infty}(x,y) = -2a\mu^2\frac{x_1x_2}{x_1^2+x_2^2},$$

$$D_{31}^{\infty}(x,y) = a\left\{-4h^2\mu(\lambda+2\mu)\frac{x_1}{x_1^2+x_2^2} + \mu^2x_1[1+\ln(x_1^2+x_2^2)]\right\},$$

$$D_{12}^{\infty}(x,y) = -2a\mu^2\frac{x_1x_2}{x_1^2+x_2^2},$$

$$D_{22}^{\infty}(x,y) = -a\mu^2\left\{\frac{-x_1^2+x_2^2}{x_1^2+x_2^2} + 2 + \ln(x_1^2+x_2^2)\right\},$$

$$D_{32}^{\infty}(x,y) = a\left\{-4h^2\mu(\lambda+2\mu)\frac{x_2}{x_1^2+x_2^2} + \mu^2x_2[1+\ln(x_1^2+x_2^2)]\right\},$$

$$\begin{aligned} D_{13}^{\infty}(x,y) = -a\mu^2\Bigg\{ & \frac{(-x_1^2+x_2^2)y_1 - 2x_1x_2y_2}{x_1^2+x_2^2} \\ & + x_1[1+\ln(x_1^2+x_2^2)] - y_1[2+\ln(x_1^2+x_2^2)]\Bigg\}, \end{aligned}$$

$$D_{23}^{\infty}(x,y) = -a\mu^2\Bigg\{\frac{(x_1^2-x_2^2)y_2 - 2x_1x_2y_1}{x_1^2+x_2^2} + x_2[1+\ln(x_1^2+x_2^2)] - y_2[2+\ln(x_1^2+x_2^2)]\Bigg\},$$

$$D_{33}^{\infty}(x,y) = a\Bigg\{-4h^2\mu(\lambda+3\mu) + 4h^2\mu(\lambda+2\mu)\frac{x_1y_1+x_2y_2}{x_1^2+x_2^2} + \tfrac{1}{2}\mu^2(x_1^2+x_2^2)\ln(x_1^2+x_2^2) - 2h^2\mu(\lambda+2\mu)\ln(x_1^2+x_2^2) - \mu^2(x_1y_1+x_2y_2)[1+\ln(x_1^2+x_2^2)]\Bigg\},$$

where

$$a = [8\pi h^2\mu^2(\lambda+2\mu)]^{-1}.$$

Obviously, in view of (5.5), matrix P also splits into a sum of two terms, $P^{\mathscr{A}}$ and P^{∞}, constructed from (5.3) with $D^{\mathscr{A}}$ and D^{∞}, respectively.

It is easily verified that, for $x \neq y$,

$$AD^{\infty} = AD^{\mathscr{A}} = 0 \quad \text{in } S^+ \cup S^-$$

and

$$P^{\infty} = 0, \quad P = P^{\mathscr{A}}.$$

Using the expressions of D_{ij}^{∞} converted to polar coordinates, we find that for a vector function ψ on ∂S,

$$\begin{aligned}
&\big(D^{\infty}(x_p,y)\psi(y)\big)_1 \\
&\quad = -a\mu^2\{[r(2\ln r+1)\cos\theta]\psi_3 \\
&\qquad + [2(\ln r+1)+\cos(2\theta)](\psi_1 - y_1\psi_3) + \sin(2\theta)(\psi_2 - y_2\psi_3)\}, \\
&\big(D^{\infty}(x_p,y)\psi(y)\big)_2 \\
&\quad = -a\mu^2\{[(2\ln r+1)\sin\theta]\psi_3 + \sin(2\theta)(\psi_1 - y_1\psi_3) \\
&\qquad + [2(\ln r+1)-\cos(2\theta)](\psi_2 - y_2\psi_3)\}, \\
&\big(D^{\infty}(x_p,y)\psi(y)\big)_3 \\
&\quad = a\{[\mu^2r^2\ln r - 4h^2\mu(\lambda+2\mu)\ln r - 4h^2\mu(\lambda+3\mu)]\psi_3
\end{aligned}$$

$$+ [\mu^2 r(2\ln r + 1) - 4h^2\mu(\lambda + 2\mu)r^{-1}]$$
$$\times [(\cos\theta)(\psi_1 - y_1\psi_3) + (\sin\theta)(\psi_2 - y_2\psi_3)]\}.$$

These formulas show that if

$$\begin{aligned} &\int_{\partial S} (f^{(\alpha)})^{\mathsf{T}}\psi \, ds = \int_{\partial S} [\psi_\alpha(y) - y_\alpha\psi_3(y)] \, ds(y) = 0, \quad \alpha = 1, 2, \\ &\int_{\partial S} (f^{(3)})^{\mathsf{T}}\psi \, ds = \int_{\partial S} \psi_3(y) \, ds(y) = 0, \end{aligned} \tag{5.6}$$

then

$$\int_{\partial S} \big(D(x_p, y)\big)^{\infty}\psi(y) \, ds(y) = 0,$$

which means that

$$\int_{\partial S} D(x, y)\psi(y) \, ds(y) = \int_{\partial S} \big(D(x, y)\big)^{\mathscr{A}}\psi(y) \, ds(y). \tag{5.7}$$

The exterior Dirichlet problem consists in finding $u \in C^2(S^-) \cap C^1(\bar{S}^-)$ that satisfies

$$\begin{aligned} &Au = 0 \quad \text{in } S^-, \\ &u = \mathscr{D} \quad \text{on } \partial S, \\ &u \in \mathscr{A}^*, \end{aligned} \tag{5.8}$$

where $\mathscr{D}$ is a 3×1 vector function prescribed on ∂S.

The next two statements are proved in [Co16].

Theorem 1 (Somigliana Representation Formula) *If $u \in \mathscr{A}$ is a solution of $Au = 0$ in S^-, then*

$$u(x) = -\int_{\partial S} \{D(x, y)Tu(y) - P(x, y)\mathscr{D}(y)\} \, ds(y), \quad x \in S^-,$$

$$0 = -\int_{\partial S} \{D(x, y)Tu(y) - P(x, y)\mathscr{D}(y)\} \, ds(y), \quad x \in S^+.$$

Theorem 2 *Problem (5.8) has a unique solution u for any $\mathscr{D} \in C^{1,\alpha}(\partial S)$.*

Given the definition of $\mathscr{A}^*$, we have the decomposition

$$u = u^{\mathscr{A}} + Fc, \quad u^{\mathscr{A}} \in \mathscr{A}. \tag{5.9}$$

It is obvious that $u^{\mathscr{A}}$ is the (unique) solution of the Dirichlet boundary value problem

$$\begin{aligned} Au^{\mathscr{A}} &= 0 \quad \text{in } S^-, \\ u^{\mathscr{A}} &= \mathscr{D} - Fc \quad \text{on } \partial S, \\ u^{\mathscr{A}} &\in \mathscr{A}, \end{aligned}$$

so, by Theorem 1,

$$\begin{aligned} u^{\mathscr{A}}(x) = &-\int_{\partial S} \{D(x,y)Tu^{\mathscr{A}}(y) - P(x,y)(\mathscr{D} - Fc)(y)\}\, ds(y), \quad x \in S^-, \\ 0 = &-\int_{\partial S} \{D(x,y)Tu^{\mathscr{A}}(y) - P(x,y)(\mathscr{D} - Fc)(y)\}\, ds(y), \quad x \in S^+. \end{aligned} \tag{5.10}$$

Since $TFc = 0$, we have

$$Tu = Tu^{\mathscr{A}} + TFc = Tu^{\mathscr{A}}.$$

Also, since $\psi = Tu$ is the Neumann boundary data of the solution, it follows that (see [Co16])

$$\int_{\partial S} (f^{(i)})^{\mathsf{T}} \psi \, ds = 0, \quad i = 1, 2, 3, \tag{5.11}$$

which is (5.6). Consequently, using (5.7) and (5.9), and the equality [Co16]

$$\int_{\partial S} P(x,y)(Fc)(y)\, ds(y) = \begin{cases} -(Fc)(x), & x \in S^+, \\ 0, & x \in S^-, \end{cases}$$

we bring (5.10) to the form

$$u^{\mathscr{A}}(x) = -\int_{\partial S} \{D^{\mathscr{A}}(x,y)\psi(y) - P(x,y)\mathscr{D}(y)\}\, ds(y), \quad x \in S^-, \tag{5.12}$$

$$0 = -\int_{\partial S} \{D^{\mathscr{A}}(x,y)\psi(y) - P(x,y)\mathscr{D}(y)\}\, ds(y) + (Fc)(x), \quad x \in S^+. \tag{5.13}$$

5.4 Generalized Fourier Series Method

The computational algorithm is constructed by means of an auxiliary simple, closed, C^2–curve ∂S_* lying strictly inside S^+.

Let $\{x^{(k)},\ k = 1, 2, \ldots\}$ be a set of points densely distributed on ∂S_*, and let

$$\mathscr{G} = \{\varphi^{(ik)},\ i = 1, 2, 3,\ \ k = 0, 1, 2, \ldots\}$$

be the set of vector functions on ∂S defined by

$$\varphi^{(i0)}(x) = f^{(i)}(x), \quad i = 1, 2, 3,$$
$$\varphi^{(ik)}(x) = (D^{\mathscr{A}})^{(i)}(x, x^{(k)}), \quad i = 1, 2, 3,\ \ k = 1, 2, \ldots.$$

The functions $\varphi^{(ik)}$, $k = 1, 2, \ldots$, can also be expressed in terms of the rows of $D^{\mathscr{A}}$. From the symmetry (5.2) we deduce that

$$D^{\mathscr{A}}(y, x) = (D^{\mathscr{A}})^{\mathsf{T}}(x, y),$$

where the left-hand side is the class $\mathscr{A}$ component of the asymptotic expansion with respect to x after the points x and y have been interchanged in D. Then

$$(D^{\mathscr{A}})^{(i)}(y, x) = \left((D^{\mathscr{A}})^{\mathsf{T}}\right)^{(i)}(x, y) = (D^{\mathscr{A}})_{(i)}(x, y),$$

which leads to

$$\varphi^{(ik)}(x) = (D^{\mathscr{A}})^{(i)}(x, x^{(k)}) = \left((D^{\mathscr{A}})_{(i)}(x^{(k)}, x)\right)^{\mathsf{T}}. \tag{5.14}$$

Theorem 3 *$\mathscr{G}$ is linearly independent on ∂S and complete in $L^2(\partial S)$.*

Proof Suppose that

$$\sum_{j=1}^{3}\sum_{k=0}^{N}\alpha_{jk}\varphi^{(jk)}(x) = 0, \quad x \in \partial S$$

for some positive integer N and real numbers α_{ik}, $i = 1, 2, 3$, $k = 0, 1, 2, \ldots, N$. Then the function

$$g(x) = \sum_{i=1}^{3}\sum_{k=0}^{N}\alpha_{ik}\varphi^{(ik)}(x) = \sum_{i=1}^{3}\alpha_{i0}f^{(i)}(x) + \sum_{i=1}^{3}\sum_{k=1}^{N}\alpha_{ik}(D^{\mathscr{A}})^{(i)}(x, x^{(k)})$$

is the unique solution of the homogeneous Dirichlet problem

$$\begin{aligned} Ag &= 0 \quad \text{in } S^-, \\ g &= 0 \quad \text{on } \partial S, \\ g &\in \mathscr{A}^*, \end{aligned}$$

so $g = 0$ in $\bar{S}^-$. This implies [Co16] that

$$\alpha_{i0} = 0, \quad i = 1, 2, 3,$$

and

$$\sum_{i=1}^{3} \sum_{k=1}^{N} \alpha_{ik} (D^{\mathscr{A}})^{(i)}(x, x^{(k)}) = 0, \quad x \in S^-. \tag{5.15}$$

According to the asymptotic expansion of $D^{\mathscr{A}}$ in [Co16], the entries $D_{3j}^{\mathscr{A}}(x, x^{(k)})$ become infinite as $|x| \to \infty$. This contradicts (5.15) unless

$$\alpha_{ik} = 0, \quad i = 1, 2, 3, \ \ k = 1, 2, \ldots, N,$$

which confirms the linear independence of the set $\mathscr{G}$.

Next, let $q \in L^2(\partial S)$ be such that

$$\int_{\partial S} (\varphi^{(i0)})^{\mathsf{T}} q \, ds = \int_{\partial S} (f^{(i)})^{\mathsf{T}} q \, ds = 0, \quad i = 1, 2, 3, \tag{5.16}$$

$$\int_{\partial S} (\varphi^{(ik)})^{\mathsf{T}} q \, ds = 0, \quad i = 1, 2, 3, \ \ k = 1, 2, \ldots. \tag{5.17}$$

Using (5.14), we see that (5.17) is the same as

$$\int_{\partial S} (D^{\mathscr{A}})_{(i)}(x^{(k)}, y) q(y) \, ds(y) = 0, \quad i = 1, 2, 3, \ \ k = 1, \ 2, \ \ldots. \tag{5.18}$$

The rest of the argument in this proof makes use of the properties of layer potentials with an L^2-density, discussed in [Co16]. Thus, by (5.17) and (5.7), the single-layer potential

$$(Vq)(x) = (Vq)^{\mathscr{A}}(x) = \int_{\partial S} D^{\mathscr{A}}(x, \ y) q(y) \, ds(y), \quad x \in S^+ \cup S^-,$$

written in this form on the basis of (5.16), is continuous on ∂S_*. Given (5.18) and the fact that the points $x^{(k)}$, $k = 1, 2, \ldots$, are densely distributed on ∂S_*, we deduce that Vq is the unique solution of the Dirichlet problem

$$A(Vq) = 0 \quad \text{in } S_*^+,$$
$$Vq = 0 \quad \text{on } \partial S_*,$$

which implies that

$$Vq = 0 \quad \text{in } \bar{S}_*^+.$$

The analyticity of Vq in $S^+ \cup S^-$ now yields

$$Vq = 0 \quad \text{in } \bar{S}^+.$$

Hence,

$$\big(T(Vq)\big)^+ = 0,$$

or, equivalently [Co16],

$$\tfrac{1}{2}\, q(x) + \int_{\partial S} T(\partial_x) D(x, y) q(y)\, ds(y) = 0 \quad \text{for a.a. } x \in \partial S,$$

where the integral is understood as principal value. This leads to the conclusion [Co16] that $q \in C^{0,\alpha}(\partial S)$ and, therefore, it is continuous in $\mathbb{R}^2$ and satisfies

$$A(Vq) = 0 \quad \text{in } S^-,$$
$$Vq = 0 \quad \text{on } \partial S,$$
$$Vq \in \mathscr{A}.$$

By Theorem 1,

$$Vq = 0 \quad \text{in } \bar{S}^-,$$

so

$$(T(Vq))^- = 0 \quad \text{on } \partial S.$$

Combining the limiting values $\big(T(Vq)\big)^{\pm}(x)$ of $\big(T(Vq)\big)(x)$ as x approaches ∂S from within S^+ and S^- (see [Co16]), we conclude that

$$q = 0,$$

which shows that the set $\mathscr{G}$ is complete in the Hilbert space $L^2(\partial S)$. □

The elements of $\mathscr{G}$ are re-ordered as the sequence

$$\varphi^{(10)},\ \varphi^{(20)},\ \varphi^{(30)},\ \varphi^{(11)},\ \varphi^{(21)},\ \varphi^{(31)},\ \varphi^{(12)},\ \varphi^{(22)},\ \varphi^{(32)},\ \ldots$$

and re-indexed as

$$\mathscr{G} = \{\varphi^{(1)},\ \varphi^{(2)},\ \varphi^{(3)},\ \varphi^{(4)},\ \varphi^{(5)},\ \varphi^{(6)},\ \varphi^{(7)},\ \varphi^{(8)},\ \varphi^{(9)},\ \ldots\},$$

where $\varphi^{(j)}$ in the new sequence is the same as $\varphi^{(ik)}$ in the original one with

$$j = i + 3k,\ \ i = 1, 2, 3,\ \ k = 0, 1, 2, \ldots. \tag{5.19}$$

5.5 Computational Procedure

Since the $x^{(k)}$ are points in S^+, from (5.13) it follows that

$$\int_{\partial S} (D^{\mathscr{A}})_{(i)}(x^{(k)}, x)\psi(x)\, ds(x) = \int_{\partial S} P_{(i)}(x^{(k)}, x)\mathscr{D}(x)\, ds(x) + (Fc)_i(x^{(k)}). \tag{5.20}$$

Given that $\mathscr{G}$ is a complete set, we may consider an expansion of the form

$$\psi = \sum_{h=1}^{\infty} p_h \varphi^{(h)}.$$

Replacing (5.20) and truncating after $n = 3N + 3$ terms, where N is the number of points $x^{(k)}$ on ∂S_*, we obtain the approximate equality

$$\begin{aligned}
&\sum_{h=1}^{3N+3} p_h \int_{\partial S} (D^{\mathscr{A}})_{(i)}(x^{(k)}, x)\varphi^{(h)}(x)\, ds(x) \\
&\qquad = \sum_{h=1}^{3N+3} p_h \int_{\partial S} (\varphi^{(j)})^{\mathsf{T}}(x)\varphi^{(h)}(x)\, ds(x) \\
&\qquad = \int_{\partial S} P_{(i)}(x^{(k)}, x)\mathscr{D}(x)\, ds(x) + \sum_{l=1}^{3} c_l f_i^{(l)}(x^{(k)}), \\
&\qquad\qquad i = 1, 2, 3,\ \ k = 1, 2, \ldots, N,\ \ j = i + 3k = 4, 5, \ldots, 3N + 3,
\end{aligned} \tag{5.21}$$

which gives rise to the approximation

$$\psi^{(n)} = \sum_{h=1}^{n} p_h \varphi^{(h)}, \quad n = 3N + 3, \tag{5.22}$$

satisfying

$$\lim_{n\to\infty} \|\psi - \psi_n\| = 0. \tag{5.23}$$

Taking the re-indexing (5.19) into account and making the notation

$$M_{jh} = \int_{\partial S} (\varphi^{(j)})^{\mathsf{T}}(x)\varphi^{(h)}(x)\, ds(x),$$

$$\beta_j = \int_{\partial S} P_{(i)}(x^{(k)}, x)\mathscr{D}(x)\, ds(x),$$

$$\gamma_{jl} = f_i^{(l)}(x^{(k)}),$$

we re-write (5.21) in the form

$$\sum_{h=1}^{3N+3} M_{jh} p_h - \sum_{l=1}^{3} \gamma_{jl} c_l = \beta_j, \quad j = 4, 5, \ldots, 3N + 3. \tag{5.24}$$

This is a linear system of $3N$ equations in $3N + 6$ unknowns p_h and c_l. Consequently, we need six additional equations.

Three of these equations are constructed by choosing a point $\hat{x}$ in S^+, arbitrary but distinct from the $x^{(k)}$ and the origin. Then (5.13) yields the system

$$\sum_{h=1}^{3N+3} p_h \int_{\partial S} (D^{\mathscr{A}})_{(i)}(\hat{x}, x)\varphi^{(h)}(x)\, ds(x)$$

$$= \int_{\partial S} P_{(i)}(\hat{x}, x)\mathscr{D}(x)\, ds(x) + \sum_{l=1}^{3} c_l f_i^l(\hat{x}), \quad i = 1, 2, 3,$$

or, with the notation

$$\hat{M}_{ih} = \int_{\partial S} (D^{\mathscr{A}})_{(i)}(\hat{x}, x)\varphi^{(h)}(x)\, ds(x),$$

$$\hat{\beta}_i = \int_{\partial S} P_{(i)}(\hat{x}, x)\mathscr{D}(x)\, ds(x),$$

$$\hat{\gamma}_{il} = \sum_{l=1}^{3} f_i^l(\hat{x}),$$

the system

$$\sum_{h=1}^{n} \hat{M}_{ih} c_h - \sum_{l=1}^{3} \hat{\gamma}_{il} a_l = \hat{\beta}_i, \quad i = 1, 2, 3. \tag{5.25}$$

Since $\psi = Tu$ satisfies (5.11), it is reasonable to ask the approximation (5.22) to do the same. Obviously, this condition translates as

$$\sum_{h=1}^{3N+3} p_h \int_{\partial S} (\varphi^{(h)})^{\mathsf{T}} \varphi^{(i)}\, ds = 0, \quad i = 1, 2, 3. \tag{5.26}$$

Equations (5.24), (5.25), and (5.26) form a non-singular system for the computation of the $3N + 3$ coefficients p_h and the three coefficients c_l, which determine $\psi^{(n)}$ and the rigid displacement approximation

$$(Fc)^{(3N+3)} = Fc^{(3N+3)}.$$

Representation formula (5.12) indicates that we should now define

$$(u^{\mathscr{A}})^{(n)}(x) = -\int_{\partial S} D^{\mathscr{A}}(x, y)\psi^{(n)}(y)\, ds(y) + \int_{\partial S} P(x, y)\mathscr{D}(y)\, ds(y), \quad x \in S^-, \tag{5.27}$$

and

$$u^{(n)}(x) = (u^{\mathscr{A}})^{(n)}(x) + F(x)c^{(n)}, \quad x \in S^-. \tag{5.28}$$

In [Co16], the exact rigid displacement Fc component of u is given by an expression involving $\mathscr{D}$ and the null space properties of a specific boundary integral operator. This suggests that we may consider an alternative definition of $u^{(n)}$, namely

$$u^{(n)}(x) = \left(u^{\mathscr{A}}\right)^{(n)}(x) + (Fc)(x), \quad x \in S^-. \tag{5.29}$$

Equality (5.29) is the form of choice in analytic arguments, and (5.28) in numerical computation.

Theorem 4 *The vector function $u^{(n)}$ defined by (5.29) is an approximation of the solution u of problem (5.1) in the sense that $u^{(n)} \to u$ uniformly on any closed and bounded subdomain S' of S^-.*

Proof By (5.9), (5.29), (5.12), and (5.27),

$$\begin{aligned} u(x) - u^{(n)}(x) &= u^{\mathscr{A}}(x) - \left(u^{\mathscr{A}}\right)^{(n)}(x) \\ &= -\int_{\partial S} D^{\mathscr{A}}(x, y)\psi(y)\, ds(y) + \int_{\partial S} D^{\mathscr{A}}(x, y)\psi^{(n)}(y)\, ds(y) \\ &= -\int_{\partial S} D^{\mathscr{A}}(x, y)\left[\psi(y) - \psi^{(n)}(y)\right] ds(y), \quad x \in S^-, \end{aligned}$$

so

$$|u(x) - u^{(n)}(x)| \le \sum_{i=1}^{3} \|(D^{\mathscr{A}})_{(i)}(x, \cdot)\|\, \|\psi - \psi^{(n)}\|, \quad x \in S'.$$

Since S' is closed and bounded, the $(D^{\mathscr{A}})_{(i)}$ are uniformly bounded on it [Co16], and the statement of the assertion now follows from (5.23). □

The numerical procedure described above is referred to as the row reduction method.

5.6 Numerical and Graphical Illustration: Known Solution

Let S^+ be the disk of radius 1 centered at origin, let ∂S_* be the circle concentric with ∂S and of radius $1/2$, and let the physical parameters of the elastic plate material, after suitable rescaling and non-dimensionalization, be

$$h = 0.5, \quad \lambda = \mu = 1.$$

Remark 1 Our choice for ∂S_* is motivated by the fact that if this curve is too far away from ∂S, then the sequence $\mathscr{G}$ becomes "less linearly independent". On the other hand, if it is too close to ∂S, then $\mathscr{G}$ develops increasing sensitivity to the singularities of matrices D and P on the boundary.

Remark 2 The approximation accuracy depends on the selection of the points $\{x^{(k)}\}$. In the interest of computational symmetry, we decided to space these points uniformly around ∂S_*; specifically, for $N = 1, 2, \ldots,$

$$\{x^{(k)} : k = 1, 2, \ldots, N\}_{\text{Cartesian}} = \left\{ \left(\tfrac{1}{2}, \frac{2\pi k}{N} \right) : k = 1, 2, \ldots, N \right\}_{\text{Polar}}.$$

It is clear that $\{x^{(k)}\}_{k=1}^{\infty}$ is the set of all points on ∂S_* whose polar angle is of the form $2\pi a$, where a is any rational number such that $0 < a \leq 1$.

Remark 3 We have performed floating-point computation with machine precision of approximately 16 digits. The most sensitive part of the process is the evaluation of integrals in the inner products, for which we set a target of 11 significant digits.

We prescribe the boundary condition (in polar coordinates on ∂S)

$$\mathscr{D}(x_p) = \begin{pmatrix} 2 + 2\cos\theta + 3\sin\theta - \sin(2\theta) + 2\cos(3\theta) - 7\sin(3\theta) \\ \qquad + 2\cos(4\theta) + 4\sin(4\theta) \\ -4 + 3\cos\theta + 2\sin\theta - 3\cos(2\theta) + 2\sin(2\theta) + 7\cos(3\theta) \\ \qquad + 2\sin(3\theta) - 4\cos(4\theta) + 2\sin(4\theta) \\ -1 - \cos\theta + 5\sin\theta + \cos(2\theta) - 14\sin(2\theta) + 7\cos(3\theta) \\ \qquad + 14\sin(3\theta) \end{pmatrix},$$

which generates the exact solution (in S^-)

$$u(x) = \begin{pmatrix} 2 + (x_1^2 + x_2^2)^{-1}(2x_1 + 3x_2) \\ \quad + (x_1^2 + x_2^2)^{-2}(-2x_1x_2 - 9x_1^2x_2 + 3x_2^3) \\ \quad\quad + (x_1^2 + x_2^2)^{-3}(2x_1^3 - 12x_1^2x_2 - 6x_1x_2^2 + 4x_2^3 + 2x_1^4 \\ \quad\quad\quad + 16x_1^3x_2 - 12x_1^2x_2^2 - 16x_1x_2^3 + 2x_2^4) \\ -4 + (x_1^2 + x_2^2)^{-1}(3x_1 + 2x_2) \\ \quad + (x_1^2 + x_2^2)^{-2}(-3x_1^2 + 4x_1x_2 + 3x_2^2 + 3x_1^3 - 9x_1x_2^2) \\ \quad\quad + (x_1^2 + x_2^2)^{-3}(4x_1^3 + 6x_1^2x_2 - 12x_1x_2^2 - 2x_2^3 - 4x_1^4 \\ \quad\quad\quad + 8x_1^3x_2 + 24x_1^2x_2^2 - 8x_1x_2^3 - 4x_2^4) \\ -1 - \ln(x_1^2 + x_2^2) - 2x_1 + 4x_2 + (x_1^2 + x_2^2)^{-1}(x_1 + x_2 - 6x_1x_2) \\ \quad + (x_1^2 + x_2^2)^{-2}(x_1^2 - 22x_1x_2 - x_2^2 + x_1^3 + 6x_1^2x_2 - 3x_1x_2^2 - 2x_2^3) \\ \quad\quad + (x_1^2 + x_2^2)^{-3}(6x_1^3 + 36x_1^2x_2 - 18x_1x_2^2 - 12x_2^3) \end{pmatrix},$$

or, in polar coordinates,

$$u(x_p) = \begin{pmatrix} 2 + r^{-1}[2\cos\theta + 3\sin\theta - 3\sin(3\theta)] \\ \quad + r^{-2}[-\sin(2\theta) + 2\cos(4\theta) + 4\sin(4\theta)] \\ \quad\quad + r^{-3}[2\cos(3\theta) - 4\sin(3\theta)] \\ -4 + r^{-1}[3\cos\theta + 2\sin\theta + 3\cos(3\theta)] \\ \quad + r^{-2}[-3\cos(2\theta) + 2\sin(2\theta) - 4\cos(4\theta) + 2\sin(4\theta)] \\ \quad\quad + r^{-3}[4\cos(3\theta) + 2\sin(3\theta)] \\ r(-2\cos\theta + 4\sin\theta) - 2\ln r - 1 \\ \quad + r^{-1}[\cos\theta + \sin\theta + \cos(3\theta) + 2\sin(3\theta)] \\ \quad\quad + r^{-2}[\cos(2\theta) - 11\sin(2\theta)] \\ \quad\quad\quad + r^{-3}[6\cos(3\theta) + 12\sin(3\theta)] \end{pmatrix}.$$

This solution contains the class $\mathscr{A}$ vertical translation $(0, 0, -2)^{\mathsf{T}}$ and the additional rigid displacement

$$f(x) = (2f^{(1)} - 4f^{(2)} + f^{(3)})(x) = (2, -4, 1 - 2x_1 + 4x_2)^{\mathsf{T}}.$$

If f is eliminated, then

$$(\mathscr{D} - f)(x_p) = \begin{pmatrix} 2\cos\theta + 3\sin\theta - \sin(2\theta) + 2\cos(3\theta) - 7\sin(3\theta) \\ \quad + 2\cos(4\theta) + 4\sin(4\theta) \\ 3\cos\theta + 2\sin\theta - 3\cos(2\theta) + 2\sin(2\theta) + 7\cos(3\theta) \\ \quad + 2\sin(3\theta) - 4\cos(4\theta) + 2\sin(4\theta) \\ -2 + \cos\theta + \sin\theta + \cos(2\theta) - 14\sin(2\theta) + 7\cos(3\theta) \\ \quad + 14\sin(3\theta) \end{pmatrix}$$

and

$$u^{\mathscr{A}}(x_p) = \begin{pmatrix} r^{-1}[2\cos\theta + 3\sin\theta - 3\sin(3\theta)] \\ \quad + r^{-2}[-\sin(2\theta) + 2\cos(4\theta) + 4\sin(4\theta)] \\ \quad\quad + r^{-3}[2\cos(3\theta) - 4\sin(3\theta)] \\ r^{-1}[3\cos\theta + 2\sin\theta + 3\cos(3\theta)] \\ \quad + r^{-2}[-3\cos(2\theta) + 2\sin(2\theta) - 4\cos(4\theta) + 2\sin(4\theta)] \\ \quad\quad + r^{-3}[4\cos(3\theta) + 2\sin(3\theta)] \\ -2\ln r - 2 + r^{-1}[\cos\theta + \sin\theta + \cos(3\theta) + 2\sin(3\theta)] \\ \quad\quad + r^{-2}[\cos(2\theta) - 11\sin(2\theta)] \\ \quad\quad\quad + r^{-3}[6\cos(3\theta) + 12\sin(3\theta)] \end{pmatrix}.$$

Using the polar form of the solution, we can verify that $u^{\mathscr{A}} = u - Fc$ fits the far-field pattern (5.4) with the coefficients

$$m_0 = 3, \quad m_1 = 1, \quad m_2 = 1, \quad m_3 = -2, \quad m_4 = 3, \quad m_5 = 0, \quad m_6 = 1,$$
$$m_7 = -2, \quad m_8 = 1, \quad m_9 = -2, \quad m_{10} = 1, \quad m_{11} = -11, \quad m_{12} = 1.$$

The entries of the approximation $(u^{\mathscr{A}})^{(57)}$ computed from $\psi^{(57)}$ (with $N = 18$ points $x^{(k)}$ on ∂S_*) for $r \geq 1.01$, $0 \leq \theta < 2\pi$, and those of $\mathscr{D} - f$ on ∂S are graphed in Fig. 5.1.

We took $r \geq 1.01$ to avoid the influence of the singularities of $D(x, y)$ and $P(x, y)$ for $x \in S^-$ very close to $y \in \partial S$. This problem can be mitigated by increasing the floating-point accuracy in the vicinity of the boundary, but can never be completely eliminated.

The graphs of the entries of $(u^{\mathscr{A}})^{(57)}$ computed from $\psi^{(57)}$ for $1.01 < r \leq 100$, $0 \leq \theta < 2\pi$, in Fig. 5.2, illustrate the class $\mathscr{A}$ behavior of the solution away from the boundary. These graphs have been truncated for better visualization.

Figure 5.3 exhibits the graphs of the entries of the actual error $(u^{\mathscr{A}})^{(57)} - u^{\mathscr{A}}$. Their approximation is 3–5 digits of accuracy near ∂S, but improves significantly away from the boundary.

Figure 5.4 shows the boundary data function $\mathscr{D}$ in polar coordinates.

The components of a typical vector function $\varphi^{(i)}$ are graphed (in polar coordinates) in Fig. 5.5.

The entries of $\psi^{(57)}$ as functions of the polar angle θ are displayed in Fig. 5.6.

Figure 5.7 shows the graphs of the error $\psi^{(57)} - Tu$ in terms of θ.

Finally, the relative L^2-error

$$\frac{\|\psi^{(3N+3)} - Tu\|}{\|Tu\|}$$

as a function of N is shown with the best least square linear fit in Fig. 5.8, where the vertical axis is displayed logarithmically to base 10. This plot indicates that the relative error decreases exponentially as N increases. For this example, the relative error is

$$1488.36 \times 10^{-0.301277\,N}.$$

Figure 5.9 exhibits the relative error in the computation of the vector c that characterizes the rigid displacement in the solution.

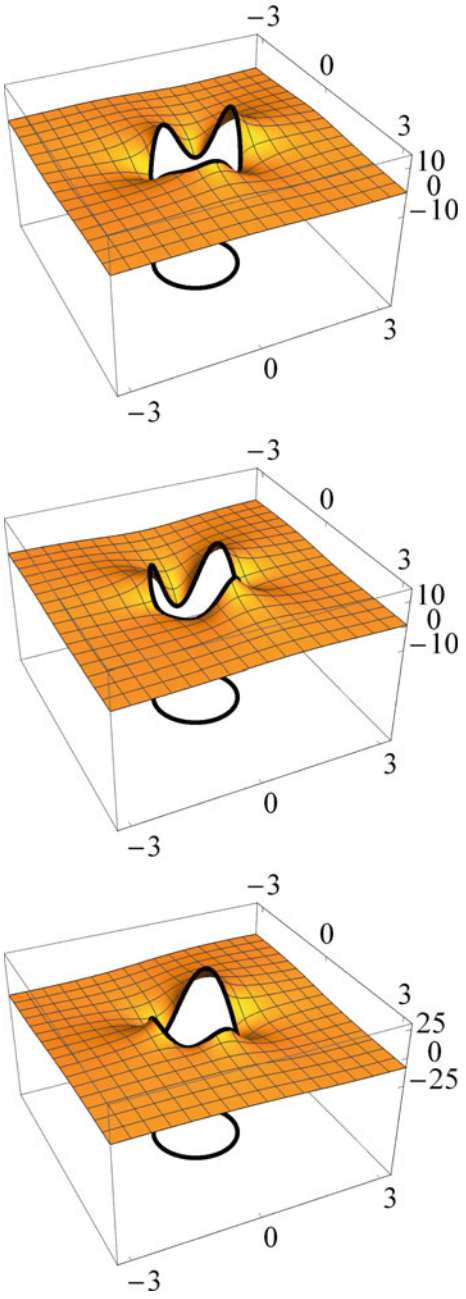


Fig. 5.1 Components of $(u^{\mathscr{A}})^{(57)}$ and $\mathscr{D} - f$

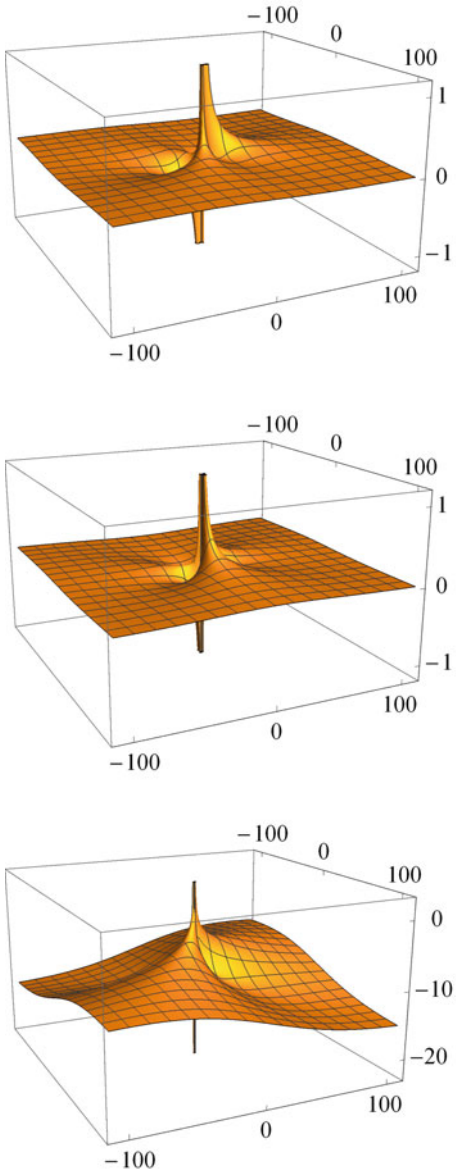


Fig. 5.2 Components of $(u^{\mathscr{A}})^{(57)}$ away from the origin

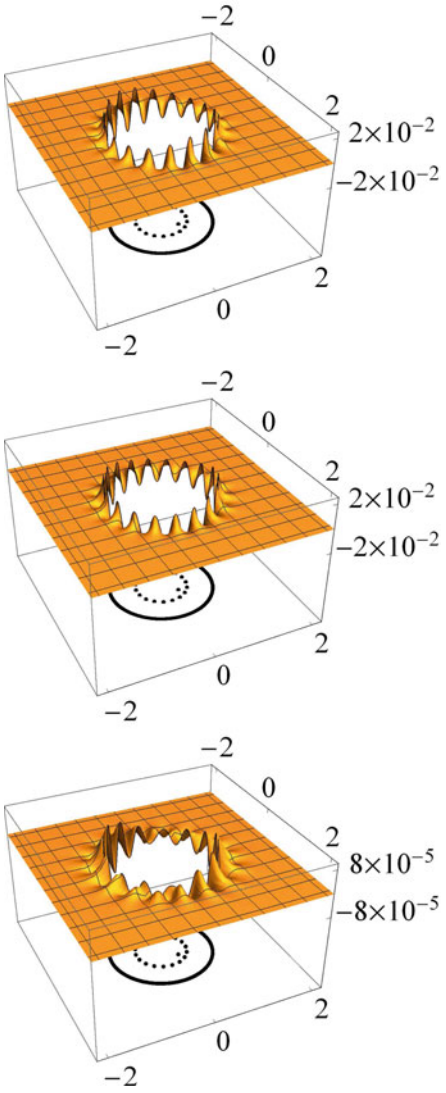


Fig. 5.3 Components of the error $(u^{\mathscr{A}})^{(57)} - u^{\mathscr{A}}$

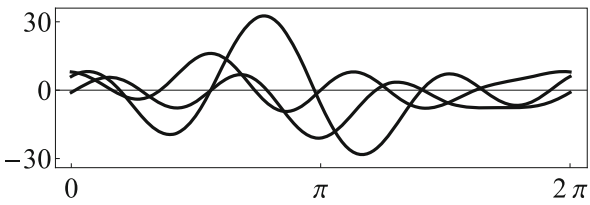


Fig. 5.4 Components of $\mathscr{D}$ in polar coordinates

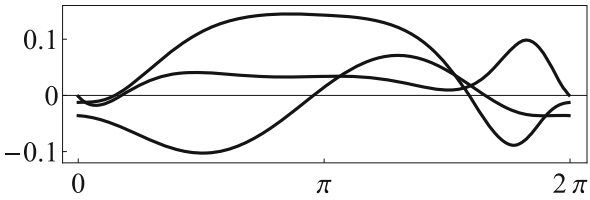


Fig. 5.5 Components of a typical $\varphi^{(i)}$ in polar coordinates

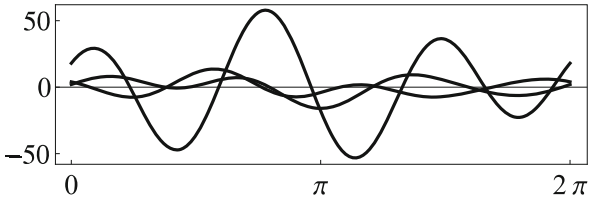


Fig. 5.6 Components of $\psi^{(57)}$ in polar coordinates

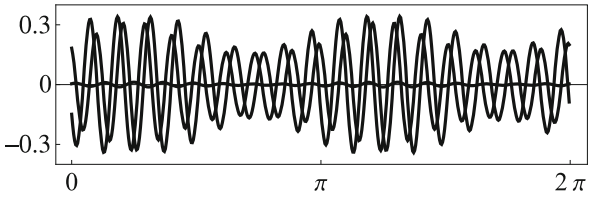


Fig. 5.7 Components of $\psi^{(57)} - Tu$ in polar coordinates

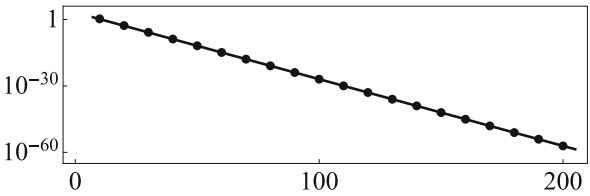


Fig. 5.8 Relative error $\|\psi^{(3N+3)} - Tu\|/\|Tu\|$ as a function of N

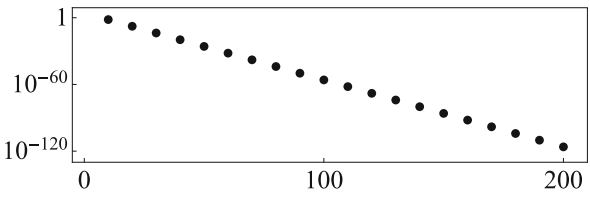


Fig. 5.9 Relative error $\|c^{(3N+3)} - c\|/\|c\|$ as a function of N

5.7 Numerical and Graphical Illustration: Unknown Solution

With the same choice of auxiliary curve and plate geometric and physical parameters as in Sect. 5.6, we consider problem (5.8) with the boundary condition vector function

$$\mathscr{D}(x_p) = \begin{pmatrix} 1.73 + \cos\theta - 4\sin\theta - 7\cos(3\theta) \\ \qquad + 14\sin(3\theta) - 4\sin(4\theta) \\ 2.51 - 4\cos\theta - \sin\theta + 4\cos(2\theta) \\ \qquad - 14\cos(3\theta) - 7\sin(3\theta) + 4\cos(4\theta) \\ -0.43 - 1.73\cos\theta - 4.51\sin\theta \\ \qquad - 7\cos(2\theta) + 21\sin(2\theta) - 14\sin(3\theta) \end{pmatrix}.$$

The exact solution of the problem is not known in this case.

The comments made in Remark 1 remain valid here as well.

Applying the row reduction method described in Sect. 5.5, we find that the approximation of the rigid displacement computed with $N = 20$ points on ∂S_* is

$$f^{(63)} = 1.73205\, f^{(1)} + 2.50999\, f^{(2)} - 0.430001\, f^{(3)}. \tag{5.30}$$

Our procedure requires $f^{(63)}$ to be subtracted from $\mathscr{D}$ when we approximate $u^{\mathscr{A}}$.

Figure 5.10 shows the graphs of the components of $(u^{\mathscr{A}})^{(63)}$ obtained from $\psi^{(63)}$ for $r \geq 1.01$, $0 \leq \theta < 2\pi$, those of the components of $\mathscr{D} - f^{(63)}$, the points $x^{(k)}$, and the auxiliary point $\hat{x} = (1/4, 1/4)$ used in the calculation.

In Fig. 5.11, we have plotted the components of $(u^{\mathscr{A}})^{(63)}$ for $1.01 \leq r \leq 100$ and $0 \leq \theta \leq 2\pi$, to illustrate the class $\mathscr{A}$ behavior of the approximation of the solution away from the boundary. The graphs have been truncated for better visualization.

The graphs (in polar coordinates) of the entries of $\mathscr{D}$ are shown in Fig. 5.12.

Figure 5.13 exhibits the graphs (in polar coordinates) of the approximation $\psi^{(63)}$.

Since the exact solution is not known in this example, we cannot perform a proper error analysis. Instead, we try to validate our method indirectly by using $\psi^{(63)}$ as the boundary condition for a Neumann problem and computing the approximate boundary trace $\hat{\mathscr{D}}^{(63)}$ of the latter. Adding $f^{(63)}$ from (5.30) to $\hat{\mathscr{D}}^{(63)}$, we expect to get close to $\mathscr{D}$. The sum of these two functions and the difference between that sum and $\mathscr{D}$ are shown in Figs. 5.14 and 5.15, respectively. The plots in Fig. 5.15 confirm the robust nature of the generalized Fourier series method. This conclusion is further strengthened by the fact that, when applying the scheme with $N = 200$, we obtain the much smaller relative error

$$\frac{\|\hat{\mathscr{D}}^{(603)} + f^{(603)} - \mathscr{D}\|}{\|\mathscr{D}\|} = 2.11107 \times 10^{-57}.$$

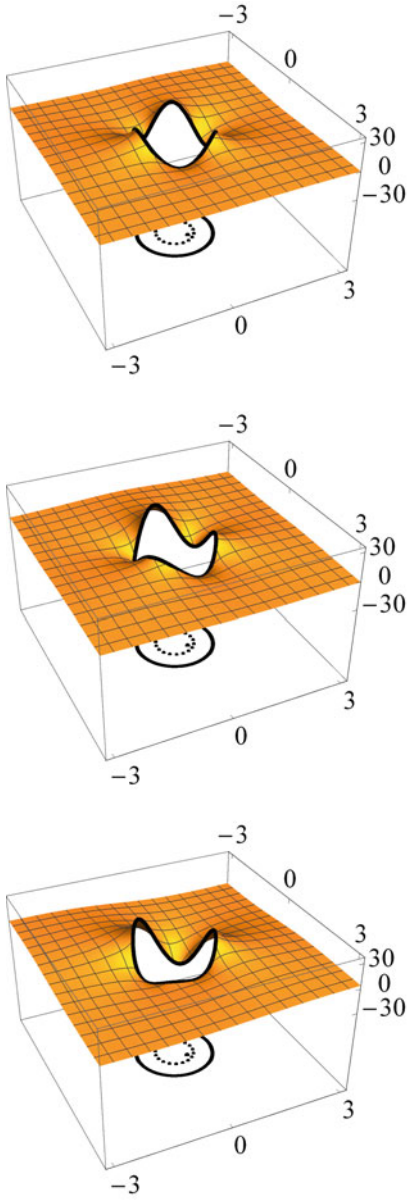


Fig. 5.10 Components of $(u^{\mathscr{A}})^{(63)}$ and $\mathscr{D} - f^{(63)}$

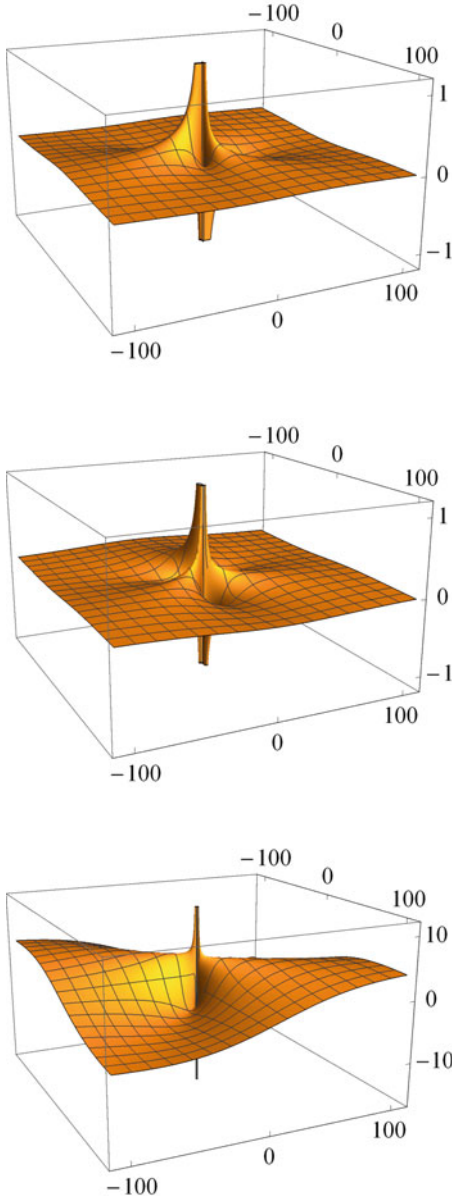


Fig. 5.11 Components of $(u^{\mathscr{A}})^{(63)}$ away from the origin

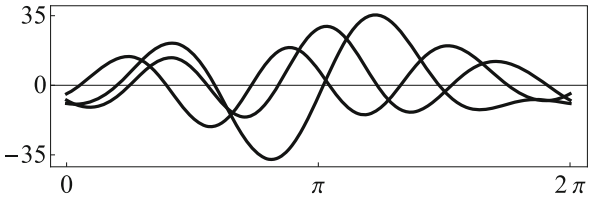


Fig. 5.12 Components of $\mathscr{D}$ in polar coordinates

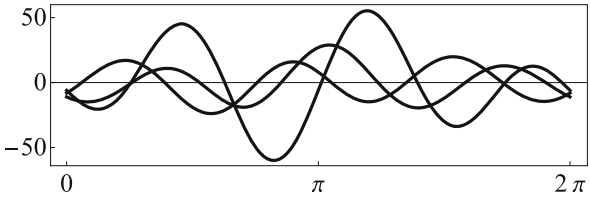


Fig. 5.13 Components of $\psi^{(63)}$ in polar coordinates

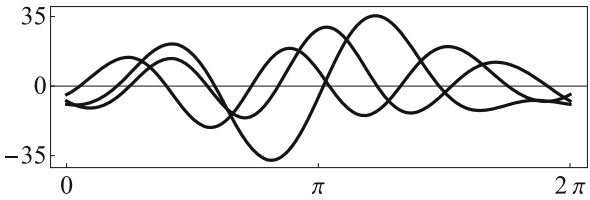


Fig. 5.14 Components of $\hat{\mathscr{D}}^{(63)} + f^{(63)}$ in polar coordinates

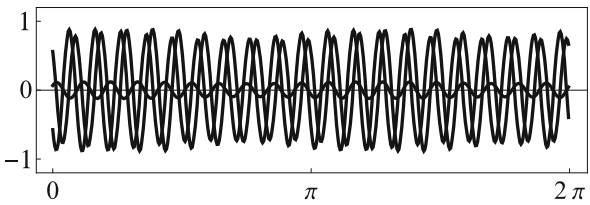


Fig. 5.15 Components of $\hat{\mathscr{D}}^{(63)} + f^{(63)} - \mathscr{D}$ in polar coordinates

References

[Co16] Constanda, C.: Mathematical Methods for Elastic Plates. Springer, London (2016)

[CoDo17a] Constanda, C., Doty, D.: Bending of elastic plates: generalized Fourier series method. In: Integral Methods in Science and Engineering: Theoretical Techniques, pp. 71–81. Birkhäuser, New York (2017)

[CoDo17b] Constanda, C., Doty, D.: The Neumann problem for bending of elastic plates. In: Proceedings of the 17th International Conference on Computational and Mathematical Methods in Science and Engineering CMMSE 2017, Cádiz, Spain, vol. II, pp. 619–622 (2017)

[CoDo18] Constanda, C., Doty, D.: Bending of elastic plates with transverse shear deformation: the Neumann problem. Math. Methods Appl. Sci. **41**, 7130–7140 (2018). https://doi.org/10.1002/mma.4704

[CoDo19a] Constanda, C., Doty, D.: The Robin problem for bending of elastic plates. Math. Methods Appl. Sci. **42**, 5639–5648 (2019). https://doi.org/10.1002/mma.5286

[CoDo19b] Constanda, C., Doty, D.: Bending of plates with transverse shear deformation: the Robin problem. Comput. Math. Methods **1**, e1015 (2019). https://doi.org/10.1002/cmm4.1015

[CoDo19c] Constanda, C., Doty, D.: Bending of elastic plates: generalized Fourier series method for the Robin problem. In: Integral Methods in Science and Engineering: Analytic Treatment and Numerical Approximations. Birkhäuser, New York, pp. 97–110 (2019)

[CoDo20] Constanda, C., Doty, D.: Analytic and numerical solutions in the theory of elastic plates. Complex Var. Elliptic Equ. **65**, 40–56 (2020). https://doi.org/10.1080/17476933.20191636789

Chapter 6
A Dirichlet Spectral Problem in Domains Surrounded by Thin Stiff and Heavy Bands

Delfina Gómez, Sergey A. Nazarov, and Maria–Eugenia Pérez-Martínez

6.1 Introduction and Statement of the Problem

Let Ω be a bounded simply connected domain of the plane $\mathbb{R}^2$ with a smooth boundary Γ and let (ν, τ) be the natural orthogonal curvilinear coordinates in a neighborhood of Γ: τ is the arc length and ν the distance along the normal vector to Γ; $\nu < 0$ inside Ω. Let ℓ denote the length of the contour Γ and $\varkappa(\tau)$ its curvature at the point τ. We assume that the domain Ω is surrounded by the thin band $\omega_\varepsilon = \{x : 0 < \nu < \varepsilon h\}$ where $\varepsilon > 0$ is a small parameter and h is a positive constant. Let Ω_ε be the domain $\Omega_\varepsilon = \Omega \cup \omega_\varepsilon \cup \Gamma$ and $\Gamma_\varepsilon = \{x : \nu = \varepsilon h\}$ the boundary of Ω_ε (see Fig. 6.1).

We consider the spectral Dirichlet problem in Ω_ε for a second order differential operator with piecewise constant coefficients:

$$-A\Delta_x U^\varepsilon = \lambda^\varepsilon U^\varepsilon \qquad \text{in } \Omega, \tag{6.1a}$$

$$-a\varepsilon^{-t}\Delta_x u^\varepsilon = \lambda^\varepsilon \varepsilon^{-t-m} u^\varepsilon \qquad \text{in } \omega_\varepsilon, \tag{6.1b}$$

$$U^\varepsilon = u^\varepsilon \qquad \text{on } \Gamma, \tag{6.1c}$$

$$\varepsilon^t A\partial_\nu U^\varepsilon = a\partial_\nu u^\varepsilon \qquad \text{on } \Gamma, \tag{6.1d}$$

$$u^\varepsilon = 0 \qquad \text{on } \Gamma_\varepsilon. \tag{6.1e}$$

D. Gómez · M.-E. Pérez-Martínez (✉)
Universidad de Cantabria, Santander, Spain
e-mail: gomezdel@unican.es; meperez@unican.es

S. A. Nazarov
Saint-Petersburg State University, St. Petersburg, Russia

Institute of Problems of Mechanical Engineering RAS, St. Petersburg, Russia
e-mail: srgnazarov@yahoo.co.uk

C. Constanda (ed.), *Computational and Analytic Methods in Science and Engineering*, https://doi.org/10.1007/978-3-030-48186-5_6

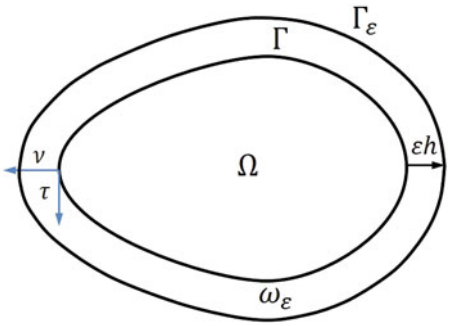


Fig. 6.1 Possible geometry of Ω_ε

Here, A and a are two positive constants, while ∂_ν denotes the derivative along the outward normal vectors ν to the curve Γ; t and m are two positive parameters.

The weak formulation of problem (6.1) reads: to find λ^ε and $\{U^\varepsilon, u^\varepsilon\} \in H_0^1(\Omega_\varepsilon) \setminus \{\{0,0\}\}$, satisfying

$$A\int_\Omega \nabla_x U^\varepsilon \cdot \nabla_x G\, dx + \frac{a}{\varepsilon^t}\int_{\omega_\varepsilon} \nabla_x u^\varepsilon \cdot \nabla_x g\, dx$$
$$= \lambda^\varepsilon \left(\int_\Omega U^\varepsilon\, G\, dx + \frac{1}{\varepsilon^{t+m}} \int_{\omega_\varepsilon} u^\varepsilon\, g\, dx \right) \quad \forall \{G, g\} \in H_0^1(\Omega_\varepsilon). \tag{6.2}$$

Here, and in what follows, we identify a function in $L^2(\Omega_\varepsilon)$ with the pair of functions $\{G, g\}$, where G stands for the restriction of the function to Ω and g for the restriction of the function to ω_ε. In particular, the eigenpairs formed by the eigenvalues λ^ε and the corresponding eigenfunctions read $(\lambda^\varepsilon, \{U^\varepsilon, u^\varepsilon\})$.

For each $\varepsilon > 0$, problem (6.2) is a standard spectral problem in the couple of spaces $H_0^1(\Omega_\varepsilon) \subset L^2(\Omega_\varepsilon)$, with a positive and discrete spectrum. Let us consider

$$0 < \lambda_1^\varepsilon \le \lambda_2^\varepsilon \le \cdots \le \lambda_k^\varepsilon \le \cdots \xrightarrow{k\to\infty} \infty \tag{6.3}$$

the sequence of eigenvalues repeated according to their multiplicity.

For each fixed $k \in \mathbb{N}$ and a small ε, we have

$$\begin{aligned} C \le \lambda_k^\varepsilon \le C_k \qquad & \text{when } m \le 2, \\ C\varepsilon^{m-2} \le \lambda_k^\varepsilon \le C_k \varepsilon^{m-2} \quad & \text{when } m > 2, \end{aligned} \tag{6.4}$$

where the positive constants C and C_k do not depend on ε, but $C_k \to \infty$ as $k \to \infty$ (see [GoEtAl20] for the proof based on the minimax principle and the Poincaré inequality). Relations in (6.4) indicate the order of magnitude of the eigenvalues of problem (6.2) for fixed k, the so-called *low frequencies*.

We study the asymptotic behavior, as $\varepsilon \to 0$, of the eigenvalues λ^ε of (6.2) and the corresponding eigenfunctions $\{U^\varepsilon, u^\varepsilon\}$. This problem is of interest, for instance, in the study of reinforcement problems for solid media and in vibrations for a two-phase system in fluid mechanics. Here, the band ω_ε is both stiffer and heavier. Parameters t and m deal with the physical characteristic of the medium and it seems natural to have a different asymptotic behavior as $\varepsilon \to 0$ for the eigenpairs $(\lambda^\varepsilon, \{U^\varepsilon, u^\varepsilon\})$ of (6.2) depending on their value.

This problem has been considered for the first time in [GoEtAl20] where the low and high frequencies are studied when $t \geq 1$ and $m > 2$ which are of order ε^{m-2} and 1, respectively. There localization effects for the eigenfunctions corresponding to low frequencies of (6.2) are shown around points τ_0 of the boundary where the curvature of Γ has a local extremum. These localization effects differ strongly in previous papers (see below).

The aim of this paper is to study, for $t = 1$ and $m > 2$, the asymptotic behavior, as $\varepsilon \to 0$, of the eigenvalues λ^ε of (6.2) and the corresponding eigenfunctions $\{U^\varepsilon, u^\varepsilon\}$ in case the curvature of Γ is constant; that is, when the domain Ω is a disk. Here, explicit computations for the eigenpairs of (6.1) can be done by means of the Bessel functions and the eigenvalues of (6.1) are the roots of certain transcendental equation (cf. Sect. 6.2). In order to describe their asymptotic behavior as $\varepsilon \to 0$, we also provide the asymptotic expansions for the low and high frequencies when $t = 1$ and $m > 2$. In contrast with the case where the curvature is not constant (considered in [GoEtAl20]), the corresponding eigenfunctions are now significant over the whole domain Ω_ε, and no localization effects arise at points of the boundary.

Let us recall the results in [GoEtAl06a, GoEtAl06b, GoEtAl11] which are close papers to the problem under consideration. There, the spectral Neumann problem is considered for different values of t and m; that is, (6.1a)–(6.1d) and

$$\partial_\nu u^\varepsilon = 0 \quad \text{on } \Gamma_\varepsilon;$$

also it is assumed that $\omega_\varepsilon = \{x : 0 > \nu > -\varepsilon h(\tau)\}$ where h is a strictly positive function of the τ variable ℓ-periodic, $h \in C^\infty(\mathbb{S}_\ell)$ where $\mathbb{S}_\ell$ stands for the circle of length ℓ. Note that ω_ε may vary with the arc length. A characterization of the limiting problems for the eigenpairs of the Neumann problem for the different values of t and m has been obtained in [GoEtAl06a] by means of asymptotic expansions. Sharp bounds for convergence rates of the eigenpairs $(\lambda^\varepsilon, \{U^\varepsilon, u^\varepsilon\})$ in the case where $t = 1$ and $m = 0$ are also given by using the so-called *inverse-direct reduction method* (cf. [Na02, Na03, LoEtAl05]). A different approach for the eigenpairs is provided in [GoEtAl06b] for the case where $t > 1$ and $m = 0$ where, in addition to the convergence, a complete asymptotic expansion for the eigenpairs has been obtained, and a connection of this problem with Wentzell problems with small parameters has been shown. Also, both papers [GoEtAl06a, GoEtAl06b] describe precise bounds for convergence rates for the low frequencies and the corresponding eigenfunctions in the cases mentioned above $m = 0$ and $t \geq 1$. We refer to [GoEtAl06a, GoEtAl06b] for further references.

Paper [GoEtAl11] deals with the Neumann problem in the case where $t = 1$ and $m > 0$, and considers the low and high frequencies which are now of order ε^m and 1, respectively. The limiting problems associated with both kinds of frequencies are obtained and information on the structure of the corresponding eigenfunctions is also provided. These problems appear independently of the geometry of the band ω_ε, but for $m > 2$ there are other limiting problems associated with the intermediate frequencies, namely eigenvalues of order ε^{m-2}, which strongly depend on this geometry: more precisely whether the function h is constant or not. Moreover, only in the case where h is not constant, the eigenfunctions corresponding to the middle frequencies are localized asymptotically in small neighborhoods of points τ_0 of the boundary where the function h presents a local maximum.

It should be pointed out that this paper contains two very different parts. These parts are in Sects. 6.2 and 6.3–6.5, respectively. In Sect. 6.2, for fixed ε, we obtain explicit formulas for the eigenelements of (6.1) when the domain Ω is a disk. In this case, using separation of variables and the Bessel functions, we obtain some explicit formulas for the eigenfunctions of (6.1) (cf. (6.11), (6.12), (6.14), (6.15)), whereas the eigenvalues of (6.1) are described by means of the roots of a certain equation (cf. (6.9)). This is valid for all t and m. But it does not provide much information on the asymptotic behavior of the eigenvalues and the eigenfunctions when $\varepsilon \to 0$.

In Sects. 6.3–6.5, by means of asymptotic expansions, we describe the behavior, as $\varepsilon \to 0$, of the eigenvalues of (6.1) and the corresponding eigenfunctions $\{U^\varepsilon, u^\varepsilon\}$. Since the curvature of Γ is constant, these asymptotics are not provided in [GoEtAl20] and here we complement the results in [GoEtAl20]. We consider the low and high frequencies in the case where $t = 1$ and $m > 2$ which are of order ε^{m-2} and 1, respectively. More precisely, in Sect. 6.3, we construct three-term asymptotic expansions of eigenvalues of (6.1) of order ε^{m-2},

$$\lambda^\varepsilon = \varepsilon^{m-2}(\lambda_0 + \varepsilon\lambda_1 + \varepsilon^2\lambda_2 + o(\varepsilon^2)),$$

and for the corresponding eigenfunctions, while the justification for these asymptotic expansions is given in Sect. 6.4. In particular, we provide estimates for the convergence rate of the eigenvalues of (6.1) of order $O(\varepsilon^{m-2})$ and also of the corresponding eigenfunctions as stated in Theorem 1. Finally, the eigenvalues of order 1, that is, the high frequencies, are considered in Sect. 6.5.

6.2 Some Explicit Computations

When the domain Ω is a disk, explicit computations for the eigenpairs of (6.1) can be performed by means of the Bessel functions. In this section, for $\varepsilon > 0$ fixed, we obtain some explicit formulas for the eigenfunctions of (6.1) (cf. (6.11), (6.12), (6.14), (6.15)), whereas the eigenvalues of (6.1) are described by means of roots of a certain transcendental equation (cf. (6.9)). We refer to [GoEtAl19] for some explicit computations for a stiff problem posed in a rectangle composed of

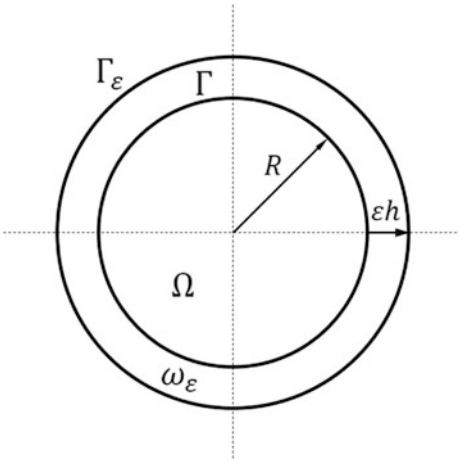


Fig. 6.2 Geometry for Ω_ε when the curvature of Γ is constant

two parts in which the stiffness constants are of different orders of magnitude. For further dimensions and explicit computations, we also refer to [Pe95], [LoPe97], and [LoEtAl03] for a stiff problem, [GoEtAl98], [GoEtAl99a], [GoEtAl99b], and [CaEtAl05] for vibrating systems with concentrated masses, and [Pe03] dealing with whispering gallery eigenmodes along interfaces.

We assume that the domain Ω is the disk with radius $R > 0$ centered at the origin, namely $\Omega = B(O, R) = \{x : \|x\| < R\}$. Thus, Γ and Γ_ε are the circles centered at the origin with radii R and $R + \varepsilon h$, respectively, and the band is the annulus $\omega_\varepsilon = \{x : R < \|x\| < R + \varepsilon h\}$ (see Fig. 6.2). To simplify, we also assume that the constant A and a in (6.1) are equal to 1.

We introduce the polar coordinates $x_1 = r\cos\theta$, $x_2 = r\sin\theta$ with $r \in [0, R + \varepsilon h)$ and $\theta \in [0, 2\pi)$ in problem (6.1) and we get

$$\begin{cases} -\partial_r^2 U^\varepsilon - r^{-1}\partial_r U^\varepsilon - r^{-2}\partial_\theta^2 U^\varepsilon = \lambda^\varepsilon U^\varepsilon & \text{for } (r,\theta) \in [0, R) \times [0, 2\pi), \\ -\partial_r^2 u^\varepsilon - r^{-1}\partial_r u^\varepsilon - r^{-2}\partial_\theta^2 u^\varepsilon = \lambda^\varepsilon \varepsilon^{-m} u^\varepsilon & \text{for } (r,\theta) \in (R, R + \varepsilon h) \times [0, 2\pi), \\ U^\varepsilon(R,\theta) = u^\varepsilon(R,\theta) & \text{for } \theta \in [0, 2\pi), \\ \varepsilon^t \partial_r U^\varepsilon(R,\theta) = \partial_r u^\varepsilon(R,\theta) & \text{for } \theta \in [0, 2\pi), \\ u^\varepsilon(R + \varepsilon h, \theta) = 0 & \text{for } \theta \in [0, 2\pi). \end{cases} \tag{6.5}$$

Using separation of variables, we look for the eigenelements $(\lambda^\varepsilon, \{U^\varepsilon, u^\varepsilon\})$ of (6.5) in the form

$$\begin{aligned} U^\varepsilon(r,\theta) &= R^\varepsilon(r)T^\varepsilon(\theta) \\ u^\varepsilon(r,\theta) &= r^\varepsilon(r)t^\varepsilon(\theta), \end{aligned} \tag{6.6}$$

for certain functions R^ε, T^ε, r^ε, and t^ε. Replacing (6.6) in (6.5) we have

$$\begin{cases} \dfrac{r^2R^{\varepsilon\prime\prime}(r)+rR^{\varepsilon\prime}(r)}{R^\varepsilon(r)}+\lambda^\varepsilon r^2=-\dfrac{T^{\varepsilon\prime\prime}(\theta)}{T^\varepsilon(\theta)}=\mu^\varepsilon & \text{for } (r,\theta)\in[0,R)\times[0,2\pi), \\ \dfrac{r^2r^{\varepsilon\prime\prime}(r)+rr^{\varepsilon\prime}(r)}{r^\varepsilon(r)}+\lambda^\varepsilon r^2\varepsilon^{-m}=-\dfrac{t^{\varepsilon\prime\prime}(\theta)}{t^\varepsilon(\theta)}=\hat\mu^\varepsilon & \text{for } (r,\theta)\in(R,R+\varepsilon h)\times[0,2\pi), \\ T^\varepsilon(0)=T^\varepsilon(2\pi), & T^{\varepsilon\prime}(0)=T^{\varepsilon\prime}(2\pi), \\ t^\varepsilon(0)=t^\varepsilon(2\pi), & t^{\varepsilon\prime}(0)=t^{\varepsilon\prime}(2\pi), \\ R^\varepsilon(R)T^\varepsilon(\theta)=r^\varepsilon(R)t^\varepsilon(\theta) & \text{for } \theta\in[0,2\pi), \\ \varepsilon^l R^{\varepsilon\prime}(R)T^\varepsilon(\theta)=r^{\varepsilon\prime}(R)t^\varepsilon(\theta) & \text{for } \theta\in[0,2\pi) \\ r^\varepsilon(R+\varepsilon h)t^\varepsilon(\theta)=0 & \text{for } \theta\in[0,2\pi), \end{cases} \tag{6.7}$$

where μ^ε and $\hat\mu^\varepsilon$ are constants to be determined. It is easy to check that the only values μ^ε and $\hat\mu^\varepsilon$ satisfying (6.7) with $U^\varepsilon(r,\theta)=R^\varepsilon(r)T^\varepsilon(\theta)\not\equiv 0$ and $u^\varepsilon(r,\theta)=r^\varepsilon(r)t^\varepsilon(\theta)\not\equiv 0$ are

$$\mu_k^\varepsilon=\hat\mu_k^\varepsilon=k^2,\ \text{with } k\in\mathbb{N}_0=\mathbb{N}\cup\{0\},$$

and, consequently,

$$\begin{aligned} T_k^\varepsilon(\theta)=t_k^\varepsilon(\theta)&=\sin(k\theta)\ \text{ for } \theta\in[0,2\pi),\ k\in\mathbb{N}, \\ T_k^\varepsilon(\theta)=t_k^\varepsilon(\theta)&=\cos(k\theta)\ \text{for } \theta\in[0,2\pi),\ k\in\mathbb{N}_0. \end{aligned}$$

Thus, for $k\in\mathbb{N}_0$ fixed, $(\lambda^\varepsilon,\{R^\varepsilon,r^\varepsilon\})$ verifies

$$r^2R^{\varepsilon\prime\prime}(r)+rR^{\varepsilon\prime}(r)+(\lambda^\varepsilon r^2-k^2)R^\varepsilon(r)=0 \quad \text{for } r\in[0,R), \tag{6.8a}$$

$$r^2r^{\varepsilon\prime\prime}(r)+rr^{\varepsilon\prime}(r)+(\lambda^\varepsilon r^2\varepsilon^{-m}-k^2)r^\varepsilon(r)=0 \quad \text{for } r\in(R,R+\varepsilon h), \tag{6.8b}$$

$$R^\varepsilon(R)=r^\varepsilon(R), \qquad \varepsilon^l R^{\varepsilon\prime}(R)=r^{\varepsilon\prime}(R), \tag{6.8c}$$

$$r^\varepsilon(R+\varepsilon h)=0. \tag{6.8d}$$

For $\varepsilon>0$ and $k\in\mathbb{N}_0$ fixed, the solutions of (6.8a)–(6.8b) that have no singularity at the origin are given by

$$\begin{aligned} R^\varepsilon(r)&=a^\varepsilon J_k(\sqrt{\lambda^\varepsilon}r) && \text{for } r\in[0,R), \\ r^\varepsilon(r)&=b^\varepsilon J_k(\sqrt{\lambda^\varepsilon\varepsilon^{-m}}r)+\tilde b^\varepsilon Y_k(\sqrt{\lambda^\varepsilon\varepsilon^{-m}}r) && \text{for } r\in(R,R+\varepsilon h), \end{aligned}$$

where $J_k(s)$ and $Y_k(s)$ denote the Bessel functions of the first and second kind, respectively, and a^ε, b^ε, $\tilde b^\varepsilon$ are some constants. Using the conditions (6.8c)–(6.8d),

it can be proved that only the values λ which are roots of the equation

$$\begin{aligned}&\Big(Y_k\big(\sqrt{\lambda\varepsilon^{-m}}(R+\varepsilon h)\big)J_k'(\sqrt{\lambda\varepsilon^{-m}}R)-J_k\big(\sqrt{\lambda\varepsilon^{-m}}(R+\varepsilon h)\big)Y_k'(\sqrt{\lambda\varepsilon^{-m}}R)\Big)J_k(\sqrt{\lambda}R)\\&\quad=\Big(Y_k\big(\sqrt{\lambda\varepsilon^{-m}}(R+\varepsilon h)\big)J_k(\sqrt{\lambda\varepsilon^{-m}}R)-J_k\big(\sqrt{\lambda\varepsilon^{-m}}(R+\varepsilon h)\big)Y_k(\sqrt{\lambda\varepsilon^{-m}}R)\Big)\\&\qquad*\varepsilon^{t+m/2}J_k'(\sqrt{\lambda}R)\end{aligned}\tag{6.9}$$

verify (6.8) with $R^\varepsilon \not\equiv 0$ or $r^\varepsilon \not\equiv 0$. Moreover, in this case, namely when λ is a root of (6.9), and if $J_k(\sqrt{\lambda}R) \neq 0$, it follows that

$$\begin{aligned}R_k^\varepsilon(r)&=\Big(Y_k\big(\sqrt{\lambda\varepsilon^{-m}}(R+\varepsilon h)\big)J_k(\sqrt{\lambda\varepsilon^{-m}}R)\\&\qquad-J_k\big(\sqrt{\lambda\varepsilon^{-m}}(R+\varepsilon h)\big)Y_k(\sqrt{\lambda\varepsilon^{-m}}R)\Big)(J_k(\sqrt{\lambda}R))^{-1}J_k(\sqrt{\lambda}r)\\r_k^\varepsilon(r)&=Y_k\big(\sqrt{\lambda\varepsilon^{-m}}(R+\varepsilon h)\big)J_k(\sqrt{\lambda\varepsilon^{-m}}r)-J_k\big(\sqrt{\lambda\varepsilon^{-m}}(R+\varepsilon h)\big)Y_k(\sqrt{\lambda\varepsilon^{-m}}r)\end{aligned}\tag{6.10}$$

Thus, any root λ of (6.9) is an eigenvalue of (6.5) and the corresponding eigenfunctions are

$$\begin{aligned}&U_k^\varepsilon(r,\theta)=\alpha^\varepsilon R_k^\varepsilon(r)\sin(k\theta)\ \text{ for } (r,\theta)\in[0,R)\times[0,2\pi),\\&u_k^\varepsilon(r,\theta)=\alpha^\varepsilon r_k^\varepsilon(r)\sin(k\theta)\quad \text{for } (r,\theta)\in[R,R+\varepsilon h)\times[0,2\pi)\end{aligned}\tag{6.11}$$

with $k \in \mathbb{N}$, and

$$\begin{aligned}&U_k^\varepsilon(r,\theta)=\alpha^\varepsilon R_k^\varepsilon(r)\cos(k\theta)\ \text{ for } (r,\theta)\in[0,R)\times[0,2\pi)\\&u_k^\varepsilon(r,\theta)=\alpha^\varepsilon r_k^\varepsilon(r)\cos(k\theta)\quad \text{for } (r,\theta)\in[R,R+\varepsilon h)\times[0,2\pi)\end{aligned}\tag{6.12}$$

with $k \in \mathbb{N}_0$, where α^ε is a constant, R_k^ε and r_k^ε are given by (6.10); see Figs. 6.3, 6.4, 6.5, and 6.6 for some examples of eigenfunctions of (6.5) with $R = 1$, $h = 2$, $\varepsilon = 0.1$ and different values of t, m, and k.

We observe that the eigenfunctions of (6.8) corresponding to the possible eigenvalues $\lambda = \lambda^\varepsilon$ (roots of (6.9)) such that $J_k(\sqrt{\lambda^\varepsilon}R) = 0$ with $k \in \mathbb{N}_0$ fixed, are not included in (6.10). Each one of these values, $\lambda = \eta_{k,j}^2 R^{-2}$ with $\{\eta_{k,j}\}_{j=1}^\infty$ zeros of the Bessel functions $J_k(s)$, is eigenvalue of (6.8) (and, consequently, of (6.5)) only for certain values of ε; those ε that satisfy the transcendental equation

$$\begin{aligned}&J_k(\eta_{k,j}\varepsilon^{-m/2})Y_k\big(\eta_{k,j}\varepsilon^{-m/2}(1+\varepsilon hR^{-1})\big)\\&-Y_k(\eta_{k,j}\varepsilon^{-m/2})J_k\big(\eta_{k,j}\varepsilon^{-m/2}(1+\varepsilon hR^{-1})\big)=0.\end{aligned}$$

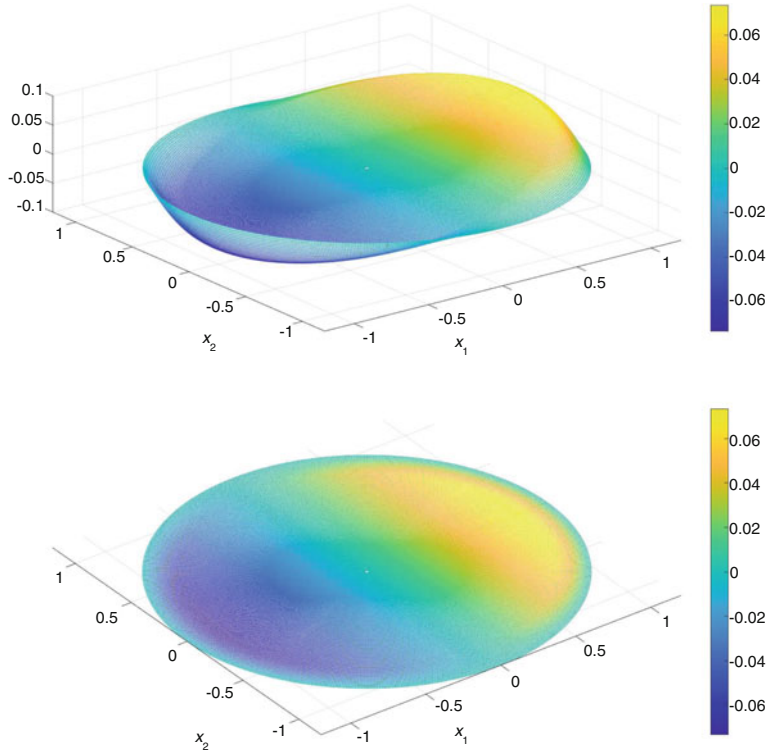


Fig. 6.3 Eigenfunction (6.12) corresponding to $\lambda^\varepsilon \approx 0.02$ for $R = 1$, $h = 2$, $t = 1$, $m = 3.5$, $\varepsilon = 0.1$, and $k = 1$, and its projection on the plane x_1x_2

In this case, $J_k'(\sqrt{\lambda^\varepsilon}R) = J_k'(\eta_{k,j}) \neq 0$ and the corresponding eigenfunctions of (6.8) are

$$\begin{aligned} R^\varepsilon_{k,j}(r) &= \Big(Y_k\big(\eta_{k,j}\varepsilon^{-m/2}(1+\varepsilon h R^{-1})\big)J_k'(\eta_{k,j}\varepsilon^{-m/2}) \\ &\quad -J_k\big(\eta_{k,j}\varepsilon^{-m/2}(1+\varepsilon h R^{-1})\big)Y_k'(\eta_{k,j}\varepsilon^{-m/2})\Big) \\ &\quad *(J_k'(\eta_{k,j}))^{-1}\varepsilon^{-t-m/2}J_k(\eta_{k,j}R^{-1}r) \\ r^\varepsilon_{k,j}(r) &= Y_k\big(\eta_{k,j}\varepsilon^{-m/2}(1+\varepsilon h R^{-1})\big)J_k(\eta_{k,j}\varepsilon^{-m/2}R^{-1}r) \\ &\quad -J_k\big(\eta_{k,j}\varepsilon^{-m/2}(1+\varepsilon h R^{-1})\big)Y_k(\eta_{k,j}\varepsilon^{-m/2}R^{-1}r) \end{aligned} \tag{6.13}$$

and, consequently, the eigenfunctions of (6.5) are

$$\begin{aligned} U^\varepsilon_{k,j}(r,\theta) &= \alpha^\varepsilon R^\varepsilon_{k,j}(r)\sin(k\theta) \text{ for } (r,\theta) \in [0,R)\times[0,2\pi), \\ u^\varepsilon_{k,j}(r,\theta) &= \alpha^\varepsilon r^\varepsilon_{k,j}(r)\sin(k\theta) \quad \text{for } (r,\theta) \in [R,R+\varepsilon h)\times[0,2\pi) \end{aligned} \tag{6.14}$$

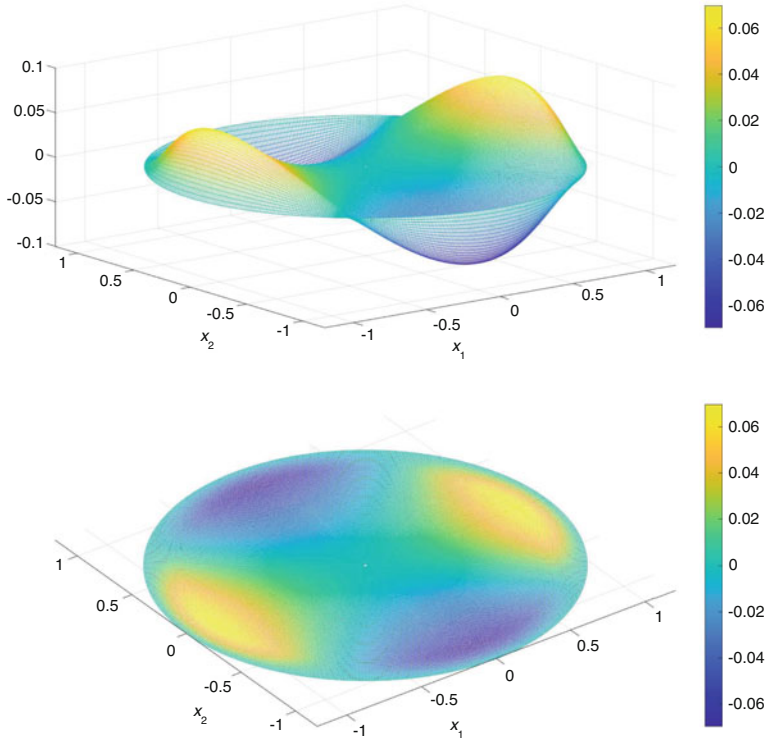


Fig. 6.4 Eigenfunction (6.12) corresponding to $\lambda^\varepsilon \approx 0.072$ for $R = 1$, $h = 2$, $t = 1$, $m = 3$, $\varepsilon = 0.1$, and $k = 2$, and its projection on the plane x_1x_2

with $k \in \mathbb{N}$, and

$$\begin{aligned} U^\varepsilon_{k,j}(r,\theta) &= \alpha^\varepsilon R^\varepsilon_{k,j}(r)\cos(k\theta) \quad \text{for } (r,\theta) \in [0,R) \times [0,2\pi) \\ u^\varepsilon_{k,j}(r,\theta) &= \alpha^\varepsilon r^\varepsilon_{k,j}(r)\cos(k\theta) \quad \text{for } (r,\theta) \in [R,R+\varepsilon h) \times [0,2\pi) \end{aligned} \tag{6.15}$$

with $k \in \mathbb{N}_0$, where α^ε is any constant, and $R^\varepsilon_{k,j}$ and $r^\varepsilon_{k,j}$ are given by (6.13); see, for example, Figs. 6.7 and 6.8.

Remark 1 Note that with the explicit formulas we do not determine the eigenvalues ordered as in the sequence (6.3).

Remark 2 Note that Figs. 6.7 and 6.8 show the graphic of some eigenfunctions associated with *high frequencies* and, further specifying, with eigenvalues λ^ε which coincide with eigenvalues of the Dirichlet problem in Ω

$$\begin{cases} -\Delta_{\mathbf{x}} U = \lambda U & \text{in } \Omega, \\ U = 0 & \text{on } \partial\Omega. \end{cases}$$

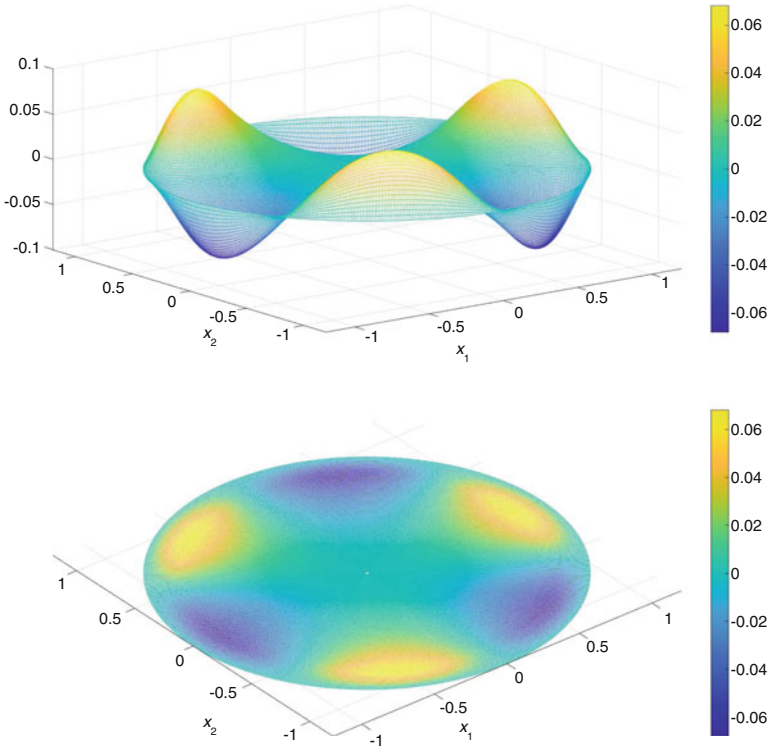


Fig. 6.5 Eigenfunction (6.12) corresponding to $\lambda^\varepsilon \approx 0.079$ for $R = 1$, $h = 2$, $t = 1$, $m = 3$, $\varepsilon = 0.1$, and $k = 3$, and its projection on the plane x_1x_2

This is in good agreement with the results obtained in Sect. 6.5 for the high frequencies. In this case, for certain sequences of ε, the corresponding eigenfunctions are close to zero in ω_ε.

6.3 Asymptotic Expansions for Low Frequencies

In this section, we study the asymptotic behavior of the eigenvalues of (6.1) of order $O(\varepsilon^{m-2})$ when $m > 2$ and $t = 1$, the so-called low frequencies, and their corresponding eigenfunctions. Different limit behaviors appear for these frequencies depending on whether the curvature $\varkappa$ of Γ, the boundary of the fixed domain Ω, is constant or not. Here, we provide asymptotic expansions for the case where the curvature is constant, namely $\varkappa(\tau) = \varkappa_0$ for all $\tau \in \mathbb{S}_\ell$, while the justification for these asymptotic expansions is given in Sect. 6.4. The asymptotic expansions and the justification for the case where $\varkappa$ is not a constant are provided in [GoEtAl20].

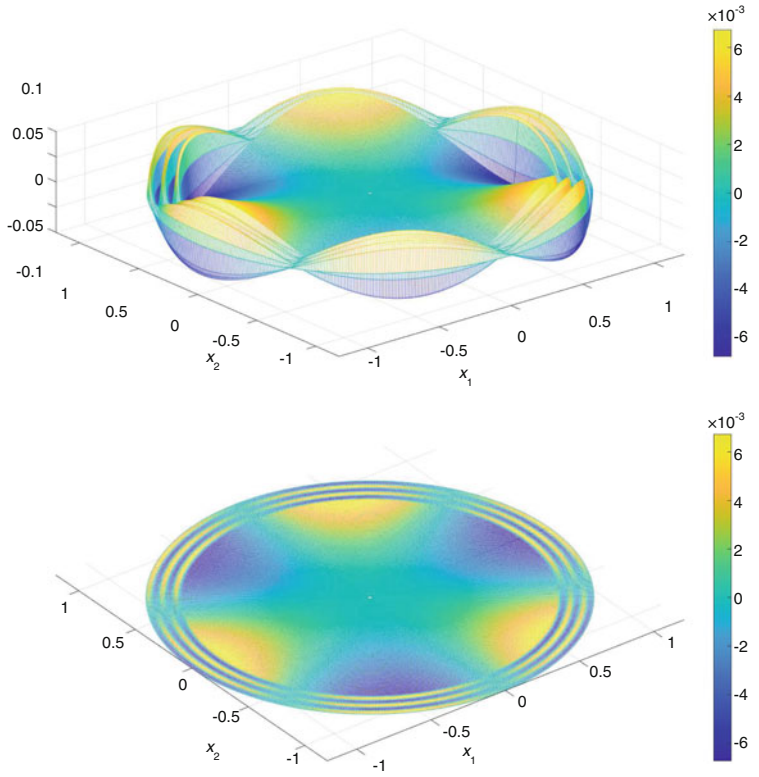


Fig. 6.6 Eigenfunction (6.12) corresponding to $\lambda^\varepsilon \approx 7.438$ for $R = 1$, $h = 2$, $t = 1$, $m = 3$, $\varepsilon = 0.1$, and $k = 3$, and its projection on the plane x_1x_2

Let us recall that in [GoEtAl20] new localization effects for the eigenfunctions of (6.2) are shown around points τ_0 of the boundary where the curvature of Γ has a local extremum. However, when the curvature is constant, we state here that the corresponding eigenfunctions are significant over the whole domain Ω_ε, and no localization effects arise at points of the boundary.

Without loss of generality, we assume again that the domain Ω is the disk with radius $R > 0$ centered at the origin, namely $\Omega = B(O, R) = \{x : \|x\| < R\}$, and the band ω_ε is the annulus $\omega_\varepsilon = \{x : R < \|x\| < R + \varepsilon h\}$.

In a neighborhood of Γ, the circle centered at the origin with radius R, let (ν, τ) be the natural orthogonal curvilinear coordinates: τ is the arc length and ν the distance along the normal vector to Γ; $\nu < 0$ inside Ω (cf. Sect. 6.1). Now, the length of the curve Γ is given by $\ell = 2\pi R$ and its curvature is the constant $\varkappa_0 = 1/R$.

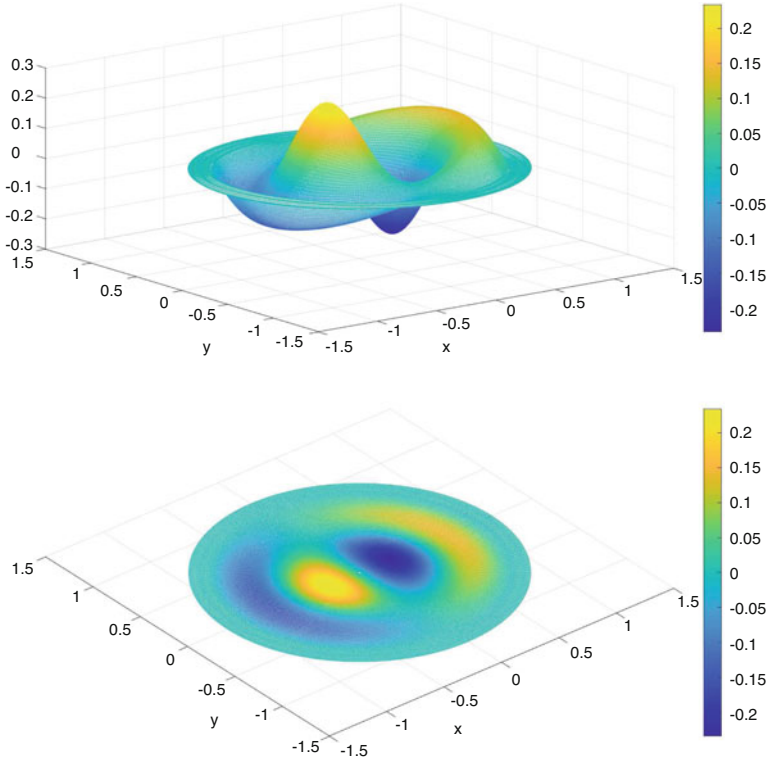


Fig. 6.7 Eigenfunction (6.15) corresponding to $\lambda^\varepsilon \approx 49.2186$ for $R = 1$, $h = 2$, $t = 1$, $m = 2$, $\varepsilon = 0.078$, and $k = 1$, and its projection on the plane x_1x_2

Also in a neighborhood of Γ, we introduce some *local variables* defined by

$$(\zeta, \tau), \quad \zeta = \frac{\nu}{\varepsilon}, \tag{6.16}$$

that transform the thin domain ω_ε into a band of length ℓ and width $O(1)$; namely, $\{(\nu, \tau) : \nu \in (0, \varepsilon h), \tau \in \mathbb{S}_\ell\}$ into $\{(\zeta, \tau) : \zeta \in (0, h), \tau \in \mathbb{S}_\ell\}$.

Since a boundary layer phenomenon appears in a neighborhood of Γ, it proves necessary to consider outer expansions for the eigenfunctions in Ω and inner expansions in a neighborhood of Γ in the local coordinates (6.16). Thus, we write the Laplace operator in curvilinear coordinates

$$\Delta_{\nu,\tau} = K(\nu, \tau)^{-1}\partial_\nu(K(\nu, \tau)\partial_\nu) + K(\nu, \tau)^{-1}\partial_\tau(K(\nu, \tau)^{-1}\partial_\tau), \tag{6.17}$$

being $K(\nu, \tau) = 1 + \nu\varkappa_0$, and introduce the local variable (6.16). Gathering the coefficients at the different powers of ε, we write

$$\Delta_{\zeta,\tau} = \varepsilon^{-2}\,\partial_\zeta^2 + \varepsilon^{-1}\,\varkappa_0\partial_\zeta - \varkappa_0^2\zeta\,\partial_\zeta + \partial_\tau^2 + \cdots; \tag{6.18}$$

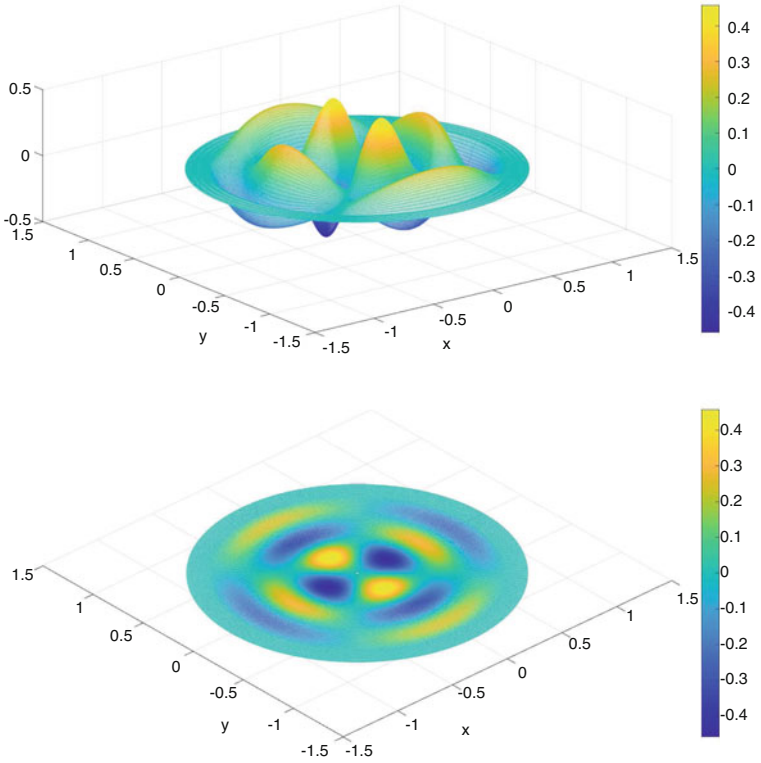


Fig. 6.8 Eigenfunction (6.15) corresponding to $\lambda^\varepsilon \approx 135.0198$ for $R = 1$, $h = 2$, $t = 1$, $m = 2$, $\varepsilon = 0.075$, and $k = 2$, and its projection on the plane x_1x_2

here and in the sequel the dots denote further asymptotic terms of different powers of ε of no use in the present analysis.

We consider an asymptotic expansion for the eigenvalues λ^ε and for the corresponding eigenfunctions $\{U^\varepsilon, u^\varepsilon\}$ in Ω and ω_ε of the form:

$$\lambda^\varepsilon = \varepsilon^{m-2}(\lambda_0 + \varepsilon\lambda_1 + \varepsilon^2\lambda_2 + \cdots), \tag{6.19}$$

$$U^\varepsilon(x) = V(x) + \varepsilon \mathsf{V}_1(x) + \varepsilon^2 \mathsf{V}_2(x) + \cdots, \qquad x \in \Omega, \tag{6.20}$$

$$u^\varepsilon(\zeta, \tau) = v_0(\zeta, \tau) + \varepsilon v_1(\zeta, \tau) + \varepsilon^2 v_2(\zeta, \tau) + \cdots, \qquad \zeta \in [0, h), \tau \in \mathbb{S}_\ell, \tag{6.21}$$

respectively, where v_i are ℓ–periodic functions in τ. Besides, we assume that at least one of the functions V or v_0 in (6.20)–(6.21) are different from zero (see normalization in (6.43) and convergence (6.45)).

By replacing expansions (6.19)–(6.21) in problem (6.1), after considering equations (6.18), we collect coefficients of the same powers of ε. In a first step, we see that the leading terms in (6.19)–(6.21) satisfy the following problem:

$$-A\Delta_x V = 0 \quad \text{in } \Omega, \tag{6.22}$$

$$-a\partial_\zeta^2 v_0 = \lambda_0 v_0, \quad \zeta \in (0,h),\ \tau \in \mathbb{S}_\ell, \tag{6.23}$$

$$V = v_0 \quad \text{on } \Gamma, \tag{6.24}$$

$$a\partial_\zeta v_0(0,\tau) = 0, \quad \tau \in \mathbb{S}_\ell, \tag{6.25}$$

$$v_0(h,\tau) = 0, \quad \tau \in \mathbb{S}_\ell. \tag{6.26}$$

From (6.23), (6.25), and (6.26), we deduce that λ_0 is an eigenvalue of

$$\begin{cases} -ay_0'' = \lambda_0 y_0 & \zeta \in (0,h), \\ y_0'(0) = y_0(h) = 0 \end{cases} \tag{6.27}$$

and

$$v_0(\zeta,\tau) = y_0(\zeta)v(\tau), \quad \zeta \in (0,h),\ \tau \in \mathbb{S}_\ell, \tag{6.28}$$

where $y_0(\zeta)$ is an eigenfunction of (6.27) corresponding to λ_0 and $v(\tau)$, at this stage, is an arbitrary function of τ. Obviously, by assumption, $v(\tau)$ does not vanish on $\mathbb{S}_\ell$ (see (6.22), (6.24), and (6.28)). It is clear that the eigenvalues of (6.27) are given by

$$\lambda_{0,k} = \frac{a(2k-1)^2\pi^2}{4h^2} \qquad \text{for } k = 1,2,\ldots \tag{6.29}$$

and the corresponding eigenfunctions can be chosen to be

$$y_{0,k}(\zeta) = \cos\left(\frac{(2k-1)\pi}{2h}\zeta\right) \qquad \text{for } k = 1,2,\ldots, \tag{6.30}$$

while $v(\tau)$ in (6.28) has to be determined.

In a second step, we obtain the problem

$$-a\partial_\zeta^2 v_1 - a\varkappa_0\,\partial_\zeta v_0 = \lambda_0 v_1 + \lambda_1 v_0, \quad \zeta \in (0,h),\ \tau \in \mathbb{S}_\ell, \tag{6.31}$$

$$a\partial_\zeta v_1(0,\tau) = 0, \quad \tau \in \mathbb{S}_\ell, \tag{6.32}$$

$$v_1(h,\tau) = 0, \quad \tau \in \mathbb{S}_\ell. \tag{6.33}$$

Since $v_0(\zeta,\tau) = y_0(\zeta)v(\tau)$ verifies (6.23), (6.25), and (6.26), the compatibility condition for the non-homogeneous problem (6.31)–(6.33) in the ζ–variable reads

$$-a\varkappa_0 v(\tau)\int_0^h y_0'(\zeta)y_0(\zeta)\,d\zeta = \lambda_1 v(\tau)\int_0^h y_0(\zeta)^2\,d\zeta, \quad \tau\in\mathbb{S}_\ell.$$

Moreover, by (6.30), $\int_0^h y_0^2\,d\zeta = \frac{h}{2}$ and $\int_0^h y_0' y_0\,d\zeta = \frac{1}{2}(y_0(h)^2 - y_0(0)^2) = -\frac{1}{2}$, and consequently we have

$$\lambda_1 = \frac{a\varkappa_0}{h}. \tag{6.34}$$

In addition, the functions v_1 satisfying (6.31) and (6.32) can be written in the form

$$v_1(\zeta,\tau) = \varkappa_0 v(\tau)y_1(\zeta), \quad \zeta\in(0,h),\ \tau\in\mathbb{S}_\ell,$$

where $y_1(\zeta)$ is a solution of

$$\begin{cases} -ay_1'' - \lambda_0 y_1 = ay_0' + \dfrac{a}{h}y_0 & \zeta\in(0,h),\\ y_1'(0) = y_1(h) = 0. \end{cases} \tag{6.35}$$

In fact, for each fixed eigenpair (λ_0, y_0) of (6.27), we can choose the solution $y_1(\zeta)$ above to be the unique solution which satisfies $\int_0^h y_1(\zeta)y_0(\zeta)\,d\zeta = 0$, and then, for $(\lambda_0, y_0) = (\lambda_{0,k}, y_{0,k})$ verifying (6.29) and (6.30), we have $v_1(\zeta,\tau) = v_{1,k}(\zeta,\tau)$ defined by

$$v_{1,k}(\zeta,\tau) = \varkappa_0 v(\tau)y_{1,k}(\zeta) \tag{6.36}$$

where

$$\begin{aligned} y_{1,k}(\zeta) = &-\left(\frac{\zeta}{2} + \frac{(8-(2k-1)^2\pi^2)h}{4(2k-1)^2\pi^2}\right)\cos\left(\frac{(2k-1)\pi}{2h}\zeta\right)\\ &-\frac{1}{(2k-1)\pi}(\zeta-h)\sin\left(\frac{(2k-1)\pi}{2h}\zeta\right), \end{aligned}$$

for $k = 1, 2\ldots.$

Following the process, in the next step, we have the problem for v_2:

$$\begin{aligned} &-a\partial_\zeta^2 v_2 - a\varkappa_0\,\partial_\zeta v_1 + a\varkappa_0^2\,\zeta\,\partial_\zeta v_0 - a\partial_\tau^2 v_0\\ &= \lambda_0 v_2 + \lambda_1 v_1 + \lambda_2 v_0, \quad \zeta\in(0,h),\ \tau\in\mathbb{S}_\ell, \end{aligned} \tag{6.37}$$

$$a\partial_\zeta v_2(0,\tau) = A\partial_\nu V(0,\tau), \quad \tau\in\mathbb{S}_\ell, \tag{6.38}$$

$$v_2(h,\tau) = 0, \quad \tau\in\mathbb{S}_\ell. \tag{6.39}$$

Since v_0 verifies (6.23), (6.25), and (6.26), the compatibility condition for the non-homogeneous problem (6.37)–(6.39) reads

$$A\partial_\nu V(0,\tau)y_0(0) - a\int_0^h (\varkappa_0\,\partial_\zeta v_1(\zeta,\tau) - \varkappa_0^2\,\zeta\,\partial_\zeta v_0(\zeta,\tau) + \partial_\tau^2 v_0(\zeta,\tau))y_0(\zeta)\,d\zeta$$
$$= \lambda_1\int_0^h v_1(\zeta,\tau)y_0(\zeta)\,d\zeta + \lambda_2\int_0^h v_0(\zeta,\tau)y_0(\zeta)\,d\zeta.$$

Now, by virtue of (6.28), (6.30), (6.34), and (6.36) we get

$$A\partial_\nu V(0,\tau) + a\varkappa_0^2 d_k v(\tau) - v''(\tau)\frac{ah}{2} = \lambda_2 v(\tau)\frac{h}{2},$$

where d_k denotes the constant

$$d_k = -\frac{h(4+(2k-1)^2\pi^2)}{8(2k-1)^2\pi^2}. \tag{6.40}$$

The last compatibility condition gives a boundary condition for the function V in Eq. (6.22). Thus, due to (6.24), (λ_2, V) in (6.19)–(6.20) is an eigenpair of the spectral problem

$$\begin{cases} -A\Delta_x V = 0 & \text{in } \Omega, \\ A\partial_\nu V = \dfrac{ah}{2}\partial_\tau^2 V - a\varkappa_0^2 d_k V + \lambda_2\dfrac{h}{2}V & \text{on } \Gamma. \end{cases} \tag{6.41}$$

Note that the spectral parameter and the second order tangential derivatives appear in the boundary condition on Γ (cf. [GoEtAl06a]).

We consider the weak formulation (which does not depend on λ_0): to find μ and $V \in \mathscr{H}^{1,1}(\Omega,\Gamma)\setminus\{0\}$, such that

$$A\int_\Omega \nabla_x V\cdot\nabla_x W\,dx + \frac{ah}{2}\int_\Gamma \partial_\tau V\partial_\tau W\,d\tau = \mu\frac{h}{2}\int_\Gamma VW\,d\tau \quad \forall W\in\mathscr{H}^{1,1}(\Omega,\Gamma), \tag{6.42}$$

where the functional space $\mathscr{H}^{1,1}(\Omega,\Gamma)$ is defined as the completion of $C^\infty(\overline{\Omega})$ with respect to the norm

$$\|W\|_{\mathscr{H}^{1,1}(\Omega,\Gamma)} = \left(\|W\|^2_{H^1(\Omega)} + \|W\|^2_{H^1(\Gamma)}\right)^{1/2}.$$

Since adding the term $\frac{h}{2}\int_\Gamma VW\,d\tau$ to the left hand side of (6.42) one gets a scalar product in $\mathscr{H}^{1,1}(\Omega,\Gamma)$ and the embeddings of this space in $L^2(\Omega)$ and $L^2(\Gamma)$ are compact, we can write an eigenvalue problem for a non-negative, symmetric, and

compact operator $\mathscr{A}$ on $\mathscr{H}^{1,1}(\Omega, \Gamma)$ defined by

$$(\mathscr{A}U, W) = \frac{h}{2}\int_{\Gamma} UW\, d\tau \quad \forall U, W \in \mathscr{H}^{1,1}(\Omega, \Gamma),$$

whose eigenvalues are 0, with the corresponding eigenspace $\{U \in \mathscr{H}^{1,1}(\Omega, \Gamma) : U = 0 \text{ on } \Gamma\}$, and $(\mu + 1)^{-1}$ with finite multiplicity and μ an eigenvalue of (6.42). Therefore, (6.42) has a discrete spectrum which we denote by $\{\mu_p\}_{p=1}^{\infty}$. Thus, (6.41) also has a discrete spectrum given by $\mu_p + 2a\varkappa_0^2 d_k/h$ with d_k defined by (6.40), which depends on λ_0.

Hence, we have found the double sequences for the low frequencies

$$\lambda^{\varepsilon} \sim \varepsilon^{m-2}\frac{a(2k-1)^2\pi^2}{4h^2} + \varepsilon^{m-1}\frac{a\varkappa_0}{h} + \varepsilon^m\left(\mu_p + \frac{2a\varkappa_0^2 d_k}{h}\right), \qquad k,\ p = 1, 2\ldots$$

for which the corresponding eigenfunctions $\{U^{\varepsilon}, u^{\varepsilon}\}$ have asymptotics in Ω_{ε} given by

$$U^{\varepsilon}(x) \sim V^p(x) \qquad x \in \Omega,$$

and

$$u^{\varepsilon}(v, \tau) \sim v_{0,k}^p(\frac{v}{\varepsilon}, \tau) + \varepsilon v_{1,k}^p(\frac{v}{\varepsilon}, \tau), \qquad x \in \omega_{\varepsilon}$$

where V^p is an eigenfunction of problem (6.42) corresponding to μ_p, $v_{0,k}^p(\zeta, \tau) = y_{0,k}(\zeta)V^p(0, \tau)$, $y_{0,k}$ is an eigenfunction of problem (6.27) corresponding to the eigenvalue $a(2k-1)^2\pi^2h^{-2}4^{-1}$ (cf. (6.30)), and $v_{1,k}^p$ is given by (6.36) for $v(\tau) \equiv V^p(0, \tau)$.

6.4 On Convergence for Low Frequencies

In this section, we justify the asymptotic expansions in Sect. 6.3 up to a certain degree which can be improved by constructing higher order terms in (6.19)–(6.21). In particular, we provide estimates which establish the closeness of the eigenvalues of (6.1) and

$$\lambda_0 + \varepsilon\frac{a\varkappa_0}{2} + \varepsilon^2\lambda_2,$$

where (λ_0, λ_2) are pairs of eigenvalues of (6.27) and (6.41) (see (6.19) and (6.44)). When justifying asymptotics (6.20) and (6.21) for the eigenfunctions $\{U^{\varepsilon}, u^{\varepsilon}\}$, we deal with groups of eigenfunctions corresponding to eigenvalues λ^{ε} of (6.1)

verifying (6.44) and the approaches hold in the topology of $H^1(\Omega_\varepsilon)$ given by the scalar product (6.43), in the way stated by the Theorem 1 (see (6.54)). Throughout the section we assume that $t = 1$ and $m > 2$.

Let us note that choosing a suitable normalization for the eigenfunctions (cf. (6.43)) is essential in order to obtain the results throughout the section.

We first introduce some notations and results for further use. For each $\varepsilon > 0$, let us denote by $\mathscr{H}^\varepsilon$ the space $H_0^1(\Omega_\varepsilon)$ with the scalar product

$$(\{U,u\},\{V,v\})_{\mathscr{H}^\varepsilon} = \varepsilon^2 A\int_\Omega \nabla_x U\cdot\nabla_x V\,dx + \varepsilon a\int_{\omega_\varepsilon}\nabla_x u\cdot\nabla_x v\,dx \tag{6.43}$$

for all $\{U,u\},\{V,v\} \in H_0^1(\Omega_\varepsilon)$. Let $\mathscr{A}^\varepsilon$ be a positive, compact, and symmetric operator on $\mathscr{H}^\varepsilon$ defined by

$$(\mathscr{A}^\varepsilon\{U,u\},\{G,g\})_{\mathscr{H}^\varepsilon} = \varepsilon^m\int_\Omega UG\,dx + \frac{1}{\varepsilon}\int_{\omega_\varepsilon} ug\,dx \quad \forall\{U,u\},\{G,g\}\in H_0^1(\Omega_\varepsilon).$$

It is clear that the eigenvalues of $\mathscr{A}^\varepsilon$ are $\{\varepsilon^{m-2}/\lambda_k^\varepsilon\}_{k=1}^\infty$ where $\{\lambda_k^\varepsilon\}_{k=1}^\infty$ are the eigenvalues of (6.1).

In order to prove convergence results, we use a classical result on "almost eigenvalues and eigenvectors" from the spectral perturbation theory, namely Lemma 1 (see [ViLy57] and Chapter 6 in [BiSo87] for the proof).

Lemma 1 *Let $A : H \longrightarrow H$ be a linear, self-adjoint, positive, and compact operator on a separable Hilbert space H. Let $u \in H$, with $\|u\|_H = 1$ and $\lambda,\ r > 0$ such that $\|Au - \lambda u\|_H \le r$. Then, there exists an eigenvalue λ_i of the operator A satisfying the inequality $|\lambda - \lambda_i| \le r$. Moreover, for any $r^* > r$ there is $u^* \in H$, with $\|u^*\|_H = 1$, u^* belonging to the eigenspace associated with all the eigenvalues of the operator A lying on the segment $[\lambda - r^*, \lambda + r^*]$ and such that*

$$\|u - u^*\|_H \le \frac{2r}{r^*}.$$

Theorem 1 *Let (λ_0, y_0) be an eigenpair of (6.27) such that $\|y_0\|^2_{L^2(0,h)} = h/2$. Let y_1 be the solution of (6.35) such that $\int_0^h y_1 y_0\,d\zeta = 0$. Let (λ_2, V) be an eigenpair of (6.41) where*

$$d_k = y_0(0)^{-1}\left(\int_0^h \zeta y_0' y_0\,d\zeta - \int_0^h y_1' y_0\,d\zeta\right),$$

(cf. (6.40)) and such that $\|V\|^{-2}_{L^2(\Gamma)} = h/2$. Let us consider $v_0(\zeta,\tau) = y_0(\zeta)V(0,\tau)$ for $(\zeta,\tau)\in\omega$. Then, if $t = 1$ and $m > 2$, there are eigenvalues $\lambda^\varepsilon_{k(\varepsilon)}$ of problem

(6.1) such that

$$\left| \frac{\lambda^{\varepsilon}_{k(\varepsilon)}}{\varepsilon^{m-2}} - \lambda_0 - \varepsilon \frac{a\varkappa_0}{h} - \varepsilon^2 \lambda_2 \right| \le C(\varepsilon^3 + \varepsilon^{m-1}) \tag{6.44}$$

where C is a constant independent of ε. Moreover, there is a linear combination of eigenfunctions $\{\tilde{U}^{\varepsilon}, \tilde{u}^{\varepsilon}\} \in H^1(\Omega_\varepsilon)$, $\{\tilde{U}^{\varepsilon}, \tilde{u}^{\varepsilon}\}$ corresponding to eigenvalues $\lambda^{\varepsilon}_{k(\varepsilon)}$ of (6.1) which satisfy $\lambda^{\varepsilon}_{k(\varepsilon)} \varepsilon^{2-m} \in [\lambda_0 - K\varepsilon^{\vartheta}, \lambda_0 + K\varepsilon^{\vartheta}]$ with $K > 0$ and $0 < \vartheta < \min(2, m-2)$, $\|\{\tilde{U}^{\varepsilon}, \tilde{u}^{\varepsilon}\}\|_{\mathscr{H}^{\varepsilon}} = 1$, such that

$$\|\tilde{U}^{\varepsilon} - \alpha^{\varepsilon} V\|_{H^1(\Omega)} + \|\tilde{\mathfrak{u}}^{\varepsilon} - \alpha^{\varepsilon} v_0\|_{H^1(\omega)} \le C\varepsilon^{\min(2-\vartheta, m-2-\vartheta, 1)}, \tag{6.45}$$

where $\tilde{\mathfrak{u}}^{\varepsilon}(\zeta, \tau) \equiv \tilde{u}^{\varepsilon}(x)$ for $(\zeta, \tau) \in \omega$, and α^{ε} is a well determined constant (see (6.46) and (6.55)), $\alpha^{\varepsilon} \to \lambda_0^{-1/2}$, as $\varepsilon \to 0$. As a consequence, $\tilde{U}^{\varepsilon}$ ($\tilde{\mathfrak{u}}^{\varepsilon}$, respectively) converge towards αV (αv_0, respectively) in $H^1(\Omega)$ ($H^1(\omega)$, respectively) as $\varepsilon \to 0$, the constant α being $\alpha = \lambda_0^{-1/2}$.

Proof Let $\lambda_0, \lambda_2, y_0, y_1, V$, and v_0 be as the theorem states. Let us define $v_1(\zeta, \tau) = \varkappa_0 y_1(\zeta) V(0, \tau)$ for $(\zeta, \tau) \in \omega$. Note that $v_1 \in H^1(\omega)$ verifies (6.31)–(6.33) for $\lambda_1 = \frac{a\varkappa_0}{h}$. Let us consider $v_2 \in H^1(\omega)$ satisfying periodic conditions on $\tau = 0$ and $\tau = \ell$ and verifying problem (6.37)–(6.39); v_1, v_2 are determined uniquely by prescribing the orthogonality conditions

$$\int_0^h v_i(\zeta, \tau) y_0(\zeta)\, d\zeta = 0 \quad \text{for } i = 1, 2.$$

Note that v_0, v_1, v_2, and V are smooth functions, in particular, $v_0, v_1, v_2 \in H^2(\omega)$ and $V \in H^2(\Omega)$.

For sufficiently small ε, we consider the function $\{W^{\varepsilon}, w^{\varepsilon}\}$ defined by

$$\begin{cases} W^{\varepsilon}(x) = V(x) + \varepsilon P v_1(x) + \varepsilon^2 P v_2(x) & \text{if } x \in \Omega, \\ w^{\varepsilon}(\nu, \tau) = v_0(\nu/\varepsilon, \tau) + \varepsilon v_1(\nu/\varepsilon, \tau) + \varepsilon^2 v_2(\nu/\varepsilon, \tau) & \text{if } 0 \le \nu \le \varepsilon h,\ \tau \in \mathbb{S}_\ell, \end{cases} \tag{6.46}$$

where $P : H^2(\omega) \to H^2(\Omega)$ is the continuous operator such that, for any $v \in H^2(\omega)$, Pv is a harmonic function and $(Pv)|_{\Gamma} = v(0, \tau)$. It is clear that $\{W^{\varepsilon}, w^{\varepsilon}\} \in H^1_0(\Omega_\varepsilon)$. In order to apply Lemma 1 we first prove the estimate

$$\begin{aligned} &\left| \left(\mathscr{A}^{\varepsilon} \{\tilde{W}^{\varepsilon}, \tilde{w}^{\varepsilon}\} - \frac{1}{\lambda_0 + \varepsilon\lambda_1 + \varepsilon^2 \lambda_2} \{\tilde{W}^{\varepsilon}, \tilde{w}^{\varepsilon}\}, \{W, w\} \right)_{\mathscr{H}^{\varepsilon}} \right| \\ &\quad \le C \left(\varepsilon^3 + \varepsilon^{m-1} \right) \|\{W, w\}\|_{\mathscr{H}^{\varepsilon}} \quad \forall \{W, w\} \in \mathscr{H}^{\varepsilon}, \end{aligned} \tag{6.47}$$

where $\{\tilde{W}^{\varepsilon}, \tilde{w}^{\varepsilon}\} = \{W^{\varepsilon}, w^{\varepsilon}\} \|\{W^{\varepsilon}, w^{\varepsilon}\}\|^{-1}_{\mathscr{H}^{\varepsilon}}$ and C is a constant independent of ε.

Considering definitions of $\mathscr{A}^\varepsilon$ and the scalar product $(\cdot,\cdot)_{\mathscr{H}^\varepsilon}$ in (6.43), we write

$$\begin{aligned}(\lambda_0+\varepsilon\lambda_1+\varepsilon^2\lambda_2)&\left(\mathscr{A}^\varepsilon\{W^\varepsilon,w^\varepsilon\}-\frac{1}{\lambda_0+\varepsilon\lambda_1+\varepsilon^2\lambda_2}\{W^\varepsilon,w^\varepsilon\},\{W,w\}\right)_{\mathscr{H}^\varepsilon}\\ &=(\lambda_0+\varepsilon\lambda_1+\varepsilon^2\lambda_2)\left(\varepsilon^m\int_\Omega W^\varepsilon W\,dx+\frac{1}{\varepsilon}\int_{\omega_\varepsilon} w^\varepsilon w\,dx\right)\\ &\quad-\varepsilon^2 A\int_\Omega \nabla_x W^\varepsilon\cdot\nabla_x W\,dx-\varepsilon a\int_{\omega_\varepsilon}\nabla_x w^\varepsilon\cdot\nabla_x w\,dx. \end{aligned}\tag{6.48}$$

For the integrals in ω_ε, we perform the change of variables (6.16), and we denote by ω the band $(0,h)\times\mathbb{S}_\ell$. Let $K_\varepsilon(\zeta,\tau)$ be the Jacobian of the transformation from (x_1,x_2) to (ν,τ) in the (ζ,τ) variables, namely $K_\varepsilon(\zeta,\tau)=1+\varepsilon\zeta\varkappa_0$. Thus, the integrals

$$\frac{1}{\varepsilon}\int_{\omega_\varepsilon} w^\varepsilon w\,dx\quad\text{and}\quad\varepsilon\int_{\omega_\varepsilon}\nabla_x w^\varepsilon\cdot\nabla_x w\,dx\tag{6.49}$$

in (6.48) read

$$\int_\omega w^\varepsilon w\,K_\varepsilon\,d\zeta d\tau\quad\text{and}\quad\int_\omega\partial_\zeta w^\varepsilon\partial_\zeta w K_\varepsilon d\zeta d\tau+\varepsilon^2\int_\omega\partial_\tau w^\varepsilon\partial_\tau w\,K_\varepsilon^{-1}\,d\zeta d\tau\tag{6.50}$$

respectively, where now w^ε and w are written in the variables (ζ,τ).

Now, taking into account the definition (6.46) of $\{W^\varepsilon,w^\varepsilon\}$, the change (6.16), and the formulas above (6.49) and (6.50), we consider the decomposition

$$\begin{aligned}&(\lambda_0+\varepsilon\lambda_1+\varepsilon^2\lambda_2)\left(\mathscr{A}^\varepsilon\{W^\varepsilon,w^\varepsilon\}-\frac{1}{\lambda_0+\varepsilon\lambda_1+\varepsilon^2\lambda_2}\{W^\varepsilon,w^\varepsilon\},\{W,w\}\right)_{\mathscr{H}^\varepsilon},\\ &=I_1+I_2+I_3+I_4+I_5+I_6\end{aligned}$$

where

$$I_1=\lambda_0\int_\omega v_0 w\,d\zeta d\tau-a\int_\omega\partial_\zeta v_0\partial_\zeta w\,d\zeta d\tau\,,$$

$$\begin{aligned}I_2=\varepsilon\Big(&\lambda_0\int_\omega v_0 w\zeta\varkappa_0\,d\zeta d\tau+\lambda_0\int_\omega v_1 w\,d\zeta d\tau+\lambda_1\int_\omega v_0 w\,d\zeta d\tau\\ &-a\int_\omega\partial_\zeta v_0\partial_\zeta w\zeta\varkappa_0\,d\zeta d\tau-a\int_\omega\partial_\zeta v_1\partial_\zeta w\,d\zeta d\tau\Big)\,,\end{aligned}$$

$$I_3 = \varepsilon^2 \left(\lambda_0 \int_\omega v_1 w \zeta \varkappa_0 \, d\zeta d\tau + \lambda_0 \int_\omega v_2 w \, d\zeta d\tau + \lambda_1 \int_\omega v_0 w \zeta \varkappa_0 \, d\zeta d\tau \right)$$
$$+ \lambda_1 \int_\omega v_1 w \, d\zeta d\tau + \lambda_2 \int_\omega v_0 w \, d\zeta d\tau - A \int_\Omega \nabla V \cdot \nabla W \, dx$$
$$- a \int_\omega \partial_\tau v_0 \partial_\tau w \, d\zeta d\tau - a \int_\omega \partial_\zeta v_1 \partial_\zeta w \zeta \varkappa_0 \, d\zeta d\tau - a \int_\omega \partial_\zeta v_2 \partial_\zeta w \, d\zeta d\tau \Big),$$

$$I_4 = \varepsilon^3 \Big(\lambda_0 \int_\omega v_2 w \zeta \varkappa_0 \, d\zeta d\tau + \lambda_1 \int_\omega v_1 w \zeta \varkappa_0 \, d\zeta d\tau + \lambda_2 \int_\omega v_0 w \zeta \varkappa_0 \, d\zeta d\tau$$
$$+ \lambda_1 \int_\omega v_2 w K_\varepsilon \, d\zeta d\tau + \lambda_2 \int_\omega (v_1 + \varepsilon v_2) w K_\varepsilon \, d\zeta d\tau - a \int_\omega \partial_\zeta v_2 \partial_\zeta w \zeta \varkappa_0 \, d\zeta d\tau \Big),$$

$$I_5 = -a\varepsilon^2 \int_\omega \partial_\tau v_0 \partial_\tau w (K_\varepsilon^{-1} - 1) \, d\zeta d\tau - a\varepsilon^3 \int_\omega \partial_\tau (v_1 + \varepsilon v_2) \partial_\tau w K_\varepsilon^{-1} \, d\zeta d\tau,$$

$$I_6 = (\lambda_0 + \varepsilon \lambda_1 + \varepsilon^2 \lambda_2) \varepsilon^m \int_\Omega W^\varepsilon W \, dx - A\varepsilon^3 \int_\Omega \nabla_x (P v_1 + \varepsilon P v_2) \cdot \nabla_x W \, dx \, .$$

Then, we prove the estimate (6.47) for each I_i above.

Indeed, the fact that v_0, v_1, v_2, V satisfy (6.23)–(6.26), (6.31)–(6.33), (6.37)–(6.39), and (6.41), respectively, leads us to $I_1 = I_2 = I_3 = 0$. In addition, on account of (6.43), the change of variables (6.16) (see (6.49) and (6.50)) and the boundary condition on Γ_ε, we have

$$|I_4| \le C\varepsilon^3 \|\{W, w\}\|_{\mathscr{H}^\varepsilon}.$$

As regards I_5, integrating by parts in ω and taking into account the smoothness of v_0, v_1, and v_2 and the definition of K_ε and of the scalar product (6.43) yields

$$|I_5| \le C\varepsilon^3 [\|\partial_\tau^2 v_0\|_{L^2(\omega)} \|w\|_{L^2(\omega)} + \|\partial_\tau v_0\|_{L^2(\omega)} \|w\|_{L^2(\omega)} + \|\partial_\tau^2 v_1\|_{L^2(\omega)} \|w\|_{L^2(\omega)} +$$
$$\|\partial_\tau v_1\|_{L^2(\omega)} \|w\|_{L^2(\omega)} + \|\partial_\tau v_2\|_{L^2(\omega)} \|\partial_\tau w\|_{L^2(\omega)}] \le C\varepsilon^3 \|\{W, w\}\|_{\mathscr{H}^\varepsilon}.$$

In order to obtain bounds for I_6, we take into account the fact that $W = w$ on Γ and $w = 0$ on Γ_ε, and the trace inequalities

$$\|W\|^2_{L^2(\Gamma)} = \Big(\int_\Gamma \int_0^h \partial_\nu w \, d\nu d\tau \Big)^2 \le C\varepsilon \|\nabla_x w\|^2_{L^2(\omega_\varepsilon)} \le C \|\{W, w\}\|^2_{\mathscr{H}^\varepsilon}$$
$$\forall \{W, w\} \in H_0^1(\Omega_\varepsilon). \tag{6.51}$$

and

$$\|\partial_v U\|_{L^2(\Gamma)} \leq C\|U\|_{H^2(\Omega)} \quad \forall U \in H^2(\Omega). \tag{6.52}$$

Then, integrating by parts in Ω and using definition of the operator P, the fact that $W = w$ on Γ, the definition (6.43), formula (6.51) and (6.52), we obtain

$$\begin{aligned} |I_6| &\leq C[\varepsilon^m \|W^\varepsilon\|_{L^2(\Omega)}\|W\|_{L^2(\Omega)} + \varepsilon^3\|\partial_v(Pv_1)\|_{L^2(\Gamma)}\|W\|_{L^2(\Gamma)} \\ &\quad +\varepsilon^4\|\nabla_x(Pv_2)\|_{L^2(\Omega)}\|\nabla_x W\|_{L^2(\Omega)}] \\ &\leq C(\varepsilon^{m-1} + \varepsilon^3)\|\{W, w\}\|_{\mathscr{H}^\varepsilon}. \end{aligned}$$

Finally, considering the local coordinates (6.16), we verify that $\|\{W^\varepsilon, w^\varepsilon\}\|^2_{\mathscr{H}^\varepsilon} \to \|\partial_\zeta v_0\|^2_{L^2(\omega)} = \lambda_0$ as $\varepsilon \to 0$, and we have proved that the estimate (6.47) holds for sufficiently small ε.

Now, we apply Lemma 1 for $H = \mathscr{H}^\varepsilon$, $A = \mathscr{A}^\varepsilon$, $\lambda = (\lambda_0 + \varepsilon\lambda_1 + \varepsilon^2\lambda_2)^{-1}$ and $u = \{\tilde{W}^\varepsilon, \tilde{w}^\varepsilon\}$ and $r = C(\varepsilon^3 + \varepsilon^{m-1})$ which provides, for sufficiently small ε, at least one eigenvalue $\lambda^\varepsilon_{k(\varepsilon)}$ of (6.1) verifying $|(\lambda^\varepsilon_{k(\varepsilon)}\varepsilon^{2-m})^{-1} - (\lambda_0 + \varepsilon\lambda_1 + \varepsilon^2\lambda_2)^{-1}| \leq C(\varepsilon^3 + \varepsilon^{m-1})$, and consequently, we deduce (6.44). Moreover, if we take, for instance, $r^* = \varepsilon^\vartheta$ with $0 < \vartheta < \min(2, m-2)$, Lemma 1 also provides a function $\{\tilde{U}^\varepsilon, \tilde{u}^\varepsilon\} \in \mathscr{H}^\varepsilon$, with $\|\{\tilde{U}^\varepsilon, \tilde{u}^\varepsilon\}\|_{\mathscr{H}^\varepsilon} = 1$, $\{\tilde{U}^\varepsilon, \tilde{u}^\varepsilon\}$ belonging to the eigenspace associated with all the eigenvalues $(\lambda^\varepsilon_{k(\varepsilon)}\varepsilon^{2-m})^{-1}$ of $\mathscr{A}^\varepsilon$ contained in

$$[(\lambda_0 + \varepsilon\lambda_1 + \varepsilon^2\lambda_2)^{-1} - \varepsilon^\vartheta, (\lambda_0 + \varepsilon\lambda_1 + \varepsilon^2\lambda_2)^{-1} + \varepsilon^\vartheta], \tag{6.53}$$

such that

$$\|\{\tilde{U}^\varepsilon, \tilde{u}^\varepsilon\} - \alpha^\varepsilon\{W^\varepsilon, w^\varepsilon\}\|_{\mathscr{H}^\varepsilon} \leq C(\varepsilon^{3-\vartheta} + \varepsilon^{m-1-\vartheta}) \tag{6.54}$$

is satisfied where

$$\alpha^\varepsilon = \|\{W^\varepsilon, w^\varepsilon\}\|^{-1}_{\mathscr{H}^\varepsilon}. \tag{6.55}$$

Now, from (6.46) and (6.43) and (6.16), we conclude that, for $m > 2$,

$$\|\tilde{\mathfrak{u}}^\varepsilon - \alpha^\varepsilon v_0\|_{L^2(\omega)} + \|\partial_\zeta(\tilde{\mathfrak{u}}^\varepsilon - \alpha^\varepsilon v_0)\|_{L^2(\omega)} \leq C\varepsilon^{\min(3-\vartheta, m-1-\vartheta)}$$

and

$$\|\partial_\tau(\tilde{\mathfrak{u}}^\varepsilon - \alpha^\varepsilon v_0)\|_{L^2(\omega)} + \|\nabla_x(\tilde{U}^\varepsilon - \alpha^\varepsilon V)\|_{L^2(\Omega)} \leq C\varepsilon^{\min(2-\vartheta, m-2-\vartheta, 1)}.$$

Finally, since $\tilde{U}^\varepsilon|_\Gamma = \tilde{\mathfrak{u}}^\varepsilon(0, \tau)$ and $V|_\Gamma = v_0(0, \tau)$, Friedrichs' inequality for $\tilde{U}^\varepsilon$ in Ω and the trace inequality for $\tilde{\mathfrak{u}}^\varepsilon$ in ω lead us to assert estimate (6.45) and

$\tilde{U}^\varepsilon$ ($\tilde{\mathfrak{u}}^\varepsilon$, respectively) converge towards αV (αv_0, respectively) in $H^1(\Omega)$ ($H^1(\omega)$, respectively) as $\varepsilon \to 0$ being $\alpha = \lambda_0^{-1/2}$. Therefore, the theorem is proved.

6.5 High Frequencies

As occurs in many singularly perturbed problems (see, for instance, [LoPe97, GoEtAl99a, GoEtAl04, LoEtAl05, GoEtA111, GoEtAl20]), there are sequences of eigenvalues of (6.1), $\lambda^\varepsilon = \lambda^\varepsilon_{k(\varepsilon)}$ with $k(\varepsilon) \to \infty$, of order ε^β for some $\beta < m - 2$, whose corresponding eigenfunctions suitably normalized do not vanish asymptotically. Here, we focus our attention on the eigenvalues of (6.1) of order 1, the so-called high frequencies.

Throughout this section we consider the case where $m > 0$. We first obtain the limiting problem associated with the eigenvalues λ^ε of (6.1) of order 1 by means of asymptotic expansions. Later on, we show that the eigenvalues λ^ε asymptotically close to eigenvalues of the Dirichlet problem in Ω give rise to global vibrations in the way stated by Theorem 2. It should be noted that convergence results hold for all $m > 0$, while some restrictions and extensions for the asymptotic expansions for certain values of m are in Remark 3.

For $m > 2$ (see Remark 3 for $m \in (0, 2]$), we assume an asymptotic expansion for the eigenvalues λ^ε and for the corresponding eigenfunctions $\{U^\varepsilon, u^\varepsilon\}$ in Ω and ω_ε of the form:

$$\lambda^\varepsilon = \lambda_0 + \varepsilon\lambda_1 + \varepsilon^2\lambda_2 + \cdots \tag{6.56}$$

$$U^\varepsilon(x) = V(x) + \varepsilon \mathsf{V}_1(x) + \varepsilon^2 \mathsf{V}_2(x) + \cdots, \qquad x \in \Omega, \tag{6.57}$$

$$u^\varepsilon(\zeta, \tau) = v_0(\zeta, \tau) + \varepsilon v_1(\zeta, \tau) + \varepsilon^2 v_2(\zeta, \tau) + \cdots, \qquad \zeta \in [0, h), \tau \in \mathbb{S}_\ell, \tag{6.58}$$

respectively, where (ζ, τ) are the local coordinates given by (6.16), and v_i are ℓ–periodic functions in τ. Besides, we suppose that at least one of the functions V or v_0 in (6.57) and (6.58) is different from zero.

By replacing (6.56), (6.57), and (6.58) in (6.1), on account of (6.18), we have that the leading terms in the asymptotic expansions satisfy the equations

$$\begin{aligned} -A\Delta_x V &= \lambda_0 V \quad \text{in } \Omega, \\ 0 &= \lambda_0 v_0, \quad \zeta \in (0, h), \tau \in \mathbb{S}_\ell, \\ V &= v_0 \quad \text{on } \Gamma. \end{aligned}$$

Hence, $\lambda_0 = 0$ or $v_0 \equiv 0$. Since we are dealing with the eigenvalues of order 1, we consider the case where $\lambda_0 \neq 0$, and consequently we have that (λ_0, V) is an

eigenpair of the Dirichlet problem

$$\begin{cases} -A\Delta_x V = \lambda_0 V & \text{in } \Omega, \\ V = 0 & \text{on } \Gamma. \end{cases} \tag{6.59}$$

As is known, problem (6.59) has a discrete spectrum, which can be computed in case the domain Ω is the disk with radius R centered at the origin using polar coordinates and separation of variables. Indeed, the eigenvalues are given by $\lambda_0 = A\eta_{k,j}^2 R^{-2}$ where $\{\eta_{k,j}\}_{j=1}^{\infty}$ zeros of the Bessel functions $J_k(s)$ of first kind with $k \in \mathbb{N}_0$. The corresponding eigenfunctions are

$$V(r,\theta) = \alpha J_k(\sqrt{A^{-1}\lambda_0}\, r)\sin(k\theta) = \alpha J_k(\eta_{k,j}R^{-1}r)\sin(k\theta), \quad k \in \mathbb{N},$$

$$V(r,\theta) = \alpha J_k(\sqrt{A^{-1}\lambda_0}\, r)\cos(k\theta) = \alpha J_k(\eta_{k,j}R^{-1}r)\cos(k\theta), \quad k \in \mathbb{N}_0,$$

for $(r,\theta) \in [0, R) \times [0, 2\pi)$, α being any constant (compare these expressions with formulae (6.14) and (6.15), respectively, in the fixed domain $[0, R) \times [0, 2\pi)$).

As outlined for the asymptotics of the eigenfunctions corresponding to the low frequencies, an appropriate normalization for the eigenfunctions must be prescribed to obtain convergence for the high frequencies. We denote by $\mathfrak{H}^\varepsilon$ the space $H_0^1(\Omega_\varepsilon)$ with the scalar product

$$(\{W, w\}, \{G, g\})_{\mathfrak{H}^\varepsilon} = A\int_\Omega \nabla_x W \cdot \nabla_x G\, dx + \frac{a}{\varepsilon^t}\int_{\omega_\varepsilon} \nabla_x w \cdot \nabla_x g\, dx \tag{6.60}$$

for all $\{W, w\}, \{G, g\} \in H_0^1(\Omega_\varepsilon)$.

Next, we state the convergence of sequences of eigenvalues of (6.1) towards those of (6.59) and give bounds for the convergence rates for the eigenvalues and eigenfunctions (cf. (6.61) and (6.62), respectively). The proof of this result can be found in [GoEtAl20] in a much more general framework.

Theorem 2 *Let (λ_0, V) be an eigenpair of the Dirichlet problem (6.59) such that $\|V\|_{L^2(\Omega)} = 1$. Then, for $m > 0$ and $t \geq 1$, there are eigenvalues $\lambda^\varepsilon_{k(\varepsilon)}$ of problem (6.1) such that*

$$|\lambda^\varepsilon_{k(\varepsilon)} - \lambda_0| \leq C\varepsilon, \tag{6.61}$$

where C is a constant independent of ε. In addition, there is a linear combination of eigenfunctions $\{\widetilde{U}^\varepsilon, \widetilde{u}^\varepsilon\} \in H_0^1(\Omega_\varepsilon)$, $\{\widetilde{U}^\varepsilon, \widetilde{u}^\varepsilon\}$ corresponding to the eigenvalues $\lambda^\varepsilon_{k(\varepsilon)}$ of (6.1) in the segment $[\lambda_0 - K\varepsilon^\vartheta, \lambda_0 + K\varepsilon^\vartheta]$ with $K > 0$ and $0 < \vartheta < 1$, $\|\{\widetilde{U}^\varepsilon, \widetilde{u}^\varepsilon\}\|_{\mathfrak{H}^\varepsilon} = 1$, such that

$$\|\widetilde{U}^\varepsilon - \lambda_0^{-1/2} V\|_{H^1(\Omega)} \leq C\varepsilon^{1-\vartheta}. \tag{6.62}$$

Remark 3 It should be noted that the technique of asymptotic expansions throughout this section also applies in the case where $m \in (0, 2)$ and we obtain the same limit problem (6.59). In this case we need to use further terms of the asymptotic expansions of u^ε in ω_ε. As a matter of fact, for $m \neq 1$ the expansion (6.21) must be suitably modified by introducing other terms for different powers of ε, namely of the order ε^p, with $p > 0$, $p \notin \mathbb{N}$, depending on the particular value of m. Moreover, for $m \in (0, 2)$, the convergence of the kth eigenvalue of (6.2), when $\varepsilon \to 0$, towards the kth eigenvalue of (6.59) holds following the technique in [GoEtAl11] (see also [GoEtAl04]).

In the case where $m = 2$, the asymptotic expansions (6.56)–(6.58) and (6.19)–(6.21) provide two possibilities for λ_0 that we state here without a proof. One is λ_0 to be an eigenvalue of (6.59) and the other is λ_0 to be an eigenvalue of (6.27). However, it remains to identify the eigenfunctions in (6.57), (6.58) and (6.20), (6.21) which involve different normalization (see norms (6.43) and (6.60) to compare). This case remains as an open problem.

Acknowledgments This work has partially been supported by the Spanish MICINN grant PGC2018-098178-B-I00 and by Russian Foundation of Basic Research 18-01-00325.

References

[BiSo87] Birman, M.S., Solomyak, M.Z.: Spectral Theory of Self-Adjoint Operators in Hilbert Space. Reidel Publ. Company, Dordrecht (1987)

[CaEtAl05] Caínzos, J., Vilasánchez, M., Pérez, E.: Comportamiento asintótico de soluciones de un problema espectral de Neumann (1-D) con masa concentrada. In: XIX CEDYA / IX Congress of Applied Mathematics, Universidad Carlos III de Madrid, Madrid, p. 5 (2005)

[GoEtAl04] Golovaty, Y.D., Gómez, D., Lobo, M., Pérez, E.: On vibrating membranes with very heavy thin inclusions. Math. Models Methods Appl. Sci. **14**(7), 987–1034 (2004)

[GoEtAl98] Gómez, D., Lobo, M., Pérez, E.: Sobre vibraciones de un sistema con una masa concentrada. In: XV CEDYA/V Congress of Applied Mathematics, Universidad de Vigo, Vigo, pp. 453–458 (1998)

[GoEtAl99a] Gómez, D., Lobo, M., Pérez, E.: On the eigenfunctions associated with the high frequencies in systems with a concentrated mass. J. Math. Pures Appl. **78**, 841–865 (1999)

[GoEtAl99b] Gómez, D., Lobo, M., & Pérez, E.: Vibrations of structures containing very heavy kernels: high frequencies. In: Vibration, Noise and Structural Dynamics'99, Staffordshire University, pp. 212–220 (1999)

[GoEtAl06a] Gómez, D., Lobo, M., Nazarov, S.A., Pérez, E.: Spectral stiff problems in domains surrounded by thin bands: asymptotic and uniform estimates for eigenvalues. J. Math. Pures Appl. **85**, 598–632 (2006)

[GoEtAl06b] Gómez, D., Lobo, M., Nazarov, S.A., Pérez, E.: Asymptotics for the spectrum of the Wentzell problem with a small parameter and other related stiff problems. J. Math. Pures Appl. **86**, 369–402 (2006)

[GoEtAl11] Gómez, D., Nazarov, S.A., Pérez. E.: Spectral stiff problems in domains surrounded by thin stiff and heavy bands: local effects for eigenfunctions. Netw. Heterog. Media **6**, 1–35 (2011)

[GoEtAl19] Gómez, D., Navazo-Esteban, S., Pérez-Martínez, M.-E.: A stiff problem: stationary waves and approximations. In: Constanda, C., Harris, P. (eds.) Integral Methods in Science and Engineering. Analytic Treatment and Numerical Approximations, Birhäuser, Switzerland (2019), pp. 133–148.

[GoEtAl20] Gómez, D., Nazarov, S.A., Pérez-Martínez, M.-E.: Localization effects for Dirichlet problems in domains surrounded by thin stiff and heavy bands (Submitted)

[LoPe97] Lobo, M., Pérez, E.: High frequency vibrations in a stiff problem. Math. Models Methods Appl. Sci. **7**, 291–311 (1997)

[LoEtAl03] Lobo, M., Nazarov, S., Pérez, E.: Asymptotically sharp uniform estimates in a scalar stiff problem. Comptes Rendues de Mecanique **331**, 325–330 (2003)

[LoEtAl05] Lobo, M., Nazarov, S., Pérez, E.: Eigenoscillations of contrasting nonhomogeneous elastic bodies. Asymptotic and uniform estimates for eigenvalues. IMA J. Appl. Math. **70**, 419–458 (2005)

[Na02] Nazarov, S.A.: Asymptotic Theory of Thin Plates and Rods. Vol.1. Dimension Reduction and Integral Estimates (in Russian). Nauchnaya Kniga, Novosibirck (2002)

[Na03] Nazarov, S.A: Uniform estimates of remainders in asymptotic expansions of solutions to the problem on eigen-oscillations of a piezoelectric plate. Probl. Mat. Analiz. **25**, 99–188 (2003). (English transl.: J. Math. Sci., 114(5):1657–1725, 2003)

[Pe95] Pérez, E.: Altas frecuencias en un problema "stiff" relativo a las vibraciones de una cuerda. In: XIV CEDYA/IV Congress of Applied Mathematics, Universidad de Barcelona, Barcelona (1995), 7pp. (electronic)

[Pe03] Pérez, E.: On the whispering gallery modes on interfaces of membranes composed of two materials with very different densities. Math. Models Methods Appl. Sci. **13**, 75–98 (2003)

[ViLy57] Vishik, M.I., Lyusternik. L.A.: Regular degeneration and boundary layer for linear differential equations with small parameter. Uspekhi Mat. Nauk. **12**, 3–122 (1957). (English transl.: Am. Math. Soc. Transl. Ser. 2, 20:239–364, 1962)

Chapter 7
Spectral Homogenization Problems in Linear Elasticity with Large Reaction Terms Concentrated in Small Regions of the Boundary

Delfina Gómez, Sergey A. Nazarov, and Maria-Eugenia Pérez-Martínez

7.1 Introduction

In this paper, we study the asymptotic behavior of the vibrations of an elastic body which has very large surface reaction terms concentrated in small regions. We assume that the elastic material fills the domain Ω of the upper half-space $\mathbb{R}^{3+}$, and a part Σ of its surface lies on the plane $\{x_3 = 0\}$ and contains small regions T^ε of size r_ε, at a distance ε between them (cf. Fig. 7.1). The boundary conditions are of Winkler–Robin type on T^ε. Outside, the surface Σ is traction-free while the rest of the surface $\partial\Omega \setminus \Sigma$ is assumed to be fixed. Here ε and r_ε are two small parameters $r_\varepsilon \ll \varepsilon \ll 1$.

As is well known, from the mechanical viewpoint, the small regions behave as "springs" and the elastic coefficients of these springs are defined through the so-called *Robin reaction matrix*, which we denote by $\beta(\varepsilon)M(x)$. Matrix $M(x)$ depends on the point where *the reaction regions* T^ε are placed, while the parameter $\beta(\varepsilon)$, which is referred to as *the reaction parameter*, can range from very small to very large. The reaction regions T^ε are assumed to be domains of the plane $\mathbb{R}^2$ homothetics of a fixed domain T, with a Lipschitz boundary. Analyzing the different relations between the three parameters of the problem, ε, r_ε and $\beta(\varepsilon)$, is crucial to detect several behaviors of the vibrations of the structure. We consider the associated spectral problem and we address the asymptotic behavior of the eigenvalues and

D. Gómez · M.-E. Pérez-Martínez (✉)
Universidad de Cantabria, Santander, Spain
e-mail: gomezdel@unican.es; meperez@unican.es

S. A. Nazarov
Saint-Petersburg State University, St. Petersburg, Russia

Institute of Problems of Mechanical Engineering RAS, St. Petersburg, Russia
e-mail: srgnazarov@yahoo.co.uk

C. Constanda (ed.), *Computational and Analytic Methods in Science and Engineering*, https://doi.org/10.1007/978-3-030-48186-5_7

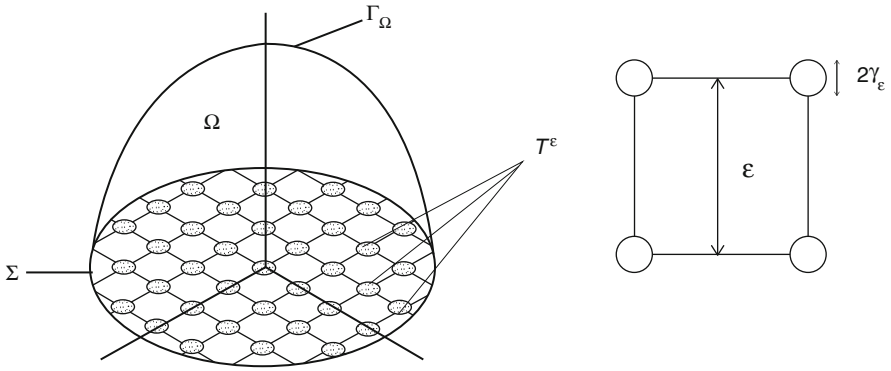


Fig. 7.1 Geometrical configuration of the problem

eigenfunctions when $\varepsilon \to 0$. Namely, we obtain spectral homogenized problems depending on the relations between ε, r_ε and $\beta(\varepsilon)$.

The stationary problem, for an isotropic homogeneous media, cf. (7.54), and a surface Σ which is stuck to the plane along the regions T^ε, has been studied in [LoPe87, LoPe88] and [BrEtAl90], where a *critical size* $O(\varepsilon^2)$ of the stuck regions appears (cf. (7.1) with $r_0 > 0$), which is somehow *classical* in the literature of the applied mathematics. For this size, the asymptotic behavior of the solution is intermediate between the extreme cases. Namely, for $r_\varepsilon \gg \varepsilon^2$ the stuck regions are large enough and the body behaves as if the whole Σ is stuck to the plane; for $r_\varepsilon \ll \varepsilon^2$ the stuck regions are very small and the surface behaves as if it would be free, while for $r_\varepsilon = O(\varepsilon^2)$ a spring/Robin boundary condition is asymptotically imposed as an intermediate condition between Dirichlet and Neumann. It contains the so-called *strange term* and links stresses and displacements, the elastic coefficients of this spring being given by a constant matrix, the so-called *capacity matrix* (cf. (7.20) and (7.21) for $W^{l,x} \equiv W^l$ independent of x).

Here, we deal with a different problem, and obtain the above-mentioned homogenized problems for a particular relation between the parameters. As a matter of fact, in addition to the critical size, a *critical relation for parameters* appears (cf. (7.3) with $\beta^* > 0$) which also provides the asymptotic behavior of solutions different from extreme cases. Now, several kinds of capacity matrices arise, which are obtained from the microstructure of the problem and depend on the macroscopic variable. This dependence is due to both the nonhomogeneous media Ω and the non-constant Robin matrix $M(x)$.

Notice that other different boundary homogenization problems in linear elasticity have been addressed in the literature. Let us mention [NgSa85] and [GrMiOr15], which consider stationary homogenization problems for the elasticity system in a perforated media along a plane, the size r_ε of the perforations in the plane being $r_\varepsilon = O(\varepsilon)$. Also, [BrLoPe90, LoPe92] and [JaAdBr00] consider cylindrical bodies, the regions where the displacements vanish being thin bands rolled around the body;

a different critical size and capacity matrix appear on account of the geometrical configuration of the domain. For the case of a certain non-periodical distribution of the regions T^ε, for extreme cases, let us mention [OlCh93] and [OlCh96]. For a strongly oscillating boundary, see [Na08]. For the Stokes fluid problem in a perforated domain along a plane, let us mention the works [Al83, Br92, BrEtAl16] and [GoEtAl18] where, also, a so-called *Stokes capacity matrix* appears on the transmission condition on Σ when $r_\varepsilon = O(\varepsilon^2)$; see also [SaSa82] for various effects on the perforated wall. We mention [Sa85] and [Co87] for fluids flow in media with perforated walls when $r_\varepsilon = O(\varepsilon)$. For critical relations between parameters in several fluid homogenization problems, see, e.g., [Co85] and [CiDoEn96]; also, see [GoEtAl18] and [GoLoPe19] for further references in this connection.

Other papers addressing homogenization problems for the elasticity operator, with the same geometrical configuration here considered, are [IoOnVel05] when $r_\varepsilon \ll \varepsilon$ and [GoNaPe18] for $r_\varepsilon = O(\varepsilon)$. Both consider spectral problems with alternating boundary conditions of Steklov type and, consequently, they differ greatly from the problem here considered, the results also being very different.

All these works belong to a large class of boundary homogenization problems studied for a long time in the literature of applied mathematics for several operators. We mention a few related to scalar problems such as [Sa82, Mu85, De87, Pi87, CiMu97, LoEtAl97, MaKh06, GoPeSh12] and [GoEtAl19], some of which have introduced keywords such as critical sizes and strange terms (see [CiMu97, MaKh06] and [GoEtAl19, GoPeSh12] for further references, in this connection). [GoPeSh12, GoPeSh13] and [GoEtAl13] also address spectral problems for the Laplacian with large parameters on the boundary of the perforations. See [GoPeSh12, GoEtAl18, GoLoPe19] and [GoEtAl19] for an extensive bibliography on homogenization for perforated domains along manifolds and large adsorption parameters.

Let us introduce parameters r_0, β^0 and β^* which play an important role in the description of the homogenized problems. They are defined by three limits:

$$\lim_{\varepsilon\to 0} \frac{r_\varepsilon}{\varepsilon^2} = r_0, \tag{7.1}$$

$$\lim_{\varepsilon\to 0} r_\varepsilon \beta(\varepsilon) = \beta^0, \tag{7.2}$$

and

$$\lim_{\varepsilon\to 0} \frac{\beta(\varepsilon) r_\varepsilon^2}{\varepsilon^2} = \beta^*. \tag{7.3}$$

In the case where $r_0 > 0$, we deal with the *classical critical size* of the reaction regions T^ε mentioned above. Equation (7.2) provides a relation between sizes of reaction regions and the reaction parameter which is important to determine the *local problems*. The case where $\beta^* > 0$ is referred to as *critical relation between*

the parameters. It occurs when the total area of the reaction regions $O(\varepsilon^{-2}r_\varepsilon^2)$ multiplied by the parameter of reaction $\beta(\varepsilon)$ is of order 1.

The most critical situation happens when $r_0 > 0$ and $\beta^0 > 0$ which also amounts to $r_0 > 0$ and $\beta^* > 0$. In this case, the strange term has a character which is completely different from that obtained in the literature. It contains a so-called *extended capacity matrix* $\mathscr{C}^e(x)$, cf. (7.16), which depends on the Robin matrix $M(x)$ in a non-trivial way. To have an idea of this dependence, one may compare with the scalar case: although the homogenized problem remains linear, the dependence on $M(x)$ would be nonlinear (cf., e.g., [GoPeSh12] and [GoEtAl18]). Matrix $\mathscr{C}^e(x)$ also depends on the parameter β^0, cf. (7.16), and (7.17).

The rest of the critical relations between parameters, for which a spring type boundary condition intermediate between Dirichlet and Neumann is obtained, deal with $r_0 > 0$ and $\beta^0 = +\infty$ or $\beta^* > 0$ and $r_0 = +\infty$. The first one, $r_0 > 0$ and $\beta^0 = +\infty$ (also $\beta^* = +\infty$), asymptotically amounts to regions T^ε stuck to the plane because of the large reaction parameter and, consequently, the spring boundary condition ignores $M(x)$. It contains a new *capacity matrix* $\mathscr{C}(x)$, which depends on the macroscopic variable x, but it is independent of $M(x)$. When the media in the original problem (7.8) is isotropic and homogeneous, cf. (7.53), this matrix coincides with that obtained in [LoPe87] and [LoPe88]. The second relation $\beta^* > 0$ and $r_0 = +\infty$, always keeping $r_\varepsilon = O(\beta(\varepsilon)^{-1/2}\varepsilon)$, provides an averaged spring type condition on Σ where the Robin reaction matrix is $M(x)$ multiplied by the averaged constant $\beta^*|T|$, cf. (7.22). Let us refer to [NaSoSp10] for other extended capacity matrices in very different problems.

Finally, the structure of the paper is as follows. Section 7.2 contains the setting of the spectral homogenization problem and some a priori estimates for the eigenvalues which are useful for the asymptotic analysis. Section 7.3 contains the list of spectral homogenized problems with the corresponding stationary local problems (cf. Fig. 7.2); the macroscopic variable appears as a parameter in these local problems. We obtain all these problems in Sect. 7.4 using asymptotic expansions and matching principles. Section 7.5 addresses the setting of the homogenized and local problems, in the suitable Sobolev spaces, when the media is isotropic (see Remark 7.1, in this connection).

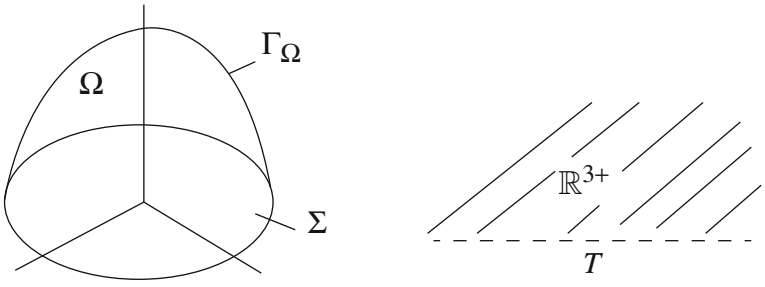


Fig. 7.2 The domains of setting for homogenized and local problems

7.2 Setting of the Problem

Let Ω be an open bounded domain of $\mathbb{R}^3$ situated in the upper half-space $\mathbb{R}^{3+} = \{x \in \mathbb{R}^3 : x_3 > 0\}$, with a Lipschitz boundary $\partial\Omega$. Let Σ be the part of the boundary in contact with the plane $\{x_3 = 0\}$ which is assumed to be non-empty and let Γ_Ω be the rest of the boundary of Ω: $\partial\Omega = \overline{\Gamma}_\Omega \cup \overline{\Sigma}$. Let T denote an open bounded domain of the plane $\{x_3 = 0\}$ with a Lipschitz boundary. Without any restriction, we can assume that both Σ and T contain the origin of coordinates.

Let ε be a small parameter $\varepsilon \ll 1$. Let r_ε be an order function such that $r_\varepsilon \ll \varepsilon$. For $\mathrm{k} = (k_1, k_2) \in \mathbb{Z}^2$, we denote by $\widetilde{x}^\varepsilon_\mathrm{k}$ the point of the plane $\{x_3 = 0\}$ of coordinates $\widetilde{x}^\varepsilon_\mathrm{k} = (k_1\varepsilon, k_2\varepsilon, 0)$, and by $T^\varepsilon_{\widetilde{x}_\mathrm{k}}$ the homothetic domain of T of ratio r_ε after translation to the point $\widetilde{x}^\varepsilon_\mathrm{k}$:

$$T^\varepsilon_{\widetilde{x}_\mathrm{k}} = \widetilde{x}^\varepsilon_\mathrm{k} + r_\varepsilon T\,.$$

If there is no ambiguity, we shall write $\widetilde{x}_\mathrm{k}$ instead of $\widetilde{x}^\varepsilon_\mathrm{k}$, and T^ε instead of $T^\varepsilon_{\widetilde{x}_\mathrm{k}}$ and $\widetilde{x}^\varepsilon_\mathrm{k}$ is referred to as the center of $T^\varepsilon_{\widetilde{x}_\mathrm{k}}$.

In this way, for a fixed ε, we have constructed a grid of squares in the plane $\{x_3 = 0\}$ whose vertices are the centers of the regions T^ε. Let $\mathscr{J}^\varepsilon$ denote $\mathscr{J}^\varepsilon = \{\mathrm{k} \in \mathbb{Z}^2 : T^\varepsilon_{\widetilde{x}_\mathrm{k}} \subset \Sigma\}$, while N_ε denotes the number of elements of $\mathscr{J}^\varepsilon$:

$$N_\varepsilon \approxeq \frac{|\Sigma|}{\varepsilon^2} = O(\varepsilon^{-2}).$$

Finally, if no confusion arises, we denote by $\bigcup T^\varepsilon$ the union of all the T^ε contained in Σ, namely,

$$\bigcup T^\varepsilon \equiv \bigcup_{\mathrm{k}\in\mathscr{J}^\varepsilon} T^\varepsilon_{\widetilde{x}_\mathrm{k}}\,.$$

Also, in what follows $x = (x_1, x_2, x_3)$ denotes the usual Cartesian coordinates, while by $\hat{x} = (x_1, x_2)$ we refer to the two first components of $x \in \mathbb{R}^3$.

The geometrical configuration in the plane is analogous to that in [Sa82, Mu85] for scalar problems and that in [LoPe87, LoPe88, BrEtAl90, IoOnVel05] and [GoNaPe18] for the elasticity system.

Under the basis that the domain Ω is filled by an elastic material, for $i, j, k, l = 1, 2, 3$, we denote by $a_{ijkl}(x)$ the elastic coefficients of the material, which are assumed to be continuous functions defined in $\overline{\Omega}$ and satisfy the standard symmetry and coercivity properties (cf., e.g., [OlShYo92] and [Te79])

$$a_{ijkl}(x) = a_{jikl}(x) = a_{klij}(x), \quad i, j, k, l = 1, 2, 3, \quad \forall x \in \overline{\Omega},$$

and

$$\exists \alpha_1 > 0 \,:\, a_{ijkl}(x)\xi_{ij}\xi_{kl} \geq \alpha_1 \xi_{ij}\xi_{ij}, \quad \forall \xi \ 3\times 3 - \text{matrix} : \xi_{ij} = \xi_{ji}, \\ i,j = 1,2,3, \quad \forall x \in \overline{\Omega}. \tag{7.4}$$

Also, for a given displacement vector $u(x) = (u_1(x), u_2(x), u_3(x))$ we use the standard notations for stress and strain tensors $\sigma(u)$ and $e(u)$; namely, we denote by $(\sigma_{ij}(u))_{i,j=1,2,3}$ the stress tensor which is related to the strain tensor $(e_{ij}(u))_{i,j=1,2,3}$ by the Hooke's law

$$\sigma_{ij}(u) = a_{ijkl}(x)e_{kl}(u), \tag{7.5}$$

where

$$e_{kl}(u) = \frac{1}{2}\Big(\frac{\partial u_k}{\partial x_l} + \frac{\partial u_l}{\partial x_k}\Big). \tag{7.6}$$

Above, and in what follows, we use the convention of summation convention over repeated indexes.

In connection with the reaction coefficients in the small regions T^ε, let us introduce a symmetric and positive definite 3×3-matrix, $M_{ij} \in C(\overline{\Sigma})$:

$$\exists \alpha_2 > 0 \,:\, M_{ij}(x_1,x_2,0)\xi_i\xi_j \geq \alpha_2 \xi_i^2, \quad \forall \xi \in \mathbb{R}^3, \quad \forall (x_1,x_2,0) \in \overline{\Sigma}. \tag{7.7}$$

Let us consider the spectral problem

$$\begin{cases} -\dfrac{\partial \sigma_{ij}^\varepsilon}{\partial x_j} = \lambda^\varepsilon u_i^\varepsilon & \text{in } \Omega\,, \\ u^\varepsilon = 0 & \text{on } \Gamma_\Omega\,, \\ \sigma_{ij}^\varepsilon n_j = 0 & \text{on } \Sigma \setminus \bigcup T^\varepsilon, \\ \sigma_{ij}^\varepsilon n_j + \beta(\varepsilon) M_{ij} u_j^\varepsilon = 0 & \text{on } \bigcup T^\varepsilon\,, \end{cases} \qquad i = 1,2,3 \tag{7.8}$$

where λ^ε denotes the spectral parameter, and $u^\varepsilon = (u_1^\varepsilon, u_2^\varepsilon, u_3^\varepsilon)$ the corresponding eigenvector. u^ε is related to stress and strain tensors by (7.5) and (7.6), respectively. In particular, in (7.8) we have denoted by

$$\sigma_{ij}^\varepsilon \equiv \sigma_{ij}(u^\varepsilon) = a_{ijkl}e_{kl}(u^\varepsilon),$$

while n stands for the unit outer normal to Ω along Σ, namely, $n = (0,0,-1)$. The parameter $\beta(\varepsilon)$ arising in the equations on T^ε, linking stresses and displacements, is a positive parameter which is referred to as *the Robin/Winkler coefficient of reaction* (the reaction parameter, in short). It can range from very large to very small or it can be of order 1.

Above, we represent a spectral problem associated with a Winkler bed (the foundation) composed by a block of an elastic material, which has a part of its boundary Γ_Ω clamped to a rigid support, while the part in contact with the plane $\{x_3 = 0\}$ lies partially on a series of "springs" with the elastic coefficients $\beta(\varepsilon)M(x)$. Outside these springs, *the reaction regions* $\bigcup T^\varepsilon$, the surface is traction-free, cf., e.g., [At84] for scalar models in the framework of variational inequalities.

Throughout the paper, based on asymptotic expansions, we address the asymptotic behavior of $(\lambda^\varepsilon, u^\varepsilon)$ as $\varepsilon \to 0$, depending on the different values of r_0, β^0 and β^* in (7.1), (7.2) and (7.3), respectively.

7.2.1 The Spectrum and the Estimates for Eigenvalues

Let us denote by $\mathbf{V}$ the space $\{v \in (H^1(\Omega))^3 : v = 0 \text{ on } \Gamma_\Omega\}$ with the norm generated by the scalar product

$$(u, v)_{\mathbf{V}} = \int_\Omega e_{ij}(u)e_{ij}(v)\,dx\,. \tag{7.9}$$

On account of the Korn's inequality, (7.9) defines a norm in $\mathbf{V}$.

For fixed $\varepsilon > 0$, the weak formulation of problem (7.8) reads: find $\lambda^\varepsilon \in \mathbb{R}$, $u^\varepsilon \in \mathbf{V}$, $u^\varepsilon \neq 0$, satisfying

$$\int_\Omega \sigma_{ij}(u^\varepsilon)e_{ij}(v)\,dx + \beta(\varepsilon)\int_{\bigcup T^\varepsilon} M_{ij}u_i^\varepsilon v_j\,d\hat{x} = \lambda^\varepsilon \int_\Omega u_i^\varepsilon v_i\,dx,\ \ \forall v \in \mathbf{V}. \tag{7.10}$$

On account of (7.4) and (7.7), the left-hand side of (7.10) defines a bilinear, symmetric continuous and coercive form on $\mathbf{V} \subset (L^2(\Omega))^3$. Consequently, (7.10) has the discrete spectrum:

$$0 < \lambda_1^\varepsilon \le \lambda_2^\varepsilon \le \cdots \le \lambda_n^\varepsilon \le \cdots \xrightarrow{n\to\infty} +\infty, \tag{7.11}$$

where we have adopted the convention of repeated eigenvalues according to their multiplicities. The corresponding eigenfunctions form a basis in $\mathbf{V}$ and $(L^2(\Omega))^3$, and we assume that they are subject to the orthonormalization condition

$$\left(u^{n,\varepsilon}, u^{n,\varepsilon}\right)_{(L^2(\Omega))^3} = \delta_{n,m}. \tag{7.12}$$

The following lemma gives bounds for the eigenvalues of (7.8).

Lemma 7.1 *For each fixed $n \in \mathbb{N}$, there exist C and C_n constants independent of ε such that*

$$0 < C \leq \lambda_n^\varepsilon \leq C_n, \qquad \forall \varepsilon > 0. \tag{7.13}$$

Proof The left-hand side of the inequality holds since

$$\lambda_n^\varepsilon \geq \lambda_1^\varepsilon = \frac{\int\limits_\Omega \sigma_{ij}(u^{1,\varepsilon}) e_{ij}(u^{1,\varepsilon})\, dx + \beta(\varepsilon) \int\limits_{\bigcup T^\varepsilon} M_{ij} u_i^{1,\varepsilon} u_j^{1,\varepsilon}\, d\hat{x}}{\int\limits_\Omega u_i^{1,\varepsilon} u_i^{1,\varepsilon}\, dx} \geq C,$$

where $u^{1,\varepsilon}$ denotes an eigenvector corresponding to the first eigenvalue λ_1^ε, and the last inequality is a consequence of the Poincaré and Korn inequalities along with (7.4) and (7.7).

On the other hand, according to the minimax principle,

$$\lambda_n^\varepsilon = \min_{E_n \subset \mathbf{V}} \max_{v \in E_n, v \not\equiv 0} \frac{\int\limits_\Omega \sigma_{ij}(v) e_{ij}(v)\, dx + \beta(\varepsilon) \int\limits_{\bigcup T^\varepsilon} M_{ij} v_i v_j\, d\hat{x}}{\int\limits_\Omega v_i v_i dx}, \tag{7.14}$$

where the minimum is taken over the set of all the subspaces E_n of $\mathbf{V}$ of dimension n. Let us take the particular space E_n^* generated by the eigenvectors $[u^{1,0}, u^{2,0}, \cdots, u^{n,0}]$ of the Dirichlet problem in Ω (cf. (7.24)), namely the eigenvectors associated with the eigenvalues corresponding with $\{\lambda_1^0, \lambda_2^0, \cdots \lambda_n^0\}$ in the series (7.58). In (7.14) we write

$$\lambda_n^\varepsilon \leq \max_{v \in E_n^*, v \not\equiv 0} \frac{\int\limits_\Omega \sigma_{ij}(v) e_{ij}(v)\, dx}{\int\limits_\Omega v_i v_i dx} = \lambda_n^0,$$

and the right-hand side of (7.13) is proved.

7.3 The Spectral Homogenized Problems

In order to make the reading of the paper easier, in this section, we state all the spectral homogenized problems which depend on the different relations between parameters. We also state the local problems that allow us to describe the strange terms in the homogenized problems. We obtain all these problems in Sect. 7.4, by using the technique of matched asymptotic expansions.

(P1) In the most critical situation where $\beta^0 > 0$ and $r_0 > 0$, the homogenized problem reads

$$\begin{cases} -\dfrac{\partial \sigma_{ij,x}(u^0)}{\partial x_j} = \lambda^0 u_i^0 & \text{in } \Omega \\ u^0 = 0 & \text{on } \Gamma_\Omega \\ \sigma_{ij,x}(u^0) n_j + r_0 \mathscr{C}^e_{ij} u_j^0 = 0 & \text{on } \Sigma \end{cases}, \qquad i = 1, 2, 3, \tag{7.15}$$

where the matrix $\mathscr{C}^e = (\mathscr{C}^e_{il})_{i,l=1,2,3}$ is defined as

$$\mathscr{C}^e_{il}(\hat{x}) = \int_T \sigma^{\hat{x}}_{i3,y}(W^{l,M,\hat{x}}) d\hat{y}, \tag{7.16}$$

$W^{l,M,\hat{x}}$ being the solution of the $M(\hat{x})$-*dependent local problem*

$$\begin{cases} -\dfrac{\partial \sigma^{\hat{x}}_{ij,y}(W^{l,M,\hat{x}})}{\partial y_j} = 0 & \text{in } \mathbb{R}^{3+} \\ \sigma^{\hat{x}}_{ij,y}(W^{l,M,\hat{x}}) n_j = 0 & \text{on } \{y_3 = 0\} \setminus T \\ \sigma^{\hat{x}}_{ij,y}(W^{l,M,\hat{x}}) n_j - \beta^0 M_{ij}(\hat{x})(\mathrm{e}^l_j - W^{l,M,\hat{x}}_j) = 0 & \text{on } T \\ W^{l,M,\hat{x}}(y) \longrightarrow 0 & \text{as } |y| \to \infty,\ y_3 > 0 \end{cases}, \quad i = 1, 2, 3. \tag{7.17}$$

Above, and in what follows, variable y is an auxiliary variable in $\mathbb{R}^3$, cf. (7.28), and lower indexes x or y in the components of the stress and strain tensors mean the variable for derivation. The upper index $\hat{x}$ is a parameter which refers to the elastic homogeneous media with constant elastic coefficients $a_{ijkl}(\hat{x})$. Namely, in (7.16) and (7.17),

$$\sigma^{\hat{x}}_{ij,y}(V) = a_{ijkl}(\hat{x}) e_{kl,y}(V). \tag{7.18}$$

Also, e^l stands for the unitary vector in the y_l-direction, while $l = 1, 2, 3$.

(P2) For the critical size $r_0 > 0$, when $\beta^0 = +\infty$, the homogenized problem reads

$$\begin{cases} -\dfrac{\partial \sigma_{ij,x}(u^0)}{\partial x_j} = \lambda^0 u_i^0 & \text{in } \Omega \\ u^0 = 0 & \text{on } \Gamma_\Omega \\ \sigma_{ij,x}(u^0) n_j + r_0 \mathscr{C}_{ij} u_j^0 = 0 & \text{on } \Sigma \end{cases}, \qquad i = 1, 2, 3, \tag{7.19}$$

where the matrix $\mathscr{C} = (\mathscr{C}_{il})_{i,l=1,2,3}$ is defined as

$$\mathscr{C}_{il}(\hat{x}) = -\left\langle \sigma_{i3,y}(W^{l,\hat{x}}), 1 \right\rangle_{H^{-1/2}(T) \times H^{1/2}(T)} \tag{7.20}$$

$W^{l,\hat{x}}$ being the solution of the $\hat{x}$-*dependent local problem*

$$\begin{cases} -\dfrac{\partial \sigma^{\hat{x}}_{ij,y}(W^{l,\hat{x}})}{\partial y_j} = 0 & \text{in } \mathbb{R}^{3+} \\ \sigma^{\hat{x}}_{ij,y}(W^{l,\hat{x}})n_j = 0 & \text{on } \{y_3 = 0\} \setminus T \\ W^{l,\hat{x}}(y) = \mathrm{e}^l & \text{on } T \\ W^{l,\hat{x}}(y) \longrightarrow 0 & \text{as } |y| \to \infty,\ y_3 > 0 \end{cases}, \qquad i = 1, 2, 3. \tag{7.21}$$

$\sigma^{\hat{x}}_{ij,y}$ and e^l in (7.20) and (7.21) are defined as in the previous item (cf. (7.18)).

(P3) For the critical relation where $\beta^* > 0$ with $r_0 = +\infty$, the homogenized problem reads

$$\begin{cases} -\dfrac{\partial \sigma_{ij,x}(u^0)}{\partial x_j} = \lambda^0 u^0_i & \text{in } \Omega \\ u^0 = 0 & \text{on } \Gamma_\Omega \\ \sigma_{ij,x}(u^0)n_j + \beta^*|T|M_{ij}u^0_j = 0 & \text{on } \Sigma \end{cases}, \qquad i = 1, 2, 3. \tag{7.22}$$

Note that, in this case, to each reaction parameter $\beta(\varepsilon)$ corresponds *a new critical size* of the reaction regions $r_\varepsilon = O(\varepsilon\beta(\varepsilon)^{-1/2})$, while to each size r_ε corresponds *a critical reaction parameter* $\beta(\varepsilon) = O(\varepsilon^2 r_\varepsilon^{-2})$. The reaction matrix $\beta^*|T|M$ which appears in the boundary condition on Σ is referred to as *averaged reaction matrix*.

(P4) For the extreme cases where $\beta^* = 0$ or $r_0 = 0$, the homogenized problem is

$$\begin{cases} -\dfrac{\partial \sigma_{ij,x}(u^0)}{\partial x_j} = \lambda^0 u^0_i & \text{in } \Omega \\ u^0 = 0 & \text{on } \Gamma_\Omega \\ \sigma_{ij,x}(u^0)n_j = 0 & \text{on } \Sigma \end{cases}, \qquad i = 1, 2, 3. \tag{7.23}$$

(P5) For the extreme case where $r_0 = +\infty$ and, $\beta^0 > 0$, or $\beta^0 = +\infty$, or $\beta^0 = 0$ and $\beta^* = +\infty$, the homogenized problem is the Dirichlet eigenvalue problem:

$$\begin{cases} -\dfrac{\partial \sigma_{ij,x}(u^0)}{\partial x_j} = \lambda^0 u^0_i & \text{in } \Omega \\ u^0 = 0 & \text{on } \partial\Omega \end{cases}, \qquad i = 1, 2, 3. \tag{7.24}$$

The discreteness of the spectrum of problems (7.23) and (7.24) is well known in the literature, that of problem (7.22) follows as that of (7.8), with minor modifications, while that of (7.15) and (7.19) is a consequence of the properties of matrices (7.16) and (7.20), respectively. That is, the discreteness of these spectra

is linked to the setting of problems (7.17) and (7.21) as well as to the properties of their respective solutions. This is addressed in Sect. 7.5.

7.4 Asymptotic Expansions

Taking into account (7.13) and (7.12), for each $n = 1, 2, \cdots$ we consider the asymptotic expansions for the eigenvalue $\lambda^\varepsilon \equiv \lambda_n^\varepsilon$ and the corresponding eigenvector $u^\varepsilon \equiv u^{n,\varepsilon}$ of (7.8) as follows.

Assume

$$\lambda^\varepsilon = \lambda^0 + \cdots, \tag{7.25}$$

and an outer expansion for the eigenfunction

$$u^\varepsilon(x) = u^0(x) + \cdots, \quad \text{in } \Omega \cap \{x_3 > d\} \quad \forall d > 0, \tag{7.26}$$

which in fact, is supposed to be valid for x "far" from regions the $T^\varepsilon_{\widetilde{x}_k}$, namely, at a distance $\rho \gg r_\varepsilon$ from the center $\widetilde{x}_k$. In addition, we assume a local expansion in a neighborhood of each reaction region $T^\varepsilon_{\widetilde{x}_k}$

$$u^\varepsilon(x) = V^0(y) + \cdots \quad \text{for } y \in \overline{\mathbb{R}^{3+}}. \tag{7.27}$$

Above, and in what follows, we denote by

$$y = \frac{x - \widetilde{x}_k}{r_\varepsilon} \tag{7.28}$$

the local variable in a neighborhood of each center $\widetilde{x}_k$, $k \in \mathcal{J}^\varepsilon$, and by dots we denote regular terms in the asymptotic series containing lower order functions of ε that we are not using in our analysis.

By matching the local and outer expansions for u^ε, at the first order, we can write

$$\lim_{|y|\to\infty} V^0(y) = \lim_{x\to\widetilde{x}_k} u^0(x). \tag{7.29}$$

By replacing (7.25) and (7.26) in (7.8) we obtain the following equations for u^0:

$$\begin{cases} -\dfrac{\partial \sigma_{ij,x}(u^0)}{\partial x_j} = \lambda^0 u_i^0 & \text{in } \Omega, \\ u^0 = 0 & \text{on } \Gamma_\Omega, \end{cases} \tag{7.30}$$

plus some boundary condition on Σ to be determined. In order to do this, we first determine $V^0(y)$ in the local expansion (7.27). Taking derivatives in (7.8) with

respect to y, cf. (7.28), we replace (7.25) and (7.27) in (7.8), and take into account the continuity of the elastic coefficients $a_{ijkl}(x)$ and $M_{ij}(x)$, and (7.29). Then, we obtain that V^0 satisfies

$$\begin{cases} -\dfrac{\partial \sigma^{\mathrm{k}}_{ij,y}(V^0)}{\partial y_j} = 0 & \text{in } \mathbb{R}^{3+}, \\ \sigma^{\mathrm{k}}_{ij,y}(V^0) n_j = 0 & \text{on } \{y_3 = 0\} \setminus T, \\ \dfrac{1}{r_\varepsilon} \sigma^{\mathrm{k}}_{ij,y}(V^0) n_j + \beta(\varepsilon) M_{ij}(\widetilde{x}_{\mathrm{k}}) V^0_j = 0 & \text{on } T, \\ V^0(y) \longrightarrow u^0(\widetilde{x}_{\mathrm{k}}) & \text{as } |y| \to \infty,\ y_3 > 0. \end{cases} \tag{7.31}$$

Above, and in what follows, for simplicity, we write the upper index k in the strain tensor (7.18) when $\hat{x} \equiv \widetilde{x}_{\mathrm{k}}$, namely,

$$\sigma^{\mathrm{k}}_{ij,y}(V) = a_{ijkl}(\widetilde{x}_{\mathrm{k}}) e_{kl,y}(V). \tag{7.32}$$

We also note that the solution of (7.31) strongly depends on ε both because of the condition at infinity and because of the Robin boundary condition

$$\sigma^{\mathrm{k}}_{ij,y}(V^0) n_j + r_\varepsilon \beta(\varepsilon) M_{ij}(\widetilde{x}_{\mathrm{k}}) V^0_j = 0 \quad \text{on } T. \tag{7.33}$$

Therefore, considering the three possible limits in (7.2), $\beta^0 = 0$, $\beta^0 = +\infty$, or $\beta^0 > 0$, asymptotically, we obtain three different boundary conditions on T for V^0:

(a) If $\beta^0 = 0$, then, $\sigma^{\mathrm{k}}_{ij,y}(V^0) n_j \approx 0$ on T, which implies $V^0(y) \approx u^0(\widetilde{x}_{\mathrm{k}})$, $\forall y \in \mathbb{R}^{3+}$,
(b) If $\beta^0 = +\infty$, then, $V^0(y) \approx 0$ on T,
(c) If $\beta^0 > 0$, then, $\sigma^{\mathrm{k}}_{ij,y}(V^0) n_j + \beta^0 M_{ij}(\widetilde{x}_{\mathrm{k}}) V^0_j \approx 0$ on T.

We observe that only in the case where $\beta^0 > 0$, the dependence on the centers of the reaction regions cannot be avoided. However, V^0 can be written as a linear combination of solutions of three linear elasticity problems in the half-space $\mathbb{R}^{3+}$. These *local problems* avoid the dependence on the first term of the outer expansion (7.26), and are considered in Sect. 7.4.1.

7.4.1 The Stationary Local Problems

We decompose V^0 as follows:

$$V^0(y) \approx u^0_l(\widetilde{x}_{\mathrm{k}})(\mathrm{e}^l - W^l(y)), \tag{7.34}$$

where, for $l = 1, 2, 3$, W^l is the solution of different problems depending on the values of β^0 in (7.2):

(a) $W^l \equiv 0$ when $\beta^0 = 0$.
(b) When $\beta^0 = +\infty$, then, $W^l \equiv W^{l,\widetilde{x}_k}$ is the solution of

$$\begin{cases} -\dfrac{\partial \sigma^k_{ij,y}(W^{l,\widetilde{x}_k})}{\partial y_j} = 0 & \text{in } \mathbb{R}^{3+}, \\ \sigma^k_{ij,y}(W^{l,\widetilde{x}_k}) n_j = 0 & \text{on } \{y_3 = 0\} \setminus T, \\ W^{l,\widetilde{x}_k}(y) = \mathrm{e}^l & \text{on } T, \\ W^{l,\widetilde{x}_k}(y) \longrightarrow 0 & \text{as } |y| \to \infty,\ y_3 > 0. \end{cases} \tag{7.35}$$

Notice that for a homogeneous media in (7.8), problem (7.35) does not depend on the parameter $\widetilde{x}_k$.

(c) When $\beta^0 > 0$, then, $W^l \equiv W^{l,M,\widetilde{x}_k}$ is the solution of

$$\begin{cases} -\dfrac{\partial \sigma^k_{ij,y}(W^{l,M,\widetilde{x}_k})}{\partial y_j} = 0 & \text{in } \mathbb{R}^{3+}, \\ \sigma^k_{ij,y}(W^{l,M,\widetilde{x}_k}) n_j = 0 & \text{on } \{y_3 = 0\} \setminus T, \\ \sigma^k_{ij,y}(W^{l,M,\widetilde{x}_k}) n_j - \beta^0 M_{ij}(\widetilde{x}_k)(\mathrm{e}^l_j - W^{l,\widetilde{x}_k}_j) = 0 & \text{on } T, \\ W^{l,M,\widetilde{x}_k}(y) \to 0 & \text{as } |y| \to \infty,\ y_3 > 0. \end{cases} \tag{7.36}$$

In (7.36) there is a nonhomogeneous Robin condition on T which depends on the center of the regions $T^\varepsilon_{\widetilde{x}_k}$. Here, even if the media in (7.8) is homogeneous, we obtain a parametric family of three local problems, the parameter of dependence being $\widetilde{x}_k$. More specifically, in (7.36) the elastic constants of the medium and those of the spring depend on $\widetilde{x}_k$: $a_{ijkl}(\widetilde{x}_k)$ and $M_{ij}(\widetilde{x}_k)$, respectively.

7.4.2 The Boundary Condition on Σ

Considering (7.30), in order to obtain the boundary condition on Σ for u^0, we perform an integration by parts over the equilibrium equations in *coin-like domains*, neglecting the stresses across the lateral surface. We define one of these domains as follows. Let us consider Σ_1 an open domain contained in Σ such that $\partial \Sigma_1$ does not touch any region $T^\varepsilon_{\widetilde{x}_k}$. Let $\delta(\varepsilon)$ be positive, $r_\varepsilon \ll \delta(\varepsilon) \ll 1$. We consider the coin-like domain

$$\Omega^{\delta(\varepsilon)}_{\Sigma_1} = \Omega \cap (\Sigma_1 \times (0, \delta(\varepsilon))). \tag{7.37}$$

Let $\Gamma_{\delta(\varepsilon)}$ denote the lateral boundary of $\Omega_{\Sigma_1}^{\delta(\varepsilon)}$ in such a way that

$$\partial\Omega_{\Sigma_1}^{\delta(\varepsilon)} = \overline{\Gamma_{\delta(\varepsilon)}} \cup \overline{\Sigma_1^{\delta(\varepsilon)}} \cup \overline{\Sigma_1}, \tag{7.38}$$

where $\Sigma_1^{\delta(\varepsilon)}$ denotes the set $\{x : (x_1, x_2, 0) \in \Sigma_1,\ x_3 = \delta(\varepsilon)\}$. On $\Sigma_1^{\delta(\varepsilon)}$, we are "far" from the reaction regions $T_{\widetilde{x}_k}^{\varepsilon}$ and (7.26) hold. "Near" each region T^{ε}, we need to use the local expansion, which in terms of the macroscopic variable reads

$$u^{\varepsilon}(x) = V^0\big((x - \widetilde{x}_k)r_{\varepsilon}^{-1}\big) + \cdots.$$

In particular, on each reaction region $T_{\widetilde{x}_k}^{\varepsilon}$ we have (cf. (7.28) and (7.32))

$$\sigma_{i3,x}(u^{\varepsilon}) = \sigma_{i3,x}(V^0(y)) \approx a_{i3kh}(\widetilde{x}_k)\frac{1}{r_{\varepsilon}}e_{kh,y}(V^0(y)) + \cdots = \frac{1}{r_{\varepsilon}}\sigma_{i3,y}^{k}(V^0(y)) + \cdots \tag{7.39}$$

Now, we multiply the divergence vector in (7.8) by e^i and apply the Green formula over $\Omega_{\Sigma_1}^{\delta(\varepsilon)}$ (cf. (7.37) and (7.38)) to obtain

$$\int_{\Sigma_1 \cap \bigcup T_{\widetilde{x}_k}^{\varepsilon}} \sigma_{i3,x}(u^{\varepsilon})d\hat{x} = \int_{\Omega_{\Sigma_1}^{\delta(\varepsilon)}} \lambda^{\varepsilon} u_i^{\varepsilon} dx + \int_{\Gamma_{\delta(\varepsilon)}} \sigma_{ij,x}(u^{\varepsilon})n_j d\Gamma_{\delta} + \int_{\Sigma_1^{\delta(\varepsilon)}} \sigma_{i3,x}(u^{\varepsilon})d\hat{x}. \tag{7.40}$$

We observe that, by construction (cf. (7.13) and (7.12)), the two first integrals on the right-hand side of (7.40) converge toward zero as $\varepsilon \to 0$. For the other integral, we use the approximation (7.26), namely

$$\sigma_{i3,x}(u^{\varepsilon})\Big|_{x_3=\delta(\varepsilon)} = \sigma_{i3,x}(u^0)\Big|_{x_3=0} + \cdots.$$

Therefore, introducing this and (7.39) in (7.40), and performing the change of variable (7.28), we write

$$\begin{aligned}\int_{\Sigma_1} \sigma_{i3,x}(u^0)d\hat{x} &= \lim_{\varepsilon\to 0} \sum_{\widetilde{x}_k \in \Sigma_1} \int_{T_{\widetilde{x}_k}^{\varepsilon}} \sigma_{i3,x}\big(V^0\big(\frac{x - \widetilde{x}_k}{r_{\varepsilon}}\big)\big)d\hat{x} \\ &= \lim_{\varepsilon\to 0} r_{\varepsilon} \sum_{\widetilde{x}_k \in \Sigma_1} \int_{T} \sigma_{i3,y}^{k}(V^0(y))d\hat{y}. \end{aligned} \tag{7.41}$$

Considering the case where $\beta^0 > 0$ or $\beta^0 = +\infty$, cf. (7.2), in order to introduce the decomposition (7.34), we notice that from the Green formula, the integrals on T

(similarly on $T^{\varepsilon}_{\widetilde{x}_k}$) in (7.41) are in fact the duality products

$$\left\langle \sigma^{\mathrm{k}}_{i3,y}(V^0)\big|_T\ ,\ 1\right\rangle_{H^{-1/2}(T)\times H^{1/2}(T)},$$

which is suitable, taking into account that $\sigma^{\mathrm{k}}_{i3,y}(W^l)\big|_{\{y_3=0\}}$ is a distribution of compact support on $\overline{T}$.

Consequently, (7.34) and (7.41) lead us to

$$\begin{aligned}\int_{\Sigma_1}\sigma_{i3,x}(u^0)d\hat{x} &= \lim_{\varepsilon\to 0} r_\varepsilon \sum_{\widetilde{x}_k\in\Sigma_1} u^0_l(\widetilde{x}_{\mathrm{k}})\left\langle \sigma^{\mathrm{k}}_{i3,y}(\mathrm{e}^l - W^l)\ ,\ 1\right\rangle_{H^{-1/2}(T)\times H^{1/2}(T)}\\ &= -\lim_{\varepsilon\to 0} r_\varepsilon \sum_{\widetilde{x}_k\in\Sigma_1} \mathscr{B}_{il}(\widetilde{x}_{\mathrm{k}})u^0_l(\widetilde{x}_{\mathrm{k}}),\end{aligned} \tag{7.42}$$

where, $W^l \equiv W^{l,M,\widetilde{x}_{\mathrm{k}}}$ is the solution of (7.36) when $\beta^0 > 0$ and $W^l \equiv W^{l,\widetilde{x}_{\mathrm{k}}}$ is the solution of (7.35) when $\beta^0 = +\infty$. In addition, we have introduced the matrix $\mathscr{B}(\widetilde{x}_{\mathrm{k}}) = \big(\mathscr{B}_{il}(\widetilde{x}_{\mathrm{k}})\big)_{i,l=1,2,3}$ which is defined as follows:

$$\mathscr{B}_{il}(\widetilde{x}_{\mathrm{k}}) = -\int_T \sigma^{\mathrm{k}}_{i3,y}(\mathrm{e}^l - W^{l,M,\widetilde{x}_{\mathrm{k}}})d\hat{y} = \int_T \sigma^{\mathrm{k}}_{i3,y}(W^{l,M,\widetilde{x}_{\mathrm{k}}})d\hat{y}, \qquad \text{when } \beta^0 > 0, \tag{7.43}$$

and

$$\begin{aligned}\mathscr{B}_{il}(\widetilde{x}_{\mathrm{k}}) &= -\left\langle \sigma^{\mathrm{k}}_{i3,y}(\mathrm{e}^l - W^{l,\widetilde{x}_{\mathrm{k}}})\ ,\ 1\right\rangle_{H^{-1/2}(T)\times H^{1/2}(T)}\\ &= \left\langle \sigma^{\mathrm{k}}_{i3,y}(W^{l,\widetilde{x}_{\mathrm{k}}})\ ,\ 1\right\rangle_{H^{-1/2}(T)\times H^{1/2}(T)}, \qquad \text{when } \beta^0 = +\infty.\end{aligned} \tag{7.44}$$

That is, $\mathscr{B}(\widetilde{x}_{\mathrm{k}}) = \mathscr{C}^e(\widetilde{x}_{\mathrm{k}})$ when $\beta^0 > 0$, cf. (7.16), while $\mathscr{B}(\widetilde{x}_{\mathrm{k}}) = \mathscr{C}(\widetilde{x}_{\mathrm{k}})$ when $\beta^0 = +\infty$, cf. (7.20).

In both cases, assuming a continuous dependence of $\mathscr{B}$ on the parameter $\hat{x} \in \Sigma$ (cf. [GoNaPe20]), and under the assumption that $r_0 > 0$ in (7.1) we obtain that (7.42) reads

$$\int_{\Sigma_1}\sigma_{i3,x}(u^0)d\hat{x} = -r_0\int_{\Sigma_1}\mathscr{B}_{il}(\hat{x})u^0_l(\hat{x},0)\,d\hat{x}. \tag{7.45}$$

Here, it is self-evident that the definition that we use for $\mathscr{B}_{il}(\hat{x})$ for $\hat{x} \in \Sigma$, is that in (7.43) and (7.44) by replacing $\widetilde{x}_{\mathrm{k}}$ by $\hat{x}$ in all the involved functions.

Obviously, when $\beta^0 > 0$ or $\beta^0 = +\infty$ and $r_0 = 0$, the condition

$$\int_{\Sigma_1}\sigma_{i3,x}(u^0)d\hat{x} = 0 \tag{7.46}$$

is asymptotically imposed.

The reasoning above must be slightly modified in the case where $\beta^0 > 0$ or $\beta^0 = +\infty$ and $r_0 = +\infty$ as follows. When $r_0 = +\infty$, we multiply both sides of the equality in (7.40) by $\varepsilon^2 r_\varepsilon^{-1}$. Since $\varepsilon^2 r_\varepsilon^{-1} \to 0$ as $\varepsilon \to 0$, the reasoning in (7.41)–(7.45) gives

$$0 = \lim_{\varepsilon\to 0} \sum_{\widetilde{x}_k \in \Sigma_1} \varepsilon^2 \int_{T^\varepsilon_{\widetilde{x}_k}} \sigma^{\mathrm{k}}_{i3,y}(V^0(y)) d\hat{y} = -\int_{\Sigma_1} \mathscr{B}_{il}(\hat{x}) u^0_l(\hat{x}, 0)\, d\hat{x}, \tag{7.47}$$

and by the properties of matrix $\mathscr{B}_{il}$ (cf. Sect. 7.5), we deduce that $\int_{\Sigma_1} u^0_l d\hat{x} = 0$.

Finally, in the case where $r_0 = +\infty$ and $\beta^0 = 0$, we can distinguish between the cases when $\beta^* > 0$ or $\beta^* = 0$ in (7.3). Indeed, recalling the relation (7.33), we rewrite (7.41) as follows:

$$\begin{aligned}\int_{\Sigma_1} \sigma_{i3,x}(u^0) d\hat{x} &= \lim_{\varepsilon\to 0} \sum_{\widetilde{x}_k \in \Sigma_1} \int_T r_\varepsilon \sigma^{\mathrm{k}}_{i3,y}(V^0(y)) d\hat{y} \\ &= -\lim_{\varepsilon\to 0} \beta(\varepsilon) r_\varepsilon^2 \sum_{\widetilde{x}_k \in \Sigma_1} \int_T M_{il}(\widetilde{x}_{\mathrm{k}}) V^0_l(y) d\hat{y}.\end{aligned}$$

Then, on account of item (a) below (7.33) and (7.34), we get

$$\int_{\Sigma_1} \sigma_{i3,x}(u^0) d\hat{x} = -\lim_{\varepsilon\to 0} \beta(\varepsilon) r_\varepsilon^2 \varepsilon^{-2} \sum_{\widetilde{x}_k \in \Sigma_1} \varepsilon^2 M_{il}(\widetilde{x}_{\mathrm{k}}) u^0_l(\widetilde{x}_{\mathrm{k}}) \int_T d\hat{y}. \tag{7.48}$$

Now, when $\beta^* = 0$ in (7.3), (7.48) gives (7.46), while, when $\beta^* > 0$ in (7.3), (7.48) gives

$$\int_{\Sigma_1} \sigma_{i3,x}(u^0) d\hat{x} = -\beta^* |T| \int_{\Sigma_1} M_{il}(\hat{x}) u^0_l(\hat{x}, 0) d\hat{x}\,. \tag{7.49}$$

Obviously, in the two last cases β^0 must be $\beta^0 = 0$ to somehow compensate $r_0 = +\infty$. Also, when $\beta^0 = 0$, $r_0 = +\infty$ and $\beta^* = +\infty$, we follow the idea of (7.47).

Gathering all the possible limit behavior for parameters, cf. (7.1)–(7.3), on account of the somewhat arbitrary choice of $\Sigma_1 \subset \Sigma$, from (7.45)–(7.47) and (7.49) we obtain the following boundary conditions on Σ to be added to (7.30) in order to determine the first terms of the asymptotic expansions (7.25) and (7.26), namely the pair (λ^0, u^0).

(P1) If $\beta^0 > 0$ and $r_0 > 0$, then we have $\sigma_{ij,x}(u^0) n_j + r_0 \mathscr{C}^e_{il} u^0_l = 0$ on Σ. This gives that (λ^0, u^0) is an eigenpair of problem (7.15).

(P2) If $\beta^0 = +\infty$ and $r_0 > 0$, then we have $\sigma_{ij,x}(u^0)n_j + r_0\mathscr{C}_{il}u_l^0 = 0$ on Σ. This gives that (λ^0, u^0) is an eigenpair of problem (7.19).

(P3) If $\beta^* > 0$ and $r_0 = +\infty$, then, $\sigma_{ij,x}(u^0)n_j + \beta^*|T|M_{il}u_l^0 = 0$ on Σ. Here, obviously, $\beta^0 = 0$. This gives that (λ^0, u^0) is an eigenpair of problem (7.22).

(P4) If $\beta^* = 0$ or $r_0 = 0$, then $\sigma_{ij,x}(u^0)n_j = 0$ on Σ. Note that $\beta^* = 0$ also contains the case where $\beta^0 = 0$ and $r_0 > 0$. This gives that (λ^0, u^0) is an eigenpair of problem (7.23).

(P5) If $\beta^0 > 0$ or $\beta^0 = +\infty$, and $r_0 = +\infty$, then, $u^0 = 0$ on Σ. This gives that (λ^0, u^0) is an eigenpair of problem (7.24). The same holds when $\beta^0 = 0$, $r_0 = +\infty$ and $\beta^* = +\infty$.

7.5 Abstract Framework for Local and Homogenized Problems

In this section, we provide the variational formulations for local and homogenized problems. When dealing with problems (7.15) and (7.19) that contain the capacity matrix, we need to restrict ourselves to the case of isotropic media, cf. (7.54). The reason is that, in order to show the symmetry and positivity of the capacity matrices, we need a thorough study of the behavior at infinity of the solutions $W^{l,M,\hat{x}}$ and $W^{l,\hat{x}}$ of the corresponding local problems. This is performed in Sect. 7.5.1 for isotropic media, leaving the study of anisotropic media for a forthcoming publication, cf. [Na99, GoNaPe20].

As regards the local problem (7.35), in the general case of a nonhomogeneous and anisotropic material, cf. (7.5), the centers $\widetilde{x}_k$ become parameters arising in the stress tensor (7.32), and we have a parametric family of local problems (7.35) whose solutions satisfy the equilibrium equations for a homogeneous media filling the half-space $\mathbb{R}^{3+}$. The same applies to (7.21) and $\hat{x} \in \Sigma$. The proof of the existence and uniqueness of solution of (7.21) follows the scheme in [LoPe88]. For the sake of completeness, we introduce the result here below.

Let $\mathscr{D}(\overline{\mathbb{R}^{3+}})$ be the space of functions that are restrictions to $\overline{\mathbb{R}^{3+}}$ of the elements of $\mathscr{D}(\mathbb{R}^3)$, and let $\mathscr{D}_T(\overline{\mathbb{R}^{3+}})$ be the space of functions of $\mathscr{D}(\overline{\mathbb{R}^{3+}})$ such that they vanish in a neighborhood of $\overline{T}$. Let us define the functional spaces $\mathscr{W}$ and $\mathscr{W}_0$ as the completion of $(\mathscr{D}(\overline{\mathbb{R}^{3+}}))^3$ and $(\mathscr{D}_T(\overline{\mathbb{R}^{3+}}))^3$ respectively, with the norm

$$\Big(\sum_{i,j=1}^{3} \big\|e_{ij,y}(U); L^2(\mathbb{R}^{3+})\big\|^2\Big)^{1/2}. \tag{7.50}$$

For each $l = 1, 2, 3$, we take a function $\Psi^l \in (\mathscr{D}(\overline{\mathbb{R}^{3+}}))^3$ such that $\Psi^l = \mathrm{e}^{\mathrm{l}}$ in a neighborhood of $\overline{T}$. Then, there is a unique solution $W^{l,\hat{x}} \in \Psi^l + \mathscr{W}_0$ satisfying

$$\int_{\mathbb{R}^{3+}} \sigma_{ij,y}^{\hat{x}}(W^{l,\hat{x}})e_{ij,y}(V)dy = 0 \quad \forall V \in \mathscr{W}_0. \tag{7.51}$$

This is a weak formulation of problem $(7.21)_1$–$(7.21)_3$. The precise behavior at infinity of $W^{l,\hat{x}}$ for the isotropic media (7.54)) is a consequence of (7.57) (cf. Lemma 7.2).

As regards the setting of the local problems (7.17), in the suitable Sobolev spaces, this follows the scheme below.

Consider the space $\mathscr{V}$ completion of $(\mathscr{D}(\overline{\mathbb{R}^{3+}}))^3$ with respect to the norm

$$\|U\|_{\mathscr{V}} = \left(\sum_{i,j=1}^{3} \|e_{ij,y}(U); L^2(\mathbb{R}^{3+})\|^2 + \sum_{i=1}^{3} \|U_i; L^2(T)\|^2 \right)^{1/2}.$$

Due to the Korn's inequality over bounded domains, the continuous embedding $\mathscr{V} \subset (H^1_{loc}(\mathbb{R}^{3+}))^3$ holds.

Let us define the bilinear, symmetric, continuous, and coercive form on $\mathscr{V}$:

$$a_{\hat{x}}(U,V) = \int_{\mathbb{R}^{3+}} \sigma^{\hat{x}}_{ij,y}(U) e_{ij,y}(V)\, dy + \beta^0 M_{ij}(\hat{x}) \int_T U_i V_j\, d\hat{y} \quad \forall U, V \in \mathscr{V},$$

which, on account of (7.4) and (7.7), defines a norm in $\mathscr{V}$ equivalent to $\|\cdot\|_{\mathscr{V}}$; $\hat{x} \in \Sigma$ being a parameter, cf. (7.18). Also, for $l = 1, 2, 3$, and $\hat{x} \in \Sigma$, let us consider the linear continuous functional on $\mathscr{V}$:

$$F_{l,\hat{x}}(U) = \beta^0 M_{il}(\hat{x}) \int_T U_i\, d\hat{y} \quad \forall U \in \mathscr{V}.$$

Then, the Riesz theorem ensures that there exists a unique function $W^{l,M,\hat{x}} \in \mathscr{V}$ satisfying

$$a_{\hat{x}}(W^{l,M,\hat{x}}, V) = F_{l,\hat{x}}(V) \quad \forall V \in \mathscr{V}, \tag{7.52}$$

which is a weak formulation of the problem $(7.17)_1$–$(7.17)_3$, $l = 1, 2, 3$. The condition at infinity in $(7.17)_4$ is provided below, in Sect. 7.5.1, when the original media is isotropic, cf. (7.54).

7.5.1 The Case of an Isotropic Medium

The existence and uniqueness of solution of (7.21) as well as its precise behavior at infinity, has been considered in [LoPe88] when the medium filling the domain Ω is isotropic and homogeneous; that is, when (7.5) reads

$$\sigma_{ij,x}(u) = \lambda \delta_{ij} e_{kk,x}(u) + 2\mu e_{ij,x}(u), \tag{7.53}$$

λ and μ being the Lamé coefficients. In this case, the three local problems (7.21), $l = 1, 2, 3$, are independent of $\hat{x}$. However, the technique in [LoPe88], which uses the Green tensor for an isotropic and homogeneous medium filling the half-space, also can be applied to the case of an isotropic media (7.54), as we do below. Indeed, we note that for

$$\sigma_{ij,x}(u) = \lambda(x)\delta_{ij}e_{kk,x}(u) + 2\mu(x)e_{ij,x}(u), \tag{7.54}$$

the Lamé coefficients appearing in (7.21) are also constants:

$$\sigma^{\hat{x}}_{ij,y}(U) = \lambda(\hat{x})\delta_{ij}e_{kk,y}(U) + 2\mu(\hat{x})e_{ij,y}(U). \tag{7.55}$$

In addition, the Lamé coefficients, the Young modulus E, and the Poisson coefficient ς are related by:

$$\lambda(\hat{x}) = \frac{E(\hat{x})\varsigma(\hat{x})}{(1+\varsigma(\hat{x}))(1-2\varsigma(\hat{x}))}, \quad \mu(\hat{x}) = \frac{E(\hat{x})}{2(1+\varsigma(\hat{x}))}$$

cf., e.g., [LaLi90] and [Te83].

First, let us introduce the Green tensor $G^{\hat{x}}$ (see Section I.8 of [LaLi90]):

$$G^{\hat{x}}_{11} = g(\hat{x})\left(\frac{2(1-\varsigma(\hat{x}))\rho+\xi_3}{\rho(\rho+\xi_3)} + \frac{\xi_1^2(2\rho(\varsigma(\hat{x})\rho+\xi_3)+\xi_3^2)}{\rho^3(\rho+\xi_3)^2}\right),$$

$$G^{\hat{x}}_{22} = g(\hat{x})\left(\frac{2(1-\varsigma(\hat{x}))\rho+\xi_3}{\rho(\rho+\xi_3)} + \frac{\xi_2^2(2\rho(\varsigma(\hat{x})\rho+\xi_3)+\xi_3^2)}{\rho^3(\rho+\xi_3)^2}\right),$$

$$G^{\hat{x}}_{12} = g(\hat{x})\left(\frac{\xi_1\xi_2(2\rho(\varsigma(\hat{x})\rho+\xi_3)+\xi_3^2)}{\rho^3(\rho+\xi_3)^2}\right), \; G^{\hat{x}}_{13} = g(\hat{x})\left(\frac{\xi_1\xi_3}{\rho^3} - \frac{(1-2\varsigma(\hat{x}))\xi_1}{\rho(\rho+\xi_3)}\right),$$

$$G^{\hat{x}}_{23} = g(\hat{x})\left(\frac{\xi_2\xi_3}{\rho^3} - \frac{(1-2\varsigma(\hat{x}))\xi_2}{\rho(\rho+\xi_3)}\right), \quad G^{\hat{x}}_{33} = g(\hat{x})\left(\frac{\xi_3^2}{\rho^3} + \frac{2(1-\varsigma(\hat{x}))}{\rho}\right),$$

where

$$g(\hat{x}) = \frac{1+\varsigma(\hat{x})}{2\pi E(\hat{x})}, \quad \rho = \sqrt{\xi_1^2+\xi_2^2+\xi_3^2}, \quad (\xi_1,\xi_2,\xi_3) \in \mathbb{R}^{3+}, \quad G^{\hat{x}}_{ij} = G^{\hat{x}}_{ji}.$$

Then, we consider the solution $W^{l,M,\hat{x}} \in \mathscr{V}$ of (7.52), and the normal component of the associated stress tensor on the plane $\{y_3 = 0\}$, $\sigma^{l,\hat{x}}$, which has a compact support on $\overline{T}$:

$$(\sigma_1^{l,\hat{x}}, \sigma_2^{l,\hat{x}}, \sigma_3^{l,\hat{x}}) := (\sigma^{\hat{x}}_{13,y}(W^{l,M,\hat{x}}), \sigma^{\hat{x}}_{23,y}(W^{l,M,\hat{x}}), \sigma^{\hat{x}}_{33,y}(W^{l,M,\hat{x}})),$$

and, according to Section 4 in [LoPe88], we can write $W^{l,M,\hat{x}} = G^{\hat{x}} * \sigma^{l,\hat{x}}$ whose components are

$$W_i^{l,M,\hat{x}} = G_{ij}^{\hat{x}} * \sigma_{j3,y}^{\hat{x}}(W^{l,M,\hat{x}}), \quad i = 1, 2, 3.$$

Further specifying, for each $(y_1, y_2, y_3) \in \mathbb{R}^{3+}$, we write

$$W_i^{l,M,\hat{x}}(y_1, y_2, y_3) = \int_T \sigma_j^{l,\hat{x}}(\xi_1, \xi_2) G_{ij}^{\hat{x}}(y_1 - \xi_1, y_2 - \xi_2, y_3) d\xi_1 d\xi_2. \tag{7.56}$$

In the case of an isotropic media and $W^{l,\hat{x}}$ the solution of (7.51), the representation above reads

$$W_i^{l,\hat{x}}(y_1, y_2, y_3) = \left\langle \sigma_j^{l,\hat{x}}, \quad G_{ij}^{\hat{x}}(y_1 - \cdot, y_2 - \cdot, y_3) \right\rangle_{H^{-1/2}(T) \times H^{1/2}(T)}, \tag{7.57}$$

see Theorem 4.1 in [LoPe88] for the result.

Finally, using (7.56), (7.57) and the reasoning of Proposition 4.1 in [LoPe88], we state the following lemma.

Lemma 7.2 *For fixed $\hat{x} \in \Sigma$ and the isotropic media in (7.55), there is a positive constant $C_{\hat{x}}$ such that for any $y \in \mathbb{R}^{3+}$, $l = 1, 2, 3$, and $W^l = W^{l,M,\hat{x}}$ or $W^l = W^{l,M,\hat{x}}$, the following estimates hold*

$$|W_i^l(y)| \leq C_{\hat{x}} \left(\frac{1}{d(y, \overline{T})} + \frac{1}{d(y, \overline{T})^2} \right), \quad i = 1, 2, 3.$$

$$\left| \frac{\partial W_i^l}{\partial y_j}(y) \right| \leq C_{\hat{x}} \left(\frac{1}{d(y, \overline{T})} + \frac{1}{d(y, \overline{T})^2} \right), \quad i, j = 1, 2, 3.$$

7.5.2 Setting of the Homogenized Problems

As it has been outlined in Sect. 7.3, the variational formulation in terms of bilinear, continuous, and coercive forms on a couple of Hilbert spaces with a compact and dense embedding $\mathbf{V} \subset \mathbf{H}$ is classical for problems (7.23) and (7.24) (cf., e.g., [SaSa89]). The variational formulation for (7.22) holds as that for (7.8), on account of (7.4) and (7.7). Let us show that it is also classical for problems (7.15) and (7.19) when the media is isotropic, cf. (7.54).

For each fixed $\hat{x} \in \Sigma$, the fact that matrix $\mathscr{C}(\hat{x})$, cf. (7.20), is symmetric and positive definite has been shown in [LoPe88], as a consequence of the equality

$$\mathscr{C}_{il}(\hat{x}) = \left\langle \sigma_{ij,y} n_j(W^{l,\hat{x}}), 1 \right\rangle_{H^{-1/2}(T) \times H^{1/2}(T)} = \int_{\mathbb{R}^{3+}} \sigma_{pj,y}(W^{l,\hat{x}}) e_{pj,y}(W^{i,\hat{x}}) \, dy.$$

We show that the matrix $\mathscr{C}^e(\hat{x})$, cf. (7.16), is a symmetric and positive definite matrix by performing an integration by parts as follows: For each fixed l, we multiply the divergence vector in (7.17) by $W^{i,M,\hat{x}}$. Then, taking integrals over a half-ball $B(0,R)\cap\mathbb{R}^{3+}$, we apply the Green formula and take limits as $R\to\infty$; because of Lemma 7.2, we obtain the chain of equalities

$$\int_{\mathbb{R}^{3+}} \sigma^{\hat{x}}_{pj,y}(W^{l,M,\hat{x}})e_{pj,y}(W^{i,M,\hat{x}})\,dy+\beta^0 M_{pj}(\hat{x})\int_T (\mathrm{e}^l-W^{l,M,\hat{x}})_j(\mathrm{e}^i-W^{i,M,\hat{x}})_p\,d\hat{y}$$

$$=-\beta^0 M_{ij}(\hat{x})\int_T (\mathrm{e}^l-W^{l,M,\hat{x}})_j\,d\hat{y}=-\int_T \sigma^{\hat{x}}_{ij,y}(W^{l,M,\hat{x}})n_j d\hat{y}=\mathscr{C}^e_{ij}(\hat{x}).$$

In this way, the symmetry of $\mathscr{C}^e$ comes from the symmetry of the coefficients a_{ijkl} and M_{ij}, while the positiveness of $\mathscr{C}^e$ is due to (7.4) and (7.7). Indeed, it is simple to verify that, $\forall\bar{\alpha}\in\mathbb{R}^{3+}$,

$$\bar{\alpha}\mathscr{C}^e\bar{\alpha}^\top=\int_{\mathbb{R}^{3+}}\sigma^{\hat{x}}_{pj,y}(\alpha_l W^{l,M,\hat{x}})e_{pj,y}(\alpha_l W^{l,M,\hat{x}})\,dy$$

$$+\beta^0 M_{pj}(\hat{x})\int_T \alpha_l(\mathrm{e}^l-W^{l,M,\hat{x}})_j\alpha_l(\mathrm{e}^l-W^{l,M,\hat{x}})_p\,d\hat{y}.$$

Consequently, $\bar{\alpha}\mathscr{C}\bar{\alpha}^\top\geq 0$ and $\bar{\alpha}\mathscr{C}\bar{\alpha}^\top=0$ implies $\bar{\alpha}=0$ as a consequence of the behavior at the infinity for the functions $W^{l,M,\hat{x}}$. Here, $\bar{\alpha}=(\alpha_1,\alpha_2,\alpha_3)$ and $\top$ stands for transposition.

Finally, we state the weak formulation of each homogenized problem.

- For problem (7.15): find $\lambda^0\in\mathbb{R}$, $u^0\in\mathbf{V}$, $u^0\neq 0$, satisfying

$$\int_\Omega \sigma_{ij,x}(u^0)e_{ij,x}(v)\,dx+r_0\int_\Sigma \mathscr{C}^e_{ij}(\hat{x})u^0_i v_j\,d\hat{x}=\lambda^0\int_\Omega u^0_i v_i\,dx,\quad \forall v\in\mathbf{V}.$$

- For problem (7.19): find $\lambda^0\in\mathbb{R}$, $u^0\in\mathbf{V}$, $u^0\neq 0$, satisfying

$$\int_\Omega \sigma_{ij,x}(u^0)e_{ij,x}(v)\,dx+r_0\int_\Sigma \mathscr{C}_{ij}(\hat{x})u^0_i v_j\,d\hat{x}=\lambda^0\int_\Omega u^0_i v_i\,dx,\quad \forall v\in\mathbf{V}.$$

- For problem (7.22): find $\lambda^0\in\mathbb{R}$, $u^0\in\mathbf{V}$, $u^0\neq 0$, satisfying

$$\int_\Omega \sigma_{ij,x}(u^0)e_{ij,x}(v)\,dx+\beta^*|T|\int_\Sigma M_{ij}(\hat{x})u^0_i v_j\,d\hat{x}=\lambda^0\int_\Omega u^0_i v_i\,dx,\quad \forall v\in\mathbf{V}.$$

- For problem (7.23): find $\lambda^0 \in \mathbb{R}$, $u^0 \in \mathbf{V}$, $u^0 \neq 0$, satisfying

$$\int_{\Omega} \sigma_{ij,x}(u^0) e_{ij,x}(v)\, dx = \lambda^0 \int_{\Omega} u_i^0 v_i\, dx, \quad \forall v \in \mathbf{V}.$$

- For problem (7.24): find $\lambda^0 \in \mathbb{R}$, $u^0 \in (H_0^1(\Omega))^3$, $u^0 \neq 0$, satisfying

$$\int_{\Omega} \sigma_{ij,x}(u^0) e_{ij}(v)\, dx = \lambda^0 \int_{\Omega} u_i^0 v_i\, dx, \quad \forall v \in (H_0^1(\Omega))^3.$$

On account of the Korn's inequality all these problems have a discrete spectrum:

$$0 < \lambda_1^0 \leq \lambda_2^0 \leq \cdots \leq \lambda_n^0 \leq \cdots \xrightarrow{n\to\infty} +\infty, \tag{7.58}$$

while we can chose the corresponding eigenfunctions $\{u^{n,0}\}_{n=1}^{\infty}$ to form an orthonormal basis in $(L^2(\Omega))^3$.

Remark 7.1 In order to show the convergence of the eigenvalues in (7.11) and of the corresponding eigenfunctions, as $\varepsilon \to 0$, toward those of the homogenized problems, we need to show that the constant $C_{\hat{x}}$ arising in Lemma 7.2 is independent of $\hat{x}$. This will be proved in a forthcoming publication (cf. [Na99] and [GoNaPe20]), as a consequence of certain smooth dependence of the solutions $W^{l,M,\hat{x}}$ and $W^{l,\hat{x}}$ on the parameter $\hat{x}$. Obviously, this result holds true when the initial media in (7.8) is isotropic and homogeneous, cf. (7.53), and also the Robin matrix M is constant. Indeed, for a homogeneous and isotropic media $W^{l,\hat{x}} \equiv W^l$ does not depend on $\hat{x}$. In addition, if M is a constant matrix, $W^{l,M,\hat{x}} \equiv W^{l,M}$ is also independent of $\hat{x}$.

Similarly, in [GoNaPe20] we provide the precise decay of the solution of the local problems (7.21) and (7.17) when the original medium is anisotropic. This ends the correct setting of problems (7.15) and (7.19) in nonhomogeneous and anisotropic media.

Acknowledgments This work has been partially supported by Spanish MICINN grant PGC2018-098178-B-I00, Russian Foundation of Basic Research grant 18-01-00325, and the Convenium Banco Santander—Universidad de Cantabria 2018.

References

[Al83] Allaire, G.: Homogenization of the Navier-Stokes equations in open sets perforated with tiny holes II. Non critical size of the holes for a volume distribution of holes and a surface distribution of holes. Arch. Ration. Mech. Anal. **113**, 261–298 (1983)

[At84] Attouch, H.: Variational Convergence for Functions and Operators. Applicable Mathematics Series. Pitman, London (1984)

[Br92] Brillard, A.: Asymptotic flow of a viscous and incompressible fluid through a plane sieve. In: Progress in Partial Differential Equations: Calculus of Variations, Applications. Pitman Research Notes in Mathematics Series, vol. 267, pp. 158–172. Longman Scientific & Technical, Harlow (1992)

[BrEtAl90] Brillard, A., Lobo, M., Pérez, E.: Homogénéisation de Frontières par epi-convergence en élasticité linéare. RAIRO Modél. Math. Anal. Numér. **24**, 5–26 (1990)

[BrEtAl16] Brillard, A., Gómez, D., Lobo, M., Pérez, E., Shaposhnikova, T.A.: Boundary homogenization in perforated domains for adsorption problems with an advection term. Appl. Anal. **95**, 218–237 (2016)

[BrLoPe90] Brillard, A., Lobo, M., Pérez, E.: Un probléme d'homogénéisation de frontière en élasticité linéare pour un corps cylindrique. C.R. Acad. Sci. Paris Sér. II Méc. Phys. Chim. Sci. Univers Sci. Terre **311**, 15–20 (1990)

[CiDoEn96] Cioranescu, D., Donato, P., Ene, H.: Homogenization of the Stokes problem with non-homogeneous slip boundary conditions. Math. Methods Appl. Sci. **19**, 857–881 (1996)

[CiMu97] Cioranescu, D., Murat, F.: A strange term coming from nowhere. In: Topics in the Mathematical Modelling of Composite Materials. Progress in Nonlinear Differential Equations and Their Applications, vol. 31, pp. 45–93. Birkäuser, Boston (1997)

[Co85] Conca, C.: On the application of the homogenization theory to a class of problems arising in fluid mechanics. J. Math. Pures Appl. **64**, 31–75 (1985)

[Co87] Conca, C.: Étude d'un fluide traversant une paroi perforée. I. Comportement limite près de la paroi. J. Math. Pures Appl. **66**, 1–43 (1987)

[De87] Del Vecchio, T.: The thick Neumann's sieve. Ann. Mat. Pura Appl. **147**, 363–402 (1987)

[GoEtAl13] Gómez, D., Pérez, E., Shaposhnikova, T.A.: Spectral boundary homogenization problems in perforated domains with Robin boundary conditions and large parameters. In: Integral Methods in Science and Engineering, pp. 155–174. Birkhäuser/Springer, New York (2013)

[GoEtAl18] Gómez, D., Lobo, L., Pérez, E., Sanchez-Palencia, E.: Homogenization in perforated domains: a Stokes grill and an adsorption process. Appl. Anal. **97**, 2893–2919 (2018)

[GoEtAl19] Gómez, D., Pérez, E., Podolskiy, A.V., Shaposhnikova, T.A.: Homogenization of variational inequalities for the p-Laplace operator in perforated media along manifolds. Appl. Math. Optim. **79**, 695–713 (2019)

[GoLoPe19] Gómez, D., Lobo, M., Pérez-Martínez, M.-E.: Asymptotics for models of non-stationary diffusion in domains with a surface distribution of obstacles. Math. Methods Appl. Sci. **42**, 403–413 (2019)

[GoNaPe18] Gómez, D., Nazarov, S.A., Pérez, E.: Homogenization of Winkler-Steklov spectral conditions in three-dimensional linear elasticity. Z. Angew. Math. Phys. **69**(2), article 35, 23 pp. (2018)

[GoNaPe20] Gómez, D., Nazarov, S.A., Pérez, E.: Asymptotics for spectral problems with rapidly alternating boundary conditions on a strainer Winkler foundation. Submitted, (2020)

[GoPeSh12] Gómez, D., Pérez, E., Shaposhnikova, T.A.: On homogenization of nonlinear Robin type boundary conditions for cavities along manifolds and associated spectral problems. Asymptot. Anal. **80**, 289–322 (2012)

[GoPeSh13] Gómez, D., Pérez, E., Shaposhnikova, T.A.: On correctors for spectral problems in the homogenization of Robin boundary conditions with very large parameters. Int. J. Appl. Math. **26**, 309–320 (2013)

[GrMiOr15] Griso, G., Migunova, A., Orlik, J.: Homogenization via unfolding in periodic layer with contact. Asymptot. Anal. **99**, 23–52 (2015)

[IoOnVel05] Ionescu, I., Onofrei, D., Vernescu, B.: Γ-Convergence for a fault model with slip-weakening friction and periodic barriers. Quart. Appl. Math. **63**(4), 747–778 (2005)

[JaAdBr00] El Jarroudi, M., Addou, A., Brillard, A.: Asymptotic analysis and boundary homogenization in linear elasticity. Math. Methods Appl. Sci. **23**, 655–683 (2000)

[LaLi90] Landau, L., Lifchitz, E.: Théorie de l'Élasticité. Physique Théorique. Tome 7. Mir, Moscow (1990)

[LoEtAl97] Lobo, M., Oleinik, O.A., Pérez, M.E., Shaposhnikova, T.A.: On homogenization of solutions of boundary value problems in domains, perforated along manifolds. Ann. Scuola Norm. Sup. Pisa Cl. Sci. 4^e série **25**, 611–629 (1997)

[LoPe87] Lobo, M., Pérez, E.: Comportement asymptotique d'un corps élastique dont une surface présente de petites zones de collage. C.R. Acad. Sci. Paris Sér. II Méc. Phys. Chim. Sci. Univers. Sci. Terre **304**, 195–198 (1987)

[LoPe88] Lobo, M., Pérez, E.: Asymptotic behaviour of an elastic body with a surface having small stuck regions. RAIRO Modél. Math. Anal. Numér. **22**, 609–624 (1988)

[LoPe92] Lobo, M., Pérez, E.: Boundary homogenization of certain elliptic problems for cylindrical bodies. Bull. Sci. Math. **116**, 399–426 (1992)

[MaKh06] Marchenko, V.A., Khruslov, E.Ya.: Homogenization of Partial Differential Equations. Birkhäuser, Boston (2006)

[Mu85] Murat, F.: The Neumann sieve. In: Nonlinear Variational Problems (Isola d'Elba, 1983). Research Notes in Mathematics, vol. 127, pp. 24–32. Pitman, Boston (1985)

[Na99] Nazarov, S.A.: Polynomial property of selfadjoint elliptic boundary value problems, and the algebraic description of their attributes. Uspekhi Mat. Nauk **54**, 77–142 (1999). English translation: Russ. Math. Surv. **54**, 947–1014 (1999)

[Na08] Nazarov, S.A.: Asymptotics of solutions and modeling of the elasticity problems in a domain with the rapidly oscillating boundary. Math. Izvestiya **72**(3), 509–564 (2008)

[NaSoSp10] Nazarov, S.A., Sokolowski, J., Specovius-Neugebauer, M.: Polarization matrices in anisotropic heterogeneous elasticity. Asymptot. Anal. **68**(4), 189–221 (2010)

[NgSa85] Nguetseng, G., Sanchez-Palencia, E.: Stress concentration for defects distributed near a surface. In: Local Effects in the Analysis of Structures. Studies in Applied Mechanics, vol. 12, pp. 55–74. Elsevier, Amsterdam (1985)

[OlCh93] Oleinik, O.A., Chechkin, G.: On boundary value problems for elliptic equations with rapidly changing type of boundary conditions. Uspekhi Mat. Nauk **48**, 163–164 (1993). English translation: Russ. Math. Surv. **48**, 173–175 (1993)

[OlCh96] Oleinik, O.A., Chechkin, G.: On asymptotics of solutions and eigenvalues of the boundary value problem with rapidly alternating boundary conditions for the system of elasticity. Atti Accad. Naz. Lincei Cl. Sci. Fis. Mat. Natur. Rend. Lincei (9) Mat. Appl. **7**, 5–15 (1996)

[OlShYo92] Oleinik, O.A., Shamaev, A.S., Yosifian G.A.: Mathematical Problems in Elasticity and Homogenization. North-Holland, London (1992)

[Pi87] Picard, C.: Analyse limite d'équations variationnelles dans un domaine contenant une grille. RAIRO Modél. Math. Anal. Numér. **21**, 293–326 (1987)

[SaSa82] Sanchez-Hubert, J., Sanchez-Palencia, E.: Acoustic fluid flow through holes and permeability of perforated walls. J. Math. Anal. Appl. **87**, 427–453 (1982)

[SaSa89] Sanchez-Hubert, J., Sanchez-Palencia, E.: Vibration and Coupling of Continuous Systems. Asymptotic Methods. Springer, Heidelberg (1989)

[Sa82] Sanchez-Palencia, E.: Boundary value problems in domains containing perforated walls. In: Nonlinear Partial Differential Equations and Their Applications. Collège de France Seminar, vol. III. Research Notes in Mathematics, vol. 70, pp. 309–325. Pitman, Boston (1982)

[Sa85] Sanchez-Palencia, E.: Un problème d'ecoulement lent d'un fluide incompressible au travers d'une paroi finement perforée. In: Homogenization Methods: Theory and Applications in Physics. Collect. Dir. Études Rech. Élec. France, vol. 57, pp. 371–400. Eyrolles, Paris (1985)

[Te79] Temam, R.: Navier-Stokes Equations. Theory and Numerical Analysis. Studies in Mathematics and Its Applications, vol. 2. North-Holland, Amsterdam (1979)

[Te83] R. Temam, Problèmes Mathématiques en Plasticité. Gautier Villars, Paris (1983)

Chapter 8
The Mathematical Modelling of the Motion of Biological Cells in Response to Chemical Signals

Paul J. Harris

8.1 Introduction

The motion of biological cells in response to chemical signals in the medium in which they are immersed has been observed in many experiments, such as those reported in [MaEtAl00, HoKa07, NiEtAl07, La16] for example. It is thought that the cells can sense and move in the direction in which the concentration of the chemical is increasing. This process is known as chemotaxis. In many cases the cells themselves secrete the chemical signal, often in the form of a protein, in order to signal and attract other cells that are nearby. The cells then come together to form clusters and ultimately form organs and other large structures.

Over the years a number of mathematical models for simulating how cells cluster together by following chemical signals have been proposed. The most common type of model used to solve this problem is the Keller–Segel model which calculates the relative concentrations of both the cells and the signalling chemical and simulates how the concentration of the cells changes in response to changes in the concentrations of the chemical signal (see [ChEtAl12, GaZa98, KeSe71, LaSc74] and the references therein). The concentrations of both the cells and the chemical signal are modelled using coupled diffusion-reaction equations (see [DeEtAl16, IsZa16, RiEtAl16] and the references therein). The advantage of such models is that it is relatively cheap to implement a numerical method for finding the solution of the governing equations. However, these models are not capable of simulating the motion and behaviour of individual cells.

Alternative models which consider the motion of the individual cells have also been developed. Harris [Ha17] proposed a simple method which models the motion

P. J. Harris (✉)
The University of Brighton, Brighton, UK
e-mail: p.j.harris@brighton.ac.uk

C. Constanda (ed.), *Computational and Analytic Methods in Science and Engineering*, https://doi.org/10.1007/978-3-030-48186-5_8

of each individual cell by treating each cell as a rigid particle. The spread of the chemical signal is modelled using a closed-form solution to the diffusion equation. The gradient of the concentration of the chemical is treated as a force acting on the cells and causing them to move in the direction in which the concentrations are increasing. This model is discussed in Sect. 8.3.

A more sophisticated model which uses a combined boundary element and finite element method to simulate both the fluid motion and the spread of the chemical signal has been presented in [Ha19] and is discussed in Sect. 8.4. However, this method is very computationally expensive and a new method that uses just the boundary element method to model the motion, based on the model introduced in [Ha18], is discussed in Sect. 8.5.

8.2 Description of the Problem

A complete mathematical model of the motion of a biological cell in response to a chemical signal has to address a number of processes that are taking place simultaneously. However each of these processes can potentially require a sophisticated mathematical model to fully describe what is happening. In order to reduce the complexity of the problem, the models considered here will make simplifying assumptions about some of the stages.

1. *The process by which the cell manufactures the chemical signal.*
   How a cell manufactures the chemical will not be considered in any detail in this work as a complicated mathematical model would be needed to simulate the process. Instead, it will be assumed that either a cell produces the chemical over a very short time period so that it appears to manufacture it spontaneously, or it will be treated as a source term in the differential equation that models how the chemical spreads from the cells. Mathematically the former is easier to model as it is only necessary to model the spread of a fixed mass of the chemical from the secreting cell. However, the latter is more realistic as it simulates the cell manufacturing the chemical over a time interval rather than producing the chemical instantaneously.
2. *The process by which the chemical signal spreads out from the cell which is manufacturing it.*
   In cases where the cells are not moving, the spread of the chemical signal can be modelled using the linear diffusion equation. The linear diffusion was also used in the simple model proposed in Harris [Ha18] (and described in Sect. 8.3 below) where it was assumed that the motion of the fluid can be neglected. However, the cells are usually located in a fluid which is moving due to the motion of the cells themselves. In this case the spread of the chemical signal can be modelled using the convection-diffusion equation. Since the velocity term in the equation is not constant in either space or time the convection-diffusion equation needs to be solved numerically. Harris [Ha19] showed that it is possible to model the

spread in moving fluid using finite element method. However this method, which is discussed in Sect. 8.4 below, is computationally expensive. An alternative is to use a simple linear diffusion model that moves with the emitting cell and this will be discussed in Sect. 8.5.

3. *How a separate cell detects and reacts to the chemical.*
   The cells detect the presence and concentration of the chemical through receptors in their outer membrane, with the membrane moving towards regions of higher concentration. This can be modelled as a loading term similar to a pressure term acting on the boundary of the cell. This is the method used in the model described in Sect. 8.5. In the combined finite element and boundary element model given in Sect. 8.4 it is assumed that the cell has receptors throughout its interior, whilst for the simple model in Sect. 8.3 the force is simply taken as proportional to the gradient of the chemical concentrations as the centre of the cell.
4. *The exact mechanism by which a cell moves in response to the chemical.*
   A finite element model of how a cell moves has been developed in [ElEtAl12] but the use of this model here would be computationally expensive. In addition, there is some experimental evidence that on the time-scale of their overall motion, the cells essentially move as rigid bodies (see the images in [NiEtAl07] for example). Therefore, in all the models presented here the cells are treated as rigid bodies moving in response to the external forces acting on them. In addition, there may be a stochastic element to the processes by which a cell detects the chemical signal and so a cell may not follow the chemical gradient precisely. However, the stochastic element of the cell motion has not been considered here.
5. *The motion of the surrounding fluid due to the motion of the cells.*
   In the simple model (see Sect. 8.3) the motion of the fluid is neglected. In the other models considered here, a Stokes flow model is used for the fluid motion and the governing equations are solved using the boundary integral method.
6. *What happens when two (or more) cells collide.*
   When two cells (or clusters of cells) collide, they usually combine to form a larger cluster. In the simple model where the cells are modelled as rigid particles, when the cells collide they simply stick together to form larger clusters. Quantities such as the velocity of the new cluster are calculated using appropriate conservation laws. In reality, the cells deform and change shape when they collide and such effects need to be included in the more sophisticated models. This has not been considered in the boundary integral type models presented here and the simulation simply stops when two cells (or clusters) collide. A complete model of cell collisions will be developed in the future.

In the models presented in this paper it is assumed that the layer of fluid containing the cells is thin enough that the vertical variations of both the fluid motion and concentrations of the chemical signal can be neglected so that the problem is two-dimensional. This corresponds to a typical experimental situation where the motion of the cells is observed through a microscope.

8.3 A Simple Mathematical Model

This section presents a simple mathematical model for simulating how a number of small clusters of biological cells can use chemical signals to attract other nearby clusters to combine to form larger clusters. In this model the cells are assumed to be rigid particles which have a simple geometric shape such as a circle. Further, the motion of the fluid is not modelled it is assumed that the effect of the fluid on the motion of a cell can be treated as a linear damping terms in the governing equations.

8.3.1 Description of the Model

Assume that the geometric shape of every cell is the same so that they can be modelled as circles of radius r. Let (x_i, y_i) denote the coordinates of the centre of the ith cell. If the cell is part of a cluster of cells, it is assumed that its position within the cluster does not change and that its velocity and acceleration are the same as the velocity and acceleration of the cluster. It is also assumed that all the cells have the same density and hence they must all have the same mass as they are assumed to all have the same radius.

The spread of the chemical signal through the fluid in which the cells are immersed can be modelled using the linear diffusion equation

$$\frac{\partial c}{\partial t} = D\nabla^2 c \tag{8.1}$$

where D is the diffusion constant. In the problem under consideration here the chemical signal is being manufactured and emitted by the cells, so let c_i denote the concentration of the chemical emitted by the ith cell and let

$$c(\mathbf{x}, t) = \sum_{i=1}^{N} c_i(\mathbf{x}, t)$$

where N is the total number of cells present. If the ith cell spontaneously manufactures and emits an amount A_i of the chemical signal at time t_i, then the concentration which satisfies (8.1) is given by

$$c_i = \begin{cases} \dfrac{A_i}{D(t - t_i + t_\varepsilon)} \exp\left(-\dfrac{(x - \tilde{x}_i)^2 + (y - \tilde{y}_i)^2}{4D(t - t_i + t_\varepsilon)}\right) & t \geq t_i \\ 0 & t < t_i \end{cases} \tag{8.2}$$

where $(\tilde{x}_i, \tilde{y}_i)$ is the location of the cell when $t = t_i$ and t_ε is a small time parameter used to avoid computational problems if $t = t_i$.

All of the cells can sense the direction in which the concentration of the chemical is increasing. This can be modelled as a force proportional to the gradient of the concentrations acting on the cell. Therefore the force acting on the jth cell due to the chemical signal is given by

$$\mathbf{f}_j = k\nabla c(\mathbf{x}_j, t) = k\sum_{i=1}^{N} \nabla c_i(\mathbf{x}_j, t)$$

where k is a parameter to control how strongly a cell reacts to the gradient of chemical concentrations. The total force due the chemical signal acting on a cluster of cells is given by the sum of the force acting on each individual cell in the cluster. Hence, if cluster C_j contains N_j cells, then the force acting on the cluster is

$$\mathbf{F}_j = \sum_{i=1}^{N_j} \mathbf{f}_i = k\sum_{i=1}^{N_j} \left(\sum_{l=1}^{N} \nabla c_l(\mathbf{x}_i, t) \right). \tag{8.3}$$

Using Newton's second law, the acceleration $\mathbf{a}_j$ of the cluster is given by

$$N_j\, m\, \mathbf{a}_j = \mathbf{F}_j - \lambda \mathbf{v}_j \qquad \Rightarrow \qquad \frac{d\mathbf{v}_j}{dt} = \frac{1}{N_j\, m}\left(\mathbf{F}_j - \lambda \mathbf{v}_j\right)$$

where $\mathbf{v}_j$ is the velocity of the jth cluster, λ is a constant which simulates the viscous damping of the fluid and m is the mass of a single cell. If $\delta\mathbf{x}_j$ is the displacement of the cluster, then the equations of motion of the cluster are

$$\begin{aligned} \frac{\mathrm{d}(\delta\mathbf{x}_j)}{\mathrm{d}t} &= \mathbf{v}_j \\ \frac{\mathrm{d}\mathbf{v}_j}{\mathrm{d}t} &= \frac{1}{N_j\, m}\left(\mathbf{F}_j - \lambda \mathbf{v}_j\right). \end{aligned} \tag{8.4}$$

Hence if the ith cell is contained in the jth cluster, its displacement is given by $\delta\mathbf{x}_j$.

Equations (8.2)–(8.4) give a system of differential equations in time that can be solved to give the locations of the cells at different times. This system of differential equations can be solved using any suitable numerical method, and an adaptive fourth-order Runge–Kutta scheme (as described in Harris [Ha17]) has been used here. Further details of this method, and other methods that could be used to integrate the system through time, are given in one of the many text books on numerical methods, such as [At89].

The only other aspect that needs to be considered is what happens when two clusters collide. Let C_i and C_j denote the sets of cells which are contained in clusters i and j, respectively. The two clusters will have collided if there exist a cell in C_i, with its centre located at $\mathbf{x}_i$, and cell in C_j, with its centre located at $\mathbf{x}_j$ such that

$$|\mathbf{x}_i - \mathbf{x}_j| \leq 2r.$$

Once two clusters have collided they combine to form a single, larger cluster. Using the conservation of momentum, the velocity $\mathbf{v}_{\text{new}}$ of the new cluster is given by

$$\mathbf{v}_{\text{new}} = \frac{1}{N_i + N_j}\left(N_i \mathbf{v}_i + N_j \mathbf{v}_j\right)$$

where N_i denotes the number of cells in cluster i and $\mathbf{v}_i$ is the velocity of cluster i.

8.3.2 Numerical Results with the Simple Model

The simple model described above has been used to model how cells and small clusters of cells combine to form larger clusters. Here the values of parameters used are $D = 1$, $\lambda = 50$ and $k = 1$. The distances are scaled so that the cells have radius 1, the mass of each cell is 1 unit and the values given for the times are non-dimensional.

Figure 8.1 shows a typical example with 497 cells which are initially in 200 small clusters, and where the initial locations of the cells and clusters were randomly chosen. The result presented in Fig. 8.1 show that this model can be used to simulate how cells can cluster together due to chemotaxis. Further examples which explore how changing the physical parameters such as the D and λ affects how the cells behave can be found in Harris [Ha17]. In addition, an example with a much larger number of cells and an example in three space dimensions can also be found in [Ha17].

Whilst the model is relatively simple to implement, and can rapidly simulate the motion of a large number of cells, it does not model the motion of the fluid in which the cells are immersed. The fluid effects need to be included as there is some evidence from the results of using the more sophisticated models discussed in Sects. 8.4 and 8.5 that as one cell moves the resulting fluid motion can push other nearby cells causing them to move. Further, the simple model can only be used to simulate the motion of circular cells, and most cells and clusters of cells are not circular.

More sophisticated models which can simulate the motion of the fluid due to the cell motion and more accurately portray the hydrodynamic forces on a cell are considered in the following sections. However, the computational cost of using such models means that they cannot be used with as many cells or clusters as this simple model.

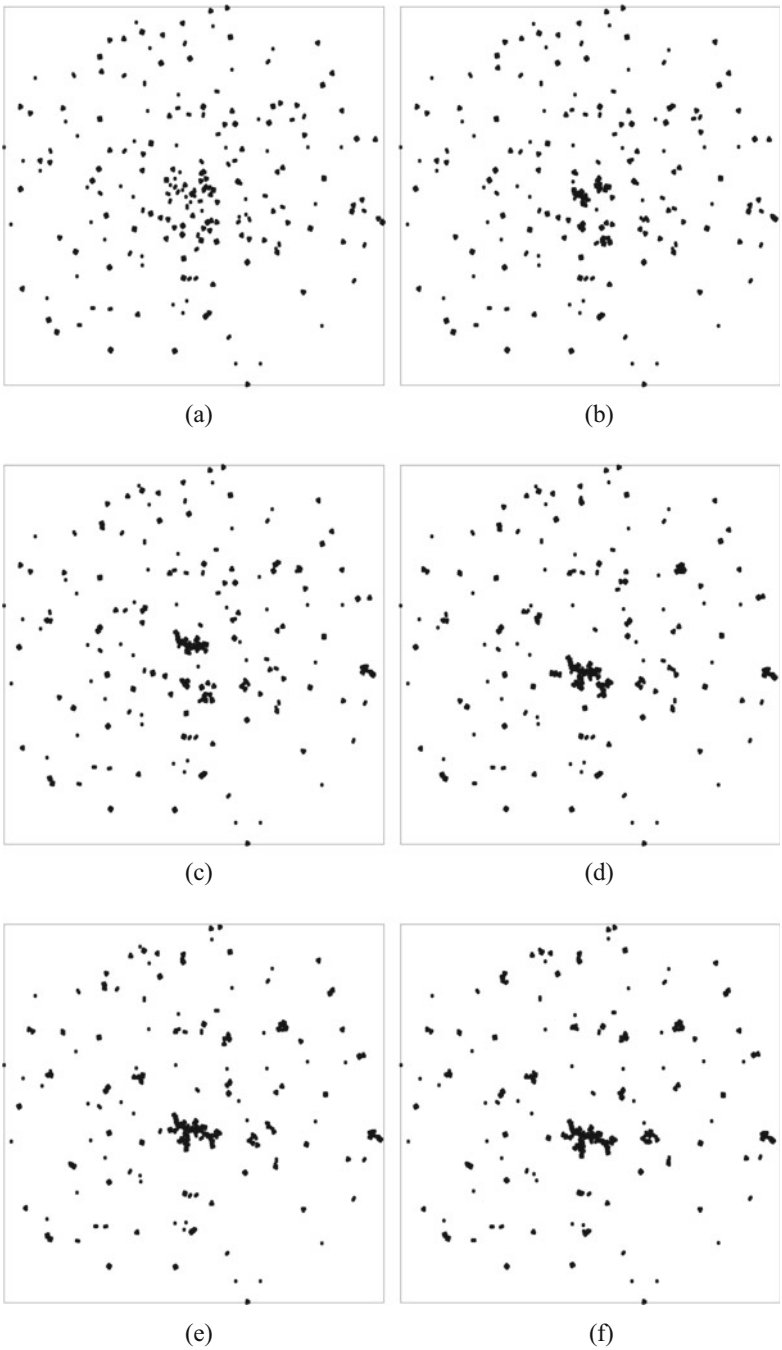


Fig. 8.1 The locations of the cells showing how they can cluster together due to chemotaxis using the simple model. (**a**) $t = 0$. (**b**) $t = 100$. (**c**) $t = 200$. (**d**) $t = 300$. (**e**) $t = 400$. (**f**) $t = 500$

8.4 Combined Finite Element and Boundary Element Methods

The model presented in Sect. 8.3 above has the advantage of being relatively simple to implement and can simulate the motion of a large number of cells. However, it is only appropriate to use this model with cells that have simple geometrical shapes (such as circles or spheres) and it does not include the motion of the fluid which surrounds the cells.

An alternative method that fully models the fluid flow as a Stokes flow has been described in Harris [Ha18] which considered the motion of cells (and clusters of cells) in response to an external chemical signal. This method was further developed in Harris [Ha19] which uses the convection-diffusion equation to model the spread of the chemical through the moving fluid. In [Ha19] the convection-diffusion is solved using the finite element method and it is this model that is described in this section. It is noted that the model presented here can be used for either simulating the motion of individual cells or for simulating the motion of clusters of cells as a cluster can be treated as being the same as a single large irregularly shaped cell.

8.4.1 Fluid Motion

Let Ω denote the region containing the cells and the surrounding fluid. Further, let Ω_i and Γ_i denote the interior and boundary of the ith cell, respectively. Finally, let Γ_0 denote the exterior boundary of the fluid region (which is needed to avoid the well-known problems with Stokes' paradox when considering a two-dimensional Stokes flow, see [Li86]) and Ω_F denote the fluid filled region inside Γ_0 and exterior to all of the cells. For convenience let $\Gamma = \bigcup_{i=0}^{N} \Gamma_i$ and $\Omega_C = \bigcup_{i=1}^{N} \Omega_i$ where N is the total number of cells.

Since the size of a cell is very small (the radius of a typical cell is of the order of 10^{-5} m) and the velocity slow (typically a cell will take minutes or hours to move a single diameter) then the Reynolds number of the flow is very small and at any instant the fluid velocity $\mathbf{u}$ can be represented as a Stokes flow

$$\begin{aligned} -\nabla p + \mu \nabla \mathbf{u} &= \mathbf{0} \\ \nabla \cdot \mathbf{u} &= 0 \end{aligned} \tag{8.5}$$

where p denotes the pressure in the fluid and μ is the dynamic viscosity. The boundary conditions for this problem are

$$\begin{aligned} \mathbf{u}(\mathbf{x}) &= \mathbf{v}_i - \omega_i \, J \, (\mathbf{x} - \mathbf{x}_i) \quad x \in \Gamma_i \\ \mathbf{u}(\mathbf{x}) &= \mathbf{0} \qquad\qquad\qquad\qquad\quad x \in \Gamma_0 \end{aligned} \tag{8.6}$$

where $\mathbf{x}_i$, $\mathbf{v}_i$ and ω_i denote the location of the centre of mass, the velocity and angular velocity of the ith cell, respectively, and

$$J = \begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix}.$$

The Stokes flow problem (8.5) subject to the boundary conditions (8.6) can be solved using the boundary integral method. It is shown in [Po92] that

$$\oint_\Gamma T(\mathbf{x}, \mathbf{x}_0)\mathbf{u}(\mathbf{x})\, \mathrm{d}\Gamma_\mathbf{x} - \oint_\Gamma G(\mathbf{x}, \mathbf{x}_0)\mathbf{F}(\mathbf{x})\, \mathrm{d}\Gamma_\mathbf{x} = \begin{cases} \mathbf{u}(\mathbf{x}_0) & \mathbf{x}_0 \in \Omega_F \\ \frac{1}{2}\mathbf{u}(\mathbf{x}_0) & \mathbf{x}_0 \in \Gamma \\ \mathbf{0} & \mathbf{x} \in \Omega_C \end{cases} \tag{8.7}$$

where $\mathbf{F}$ is the boundary force,

$$T_{ij}(\mathbf{x}, \mathbf{x}_0) = -\frac{\mathbf{r} \cdot \mathbf{n}}{\pi r^4} r_i r_j \qquad G_{ij}(\mathbf{x}, \mathbf{x}_0) = \frac{1}{4\pi\mu}\left(-\delta_{ij} \ln(r) + \frac{r_i r_j}{r^2}\right)$$

$\mathbf{r} = \mathbf{x} - \mathbf{x}_0$, $r = |\mathbf{r}|$, $\mathbf{n}$ is the unit normal to Γ directed onto the fluid domain Ω_F and δ_{ij} is the Kronecker delta function. Equation (8.7) for $\mathbf{x}_0 \in \Gamma$ yields a Fredholm integral equation of the first kind for the surface forces $\mathbf{F}$ on the whole of the boundary Γ. Once $\mathbf{F}$ has been found on Γ, (8.7) for $\mathbf{x}_0 \in \Omega_F$ can be used to find the velocity at any point in the fluid.

The boundary integral equation (8.7) for $\mathbf{x}_0 \in \Gamma$ can be solved using the boundary element method. In the work presented here a simple piecewise constant boundary element formulation has been used. Further details are not given here as a complete description of the boundary element method can be found in one of the many texts on the subject, such as [BeEtAl09].

8.4.2 *Spread of the Chemical Signal*

The concentration of a substance spreading through a moving fluid can be modelled using the convection-diffusion equation

$$\frac{\partial c}{\partial t} = \nabla \cdot (D(\mathbf{x}, t)\, \nabla c) - \nabla \cdot (c\, \mathbf{u}(\mathbf{x}, t)) + f(\mathbf{x}, t) \tag{8.8}$$

where $D(\mathbf{x}, t)$ is the diffusion parameter, $f(\mathbf{x}, t)$ is a source term and $\mathbf{u}(\mathbf{x}, t)$ is the fluid velocity. In the application under consideration here (8.8) is solved on the whole of Ω, with $D(\mathbf{x}, t)$ taking different values depending on whether that at time

t the point $\mathbf{x}$ is in the fluid or inside a cell. Hence

$$D(\mathbf{x}, t) = \begin{cases} D_F & \mathbf{x} \in \Omega_F \\ D_C & \mathbf{x} \in \Omega_C \end{cases}.$$

The source term in (8.8) can be used to simulate a cell manufacturing the chemical signal. For example, choosing

$$f(\mathbf{x}, t) = \begin{cases} 1 & \mathbf{x} \in \Omega_i \text{ and } t \leq t_i \\ 0 & \text{otherwise} \end{cases}$$

simulates the ith cell manufacturing them chemical at a unit rate over its two-dimensional area for $t \leq t_i$. This is the source term that has been used in the results presented here.

The differential equation (8.8) can be solved using the finite element method. However, the domain Ω is large compared to the size of a typical cell, and it is not computationally feasible to apply the finite element method to the whole domain. Therefore Ω is approximated by a smaller domain Ω_a. Generally, this is sufficient provided this approximate domain is chosen to be large enough that the chemical has not spread to its boundary when the end time of the simulation has been reached.

In Harris [Ha19] it is shown that the finite element approximation to (8.8) can be expressed in matrix form as

$$M\dot{\mathbf{c}} = K(t)\mathbf{c} + \mathbf{f}(t) \tag{8.9}$$

where

$$\begin{aligned} M_{ij} &= \int_{\Omega_a} \phi_i(\mathbf{x})\, \phi_j(\mathbf{x})\ \mathrm{d}\mathbf{x} \\ K_{ij}(t) &= -\int_{\Omega_a} \left[D(\mathbf{x}, t)\, \nabla\phi_j(\mathbf{x}) - \phi_j(\mathbf{x})\, \mathbf{u}(\mathbf{x}, t) \right] \cdot \nabla\phi_i(\mathbf{x})\ \mathrm{d}\mathbf{x} \\ \mathbf{f}_i &= \int_{\Omega_a} f(\mathbf{x}, t)\, \phi_i(\mathbf{x})\ \mathrm{d}\mathbf{x} \end{aligned}$$

and $\{\phi_i(\mathbf{x})\}$ denotes the set of finite element basis functions. Linear triangular basis functions have been used in this work and the full details of the finite element method, including other choices of basis functions, can be found in [ZiTa89] for example. The fluid velocity term $\mathbf{u}(\mathbf{x}, t)$ can be computed using the boundary integral equation (8.7) for $\mathbf{x_0} \in \Omega_F$ if the point $\mathbf{x}$ is in the fluid, or is taken to be the velocity of the cell if the point is inside one of the cells.

8.4.3 Time Integration

The total force acting on each cell is the sum of the fluid forces on the surface of the cell, and the reaction of the cell in the direction in which the concentration of the chemical signal is increasing over the interior of the cell. Hence the force acting on the ith cell is

$$\int_{\Omega_i} k_i \, \nabla c(\mathbf{x}, t) \; \mathrm{d}\Omega + \oint_{\Gamma_i} \mathbf{F}(\mathbf{x}, t) \; \mathrm{d}\Gamma$$

where, as before, k_i is a parameter which controls how strongly the cell reacts to the gradient of the chemical concentration. However, this may not be a realistic model of how a cell detects the gradient of the concentration of the chemical as generally it is thought that cells only have chemical receptors in their outer membranes. A more realistic way of determining the force due to the gradient of the chemical concentration is discussed and used in the next section.

Using Newton's second law the acceleration of the cell can be expressed as

$$\mathbf{a}_i(t) = \frac{\mathrm{d}\mathbf{v}_i}{\mathrm{d}t} = \frac{1}{m_i}\left[\int_{\Omega_i} k_i \, \nabla c(\mathbf{x}, t) \; \mathrm{d}\Omega + \oint_{\Gamma_i} \mathbf{F}(\mathbf{x}, t) \; \mathrm{d}\Gamma\right]$$

where m_i is the mass of the cell. Similarly, the angular acceleration of the cell can be expressed as

$$\alpha_i = \frac{\mathrm{d}\omega_i}{\mathrm{d}t} = \frac{1}{I_i}\left[\int_{\Omega_i} k_i (\mathbf{x} - \mathbf{x}_i)^T \, J \, \nabla c(\mathbf{x}, t) \; \mathrm{d}\Omega + \oint_{\Gamma_i} (\mathbf{x} - \mathbf{x}_i)^T \, J \, \mathbf{F}(\mathbf{x}, t) \; \mathrm{d}\Gamma\right]$$

where I_i is the moment of inertial of the ith cell, and recall that

$$J = \begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix}.$$

The location and velocity of each of the cells can be updated using

$$\left.\begin{aligned} \mathbf{v}_i(t+h) &= \mathbf{v}_i(t) + h \, \mathbf{a}_i(t) \\ \omega_i(t+h) &= \omega_i(t) + h \, \alpha_i(t) \\ \mathbf{x}_i(t+h) &= \mathbf{x}_i(t) + \frac{h}{2}\left[\mathbf{v}_i(t) + \mathbf{v}_i(t+h)\right] \\ \theta_i(t+h) &= \theta_i(t) + \frac{h}{2}\left[\omega_i(t) + \omega_i(t+h)\right] \end{aligned}\right\} \qquad i = 1, \ldots, N$$

where θ_i is the rotation of the ith cell about its centre of mass.

8.4.4 Numerical Results Using the Combined Model

The results presented here are for some typical examples which can be used to study how the chemical concentrations spread out from the cells and moves with the cells. A more complete set of results which also consider the accuracy of the method can be found in Harris [Ha19].

Figure 8.2 shows the results of using the combined finite element and boundary integral method for modelling the motion of two circular cells in a viscous fluid where the source term in the convection-diffusion equation (8.8) is

$$f(\mathbf{x}, t) = \begin{cases} 1 \ \mathbf{x} \in \Omega_1 \text{ and } t \leq 3 \\ 0 \text{ otherwise.} \end{cases}$$

In Fig. 8.2 the cell on the left is Ω_1 and the cell on the right is Ω_2. Here regions with high concentrations of the chemical signal a coloured red, and regions with low concentrations coloured blue. The corresponding results for two elongated cells are shown in Fig. 8.3. It is noted that Fig. 8.3 shows that if the emitting cell is not circular, then the chemical does not spread out from the cell in a circular pattern.

In the examples shown in Figs. 8.2 and 8.3 the chemical spreads out from the emitting cell, as expected. The results also show that as the second cell starts to move in response to the chemical signal, it causes the fluid to start moving and in turn this causes the first cell also to move. Hence these results show that the hydrodynamic effects of the cell motion are important and need to be included. These effects are not included in the simple model discussed in Sect. 8.3 which is a major disadvantage of the simple model.

The results also show that as the cell secreting the chemical moves and causes the fluid around it to move, the chemical signal moves with cell and the fluid. Hence, from the perspective of a point in the cell, the chemical seems to simply spread out according to the linear diffusion equation as the concentrations are moving with the cell. It might be possible to exploit this to avoid having to solve the convection-diffusion equation. This would remove the need for using the finite element method which, in turn, would greatly reduce the computational cost of the model. This approach to simulating the cell motion is discussed in Sect. 8.5 below.

8.5 Simplified Boundary Integral Model

A version of the boundary integral method was used in Harris [Ha18] to model the motion of cells due to a chemical signal which was simply present in the surrounding fluid. The results presented in [Ha18] compared well to the motion of clusters observed in experiments, although it is likely that the chemical signal was secreted by a cell (or cluster of cells) that was outside the field of view in the experiments rather than the chemical just being present in the fluid.

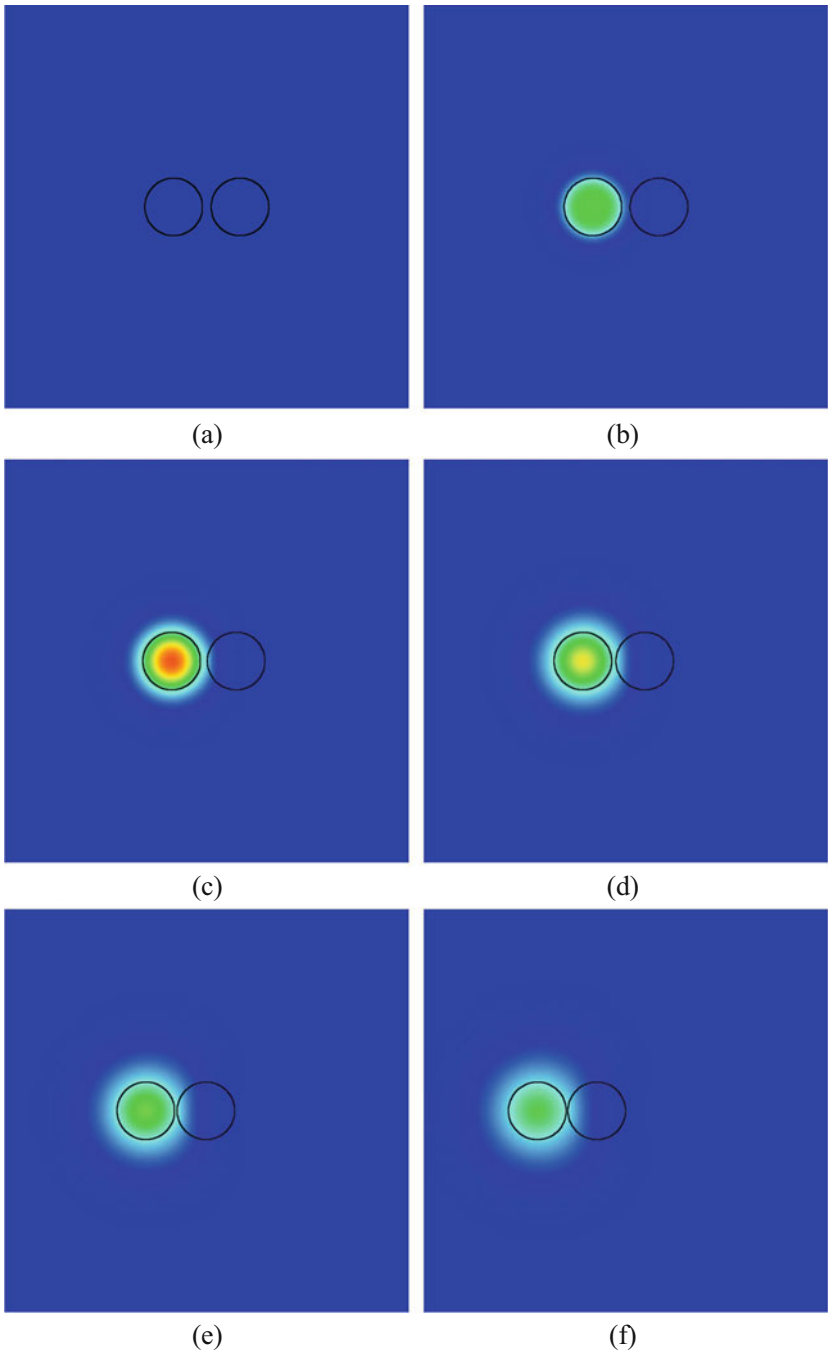


Fig. 8.2 The simulated motion of two circular cells and the concentrations of the chemical using the combined finite element and boundary integral method. (**a**) $t = 0.0$. (**b**) $t = 1.6$. (**c**) $t = 3.2$. (**d**) $t = 4.8$. (**e**) $t = 6.4$. (**f**) $t = 8.0$

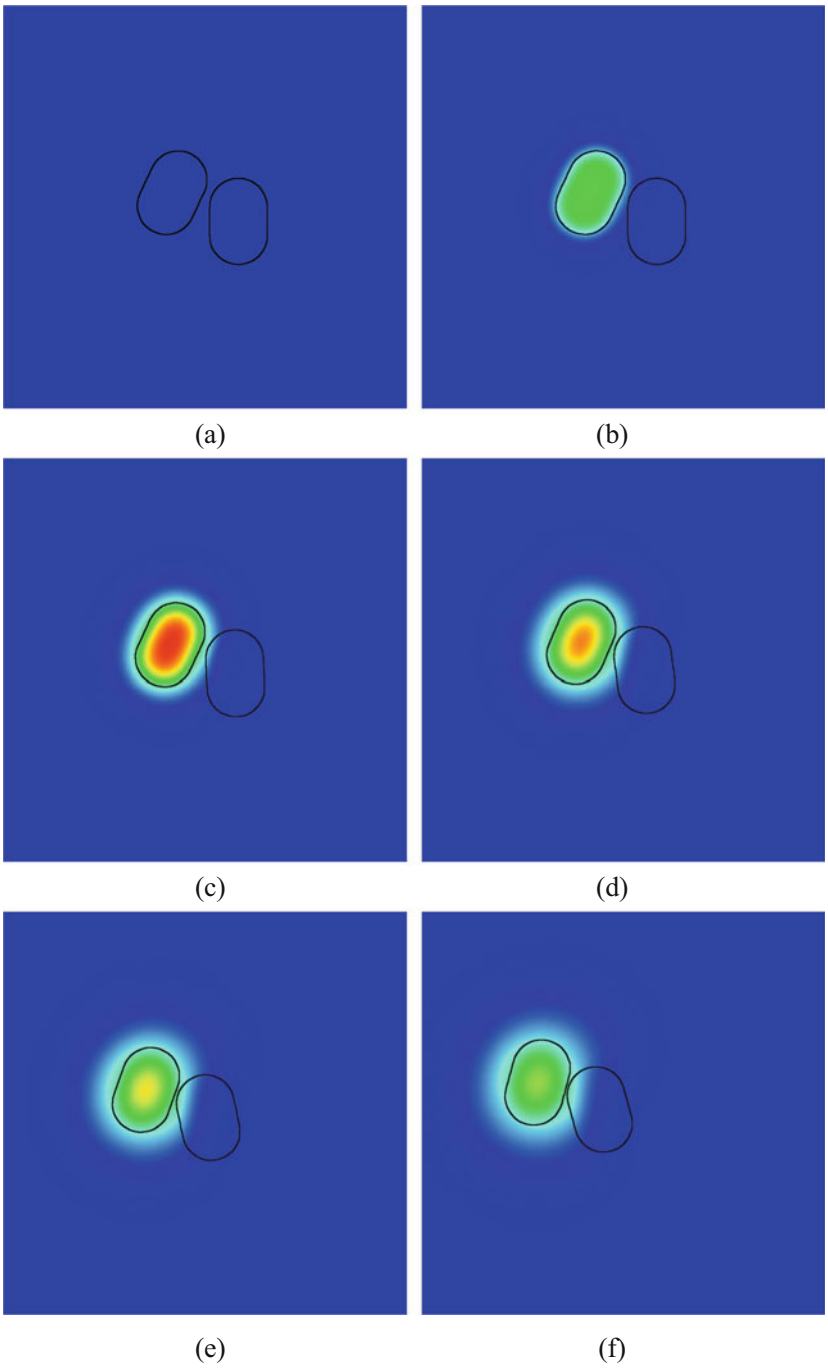


Fig. 8.3 The simulated motion of two elongated cells and the concentrations of the chemical using the combined finite element and boundary integral method. (**a**) $t = 0.0$. (**b**) $t = 1.6$. (**c**) $t = 3.2$. (**d**) $t = 4.8$. (**e**) $t = 6.4$. (**f**) $t = 8.0$

The main drawback of the combined finite element and boundary element model introduced in Sect. 8.4 is the computational cost. Since the finite element stiffness matrix appearing in (8.9) depends on the current locations of the cells it needs to be calculated at each time step. This cost is further increased as the fluid velocity is needed at every quadrature point used to find the stiffness matrix and this is found by using (8.7) for $\mathbf{x}_0 \in \Omega_F$ which requires the evaluation of the boundary integral terms. From a computational perspective, it would be better if this could be avoided.

The results presented in Sect. 8.4.4 above show that when a moving cell secretes the chemical signal the concentration of the spreading chemical moves with the cell. Hence the concentrations of the spreading chemical can be expressed in terms of a solution to the diffusion equation which moves with the cell.

In the case of a circular cell emitting the chemical, the concentration at the point (x, y) can be expressed as

$$c(x, y, t) = \frac{A}{D(t + t_\varepsilon)} \exp\left(-\frac{(x - x_i(t))^2 + (y - y_i(t))^2}{4D(t + t_\varepsilon)}\right) \tag{8.10}$$

where A is the magnitude of the chemical signal and t_ε is a small time that is used to avoid computational problems when $t = 0$. Recall that D is the diffusion parameter. However, if the cell is not circular, then the results in Fig. 8.3 above show that the chemical will not spread out in a circular pattern and so for cells which are not circular (8.10) is not the most appropriate way of representing the chemical concentrations. An alternative method that can be used with cells that are not circular is currently being developed.

Since the velocities of the cells are known at each time step, the same boundary integral methods as described in Sect. 8.4.1 can be used to calculate the hydrodynamic forces acting on the boundary of cell.

Since it is thought that cells only have receptors for the chemical in their outer membrane, it is more realistic to calculate the force due to the chemical signal by integrating over the boundary of the cell rather than over the interior of the cell. Hence in this case the acceleration and angular acceleration of the cell are given by

$$\mathbf{a}_i(t) = \frac{1}{m_i}\left[\oint_{\Gamma_i} k\, \nabla c(\mathbf{x}, t)\ \mathrm{d}\Gamma + \oint_{\Gamma_i} \mathbf{F}(\mathbf{x}, t)\ \mathrm{d}\Gamma\right]$$

and

$$\alpha_i = \frac{1}{I_i}\left[\oint_{\Gamma_i} k(\mathbf{x} - \mathbf{x}_i)^T\, J\, \nabla c(\mathbf{x}, t)\ \mathrm{d}\Gamma + \oint_{\Gamma_i} (\mathbf{x} - \mathbf{x}_i)\, J\, \mathbf{F}(\mathbf{x}, t)\ \mathrm{d}\Gamma\right]$$

respectively. Here the required gradient of the chemical concentrations can be found by differentiating (8.10). Hence the equations of motion for the cells are

$$\frac{\mathrm{d}\mathbf{v}_i}{\mathrm{d}t} = \mathbf{a}_i \qquad \frac{\mathrm{d}\mathbf{x}_i}{\mathrm{d}t} = \mathbf{v}_i$$
$$\frac{\mathrm{d}\omega_i}{\mathrm{d}t} = \alpha_i \qquad \frac{\mathrm{d}\theta_i}{\mathrm{d}t} = \omega_i$$

which can be integrated through time using a fourth-order Runge–Kutta scheme. The full details of the Runge–Kutta scheme are not given here and can be found in any text on numerical methods, such as [At89].

8.5.1 Numerical Results with the Boundary Integral Method

A test problem of five cells arranged in a cross pattern is used to validate the method. The initial configuration of the cells can be seen in Fig. 8.4a, where each of the outer cells are located the same distance from the central cell. In this example the chemical signal is secreted by the central cell, and so the motion should be symmetric in the sense that all of the outer cells should move the same distance towards the central cell at each time step. Additionally, the central cell should remain in the same position. The results in Fig. 8.4b, f show that the four outer cells move towards the central cell as expected. Figure 8.5 shows the distance that each of the four outer cells have moved towards the central cell over time. The four curves are superimposed which shows that the motion of the cells has the expected symmetry. Further, the distance moved by the central cell is of the order of 10^{-14} which is the magnitude of the rounding error in the computer used to perform the calculations and so can be considered to be zero, as expected. A second example is shown in Fig. 8.6 where the boundary integral method has been used to simulate the motion of 10 randomly placed cells.

These results show that this simple boundary integral method for the fluid mechanics along with a simple representation of the concentrations of the chemical secreted by one of the cells can be used to simulate the motion of cells due to chemotaxis. Although the combined finite element and boundary element method give a complete solution to the problem, there are major computational drawbacks to use the combined model. For example, even when a coarse finite element mesh is used the combined method can take over 24 h to simulate the motion of just two cells on a typical PC, whilst the boundary integral method for simulating the motion of the ten cells shown in Fig. 8.6 took less than 2 h on the same PC.

However, the boundary integral model, as presented here, can only be used for circular cells. The results of using the combined method for elongated cells, presented in Fig. 8.3, show that for such cells the concentrations of the chemical do not spread out in a circular pattern. A modified method which can be used for cells that are not circular is currently being developed.

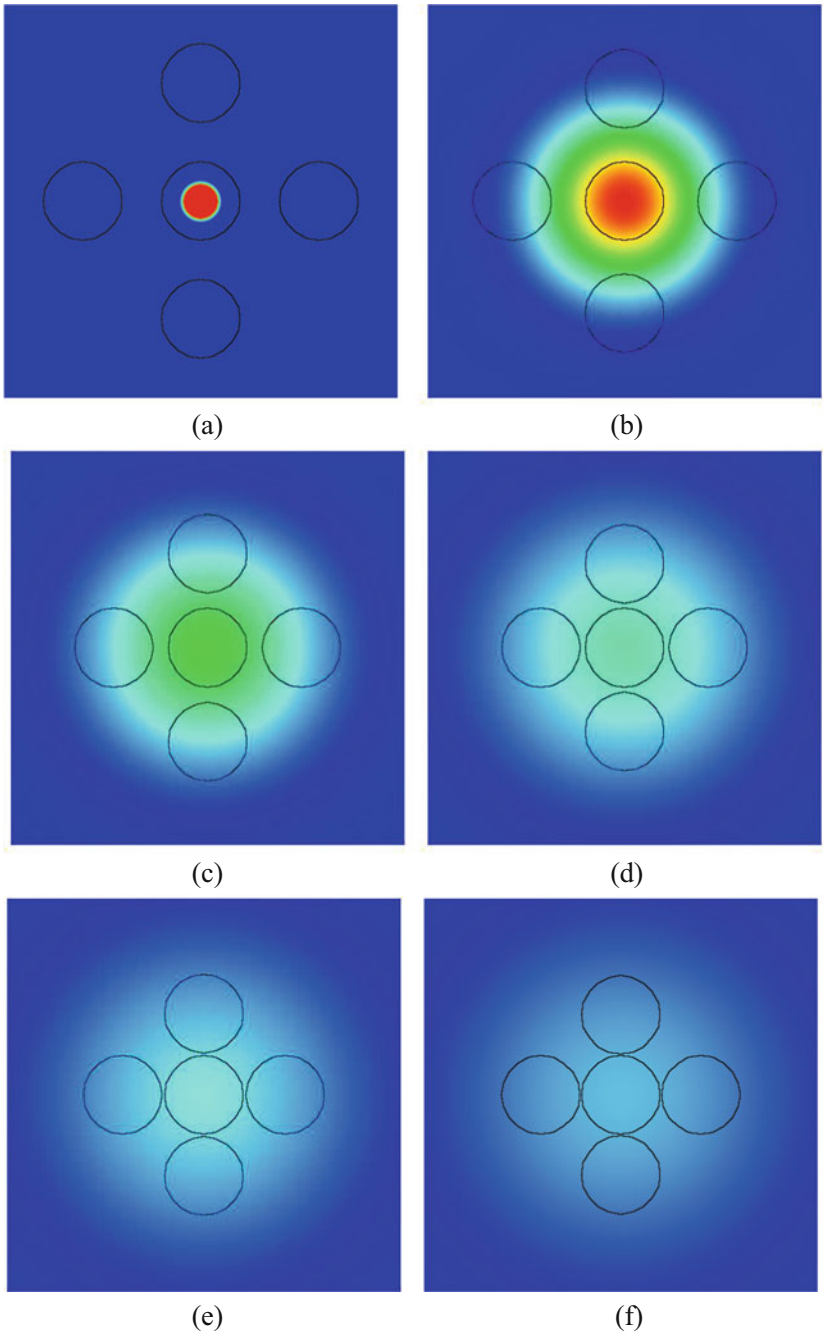


Fig. 8.4 The motion of five cells, initially arranged in a cross pattern, using the boundary element method and (8.10) for the chemical concentrations. (**a**) $t = 0$. (**b**) $t = 1$. (**c**) $t = 2$. (**d**) $t = 3$. (**e**) $t = 4$. (**f**) $t = 5$

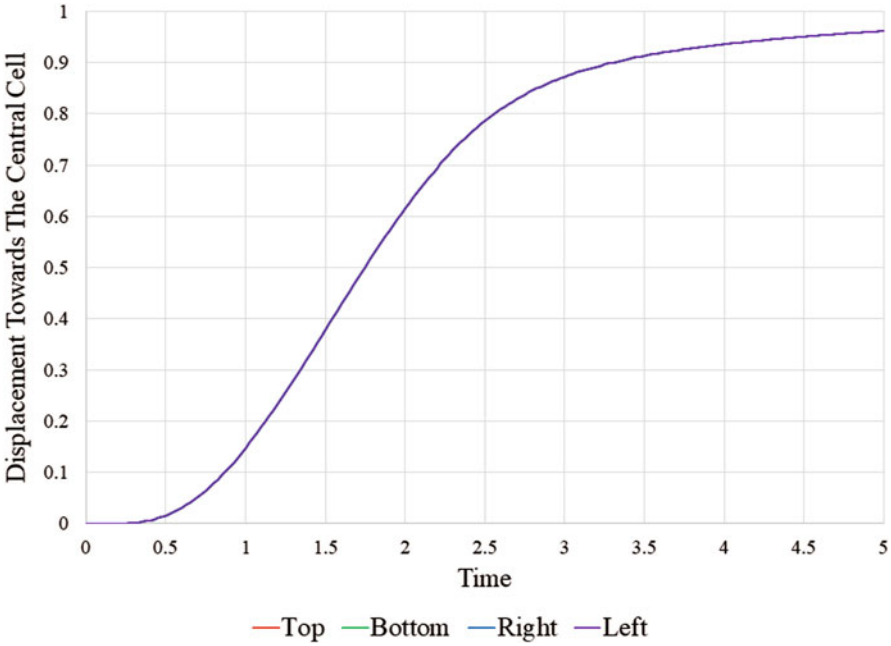


Fig. 8.5 The distance moved by each of the four outer cells towards the central cell

8.6 Conclusions and Future Work

This paper has discussed a number of methods for modelling the motion of cells due to chemotaxis. The simple model discussed in Sect. 8.3 has the advantage that it can rapidly calculate the motion of a large number of cells. The calculations for the example presented here with 497 cells initially arranged in 200 small clusters took approximately 1 h of CPU time to complete. By using this method it is possible to quickly see the patterns that the cells form, but the model does not fully include the fluid dynamics of the problem, and can only be used for circular cells.

The combined finite element and boundary element method gives a complete simulation of the cell motion and fluid flow. However, the method is very expensive computationally. The calculations with only two cells can often take over 24 h on the same computer as used for the simple model unless very coarse finite element and boundary element meshes are used. However, the method can be used for cells that are not circular, and can be used for irregularly shaped clusters of cells.

As a compromise, a model using the boundary element method to compute the fluid motion and a moving linear diffusion model for the concentrations of the chemical signal is being developed. This model has the advantage that it avoids having to use the finite element method to calculate the concentrations. The calculations for the 10 cell example given in Sect. 8.5.1 required approximately 81 min CPU time on the same computer as was used for the other examples. This

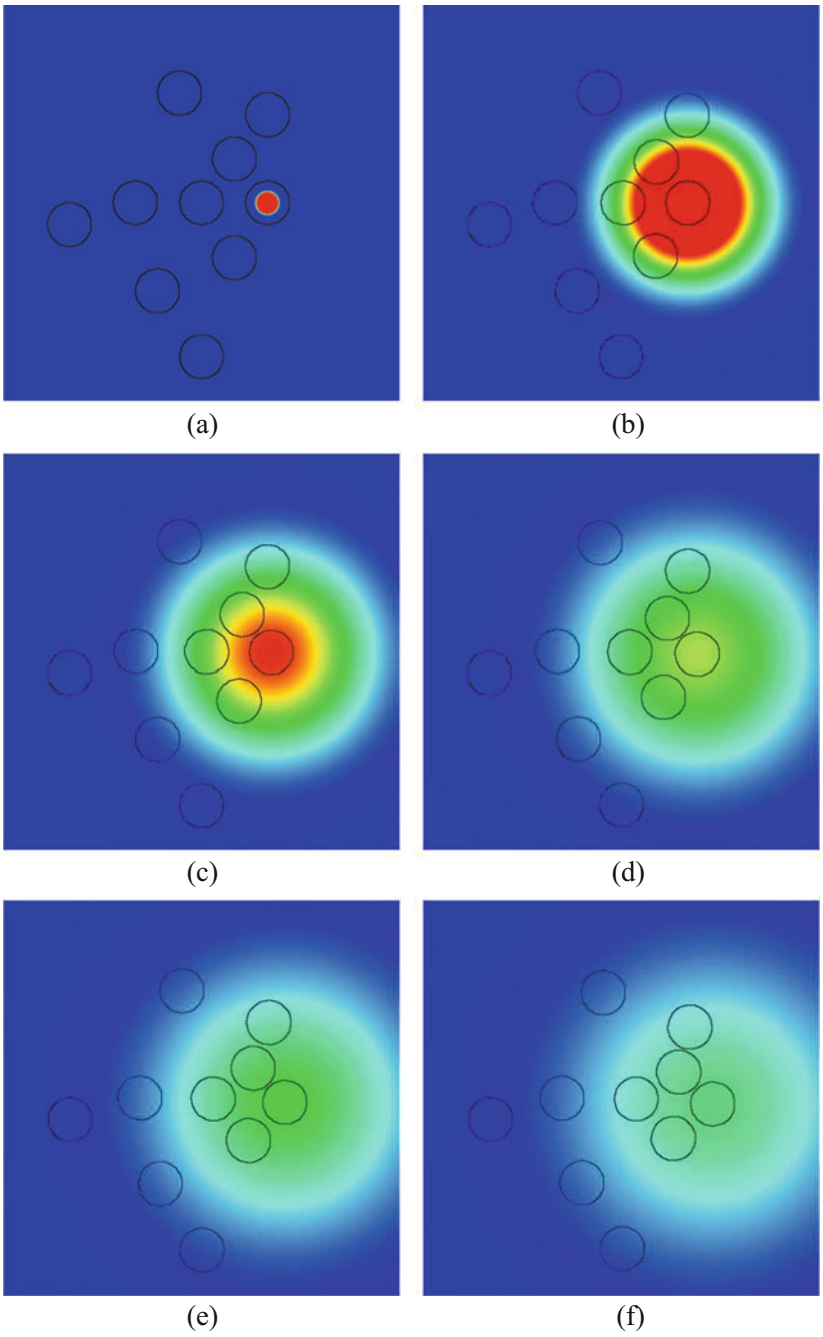


Fig. 8.6 The simulated motion of ten randomly positioned circular cells using the simple boundary integral method. (**a**) $t = 0$. (**b**) $t = 1$. (**c**) $t = 2$. (**d**) $t = 3$. (**e**) $t = 4$. (**f**) $t = 5$

shows the considerable saving in CPU time that is obtained by avoiding having to use the finite element method for calculating the chemical concentrations.

The drawback of the boundary element method described in Sect. 8.5 is that the formula used for the chemical concentrations is only valid when the secreting cell is circular, although there are no restrictions on the shapes of the other, non-emitting, cells. An alternative method for modelling the secretions from a cell which is not circular is currently begin developed.

Another deficiency of both the combined method and the boundary integral method is that they do not include any simulations of what happens when two cells or clusters of cells collide. When two cells do collide they usually deform as the parts of their membranes that are in contact stick together causing the two cells to change shape. Some of this process is analogous to the initial stages of two liquid droplets combining together to form a single droplet, so the same methods could be used to model cell collisions. A model that simulates the collision of two cells will be developed in the future and incorporated into the boundary integral model presented here.

Acknowledgments The author would like to thank Matteo Santin from Brighton Centre for Regenerative Medicine and Devices for his help and advice with some of the biological aspects of this work.

References

[At89] Atkinson, K.E.: An Introduction to Numerical Analysis, 2nd edn. Wiley, New York (1989)

[BeEtAl09] Beer, G., Smith, I.M., Duenser, C.: The Boundary Element Method with Programming. Springer, Vienna (2008)

[ChEtAl12] Chertock, A., Kurganov, A., Wang, X., Wu, Y.: On a chemotaxis model with saturated chemotactic flux. Kinet. Relat. Model **5**, 51–95 (2012)

[CrNi47] Crank, J., Nicolson, P.: A practical method for numerical evaluation of solutions of partial differential equations of the heat-conduction type. Math. Proc. Camb. Philos. Soc. **43**, 50–67 (1947)

[DeEtAl16] Deleuze, Y., Chiang, C., Thiriet, M., Sheu, T.W.H.: Numerical study of plume patterns in a chemotaxis-diffusion-convection coupling system. Comput. Fluids **126**, 58–70 (2016)

[ElEtAl12] Elliott, C.M., Stinner, B., Venkataraman, C.: Modelling cell motility and chemotaxis with evolving surface finite elements. J. R. Soc. Interface **9**, 3027–2044 (2012)

[GaZa98] Gajewski, H., Zacharias, K.: Global behaviour of a reaction - diffusion system modelling chemotaxis. Math. Nachr. **195**, 77–114 (1998)

[GrMe16] Green, E.R., Mecsas, J.: Bacterial secretion systems: an overview. Microbiol. Spectr. **4**, 1–19 (2016)

[Ha17] Harris, P.J.: A simple mathematical model of cell clustering by chemotaxis. Math. Biosci. **294**, 62–70 (2017)

[Ha18] Harris, P.J.: Modelling the motion of clusters of cells in a viscous fluid using the boundary integral method. Math. Biosci. **360**, 141–152 (2018)

[Ha19] Harris, P.J.: A combined boundary element and finite element model of cell motion due to chemotaxis. In: Constanda, C., Harris, P.J. (eds.) Integral Methods in Science and Engineering: Analytic Treatment and Numerical Approximations, pp. 163–172. Birkhäuser, Basel (2019)

[HoKa07] Hoeller, O., Kay, R.: Chemotaxis in the absence of pip3 gradients. Curr. Biol. **17**, 813–817 (2007)

[IsZa16] Islam, S., Zaman, R.: A computational modeling and simulation of spatial dynamics in biological systems. Appl. Math. Mod. **40**, 4524–4542 (2016)

[KeSe71] Keller, E.F., Segel, L.A.: Model for chemotaxis. J. Theor. Biol. **30**, 225–234 (1971)

[La16] Laganenka, L., Colin, R., Sourjik, V.: Chemotaxis towards autoinducer 2 mediates autoaggregation in Escherichia coli. Nat. Commun. **7**, 13979 (2016)

[LaSc74] Lapidus, I.R., Schiller, R.: A mathematical model for bacterial chemotaxis. Biophys. J. **14**, 825–834 (1974)

[Li86] Lighthill, M.J.: An Informal Introduction to Theoretical Fluid Mechanics. Clarendon Press, Oxford (1986)

[MaEtAl00] Malawista, s., Chevance, A., Boxer, L.: Random locomotion and chemotaxis of human blood polymorphonuclear leukocytes from a patient with Leukocyte Adhesion Deficiency-1: normal displacement in close quarters via chimneying. Cell Motil. Cytoskeleton **46**, 183–189 (2000)

[Ma99] Mazumdar, J.: The mathematics of diffusion. In: An Introduction to Mathematical Physiology and Biology. Cambridge University Press, Cambridge (1999)

[NiEtAl07] Nitta, N., Tsuchiya, T., Yamauchi, A., Tamatani, T., Kanegasaki, S.: Quantitative analysis of eosinophil chemotaxis tracked using a novel optical device–TAXIScan. J. Immunol. Methods **320**, 155–163 (2007)

[Po92] Pozrikidis, C.: Boundary Integral and Singularity Methods for Linearized Viscous Flow. Cambridge Texts in Applied Mathematics. Cambridge University Press, Cambridge (1992)

[RiEtAl16] Ritter, J., Klar, A., Schneider, F.: Partial-moment minimum-entropy models for kinetic chemotaxis equations in one and two dimensions. J. Comput. Appl. Math. **306**, 300–315 (2016)

[ZiTa89] Zienkiewicz, O.C., Taylor, R.L.: The Finite Element Method, 4th edn. McGraw-Hill Book Company Europe, London (1989)

Chapter 9
Numerical Calculation of Interior Transmission Eigenvalues with Mixed Boundary Conditions

Andreas Kleefeld and Jijun Liu

9.1 Introduction

A transmission eigenvalue problem is a non-classical boundary value problem for a specified differential operator which acts on a pair of functions $(u(x), v(x))$ in some given open and bounded domain D, where the functions $u(x)$ and $v(x)$ are coupled on the boundary $\partial D = \Gamma$. The exterior normal on Γ is denoted by ν.

A typical example arising in acoustic wave scattering is the (classical) interior transmission eigenvalue problem with specified refraction index $n(x) \in L^\infty(D)$ satisfying $\mathrm{Re}\{n(x)\} > 0$ and $\mathrm{Im}\{n(x)\} \geq 0$. It is given by

$$\begin{cases} \Delta u + k^2 u = 0\,, & x \in D\,, \\ \Delta v + k^2 n(x) v = 0\,, & x \in D\,, \\ u = v\,, \quad \frac{\partial u}{\partial \nu} = \frac{\partial u}{\partial \nu}\,, & x \in \Gamma\,, \end{cases} \tag{9.1}$$

for which one tries to find $k \in \mathbb{C}\backslash\{0\}$ such that there exists non-trivial solutions $(u, v) \in L^2(D) \times L^2(D)$ and $u - v \in H_0^2(D)$. These values k are called (classical) interior transmission eigenvalues (ITEs).

Originally, the distribution properties of the eigenvalues such as discreteness and their asymptotic behavior have been studied in detail in order to determine characteristics of media [CoMo87, CoMo88, Ki86] as well as the existence [CaGiHa10]. These properties are not trivial to derive due to the fact that the underlying eigen-

A. Kleefeld (✉)
Forschungszentrum Jülich GmbH, Jülich Supercomputing Centre, Jülich, Germany
e-mail: a.kleefeld@fz-juelich.de

J. Liu
School of Mathematics/Seu-Yau Center, Southeast University, Nanjing, China
e-mail: jjliu@seu.edu.cn

C. Constanda (ed.), *Computational and Analytic Methods in Science and Engineering*, https://doi.org/10.1007/978-3-030-48186-5_9

value problem is neither elliptic nor self-adjoint. Hence, it cannot be investigated by using standard spectral theory for differential operators. Hence, this interior transmission problem is of interest to researchers working on non-standard spectral problems. Additionally, researchers are interested in finding incident waves that do not scatter which is closely related to the interior eigenvalue problem (9.1) (see for example [GiPa13]).

Reconstruction algorithms for inverse scattering problems are, for example, the linear sampling method and the factorization method [CaCo06, KiGr08] which are not justified theoretically for wave numbers that are ITEs. Hence, researchers started to compute such exceptional values. It has been shown that ITEs can be determined from scattered data or far-field data [CaCoHa10]. Since then, a variety of new methods such as FEM [LiHuLiLi15, Su11], BEM [Kl13, Kl15], and the inside-outside-duality method [KiLe13] have appeared (see [KlPi18] for a recent and detailed overview as well as the MFS method). However, the numerical calculation of those is still an on-going and challenging research topic especially the calculation of complex-valued ITEs whose existence is still open for general scatterer.

However, motivated by a more general physical configuration, the transmission eigenvalue problems may be of a more complicated form. We consider the situation where an inhomogeneous obstacle D is located in a perfect conducting substrate D_2 with the boundary $\Gamma_2 \subset \Gamma$, while the remaining part of the boundary $\Gamma_1 = \Gamma \setminus \Gamma_2$ contacts with the surface of background dielectric medium D_1. See Fig. 9.1 for an illustration of the situation. Then the following interior transmission problem with mixed boundary condition arises (see also [YaMo14] and [LiLi16]):

$$\begin{cases} \Delta u + k^2 u = 0\,, & x \in D\,, \\ \Delta v + k^2 n v = 0\,, & x \in D\,, \\ u = v\,, \quad \frac{\partial u}{\partial \nu} = \frac{\partial v}{\partial \nu}\,, & x \in \Gamma_1\,, \ \text{(transmission condition)} \\ u = v = 0\,, & x \in \Gamma_2\,, \ \text{(hom. Dirichlet condition)} \end{cases} \tag{9.2}$$

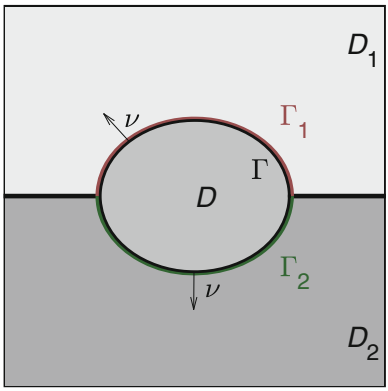


Fig. 9.1 Exemplary setup of the physical configuration

with $\Gamma = \Gamma_1 \cup \Gamma_2$ assuming $\Gamma_1 \neq \emptyset$ and $\Gamma_2 \neq \emptyset$. Here, $n(x) \neq 1$ is the real-valued index of refraction. Although the distributions and the discreteness properties have been analyzed in [LiLi16] for general complex-valued $n(x)$, the efficient numerical calculation of mixed interior transmission eigenvalues (MITEs) is still absent. This is mainly due to the presence of the mixed boundary condition on Γ which causes extra difficulties for the construction of suitable basis functions that are needed in the finite element method.

But the boundary integral equation method is a powerful tool for solving boundary value problems of partial differential equations (PDEs), especially when the problem is homogeneous and the fundamental solution to the corresponding PDE can be represented explicitly. The main advantage of solving boundary value problems by this scheme is by representing the solution in potential form, the solution can be essentially converted into the task of finding a density function defined on the boundary of the domain, and consequently the amount of computations can be dramatically decreased (see [Kl12a]). By this motivation, we propose to solve the mixed transmission eigenvalue problem (9.2) with constant refraction index $n(x) \equiv n$ (constant) in $D \subset \mathbb{R}^2$ by the boundary integral equation method since the fundamental solution can be given explicitly in analytic form.

9.1.1 Contribution Within This Chapter

First, a short summary for the existence and discreteness for the real-valued index of refraction not equal to one is given in Sect. 9.2. Although the results are a special case of [LiLi16], the sufficient conditions on the index of refraction as well as the estimates of the lower bound of positive eigenvalues can be stated more clearly. Second, a derivation of a system of boundary integral equations to solve the mixed interior transmission problem including its approximation via boundary element collocation method leading to a non-linear eigenvalue problem is given in Sect. 9.3. Lastly, extensive numerical results for the computations of mixed interior transmission eigenvalues are presented for the first time for various scatterers in two dimensions in Sect. 9.4. In addition, the corresponding eigenfunctions are shown as well. A short summary and conclusion is given in Sect. 9.5. Finally, for the special case of the unit square an alternative method is given in the Appendix to find mixed interior transmission eigenvalues.

9.2 A Review on Some Theoretical Results

To consider the numerical computations for the transmission eigenvalues of (9.2), we first need the distribution properties of the eigenvalues. Although the properties of the eigenvalues with classical boundary conditions have been thoroughly studied, the theoretical results for eigenvalues using mixed boundary conditions as studied

here are still rare (see [LiLi16]). Since our numerical scheme for computing the eigenvalues by boundary integral equations is established for the constant index of refraction $n(x) \equiv n_0$ in $\overline{D}$ for $0 < n_0 < 1$ or $n_0 > 1$, some existing results for the distributions of eigenvalues under the assumption either $n(x) \in (0, 1)$ or $n(x) > 1$ are applicable.

In this section, we give these theoretical properties of the transmission eigenvalues for $n(x) \in C(D)$ with either positive lower point $n_- > 1$ or positive upper bound $n_+ < 1$, which means that there are no zero points of $n(x) - 1$ in $\overline{D}$. These results can be considered as special cases established in [LiLi16] for both complex-valued refraction index $n(x)$ and complex-valued background medium. However, in our case with real-valued index of refraction $n(x)$, the corresponding results can be much more simplified. To state the results clearly, we introduce the Sobolev space

$$\tilde{H}_{0,1}(D) = \Big\{ w \in L^2(D), \quad \nabla w \in (L^2(D))^2, \quad \Delta w \in L^2(D), \\ w = 0 \text{ on } \Gamma, \quad \nu \cdot \nabla w = 0 \text{ on } \Gamma_1 \Big\},$$

with scalar product

$$\langle u, v \rangle_{\tilde{H}_{0,1}(D)} = (u, v)_{L^2(D)} + (\nabla u, \nabla v)_{L^2(D)} + (\Delta u, \Delta v)_{L^2(D)}$$

for two complex-valued functions u and v from $\tilde{H}_{0,1}(D)$. Then the transmission eigenvalue problem (9.2) can be restated as: Find $k \in \mathbb{C}\backslash\{0\}$ such that there exists a non-zero pair $(u, v) \in (L^2(D))^2$ satisfying $v - u \in \tilde{H}_{0,1}(D)$ and

$$\begin{cases} \Delta u + k^2 u = 0, & x \in D, \\ \Delta v + k^2 n v = 0, & x \in D, \\ u = v = 0, & x \in \Gamma_2. \text{ (hom. Dirichlet condition)} \end{cases}$$

Note that the transmission conditions in (9.2) have been incorporated in the requirement $v - u \in \tilde{H}_{0,1}(D)$. In the case $\Gamma_2 = \emptyset$, it is well-known based on the analytic Fredholm theorem that the set of ITEs is at most discrete with $+\infty$ as the only possible accumulation point. We will prove that such a property is also true for our problem (9.2) with mixed boundary condition. Therefore, define $z = v - u$ and $n_c(x) = n(x) - 1 \neq 0$ in $\overline{D}$. Since $z \in \tilde{H}_{0,1}(D)$ fulfills

$$\Delta z + k^2 z = -k^2 v n_c, \tag{9.3}$$

by deleting v, $z(x)$ satisfies the following differential equation of fourth order

$$\left(\Delta + k^2 n(x)\right) \frac{1}{n_c(x)} \left(\Delta + k^2\right) z = 0, \quad x \in D. \tag{9.4}$$

Using (9.3) and the boundary condition $v|_{\Gamma_2} = 0$, we have

$$\frac{1}{n_c(x)}\left(\Delta + k^2\right) z = 0 \qquad \text{on } \Gamma_2 .$$

With $z|_{\Gamma_2} = v|_{\Gamma_2} - u|_{\Gamma_2} = 0$ it can be further simplified to

$$\frac{1}{n_c(x)}\Delta z = 0 \qquad \text{on } \Gamma_2 .$$

Therefore, we conclude that the transmission eigenvalues $k \in \mathbb{C}\backslash\{0\}$ are those values such that there exists some non-trivial solution $z \in \tilde{H}_{0,1}(D)$ satisfying

$$\begin{cases} \left(\Delta + k^2 n(x)\right) \frac{1}{n_c(x)} \left(\Delta + k^2\right) z = 0 , & x \in D , \\ z = 0 , & x \in \Gamma = \Gamma_1 \cup \Gamma_2 , \\ \frac{1}{n_c(x)} \Delta z = 0 , & x \in \Gamma_2 , \\ \frac{\partial z}{\partial \nu} = 0 , & x \in \Gamma_1 . \end{cases} \tag{9.5}$$

The distributions of the transmission eigenvalues can be analyzed in terms of (9.5). To this end, we need the following estimate for $u \in \tilde{H}_{0,1}(D)$ (see [YaMo14]), which can be considered as a generalization of the Poincaré inequality, and can be applied to estimate the lower bound of real-valued eigenvalues in order to qualify the numerical results.

Lemma 9.1 *For any $w \in \tilde{H}_{0,1}(D)$, we obtain the estimate*

$$\|\nabla w\|^2_{L^2(D)} \leq \frac{1}{\lambda(D)} \|\Delta w\|^2_{L^2(D)} .$$

Here, $\lambda(D)$ denotes the first eigenvalue of the buckled plate eigenvalue problem given by

$$\begin{cases} -\Delta^2 w = \lambda \Delta w , & \text{in } D , \\ w = 0 , & \text{on } \Gamma = \Gamma_1 \cup \Gamma_2 , \\ \nu \cdot \nabla w = 0 , & \text{on } \Gamma_1 , \\ \Delta w = 0 , & \text{on } \Gamma_2 . \end{cases}$$

For a real-valued index of refraction $0 < n(x) \neq 1$ in $\overline{D}$, there exists two constant $n_- > 0$ and $n_+ > 0$ such that

$$\begin{aligned} n_+ \geq n(x) \geq n_- > 1 , \qquad & \text{if } n(x)|_{\overline{D}} > 1 , \\ 1 > n_+ \geq n(x) \geq n_- > 0 , \qquad & \text{if } n(x)|_{\overline{D}} < 1 , \end{aligned}$$

which ensures

$$\frac{1}{|n(x)-1|} \geq \alpha > 0\,, \quad x \in \overline{D}$$

for some small constant $\alpha > 0$. Based on Lemma 9.1, the following results state the distribution properties of the mixed interior transmission eigenvalues (MITEs).

Theorem 9.1 *For a real-valued index of refraction $0 < n(x) \neq 1$ in $\overline{D}$, we assume that*

$$\begin{aligned} 0 < \frac{1}{n_- - 1} < 1\,, &\qquad \text{if } n(x)|_{\overline{D}} > 1\,,\\ 0 < \frac{n_+}{1-n_+} < 1\,, &\qquad \text{if } n(x)|_{\overline{D}} < 1\,. \end{aligned} \tag{9.6}$$

Then the set of MITEs is at most discrete and does not accumulate at zero and all the real-valued MITEs (if they exist) are such that

$$k^2 \geq \begin{cases} \lambda(D)\frac{n_- - 2}{n_-(n_- - 1)}\,, & \text{if } n(x)|_{\overline{D}} > 1\,,\\ \lambda(D)\frac{1-2n_+}{1-n_+}\,, & \text{if } n(x)|_{\overline{D}} < 1\,, \end{cases}$$

where $\lambda(D)$ is the first eigenvalue of (9.4).

This result is just a special case of [LiLi16, Theorem 3.3] since we have

$$\begin{aligned} 0 < \frac{1}{|n(x)-1|} = \frac{1}{n(x)-1} \leq \frac{1}{n_- - 1} = \alpha\,, &\qquad \text{if } n(x)|_{\overline{D}} > 1\,,\\ 0 < \frac{1}{|n(x)-1|} = \frac{1}{1-n(x)} \leq \frac{1}{1-n_+} = \alpha\,, &\qquad \text{if } n(x)|_{\overline{D}} < 1\,. \end{aligned}$$

Therefore, we omit the details for the proof. As for the existence of mixed interior transmission eigenvalues, [LiLi16, Theorem 3.7] leads to the following result.

Theorem 9.2 *If (9.6) is replaced by the assumptions*

$$\begin{aligned} 0 < \frac{1}{n_- - 1} < \frac{1}{8}\,, &\qquad \text{if } n(x)|_{\overline{D}} > 1\,,\\ 0 < \frac{n_+}{1-n_+} < \frac{1}{8}\,, &\qquad \text{if } n(x)|_{\overline{D}} < 1\,, \end{aligned} \tag{9.7}$$

then there exists an infinite number of transmission eigenvalues with $+\infty$ as the only possible accumulation point.

Remark 9.1 The assumption (9.6) or (9.7) can be explained easily. Roughly speaking, in the case $n(x)|_{\overline{D}} \neq 1$, if the values of $n(x)$ are far away from the background

index $n_0(x) \equiv 1$, then there always exist discrete transmission eigenvalues (but not necessarily being real-valued) with $+\infty$ as the only possible accumulation point. In Theorem 9.1, we need $n(x) > n_- > 2$ for $n(x) > 1$ and $n(x) < n_+ < 1/2$ for $0 < n(x) < 1$. This condition is strengthened in Theorem 9.2 as $n(x) > n_- > 9$ for $n(x) > 1$ and $n(x) < n_+ < 1/9$ for $0 < n(x) < 1$.

Based on the theoretical results for the distributions of transmission eigenvalues for a real-valued index of refraction $n(x) \neq 1$, we consider the numerical calculation of mixed interior transmission eigenvalues for $n(x)$ being constant in $\overline{D}$ by focusing on the boundary integral equation method in the next section.

9.3 System of Boundary Integral Equations and Its Approximation

In this section, we derive a 4×4 system of boundary integral equations to solve the interior transmission problem with mixed boundary conditions.

Denote by $\Phi_k(x, y) = \mathrm{i}H_0^{(1)}(k|x-y|)/4$, $x \neq y$ the fundamental solution of the two-dimensional Helmholtz equation with wave number k. The single- and double-layer potentials for the Helmholtz equation over the surface Γ are given for $x \notin \Gamma$ by

$$\mathrm{SL}_k^{\Gamma}[\psi](x) = \int_{\Gamma} \Phi_k(x, y)\, \psi(y)\, \mathrm{d}s(y)\,,$$
$$\mathrm{DL}_k^{\Gamma}[\psi](x) = \int_{\Gamma} \partial_{\nu(y)} \Phi_k(x, y)\, \psi(y)\, \mathrm{d}s(y)\,.$$

According to Green's representation theorem (see [CoKr13, p. 17]), we have

$$u(x) = \mathrm{SL}_k^{\Gamma}[\partial_\nu u|_{\Gamma}](x) - \mathrm{DL}_k^{\Gamma}[u|_{\Gamma}](x)\,, \qquad x \in D\,. \tag{9.8}$$

Due to the fact that Γ is the disjoint union of Γ_1 and Γ_2, we can rewrite (9.8) as

$$\begin{aligned} u(x) = {} & \mathrm{SL}_k^{\Gamma_1}\left[\partial_\nu u|_{\Gamma_1}\right](x) + \mathrm{SL}_k^{\Gamma_2}\left[\partial_\nu u|_{\Gamma_2}\right](x) \\ & - \mathrm{DL}_k^{\Gamma_1}\left[u|_{\Gamma_1}\right](x) - \mathrm{DL}_k^{\Gamma_2}\left[u|_{\Gamma_2}\right](x)\,, \qquad x \in D \end{aligned} \tag{9.9}$$

and similarly we obtain

$$\begin{aligned} v(x) = {} & \mathrm{SL}_{k\sqrt{n}}^{\Gamma_1}\left[\partial_\nu v|_{\Gamma_1}\right](x) + \mathrm{SL}_{k\sqrt{n}}^{\Gamma_2}\left[\partial_\nu v|_{\Gamma_2}\right](x) \\ & - \mathrm{DL}_{k\sqrt{n}}^{\Gamma_1}\left[v|_{\Gamma_1}\right](x) - \mathrm{DL}_{k\sqrt{n}}^{\Gamma_2}\left[v|_{\Gamma_2}\right](x)\,, \qquad x \in D\,. \end{aligned} \tag{9.10}$$

By the boundary condition $u|_{\Gamma_2} = v|_{\Gamma_2} = 0$, Eqs. (9.9) and (9.10) can be simplified to

$$u(x) = \mathrm{SL}_k^{\Gamma_1}\left[\partial_\nu u|_{\Gamma_1}\right](x) + \mathrm{SL}_k^{\Gamma_2}\left[\partial_\nu u|_{\Gamma_2}\right](x) - \mathrm{DL}_k^{\Gamma_1}\left[u|_{\Gamma_1}\right](x)\,, \tag{9.11}$$

$$v(x) = \mathrm{SL}_{k\sqrt{n}}^{\Gamma_1}\left[\partial_\nu v|_{\Gamma_1}\right](x) + \mathrm{SL}_{k\sqrt{n}}^{\Gamma_2}\left[\partial_\nu v|_{\Gamma_2}\right](x) - \mathrm{DL}_{k\sqrt{n}}^{\Gamma_1}\left[v|_{\Gamma_1}\right](x)\,, \tag{9.12}$$

where $x \in D$. The boundary integral operators over the surface Γ_i evaluated at a point of Γ_j are defined as

$$\mathrm{S}_k^{\Gamma_i\to\Gamma_j}\left[\psi|_{\Gamma_i}\right](x) = \int_{\Gamma_i} \Phi_k(x,y)\psi(y)\,\mathrm{d}s(y)\,, \quad x \in \Gamma_j\,,$$

$$\mathrm{K}_k^{\Gamma_i\to\Gamma_j}\left[\psi|_{\Gamma_i}\right](x) = \int_{\Gamma_i} \partial_{\nu_i(y)}\Phi_k(x,y)\psi(y)\,\mathrm{d}s(y)\,, \quad x \in \Gamma_j\,,$$

$$\mathrm{K}_k^{\top\,\Gamma_i\to\Gamma_j}\left[\psi|_{\Gamma_i}\right](x) = \int_{\Gamma_i} \partial_{\nu_j(x)}\Phi_k(x,y)\psi(y)\,\mathrm{d}s(y)\,, \quad x \in \Gamma_j\,,$$

$$\mathrm{T}_k^{\Gamma_i\to\Gamma_j}\left[\psi|_{\Gamma_i}\right](x) = \partial_{\nu_j(x)}\int_{\Gamma_i} \partial_{\nu_i(y)}\Phi_k(x,y)\psi(y)\,\mathrm{d}s(y)\,, \quad x \in \Gamma_j\,,$$

where $i, j \in \{1, 2\}$.

9.3.1 First Boundary Integral Equation

Letting $D \ni x \to x \in \Gamma_1$ in (9.12) and (9.12) and using the jump relations (see [CoKr13, p. 39]), yields

$$u|_{\Gamma_1} = \mathrm{S}_k^{\Gamma_1\to\Gamma_1}\left[\partial_\nu u|_{\Gamma_1}\right] + \mathrm{S}_k^{\Gamma_2\to\Gamma_1}\left[\partial_\nu u|_{\Gamma_2}\right] - \left(\mathrm{K}_k^{\Gamma_1\to\Gamma_1}\left[u|_{\Gamma_1}\right] - \frac{1}{2}u|_{\Gamma_1}\right) \tag{9.13}$$

and

$$v|_{\Gamma_1} = \mathrm{S}_{k\sqrt{n}}^{\Gamma_1\to\Gamma_1}\left[\partial_\nu v|_{\Gamma_1}\right] + \mathrm{S}_{k\sqrt{n}}^{\Gamma_2\to\Gamma_1}\left[\partial_\nu v|_{\Gamma_2}\right] - \left(\mathrm{K}_{k\sqrt{n}}^{\Gamma_1\to\Gamma_1}\left[v|_{\Gamma_1}\right] - \frac{1}{2}v|_{\Gamma_1}\right)\,. \tag{9.14}$$

Taking the difference of (9.13) and (9.14) and using the boundary conditions $u|_{\Gamma_1} = v|_{\Gamma_1}$ and $\partial_\nu u|_{\Gamma_1} = \partial_\nu v|_{\Gamma_1}$, gives the first boundary integral equation

$$\begin{aligned} 0 = &\left(\mathrm{S}_k^{\Gamma_1\to\Gamma_1} - \mathrm{S}_{k\sqrt{n}}^{\Gamma_1\to\Gamma_1}\right)\left[\partial_\nu u|_{\Gamma_1}\right] + \mathrm{S}_k^{\Gamma_2\to\Gamma_1}\left[\partial_\nu u|_{\Gamma_2}\right] \\ &- \mathrm{S}_{k\sqrt{n}}^{\Gamma_2\to\Gamma_1}\left[\partial_\nu v|_{\Gamma_2}\right] - \left(\mathrm{K}_k^{\Gamma_1\to\Gamma_1} - \mathrm{K}_{k\sqrt{n}}^{\Gamma_1\to\Gamma_1}\right)\left[u|_{\Gamma_1}\right]\,. \end{aligned} \tag{9.15}$$

9.3.2 Second Boundary Integral Equation

Applying the same strategy as before for $D \ni x \to x \in \Gamma_2$ in (9.12) and (9.12), yields

$$u|_{\Gamma_2} = \mathrm{S}_k^{\Gamma_1\to\Gamma_2}\left[\partial_\nu u|_{\Gamma_1}\right] + \mathrm{S}_k^{\Gamma_2\to\Gamma_2}\left[\partial_\nu u|_{\Gamma_2}\right] - \mathrm{K}_k^{\Gamma_1\to\Gamma_2}\left[u|_{\Gamma_1}\right] \tag{9.16}$$

and

$$v|_{\Gamma_2} = \mathrm{S}_{k\sqrt{n}}^{\Gamma_1\to\Gamma_2}\left[\partial_\nu v|_{\Gamma_1}\right] + \mathrm{S}_{k\sqrt{n}}^{\Gamma_2\to\Gamma_2}\left[\partial_\nu v|_{\Gamma_2}\right] - \mathrm{K}_{k\sqrt{n}}^{\Gamma_1\to\Gamma_2}\left[v|_{\Gamma_1}\right]. \tag{9.17}$$

Taking the difference of (9.16) and (9.17), setting $u|_{\Gamma_2} = v|_{\Gamma_2} = 0$, and applying the boundary conditions $u|_{\Gamma_1} = v|_{\Gamma_1}$ and $\partial_\nu u|_{\Gamma_1} = \partial_\nu v|_{\Gamma_1}$, gives the second boundary integral equation

$$\begin{aligned} 0 = &\left(\mathrm{S}_k^{\Gamma_1\to\Gamma_2} - \mathrm{S}_{k\sqrt{n}}^{\Gamma_1\to\Gamma_2}\right)\left[\partial_\nu u|_{\Gamma_1}\right] + \mathrm{S}_k^{\Gamma_2\to\Gamma_2}\left[\partial_\nu u|_{\Gamma_2}\right] \\ &- \mathrm{S}_{k\sqrt{n}}^{\Gamma_2\to\Gamma_2}\left[\partial_\nu v|_{\Gamma_2}\right] - \left(\mathrm{K}_k^{\Gamma_1\to\Gamma_2} - \mathrm{K}_{k\sqrt{n}}^{\Gamma_1\to\Gamma_2}\right)\left[u|_{\Gamma_1}\right]. \end{aligned} \tag{9.18}$$

9.3.3 Third Boundary Integral Equation

Next, we apply the normal derivative to (9.12) and (9.12), let $D \ni x \to x \in \Gamma_1$, and use the jump relations. This yields

$$\begin{aligned} \partial_\nu u|_{\Gamma_1} = &\mathrm{K}_k^{\top\,\Gamma_1\to\Gamma_1}\left[\partial_\nu u|_{\Gamma_1}\right] + \frac{1}{2}\partial_\nu u|_{\Gamma_1} + \mathrm{K}_k^{\top\,\Gamma_2\to\Gamma_1}\left[\partial_\nu u|_{\Gamma_2}\right] \\ &- \mathrm{T}_k^{\Gamma_1\to\Gamma_1}\left[u|_{\Gamma_1}\right] \end{aligned} \tag{9.19}$$

and

$$\begin{aligned} \partial_\nu v|_{\Gamma_1} = &\mathrm{K}_{k\sqrt{n}}^{\top\,\Gamma_1\to\Gamma_1}\left[\partial_\nu v|_{\Gamma_1}\right] + \frac{1}{2}\partial_\nu v|_{\Gamma_1} + \mathrm{K}_{k\sqrt{n}}^{\top\,\Gamma_2\to\Gamma_1}\left[\partial_\nu v|_{\Gamma_2}\right] \\ &- \mathrm{T}_{k\sqrt{n}}^{\Gamma_1\to\Gamma_1}\left[v|_{\Gamma_1}\right]. \end{aligned} \tag{9.20}$$

Taking the difference of (9.19) and (9.20) and using the boundary conditions $u|_{\Gamma_1} = v|_{\Gamma_1}$ and $\partial_\nu u|_{\Gamma_1} = \partial_\nu v|_{\Gamma_1}$, gives the third boundary integral equation

$$\begin{aligned} 0 = &\left(\mathrm{K}_k^{\top\,\Gamma_1\to\Gamma_1} - \mathrm{K}_{k\sqrt{n}}^{\top\,\Gamma_1\to\Gamma_1}\right)\left[\partial_\nu u|_{\Gamma_1}\right] + \mathrm{K}_k^{\top\,\Gamma_2\to\Gamma_1}\left[\partial_\nu u|_{\Gamma_2}\right] \\ &- \mathrm{K}_{k\sqrt{n}}^{\top\,\Gamma_2\to\Gamma_1}\left[\partial_\nu v|_{\Gamma_2}\right] - \left(\mathrm{T}_k^{\Gamma_1\to\Gamma_1} - \mathrm{T}_{k\sqrt{n}}^{\Gamma_1\to\Gamma_1}\right)\left[u|_{\Gamma_1}\right]. \end{aligned} \tag{9.21}$$

9.3.4 Fourth Boundary Integral Equation

Again, we apply the normal derivative to (9.12) and (9.12), let $D \ni x \to x \in \Gamma_2$, and use the jump relations. This gives

$$\begin{aligned}\partial_\nu u|_{\Gamma_2} = {}& \mathrm{K}_k^{\top\,\Gamma_1\to\Gamma_2}\left[\partial_\nu u|_{\Gamma_1}\right] + \mathrm{K}_k^{\top\,\Gamma_2\to\Gamma_2}\left[\partial_\nu u|_{\Gamma_2}\right] + \frac{1}{2}\partial_\nu u|_{\Gamma_2} \\ & - \mathrm{T}_k^{\Gamma_1\to\Gamma_2}\left[u|_{\Gamma_1}\right] \end{aligned} \tag{9.22}$$

and

$$\begin{aligned}\partial_\nu v|_{\Gamma_2} = {}& \mathrm{K}_{k\sqrt{n}}^{\top\,\Gamma_1\to\Gamma_2}\left[\partial_\nu v|_{\Gamma_1}\right] + \mathrm{K}_{k\sqrt{n}}^{\top\,\Gamma_2\to\Gamma_2}\left[\partial_\nu v|_{\Gamma_2}\right] + \frac{1}{2}\partial_\nu v|_{\Gamma_2} \\ & - \mathrm{T}_{k\sqrt{n}}^{\Gamma_1\to\Gamma_2}\left[v|_{\Gamma_1}\right]. \end{aligned} \tag{9.23}$$

Equations (9.22) and (9.23) can be rewritten as

$$\begin{aligned}0 = {}& \mathrm{K}_k^{\top\,\Gamma_1\to\Gamma_2}\left[\partial_\nu u|_{\Gamma_1}\right] + \mathrm{K}_k^{\top\,\Gamma_2\to\Gamma_2}\left[\partial_\nu u|_{\Gamma_2}\right] - \mathrm{T}_k^{\Gamma_1\to\Gamma_2}\left[u|_{\Gamma_1}\right] \\ & - \frac{1}{2}\partial_\nu u|_{\Gamma_2} \end{aligned} \tag{9.24}$$

and

$$\begin{aligned}0 = {}& \mathrm{K}_{k\sqrt{n}}^{\top\,\Gamma_1\to\Gamma_2}\left[\partial_\nu v|_{\Gamma_1}\right] + \mathrm{K}_{k\sqrt{n}}^{\top\,\Gamma_2\to\Gamma_2}\left[\partial_\nu v|_{\Gamma_2}\right] - \mathrm{T}_{k\sqrt{n}}^{\Gamma_1\to\Gamma_2}\left[v|_{\Gamma_1}\right] \\ & - \frac{1}{2}\partial_\nu v|_{\Gamma_2} \end{aligned} \tag{9.25}$$

respectively. Taking the difference of (9.24) and (9.25) and using the boundary conditions $u|_{\Gamma_1} = v|_{\Gamma_1}$ and $\partial_\nu u|_{\Gamma_1} = \partial_\nu v|_{\Gamma_1}$, gives the fourth boundary integral equation

$$\begin{aligned}0 = {}& \left(\mathrm{K}_k^{\top\,\Gamma_1\to\Gamma_2} - \mathrm{K}_{k\sqrt{n}}^{\top\,\Gamma_1\to\Gamma_2}\right)\left[\partial_\nu u|_{\Gamma_1}\right] + \mathrm{K}_k^{\top\,\Gamma_2\to\Gamma_2}\left[\partial_\nu u|_{\Gamma_2}\right] \\ & - \mathrm{K}_{k\sqrt{n}}^{\top\,\Gamma_2\to\Gamma_2}\left[\partial_\nu v|_{\Gamma_2}\right] - \left(\mathrm{T}_k^{\Gamma_1\to\Gamma_2} - \mathrm{T}_{k\sqrt{n}}^{\Gamma_1\to\Gamma_2}\right)\left[u|_{\Gamma_1}\right] - \frac{1}{2}\partial_\nu u|_{\Gamma_2} \\ & + \frac{1}{2}\partial_\nu v|_{\Gamma_2}. \end{aligned} \tag{9.26}$$

9.3.5 System of Boundary Integral Equations

The four equations (9.15), (9.18), (9.21), and (9.26) can be written abstractly as

$$Z(k)g = 0 \tag{9.27}$$

with $Z(k)$ given by

$$\begin{pmatrix} \mathrm{S}_k^{\Gamma_1\to\Gamma_1} - \mathrm{S}_{k\sqrt{n}}^{\Gamma_1\to\Gamma_1} & \mathrm{K}_k^{\Gamma_1\to\Gamma_1} - \mathrm{K}_{k\sqrt{n}}^{\Gamma_1\to\Gamma_1} & \mathrm{S}_k^{\Gamma_2\to\Gamma_1} & \mathrm{S}_{k\sqrt{n}}^{\Gamma_2\to\Gamma_1} \\ \mathrm{S}_k^{\Gamma_1\to\Gamma_2} - \mathrm{S}_{k\sqrt{n}}^{\Gamma_1\to\Gamma_2} & \mathrm{K}_k^{\Gamma_1\to\Gamma_2} - \mathrm{K}_{k\sqrt{n}}^{\Gamma_1\to\Gamma_2} & \mathrm{S}_k^{\Gamma_2\to\Gamma_2} & \mathrm{S}_{k\sqrt{n}}^{\Gamma_2\to\Gamma_2} \\ \mathrm{K}_k^{\top\,\Gamma_1\to\Gamma_1} - \mathrm{K}_{k\sqrt{n}}^{\top\quad\Gamma_1\to\Gamma_1} & \mathrm{T}_k^{\Gamma_1\to\Gamma_1} - \mathrm{T}_{k\sqrt{n}}^{\Gamma_1\to\Gamma_1} & \mathrm{K}_k^{\top\,\Gamma_2\to\Gamma_1} & \mathrm{K}_{k\sqrt{n}}^{\top\quad\Gamma_2\to\Gamma_1} \\ \mathrm{K}_k^{\top\,\Gamma_1\to\Gamma_2} - \mathrm{K}_{k\sqrt{n}}^{\top\quad\Gamma_1\to\Gamma_2} & \mathrm{T}_k^{\Gamma_1\to\Gamma_2} - \mathrm{T}_{k\sqrt{n}}^{\Gamma_1\to\Gamma_2} & \mathrm{K}_k^{\top\,\Gamma_2\to\Gamma_2} - \frac{1}{2}I & \mathrm{K}_{k\sqrt{n}}^{\top\quad\Gamma_2\to\Gamma_2} - \frac{1}{2}I \end{pmatrix} \tag{9.28}$$

and

$$g = \begin{pmatrix} \alpha & -\beta & \gamma & -\delta \end{pmatrix}^\top ,$$

where we used the notation

$$\alpha = \partial_\nu u|_{\Gamma_1} , \ \beta = u|_{\Gamma_1} , \ \gamma = \partial_\nu u|_{\Gamma_2} , \text{ and } \delta = \partial_\nu v|_{\Gamma_2} . \tag{9.29}$$

The matrix entries in (9.28) are boundary integral operator with a specific kernel. If the operator O is of the form $\mathrm{O}^{\Gamma_i\to\Gamma_j}$ with $i \neq j$, then the kernel is smooth. Additionally, the kernel of the operator $\mathrm{S}_k^{\Gamma_1\to\Gamma_1} - \mathrm{S}_{k\sqrt{n}}^{\Gamma_1\to\Gamma_1}$ is smooth as well. The remaining entries contain a kernel with a weak singularity which is of logarithmic form (notice again that $n \neq 1$ in D). In three dimensions the situation changes slightly to a weak singularity. Hence, in both cases the system can easily be approximated numerically to high accuracy by the boundary element collocation method as developed in [KlLi11, KlLi12] which has been successfully used in [AnChAk13, KiKl12, Kl12b, Kl12c] for the three-dimensional case.

To show that the operator is Fredholm of index zero and analytic for $k \in \mathbb{C}\backslash\mathbb{R}_{\leq 0}$, one would follow the same arguments as given in [Co11, Theorem 5.3.9] or [CoHa13]. The only difficulty is to use the correct Sobolev spaces of the form $\tilde{H}^s(\Gamma_i)$ (see [Mc00] for the definition of the Sobolev spaces) or alternatively the Lions–Magenes spaces $H_{00}^s(\Gamma_i)$ as given in [LiMa72].

9.3.6 *Approximation of the System*

In this section, we shortly explain how to discretize the resulting boundary integral operator via boundary element collocation method. An extensive explanation has previously been given in [KlLi11, KlLi12] for the three-dimensional case and the two-dimensional case works conceptually similar (see also [At97] for the Laplace equation).

We consider two kinds of scatterers in two dimensions. The first class has a boundary that can be described through polar coordinates and the second class has a boundary that can be described through lines. For the first class, we define the set of points through an equidistant use of the polar angle, whereas for the second class the edges are subdivided into equal parts. The curved boundary for the scatterers of the first class is now approximated by a polygon having the previously defined points as vertices. The set of collocation points are the midpoints of each line segment having m collocation points in total. Now, the approximation of each integral over such a line segment can easily be carried out by numerical integration where we assume that the unknown function is approximated by constant interpolation at a midpoint. Note that we have at most a logarithmic singularity in the kernel if the collocation point is situated on a line segment on which we are integrating over. A Gauss–Kronrad quadrature can deal easily with such a singularity.

After the discretization, we can regard (9.28) as a non-linear eigenvalue problem of the form $\mathbf{Z}(k)\tilde{g} = 0$ with $\mathbf{Z}(k) \in \mathbb{C}^{m\times m}$ and $\tilde{g}$ the discretized version of g given by (9.29) which we solve with Beyn's algorithm [Be12] as done in [CaKr17, Kl13, Kl15, StUn12]. This algorithm uses complex-valued contour integration of the resolvent to reduce the non-linear eigenvalue problem to a linear eigenvalue problem of much smaller size based upon the famous Keldysh's Theorem. For this algorithm one has to specify a 2π-periodic contour in the complex plane and it will find all non-linear eigenvalues situated in this contour to high accuracy due to the fact that the approximation of a 2π-periodic function via the trapezoidal rule yields exponential convergence. We therefore use a circle of radius R with center $C = (c_x, c_y\mathrm{i})$ with $N = 40$ nodes for the trapezoidal rule.

9.4 Numerical Results

In this section, we present extensive numerical results for some two-dimensional scatterers although we can easily calculate them in three dimensions as shown in [Kl13, Kl15] for the classic interior transmission eigenvalues. The reason is that we can nicely present the corresponding eigenfunctions which is much more difficult in three dimensions.

The first scatterer under consideration is the unit circle $\mathcal{C}$, where we used Γ_1 as the upper half of the circle and Γ_2 as the lower half of the circle. The first five MITEs are given by 1.6818, 2.3185, 2.9533, 3.0791, and 3.1409 where we used

the index of refraction $n = 4$. After solving the non-linear eigenvalue problem (9.27) using the parameters $N = 40$, $R = 1/2$, with centers $C = (2, 0\mathrm{i})$ and $C = (3, 0\mathrm{i})$, respectively, we additionally obtain the discretized version of the functions $\partial_\nu u|_{\Gamma_1} = \partial_\nu v|_{\Gamma_1}$, $u|_{\Gamma_1} = v|_{\Gamma_1}$, $\partial_\nu u|_{\Gamma_2}$, and $\partial_\nu v|_{\Gamma_2}$. We also have $u|_{\Gamma_2} = v|_{\Gamma_2} = 0$ and hence we can insert these function approximations into (9.12) and (9.12) in order to compute the approximate solution of u and v at any point situated inside of the scatterer $\mathcal{C}$. We denote with $u^{(i)}$ and $v^{(i)}$ the approximate eigenfunctions corresponding to the i-th MITE. The absolute value of $u^{(i)}$ and $v^{(i)}$ inside $\mathcal{C}$ for the first five MITEs is shown in Fig. 9.2. Note that we were also able to find a complex-valued MITE pair $2.3596 + 0.3413\mathrm{i}$ and $2.3592 - 0.34134\mathrm{i}$. The corresponding eigenfunctions $u^{(cv)}$ and $v^{(cv)}$ are also given in Fig. 9.2. Since we used constant interpolation for the boundary element collocation method, a linear convergence rate for the eigenvalues is expected and achieved. Note that we do not know the exact MITE values. However, we observe that the error of the imaginary part of the real-valued MITE halves if we double the number of collocation nodes m. In Table 9.1 we clearly see the linear convergence order. Additionally, we get the following complex-valued MITE pair $2.7340 + 0.3801\mathrm{i}$ and $2.7334 - 0.3802\mathrm{i}$. The corresponding eigenfunctions are shown in Fig. 9.3. Next, we calculate the MITEs for an ellipse $\mathcal{E}$ with major semi-axis 1 and minor semi-axis $4/5$ using the same parameters as before. We obtain the four real-valued MITEs 1.9111, 2.4973, 3.1282, and 3.4609. If we use the major semi-axis 1 and minor semi-axis $1/2$ for the ellipse with the same parameters as before, we obtain the first four real-valued MITEs 2.7709, 3.1764, 3.7892, and 4.3916. A complex-valued MITE pair is given by $3.8947 + 0.5352\mathrm{i}$ and $3.8928 - 0.5365\mathrm{i}$. Using the minor semi-axis $3/10$ yields the four real-valued MITEs 4.5026, 4.7231, 5.2731, and 5.7279. Note that the classical interior transmission eigenvalues for various minor semi-axis are given in [CaKr17, KlPi18] and a summary of the results for the various ellipses is given in Table 9.2. In Table 9.3 we list the first four real-valued MITEs for deformed ellipses given by the parametrization $(3\cos(t)/4 + \kappa\cos(2t), \sin(t))$, $t \in [0, 2\pi)$ for $\kappa = 0, 1/10, 1/5$, and $3/10$ which has been used before in [CaKr17, KlPi18] for classical interior transmission eigenvalues. We again used $n = 4$ and 1280 collocation points and the same boundary conditions as before.

The MITEs for the unit square $\mathcal{S}$ using the index of refraction $n = 4$ with transmission conditions on the south and east part and homogeneous Dirichlet conditions on the north and west part of the boundary are given by 3.0503, 4.2622, and 5.1805 where we used 512 collocation points and $R = 1$ with the centers $C = (2, 0\mathrm{i})$, $C = (3, 0\mathrm{i})$, and $C = (4, 0\mathrm{i})$, respectively. Figure 9.4 shows the absolute value of the first three eigenfunctions $u^{(i)}$ and $v^{(i)}$ for the unit square. In Fig. 9.5 we display the first three eigenfunctions $u^{(i)}$ and $v^{(i)}$ for the unit square for the eigenvalues 2.6717, 3.6662, and 4.8367, respectively, where we used the index of refraction $n = 4$ with transmission conditions on the south part and homogeneous Dirichlet condition on the remaining edges. We again used 512 collocation points and $R = 1$ with the centers $C = (2.5, 0\mathrm{i})$, $C = (3.5, 0\mathrm{i})$, and $C = (4.5, 0\mathrm{i})$, respectively. Note that in this case it is possible to derive an analytic equation such that its zeros are the MITEs (we refer the reader to Table 9.5 in the Appendix). We

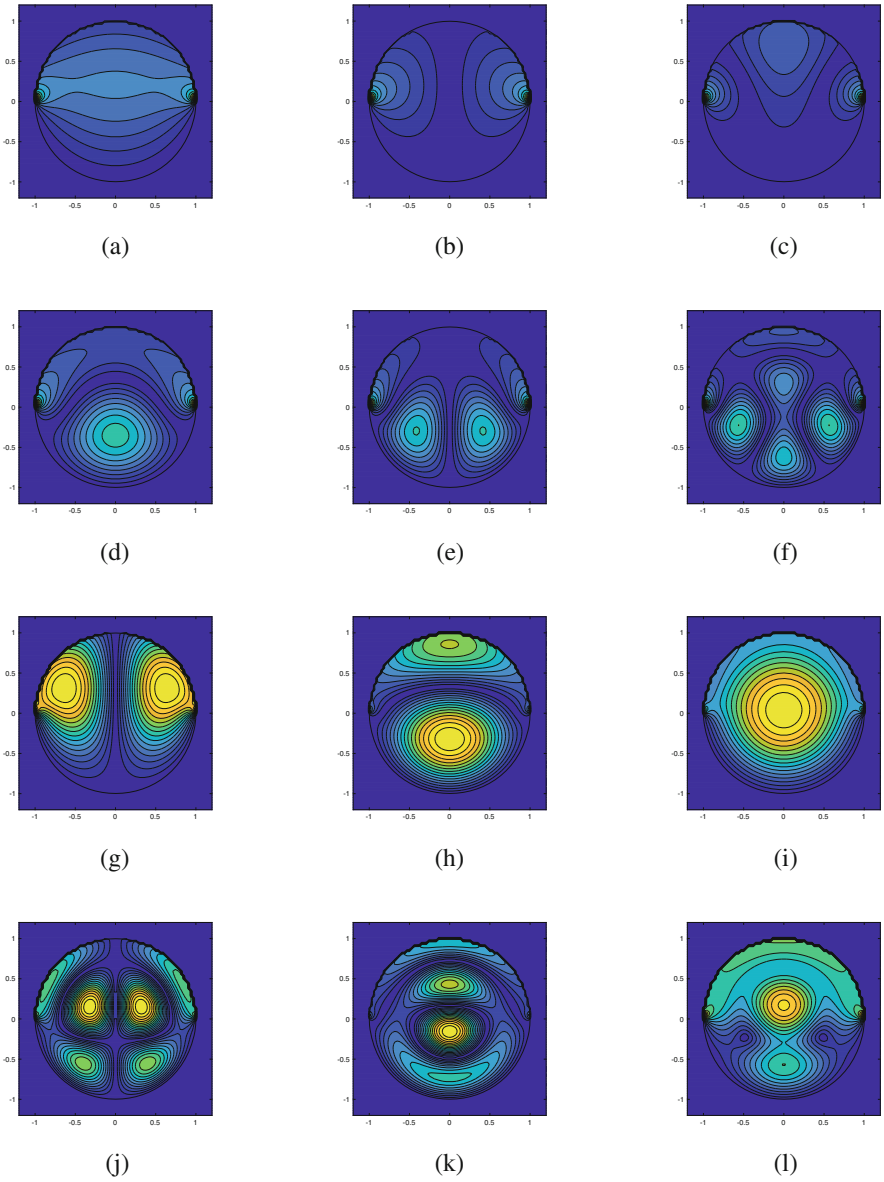


Fig. 9.2 The absolute value of the eigenfunctions u (first row and third row) and v (second row and fourth row) for the first five real-valued MITEs and one complex-valued MITE for the unit circle $\mathcal{C}$ using the index of refraction $n = 4$. The MITEs are 1.6818, 2.3185, 2.9533, 3.0791, 3.1409 and $2.3596 + 0.3413\mathrm{i}$, respectively. (**a**) $|u^{(1)}|$. (**b**) $|u^{(2)}|$. (**c**) $|u^{(3)}|$. (**d**) $|v^{(1)}|$. (**e**) $|v^{(2)}|$. (**f**) $|v^{(3)}|$. (**g**) $|u^{(4)}|$. (**h**) $|u^{(5)}|$. (**i**) $|u^{(cv)}|$. (**j**) $|v^{(4)}|$. (**k**) $|v^{(5)}|$. (**l**) $|v^{(cv)}|$

Table 9.1 Convergence of the first five real-valued MITEs for the unit circle C using the index of refraction $n = 4$

m	$k^{(1)}$	$k^{(2)}$	$k^{(3)}$	$k^{(4)}$	$k^{(5)}$
20	$1.6691 + 0.0445\mathrm{i}$	$2.3048 + 0.0268\mathrm{i}$	$2.9405 + 0.0176\mathrm{i}$	$3.1055 + 0.0063\mathrm{i}$	$3.1687 + 0.0054\mathrm{i}$
40	$1.6735 + 0.0216\mathrm{i}$	$2.3075 + 0.0132\mathrm{i}$	$2.9412 + 0.0094\mathrm{i}$	$3.0846 + 0.0037\mathrm{i}$	$3.1475 + 0.0028\mathrm{i}$
80	$1.6772 + 0.0106\mathrm{i}$	$2.3121 + 0.0066\mathrm{i}$	$2.9460 + 0.0050\mathrm{i}$	$3.0800 + 0.0020\mathrm{i}$	$3.1424 + 0.0014\mathrm{i}$
160	$1.6795 + 0.0052\mathrm{i}$	$2.3152 + 0.0033\mathrm{i}$	$2.9495 + 0.0025\mathrm{i}$	$3.0791 + 0.0010\mathrm{i}$	$3.1412 + 0.0007\mathrm{i}$
320	$1.6808 + 0.0026\mathrm{i}$	$2.3170 + 0.0017\mathrm{i}$	$2.9516 + 0.0013\mathrm{i}$	$3.0790 + 0.0005\mathrm{i}$	$3.1410 + 0.0003\mathrm{i}$
640	$1.6814 + 0.0013\mathrm{i}$	$2.3180 + 0.0008\mathrm{i}$	$2.9527 + 0.0007\mathrm{i}$	$3.0791 + 0.0003\mathrm{i}$	$3.1409 + 0.0002\mathrm{i}$
1280	$1.6818 + 0.0007\mathrm{i}$	$2.3185 + 0.0004\mathrm{i}$	$2.9533 + 0.0003\mathrm{i}$	$3.0791 + 0.0001\mathrm{i}$	$3.1409 + 0.0001\mathrm{i}$

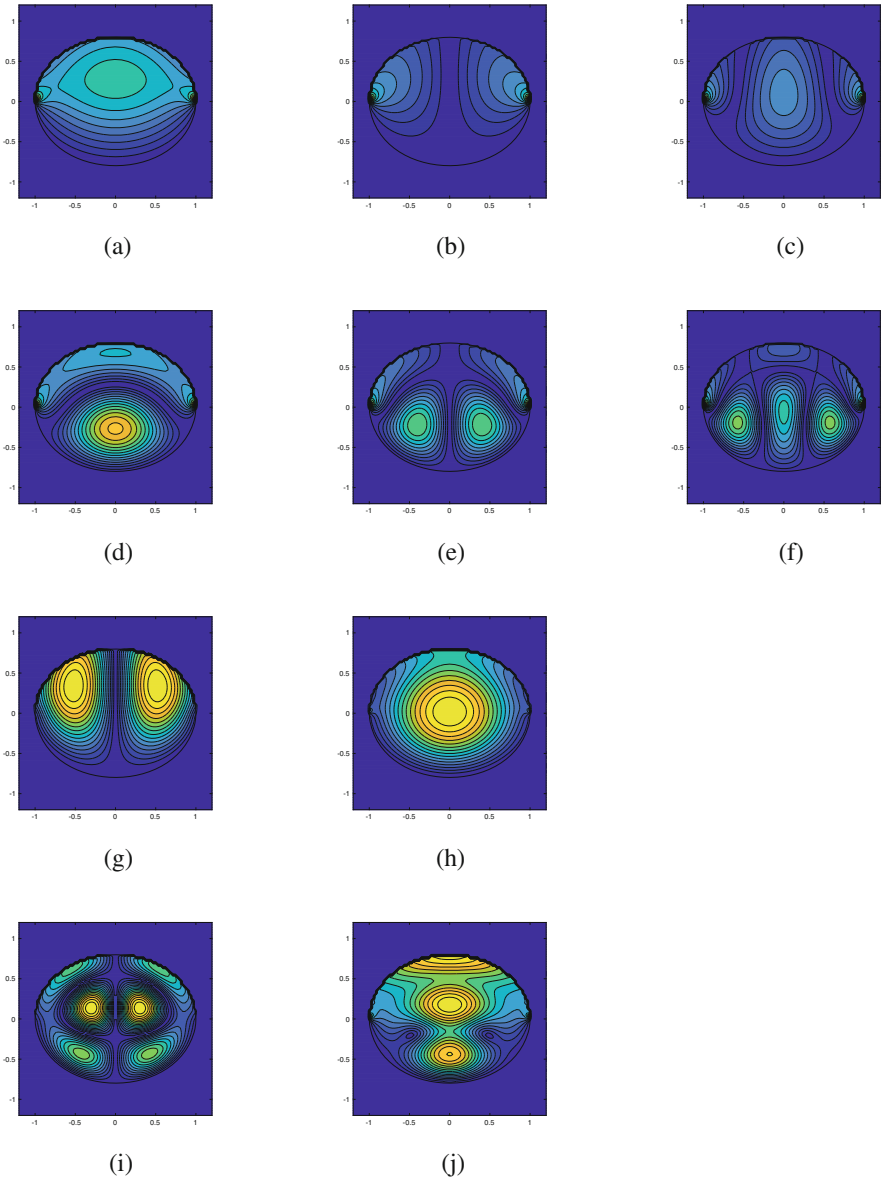


Fig. 9.3 The absolute value of the eigenfunctions u (first row and third row) and v (second row and fourth row) for the first four real-valued MITEs and one complex-valued MITE for the ellipse $\mathcal{E}$ using the index of refraction $n = 4$. The MITEs are 1.9111, 2.4973, 3.1282, 3.4609 and 2.7340+ 0.3801i, respectively. (**a**) $|u^{(1)}|$. (**b**) $|u^{(2)}|$. (**c**) $|u^{(3)}|$. (**d**) $|v^{(1)}|$. (**e**) $|v^{(2)}|$. (**f**) $|v^{(3)}|$. (**g**) $|u^{(4)}|$. (**h**) $|u^{(cv)}|$. (**i**) $|v^{(4)}|$. (**j**) $|v^{(cv)}|$

Table 9.2 The first four real-valued MITEs for ellipses with major semi-axis 1 and various minor semi-axis using the index of refraction $n = 4$

Minor semi-axis	First MITE	Second MITE	Third MITE	Fourth MITE
1	1.6818	2.3185	2.9533	3.0791
4/5	1.9111	2.4973	3.1282	3.4609
1/2	2.7709	3.1764	3.7892	4.3916
3/10	4.5026	4.7231	5.2731	5.7279

Table 9.3 The first four real-valued MITEs for deformed ellipses with various deformation parameter κ using the index of refraction $n = 4$

κ	First MITE	Second MITE	Third MITE	Fourth MITE
0	1.9626	2.8575	3.2436	3.6951
1/10	1.9755	2.8404	3.2836	3.6542
1/5	2.0122	2.8087	3.3753	3.5941
3/10	2.0674	2.7899	3.4327	3.6203

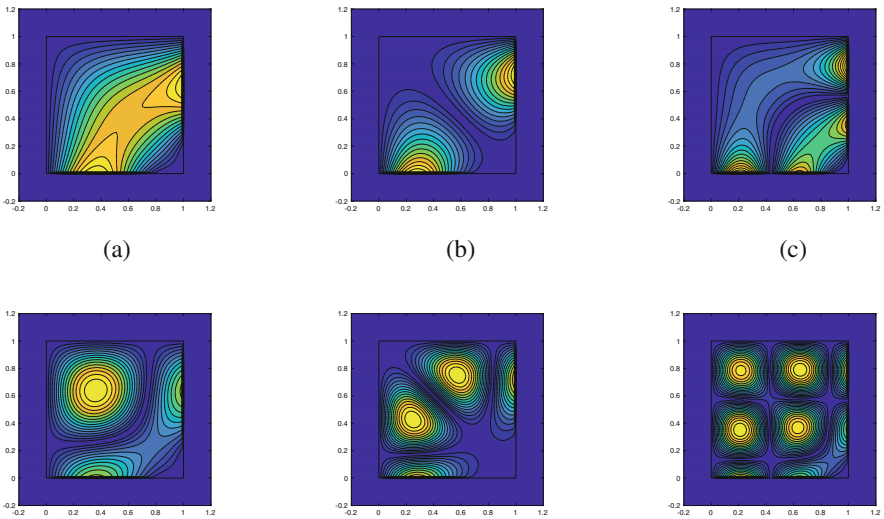


Fig. 9.4 The absolute value of the eigenfunctions u (first row) and v (second row) for the first three real-valued MITEs for the unit square $\mathcal{S}$ using the index of refraction $n = 4$ with transmission conditions on the south and east part and homogeneous Dirichlet conditions on the north and west part of the boundary. The MITEs are 3.0503, 4.2622, and 5.1805, respectively. (**a**) $|u^{(1)}|$. (**b**) $|u^{(2)}|$. (**c**) $|u^{(3)}|$. (**d**) $|v^{(1)}|$. (**e**) $|v^{(2)}|$. (**f**) $|v^{(3)}|$

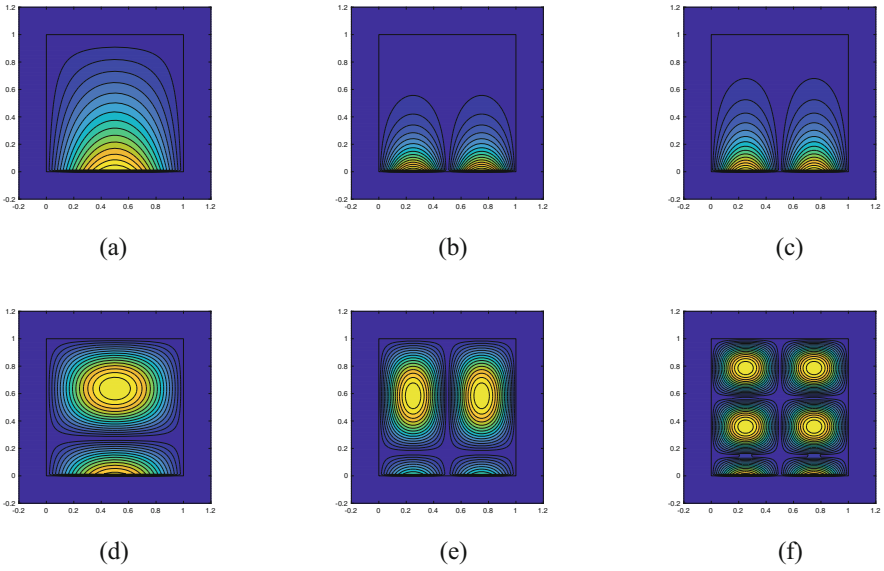


Fig. 9.5 The absolute value of the eigenfunctions u (first row) and v (second row) for the first three real-valued MITEs for the unit square $\mathcal{S}$ using the index of refraction $n = 4$ with transmission conditions on the south part and homogeneous Dirichlet conditions on the remaining edges. The MITEs are 2.6717, 3.6662, and 4.8367, respectively. (**a**) $|u^{(1)}|$. (**b**) $|u^{(2)}|$. (**c**) $|u^{(3)}|$. (**d**) $|v^{(1)}|$. (**e**) $|v^{(2)}|$. (**f**) $|v^{(3)}|$

obtain with a root finding algorithm 2.671 552 787 839 805, 3.666 034 666 514 623, and 4.836 476 632 026 555 and hence the first four digits of the reported results agree.

Next, we also present the first three real-valued MITEs for the unit square where we impose homogeneous Dirichlet condition on the west part and transmission condition on the remaining edges. We use the same parameters as before. We obtain 4.0802, 5.2285, and 5.7030. The corresponding three eigenfunctions $u^{(i)}$ and $v^{(i)}$ are given in Fig. 9.6.

To complete our numerical results, we also present numerical results for the case $0 < n < 1$ using the index of refraction $n = 1/2$. Again we use ellipses with various minor axis similar as in Table 9.3. The results are summarized in Table 9.4.

To find the MITEs we use a circle with radius $R = 1/2$ and centers $C = (3, 0\mathrm{i})$, $C = (4.5, 0\mathrm{i})$, and $C = (6, 0\mathrm{i})$ in the non-linear eigenvalue solver for the unit circle scatterer, whereas we use the centers $C = (3.5, 0\mathrm{i})$, $C = (4.5, 0\mathrm{i})$, $C = (5.5, 0\mathrm{i})$, and $C = (6.5, 0\mathrm{i})$ for the ellipse with minor semi-axis $4/5$. For the ellipse with minor semi-axis $1/2$ we use the centers $C = (5, 0\mathrm{i})$, $C = (6, 0\mathrm{i})$, $C = (7, 0\mathrm{i})$, and $C = (8.5, 0\mathrm{i})$.

As a final remark, we would like to mention that we also tried different n not satisfying (9.6) and (9.7). In all cases, we have an accumulation point at infinity and no trouble computing the MITEs. This can also be verified numerically with (9.32) for a variety of n and p.

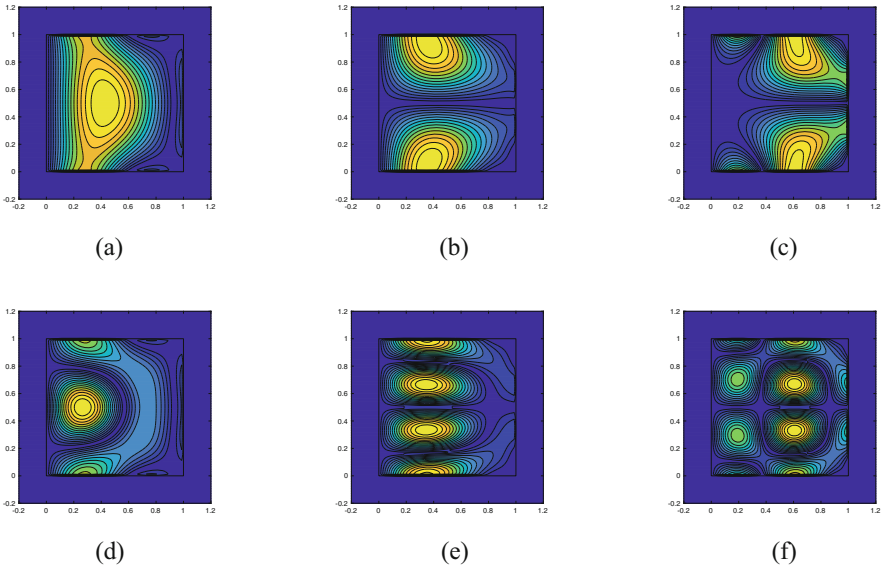


Fig. 9.6 The absolute value of the eigenfunctions u (first row) and v (second row) for the first three real-valued MITEs for the unit square $\mathcal{S}$ using the index of refraction $n = 4$ with homogeneous Dirichlet conditions on the west part and transmission conditions on the remaining edges. The MITEs are 4.0802, 5.2285, and 5.7030, respectively. (**a**) $|u^{(1)}|$. (**b**) $|u^{(2)}|$. (**c**) $|u^{(3)}|$. (**d**) $|v^{(1)}|$. (**e**) $|v^{(2)}|$. (**f**) $|v^{(3)}|$

Table 9.4 The first four real-valued MITEs for ellipses with major semi-axis 1 and various minor semi-axis using the index of refraction $n = 1/2$

Minor semi-axis	First MITE	Second MITE	Third MITE	Fourth MITE
1	3.1620	4.5193	4.6482	5.8022
4/5	3.5798	4.8518	5.5187	6.2683
1/2	5.1115	6.1186	7.3248	8.4891

Additionally, all presented numerical results are indeed mixed interior transmission eigenvalues (not mixed exterior transmission eigenvalues), since we always computed the corresponding eigenfunctions which are zero outside of the domain D. Alternatively, one could impose the additional condition on the far field as done in [CoHa13].

9.5 Summary and Conclusion

In this paper, existence and discreteness for mixed interior transmission eigenvalues for a real-valued index of refraction are reviewed and sufficient conditions on the index of refraction as well as the estimates of the lower bound of positive eigen-

values are given. A new system of boundary integral equations to solve the mixed interior transmission problem is derived. Further, it is explained how this system can be approximated via the boundary element collocation method. The resulting nonlinear eigenvalue problem is then solved with complex-valued contour integrals. Extensive numerical results for the computation of mixed interior transmission eigenvalues are provided for the first time for a variety of two-dimensional scatterers and might therefore serve as reference values for new algorithms in the future. Further, an explicit expression for mixed interior transmission eigenvalues is given for the unit square and can therefore be used to check the approximation quality of new algorithms. Moreover, the eigenfunctions are shown as well. Hence, it might be worthwhile studying the behavior of the eigenfunctions both for regular scatterers as well as scatterers with corners (see [BlLiLiWa17]). Additionally, a rigorous convergence analysis needs to be worked out in the future. In sum, this chapter might provide a fundamental basis for a further study of this interesting eigenvalue problem. One direction could be the investigation whether the inside-outside-duality method (see [KiLe13, LePe14, PeKl16]) can be applied both theoretically and practically to the mixed interior transmission problem.

Acknowledgments The first author would like to thank the *9th International Conference on Computational and Mathematical Methods in Science and Engineering (CMMSE'19)* steering committee for the opportunity to present the recent results for the numerical calculation of mixed interior transmission eigenvalues on July 2nd, 2019. Further, he would like to thank Christian Constanda for the organization of and the invitation to the special session *Integral Methods in Science and Engineering* at the *CMMSE'19* in Spain. The second author is supported by the National Natural Science Foundation of China (Grant Nos. 91730304 and 11531005).

Appendix

We consider the unit square $\tilde{\square}$ with transmission boundary condition on the north part and homogeneous Dirichlet conditions on the remaining edges. Separation of variables gives

$$u(x,y) = (A\sin(\pi px) + B\cos(\pi px))\left(Ce^{\lambda y} + De^{-\lambda y}\right)$$
$$v(x,y) = \left(\hat{A}\sin(\pi px) + \hat{B}\cos(\pi px)\right)\left(\hat{C}e^{\hat{\lambda} y} + \hat{D}e^{-\hat{\lambda} y}\right)$$

with n the given index of refraction. Here $\lambda = \sqrt{\pi^2 p^2 - k^2}$ and $\hat{\lambda} = \sqrt{\pi^2 p^2 - nk^2}$. Using the boundary condition $u = v = 0$ on the east part yields $B = \hat{B} = 0$, the boundary condition $u = v = 0$ gives $p \in \mathbb{N}$, and the boundary condition $u = v = 0$ on the south part yields $D = -C$ and $\hat{D} = -\hat{C}$. Hence, we have

$$u(x,y) = \sin(\pi px)\left(e^{\lambda y} - e^{-\lambda y}\right), \qquad v(x,y) = c\sin(\pi px)\left(e^{\hat{\lambda} y} - e^{-\hat{\lambda} y}\right)$$

Table 9.5 The first twelve real-valued MITEs for $\tilde{\square}$ using the index of refraction $n = 4$

2.671 552 787 839 805 (1)	3.666 034 666 514 623 (2)	4.836 476 632 026 555 (2)
5.037 735 005 038 399 (3)	5.735 963 618 893 019 (1)	5.883 918 727 662 464 (3)
6.294 288 796 613 341 (2)	6.516 005 567 788 862 (4)	7.024 814 731 040 726 (2)
7.038 184 014 872 755 (3)	7.160 899 902 925 930 (4)	7.695 549 983 552 737 (4)

with $c \in \mathbb{R}$ a free parameter. The first transmission condition on the north part ($y = 1$) gives

$$c = \left\{ e^{\lambda} - e^{-\lambda} \right\} \setminus \left\{ e^{\hat{\lambda}} - e^{-\hat{\lambda}} \right\} . \tag{9.30}$$

The second transmission condition on the north part yields

$$\left(e^{\lambda} + e^{-\lambda} \right) \lambda = c \left(e^{\hat{\lambda}} + e^{-\hat{\lambda}} \right) \hat{\lambda} . \tag{9.31}$$

Inserting (9.30) into (9.31) gives

$$\left(e^{\hat{\lambda}} - e^{-\hat{\lambda}} \right) \left(e^{\lambda} + e^{-\lambda} \right) \lambda - \left(e^{\lambda} - e^{-\lambda} \right) \left(e^{\hat{\lambda}} + e^{-\hat{\lambda}} \right) \hat{\lambda} = 0 . \tag{9.32}$$

Hence, the function, say $f_p(k)$, on the right-hand side has to be solved for a given $p \in \mathbb{N}$ with a root finding algorithm in order to obtain the MITE k. Note that the function can be complex-valued and therefore we have to consider separately the real and imaginary part. In Table 9.5 we summarize the highly accurate MITEs using the index of refraction $n = 4$. In parentheses we list the used parameter p.

References

[AnChAk13] Anagnostopoulos, K.A., Charalambopoulos, A., Kleefeld, A.: The factorization method for the acoustic transmission problem. Inverse Prob. **29**(11), 115015 (2013)

[At97] Atkinson, K.E.: The Numerical Solution of Integral Equations of the Second Kind. Cambridge University Press, Cambridge (1997)

[Be12] Beyn, W.-J.: An integral method for solving nonlinear eigenvalue problems. Linear Algebra Appl. **436**, 3839–3863 (2012)

[BlLiLiWa17] Blåsten, E., Li, X., Liu, H., Wang, Y.: On vanishing and localizing of transmission eigenfunctions near singular points: a numerical study. Inverse Prob. **33**(10), 105001 (2017)

[CaCo06] Cakoni, F., Colton, D.: Qualitative Methods in Inverse Scattering Theory – An Introduction. Springer, Berlin (2006)

[CaCoHa10] Cakoni, F., Colton, C., Haddar, H.: On the determination of Dirichlet or transmission eigenvalues from far field data. C. R. Math. **348**(7–8), 379–383 (2010)

[CaGiHa10] Cakoni, F., Gintides, D., Haddar, H.: The existence of an infinite discrete set of transmission eigenvalues. SIAM J. Math. Anal. **42**(1), 237–255 (2010)

[CaKr17] Cakoni, F., Kress, R.: A boundary integral equation method for the transmission eigenvalue problem. Appl. Anal. **96**(1), 23–38 (2017)

[Co11] Cossonnière, A.: Valeurs propres de transmission et leur utilisation dans l'identification d'inclusions à partir de mesures électromagnétiques. PhD thesis, Université de Toulouse (2011)

[CoHa13] Cossonnière, A., Haddar. H.: Surface integral formulation of the interior transmission problem J. Integral Equ. Appl. **25**(3), 341–376 (2013)

[CoKr13] Colton, D., Kress, R.: Inverse Acoustic and Electromagnetic Scattering Theory. Springer, Berlin (2013)

[CoMo87] Colton, D., Monk, P.: The inverse scattering problem for time-harmonic acoustic waves in a penetrable medium. Q. J. Mech. Appl. Math. **40**, 189–212 (1987)

[CoMo88] Colton, D., Monk, P.: The inverse scattering problem for time-harmonic acoustic waves in an inhomogeneous medium. Q. J. Mech. Appl. Math. **41**, 97–125 (1988)

[GiPa13] Gintides, D., Pallikarakis, N.: A computational method for the inverse transmission eigenvalue problem. Inverse Prob. **29**(10), 104010 (2013)

[Ki86] Kirsch, A.: The denseness of the far field patterns for the transmission problem. IMA J. Appl. Math. **37**(3), 213–225 (1986)

[KiGr08] Kirsch, A., Grinberg, N.: The Factorization Method for Inverse Problems. Oxford University Press, Oxford (2008)

[KiKl12] Kirsch, A., Kleefeld, A.: The factorization method for a conductive boundary condition. J. Integral Equ. Appl. **24**(4), 575–601 (2012)

[KiLe13] Kirsch A., Lechleiter, A.: The inside–outside duality for scattering problems by inhomogeneous media. Inverse Prob. **29**(10), 104011 (2013)

[Kl12a] Kleefeld, A.: The exterior problem for the Helmholtz equation with mixed boundary conditions in three dimensions. Int. J. Comput. Math. **89**(17), 2392–2409 (2012)

[Kl12b] Kleefeld, A.: A modified boundary integral equation for solving the exterior Robin problem for the Helmholtz equation in three dimensions. Appl. Math. Comput. **219**(4), 2114–2123 (2012)

[Kl12c] Kleefeld, A.: The transmission problem for the Helmholtz equation in $\mathbb{R}^3$. Comput. Methods Appl. Math., **12**(3), 330–350 (2012)

[Kl13] Kleefeld, A.: A numerical method to compute interior transmission eigenvalues. Inverse Prob. **29**(10), 104012 (2013)

[Kl15] Kleefeld, A.: Numerical methods for acoustic and electromagnetic scattering: Transmission boundary-value problems, interior transmission eigenvalues, and the factorization method. Habilitation Thesis, Brandenburg University of Technology Cottbus-Senftenberg (2015)

[KlLi11] Kleefeld, A., Lin, T.-C.: The nonlinear Landweber method applied to an inverse scattering problem for sound-soft obstacles in 3D. Comput. Phys. Commun. **182**(12), 2550–2560 (2011)

[KlLi12] Kleefeld, A., Lin, T.-C.: Boundary element collocation method for solving the exterior Neumann problem for Helmholtz's equation in three dimensions. Electron. Trans. Numer. Anal. **39**, 113–143 (2012)

[KlPi18] Kleefeld, A., Pieronek, L.: The method of fundamental solutions for computing acoustic interior transmission eigenvalues. Inverse Prob. **34**(3), 035007 (2018)

[LePe14] Lechleiter, A., Peters, S.: Analytical characterization and numerical approximation of interior eigenvalues for impenetrable scatterers from far fields. Inverse Prob. **30**(4), 045006 (2014)

[LiHuLiLi15] Li, T.-X., Huang, W.-Q., Lin, W.-W., Liu, J.-J.: On spectral analysis and a novel algorithm for transmission eigenvalue problems. J. Sci. Comput. **64**(1), 83–108 (2015)

[LiLi16] Li, T.-X., Liu, J.-J.: Transmission eigenvalue problem for inhomogeneous absorbing media with mixed boundary condition. Sci. China Math. **59**(6), 1081–1094 (2016)

[LiMa72] Lions, J.-L., Magenes, E.: Non-homogeneous Boundary Value Problems and Applications. Springer, Heidelberg (1972)

[Mc00] McLean, W.: Strongly Elliptic Systems and Boundary Integral Operators. Cambridge University Press, Cambridge (2000)

[PeKl16] Peters, S., Kleefeld, A.: Numerical computations of interior transmission eigenvalues for scattering objects with cavities. Inverse Prob. **32**(4), 045001 (2016)

[StUn12] Steinbach, O., Unger, G.: Convergence analysis of a Galerkin boundary element method for the Dirichlet Laplacian eigenvalue problem. SIAM J. Numer. Anal. **50**(2), 710–728 (2012)

[Su11] Sun, J.: Iterative methods for transmission eigenvalues. SIAM J. Numer. Anal. **49**(5), 1860–1874 (2011)

[YaMo14] Yang, F., Monk, P.: The interior transmission problem for regions on a conducting surface. Inverse Prob. **30**(1), 015007 (2014)

Chapter 10
An Inequality for Hölder Continuous Functions Generalizing a Result of Carlo Miranda

Massimo Lanza de Cristoforis

10.1 Introduction

This paper concerns an inequality which can be used to prove that a continuously differentiable real-valued function

$$u : \Omega \to \mathbb{R}$$

defined on an open subset Ω of $\mathbb{R}^n$ is actually α-Hölder continuous for a given $\alpha \in]0, 1]$, and it develops an idea which has already been exploited by Agmon et al. [AgDoNi59], and then by Miranda [Mi65] to prove regularity statements for layer potentials.

If $\Omega =]0, 1[$, then an elementary sufficient condition for u to be α-Hölder continuous is that

$$|t|^{1-\alpha}|u'(t)| \quad \text{be bounded in } t \in]0, 1[.$$

Then Agmon et al. [AgDoNi59, p. 717] have observed that if

$$\Omega = \mathbb{B}_{n-1}(0, 1)\times]0, 1[,$$

The original version of this chapter was revised: Family name of the author has been updated. A correction to this chapter is available at https://doi.org/10.1007/978-3-030-48186-5_13

M. Lanza de Cristoforis (✉)
Dipartimento di Matematica 'Tullio Levi-Civita', Università degli Studi di Padova, Padova, Italy
e-mail: mldc@math.unipd.it

C. Constanda (ed.), *Computational and Analytic Methods in Science and Engineering*, https://doi.org/10.1007/978-3-030-48186-5_10

then a sufficient condition for the α-Hölder continuity of $u \in C^1(\Omega)$ is that

$$|t|^{1-\alpha}|Du(x,t)| \qquad \text{is bounded in } (x,t) \in \mathbb{B}_{n-1}(0,1)\times]0,1[\,,$$

where $Du(x,t)$ is the Jacobian matrix of u at the point (x,t). Miranda [Mi65] has considered the case in which Ω is of class $C^{1,\alpha}$, a case in which Ω is locally around its boundary points the translation of the rotation of the strict hypograph of a function

$$\gamma \in C^{1,\alpha}(\overline{\mathbb{B}_{n-1}(0,r)},]-\delta,\delta[)$$

for some $r, \delta \in]0,+\infty[$. Then Miranda has observed that a sufficient condition for the α-Hölder continuity of $u \in C^1(\text{hypograph}_s(\gamma))$ is that

$$|\gamma(\eta)-t|^{1-\alpha}|Du(\eta,t)| \qquad \text{is bounded in } (\eta,t) \in \text{hypograph}_s(\gamma)\,, \tag{10.1}$$

where

$$\text{hypograph}_s(\gamma) \equiv \{(\eta,y) \in \mathbb{B}_{n-1}(0,r)\times]-\delta,\delta[:\ y < \gamma(\eta)\}$$

is the strict hypograph of γ. As we have said above, Miranda has exploited such inequality in order to prove a regularity result for a layer potential in case Ω is of class $C^{1,\alpha}$ for some $\alpha \in]0,1[$. In an effort to simplify the proof of Miranda and to refine his results, we wish to prove an inequality which on the one hand generalizes the above inequality of Miranda, and which on the other hand could be "intrinsic," in the sense that it would not be expressed in terms of local coordinates. As a bypass product, Miranda's ideas become easier to understand. One "intrinsic" inequality for a (bounded) $u \in C^1(\Omega)$ to be α-Hölder continuous is the following:

$$Du(y)(\text{dist}(y,\partial\Omega))^{1-\alpha} \quad \text{is bounded in } y \in \Omega\,, \tag{10.2}$$

an inequality which holds in uniform domains (personal communication of Aikawa [Ai19]). Such inequality or actually its variant

$$|Du(z)|(1-|z|^2)^{1-\alpha} \ \text{ is bounded in } z \in B_{\mathbb{C}^n}(0,1)$$

has been used in the analysis of spaces of analytic functions in the unit ball $B_{\mathbb{C}^n}(0,1)$ in $\mathbb{C}^n$ with center at 0 (cf. e.g., Zhu [Zh05, §7.2]). One could probably try with it, but the disadvantage is that $C^{1,1}$ is known to be the minimal regularity (in the Schauder scale) to have a unique projection ξ_y on $\partial\Omega$, i.e., a unique point $\xi_y \in \partial\Omega$ such that

$$|y-\xi_y| = \text{dist}(y,\partial\Omega)\,,$$

for all $y \in \mathbb{R}^n \setminus \partial\Omega$ close to $\partial\Omega$, and we are interested into the case $C^{1,\alpha}$ with $\alpha \in]0,1[$. In this sense, it is hard to see how its use may simplify/improve the proof of Miranda [Mi65] for layer potentials.

Since all functions $u \in C^1(\Omega)$ are locally Lipschitz continuous, the α-Hölder continuity of u follows by the α-Hölder continuity of u close to the boundary, and thus it suffices to prove an inequality of the α-Hölder continuity of u close to the boundary.

Thus we face the problem of choosing a neighborhood of the boundary, and we want to choose it in such a way that such a neighborhood be "globally parametrized." One may think of considering a set of the form

$$N(t_1) \equiv \{x + t\nu_\Omega(x) : t \in]-t_1, t_1[, x \in \partial\Omega\},$$

where ν_Ω is the outward unit normal to $\partial\Omega$ and $t_1 > 0$ is small enough, but the problem with using $N(t_1)$ is that unfortunately the outward unit normal is only of class $C^{0,\alpha}$, and thus not of class $C^{0,1}$ as required for the map $x + t\nu_\Omega(x)$ of the variable (x, t) to be injective and open, to infer that $N(t_1)$ is actually an open neighborhood of $\partial\Omega$ and to carry out our proof of Hölder continuity of u. So the idea here in case Ω is of class $C^{1,\alpha}$, or perhaps only a Lipschitz set is to replace the vector field ν_Ω by a vector field a such that the set

$$A(t_1) \equiv \{x + ta(x) : t \in]-t_1, t_1[, x \in \partial\Omega\}$$

be an open neighborhood of $\partial\Omega$ and such that the map $x+ta(x)$ of the variable (x, t) be Lipschitz continuous, injective and open and to prove a variant of Miranda's inequality (10.1) in the set

$$A(t_1)^+ \equiv \{x + ta(x) : t \in]-t_1, 0[, x \in \partial\Omega\}.$$

We formulate a body of appropriate assumptions on a in Definition 10.4 of outer nontangential unit vector field, and we note that such assumptions do not involve the outward unit normal ν_Ω. Then we prove that to check the α-Hölder continuity of a function $u \in C^1(\Omega)$ in case Ω is a bounded open Lipschitz set, it suffices to show the existence of a finite upper bound for

$$|t|^{1-\alpha}|Du(x + ta(x))|$$

when x belongs to $\partial\Omega$ and when t belongs to $]-t_1, 0[$ (see Proposition 10.1). By exploiting such an inequality, one can simplify the proof of the result on layer potentials of Miranda [Mi65] and weaken its assumptions on the kernel. This type of application appears in the monograph [DaLaMu19, Ch. 2] with Dalla Riva and Musolino.

In [La19], we discuss the existence of a nontangential unit vector field as a for a bounded open Lipschitz set.

In [La19] we also show that the existence of an outer nontangential unit vector field as a for a bounded open Lipschitz set implies that the scalar product $a(x) \cdot \nu_\Omega(x)$ is bounded from below by a positive constant which is independent of the point $x \in \partial\Omega$ at which the outward unit normal $\nu_\Omega(x)$ exists and we note that Fichera [Fi55, pp. 207–208] had already recognized the importance of the

existence of a vector field a such that the essential infimum of $a \cdot \nu_\Omega$ is bounded away from 0 for sets with a piecewise smooth boundary in the analysis of boundary value problems for systems of partial differential equations, and in particular for the analysis of boundary value problems with unilateral constraints (cf. Fichera [Fi84, p. 413]). Then for Lipschitz sets, we mention Grisvard [Gr85, Lem 1.5.1.9]. The main difference of our conditions on a and those of Fichera and Grisvard is that our conditions are completely independent of the outward unit normal ν_Ω (which exists only almost everywhere).

10.2 Preliminaries and Notation

We denote the norm on a normed space $\mathscr{X}$ by $\|\cdot\|_{\mathscr{X}}$. Let $\mathscr{X}$ and $\mathscr{Y}$ be normed spaces. We endow the space $\mathscr{X} \times \mathscr{Y}$ with the norm defined by $\|(x, y)\|_{\mathscr{X}\times\mathscr{Y}} \equiv \|x\|_{\mathscr{X}} + \|y\|_{\mathscr{Y}}$ for all $(x, y) \in \mathscr{X} \times \mathscr{Y}$, while we use the Euclidean norm for $\mathbb{R}^n$.

The symbol $\mathbb{N}$ denotes the set of natural numbers including 0. Let $\mathbb{D} \subseteq \mathbb{R}^n$. Then $\overline{\mathbb{D}}$ denotes the closure of $\mathbb{D}$, and $\partial\mathbb{D}$ denotes the boundary and $\mathrm{diam}(\mathbb{D})$ denotes the diameter of $\mathbb{D}$, and "dist" is short for "distance." If Ω is an open subset of $\mathbb{R}^n$, then we set

$$\Omega^- \equiv \mathbb{R}^n \setminus \overline{\Omega}\,.$$

Let $n \in \mathbb{N} \setminus \{0\}$. We denote by $\mathbb{O}_n(\mathbb{R})$ the set of $n \times n$ orthogonal matrices with real entries. Let A be a matrix. Then A^t denotes the transpose matrix of A.

The symbol $|\cdot|$ denotes the Euclidean modulus in $\mathbb{R}^n$. For all $R \in]0, +\infty[$, $x \in \mathbb{R}^n$, x_j denotes the j-th coordinate of x, and we set

$$\mathbb{B}_n(x, R) \equiv \{y \in \mathbb{R}^n : |x - y| < R\}\,.$$

Let Ω be an open subset of $\mathbb{R}^n$. The space of m times continuously differentiable real-valued functions on Ω is denoted by $C^m(\Omega)$. Let $f \in C^m(\Omega)$. Then Df denotes the Jacobian matrix of f. Let $\eta \equiv (\eta_1, \ldots, \eta_n) \in \mathbb{N}^n$, $|\eta| \equiv \eta_1 + \ldots + \eta_n$. Then $D^\eta f$ denotes $\frac{\partial^{|\eta|} f}{\partial x_1^{\eta_1} \ldots \partial x_n^{\eta_n}}$. The subspace of $C^m(\Omega)$ of those functions f whose derivatives $D^\eta f$ of order $|\eta| \leq m$ can be extended with continuity to $\overline{\Omega}$ is denoted $C^m(\overline{\Omega})$. The subspace of $C^m(\overline{\Omega})$ whose functions have m-th order derivatives that are Hölder continuous with exponent $\alpha \in]0, 1]$ is denoted $C^{m,\alpha}(\overline{\Omega})$ (cf. e.g., Gilbarg and Trudinger [GiTr83]). Let $\mathbb{D} \subseteq \mathbb{R}^n$. Then $C^{m,\alpha}(\overline{\Omega}, \mathbb{D})$ denotes $\left\{f \in \left(C^{m,\alpha}(\overline{\Omega})\right)^n : f(\overline{\Omega}) \subseteq \mathbb{D}\right\}$.

Now let Ω be a bounded open subset of $\mathbb{R}^n$. If $f \in C^{0,\alpha}(\overline{\Omega})$, then its Hölder constant $|f : \overline{\Omega}|_\alpha$ is defined as $\sup\left\{\frac{|f(x)-f(y)|}{|x-y|^\alpha} : x, y \in \overline{\Omega}, x \neq y\right\}$, and we also write $\mathrm{Lip}(f) \equiv |f : \overline{\Omega}|_1$. Then $C^m(\overline{\Omega})$ and $C^{m,\alpha}(\overline{\Omega})$ are endowed with their usual norm and are well known to be Banach spaces (cf. e.g., Troianiello [Tr87, §1.2.1]).

We now wish to introduce the well-known definition of a set which is locally around each of its boundary points a "rotated" strict hypograph of a continuous function. We do so by requiring that if p is a boundary point of Ω, then the translated set $\Omega - p$ can be rotated around the origin by means of a rotation R (or more generally by $R \in \mathbb{O}_n(\mathbb{R})$) so that the rotated set $R(\Omega - p)$ equals the strict hypograph of a continuous function γ, at least around the point 0. To do so, we need to introduce the following, which is in the wake of a corresponding terminology of Burenkov [Bu98] for the analysis of Sobolev spaces on domains.

Definition 10.1 Let $n \in \mathbb{N} \setminus \{0, 1\}$. Let Ω be an open subset of $\mathbb{R}^n$. Let $p \in \partial\Omega$, $R \in \mathbb{O}_n(\mathbb{R})$, $r, \delta \in]0, +\infty[$. We say that the set

$$C(p, R, r, \delta) \equiv p + R^t(\mathbb{B}_{n-1}(0, r)\times] - \delta, \delta[)\,,$$

is a coordinate cylinder for Ω around p, provided that the intersection

$$R(\Omega - p) \cap (\mathbb{B}_{n-1}(0, r)\times] - \delta, \delta[)$$

equals the strict hypograph of a continuous function γ from $\mathbb{B}_{n-1}(0, r)$ to $] - \delta, \delta[$ which vanishes at 0 and such that $|\gamma(\eta)| < \delta/2$ for all $\eta \in \mathbb{B}_{n-1}(0, r)$, i.e., provided that there exists $\gamma \in C^0(\mathbb{B}_{n-1}(0, r),] - \delta, \delta[)$ such that

$$\begin{aligned} R(\Omega - p) \cap (\mathbb{B}_{n-1}(0, r)\times] - \delta, \delta[) & \\ = \{(\eta, y) \in \mathbb{B}_{n-1}(0, r)\times] - \delta, \delta[:\ y < \gamma(\eta)\} & \\ \equiv \text{hypograph}_s(\gamma)\,, & \\ |\gamma(\eta)| < \delta/2 \qquad \forall \eta \in \mathbb{B}_{n-1}(0, r)\,, \qquad \gamma(0) = 0\,. & \end{aligned} \tag{10.3}$$

Given a coordinate cylinder $C(p, R, r, \delta)$ for Ω around p, the corresponding function γ is uniquely determined and

$$\gamma(\eta) = \sup\{y \in] - \delta, \delta[:\ (\eta, y) \in R(\Omega - p) \cap (\mathbb{B}_{n-1}(0, r)\times] - \delta, \delta[)\}$$

for all $\eta \in \mathbb{B}_{n-1}(0, r)$. We also note that the continuity of γ implies that

$$\begin{aligned} R((\partial\Omega) - p) \cap (\mathbb{B}_{n-1}(0, r)\times] - \delta, \delta[) & \\ = \{(\eta, y) \in \mathbb{B}_{n-1}(0, r)\times] - \delta, \delta[:\ y = \gamma(\eta)\} & \\ \equiv \text{graph}(\gamma)\,. & \end{aligned} \tag{10.4}$$

We say that γ is the function which represents $\partial\Omega$ in the coordinate cylinder $C(p, R, r, \delta)$ as a graph and that the function ψ_p from $\mathbb{B}_{n-1}(0, r)$ to $\mathbb{R}^n$ defined by

$$\psi_p(\eta) \equiv p + R^t \begin{pmatrix} \eta \\ \gamma(\eta) \end{pmatrix} \qquad \forall \eta \in \mathbb{B}_{n-1}(0, r)\,, \tag{10.5}$$

is the parametrization of $\partial\Omega$ around p in the coordinate cylinder $C(p, R, r, \delta)$.

Since the continuous function γ induces the homeomorphism $(\cdot, \gamma(\cdot))$ from its domain onto its graph, the map ψ_p is a homeomorphism of $\mathbb{B}_{n-1}(0, r)$ onto $\psi_p(\mathbb{B}_{n-1}(0, r)) = (\partial\Omega) \cap C(p, R, r, \delta)$.

We also note that $\mathrm{hypograph}_s(\gamma)$ is easily seen to be path connected and that accordingly

$$\Omega \cap C(p, R, r, \delta) = p + R^t(\mathrm{hypograph}_s(\gamma))$$

is path connected. Hence, $\Omega \cap C(p, R, r, \delta)$ is contained in at most one connected component of Ω. It is sometimes useful to know that by shrinking r we still obtain a coordinate cylinder around the point p. More precisely, we have the following.

Remark 10.1 If $C(p, R, r, \delta)$ is a coordinate cylinder around the point p of $\partial\Omega$, then also $C(p, R, \rho, \delta)$ is a coordinate cylinder around the point p of $\partial\Omega$ for each $\rho \in]0, r[$, and the restriction $\gamma_{|\mathbb{B}_{n-1}(0,\rho)}$ represents $\partial\Omega$ in $C(p, R, \rho, \delta)$ as a graph.

In order to compactify our notation, we find convenient to set

$$C^{0,0} \equiv C^0 . \tag{10.6}$$

We are now ready to introduce the following.

Definition 10.2 Let $n \in \mathbb{N} \setminus \{0, 1\}$. Let $\alpha \in [0, 1]$. We say that an open subset Ω of $\mathbb{R}^n$ is a local strict hypograph of class $C^{0,\alpha}$ provided that for every point $p \in \partial\Omega$, there exist $R \in \mathbb{O}_n(\mathbb{R})$ and $r, \delta \in]0, +\infty[$ such that $C(p, R, r, \delta)$ is a coordinate cylinder for Ω around p and that the corresponding function γ which represents $\partial\Omega$ as a graph in $C(p, R, r, \delta)$ is of class $C^{0,\alpha}(\overline{\mathbb{B}_{n-1}(0, r)})$. (Here we understand that γ has a unique extension to $\overline{\mathbb{B}_{n-1}(0, r)}$ that is of class $C^{0,\alpha}(\overline{\mathbb{B}_{n-1}(0, r)})$ and that we still denote with the same symbol γ.)

One could show that if a bounded open subset Ω of $\mathbb{R}^n$ is a local strict hypograph of class C^0, then the number of connected components of Ω and of its exterior Ω^- is necessarily finite (cf. e.g., [DaLaMu19]). Then we have the following "folklore" statement. For the proof, we refer to the Appendix.

Lemma 10.1 *Let $n \in \mathbb{N} \setminus \{0, 1\}$. Let $\alpha \in [0, 1]$. Let Ω be a bounded open local strict hypograph of class $C^{0,\alpha}$. Let $r_*, \delta_* \in]0, +\infty[$. Then there exist $r \in]0, r_*[$, $\delta \in]0, \delta_*[$, $r < \delta$ such that for each $x \in \partial\Omega$ there exists $R_x \in \mathbb{O}_n(\mathbb{R})$ such that $C(x, R_x, r, \delta)$ is a coordinate cylinder for Ω around x and the corresponding function γ_x satisfies the inequality*

$$\sup_{x\in\partial\Omega} \|\gamma_x\|_{C^{0,\alpha}(\overline{\mathbb{B}_{n-1}(0,r)})} < +\infty .$$

10.3 Definition of Outer Nontangential Unit Vector Field and Introduction of a Neighborhood of the Boundary

If Ω is a bounded open subset of class C^1, then one can exploit the continuity of the outward unit normal ν_Ω and prove that if $\vartheta \in]0, 1[$, then there exist a vector field a of class C^1 in the closure of a neighborhood U of $\partial\Omega$ and $\tau \in]0, +\infty[$ such that the following conditions hold

$$\begin{aligned} &|a(x)| = 1 \qquad \forall x \in \partial\overline{U}\,, \\ &\sup_{\partial\Omega} |a - \nu_\Omega| < \vartheta\,, \\ &\inf_{\partial\Omega} a \cdot \nu_\Omega > 1 - \frac{\vartheta^2}{2} > 1 - \vartheta\,, \\ &|a(x) \cdot (y - x)| \le \vartheta |x - y| \qquad \forall x, y \in \partial\Omega\,,\ |x - y| < \tau\,, \end{aligned} \tag{10.7}$$

(see reference [DaLaMu19, Ch. 2] with Dalla Riva and Musolino). Then one can exploit the properties of a in (10.7) and prove that there exists $t_1 \in]0, +\infty[$ such that the following statements hold.

(i) The map Ψ from $(\partial\Omega)\times]-t_1, t_1[$ to $\mathbb{R}^n$ defined by

$$\Psi(x, t) = x + ta(x) \qquad \forall (x, t) \in (\partial\Omega)\times]-t_1, t_1[\,,$$

is injective, and the set

$$A(t_2) \equiv \Psi((\partial\Omega)\times]-t_2, t_2[) = \{x + ta(x) : t \in]-t_2, t_2[, x \in \partial\Omega\}$$

is an open neighborhood of $\partial\Omega$ for all $t_2 \in]0, t_1]$

(ii)

$$x + ta(x) \in \Omega \qquad \forall t \in]-t_1, 0[\,, \qquad x + ta(x) \in \Omega^- \qquad \forall t \in]0, t_1[\,,$$

(cf. [DaLaMu19, Ch. 2]). We now wonder whether the existence of a vector field a as in (10.7) around the boundary of an open set of class at least $C^{0,1}$ implies the existence of $t_1 \in]0, +\infty[$ so that the map Ψ satisfies the conditions (i) and (ii), which we need to prove a generalization of Miranda's inequality (10.1) in the set $A(t_2) \cap \Omega$ for $t_2 \in]0, t_1]$ small enough.

Since sets of class $C^{0,1}$ do not necessarily have an outward normal at all points of the boundary, we need to formulate assumptions on a which do not involve the outward unit normal. To do so, we need the following well-known definition.

Definition 10.3 Let $n \in \mathbb{N} \setminus \{0, 1\}$. Let Ω be an open subset of $\mathbb{R}^n$. Let $x \in \partial\Omega$. Let $v \in \mathbb{R}^n \setminus \{0\}$.

(i) We say that v points to the interior of Ω at x provided that there exists $\varsigma_{x,v} > 0$ such that

$$x + tv \in \Omega \qquad \forall t \in]0, \varsigma_{x,v}[.$$

(ii) We say that v points to the exterior of Ω at x provided that there exists $\varsigma_{x,v} > 0$ such that

$$x + tv \in \mathbb{R}^n \setminus \overline{\Omega} \qquad \forall t \in]0, \varsigma_{x,v}[.$$

In general, we cannot expect that a unit vector v points either to the interior or to the exterior of Ω, but this may happen under certain assumptions. Then we are ready to introduce the following definition, which does not involve the existence of the outward unit normal.

Definition 10.4 Let $n \in \mathbb{N} \setminus \{0, 1\}$. Let Ω be a bounded open subset of $\mathbb{R}^n$. Let $\vartheta \in]0, 1[$. We say that a map a from $\partial\Omega$ to $\mathbb{R}^n$ is an outer nontangential unit vector field for Ω with parameter ϑ provided that the following conditions are satisfied.

(i) $|a(x)| = 1$ for all $x \in \partial\Omega$.
(ii) There exists $\tau \in]0, +\infty[$ such that

$$|a(x) \cdot (y - x)| \leq \vartheta |y - x| \qquad \forall x, y \in \partial\Omega \text{ such that } |x - y| < \tau .$$

(iii) $a(x)$ points to the exterior of Ω at all points $x \in \partial\Omega$ and $-a(x)$ points to the interior of Ω at all points $x \in \partial\Omega$.

An outer nontangential unit vector field for Ω with parameter ϑ can play the role of a normal in the definition of a tubular neighborhood of $\partial\Omega$, as the following statement shows (see also reference [LaRo08] with Rossi for a related result which holds under stronger assumptions on Ω).

Lemma 10.2 *Let Ω be a bounded open strict local hypograph of class C^0. Let $\vartheta \in]0, 1[$. Let $n \in \mathbb{N} \setminus \{0, 1\}$. Let $a \in C^{0,1}(\partial\Omega, \mathbb{R}^n)$ be an outer nontangential unit vector field for Ω with parameter ϑ.*

Then there exists $t_1 \in]0, +\infty[$ such that the following statements hold.

(i) The map Ψ from $(\partial\Omega)\times]-t_1, t_1[$ to $\mathbb{R}^n$ defined by

$$\Psi(x, t) = x + ta(x) \qquad \forall (x, t) \in (\partial\Omega)\times]-t_1, t_1[,$$

is injective, and the set

$$A(t_2) \equiv \Psi((\partial\Omega)\times]-t_2, t_2[) = \{x + ta(x) : t \in]-t_2, t_2[, x \in \partial\Omega\}$$

is an open neighborhood of $\partial\Omega$ for all $t_2 \in]0, t_1]$

(ii) We have

$$x + ta(x) \in \Omega \qquad \forall t \in]-t_1, 0[\,, \qquad x + ta(x) \in \Omega^- \qquad \forall t \in]0, t_1[\,.$$

Proof Let τ be as in Definition 10.4(ii). By Lemma 10.1, there exist $r, \delta \in]0, +\infty[$ such that

$$r < \delta \leq \tau/2\,,$$

and such that for each $x \in \partial\Omega$ there exists $R_x \in \mathbb{O}_n(\mathbb{R})$, such that $C(x, R_x, r, \delta)$ is a coordinate cylinder for Ω around x with $\gamma_x \in C^0(\overline{\mathbb{B}_{n-1}(0, r)},]-\delta, \delta[)$ such that $\gamma_x(0) = x$, $|\gamma_x| < \delta/2$ as function which represents $\partial\Omega$ in $C(x, R_x, r, \delta)$ as a graph. We first show the existence of t_1 as in (i). Assume by contradiction that Ψ is not injective for any choice of t_1. Then for each $j \in \mathbb{N}$ there exist pairs (x'_j, t'_j), (x''_j, t''_j) in $\partial\Omega \times]-2^{-j}, 2^{-j}[$ such that

$$(x'_j, t'_j) \neq (x''_j, t''_j)\,, \qquad \Psi(x'_j, t'_j) = \Psi(x''_j, t''_j)\,. \tag{10.8}$$

In particular,

$$x'_j \neq x''_j \qquad \forall j \in \mathbb{N}\,.$$

Indeed, if $x'_j = x''_j$, then $a(x'_j) = a(x''_j)$ and the equality

$$x'_j + t'_j a(x'_j) = x''_j + t''_j a(x''_j)$$

implies that

$$t'_j a(x'_j) = t''_j a(x'_j)\,,$$

and thus $t'_j = t''_j$, contrary to $(x'_j, t'_j) \neq (x''_j, t''_j)$. Possibly selecting a subsequence, we can assume that there exist $x', x'' \in \partial\Omega$ such that

$$x' = \lim_{j\to\infty} x'_j\,, \qquad x'' = \lim_{j\to\infty} x''_j\,.$$

Since a is continuous and $\lim_{j\to\infty} t'_j = 0$, the second equality of (10.8) implies that $x' = x'' \equiv \tilde{x}$. By our contradiction assumption (10.8), we have the equality

$$0 = x''_j + t''_j a(x''_j) - x'_j - t'_j a(x'_j) \qquad \forall j \in \mathbb{N}\,,$$

which we rewrite as

$$0 = (x''_j - x'_j) + (t''_j - t'_j)a(x''_j) + t'_j(a(x''_j) - a(x'_j)) \qquad \forall j \in \mathbb{N}\,. \tag{10.9}$$

Then by dividing by $|x'_j - x''_j|$, we obtain

$$\frac{t''_j - t'_j}{|x''_j - x'_j|} a(x''_j) = -\frac{x''_j - x'_j}{|x''_j - x'_j|} - t'_j \frac{a(x''_j) - a(x'_j)}{|x''_j - x'_j|} \qquad \forall j \in \mathbb{N},$$

and thus the triangular inequality implies that

$$\left| \frac{t''_j - t'_j}{|x''_j - x'_j|} \right| \leq 1 + |t'_j| \mathrm{Lip}(a) \qquad \forall j \in \mathbb{N}. \tag{10.10}$$

Next we go back to the above equality (10.9) and take the scalar product with $\frac{x''_j - x'_j}{|x''_j - x'_j|^2}$ and obtain

$$0 = 1 + \frac{t''_j - t'_j}{|x''_j - x'_j|} a(x''_j) \cdot \frac{x''_j - x'_j}{|x''_j - x'_j|} + t'_j \frac{a(x''_j) - a(x'_j)}{|x''_j - x'_j|} \cdot \frac{x''_j - x'_j}{|x''_j - x'_j|} \qquad \forall j \in \mathbb{N}.$$

In order to exploit condition (ii) of Definition 10.4, we choose $j_0 \in \mathbb{N}$ so that

$$|x''_j - x'_j| \leq |x''_j - \tilde{x}| + |\tilde{x} - x'_j| \leq \tau \qquad \forall j_0 \leq j \in \mathbb{N},$$

and thus inequality (10.10) and the above equality imply that

$$1 \leq (1 + |t'_j| \mathrm{Lip}(a))\vartheta + |t'_j| \mathrm{Lip}(a) \qquad \forall j_0 \leq j \in \mathbb{N}.$$

Since $\lim_{j \to \infty} t'_j = 0$, we obtain $1 \leq \vartheta$, a contradiction. Hence, there exists $t_1 \in]0, +\infty[$ such that Ψ is injective on $(\partial\Omega) \times]-t_1, t_1[$.

Next we turn to show that $A(t_2)$ is open for all $t_2 \in]0, t_1]$. Let $t_2 \in]0, t_1]$, $(x^\sharp, t^\sharp) \in A(t_2)$. Since Ψ is injective, then the composition $\Psi^\sharp$ of Ψ with the continuous and injective map $(x^\sharp + R^t_{x^\sharp}(\eta, \gamma(\eta))^t, t)$ of the variable (η, t), i.e., the map

$$\begin{aligned} \Psi^\sharp(\eta, t) &\equiv \Psi(x^\sharp + R^t_{x^\sharp}(\eta, \gamma(\eta))^t, t) \\ &= x^\sharp + R^t_{x^\sharp}(\eta, \gamma(\eta))^t + t a(x^\sharp + R^t_{x^\sharp}(\eta, \gamma(\eta))^t) \end{aligned}$$

for all $(\eta, t) \in \mathbb{B}_{n-1}(0, r) \times]-t_2, t_2[$, is continuous and injective. Then the Theorem of Invariance of Domain (cf. e.g., Deimling [De85, Thm. 4.3]) implies that $\Psi^\sharp(\mathbb{B}_{n-1}(0, r) \times]-t_2, t_2[)$ is an open subset of $\mathbb{R}^n$. Since

$$(x^\sharp, t^\sharp) \in \Psi^\sharp(\mathbb{B}_{n-1}(0, r) \times]-t_2, t_2[) \subseteq A(t_2),$$

it follows that $(x^\sharp, t^\sharp)$ is interior to $A(t_2)$. Hence, $A(t_2)$ is open.

Next we turn to prove statement (ii). Let $x \in \partial\Omega$. Then we set

$$A(x,t_1)^+ \equiv \{x + ta(x) : t \in]-t_1, 0[\}\,, \qquad A(x,t_1)^- \equiv \{x + ta(x) : t \in]0, t_1[\}\,.$$

Since Ψ is injective, we have

$$\Psi(\{x\}\times]-t_1, t_1[) \cap \partial\Omega = \{x\}\,.$$

Indeed, if there exists $t \in]-t_1, t_1[$ such that

$$y \equiv x + ta(x) \in \partial\Omega\,,$$

then

$$\Psi(y,0) = \Psi(x,t)$$

and the injectivity of Ψ implies that $x = y$, $t = 0$. Hence,

$$A(x,t_1)^{\pm} \subseteq \mathbb{R}^n \setminus \partial\Omega\,.$$

Since the arcs $A(x,t_1)^{\pm}$ cannot intersect $\partial\Omega$, they cannot contain both points of Ω and of Ω^-. We now show that

$$A(x,t_1)^+ \subseteq \Omega\,, \qquad A(x,t_1)^- \subseteq \Omega^-\,,$$

and we turn to prove the former inclusion. By assumption, $-a(x)$ points to the interior of Ω and thus there exists $\varsigma_{x,-a(x)} > 0$ such that

$$x + ta(x) \in \Omega \qquad \forall t \in]-\varsigma_{x,-a(x)}, 0[\,,$$

and thus in particular for all $t \in]-\min\{\varsigma_{x,-a(x)}, t_1\}, 0[$. Hence, the connected set $A(x,t_1)^+$ contains points of Ω and thus, as we have shown above, $A(x,t_1)^+$ cannot contain points of Ω^- and we must have $A(x,t_1)^+ \subseteq \Omega$. Then the inclusion $A(x,t_1)^- \subseteq \mathbb{R}^n \setminus \overline{\Omega}$ can be proved similarly. □

10.4 An Inequality for Hölder Continuous Functions on Local Hypographs

Next we introduce the following two preliminary technical statements.

Lemma 10.3 *Let $n \in \mathbb{N} \setminus \{0,1\}$. Let $\vartheta \in]0,1[$. If $v, w \in \mathbb{R}^n$ and if*

$$|v \cdot w| \le \vartheta |v|\,|w|\,,$$

then $|v+w|^2 \geq (1-\vartheta)(|v|^2+|w|^2)+\vartheta(|v|-|w|)^2$.

Proof It suffices to note that

$$\begin{aligned}|v+w|^2 &= |v|^2+|w|^2+2v\cdot w\\ &\geq |v|^2+|w|^2-2|v\cdot w| \geq |v|^2+|w|^2-2\vartheta|v||w|\\ &= (1-\vartheta)(|v|^2+|w|^2)+\vartheta(|v|^2+|w|^2-2|v||w|)\\ &= (1-\vartheta)(|v|^2+|w|^2)+\vartheta(|v|-|w|)^2\,.\end{aligned}$$

□

Lemma 10.4 *Let* $n \in \mathbb{N}\setminus\{0,1\}$. *Let* Ω *be a bounded open Lipschitz subset of* $\mathbb{R}^n$. *Let* $\vartheta \in]0,1[$. *Let* $a \in C^{0,1}(\partial\Omega,\mathbb{R}^n)$ *be an outer nontangential unit vector field for* Ω *with parameter* ϑ.

Let $\tau \in]0,+\infty[$ *satisfy the condition (ii) of Definition 10.4. Let* $t_1 \in]0,+\infty[$ *be as in Lemma 10.2. Let* $p \in \partial\Omega$. *Let* $r,\ \delta \in]0,+\infty[$,

$$2\delta < \min\{\tau/2, t_1\} \qquad r < \delta\,,$$

and $R_p \in \mathbb{O}_n(\mathbb{R})$ *be such that* $C(p,R_p,r,\delta)$ *is a coordinate cylinder for* Ω *around* p. *Let* $\gamma_p \in C^{0,1}(\overline{\mathbb{B}_{n-1}(0,r)})$ *represent* $\partial\Omega$ *in* $C(p,R_p,r,\delta)$ *as a graph. Let*

$$C(p,R_p,r,\delta)\cap\Omega \subseteq A(t_1)^+ \equiv \{x+ta(x) : t \in]-t_1,0[, x\in\partial\Omega\}\,.$$

Let

$$t_2 \in \left]0, \min\left\{\frac{r}{4}, \frac{(1-\vartheta)^{1/2}}{2\sqrt{2}(\mathrm{Lip}(a)+1)}, \frac{t_1}{2}\right\}\right[\,.$$

If $\alpha \in]0,1]$, *then there exists* $B \in]0,+\infty[$ *such that*

$$|f : [p+R_p^t(\mathbb{B}_{n-1}(0,r/4)\times]-\delta/4,\delta/4[)]\cap A(t_2)^+|_\alpha \leq BM_{t_1,\alpha}(f) \qquad (10.11)$$

for all $f \in C^1(\Omega)$ *such that*

$$M_{t_1,\alpha}(f) \equiv \sup\{|t|^{1-\alpha}|Df(x+ta(x))| : (x,t)\in(\partial\Omega)\times]-t_1,0[\} < +\infty\,. \qquad (10.12)$$

Proof If $M_{t_1,\alpha}(f)=0$, then $Df=0$ on $A(t_1)^+$ and accordingly on the connected set $C(p,R_p,r,\delta)\cap\Omega$ and f is constant on $C(p,R_p,r,\delta)\cap\Omega$. Hence, the inequality (10.11) holds true. Thus we can assume that $M_{t_1,\alpha}(f)>0$.

In order to estimate the Hölder constant of f as in (10.11), we take two arbitrary points p', $p''\in[p+R_p^t(\mathbb{B}_{n-1}(0,r/4)\times]-\delta/4,\delta/4[)]\cap A(t_2)^+$ and we turn to estimate $|f(p')-f(p'')|$. In order to exploit condition (10.12) on f, we plan to to

define a Lipschitz arc in $A(t_1)^+$ with endpoints p' and p''. Lemma 10.2 implies that there exist $x', x'' \in \partial\Omega$, $t', t'' \in]-t_2, 0[$ such that

$$p' = x' + t'a(x')\,, \qquad p'' = x'' + t''a(x'')\,.$$

By the triangular inequality and by the equality $|a(x')| = 1$, and by the inequality $t_2 < r/4 < \delta/4$, we have

$$\begin{aligned} x' \in \mathbb{B}_n(p', t_2) &\subseteq [p + R_p^t(\mathbb{B}_{n-1}(0, r/4)\times] - \delta/4, \delta/4[)] + \mathbb{B}_n(0, t_2) \\ &\subseteq [p + R_p^t(\mathbb{B}_{n-1}(0, r/2)\times] - \delta/2, \delta/2[)]\,, \end{aligned}$$

and similarly, $x'' \in [p + R_p^t(\mathbb{B}_{n-1}(0, r/2)\times] - \delta/2, \delta/2[)]$. Since $x', x'' \in \partial\Omega$, then there exist $\xi', \xi'' \in \mathbb{B}_{n-1}(0, r/2)$ such that

$$\begin{aligned} x' &= \phi_p(\xi') = p + R_p^t(\xi', \gamma_p(\xi'))^t\,, \\ x'' &= \phi_p(\xi'') = p + R_p^t(\xi'', \gamma_p(\xi''))^t\,. \end{aligned}$$

Next we plan to introduce a Lipschitz arc on $\partial\Omega$ with endpoints x' and x''. To do so, we note that the convexity of $\mathbb{B}_{n-1}(0, r/2)$ implies that

$$y(s) \equiv \xi' + s(\xi'' - \xi') \in \mathbb{B}_{n-1}(0, r/2) \qquad \forall s \in [0, 1]\,.$$

Since γ_p is Lipschitz continuous, the arc

$$\psi_{x',x''}(s) \equiv p + R_p^t(y(s), \gamma_p(y(s))) \qquad \forall s \in [0, 1]$$

in $\partial\Omega$ is rectifiable and

$$\text{length}(\psi_{\xi',\xi''}) \le \int_0^1 \sqrt{|\xi' - \xi''|^2 + \text{Lip}\,(\gamma_p)^2|\xi' - \xi''|^2}\,ds \le c_{\gamma_p}|\xi' - \xi''|$$

where

$$c_{\gamma_p} \equiv \sqrt{1 + \text{Lip}\,(\gamma_p)^2}\,.$$

Next we introduce a Lipschitz arc in $A^+(t_2)$ with endpoints p' and p''. By the convexity of $]-t_2, 0[$, we have

$$(1-s)t' + st'' \in]-t_2, 0[\qquad \forall s \in [0, 1]\,,$$

and thus we have

$$\Gamma(s) \equiv \psi_{x',x''}(s) + [(1-s)t' + st'']a(\psi_{x',x''}(s)) \in A^+(t_2) \qquad \forall s \in [0, 1]\,.$$

Next we note that

$$\begin{aligned}\Gamma(s) &\in [p + R_p^t(\mathbb{B}_{n-1}(0, r/2)\times] - \delta/2, \delta/2[)] + \mathbb{B}_n(0, t_2)\\ &\subseteq [p + R_p^t(\mathbb{B}_{n-1}(0, 3r/4)\times] - 3\delta/4, 3\delta/4[)] \quad \forall s \in [0, 1]\,,\\ \Gamma(0) &= p'\,, \qquad \Gamma(1) = p''\,.\end{aligned}$$

Moreover,

$$\begin{aligned}\mathrm{length}(\Gamma) &\leq \mathrm{length}(\psi_{\xi',\xi''}) + |t' - t''| + |t_2|\mathrm{Lip}\,(a)\mathrm{length}(\psi_{\xi',\xi''})\\ &\leq c_{\gamma_p}|\xi' - \xi''| + |t' - t''| + |t_2|\mathrm{Lip}\,(a)c_{\gamma_p}|\xi' - \xi''|\\ &\leq |x' - x''|c_{\gamma_p}(1 + |t_2|\mathrm{Lip}\,(a)) + |t' - t''|\,.\end{aligned}$$

Next we set

$$d_1 \equiv \min\{|p' - p''|, r/4\}$$

and we exploit the curve Γ in order to define a curve in $A(t_1)^+$ with endpoints

$$x' + (t' - d_1)a(x')\,, \qquad x'' + (t'' - d_1)a(x'')\,.$$

To do so, we set

$$\Gamma_{d_1}(s) \equiv \Gamma(s) - d_1 a(\psi_{x',x''}(s)) \qquad \forall s \in [0, 1]\,.$$

Since

$$\begin{aligned}&(1 - s)t' + st'' - d_1 \in]-t_2 - d_1, 0[\\ &\quad \subseteq]-(t_1/2) - (r/4), 0[\subseteq]-(t_1/2) - (t_1/8), 0[\subseteq]-t_1, 0[\qquad \forall s \in]0, 1[\,,\end{aligned}$$

we have

$$\Gamma_{d_1}(s) \in A(t_1)^+ \qquad \forall s \in]0, 1[\,.$$

Then we have

$$\Gamma_{d_1}(s) \in [p + R_p^t(\mathbb{B}_{n-1}(0, 3r/4)\times] - 3\delta/4, 3\delta/4[)] + \mathbb{B}_n(0, d_1) \subseteq C(p, R_p, r, \delta)\,,$$

for all $s \in [0, 1]$ and

$$\Gamma_{d_1}(0) = x' + (t' - d_1)a(x')\,, \qquad \Gamma_{d_1}(1) = x'' + (t'' - d_1)a(x'')\,.$$

Moreover,

$$\begin{aligned}\text{length}(\Gamma_{d_1}) &\le \text{length}(\Gamma) + d_1 \text{Lip}(a) \text{length}(\psi_{\xi',\xi''}) \\ &\le (1 + d_1 \text{Lip}(a)) \left[|x' - x''| c_{\gamma_p} (1 + |t_2| \text{Lip}\,(a)) + |t' - t''| \right].\end{aligned}$$

Next we note that

$$\begin{aligned}x' + (t' - sd_1)a(x') \in [p + R_p^t(\mathbb{B}_{n-1}(0, 3r/4) \times] - 3\delta/4, 3\delta/4[)] + \mathbb{B}_n(0, d_1) \\ \subseteq C(p, R_p, r, \delta) \qquad \forall s \in [0, 1]\,,\end{aligned}$$

and similarly

$$x'' + (t'' - sd_1)a(x'') \in C(p, R_p, r, \delta) \qquad \forall s \in [0, 1]\,.$$

Then by the memberships

$$\begin{aligned}t' - sd_1, t'' - sd_1 \in]-t_2 - d_1, 0[\\ \subseteq]-(t_1/2) - (r/4), 0[\subseteq]-(t_1/2) - (t_1/8), 0[\subseteq]-t_1, 0[\qquad \forall s \in [0, 1]\,,\end{aligned}$$

we have

$$x' + (t' - sd_1)a(x') \in A(t_1)^+\,, \quad x'' + (t'' - sd_1)a(x'') \in A(t_1)^+ \qquad \forall s \in [0, 1]\,.$$

We now wish to estimate $|x' - x''|$ and $|t' - t''|$ in terms of $|p' - p''|$. There is no loss of generality in assuming that

$$t'' < t'\,.$$

By assumption, we know that

$$\left| a(x) \cdot \frac{y - x}{|y - x|} \right| \le \vartheta \qquad \forall x, y \in C(p, R_p, r, \delta) \cap (\partial\Omega)\,,\ x \ne y\,.$$

Indeed, $|x - y| \le 2r + \delta < \tau$ for all $x, y \in C(p, R_p, r, \delta) \cap (\partial\Omega)$. Then Lemma 10.3 implies that

$$\begin{aligned}|p' - p''| &= |(x' + t'a(x')) - (x'' + t''a(x''))| \\ &= |(x' - x'') + (t' - t'')a(x') + t''(a(x') - a(x''))| \\ &\ge |(x' - x'') + (t' - t'')a(x')| - |t''|\,|a(x') - a(x'')| \\ &\ge (1 - \vartheta)^{1/2} \sqrt{|x' - x''|^2 + |t' - t''|^2 |a(x')|^2} - t_2 \text{Lip}(a) |x' - x''|\,.\end{aligned}$$

Then by the elementary inequality

$$a_1 + a_2 \leq \sqrt{2}(a_1^2 + a_2^2)^{1/2} \qquad \forall a_1, a_2 \in [0, +\infty[\,,$$

we have

$$\begin{aligned}
&(1-\vartheta)^{1/2}\sqrt{|x'-x''|^2 + |t'-t''|^2|a(x')|^2} - t_2 \mathrm{Lip}(a)|x'-x''| \\
&\quad \geq \frac{(1-\vartheta)^{1/2}}{\sqrt{2}}(|x'-x''| + |t'-t''|) - t_2\mathrm{Lip}(a)|x'-x''| \\
&\quad = \left(\frac{(1-\vartheta)^{1/2}}{\sqrt{2}} - t_2\mathrm{Lip}(a)\right)|x'-x''| + \frac{(1-\vartheta)^{1/2}}{\sqrt{2}}|t'-t''| \\
&\quad \geq \frac{(1-\vartheta)^{1/2}}{2\sqrt{2}}|x'-x''| + \frac{(1-\vartheta)^{1/2}}{\sqrt{2}}|t'-t''|\,.
\end{aligned}$$

Indeed, $t_2 \leq \frac{(1-\vartheta)^{1/2}}{2\sqrt{2}(1+\mathrm{Lip}(a))}$. Then we have

$$|t'-t''| \leq \frac{\sqrt{2}}{(1-\vartheta)^{1/2}}|p'-p''|\,, \qquad |x'-x''| \leq \frac{2\sqrt{2}}{(1-\vartheta)^{1/2}}|p'-p''|\,. \tag{10.13}$$

We now assume that $f \in C^1(\Omega)$ satisfies condition (10.12) and we turn to estimate

$$\begin{aligned}
|f(p') - f(p'')| &= |f(x'+t'a(x')) - f(x''+t''a(x''))| \qquad (10.14)\\
&\leq |f(x'+t'a(x')) - f(x'+(t'-d_1)a(x'))| \\
&\quad + |f(x'+(t'-d_1)a(x')) - f(x''+(t''-d_1)a(x''))| \\
&\quad + |f(x''+(t''-d_1)a(x'')) - f(x''+t''a(x''))| \\
&\leq \int_{t'-d_1}^{t'} \left|a(x')\cdot Df(x'+sa(x'))\right| ds + \int_0^1 |Df(\Gamma_{d_1}(s))|\,|\Gamma'_{d_1}(s)|\,ds \\
&\quad + \int_{t''-d_1}^{t''} \left|a(x'')\cdot Df(x''+sa(x''))\right| ds \\
&\leq \int_{t'-d_1}^{t'} M_{t_1,\alpha}(f)|s|^{\alpha-1}\,ds \\
&\quad + \int_0^1 M_{t_1,\alpha}(f)|(1-s)t' + st'' - d_1|^{\alpha-1}|\Gamma'_{d_1}(s)|\,ds \\
&\quad + \int_{t''-d_1}^{t''} M_{t_1,\alpha}(f)|s|^{\alpha-1}\,ds\,.
\end{aligned}$$

Since

$$(1-s)t' + st'' - d_1 \in]-t_2 - d_1, -d_1[\qquad \forall s \in [0,1]\,,$$

we have

$$\begin{aligned} |f(p') - f(p'')| &\le M_{t_1,\alpha}(f) \int_{|t'|}^{|t'|+d_1} s^{\alpha-1}\,ds \\ &+ \text{length}(\Gamma_{d_1}) M_{t_1,\alpha}(f) d_1^{\alpha-1} + M_{t_1,\alpha}(f) \int_{|t''|}^{|t''|+d_1} s^{\alpha-1}\,ds \\ &\le M_{t_1,\alpha}(f) \left\{ \frac{d_1^\alpha}{\alpha} + |p'-p''| d_1^{\alpha-1}(1 + d_1 \text{Lip}(a)) \right. \\ &\left. \times \left[\frac{2\sqrt{2}}{(1-\vartheta)^{1/2}} c_{\gamma_p}(1 + |t_1|\text{Lip}(a)) + \frac{\sqrt{2}}{(1-\vartheta)^{1/2}} \right] + \frac{d_1^\alpha}{\alpha} \right\}, \end{aligned} \tag{10.15}$$

(cf. (10.13)). Next we observe that

$$\begin{aligned} \frac{|p'-p''|}{d_1^{1-\alpha}} &= \frac{|p'-p''|}{\min^{1-\alpha}\{|p'-p''|, r/4\}} \\ &\le \begin{cases} |p'-p''|^\alpha & \text{if } |p'-p''| \le r/4\,, \\ \frac{|p'-p''|}{(r/4)^{1-\alpha}} \le |p'-p''|^\alpha \frac{\text{diam}(\Omega)^{1-\alpha}}{(r/4)^{1-\alpha}} & \text{if } |p'-p''| \ge r/4\,. \end{cases} \end{aligned}$$

Hence, inequality (10.15) implies the validity of inequality (10.11) and the proof is complete. □

Proposition 10.1 *Let $n \in \mathbb{N} \setminus \{0,1\}$. Let Ω be a bounded open Lipschitz subset of $\mathbb{R}^n$. Let $\vartheta \in]0,1[$. Let $a \in C^{0,1}(\partial\Omega, \mathbb{R}^n)$ be an outer nontangential unit vector field for Ω with parameter ϑ.*

Let $t_1 \in]0,+\infty[$ be as in Lemma 10.2. Let $\alpha \in]0,1]$. Then there exist $B \in]0,+\infty[$, and a compact subset K of Ω such that

$$\sup_{\Omega} |f| + |f : \Omega|_\alpha \le B \max\left\{ \sup_K |f|, \sup_K |Df|, M_{t_1,\alpha}(f) \right\}$$

for all $f \in C^1(\Omega)$ such that

$$M_{t_1,\alpha}(f) \equiv \sup\{|t|^{1-\alpha} |Df(x + ta(x))| : (x,t) \in (\partial\Omega) \times]-t_1, 0[\} < +\infty\,. \tag{10.16}$$

Proof Let $\tau \in]0,+\infty[$ satisfy the condition (ii) of Definition 10.4. By our assumption on Ω, for each point $x \in \partial\Omega$ there exists a coordinate cylinder $C(x, R_x, r_x, \delta_x)$ for Ω around x and if $r_x + \delta_x$ is less than the distance between

$\partial\Omega$ and $\mathbb{R}^n \setminus A(t_1)$, then we have $C(x, R_x, r_x, \delta_x) \subseteq A(t_1)$. Thus we now choose $r_*, \delta_* \in]0, +\infty[$ such that

$$r_* + \delta_* < \operatorname{dist}(\partial\Omega, \mathbb{R}^n \setminus A(t_1)) .$$

Since we plan to invoke Lemma 10.4, we also assume that

$$2\delta_* < \min\{\tau/2, t_1\} .$$

Then by Lemma 10.1, there exist $r \in]0, r_*[$, $\delta \in]0, \delta_*[$, $r < \delta$ such that if $x \in \partial\Omega$, then there exists $R_x \in \mathbb{O}_n(\mathbb{R})$ such that $C(x, R_x, r, \delta)$ is a coordinate cylinder for Ω around x and the corresponding function γ_x satisfies the inequality

$$\sup_{x\in\partial\Omega} \|\gamma_x\|_{C^{0,1}(\overline{\mathbb{B}_{n-1}(0,r)})} < +\infty . \tag{10.17}$$

Since $r + \delta < \operatorname{dist}(\partial\Omega, \mathbb{R}^n \setminus A(t_1))$, we have

$$C(x, R_x, r, \delta) \cap \Omega \subseteq A(t_1)^+ \equiv A(t_1) \cap \Omega \qquad \forall x \in \partial\Omega .$$

Since $\partial\Omega$ is compact, there exists a finite family $\{x^{(j)}\}_{j=1}^m$ of points of $\partial\Omega$ such that

$$\partial\Omega \subseteq \bigcup_{j=1}^m [x^{(j)} + R^t_{x^{(j)}}(\mathbb{B}_{n-1}(0, r/4)\times] - \delta/4, \delta/4[)] ,$$

and we note that the right-hand side is an open neighborhood of $\partial\Omega$. We now set

$$\mu \equiv \min_{j=1,\dots,m} \left\{ r/4, \frac{(1-\vartheta)^{1/2}}{2\sqrt{2}(\operatorname{Lip}(a)+1)}, \frac{t_1}{2}, \right.$$
$$\left. \operatorname{dist}\left(\partial\Omega, \Omega \setminus \bigcup_{j=1}^m [x^{(j)} + R^t_{x^{(j)}}(\mathbb{B}_{n-1}(0, r/4)\times] - \delta/4, \delta/4[)] \right) \right\} ,$$

and we choose $t_2 \in]0, \mu[$. In particular, we have

$$A(t_2)^+ \subseteq A(t_2) \subseteq \{x \in \Omega : \operatorname{dist}(x, \partial\Omega) < \mu\}$$
$$\subseteq \bigcup_{j=1}^m [x^{(j)} + R^t_{x^{(j)}}(\mathbb{B}_{n-1}(0, r/4)\times] - \delta/4, \delta/4[)] .$$

Then Lemma 10.4 implies that there exists $B_j \in]0, +\infty[$ such that

$$|f : [x^{(j)} + R^t_{x^{(j)}}(\mathbb{B}_{n-1}(0, r/4)\times] - \delta/4, \delta/4[)] \cap A(t_2)^+|_\alpha \le B_j M_{t_1,\alpha}(f) , \tag{10.18}$$

for all $j \in \{1, \dots, m\}$ and for all $f \in C^1(\Omega)$ such that $M_{t_1,\alpha}(f) < +\infty$. Now let Ω_1 be an open subset of class C^∞ of Ω such that

$$\overline{\Omega \setminus A(t_2)^+} \subseteq \Omega_1 \subseteq \overline{\Omega_1} \subseteq \Omega\,,$$

(cf. Lemma 10.5 of the Appendix). Since Ω_1 is of class C^∞, we know that there exists $c_\alpha[\Omega_1] \in]0, +\infty[$ such that

$$|f : \overline{\Omega_1}|_\alpha \leq c_\alpha[\Omega_1] \sup\left\{\sup_{\overline{\Omega_1}} |f|, \sup_{\overline{\Omega_1}} |Df|\right\}\,, \tag{10.19}$$

for all $f \in C^1(\overline{\Omega_1})$ (cf. e.g., [La91, §2], reference [DaLaMu19, Ch. 2] with Dalla Riva and Musolino). We now take $f \in C^1(\Omega)$ such that $M_{t_1,\alpha}(f) < +\infty$ and we turn to estimate $|f : \Omega|_\alpha$. To do so, we observe that

$$\overline{\Omega} \subseteq \Omega_1 \cup \bigcup_{j=1}^m \left\{[x^{(j)} + R^t_{x^{(j)}}(\mathbb{B}_{n-1}(0, r/4)\times] - \delta/4, \delta/4[)] \cap A(t_2)\right\}\,. \tag{10.20}$$

Let Λ be a Lebesgue number corresponding to the open cover of $\overline{\Omega}$ in the right-hand side. We can clearly assume that

$$\Lambda < \delta/4\,.$$

If $p', p'' \in \Omega$ and $|p' - p''| \leq \Lambda$, then both p' and p'' belong to at least one of the open sets in the right-hand side of (10.20) and thus inequalities (10.18) and (10.19) imply that

$$|f(p') - f(p'')| \leq \max\left\{\tilde{B} M_{t_1,\alpha}(f), c_\alpha[\Omega_1] \sup_{\overline{\Omega_1}} |f|, c_\alpha[\Omega_1] \sup_{\overline{\Omega_1}} |Df|\right\} |p' - p''|^\alpha\,, \tag{10.21}$$

where

$$\tilde{B} \equiv \max_{j\in\{1,\dots,m\}} B_j\,.$$

In order to estimate $|f(p') - f(p'')|$ in case $|p' - p''| > \Lambda$, we need to estimate $\sup_\Omega |f|$. Indeed,

$$|f(p') - f(p'')| \leq \frac{2 \sup_\Omega |f|}{\Lambda^\alpha} |p' - p''|^\alpha$$

whenever $|p' - p''| > \Lambda$. To do so, we note that

$$x^{(j)} - \frac{t_2}{2} a(x^{(j)}) \in A(t_2)^+ \subseteq \Omega \qquad \forall j \in \{1, \dots, m\}\,,$$

and that our assumptions $t_2 < \mu < r/4$, $r < \delta$ imply that

$$x^{(j)} - \frac{t_2}{2}a(x^{(j)}) \in \mathbb{B}_n(x^{(j)}, r/8) \subseteq x^{(j)} + R^t_{x^{(j)}}(\mathbb{B}_{n-1}(0, r/4)\times]-\delta/4, \delta/4[)$$

for all $j \in \{1, \dots, m\}$. By Lemma 10.5 of the Appendix, there exists an open subset Ω_2 of class C^∞ of Ω such that

$$\overline{\Omega_1} \cup \left\{ x^{(j)} - \frac{t_2}{2}a(x^{(j)}) : \ j \in \{1, \dots, m\} \right\} \subseteq \Omega_2 \subseteq \overline{\Omega_2} \subseteq \Omega \, .$$

If $p \in \Omega \setminus \overline{\Omega_2}$, then $p \in \Omega \setminus \overline{\Omega_1}$ and

$$p \in A(t_2)^+ \subseteq \bigcup_{j=1}^{m} [x^{(j)} + R^t_{x^{(j)}}(\mathbb{B}_{n-1}(0, r/4)\times]-\delta/4, \delta/4[)]$$

and accordingly, there exists $\tilde{j} \in \{1, \dots, m\}$ such that

$$p \in A(t_2)^+ \cap [x^{(\tilde{j})} + R^t_{x^{(\tilde{j})}}(\mathbb{B}_{n-1}(0, r/4)\times]-\delta/4, \delta/4[)] \, .$$

Since both p and $x^{(\tilde{j})} - \frac{t_2}{2}a(x^{(\tilde{j})})$ belong to

$$A(t_2)^+ \cap [x^{(\tilde{j})} + R^t_{x^{(\tilde{j})}}(\mathbb{B}_{n-1}(0, r/4)\times]-\delta/4, \delta/4[)]$$

and $x^{(\tilde{j})} - \frac{t_2}{2}a(x^{(\tilde{j})})$ belongs to Ω_2, we have

$$\begin{aligned} |f(p)| &\leq |f(p) - f(x^{(\tilde{j})} - \frac{t_2}{2}a(x^{(\tilde{j})}))| + |f(x^{(\tilde{j})} - \frac{t_2}{2}a(x^{(\tilde{j})}))| \\ &\leq \tilde{B} M_{t_1,\alpha}(f) \left| p - x^{(\tilde{j})} - \frac{t_2}{2}a(x^{(\tilde{j})}) \right|^\alpha + \sup_{\overline{\Omega_2}} |f| \, . \end{aligned}$$

Since both p and $x^{(\tilde{j})} - \frac{t_2}{2}a(x^{(\tilde{j})})$ belong to

$$x^{(\tilde{j})} + R^t_{x^{(\tilde{j})}}(\mathbb{B}_{n-1}(0, r/4)\times]-\delta/4, \delta/4[)$$

that has a diameter less than or equal to $2(r/4) + 2(\delta/4) < \delta$, we have

$$|f(p)| \leq \tilde{B} M_{t_1,\alpha}(f)\delta^\alpha + \sup_{\overline{\Omega_2}} |f| \, .$$

If instead $p \in \overline{\Omega_2}$, we certainly have $|f(p)| \leq \sup_{\overline{\Omega_2}} |f|$. Hence,

$$\sup_{\Omega} |f| \leq \tilde{B} M_{t_1,\alpha}(f)\delta^{\alpha} + \sup_{\overline{\Omega_2}} |f|$$

and

$$\begin{aligned} &|f : \Omega|_{\alpha} \\ &\leq \max\left\{ \max\left\{ \tilde{B} M_{t_1,\alpha}(f), c_{\alpha}[\Omega_1] \sup_{\overline{\Omega_1}} |f|, c_{\alpha}[\Omega_1] \sup_{\overline{\Omega_1}} |Df| \right\}, \frac{2}{\Lambda^{\alpha}} \sup_{\Omega} |f| \right\}. \end{aligned}$$

Then by taking $K = \overline{\Omega_2}$, we conclude that B as in the statement does exist. □

Appendix

Proof of Lemma 10.1 If $x \in \partial\Omega$, then there exists a coordinate cylinder $C(x, R_x, r_x, \delta_x)$ around x and a corresponding function $\tilde{\gamma}_x$ which represents $\partial\Omega$ in $C(x, R_x, r_x, \delta_x)$. Now let

$$\delta'_x \in]0, \min\{\delta_x, \delta_*\}[\,.$$

Since $\tilde{\gamma}_x$ is continuous, there exists $r'_x \in]0, +\infty[$ such that

$$r'_x < \min\{r_x, r_*\}\,, \qquad |\tilde{\gamma}_x(\eta)| < \frac{1}{2}\delta'_x \qquad \forall \eta \in \mathbb{B}_{n-1}(0, r'_x)\,.$$

Then $C(x, R_x, r'_x, \delta'_x)$ is a coordinate cylinder for Ω around x and the restriction $\tilde{\gamma}_{x|\mathbb{B}_{n-1}(0,r'_x)}$ represents $\partial\Omega$ in $C(x, R_x, r'_x, \delta'_x)$. By definition of $C^{0,\alpha}$-norm, we have

$$\|\tilde{\gamma}_{x|\mathbb{B}_{n-1}(0,r'_x)}\|_{C^{0,\alpha}(\overline{\mathbb{B}_{n-1}(0,r'_x)})} \leq \|\tilde{\gamma}_{x|\mathbb{B}_{n-1}(0,r_x)}\|_{C^{0,\alpha}(\overline{\mathbb{B}_{n-1}(0,r_x)})}\,.$$

Since

$$\left\{x + R^t_x\left(\mathbb{B}_{n-1}(0, r'_x/4)\times]-\delta'_x, \delta'_x[\right)\right\}_{x\in\partial\Omega}$$

is an open cover of $\partial\Omega$ and $\partial\Omega$ is compact there exists a finite family $\{x^{(j)}\}_{j=1}^{k}$ of points of $\partial\Omega$ such that

$$\partial\Omega \subseteq \bigcup_{j=1}^{k}\left[x^{(j)} + R^t_{x^{(j)}}\left(\mathbb{B}_{n-1}(0, r'_{x^{(j)}}/4)\times]-\delta'_{x^{(j)}}, \delta'_{x^{(j)}}[\right)\right].$$

To shorten our notation, we find convenient to set

$$R_j \equiv R_{x^{(j)}}\,, \qquad r'_j \equiv r'_{x^{(j)}} \qquad \delta'_j \equiv \delta'_{x^{(j)}}\,, \qquad \gamma_j \equiv \tilde{\gamma}_{x^{(j)}|\mathbb{B}_{n-1}(0,r'_{x^{(j)}})}$$

for all $j \in \{1,\dots,k\}$. Next we choose

$$\delta \in]0, \frac{1}{4}\min_{j\in\{1,\dots,k\}} \delta'_j[\,.$$

By the uniform continuity of the functions γ_j, there exists

$$r \in]0, \frac{1}{4}\min_{j\in\{1,\dots,k\}} r'_j[$$

such that

$$|\gamma_j(\eta_1) - \gamma_j(\eta_2)| < \delta/2 \qquad \text{whenever } \eta_1, \eta_2 \in \mathbb{B}_{n-1}(0, r_j)\,,\ |\eta_1 - \eta_2| < r \tag{10.22}$$

for all $j \in \{1,\dots,k\}$. Next we fix an arbitrary $x \in \partial\Omega$ and we define a coordinate cylinder for Ω around x. Let $j \in \{1,\dots,k\}$ be such that

$$x \in x^{(j)} + R_j^t\left(\mathbb{B}_{n-1}(0, r'_j/4)\times]-\delta'_j, \delta'_j[\right)\,.$$

Then there exists $\eta_x \in \mathbb{B}_{n-1}(0, r'_j/4)$ such that

$$x = x^{(j)} + R_j^t(\eta_x, \gamma_j(\eta_x))^t\,.$$

Since $r < r'_j/4$, we have

$$\mathbb{B}_{n-1}(\eta_x, r) \subseteq \mathbb{B}_{n-1}(0, r'_j/2)\,.$$

Since $\delta < \delta'_j/4$, we have

$$]\gamma_j(\eta_x) - \delta, \gamma_j(\eta_x) + \delta[\subseteq]-(\delta'_j/2) - \delta, (\delta'_j/2) + \delta[\subseteq]-(3\delta'_j/4), (3\delta'_j/4)[\,.$$

Next we set

$$\gamma_x(\eta) \equiv \gamma_j(\eta_x + \eta) - \gamma_j(\eta_x) \qquad \forall \eta \in \mathbb{B}_{n-1}(0, r)\,,$$

and we claim that

$$C(x, R_j, r, \delta)$$

is a coordinate cylinder for Ω around x and that γ_x represents $\partial\Omega$ in $C(x, R_j, r, \delta)$. To do so, we observe that

$$
\begin{aligned}
& R_j(\Omega - x) \cap (\mathbb{B}_{n-1}(0, r)\times]-\delta, \delta[) \qquad (10.23)\\
&= R_j\left(\Omega - x^{(j)} - R_j^t(\eta_x, \gamma_j(\eta_x))^t\right) \cap (\mathbb{B}_{n-1}(0, r)\times]-\delta, \delta[)\\
&= \left(R_j(\Omega - x^{(j)}) - (\eta_x, \gamma_j(\eta_x))^t\right)\\
&\qquad\qquad \cap \left((\mathbb{B}_{n-1}(\eta_x, r)\times]-\delta + \gamma_j(\eta_x), \gamma_j(\eta_x) + \delta[) - (\eta_x, \gamma_j(\eta_x))^t\right)\\
&= R_j(\Omega - x^{(j)}) \cap \left(\mathbb{B}_{n-1}(\eta_x, r)\times]-\delta + \gamma_j(\eta_x), \gamma_j(\eta_x) + \delta[\right) - (\eta_x, \gamma_j(\eta_x))^t\\
&= R_j(\Omega - x^{(j)}) \cap \left(\mathbb{B}_{n-1}(0, r_j')\times]-\delta_j', \delta_j'[\right)\\
&\qquad\qquad \cap \left(\mathbb{B}_{n-1}(\eta_x, r)\times]-\delta + \gamma_j(\eta_x), \gamma_j(\eta_x) + \delta[\right) - (\eta_x, \gamma_j(\eta_x))\\
&= (\text{hypograph}_s(\gamma_j))\\
&\qquad\qquad \cap \left(\mathbb{B}_{n-1}(\eta_x, r)\times]-\delta + \gamma_j(\eta_x), \gamma_j(\eta_x) + \delta[\right) - (\eta_x, \gamma_j(\eta_x))\,.
\end{aligned}
$$

Next we observe that

$$
\begin{aligned}
& \text{hypograph}_s(\gamma_x) \qquad (10.24)\\
&= \{(\eta, y) \in \mathbb{B}_{n-1}(0, r)\times]-\delta, \delta[:\ y < \gamma_x(\eta)\}\\
&= \left\{(\eta, y) \in \mathbb{B}_{n-1}(0, r)\times]-\delta, \delta[:\ y + \gamma_j(\eta_x) < \gamma_j(\eta_x + \eta)\right\}\\
&= \left\{(\eta, y) \in \mathbb{B}_{n-1}(0, r)\times]-\delta, \delta[:\ (\eta + \eta_x, y + \gamma_j(\eta_x)) \in \text{hypograph}_s(\gamma_j)\right\}\\
&= \left\{(\eta, y) \in \mathbb{B}_{n-1}(0, r)\times]-\delta, \delta[:\ (\eta, y) \in \text{hypograph}_s(\gamma_j) - (\eta_x, \gamma_j(\eta_x))\right\}\\
&= \left\{(\eta, y) \in \left(\mathbb{B}_{n-1}(\eta_x, r)\times]-\delta + \gamma_j(\eta_x), \gamma_j(\eta_x) + \delta, [\right) - (\eta_x, \gamma_j(\eta_x)) :\right.\\
&\qquad\qquad \left.(\eta, y) \in \text{hypograph}_s(\gamma_j) - (\eta_x, \gamma_j(\eta_x))\right\}\\
&= \left[\text{hypograph}_s(\gamma_j)\right.\\
&\qquad\qquad \left.\cap \left(\mathbb{B}_{n-1}(\eta_x, r)\times]-\delta + \gamma_j(\eta_x), \gamma_j(\eta_x) + \delta, [\right)\right] - (\eta_x, \gamma_j(\eta_x))\,.
\end{aligned}
$$

Then by combining (10.23) and (10.24) we obtain

$$
R_j(\Omega - x) \cap (\mathbb{B}_{n-1}(0, r)\times]-\delta, \delta[) = \text{hypograph}_s(\gamma_x)\,.
$$

By the definition of γ_x and by inequality (10.22), we have

$$\gamma_x(0) = 0\,, \qquad |\gamma_x(\eta)| < \delta/2 \qquad \forall \eta \in \mathbb{B}_{n-1}(0,r)\,.$$

Moreover, γ_x has the same regularity of γ_j and if $\alpha > 0$, we have

$$|\gamma_x : \overline{\mathbb{B}_{n-1}(0,r)}|_\alpha \leq \|\gamma_j\|_{C^{0,\alpha}(\overline{\mathbb{B}_{n-1}(0,r_j')})} \leq \sup_{l=1,\dots,k} \|\gamma_l\|_{C^{0,\alpha}(\overline{\mathbb{B}_{n-1}(0,r_l')})} < +\infty\,,$$

and thus the proof is complete. □

Lemma 10.5 *Let Ω be an open subset of $\mathbb{R}^n$. Let K be a compact subset of Ω. Then there exists an open bounded subset Ω_1 of Ω of class C^∞ such that*

$$K \subseteq \Omega_1 \subseteq \overline{\Omega_1} \subseteq \Omega\,.$$

If we further assume that K is connected, then we can take Ω_1 to be connected.

For a proof, we refer to [DaLaMu19, Ch. 2], which contains a proof due to G. De Marco.[1]

Acknowledgments The author is indebted to Prof H. Aikawa for pointing out the validity of inequality (10.2) in uniform domains (see [Ai19]), to Prof. A. Cialdea for pointing out the references [Fi55], [Fi84], and to Prof. M. Dalla Riva and to Dr. P. Musolino for a number of suggestions on the manuscript.

The author acknowledges the support of "Gruppo Nazionale per l'Analisi Matematica, la Probabilità e le loro Applicazioni," of "INdAM" and of the project "Singular perturbation problems for the heat equation in a perforated domain:BIRD168373/16" of the University of Padova.

References

[AgDoNi59] Agmon, S., Douglis, A., Nirenberg, L.: Estimates near the boundary for solutions of elliptic partial differential equations satisfying general boundary conditions I. Commun. Pure Appl. Math. **12**, 623–727 (1959)

[Ai19] Aikawa, H.: Personal communication (2019)

[Bu98] Burenkov, V.I.: Sobolev spaces on domains. Teubner-Texte zur Mathematik [Teubner Texts in Mathematics], 137. B. G. Teubner Verlagsgesellschaft mbH, Stuttgart (1998)

[DaLaMu19] Dalla Riva, M., Lanza de Cristoforis, M., Musolino, P.: A Functional Analytic Approach to singularly perturbed boundary value problems. Book draft (2019)

[De85] Deimling, K.: Nonlinear Functional Analysis. Springer, Berlin (1985)

[Fi55] Fichera, G.: Alcuni recenti sviluppi della teoria dei problemi al contorno per le equazioni alle derivate parziali lineari (Italian). In: Convegno Internazionale sulle Equazioni Lineari alle Derivate Parziali, Trieste, 1954, pp. 174–227. Edizioni Cremonese, Roma (1955)

[1] Professor at the University of Padova, Italy.

[Fi84] Fichera, G.: Boundary value problems of elasticity with unilateral constraints. In: Mechanics of Solids. Encyclopedia of Physics, vol. II, pp. 391–424. Springer, Berlin (1984)

[GiTr83] Gilbarg, D., Trudinger, N.S.: Elliptic Partial Differential Equations of Second Order. Springer, Berlin (1983)

[Gr85] Grisvard, P.: Elliptic Problems in Nonsmooth Domains. Pitman (Advanced Publishing Program), Boston (1985)

[La91] Lanza de Cristoforis, M.: Properties and pathologies of the composition and inversion operators in Schauder spaces. Rend. Accad. Naz. Sci. XL Mem. Mat. (5) **15**, 93–109 (1991)

[La19] Lanza de Cristoforis, M.: Lipschitz sets and outer nontangential vector fields. Typewritten manuscript (2020)

[LaRo08] Lanza de Cristoforis, M., Rossi, L.: Real analytic dependence of simple and double layer potentials for the Helmholtz equation upon perturbation of the support and of the density. In: Kilbas, A.A., Rogosin, S.V. (eds.) Analytic Methods of Analysis and Differential Equations, AMADE 2006, pp. 193–220. Cambridge Scientific Publishers, Cambridge (2008)

[Mi65] Miranda, C.: Sulle proprietà di regolarità di certe trasformazioni integrali. Atti Accad. Naz. Lincei Mem. Cl. Sci. Fis. Mat. Natur. Sez I **7**, 303–336 (1965)

[Tr87] Troianiello, G.M.: Elliptic Differential Equations and Obstacle Problems. Plenum Press, New York (1987)

[Zh05] Zhu, K.: Spaces of Holomorphic Functions in the Unit Ball. Springer, New York (2005)

Chapter 11
Two-Phase Three-Component Flow in Porous Media: Mathematical Modeling of Dispersion-Free Pressure Behavior

Luara K. S. Sousa, Luana C. M. Cantagesso, Adolfo P. Pires, and Alvaro M. M. Peres

11.1 Introduction

Expected ultimate recovery from oil reservoirs ranges between 10–40% under primary depletion. Enhanced oil recovery (EOR) methods are designed to increase the final recovery of a particular field by the injection of specific fluids into the reservoir. Most of the EOR techniques can be classified into three major groups according to their most important physical-chemical mechanism: Thermal (hot waterflooding, steam drive, or in situ combustion), Chemical (alkaline flooding, surfactant flooding, or micellar polymer flooding), or Solvent (miscible or partially miscible carbon dioxide, hydrocarbon, nitrogen, or natural gas injection) [La89].

Regardless of the chosen method, an EOR project requires significant additional investments. Thus, selecting the right EOR method for a given field is an important step for the project's technical and financial success. Given the large amount of available EOR techniques, it is not feasible running full field compositional 3-D numerical simulation for each EOR method. It is necessary to perform a previous screening to select the most promising techniques based on semi-analytical solutions that are simple enough to generate fast results but accurate to model the incremental oil production and injection pressures required. This work describes a general procedure for obtaining these solutions by limiting the geometry to a single dimension and representing the fluid by at most three components. To exemplify the proposed procedure, we detail the solution for three usual oil recovery techniques such as waterflooding, polymer flooding, and miscible flooding.

Next section presents the mathematical model for a general case, its main hypothesis and solution procedure; followed by a detailed solution for each of the selected applications. Next, some concluding remarks.

L. K. S. Sousa · L. C. M. Cantagesso · A. P. Pires (✉) · A. M. M. Peres
Universidade Estadual do Norte Fluminense, Macaé, RJ, Brazil

C. Constanda (ed.), *Computational and Analytic Methods in Science and Engineering*, https://doi.org/10.1007/978-3-030-48186-5_11

11.2 Physical and Mathematical Model

In this work we solve the one-dimensional two-phase three-component flow in an infinite porous media. We consider oil displacement by the injection of an enhanced oil recovery fluid at constant rate q_{inj}. The porous media is saturated with a liquid (oil) phase at the initial state. After the beginning of the injection, three different regions appear: a single-phase injected fluid region followed by a two-phase region (displaced and displacing phases) where mass transfer may take place, and a single-phase liquid region (Fig. 11.1).

The following hypothesis will be adopted in our mathematical model:

- Isothermal flow.
- Homogeneous porous media.
- Incompressible rock and fluid system for injection and two-phase regions.
- Slightly compressible rock and fluid for original liquid phase region.
- No chemical reactions.
- Gravity, dispersion, and capillary effects are negligible.
- Darcy's law is valid.

Under these assumptions, the i-th component mass conservation is given by [La89]:

$$\frac{\partial \left(\varphi(\sum_{j=1}^{n_p}(\rho_j S_j \omega_{ij})) + (1-\varphi)\rho_s \omega_{is}\right)}{\partial t} + \frac{\partial \left(\sum_{j=1}^{n_p} \rho_j u_j \omega_{ij}\right)}{\partial x} = 0 \qquad (11.1)$$

where φ is the porosity, ρ_j is the density of phase j, S_j is the saturation of phase j, ω_{ij} is the mass fraction of component i in phase j, and u_j is the apparent velocity of phase j. Subscript s refers to the solid phase.

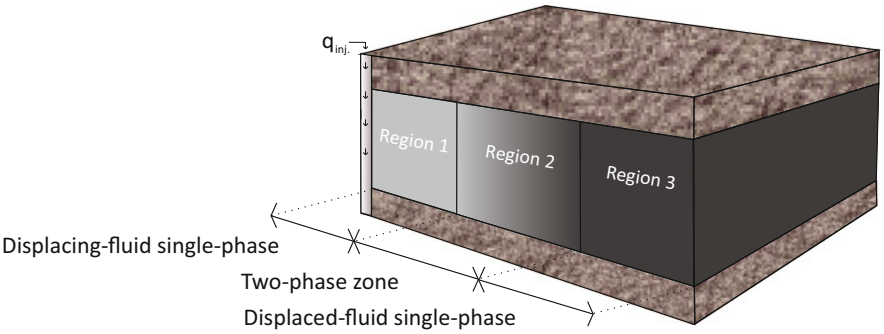


Fig. 11.1 Representation of the three regions

The fractional flow function of a phase (f_j) is defined as

$$f_j = \frac{u_j}{u_T} \Leftrightarrow u_j = f_j u_T \tag{11.2}$$

and total velocity u_T is given by the multicomponent multiphase horizontal Darcy's law:

$$u_T = -k \left[\sum_{j=1}^{N_p} \frac{k_{rj}(S_o)}{\mu_j(\vec{C})} \right] \frac{\partial p}{\partial x} \tag{11.3}$$

where p is pressure, k denotes absolute permeability, k_{rj} is phase j relative permeability, μ_j is the viscosity of phase j, S_o is the oil saturation, and $\vec{C}$ is the concentration vector.

The term inside the brackets in Eq. (11.3) represents the total mobility (λ_T), so this equation can be written as:

$$u_T = -\lambda_T(S_o, \vec{C}) k \frac{\partial p}{\partial x} \tag{11.4}$$

The problem is solved in two steps. First, the Riemann problem is solved for saturation and concentration. Then, Darcy's law is integrated over the spatial domain for pressure determination.

11.3 Applications

In this section the solution procedure described previously is applied to three different oil recovery problems: waterflooding, the most used technique of oil production; polymer flooding accounting for adsorption, a chemical method of enhanced oil recovery, and three-component miscible flooding with mass transfer between phases.

11.3.1 Waterflooding

Water injection in reservoirs is the most popular method of oil recovery, it combines low cost with ease of handling. Besides the previously presented hypothesis, in this case we consider that there is only one component in each phase (oil component in oil phase and water component in water phase) and the phases are immiscible, i.e., no mass transfer occurs.

Table 11.1 Characteristic waves for waterflooding

Eigenvalue	Shock speed
$\lambda_{S_w} = \frac{u_T}{\phi}\frac{\partial f_w}{\partial S_w}$	$D_{BL} = \frac{u_T}{\phi}\frac{[f_w]}{[S_w]}$

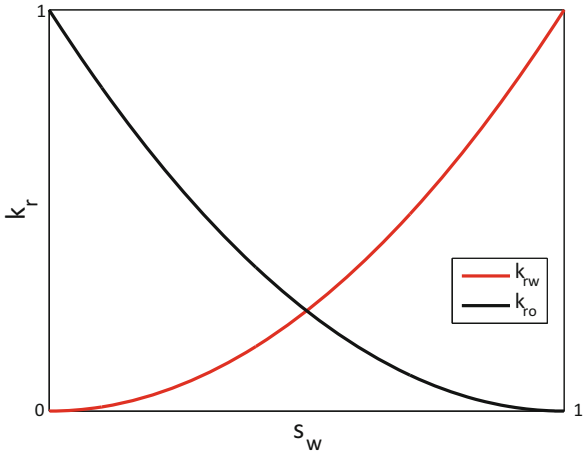


Fig. 11.2 Water-oil relative permeability

Following the described assumptions, Eq. (11.1) takes the following form for each phase:

$$\varphi \frac{\partial S_j}{\partial t} + u_T \frac{\partial f_j\left(S_j\right)}{\partial x} = 0 \tag{11.5}$$

where $j = o, w$. Since the sum of saturations equals one, only one of the equations 11.5 needs to be solved. The solution of this problem is composed by rarefaction waves and shocks (Table 11.1).

Using Corey's model [CoEtAl56] for relative permeability (Fig. 11.2), the fractional flow function for this example is presented in Fig. 11.3.

For the following initial and boundary conditions

$$\left\{ S_w\left(x, t=0\right) = S_w^{(I)} \quad S_w\left(x=0, t\right) = S_w^{(J)} \right. \tag{11.6}$$

the solution path is depicted in Fig. 11.3. The Riemann solution is composed by a saturation rarefaction wave from injection point (J) to the water saturation front (F) and a Buckley–Leverett saturation shock type from point (F) to the initial condition (I).

The water saturation versus distance profile is shown in Fig. 11.4. For this waterflooding example, the first region (single-phase water) does not appear due to the fractional flow curve value close to the injection point (J) (see Fig. 11.3). The saturation front position at any time is given by $x = D_{BL}t$, where D_{BL} is the Buckley–Leverett type shock wave speed. This plot also shows the flow rate q curve

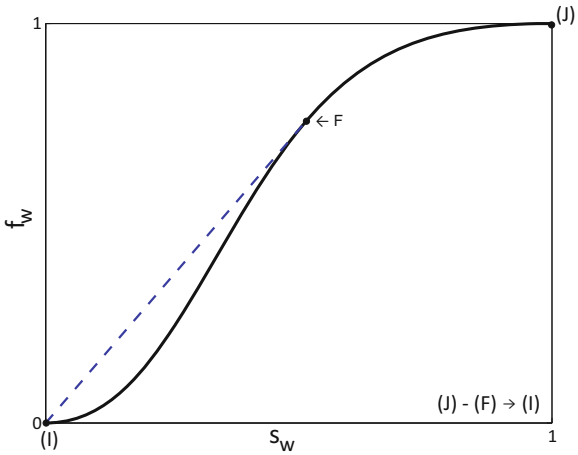


Fig. 11.3 Fractional flow function for waterflooding including solution path

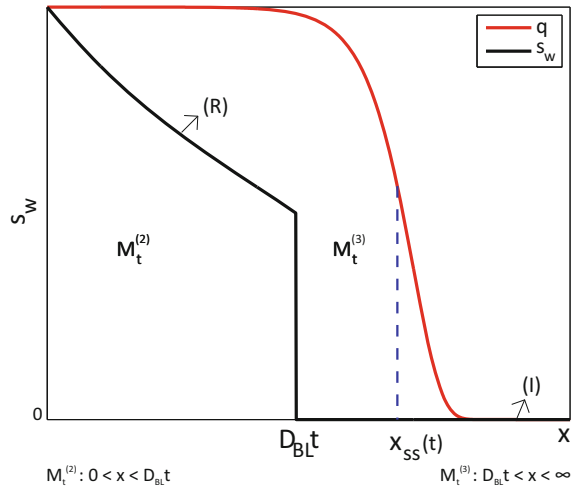


Fig. 11.4 Water saturation profile and flow rate

and the steady state front (x_{ss}). Note that as $D_{BL}t < x_{ss}(t)$, the two-phase region is within the steady-state region [Th97].

The pressure behavior at the inlet point can be calculated through the integration of Darcy's law. For the particular case of waterflooding, we find the following expression:

$$p_w\,(x=0,t) - p_i = \frac{q_{inj}}{\lambda_T^{(3)} kA} \int_0^{D_{BL}t} \left(\frac{\lambda_T^{(3)}}{\lambda_T^{(2)}\left(x',t\right)} - 1 \right) dx' + \frac{1}{\lambda_T^{(3)} kA} \int_0^{\infty} q_T\left(x',t\right) dx' \tag{11.7}$$

where A is the reservoir cross sectional area, q_T is the total flow rate, and the superscript in brackets denotes the region where total mobility was calculated.

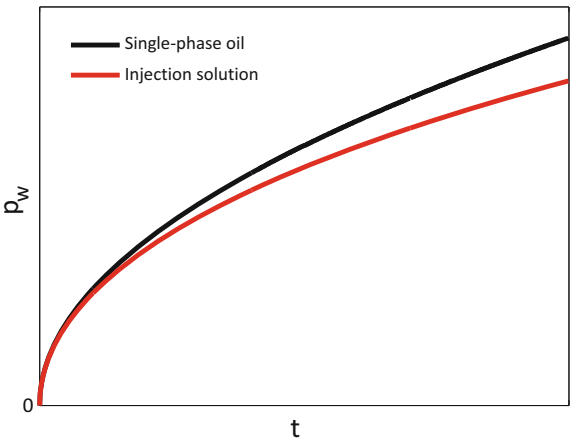


Fig. 11.5 Pressure behavior at injection point

Figure 11.5 presents the pressure behavior at the porous media injection point together with the single-phase oil solution for the same volumetric flow rate. Note that the curves separate with time, when two-phase region grows in porous media.

11.3.2 Polymer Flooding

Injection of water containing dissolved chemical components is classified as chemical enhanced oil recovery. Among others, polymers are one of the most used chemical additives, because it increases water viscosity and, as a consequence, decreases water mobility, enhancing areal sweep [La89].

Oil displacement by water containing dissolved polymers is modeled by two hyperbolic equations, representing water volume conservation and polymer mass conservation:

$$\begin{cases} \varphi \frac{\partial S_w}{\partial t} + u_T \frac{\partial f_w(S_w,\omega)}{\partial x} = 0 \\ \varphi \frac{\partial (\omega S_w + ((1-\varphi)/\varphi)(\rho_s/\rho_w)\omega_s)}{\partial t} + u_T \frac{\partial (\omega f_w(S_w,\omega))}{\partial x} = 0 \end{cases} \tag{11.8}$$

where ω is the polymer concentration in water phase and ω_s is the amount adsorbed. For the sake of simplicity, we recast the adsorbed concentration as:

$$a\,(\omega) = \frac{1-\varphi}{\varphi} \frac{\rho_s}{\rho_w} \omega_s \tag{11.9}$$

and system (11.8) becomes

$$\begin{cases} \varphi \frac{\partial S_w}{\partial t} + u_T \frac{\partial f_w(S_w,\omega)}{\partial x} = 0 \\ \varphi \frac{\partial (\omega S_w + a(\omega))}{\partial t} + u_T \frac{\partial (\omega f_w(S_w,\omega))}{\partial x} = 0 \end{cases} \tag{11.10}$$

Note that for polymer flooding, the fractional flow function depends on water saturation and polymer concentration in water phase. In order to close the system (11.10), we need a relation for the thermodynamic equilibrium between the polymer concentration in water and solid phase, the so-called adsorption isotherm. In this work, we chose Langmuir isotherm (Fig. 11.6) for this purpose:

$$a(\omega) = \frac{\gamma_1 \omega}{1 + \gamma_2 \omega} \tag{11.11}$$

where γ_1 and γ_2 are empirical constants.

Table 11.2 shows the characteristic waves for this problem.

Figure 11.7 presents the solution path in $(S_w \times f_w)$ plane along the two fractional flow functions, one calculated with the initial conditions (f_w^I) and the other with injection (boundary) conditions (f_w^J):

$$\begin{cases} S_w(x, t=0) = S_w^{(I)} & \omega(x, t=0) = \omega^{(I)} \\ S_w(x=0, t) = S_w^{(J)} & \omega(x=0, t) = \omega^{(J)} \end{cases} \tag{11.12}$$

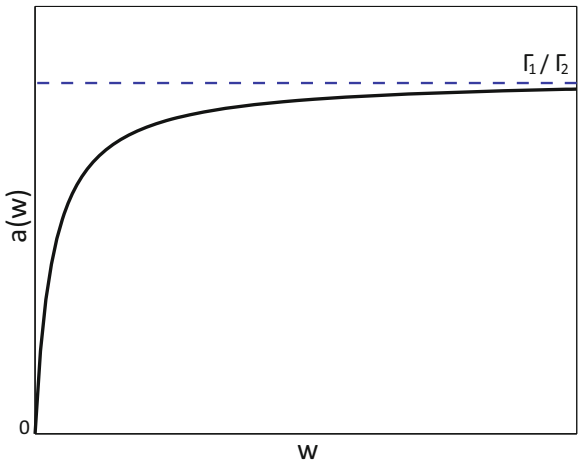


Fig. 11.6 Langmuir adsorption isotherm

Table 11.2 Characteristic waves for polymer flooding

Eigenvalues	Shock speeds
$\lambda_{S_w} = \frac{u_T}{\phi} \frac{\partial f_w}{\partial S_w}$	$D_{BL} = \frac{u_T}{\phi} \frac{[f_w]}{[S_w]}$
$\lambda_c = \frac{u_T}{\phi} \frac{f_w}{S_w + \frac{da}{d\omega}}$	$D_c = \frac{u_T}{\phi} \frac{[f_w]}{[S_w] + \frac{[a]}{[w]}}$

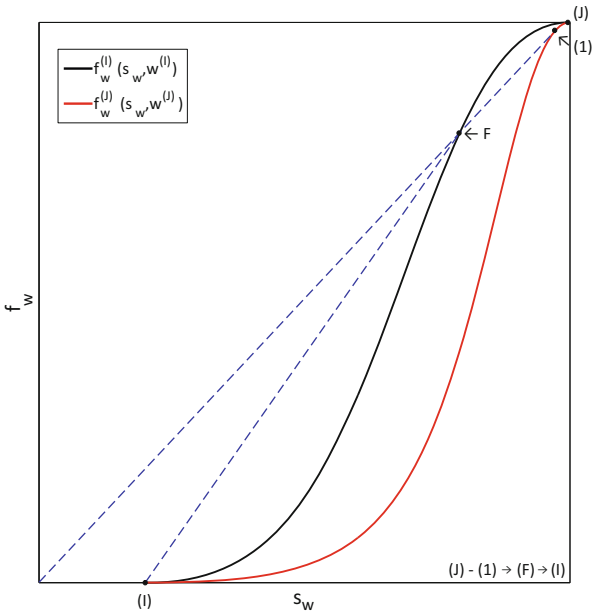


Fig. 11.7 Fractional flow for polymer flooding including solution path

Figure 11.8 zooms the high water saturation part of the solution. It starts with a saturation rarefaction wave at polymer injection concentration between points (J) and (1) followed by a jump from (1) to (F) (concentration transition). A Buckley–Leverett type shock connects points (F) and (I) after a constant state.

Water saturation and polymer concentration profiles are shown in Fig. 11.9. The two-phase region is split in two sub-regions, separated by a concentration-saturation shock at $x = D_c t$. Region $2a$ is composed by a saturation rarefaction wave with constant polymer concentration (ω^J), whereas in region $2b$ both saturation and concentration are constant (S_w^F and ω^I). At $x = D_{BL} t$ water saturation jumps from (F) to (I) (initial water saturation).

Pressure at inlet is given by

$$p_w(x=0,t) - p_i = \frac{q_{inj}}{\lambda_T^{(3)} kA}\left[\int_0^{D_c t}\left(\frac{\lambda_T^{(3)}}{\lambda_T^{(2a)}\left(x',t\right)} - 1\right)dx' + \int_{D_c t}^{D_{BL} t}\left(\frac{\lambda_T^{(3)}}{\lambda_T^{(2b)}} - 1\right)dx'\right] + \frac{1}{\lambda_T^{(3)} kA}\int_0^{\infty} q_T\left(x',t\right)dx' \qquad (11.13)$$

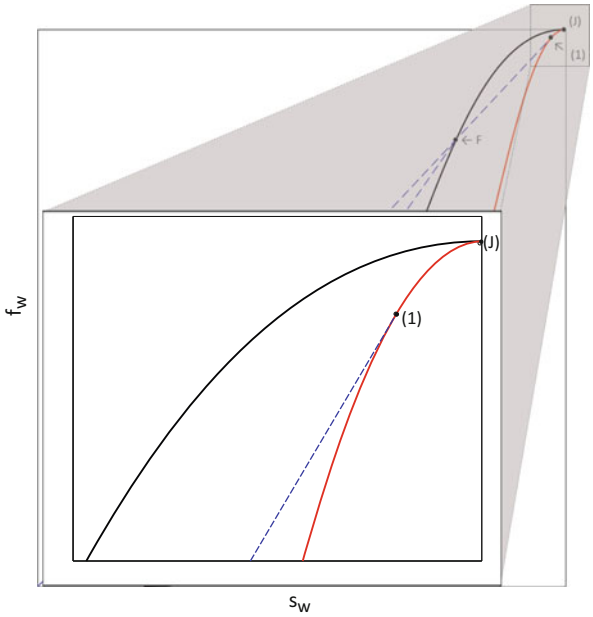


Fig. 11.8 Zoom of high water saturation region (polymer flooding)

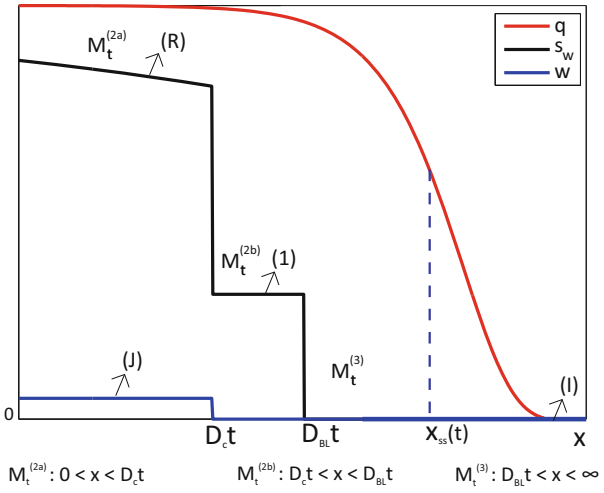


Fig. 11.9 Saturation and polymer concentration profiles for polymer flooding and flow rate

11.3.3 Miscible Flooding

In this part we present the solution for the miscible flooding problem, an enhanced oil recovery technique that is more suitable for intermediate density reservoir fluids. Recently, due to environmental concerns, carbon dioxide has become the most used injection fluid.

For the case of three-component two-phase (gas and liquid) miscible flow, the conservation law for each component is given by

$$\frac{\partial \left(\varphi(\rho_l S_l \omega_{il} + \rho_g S_g \omega_{ig})\right)}{\partial t} + \frac{\partial \left(\rho_l u_l \omega_{il} + \rho_g u_g \omega_{ig}\right)}{\partial x} = 0 \tag{11.14}$$

for $i = 1, 2, 3$.

For this problem, we will also consider that Amagat's law [PrEtAl86] is valid and that the pure component density is the same for all phases. From Amagat's law we find:

$$\rho_j \omega_{ij} = \rho_i c_{ij} \tag{11.15}$$

where c_{ij} is the volumetric concentration of component i in phase j and ρ_i is the pure component i density at system pressure and temperature. Applying Amagat's law (Eq. (11.15)) and the constant pure component density hypothesis, Eq. (11.14) becomes

$$\varphi \frac{\partial \left(S_l c_{il} + S_g c_{ig}\right)}{\partial t} + u_T \frac{\partial \left(f_l c_{il} + f_g c_{ig}\right)}{\partial x} = 0 \tag{11.16}$$

We define total concentration (C_i) and total flow (F_i) variables for component i as:

$$C_i = \sum_{j=1}^{n_p} S_j c_{ij} \tag{11.17}$$

$$F_i = \sum_{j=1}^{n_p} f_j c_{ij} \tag{11.18}$$

Applying Eqs. (11.17) and (11.18) in Eq. (11.16), the following hyperbolic system of equations is found:

$$\begin{cases} \varphi \dfrac{\partial C_2}{\partial t} + u_T \dfrac{\partial F_2}{\partial x} = 0 \\ \varphi \dfrac{\partial C_3}{\partial t} + u_T \dfrac{\partial F_3}{\partial x} = 0 \end{cases} \tag{11.19}$$

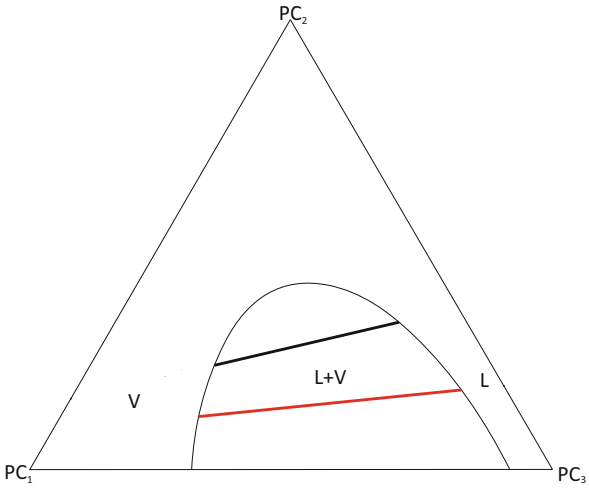


Fig. 11.10 Ternary diagram

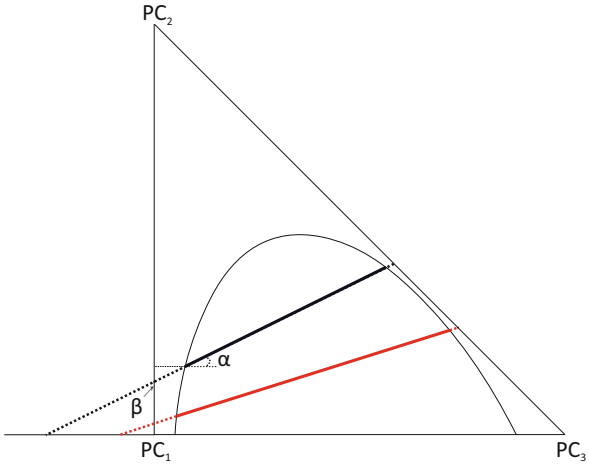


Fig. 11.11 Ternary diagram in Cartesian coordinates

The solution of the hyperbolic system (11.19) depends on the phase equilibrium conditions at system pressure and temperature. For a three-component fluid, the thermodynamic equilibrium at a fixed pressure and temperature is represented by a ternary diagram (Fig. 11.10) [Pi05], where L and V denote the vapor (gas) and liquid (oil) phases, respectively. If the overall composition of a fluid lays inside the two-phase envelop ($L + V$ region), the vapor and liquid equilibrium compositions are determined by the intercepts of the tie lines with the binodal curve [Or07]. The three-component two-phase equilibrium can also be presented in Cartesian coordinates (Fig. 11.11). In this case, the tie lines can be parameterized by two thermodynamic geometric variables α and β [Be93], given by

$$\alpha = \frac{c_{2l} - c_{2g}}{c_{3l} - c_{3g}}, \tag{11.20}$$

$$\beta = c_{2g} - \alpha c_{3g} \tag{11.21}$$

The variable α represents the tie line slope, whereas the β value is the intercept of the tie line with vertical axis. Using an equation of state, one may obtain several tie lines for a given fluid. Each tie line yields a pair (α,β), and a relationship between α and β can be built ($\alpha = \alpha(\beta)$).

In terms of the above defined geometric variables, system (11.19) becomes:

$$\begin{cases} \varphi \dfrac{\partial C}{\partial t} + u_T \dfrac{\partial F}{\partial x} = 0 \\ \varphi \dfrac{\partial (\alpha C + \beta)}{\partial t} + u_T \dfrac{\partial (\alpha F + \beta)}{\partial x} = 0 \end{cases} \tag{11.22}$$

where the simplified notation $C = C_3$ and $F = F_3$ is adopted.

The characteristic waves of this problem are found in Table 11.3.

The initial and boundary conditions for this problem are as follows:

$$\begin{cases} C\,(x, t = 0) = C^{(I)} & \beta\,(x, t = 0) = \beta^{(I)} \\ C\,(x = 0, t) = C^{(J)} & \beta\,(x = 0, t) = \beta^{(J)} \end{cases} \tag{11.23}$$

The top part of Fig. 11.12 shows the binodal curve and the tie lines of the initial and injected fluid compositions. The lower portion of Fig. 11.12 presents the solution of this problem in $F \times C$ space. The solution path is given by : $(J) \rightarrow (1) - (2) \rightarrow (3) \rightarrow (I)$, where $\rightarrow$ denotes a shock wave and $-$ indicates a rarefaction wave. The solution begins at injection conditions (J), which corresponds to single-phase gas (region 1), connected to point (1) in the two-phase region (region 2) through a concentration shock. From point (1) there is a concentration rarefaction wave up to point (2). Next, there is a concentration-β shock linking points (2) and (3), which is connected to initial conditions (I) by a concentration shock. Figure 11.13 shows a zoom of the solution near the injection point.

The gas saturation versus distance for the partially miscible gas injection is shown in Fig. 11.14. Three regions appear; a small single-phase gas region near the injection point, followed by a two-phase region and a single-phase original reservoir

Table 11.3 Characteristic waves for miscible flooding

Eigenvalues	Shock speeds
$\lambda_c = \frac{u_T}{\phi} \frac{\partial F}{\partial C}$	$D_c = D_{BL} = \frac{u_T}{\phi} \frac{[F]}{[C]}$
$\lambda_\beta = \frac{u_T}{\phi} \frac{F \frac{d\alpha}{d\beta} + 1}{C \frac{d\alpha}{d\beta} + 1}$	$D_\beta = \frac{u_T}{\phi} \frac{F^{\pm} + \frac{[\beta]}{[\alpha]}}{C^{\pm} + \frac{[\beta]}{[\alpha]}}$

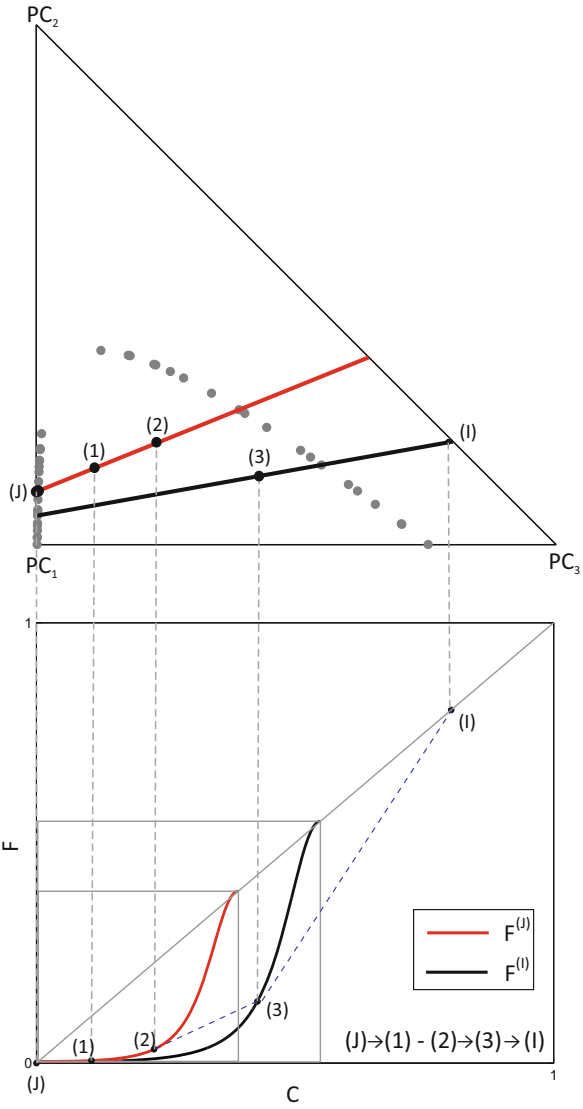


Fig. 11.12 Solution path for miscible flooding including solution path

fluid region. In the two-phase region, there is a gas saturation rarefaction region and a constant gas concentration region. A Buckley–Leverett type concentration shock connects the two-phase region and the single-phase oil region. This shock is located at $x = D_{BL}t$.

The total mobility versus distance can be obtained from the gas saturation profile, which allows a straightforward computation of the injection pressure from the following expression:

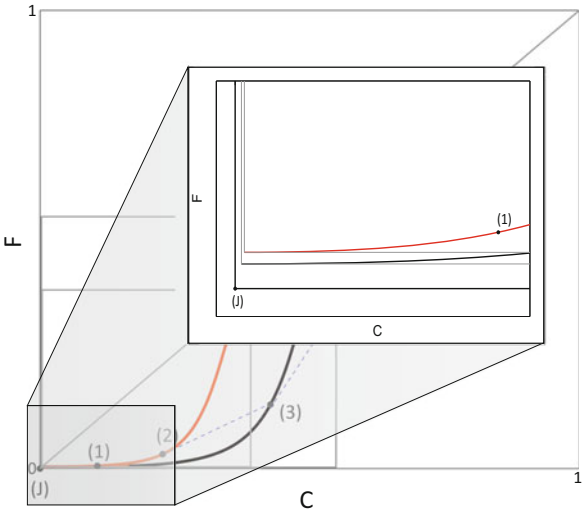


Fig. 11.13 Zoom of solution path for miscible flooding close to injection conditions

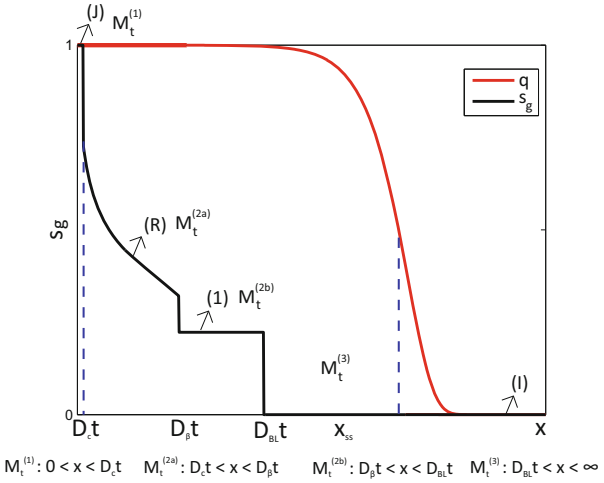


Fig. 11.14 Gas saturation profile for miscible flooding and flow rate

$$p_w(x=0,t) - p_i = \frac{q_{inj}}{\lambda_T^{(3)} kA}\left[\int_0^{D_c t}\left(\frac{\lambda_T^{(3)}}{\lambda_T^{(1)}} - 1\right)dx' + \int_{D_c t}^{D_\beta t}\left(\frac{\lambda_T^{(3)}}{\lambda_T^{(2a)}\left(x', t\right)} - 1\right)dx' + \int_{D_\beta t}^{D_{BL} t}\left(\frac{\lambda_T^{(3)}}{\lambda_T^{(2b)}} - 1\right)dx'\right] + \frac{1}{\lambda_T^{(3)} kA}\int_0^{\infty} q_T\left(x', t\right)dx' \qquad (11.24)$$

For gas injection the mobility ratio is greater than 1, thus all terms but the last in the right-hand side of Eq. (11.24) are negative. Therefore, the injection pressure at any time is lower than the pressure required for single-phase oil flow for the same volumetric rate.

11.4 Summary and Conclusions

In this work a general procedure for the pressure calculation during EOR processes in infinite reservoirs is presented. This technique was applied for the different oil recovery techniques: waterflooding, polymer flooding, and partially miscible flooding. The solution is divided into three regions: a single-phase injected fluid region (beginning at porous media inlet) followed by a two-phase region (displaced and displacing phases) where mass transfer may take place, and a single-phase original reservoir liquid region. We considered the injected fluid region and two-phase region incompressible; whereas the original reservoir fluid region was taken as slightly compressible. The saturation and concentration profiles were obtained using the method of characteristics; and the pressure profile through two-phase Darcy's law integration. The solutions developed are useful for screening the most suitable enhanced oil recovery technique for a particular field.

Acknowledgments This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior—Brasil (CAPES)—Finance Code 001. Also, the authors wish to express their gratitude for the financial support provided by Petrobras SIGER Research Network and by the Universidade Estadual do Norte Fluminense (UENF).

References

[Be93] Bedrikovetsky, P.G.: Mathematical Theory of Oil and Gas Recovery. Kluwer Academic Publishers, London (1993)

[CoEtAl56] Corey, A.T., Rathjens, C.H., Henderson, J.H., Wyllie, M.R.J.: Three-phase relative permeability. J. Can. Petrol. Technol. **8**, 63–65 (1956)

[La89] Lake, W.L.: Enhanced Oil Recovery. Prentice-Hall, Englewood Cliffs, NJ (1989)

[Or07] Orr Jr., F.M.: Theory of Gas Injection Processes. Tie-Line Publications, Copenhagen (2007)

[Pi05] Pires, A.P., Bedrikovetsky, P.G.: Analytical modeling of 1D n-component miscible displacement of ideal fluids. In: SPE Latin American and Caribbean Petroleum Engineering. SPE, Rio de Janeiro (2005). SPE 94855

[PrEtAl86] Prausnitz, J.M., Lichtenthaler, R.N., Azevedo, E.G.: Molecular Thermodynamics of Fluid-Phase Equilibria. Prentice-Hall, Englewood Cliffs (1986)

[Th97] Thompson, L.G., Reynolds, A.C.: Well testing for radially heterogeneous reservoirs under single and multiphase flow conditions. SPE Form. Eval. **12**, 57–64 (1997). SPE 30577

Chapter 12
Error Analysis and the Role of Permutation in Dynamic Iteration Schemes

Barbara Zubik-Kowal

12.1 Introduction

In this chapter, we investigate a class of dynamic iteration schemes applied to systems written in the form

$$\frac{d}{dt}x(t) = (L + D + \xi U)x(t) + g(t), \quad t \in [0, T], \tag{12.1}$$

where ξ is a given positive parameter, $g(t)$ is a given source function, and L and U are lower and upper triangular matrices, respectively. Specifically, we let the entries on the main diagonal of L and U be identically 0 and let D be a diagonal matrix. System (12.1) is supplemented by the initial condition

$$x(0) = x_0. \tag{12.2}$$

Note that after applying the Gauss–Seidel waveform relaxation technique, it is possible to obtain more than one dynamic iteration scheme depending on the ordering of the differential equations in (12.1). For example, one possible scheme obtained after the application of Gauss–Seidel waveform relaxation to (12.1) is written in the form

$$\frac{d}{dt}x^{(k+1)}(t) = (L + D)x^{(k+1)}(t) + \xi U x^{(k)}(t) + g(t), \quad t \in [0, T], \tag{12.3}$$

B. Zubik-Kowal (✉)
Department of Mathematics, Boise State University, Boise, ID, USA
e-mail: zubik@math.boisestate.edu

C. Constanda (ed.), *Computational and Analytic Methods in Science and Engineering*, https://doi.org/10.1007/978-3-030-48186-5_12

where $k = 0, 1, 2, \ldots$ and the successive dynamic iterates $x^{(k)}(t)$ are initiated from an arbitrary starting function $x^{(0)}(t)$ defined over the entire interval $[0, T]$.

Scheme (12.3) is obtained if we apply Gauss–Seidel waveform relaxation to the equations in (12.1) written in the form they are currently in, without reordering. Therefore, the corresponding initial condition for (12.3) is given by the initial vector from (12.2) and written in the form

$$x^{(k+1)}(0) = x_0. \tag{12.4}$$

On the other hand, following only a change in the order in which the differential equations in (12.1) are written, we obtain a modified dynamic iteration scheme, which is distinctly different from the previous one. This is the case even though the same Gauss–Seidel waveform relaxation technique is applied. We now formulate another dynamic iteration scheme as follows. Let the matrices $\mathbb{L}$, $\mathbb{D}$, $\mathbb{U}$ represent the matrix decomposition of (12.1) following some reordering of entire rows of the matrices L, D, and U, respectively (the same reordering is applied to each matrix). For example, if $L+D+U = \left[a_{ij}\right]_{i,j=1}^{n}$, then a possible reordering gives $\mathbb{L}+\mathbb{D}+\mathbb{U} = \left[b_{ij}\right]_{i,j=1}^{n}$, where $b_{ij} = a_{n+1-i,n+1-j}$ and n is the size of the system. That is, the rows of $\mathbb{L} + \mathbb{D} + \mathbb{U}$ and the rows of $L + D + U$ form opposite sequences. Then, an alternative dynamic iteration scheme (as opposed to (12.3)) can be written in the following form,

$$\frac{d}{dt}y^{(k+1)}(t) = (\xi\mathbb{L} + \mathbb{D})y^{(k+1)}(t) + \mathbb{U}y^{(k)}(t) + g(t). \tag{12.5}$$

Rather than being supplemented by (12.4), system (12.5) is supplemented by the alternative initial condition written in the form

$$y(0) = y_0,$$

where the components of the initial vector y_0 are ordered oppositely to those of the initial vector x_0.

Note that even though (12.5) is also obtained by the application of the Gauss–Seidel waveform relaxation technique (similarly to (12.3)), schemes (12.3) and (12.5) are different dynamic iteration schemes. Particularly, they have different convergence properties. It will be demonstrated, in the following sections, that the number k of iterations required for the successive iterates $x^{(k)}(t)$ and $y^{(k)}(t)$ to converge to the exact solution $x_i(t) = y_{n+1-i}(t)$, $i = 1, 2, \ldots, n$, is different for each of the applied schemes, indicating that the choice of permutation can impact the rate of convergence.

Motivated by electrical system simulation, Lelarasmee et al. [LeEtAl82] introduced waveform relaxation techniques which were later broadly developed by many authors following applications of the technique to parallel computing environments. Examples include [Bu95] and [MiNe96], for systems of ordinary differential equa-

tions, [Bj94] and [Bj95] for systems of delay differential equations, and [ZuVa99], [Zu00], [Zu04] for systems of more general functional differential equations. However, neither these papers nor the references therein consider rearranging the sequence of the equations in a given system as a way to optimize the process by choosing the dynamic iteration scheme with the fastest rate of convergence. The influence of the rearrangements of the sequence of equations on the convergence of the resulting dynamic iteration schemes has been recently investigated in low-dimensional systems in [Zu17] and [Zu19]. In this chapter, we expand previous results by addressing this question for dimensions up to $n = 4$ and present a methods comparison using the theoretical results.

More specifically, the goal of the chapter is to analyze the errors of the successive iterates

$$\hat{e}^{(k)}(t) = x^{(k)}(t) - x(t) \tag{12.6}$$

and

$$e^{(k)}(t) = y^{(k)}(t) - y(t) \tag{12.7}$$

in order to address the question of whether or not, and why, there is a difference in how the sequences of the successive iterates $\{x^{(k)}(t)\}_{k=0}^{\infty}$ and $\{y^{(k)}(t)\}_{k=0}^{\infty}$ converge to the exact solution even though they both originate from an application of the Gauss–Seidel waveform relaxation technique applied to the same given differential system, and whether or not the reordering of the equations in the system affects the rate of convergence of the resulting dynamic iteration schemes.

12.2 Alternative Dynamic Iteration Schemes and Their Distinct Convergence Properties

In this section, we start from addressing the question of whether or not the different dynamic iteration schemes obtained from the Gauss–Seidel waveform relaxation technique applied to (12.1) converge in the same number of iterations k. As demonstrated by the investigations developed in the following sections of the chapter, we conclude that although Gauss–Seidel waveform relaxation is applied to the same system of differential equations supplemented by the same initial conditions, the different dynamic iteration schemes (obtained by reordering the sets of differential equations (12.1) into different sequences) manifest distinct convergence properties that depend on the reordering.

As an example, the following two iterative schemes

$$\begin{cases} \frac{d}{dt}x_1^{(k+1)} = a_{11}x_1^{(k+1)} + \xi a_{12}x_2^{(k)} + \xi a_{13}x_3^{(k)} + \xi a_{14}x_4^{(k)} + g_1(t), \\ \frac{d}{dt}x_2^{(k+1)} = a_{21}x_1^{(k+1)} + a_{22}x_2^{(k+1)} + \xi a_{23}x_3^{(k)} + \xi a_{24}x_4^{(k)} + g_2(t), \\ \frac{d}{dt}x_3^{(k+1)} = a_{31}x_1^{(k+1)} + a_{32}x_2^{(k+1)} + a_{33}x_3^{(k+1)} + \xi a_{34}x_4^{(k)} + g_3(t), \\ \frac{d}{dt}x_4^{(k+1)} = a_{41}x_1^{(k+1)} + a_{42}x_2^{(k+1)} + a_{43}x_3^{(k+1)} + a_{44}x_4^{(k+1)} + g_4(t), \end{cases} \tag{12.8}$$

and

$$\begin{cases} \frac{d}{dt}x_4^{(k+1)} = a_{44}x_4^{(k+1)} + a_{43}x_3^{(k)} + a_{42}x_2^{(k)} + a_{41}x_1^{(k)} + g_4(t), \\ \frac{d}{dt}x_3^{(k+1)} = \xi a_{34}x_4^{(k+1)} + a_{33}x_3^{(k+1)} + a_{32}x_2^{(k)} + a_{31}x_1^{(k)} + g_3(t), \\ \frac{d}{dt}x_2^{(k+1)} = \xi a_{24}x_4^{(k+1)} + \xi a_{23}x_3^{(k+1)} + a_{22}x_2^{(k+1)} + a_{21}x_1^{(k)} + g_2(t), \\ \frac{d}{dt}x_1^{(k+1)} = \xi a_{14}x_4^{(k+1)} + \xi a_{13}x_3^{(k+1)} + \xi a_{12}x_2^{(k+1)} + a_{11}x_1^{(k+1)} + g_1(t), \end{cases} \tag{12.9}$$

are obtained if Gauss–Seidel waveform relaxation is applied to the same differential equations but ordered oppositely. Scheme (12.8) is determined directly from (12.1) and can be written in the form

$$\begin{cases} \frac{d}{dt}x_1 = a_{11}x_1 + \xi a_{12}x_2 + \xi a_{13}x_3 + \xi a_{14}x_4 + g_1(t), \\ \frac{d}{dt}x_2 = a_{21}x_1 + a_{22}x_2 + \xi a_{23}x_3 + \xi a_{24}x_4 + g_2(t), \\ \frac{d}{dt}x_3 = a_{31}x_1 + a_{32}x_2 + a_{33}x_3 + \xi a_{34}x_4 + g_3(t), \\ \frac{d}{dt}x_4 = a_{41}x_1 + a_{42}x_2 + a_{43}x_3 + a_{44}x_4 + g_4(t), \end{cases} \tag{12.10}$$

and scheme (12.9) is determined by first writing down the differential equations (12.10) in the opposite order before Gauss–Seidel waveform relaxation is applied. Both schemes (12.8) and (12.9) are initiated from arbitrary starting functions, which in contrast to the parameter ξ, do not influence the rates of convergence of the successive iterates $x^{(k)}(t)$ to the exact solution $x(t)$, as shown in the next section. Scheme (12.8) is supplemented by the initial condition $x^{(k+1)}(0) = x_0 = (x_{0,1}, x_{0,2}, x_{0,3}, x_{0,4})^T$ and scheme (12.9) is supplemented by the initial condition given by the initial vector $(x_{0,4}, x_{0,3}, x_{0,2}, x_{0,1})^T$, where the components of x_0 are ordered as the differential equations written in (12.9).

Note that schemes (12.8) and (12.9) are different in a number of aspects. For example, one difference is that the previous iterate $x_1^{(k)}(t)$ is used in (12.9) while it is not used in (12.8). Another noteworthy example of a difference between

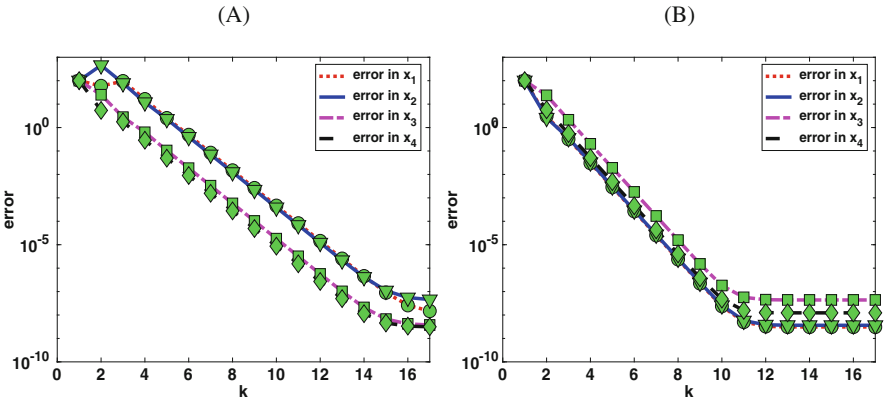


Fig. 12.1 Numerical errors arising from the application of (12.8) in panel (**a**), and from (12.9) in panel (**b**)

these schemes is that the previous iterate $x_4^{(k)}(t)$ is used in (12.8) while it is not used in (12.9). Consequently, the parameters ξa_{14}, ξa_{24}, and ξa_{34} of the governing equations serve as prefactors of the previous iterate in scheme (12.8), whereas neither of these parameters are prefactors of any of the previous iterates in scheme (12.9). These aspects give rise to differences in the rates of convergence of both schemes, as demonstrated graphically in Fig. 12.1.

As illustrated in Fig. 12.1, the schemes characterize themselves by different rates of convergence. Particularly, scheme (12.9) converges faster than scheme (12.8). Although both dynamic iteration schemes (12.8) and (12.9) originate from an application of the Gauss–Seidel waveform relaxation technique, applied to the same system of differential equations (12.10), their errors converge towards zero differently and consequently, their resulting approximations require a different number of iterations to converge towards the exact solution within a given tolerance. Namely, scheme (12.8) requires 16 iterations (as seen in Fig. 12.1, panel (a)), while scheme (12.9) requires 12 iterations (as seen in Fig. 12.1, panel (a)). More numerical examples and illustrations demonstrating various differences in convergence are presented in Sect. 12.4.

The goal of the chapter is to theoretically investigate these differences and to answer the question of what could be the possible reasons for why such similar schemes, like (12.8) and (12.9), originating from an application the Gauss–Seidel waveform relaxation technique, demonstrate convergence in a different number of iterations.

From the theoretical analysis presented in Sect. 12.3, we conclude that there is a difference in the rate of convergence of the sequences of successive iterates because of the model parameters and it is recommended to consequently reorder the given differential equations based on the values of the model parameters before applying the Gauss–Seidel waveform relaxation technique to the model equations.

Theoretical conclusions derived from the error analysis presented in Sect. 12.3 are illustrated by means of numerical experiments in Sect. 12.4. Section 12.5 is devoted to a methods comparison and finally, in Sect. 12.6, we finish with concluding remarks and plans for future work.

12.3 Error Analysis

The goal of this section is to track the role of the values of the model parameters in the propagation of errors in dynamic iteration schemes. To realize this goal, we derive formulas for the errors that feature the model parameters explicitly, in contrast to the application of matrix norms.

Let us establish the following notation. Suppose $b_{ii} \neq b_{jj}$, for $i \neq j$, $i, j = 1, 2, 3, 4$, and let

$$
\begin{aligned}
&v_1^{(1)} = 1, && v_1^{(2)} = 0, && v_1^{(3)} = 0,\\
&v_2^{(1)} = \frac{-\xi \sum_{j=1}^{1} b_{2j} v_j^{(1)}}{b_{22} - b_{11}}, && v_2^{(2)} = 1, && v_2^{(3)} = 0,\\
&v_3^{(1)} = \frac{-\xi \sum_{j=1}^{2} b_{3j} v_j^{(1)}}{b_{33} - b_{11}}, && v_3^{(2)} = \frac{-\xi \sum_{j=1}^{2} b_{3j} v_j^{(2)}}{b_{33} - b_{11}}, && v_3^{(3)} = 1,\\
&v_4^{(1)} = \frac{-\xi \sum_{j=1}^{3} b_{4j} v_j^{(1)}}{b_{44} - b_{11}}, && v_4^{(2)} = \frac{-\xi \sum_{j=1}^{3} b_{4j} v_j^{(2)}}{b_{44} - b_{11}}, && v_4^{(3)} = \frac{-\xi \sum_{j=1}^{3} b_{4j} v_j^{(3)}}{b_{44} - b_{11}}
\end{aligned}
$$

and $v_1^{(4)} = v_2^{(4)} = v_3^{(4)} = 0$, $v_4^{(4)} = 1$.

Theorem 1 *Suppose $b_{ii} \neq b_{jj}$, for $i \neq j$, and*

$$
\begin{aligned}
\eta_{ij}(t) &= v_i^{(j)}\left(e^{tb_{jj}} - e^{tb_{ii}}\right), \quad i = 2, 3, 4, \quad j = 1, 2, 3,\\
\zeta_i(t) &= v_{3+i}^{(1+i)} e^{tb_{1+i,1+i}} - v_{2+i}^{(1+i)} v_{3+i}^{(2+i)} e^{tb_{2+i,2+i}}\\
&\quad + (v_{2+i}^{(1+i)} v_{3+i}^{(2+i)} - v_{3+i}^{(1+i)}) e^{tb_{3+i,3+i}}, \quad i = 0, 1,\\
\sigma &= v_2^{(1)} v_4^{(2)} + v_3^{(1)} v_4^{(3)} - v_2^{(1)} v_3^{(2)} v_4^{(3)} - v_4^{(1)},\\
\theta(t) &= v_4^{(1)} e^{b_{11}t} - v_2^{(1)} v_4^{(2)} e^{b_{22}t} + (v_2^{(1)} v_3^{(2)} - v_3^{(1)}) v_4^{(3)} e^{b_{33}t} + \sigma e^{b_{44}t}.
\end{aligned}
$$

Then, the error $e^{(k)}(t)$ of the dynamic iteration scheme (12.9) satisfies the following relations:

$$
e_1^{(k+1)}(t) = \sum_{i=2}^{4} b_{1i} \int_0^t e^{b_{11}(t-s)} e_i^{(k)}(s) ds \tag{12.11}
$$

$$e_2^{(k+1)}(t) = \sum_{i=2}^{4} b_{1i} \int_0^t \eta_{21}(t-s) e_i^{(k)}(s) ds + \sum_{i=3}^{4} b_{2i} \int_0^t e^{b_{22}(t-s)} e_i^{(k)}(s) ds \tag{12.12}$$

$$\begin{aligned} e_3^{(k+1)}(t) = &\sum_{i=2}^{4} b_{1i} \int_0^t \zeta_0(t-s) e_i^{(k)}(s) ds + \sum_{i=3}^{4} b_{2i} \int_0^t \eta_{32}(t-s) e_i^{(k)}(s) ds \\ &+ b_{34} \int_0^t e^{b_{33}(t-s)} e_4^{(k)}(s) ds \end{aligned} \tag{12.13}$$

$$\begin{aligned} e_4^{(k+1)}(t) = &\sum_{i=2}^{4} b_{1i} \int_0^t \theta(t-s) e_i^{(k)}(s) ds + \sum_{i=3}^{4} b_{2i} \int_0^t \zeta_1(t-s) e_i^{(k)}(s) ds \\ &+ b_{34} \int_0^t \eta_{43}(t-s) e_4^{(k)}(s) ds \end{aligned} \tag{12.14}$$

Proof Subtracting the left- and right-hand side of the system

$$\frac{d}{dt} y(t) = (\xi \mathbb{L} + \mathbb{D} + \mathbb{U}) y(t) + g(t),$$

from (12.5) and using definition (12.7), we obtain the relationship

$$\frac{d}{dt} e^{(k+1)}(t) = (\xi \mathbb{L} + \mathbb{D}) e^{(k+1)}(t) + \mathbb{U} e^{(k)}(t).$$

Therefore,

$$e^{(k+1)}(t) = \int_0^t e^{(t-s)(\xi \mathbb{L} + \mathbb{D})} \mathbb{U} e^{(k)}(s) ds,$$

where

$$e^{(t-s)(\xi \mathbb{L} + \mathbb{D})} = V e^{(t-s)D} V^{-1}$$

and

$$V = \left[v_i^{(j)} \right]_{i,j=1}^{4}.$$

Then,

$$V^{-1} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ -v_2^{(1)} & 1 & 0 & 0 \\ v_2^{(1)} v_3^{(2)} - v_3^{(1)} & -v_3^{(2)} & 1 & 0 \\ \sigma & v_3^{(2)} v_4^{(3)} - v_4^{(2)} & -v_4^{(3)} & 1 \end{bmatrix},$$

and

$$Ve^{(t-s)D} = \begin{bmatrix} e^{b_{11}(t-s)} & 0 & 0 & 0 \\ v_2^{(1)} e^{b_{11}(t-s)} & e^{b_{22}(t-s)} & 0 & 0 \\ v_3^{(1)} e^{b_{11}(t-s)} & v_3^{(2)} e^{b_{22}(t-s)} & e^{b_{33}(t-s)} & 0 \\ v_4^{(1)} e^{b_{11}(t-s)} & v_4^{(2)} e^{b_{22}(t-s)} & v_4^{(3)} e^{b_{33}(t-s)} & e^{b_{44}(t-s)} \end{bmatrix}.$$

Therefore,

$$Ve^{tD}V^{-1} = \begin{bmatrix} e^{b_{11}t} & 0 & 0 & 0 \\ \eta_{21}(t) & e^{b_{22}t} & 0 & 0 \\ \zeta_0(t) & \eta_{32}(t) & e^{b_{33}t} & 0 \\ \theta(t) & \zeta_1(t) & \eta_{43}(t) & e^{b_{44}t} \end{bmatrix},$$

and the matrix $Ve^{tD}V^{-1}\mathbb{U}$ is equal to

$$\begin{bmatrix} 0 & b_{12}e^{b_{11}t} & b_{13}e^{b_{11}t} & b_{14}e^{b_{11}t} \\ 0 & b_{12}\eta_{21}(t) & b_{13}\eta_{21}(t) + b_{23}e^{b_{22}t} & b_{14}\eta_{21}(t) + b_{24}e^{b_{22}t} \\ 0 & b_{12}\zeta_0(t) & b_{13}\zeta_0(t) + b_{23}\eta_{32}(t) & b_{14}\zeta_0(t) + b_{24}\eta_{32}(t) + b_{34}e^{b_{33}t} \\ 0 & b_{12}\theta(t) & b_{13}\theta(t) + b_{23}\zeta_1(t) & b_{14}\theta(t) + b_{24}\zeta_1(t) + b_{34}\eta_{43}(t) \end{bmatrix}.$$

We now replace t by $t - s$ in the matrix, above, then multiply its first row by the vector $e^{(k)}(s) = \left(e_1^{(k)}(s), e_2^{(k)}(s), e_3^{(k)}(s), e_4^{(k)}(s)\right)^T$, integrate with respect to s over $[0, t]$, and obtain the relationship (12.11). Similarly, by multiplying the second row of the matrix by the vector $e^{(k)}(s)$ and integrating over $[0, t]$ with respect to s, we obtain the relationship (12.12). Then, using the third and fourth rows, we similarly obtain (12.13) and (12.14), respectively. □

Note that if $e^{b_{ii}t} \approx e^{b_{jj}t}$, for $i, j = 1, 2, 3, 4$, then the parameter ξ does not influence the iterative errors $e_i^{(k)}(t)$ much. This results from the fact that an application of the relationships (12.12) and (12.13) lead to

$$e_2^{(k+1)}(t) \approx \sum_{i=3}^{4} b_{2i} \int_0^t e^{b_{22}(t-s)} e_i^{(k)}(s) ds$$

and

$$e_3^{(k+1)}(t) \approx b_{34} \int_0^t e^{b_{33}(t-s)} e_4^{(k)}(s) ds,$$

which are similar to (12.11), and from (12.14), we conclude that $e_4^{(k+1)}(t)$ is small and not influenced much by the parameter ξ.

On the other hand, the parameter ξ influences the dynamic iteration scheme (12.8) more significantly. This is because, for (12.8), we get

$$\hat{e}^{(k+1)}(t) = \xi \int_0^t e^{(t-s)(L+D)} U \hat{e}^{(k)}(s) ds,$$

which implies that

$$\begin{aligned}\hat{e}^{(k+1)}(t) &= \xi^2 \int_0^t e^{(t-s)(L+D)} U \int_0^s e^{(s-\tau)(L+D)} U \hat{e}^{(k-1)}(\tau) d\tau ds = \dots \\ &= \xi^{k+1} \int_0^t e^{(t-s)(L+D)} U \int_0^s e^{(s-\tau)(L+D)} U \dots \\ &\quad \dots \int_0^p e^{(p-q)(L+D)} U \hat{e}^{(0)}(q) dq dp \dots d\tau ds.\end{aligned}$$

Therefore, eventually after $k+1$ steps, the composition of the errors $\hat{e}^{(k)}(t)$, $\hat{e}^{(k-1)}(t)$, $\dots$, $\hat{e}^{(1)}(t)$ down to $\hat{e}^{(0)}(t)$ results in the power ξ^{k+1}, which increases exponentially as k increases if $\xi > 1$. This is not the case for the errors $e^{(k)}(t)$ resulting from scheme (12.9). Consequently, if $\xi > 1$, the errors $\{e^{(k)}(t)\}_{k=0}^{\infty}$ are smaller than the corresponding errors $\{\hat{e}^{(k)}(t)\}_{k=0}^{\infty}$.

In the following section, we present numerical experiments to illustrate this conclusion and the corresponding numerical errors graphically.

12.4 Numerical Experiments and Illustrations

In this section, we present results of numerical experiments for linear systems of differential equations in order to illustrate the results of Theorem 1.

We begin by presenting the results of Figs. 12.1 and 12.2. Each figure displays numerical errors obtained by solving the same linear system via the application of the Gauss–Seidel waveform relaxation technique. Although the same iteration technique is applied to the same system of equations, evident differences are observed between the errors presented in panel (a) and panel (b) of each figure. The differences in the rates of convergence of the numerical solutions to the exact solution are caused by the fact that using different permutations of the system of differential equations before the Gauss–Seidel waveform relaxation technique is applied leads to different dynamic iteration schemes. The resulting schemes manifest different convergence rates as observed, for example, in Figs. 12.1, 12.2, and 12.3, in which the maximal errors are presented as functions of the iteration index k.

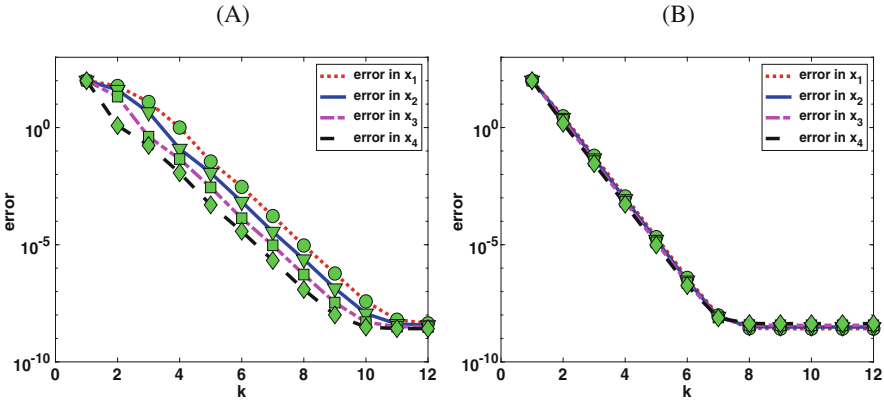


Fig. 12.2 Numerical errors resulting from the application of (12.8) in panel (**a**), and from (12.9) in panel (**b**)

The computations used to present Fig. 12.1 have been executed for system (12.1) where $\xi = 20$ and

$$L = \begin{bmatrix} 0&0&0&0\\ 1&0&0&0\\ 1&1&0&0\\ 1&1&1&0 \end{bmatrix}, \quad D = \begin{bmatrix} -102&0&0&0\\ 0&-9&0&0\\ 0&0&-101&0\\ 0&0&0&-100 \end{bmatrix}, \quad U = \begin{bmatrix} 0&1&1&1\\ 0&0&1&1\\ 0&0&0&1\\ 0&0&0&0 \end{bmatrix}. \tag{12.15}$$

For this system, we apply two different dynamic iteration schemes (12.8) and (12.9). To integrate the schemes in time, we apply BDF3. The application of BDF3 to (12.8) results in the following straightforward recursive algorithm

$$\begin{aligned}
x_{1,n+3}^{(k+1)} &= \eta_1^{-1}\Big(\tfrac{18}{11}x_{1,n+2}^{(k+1)} - \tfrac{9}{11}x_{1,n+1}^{(k+1)} + \tfrac{2}{11}x_{1,n}^{(k+1)} \\
&\quad + \tfrac{6}{11}h\big(\xi a_{12}x_{2,n+3}^{(k)} + \xi a_{13}x_{3,n+3}^{(k)} + \xi a_{14}x_{4,n+3}^{(k)} + g_{1,n+3}\big)\Big) \\
x_{2,n+3}^{(k+1)} &= \eta_2^{-1}\Big(\tfrac{18}{11}x_{2,n+2}^{(k+1)} - \tfrac{9}{11}x_{2,n+1}^{(k+1)} + \tfrac{2}{11}x_{2,n}^{(k+1)} \\
&\quad + \tfrac{6}{11}h\big(a_{21}x_{1,n+3}^{(k+1)} + \xi a_{23}x_{3,n+3}^{(k)} + \xi a_{24}x_{4,n+3}^{(k)} + g_{2,n+3}\big)\Big) \\
x_{3,n+3}^{(k+1)} &= \eta_3^{-1}\Big(\tfrac{18}{11}x_{3,n+2}^{(k+1)} - \tfrac{9}{11}x_{3,n+1}^{(k+1)} + \tfrac{2}{11}x_{3,n}^{(k+1)} \\
&\quad + \tfrac{6}{11}h\big(a_{31}x_{1,n+3}^{(k+1)} + a_{32}x_{2,n+3}^{(k+1)} + \xi a_{34}x_{4,n+3}^{(k)} + g_{3,n+3}\big)\Big) \\
x_{4,n+3}^{(k+1)} &= \eta_4^{-1}\Big(\tfrac{18}{11}x_{4,n+2}^{(k+1)} - \tfrac{9}{11}x_{4,n+1}^{(k+1)} + \tfrac{2}{11}x_{4,n}^{(k+1)} \\
&\quad + \tfrac{6}{11}h\big(a_{41}x_{1,n+3}^{(k+1)} + a_{42}x_{2,n+3}^{(k+1)} + a_{43}x_{3,n+3}^{(k+1)} + g_{4,n+3}\big)\Big).
\end{aligned} \tag{12.16}$$

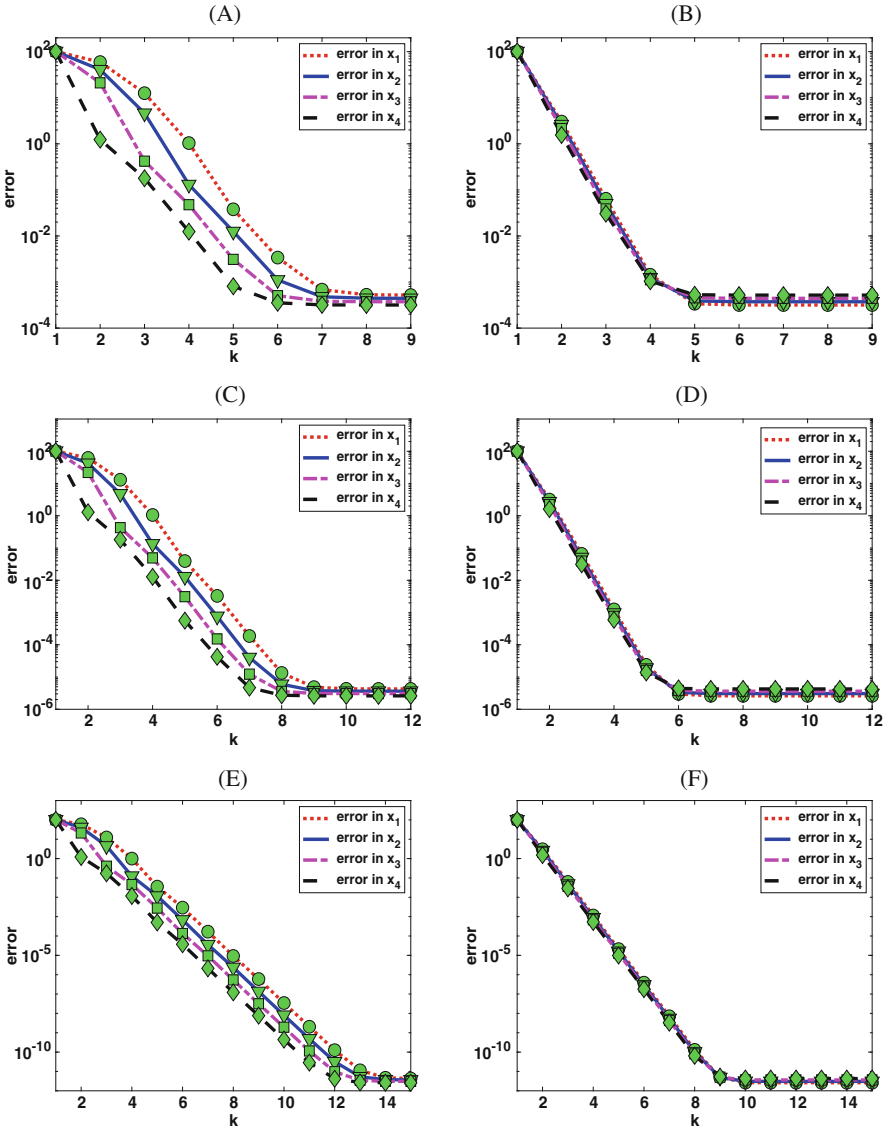


Fig. 12.3 Numerical errors resulting from the application of (12.8) in panels (**a**), (**c**), and (**e**) and from (12.9) in panels (**b**), (**d**), and (**f**). The numerical solutions in panels (**a**) and (**b**) were computed using the step size $h = 0.5$, the solutions in panels (**c**) and (**d**) were computed using $h = 0.1$, while the numerical solutions in panels (**c**) and (**d**) were computed using $h = 10^{-3}$

Here, h is the step size of the time integration, $t_n = nh$, $n = 0, 1, \ldots N$, are the corresponding grid points, $k = 0, 1, 2, \ldots$ is the iteration index, $x_{i,n}^{(k)}$ are approximations to the successive iterates $x_i^{(k)}(t_n)$ at the grid points, $\eta_i = 1 - \frac{6}{11} h a_{ii}$ and $g_{i,n+3} = g_i(t_{n+3})$, for $i = 1, 2, 3, 4$.

The algorithm (12.16) is straightforward in the sense that the approximations $x_{i,n}^{(k)}$ can be computed recursively for $k = 0, 1, 2, \ldots, n = 0, 1, \ldots, N$ without any intermediate steps. A similar recursive algorithm is obtained after the application of BDF3 to (12.9).

Note that since the current iterates $x^{k+1}(t)$ on the right-hand sides of (12.8) and (12.9) are multiplied by the lower diagonal matrices $L + D$ and $\mathbb{L} + \mathbb{D}$, respectively, and the previous iterates $x^k(t)$ are multiplied by the upper diagonal matrices U and $\mathbb{U}$, respectively, the application of the Gauss–Seidel waveform relaxation technique to (12.1) allows us to obtain the recursive algorithm (12.16) straightforwardly. Particularly, it eliminates the computationally costly necessity of determining the inverses $(L + D + \xi U)^{-1}$ and $(\xi \mathbb{L} + \mathbb{D} + \mathbb{U})^{-1}$, which would have been necessary if the technique would not have been used as an intermediate step before the application of BDF3.

The numerical errors

$$\max_{0 \le n \le N} \left| x_{i,n}^{(k)} - x_i(t_n) \right| \tag{12.17}$$

for both dynamic iteration schemes (12.8) and (12.9) applied to (12.1) defined by the matrices (12.15) are presented in Fig. 12.1a, b, respectively. The maximum errors (12.17) are plotted as functions of the iteration index k for N such that $Nh = 10$ and for $i = 1, 2, 3, 4$. Similarly, Figs. 12.2 and 12.3 present the maximum errors (12.17) for the dynamic iteration schemes (12.8) (in panels (a), (c), (e)) and (12.9) (in panels (b), (d), (f)) applied to system (12.1), where the diagonal matrix is defined by $D = \mathrm{diag}(-102, -100.9, -101, -100)$, the lower and upper triangular matrices L and U are defined as in (12.15), and $\xi = 20$.

Note that after a sufficient number of iterations, the maximum errors presented in Figs. 12.1 and 12.2 stay constant at a level of about 10^{-7}. The errors that remain constant in k (represented by the horizontal line segments) are time discretization errors. Note that the dynamic iteration schemes (12.8) and (12.9) are systems of differential equations and in order to solve them for the iterates $x^{(k)}$, they can be integrated in time t and the error resulting from the integration of (12.8) and (12.9) with respect to t is seen in the form of the errors that continue to remain constant with respect to k in Figs. 12.1 and 12.2.

The errors presented in Figs. 12.1 and 12.2 have been obtained after the integration of schemes (12.8) and (12.9) by an application of BDF3 (backward differentiation formula of order 3) with the time step $h = 10^{-2}$. Consequently, the time discretization errors seen in Figs. 12.1 and 12.2 are on the order of about 10^{-7}. Using smaller values for the time step (as seen in Fig. 12.3) or using higher order BDF methods (see Sect. 12.4, where we apply a backward differentiation formula of order 6) lead to errors that remain constant at a level that is lower than 10^{-7} after a sufficient number of iterations k.

Figure 12.3 presents two features. The first feature is the manner in which the accuracy (seen on the vertical axes) improves as the integration step size h decreases and the second feature is the improvement of the rate of convergence of the dynamic iteration schemes (seen on the horizontal axes). Panels (a) and (b) of Fig. 12.3 demonstrate numerical errors of about 10^{-4} when the step size $h = 0.5$ is used. The errors decrease down to about 10^{-6} when the step size $h = 0.1$ is used, as seen in panels (c) and (d) of Fig. 12.3. Even better improvement in the accuracy is seen in panels (e) and (f), for which the step size $h = 10^{-3}$ is used, leading to errors on the order of about 10^{-12}.

Improvement in the rate of convergence of the dynamic iteration schemes can be observed in Figs. 12.1, 12.2, and 12.3. For example, from Fig. 12.1, we observe that the maximum errors (12.17) resulting from the application of the dynamic iteration scheme (12.9) (in panel (b)) manifest a faster rate of convergence than the rate of convergence resulting from the application of scheme (12.8), in panel (a). The successive iterates $x^{(k)}(t)$ resulting from the application of the dynamic iteration scheme (12.8) converge (within the time discretization error of about 10^{-7}) to the exact solution $x(t)$ in 15 iterations. On the other hand, when we apply the Gauss–Seidel waveform relaxation technique after we change the arrangement of the equations in system (12.1) by placing them in the opposite order than that of (12.1), we obtain faster rates of convergence to the desired solution. The maximum errors (12.17) presented in panel (b), generated from the application of the dynamic iteration scheme (12.9), demonstrate the convergence of the successive iterates $x^{(k)}(t)$ to $x(t)$ in 11 iterations. This gives rise to a saving of 4 iterations in computational cost in comparison with the results for (12.8), presented in panel (a).

We derive similar conclusions from Figs. 12.2 and 12.3. For example, the maximum errors (12.17) resulting from the application of the dynamic iteration scheme (12.9), presented in panel (b) of Fig. 12.2, manifest a faster rate of convergence than that of the maximum errors resulting from the application of scheme (12.8), presented in panel (a) of Fig. 12.2. In the application of the dynamic iteration scheme (12.8), the successive iterates $x^{(k)}(t)$ converge to the exact solution $x(t)$ in 11 iterations. On the other hand, we obtain faster convergence when we apply the dynamic iteration scheme (12.9), namely this scheme converges (for the same problem) in 7 iterations. This means that we have again obtained a saving of 4 iterations in computational cost by changing only the arrangement of the equations in a given system.

12.5 Methods Comparison

In this section, we compare the method obtained by considering model parameters in dynamic iteration schemes with the solver `ode15s`. We apply the dynamic iteration schemes and the solver `ode15s` to systems of differential equations written in the form (12.1) and solve them numerically with similar accuracies (in such a way that the accuracy of the dynamic iteration schemes combined with BDF is slightly better

than the accuracy specified within the solver `ode15s`) and compare the CPU times for both of these methods. We first apply the methods to the stiff system

$$\begin{cases} \dfrac{dx_1}{dt} = -200x_1 + 0.01x_2 + g_1(t) \\ \dfrac{dx_2}{dt} = x_1 - 0.01x_2 + g_2(t), \end{cases} \tag{12.18}$$

where the forcing terms are defined by the functions $g_1(t) = p * \cos(p * t) + 200 * \sin(p * t) - 0.01 * \cos(q * t)$ and $g_2(t) = -q * \sin(q * t) - 1 * \sin(p * t) + 0.01 * \cos(q * t)$, where $p = 100$ and $q = 200$. System (12.18) is supplemented by the initial conditions $x_1(0) = x_2(0) = 0$.

The application of the Gauss–Seidel waveform relaxation technique to (12.18) results in the dynamic iteration scheme written in the form

$$\begin{cases} \dfrac{d}{dt}x_1^{(k+1)} = -200x_1^{(k+1)} + 0.01x_2^{(k)} + g_1(t) \\ \dfrac{d}{dt}x_2^{(k+1)} = x_1^{(k+1)} - 0.01x_2^{(k+1)} + g_2(t), \end{cases} \tag{12.19}$$

where $k = 0, 1, 2, \ldots$ (note that the Gauss–Seidel waveform relaxation technique generates also another dynamic iteration scheme that is different than (12.19)). Scheme (12.19) is supplemented by the initial conditions $x_1^{(k)}(0) = x_2^{(k)}(0) = 0$.

The application of BDF3 to

$$\begin{cases} \dfrac{d}{dt}x_1^{(k+1)} = a_{11}x_1^{(k+1)} + a_{12}x_2^{(k)} + g_1(t) \\ \dfrac{d}{dt}x_2^{(k+1)} = a_{21}x_1^{(k+1)} + a_{22}x_2^{(k+1)} + g_2(t), \end{cases}$$

leads to the following recursive algorithm

$$\begin{cases} x_{1,n+3}^{(k+1)} = (1 - \frac{6}{11}ha_{11})^{-1}\Big(\frac{18}{11}x_{1,n+2}^{(k+1)} - \frac{9}{11}x_{1,n+1}^{(k+1)} + \frac{2}{11}x_{1,n}^{(k+1)} \\ \qquad + \frac{6}{11}h\big(a_{12}x_{2,n+3}^{(k)} + g_1(t_{n+3})\big)\Big) \\ x_{2,n+3}^{(k+1)} = (1 - \frac{6}{11}ha_{22})^{-1}\Big(\frac{18}{11}x_{2,n+2}^{(k+1)} - \frac{9}{11}x_{2,n+1}^{(k+1)} + \frac{2}{11}x_{2,n}^{(k+1)} \\ \qquad + \frac{6}{11}h\big(a_{21}x_{1,n+3}^{(k+1)} + g_2(t_{n+3})\big)\Big), \end{cases} \tag{12.20}$$

where $k = 0, 1, 2, \ldots, n = 0, 1, \ldots, N$. Note that the algorithm (12.20) is straightforward in the sense that the approximations $x_{i,n}^{(k)}$ can be computed recursively and no intermediate steps are necessary. The algorithm (12.20) for (12.19) is defined by the parameters $a_{11} = -200$, $a_{12} = 0.01$, $a_{21} = 1$, $a_{22} = -0.01$.

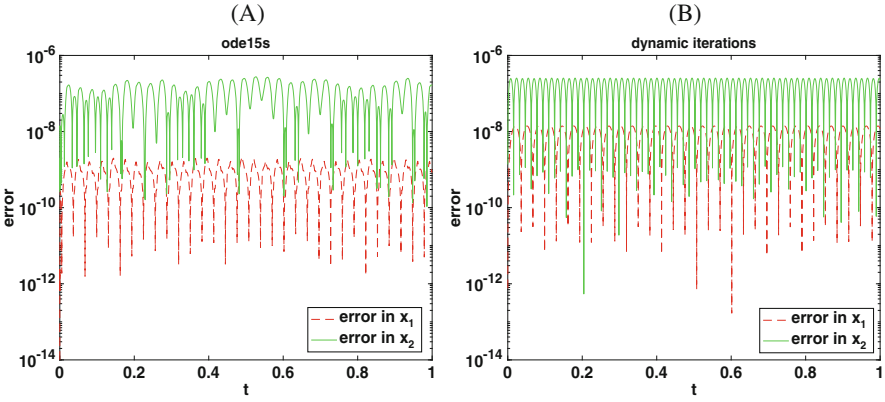


Fig. 12.4 Numerical errors resulting from the application of the recursive algorithm (12.20) (**b**) and from the solver `ode15s` (**a**)

The numerical errors

$$\left|x_1(t_n) - x_{1,n}^{(k)}\right|, \quad \left|x_2(t_n) - x_{2,n}^{(k)}\right|$$

for the recursive algorithm (12.20) are presented in Fig. 12.4b as functions of $t_n = nh$, for $h = 5 \cdot 10^{-5}$, where $n = 0, 1, 2, \ldots N$, and N is such that $Nh = 1$, and $k = 3$. Numerical errors resulting from the application of the solver `ode15s` are presented in Fig. 12.4a.

System (12.18) was solved numerically using both methods over the time interval $[0, 1]$. The numerical solution resulting from the use of the recursive algorithm (12.20) was computed at the uniformly distributed grid points $t_n = nh$, while the numerical solution resulting from the use of the solver `ode15s` was computed at the grid points that were determined from the steps selected by `ode15s`. Maximum errors were computed over the corresponding meshes generated by both methods.

The maximum error

$$\max\left\{\max_{0\le n\le N}\left|x_1(t_n) - x_{1,n}^{(k)}\right|, \max_{0\le n\le N}\left|x_2(t_n) - x_{2,n}^{(k)}\right|\right\}$$

resulting from the application of (12.20) is $2.5 \cdot 10^{-7}$ and the maximum error resulting from the application of the solver `ode15s` is $2.7 \cdot 10^{-7}$. The errors resulting from the application of both methods are comparable. However, the CPU time is 0.04 s when the recursive algorithm (12.20) is applied to (12.18) and it is 0.32 s when the solver `ode15s` is applied to the same problem, thus demonstrating that (12.20) is faster than the solver `ode15s`.

We now compare the method obtained by considering model parameters in dynamic iteration schemes with the solver `ode15s` when both methods are applied

to solve system (12.1) with L, D, and U defined by (12.15) and $\xi = 20$. We apply Theorem 1 to determine the optimal choice of the dynamic iteration scheme. Through the application of Theorem 1, one of the conclusions that we have reached in Sect. 12.3 is that, for example, the dynamic iterates $x^{(k)}(t)$ defined by scheme (12.9) converge to the exact solution $x(t)$ as $k \to \infty$ faster than the dynamic iterates defined by scheme (12.8). This faster rate of convergence of (12.9) is illustrated by means of numerical experiments in Sect. 12.4. Therefore, in this section, we consequently choose scheme (12.9) for the comparison of the two methods—the dynamic iterations and the solver `ode15s`.

In Sect. 12.4, we applied BDF3 (of order 3) to integrate both systems (12.8) and (12.9) in time t. We now apply higher order BDF6 (of order 6) to (12.9). This application leads to the following straightforward recursive algorithm

$$
\begin{aligned}
x_{4,n+6}^{(k+1)} &= \eta_4^{-1}\Big(\tfrac{360}{147}x_{4,n+5}^{(k+1)} - \tfrac{450}{147}x_{4,n+4}^{(k+1)} + \tfrac{400}{147}x_{4,n+3}^{(k+1)} - \tfrac{225}{147}x_{4,n+2}^{(k+1)} + \tfrac{72}{147}x_{4,n+1}^{(k+1)} \\
&\quad - \tfrac{10}{147}x_{4,n}^{(k+1)} + \tfrac{60}{147}h\big(a_{43}x_{3,n+6}^{(k)} + a_{42}x_{2,n+6}^{(k)} + a_{41}x_{1,n+6}^{(k)} + g_{4,n+6}\big)\Big) \\
x_{3,n+6}^{(k+1)} &= \eta_3^{-1}\Big(\tfrac{360}{147}x_{3,n+5}^{(k+1)} - \tfrac{450}{147}x_{3,n+4}^{(k+1)} + \tfrac{400}{147}x_{3,n+3}^{(k+1)} - \tfrac{225}{147}x_{3,n+2}^{(k+1)} + \tfrac{72}{147}x_{3,n+1}^{(k+1)} \\
&\quad - \tfrac{10}{147}x_{3,n}^{(k+1)} + \tfrac{60}{147}h\big(\xi a_{34}x_{4,n+6}^{(k+1)} + a_{32}x_{2,n+6}^{(k)} + a_{31}x_{1,n+6}^{(k)} + g_{3,n+6}\big)\Big) \\
x_{2,n+6}^{(k+1)} &= \eta_2^{-1}\Big(\tfrac{360}{147}x_{2,n+5}^{(k+1)} - \tfrac{450}{147}x_{2,n+4}^{(k+1)} + \tfrac{400}{147}x_{2,n+3}^{(k+1)} - \tfrac{225}{147}x_{2,n+2}^{(k+1)} + \tfrac{72}{147}x_{2,n+1}^{(k+1)} \\
&\quad - \tfrac{10}{147}x_{2,n}^{(k+1)} + \tfrac{60}{147}h\big(\xi a_{24}x_{4,n+6}^{(k+1)} + \xi a_{23}x_{3,n+6}^{(k+1)} + a_{21}x_{1,n+6}^{(k)} + g_{2,n+6}\big)\Big) \\
x_{1,n+6}^{(k+1)} &= \eta_1^{-1}\Big(\tfrac{360}{147}x_{1,n+5}^{(k+1)} - \tfrac{450}{147}x_{1,n+4}^{(k+1)} + \tfrac{400}{147}x_{1,n+3}^{(k+1)} - \tfrac{225}{147}x_{1,n+2}^{(k+1)} + \tfrac{72}{147}x_{1,n+1}^{(k+1)} \\
&\quad - \tfrac{10}{147}x_{1,n}^{(k+1)} + \tfrac{60}{147}h\big(\xi a_{14}x_{4,n+6}^{(k+1)} + \xi a_{13}x_{3,n+6}^{(k+1)} + \xi a_{12}x_{2,n+6}^{(k+1)} + g_{1,n+6}\big)\Big),
\end{aligned}
\tag{12.21}
$$

where $x_{i,n}^{(k)}$, $i = 1, 2, 3, 4$, $k = 1, 2, 3, \ldots$, are approximations to the exact solutions $x_i(t_n)$ at the grid points $t_n = nh$, for $n = 0, 1, 2, \ldots, N$, the parameters a_{ij}, $i, j = 1, 2, 3, 4$, are defined by (12.15), $\xi = 20$, $\eta_i = 1 - \frac{60}{147}ha_{ii}$ and $g_{i,n+6} = g_i(t_{n+6})$, for $i = 1, 2, 3, 4$.

The numerical errors

$$
\big|x_i(t_n) - x_{i,n}^{(k)}\big|, \quad i = 1, 2, 3, 4,
$$

resulting from the application of the recursive algorithm (12.21) are presented in Fig. 12.5b as functions of $t_n = nh$, for $h = 10^{-2}$, $n = 0, 1, 2, \ldots, N$, where N is such that $Nh = 10$, and $k = 15$. For comparison, the numerical solution resulting from the application of the solver `ode15s` was computed over the same time interval and the numerical errors are presented in Fig. 12.5a. The numerical solution resulting from the use of the solver `ode15s` was computed at the grid points determined from the steps selected by `ode15s` and the numerical error resulting from the use of `ode15s` was computed at these grid points.

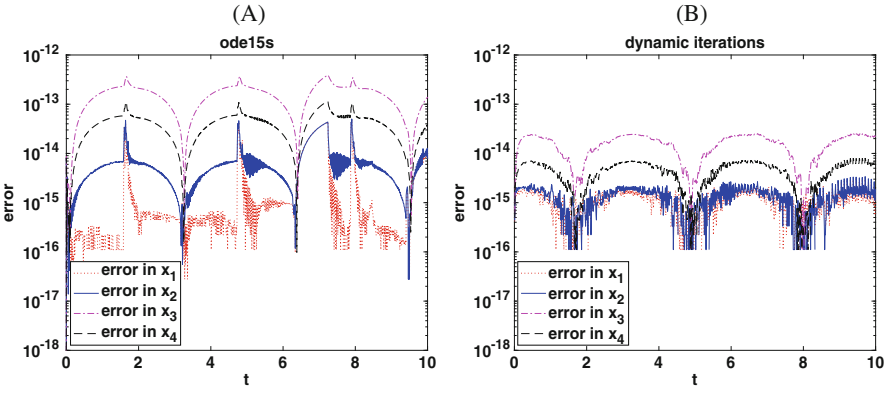


Fig. 12.5 Numerical errors resulting from the application of the recursive algorithm (12.21) (**b**) and from the solver `ode15s` (**a**)

The maximum error

$$\max_i \left\{ \max_n \left| x_i(t_n) - x_{i,n}^{(k)} \right| \right\}$$

resulting from the application of (12.21) is $2.53 \cdot 10^{-14}$ and the maximum error resulting from the application of the solver `ode15s` is $3.90 \cdot 10^{-13}$ (the maximum error resulting from `ode15s` was computed at the grid points selected by `ode15s`). The maximum error resulting from the application of (12.21) is approximately ten times smaller than the maximum error resulting from the application of the solver `ode15s`, yet the CPU time is 0.52 s when the solver `ode15s` is applied in contrast to just 0.04 s when the recursive algorithm (12.21) is applied. This demonstrates that (12.21) is more accurate and faster.

12.6 Concluding Remarks and Future Work

In this chapter, we have derived formulas for the errors of dynamic iteration schemes applied to four-dimensional systems of differential equations and concluded that the errors can decrease by rearranging the sequence of the equations in the given system. We also concluded and demonstrated that the application of the theoretical results and selection of appropriate dynamic iteration scheme gives rise to numerical solutions that are faster than the solutions computed by the variable order method widely applied to stiff differential systems. These findings are illustrated by means of numerical experiments. Future work will address the role of the model parameters in higher-dimensional systems on the rate of convergence and on the selection of the optimal dynamic iteration scheme.

References

[Bj94] Bjorhus, M.: On dynamic iteration for delay differential equations. BIT **43**, 325–336 (1994)

[Bj95] Bjorhus, M.: A note on the convergence of discretized dynamic iteration. BIT **35**, 291–296 (1995)

[Bu95] Burrage, K.: Parallel and Sequential Methods for Ordinary Differential Equations. Oxford University Press, Oxford (1995)

[LeEtAl82] Lelarasmee, E., Ruehli, A., Sangiovanni-Vincentelli, A.: The waveform relaxation method for time-domain analysis of large scale integrated circuits. IEEE Trans. CAD **1**, 131–145 (1982)

[MiNe96] Miekkala, U., Nevanlinna, O.: Iterative solution of systems of linear differential equations. Acta Numer. **5**, 259–307 (1996)

[Zu00] Zubik-Kowal, B.: Chebyshev pseudospectral method and waveform relaxation for differential and differential-functional parabolic equations. Appl. Numer. Math. **34**, 309–328 (2000)

[Zu04] Zubik-Kowal, B.: Error bounds for spatial discretization and waveform relaxation applied to parabolic functional differential equations. J. Math. Anal. Appl. **293**, 496–510 (2004)

[Zu17] Zubik-Kowal, B.: Propagation of errors in dynamic iterative schemes. In: Mikula, K., Sevcovic, D., Urban, J. (eds.) Proceedings of EQUADIFF 2017, pp. 97–106. Published by Slovak University of Technology, SPEKTRUM STU Publishing, Bratislava (2017)

[Zu19] Zubik-Kowal, B.: On the convergence of dynamic iterations in terms of model parameters. In: Constanda, C., Harris, P. (eds.) Integral methods in science and engineering, Analytic treatment and numerical approximations, pp. 219–230. Birkhäuser/Springer, Cham (2019)

[ZuVa99] Zubik-Kowal, B., Vandewalle, S.: Waveform relaxation for functional-differential equations. SIAM J. Sci. Comput. **21**, 207–226 (1999)

Correction to: An Inequality for Hölder Continuous Functions Generalizing a Result of Carlo Miranda

Massimo Lanza de Cristoforis

Correction to:
Chapter 10 in: C. Constanda (ed.),
***Computational and Analytic Methods in Science and Engineering*,**
Springer Proceedings in Complexity,
https://doi.org/10.1007/978-3-030-48186-5_10

Unfortunately the family name of the author was incorrect in the online version and it has been corrected now so that the full name appears as follows in Springer web sites.

Massimo Lanza de Cristoforis

The updated online version of this chapter can be found at
https://doi.org/10.1007/978-3-030-48186-5_10

C. Constanda (ed.), *Computational and Analytic Methods in Science and Engineering*, https://doi.org/10.1007/978-3-030-48186-5_13

Index

A
Advection–diffusion–reaction equation, 61
Angular neutron flux, 45
Asymptotic behavior, 127

B
Boundary
  element method, 3
  integral
    equations, 61
    method, 151
Bounded Lipschitz set, 197

C
Chemical signals, 151
Concentrated reaction terms, 127
Convection–diffusion equation, 151
Critical relations, 127

D
Dirichlet spectral problem, 101
Dispersion-free pressure, 223
DRBEM, 21
Dynamic iterations, 239

E
Eigenfunctions, 127
Eigenvalues, 127
Elastic plates, 75
Energy range, 45
Enhanced oil recovery, 223
Error analysis, 239
Exterior Dirichlet problem, 75

G
Gauss–Seidel waveform relaxation, 239

H
Hölder continuous functions, 197
Hypograph of a function, 197

I
Implicit integration methods, 239
Interior transmission eigenvalues, 173

L
Layer potentials, 197
Linear elasticity, 127

M
Mathematical biology, 151
Miranda, C., 197
Mixed boundary condition, 173
Motion of biological cells, 151
Multiphase flow, 223

N
Neumann eigenvalues, 1
Nonlinear eigenvalue problem, 185
Nonlinear eigenvalues, 8

C. Constanda (ed.), *Computational and Analytic Methods in Science and Engineering*, https://doi.org/10.1007/978-3-030-48186-5

P
Parametric angular neutron flux, 45
Parametric representation, 45
Permutation, 239
Point sources, 61
Porous media, 223

R
Radial basis functions, 21
Robin matrix, 127

S
Schauder space, 197
Shape optimization, 1, 11
Single-layer potential, 83
Spectral homogenization problems, 127
Stiff heavy bands, 101
Stokes flow, 151

T
Three-component flow, 223
Transient convection–diffusion–reaction, 21

U
Uniform convergence, 88

V
Variable velocity field, 21

W
Winkler foundation, 127