

Sandro Salsa

Partial Differential Equations in Action

From Modelling to Theory

Third Edition



Springer

UNITEXT – La Matematica per il 3+2

Volume 99

Editor-in-chief

A. Quarteroni

Series editors

L. Ambrosio

P. Biscari

C. Ciliberto

M. Ledoux

W.J. Runggaldier

More information about this series at <http://www.springer.com/series/5418>

Sandro Salsa

Partial Differential Equations in Action

From Modelling to Theory

Third Edition



Sandro Salsa
Dipartimento di Matematica
Politecnico di Milano
Milano, Italy

ISSN 2038-5722
UNITEXT – La Matematica per il 3+2
ISBN 978-3-319-31237-8
DOI 10.1007/978-3-319-31238-5

ISSN 2038-5757 (electronic)
ISBN 978-3-319-31238-5 (eBook)

Library of Congress Control Number: 2016932390

© Springer International Publishing Switzerland 2016

This work is subject to copyright. All rights are reserved by the Publisher, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilms or in any other physical way, and transmission or information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed.

The use of general descriptive names, registered names, trademarks, service marks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.

The publisher, the authors and the editors are safe to assume that the advice and information in this book are believed to be true and accurate at the date of publication. Neither the publisher nor the authors or the editors give a warranty, express or implied, with respect to the material contained herein or for any errors or omissions that may have been made.

Cover illustration: Simona Colombo, Giochi di Grafica, Milano, Italy
Typesetting with L^AT_EX: PTP-Berlin, Protago T_EX-Production GmbH, Germany (www.ptp-berlin.de)

This Springer imprint is published by Springer Nature
The registered company is Springer International Publishing AG Switzerland

To Anna, my wife

Preface

This book is designed as an advanced undergraduate or a first-year graduate course for students from various disciplines like applied mathematics, physics, engineering. It has evolved while teaching courses on partial differential equations (PDEs) during the last few years at the Politecnico di Milano.

The main purpose of these courses was twofold: on the one hand, to train the students to appreciate the interplay between theory and modelling in problems arising in the applied sciences, and on the other hand to give them a solid theoretical background for numerical methods, such as finite elements.

Accordingly, this textbook is divided into two parts, that we briefly describe below, writing in *italics* the relevant differences with the first edition, the second one being pretty similar.

The **first part**, Chaps. 2 to 5, has a rather elementary character with the goal of developing and studying basic problems from the macro-areas of diffusion, propagation and transport, waves and vibrations. I have tried to emphasize, whenever possible, ideas and connections with concrete aspects, in order to provide intuition and feeling for the subject.

For this part, a knowledge of advanced calculus and ordinary differential equations is required. Also, the repeated use of the method of separation of variables assumes some basic results from the theory of Fourier series, which are summarized in Appendix A.

Chapter 2 starts with the heat equation and some of its variants in which transport and reaction terms are incorporated. In addition to the classical topics, I emphasized the connections with simple stochastic processes, such as random walks and Brownian motion. This requires the knowledge of some elementary probability. It is my belief that it is worthwhile presenting this topic as early as possible, even at the price of giving up to a little bit of rigor in the presentation. An application to financial mathematics shows the interaction between probabilistic and deterministic modelling. The last two sections are devoted to two simple non linear models from flow in porous medium and population dynamics.

Chapter 3 mainly treats the Laplace/Poisson equation. The main properties of harmonic functions are presented once more emphasizing the probabilistic motivations. *I have included Perron's method of sub/super solution*, due to its renewed importance as a solution technique for fully non linear equations. The second part of this chapter deals with representation formulas in terms of potentials. In particular, the basic properties of the single and double layer potentials are presented.

Chapter 4 is devoted to first order equations and in particular to first order scalar conservation laws. The methods of characteristics and the notions of shock and rarefaction waves are introduced through a simple model from traffic dynamics. *An application to sedimentation theory illustrates the method for non convex/concave flux function*. In the last part, the method of characteristics is extended to quasilinear and fully nonlinear equations in two variables.

In Chap. 5 the fundamental aspects of waves propagation are examined, leading to the classical formulas of d'Alembert, Kirchhoff and Poisson. *A simple model of Acoustic Thermography serves as an application of Huygens principle*. In the final section, the classical model for surface waves in deep water illustrates the phenomenon of dispersion, with the help of the method of stationary phase.

The **second part** includes the two new *Chaps. 9 and 11*. In Chaps. 6 to 10 we develop Hilbert spaces methods for the variational formulation and the analysis of mainly linear boundary and initial-boundary value problems. Given the abstract nature of these chapters, I have made an effort to provide intuition and motivation about the various concepts and results, sometimes running the risk of appearing a bit wordy. The understanding of these topics requires some basic knowledge of Lebesgue measure and integration, summarized in Appendix B.

Chapter 6 contains the tools from functional analysis in Hilbert spaces, necessary for a correct variational formulation of the most common boundary value problems. The main theme is the solvability of abstract variational problems, leading to the Lax-Milgram theorem and Fredholm's alternative. Emphasis is given to the issues of compactness and weak convergence. *Section 6.10 is devoted to the fixed point theorems of Banach and of Schauder and Leray-Schauder*.

Chapter 7 is divided into two parts. The first one is a brief introduction to the theory of distributions (or generalized functions) of L. Schwartz. In the second one, the most used Sobolev spaces and their basic properties are discussed.

Chapter 8 is devoted to the variational formulation of linear elliptic boundary value problems and their solvability. The development starts with Poisson's equation and ends with general second order equations in divergence form.

In Chap. 9 I have gathered a number of applications of the variational theory of elliptic equations, in particular *to elastostatics and to the stationary Navier-Stokes equations*. Also, an application to a simple control problem is discussed.

The issue in Chap. 10, which has been *almost completely remodeled*, is the variational formulation of evolution problems, in particular of initial-boundary value problems for second order parabolic operators in divergence form and for the wave equation.

Chapter 11 contains a brief introduction to the basic concepts of the *theory of systems of first order conservation laws, in one spatial dimension*. In particular we extend from the scalar case of Chap 4, the notions of characteristics, shocks, rarefaction waves, contact discontinuity and entropy condition. The main focus is the solution of the Riemann problem.

At the end of each chapter, a number of exercises is included. Some of them can be solved by a routine application of the theory or of the methods developed in the text. Other problems are intended as a completion of some arguments or proofs in the text. Also, there are problems in which the student is required to be more autonomous. The most demanding problems are supplied with answers or hints.

Other (completely solved) exercises can be found in [17], the natural companion of this book by *S. Salsa, G. Verzini*, Springer 2015.

The order of presentation of the material is clearly a consequence of my ... prejudices. However, the exposition is flexible enough to allow substantial changes without compromising the comprehension and to facilitate a selection of topics for a one or two semester course.

In the first part, the chapters are, in practice, mutually independent, with the exception of Subsection 3.3.1 and Sect. 3.4, which presume the knowledge of Sect. 2.6.

In the second part, more attention has to be paid to the order of the arguments. The material in Sects. 6.1–6.9 and in Sect. 7.1–7.4 and 7.7–7.10 is necessary for understanding the topics in Chap. 8–10. Moreover, Chap. 9 requires also Sect. 6.10, while to cover Chap. 10, also concepts and results from Sect. 7.11 are needed.

Finally, Chap. 11 uses Subsections 4.4.2, 4.4.3 and 4.6.1.

Acknowledgments. While writing this book, during the first edition, I benefitted from comments and suggestions of many colleagues and students.

Among my colleagues, I express my gratitude to Luca Dedé, Fausto Ferrari, Carlo Pagani, Kevin Payne, Alfio Quarteroni, Fausto Saleri, Carlo Sgarra, Alessandro Veneziani, Gianmaria A. Verzini and, in particular to Cristina Cerutti, Leonede De Michele and Peter Laurence.

Among the students who have sat through my course on PDEs, I would like to thank Luca Bertagna, Michele Coti-Zelati, Alessandro Conca, Alessio Fumagalli, Loredana Gaudio, Matteo Lesinigo, Andrea Manzoni and Lorenzo Tamellini.

For the last two editions, I am particularly indebted to Leonede de Michele, Ugo Gianazza and Gianmaria Verzini for their interest, criticism and contribution. Many thanks go to Michele Di Cristo, Giovanni Molica-Bisci, Nicola Parolini Attilio Rao and Francesco Tulone for their comments and the time we spent in precious (for me) discussions. Finally, I like to express my appreciation to Francesca Bonadei and Francesca Ferrari of Springer Italia, for their constant collaboration and support.

Contents

1	Introduction	1
1.1	Mathematical Modelling	1
1.2	Partial Differential Equations	2
1.3	Well Posed Problems	5
1.4	Basic Notations and Facts	7
1.5	Smooth and Lipschitz Domains	12
1.6	Integration by Parts Formulas	15
2	Diffusion	17
2.1	The Diffusion Equation	17
2.1.1	Introduction	17
2.1.2	The conduction of heat	18
2.1.3	Well posed problems ($n = 1$)	20
2.1.4	A solution by separation of variables	23
2.1.5	Problems in dimension $n > 1$	32
2.2	Uniqueness and Maximum Principles	34
2.2.1	Integral method	34
2.2.2	Maximum principles	36
2.3	The Fundamental Solution	39
2.3.1	Invariant transformations	39
2.3.2	The fundamental solution ($n = 1$)	41
2.3.3	The Dirac distribution	43
2.3.4	The fundamental solution ($n > 1$)	47
2.4	Symmetric Random Walk ($n = 1$)	48
2.4.1	Preliminary computations	49
2.4.2	The limit transition probability	52
2.4.3	From random walk to Brownian motion	54
2.5	Diffusion, Drift and Reaction	58
2.5.1	Random walk with drift	58
2.5.2	Pollution in a channel	60
2.5.3	Random walk with drift and reaction	63

2.5.4	Critical dimension in a simple population dynamics	64
2.6	Multidimensional Random Walk	66
2.6.1	The symmetric case	66
2.6.2	Walks with drift and reaction	70
2.7	An Example of Reaction–Diffusion in Dimension $n = 3$	71
2.8	The Global Cauchy Problem ($n = 1$)	76
2.8.1	The homogeneous case	76
2.8.2	Existence of a solution	78
2.8.3	The nonhomogeneous case. Duhamel’s method	79
2.8.4	Global maximum principles and uniqueness.	82
2.8.5	The proof of the existence theorem 2.12	85
2.9	An Application to Finance	88
2.9.1	European options	88
2.9.2	An evolution model for the price S	89
2.9.3	The Black-Scholes equation	91
2.9.4	The solutions	95
2.9.5	Hedging and self-financing strategy	100
2.10	Some Nonlinear Aspects	102
2.10.1	Nonlinear diffusion. The porous medium equation	102
2.10.2	Nonlinear reaction. Fischer’s equation	105
	Problems	109
3	The Laplace Equation	115
3.1	Introduction	115
3.2	Well Posed Problems. Uniqueness	116
3.3	Harmonic Functions	118
3.3.1	Discrete harmonic functions	118
3.3.2	Mean value properties	122
3.3.3	Maximum principles	124
3.3.4	The Hopf principle	126
3.3.5	The Dirichlet problem in a disc. Poisson’s formula	127
3.3.6	Harnack’s inequality and Liouville’s theorem	131
3.3.7	Analyticity of harmonic functions	133
3.4	A probabilistic solution of the Dirichlet problem	135
3.5	Sub/Superharmonic Functions. The Perron Method	140
3.5.1	Sub/superharmonic functions	140
3.5.2	The method	142
3.5.3	Boundary behavior	143
3.6	Fundamental Solution and Newtonian Potential	147
3.6.1	The fundamental solution	147
3.6.2	The Newtonian potential	148
3.6.3	A divergence-curl system. Helmholtz decomposition formula	151
3.7	The Green Function	155
3.7.1	An integral identity	155

3.7.2	Green's function	157
3.7.3	Green's representation formula	160
3.7.4	The Neumann function	161
3.8	Uniqueness in Unbounded Domains	163
3.8.1	Exterior problems	163
3.9	Surface Potentials	166
3.9.1	The double and single layer potentials	166
3.9.2	The integral equations of potential theory	171
	Problems	174
4	Scalar Conservation Laws and First Order Equations	179
4.1	Introduction	179
4.2	Linear Transport Equation	180
4.2.1	Pollution in a channel	180
4.2.2	Distributed source	182
4.2.3	Extinction and localized source	183
4.2.4	Inflow and outflow characteristics. A stability estimate	185
4.3	Traffic Dynamics	187
4.3.1	A macroscopic model	187
4.3.2	The method of characteristics	189
4.3.3	The green light problem	191
4.3.4	Traffic jam ahead	196
4.4	Weak (or Integral) Solutions	199
4.4.1	The method of characteristics revisited	199
4.4.2	Definition of weak solution	202
4.4.3	Piecewise smooth functions and the Rankine-Hugoniot condition	205
4.5	An Entropy Condition	209
4.6	The Riemann problem	212
4.6.1	Convex/concave flux function	212
4.6.2	Vanishing viscosity method	214
4.6.3	The viscous Burgers equation	218
4.6.4	Flux function with inflection points	220
4.7	An Application to a Sedimentation Problem	224
4.8	The Method of Characteristics for Quasilinear Equations	230
4.8.1	Characteristics	230
4.8.2	The Cauchy problem	232
4.8.3	Lagrange method of first integrals	239
4.8.4	Underground flow	241
4.9	General First Order Equations	244
4.9.1	Characteristic strips	244
4.9.2	The Cauchy Problem	246
	Problems	251

5 Waves and Vibrations	259
5.1 General Concepts	259
5.1.1 Types of waves	259
5.1.2 Group velocity and dispersion relation	261
5.2 Transversal Waves in a String	264
5.2.1 The model	264
5.2.2 Energy	266
5.3 The One-dimensional Wave Equation	267
5.3.1 Initial and boundary conditions	267
5.3.2 Separation of variables	269
5.4 The d'Alembert Formula	275
5.4.1 The homogeneous equation	275
5.4.2 Generalized solutions and propagation of singularities	279
5.4.3 The fundamental solution	282
5.4.4 Nonhomogeneous equation. Duhamel's method	285
5.4.5 Dissipation and dispersion	286
5.5 Second Order Linear Equations	288
5.5.1 Classification	288
5.5.2 Characteristics and canonical form	291
5.6 The Multi-dimensional Wave Equation ($n > 1$)	296
5.6.1 Special solutions	296
5.6.2 Well posed problems. Uniqueness	298
5.7 Two Classical Models	302
5.7.1 Small vibrations of an elastic membrane	302
5.7.2 Small amplitude sound waves	306
5.8 The Global Cauchy Problem	310
5.8.1 Fundamental solution ($n = 3$) and strong Huygens' principle	310
5.8.2 The Kirchhoff formula	313
5.8.3 The Cauchy problem in dimension 2	316
5.9 The Cauchy Problem with Distributed Sources	318
5.9.1 Retarded potentials ($n = 3$)	318
5.9.2 Radiation from a moving point source	320
5.10 An Application to Thermoacoustic Tomography	324
5.11 Linear Water Waves	328
5.11.1 A model for surface waves	328
5.11.2 Dimensionless formulation and linearization	332
5.11.3 Deep water waves	334
5.11.4 Interpretation of the solution	336
5.11.5 Asymptotic behavior	338
5.11.6 The method of stationary phase	340
Problems	342

6	Elements of Functional Analysis	347
6.1	Motivations	347
6.2	Norms and Banach Spaces	353
6.3	Hilbert Spaces	358
6.4	Projections and Bases	363
6.4.1	Projections	363
6.4.2	Bases	367
6.5	Linear Operators and Duality	373
6.5.1	Linear operators	373
6.5.2	Functionals and dual space	377
6.5.3	The adjoint of a bounded operator	379
6.6	Abstract Variational Problems	382
6.6.1	Bilinear forms and the Lax-Milgram Theorem	382
6.6.2	Minimization of quadratic functionals	387
6.6.3	Approximation and Galerkin method	388
6.7	Compactness and Weak Convergence	391
6.7.1	Compactness	391
6.7.2	Compactness in $C(\bar{\Omega})$ and in $L^p(\Omega)$	392
6.7.3	Weak convergence and compactness	393
6.7.4	Compact operators	397
6.8	The Fredholm Alternative	399
6.8.1	Hilbert triplets	399
6.8.2	Solvability for abstract variational problems	402
6.8.3	Fredholm's alternative	405
6.9	Spectral Theory for Symmetric Bilinear Forms	407
6.9.1	Spectrum of a matrix	407
6.9.2	Separation of variables revisited	407
6.9.3	Spectrum of a compact self-adjoint operator	408
6.9.4	Application to abstract variational problems	411
6.10	Fixed Points Theorems	416
6.10.1	The Contraction Mapping Theorem	417
6.10.2	The Schauder Theorem	418
6.10.3	The Leray-Schauder Theorem	420
	Problems	421
7	Distributions and Sobolev Spaces	427
7.1	Distributions. Preliminary Ideas	427
7.2	Test Functions and Mollifiers	429
7.3	Distributions	433
7.4	Calculus	438
7.4.1	The derivative in the sense of distributions	438
7.4.2	Gradient, divergence, Laplacian	440
7.5	Operations with Distributions	443
7.5.1	Multiplication. Leibniz rule	443
7.5.2	Composition	444

7.5.3	Division	448
7.5.4	Convolution	449
7.5.5	Tensor or direct product	451
7.6	Tempered Distributions and Fourier Transform	454
7.6.1	Tempered distributions	454
7.6.2	Fourier transform in \mathcal{S}'	457
7.6.3	Fourier transform in L^2	460
7.7	Sobolev Spaces	461
7.7.1	An abstract construction	461
7.7.2	The space $H^1(\Omega)$	462
7.7.3	The space $H_0^1(\Omega)$	466
7.7.4	The dual of $H_0^1(\Omega)$	467
7.7.5	The spaces $H^m(\Omega)$, $m > 1$	470
7.7.6	Calculus rules	471
7.7.7	Fourier transform and Sobolev spaces	473
7.8	Approximations by Smooth Functions and Extensions	474
7.8.1	Local approximations	474
7.8.2	Extensions and global approximations	475
7.9	Traces	479
7.9.1	Traces of functions in $H^1(\Omega)$	479
7.9.2	Traces of functions in $H^m(\Omega)$	483
7.9.3	Trace spaces	484
7.10	Compactness and Embeddings	487
7.10.1	Rellich's theorem	487
7.10.2	Poincaré's inequalities	488
7.10.3	Sobolev inequality in \mathbb{R}^n	490
7.10.4	Bounded domains	492
7.11	Spaces Involving Time	494
7.11.1	Functions with values into Hilbert spaces	494
7.11.2	Sobolev spaces involving time	497
	Problems	499
8	Variational Formulation of Elliptic Problems	505
8.1	Elliptic Equations	505
8.2	Notions of Solutions	507
8.3	Problems for the Poisson Equation	509
8.3.1	Dirichlet problem	509
8.3.2	Neumann, Robin and mixed problems	512
8.3.3	Eigenvalues and eigenfunctions of the Laplace operator	517
8.3.4	An asymptotic stability result	519
8.4	General Equations in Divergence Form	521
8.4.1	Basic assumptions	521
8.4.2	Dirichlet problem	522
8.4.3	Neumann problem	527
8.4.4	Robin and mixed problems	530

8.5 Weak Maximum Principles	531
8.6 Regularity	536
Problems	544
9 Further Applications	551
9.1 A Monotone Iteration Scheme for Semilinear Equations	551
9.2 Equilibrium of a Plate	554
9.3 The Linear Elastostatic System	556
9.4 The Stokes System	561
9.5 The Stationary Navier Stokes Equations.....	566
9.5.1 Weak formulation and existence of a solution	566
9.5.2 Uniqueness	569
9.6 A Control Problem	571
9.6.1 Structure of the problem	571
9.6.2 Existence and uniqueness of an optimal pair	572
9.6.3 Lagrange multipliers and optimality conditions	574
9.6.4 An iterative algorithm	575
Problems	576
10 Weak Formulation of Evolution Problems.....	581
10.1 Parabolic Equations	581
10.2 The Cauchy-Dirichlet Problem for the Heat Equation	583
10.3 Abstract Parabolic Problems.....	586
10.3.1 Formulation	586
10.3.2 Energy estimates. Uniqueness and stability.....	589
10.3.3 The Faedo-Galerkin approximations.....	591
10.3.4 Existence	592
10.4 Parabolic PDEs	593
10.4.1 Problems for the heat equation	593
10.4.2 General Equations	596
10.4.3 Regularity	598
10.5 Weak Maximum Principles	600
10.6 The Wave Equation	602
10.6.1 Hyperbolic Equations	602
10.6.2 The Cauchy-Dirichlet problem	603
10.6.3 The method of Faedo-Galerkin	605
10.6.4 Solution of the approximate problem.....	606
10.6.5 Energy estimates	607
10.6.6 Existence, uniqueness and stability	609
Problems	611
11 Systems of Conservation Laws	615
11.1 Introduction	615
11.2 Linear Hyperbolic Systems	620
11.2.1 Characteristics	620

11.2.2 Classical solutions of the Cauchy problem	621
11.2.3 Homogeneous systems with constant coefficients. The Riemann problem	623
11.3 Quasilinear Conservation Laws	627
11.3.1 Characteristics and Riemann invariants	627
11.3.2 Weak (or integral) solutions and the Rankine-Hugoniot condition	630
11.4 The Riemann Problem	631
11.4.1 Rarefaction curves and waves. Genuinely nonlinear systems .	633
11.4.2 Solution of the Riemann problem by a single rarefaction wave	636
11.4.3 Lax entropy condition. Shock waves and contact discontinuities	638
11.4.4 Solution of the Riemann problem by a single k -shock	640
11.4.5 The linearly degenerate case	642
11.4.6 Local solution of the Riemann problem	643
11.5 The Riemann Problem for the p -system	644
11.5.1 Shock waves	644
11.5.2 Rarefaction waves	646
11.5.3 The solution in the general case	649
Problems	653
Appendix A. Fourier Series	657
A.1 Fourier Coefficients	657
A.2 Expansion in Fourier Series	660
Appendix B. Measures and Integrals	663
B.1 Lebesgue Measure and Integral	663
B.1.1 A counting problem	663
B.1.2 Measures and measurable functions	665
B.1.3 The Lebesgue integral	667
B.1.4 Some fundamental theorems	668
B.1.5 Probability spaces, random variables and their integrals	670
Appendix C. Identities and Formulas	673
C.1 Gradient, Divergence, Curl, Laplacian	673
C.2 Formulas	675
References	677
Index	681

Chapter 1

Introduction

1.1 Mathematical Modelling

Mathematical modelling plays a big role in the description of a large part of phenomena in the applied sciences and in several aspects of technical and industrial activity.

By a “mathematical model” we mean a set of equations and/or other mathematical relations capable of capturing the essential features of a complex natural or artificial system, in order to describe, forecast and control its evolution. The applied sciences are not confined to the classical ones; in addition to *physics* and *chemistry*, the practice of mathematical modelling heavily affects disciplines like *finance, biology, ecology, medicine, sociology*.

In the industrial activity (e.g. for aerospace or naval projects, nuclear reactors, combustion problems, production and distribution of electricity, traffic control, etc.) the mathematical modelling, involving first the analysis and the numerical simulation and followed by experimental tests, has become a common procedure, necessary for innovation, and also motivated by economic factors. It is clear that all of this is made possible by the enormous computational power now available.

In general, the construction of a mathematical model is based on two main ingredients:

general laws and constitutive relations.

In this book we shall deal with general laws coming from continuum mechanics and appearing as conservation or balance laws (e.g. of mass, energy, linear momentum, etc.).

The constitutive relations are of an experimental nature and strongly depend on the features of the phenomena under examination. Examples are the Fourier law of heat conduction, the Fick’s law for the diffusion of a substance or the way the speed of a driver depends on the density of cars ahead.

The outcome of the combination of the two ingredients is usually a *partial differential equation or a system of them*.

1.2 Partial Differential Equations

A partial differential equation is a relation of the following type:

$$F(x_1, \dots, x_n, u, u_{x_1}, \dots, u_{x_n}, u_{x_1 x_1}, u_{x_1 x_2} \dots, u_{x_n x_n}, u_{x_1 x_1 x_1}, \dots) = 0 \quad (1.1)$$

where the unknown $u = u(x_1, \dots, x_n)$ is a function of n variables and $u_{x_j}, \dots, u_{x_i x_j}, \dots$ are its partial derivatives. The highest order of differentiation occurring in the equation is the *order of the equation*.

A first important distinction is between *linear* and *nonlinear* equations.

Equation (1.1) is *linear* if F is linear with respect to u and all its derivatives, otherwise it is *nonlinear*.

A second distinction concerns the types of nonlinearity. We distinguish:

- *Semilinear equations* when F is nonlinear only with respect to u but is linear with respect to all its derivatives, with coefficients depending only on \mathbf{x} .
- *Quasi-linear equations* when F is linear with respect to the highest order derivatives of u , with coefficients depending only on \mathbf{x} , u and lower order derivatives.
- *Fully nonlinear equations* when F is nonlinear with respect to the highest order derivatives of u .

The theory of linear equations can be considered sufficiently well developed and consolidated, at least for what concerns the most important questions. On the contrary, the nonlinearities present such a rich variety of aspects and complications that a general theory does not appear to be conceivable. The existing results and the new investigations focus on more or less specific cases, especially interesting in the applied sciences.

To give the reader an idea of the wide range of applications we present a series of examples, suggesting one of the possible interpretations. Most of them are considered at various level of deepness in this book. In the examples, \mathbf{x} represents a space variable (usually in dimension $n = 1, 2, 3$) and t is a time variable.

We start with **linear equations**. In particular, equations (1.2)–(1.5) are fundamental and their theory constitutes a starting point for many other equations.

1. Transport equation (first order):

$$u_t + \mathbf{v} \cdot \nabla u = 0. \quad (1.2)$$

It describes for instance the transport of a solid polluting substance along a channel; here u is the concentration of the substance and \mathbf{v} is the stream speed. We consider the one-dimensional version of (1.2) in Sect. 4.2.

2. Diffusion or heat equation (second order):

$$u_t - D\Delta u = 0, \quad (1.3)$$

where $\Delta = \partial_{x_1 x_1} + \partial_{x_2 x_2} + \dots + \partial_{x_n x_n}$ is the *Laplace operator*. It describes the conduction of heat through a homogeneous and isotropic medium; u is the temper-

ature and D encodes the thermal properties of the material. Chapter 2 is devoted to the heat equation and its variants.

3. Wave equation (second order):

$$u_{tt} - c^2 \Delta u = 0. \quad (1.4)$$

It describes for instance the propagation of transversal waves of small amplitude in a perfectly elastic chord (e.g. of a violin) if $n = 1$, or membrane (e.g. of a drum) if $n = 2$. If $n = 3$ it governs the propagation of electromagnetic waves in vacuum or of small amplitude sound waves (Sect. 5.7). Here u may represent the wave amplitude and c is the propagation speed.

4. Laplace's or potential equation (second order):

$$\Delta u = 0, \quad (1.5)$$

where $u = u(\mathbf{x})$. The diffusion and the wave equations model evolution phenomena. The Laplace equation describes the corresponding *steady state*, in which the solution does not depend on time anymore. Together with its nonhomogeneous version

$$\Delta u = f,$$

called *Poisson's* equation, it plays an important role in electrostatics as well. Chapter 3 is devoted to these equations.

5. Black-Scholes equation (second order):

$$u_t + \frac{1}{2}\sigma^2 x^2 u_{xx} + rxu_x - ru = 0.$$

Here $u = u(x,t)$, $x \geq 0$, $t \geq 0$. Fundamental in mathematical finance, this equation governs the evolution of the price u of a so called *derivative* (e.g. an *European option*), based on an underlying asset (a stock, a currency, etc.) whose price is x . We meet the Black-Scholes equation in Sect. 2.9.

6. Vibrating plate (fourth order):

$$u_{tt} - \Delta^2 u = 0,$$

where $\mathbf{x} \in \mathbb{R}^2$ and

$$\Delta^2 u = \Delta(\Delta u) = \frac{\partial^4 u}{\partial x_1^4} + 2 \frac{\partial^4 u}{\partial x_1^2 \partial x_2^2} + \frac{\partial^4 u}{\partial x_2^4}$$

is the *biharmonic operator*. In the theory of linear elasticity, it models the transversal waves of small amplitude of a homogeneous isotropic plate (see Sect. 9.2 for the stationary version).

7. Schrödinger equation (second order):

$$-iu_t = \Delta u + V(\mathbf{x}) u$$

where i is the complex unit. This equation is fundamental in quantum mechanics and governs the evolution of a particle subject to a potential V . The function $|u|^2$ represents a *probability density*. We will briefly encounter the Schrödinger equation in Problem 6.6.

Let us list now some examples of **nonlinear equations**.

8. Burgers equation (quasilinear, first order):

$$u_t + cuu_x = 0 \quad (x \in \mathbb{R}).$$

It governs a one dimensional flux of a nonviscous fluid but it is used to model traffic dynamics as well. Its viscous variant

$$u_t + cuu_x = \varepsilon u_{xx} \quad (\varepsilon > 0)$$

constitutes a basic example of competition between *dissipation* (due to the term εu_{xx}) and *steepening* (shock formation due to the term cuu_x). We will discuss these topics in Sects. 4.4 and 4.5.

9. Fisher's equation (semilinear, second order):

$$u_t - D\Delta u = ru(M - u) \quad (D, r, M \text{ positive constants}).$$

It governs the evolution of a population of density u , subject to diffusion and logistic growth (represented by the right hand side). We will meet Fisher's equation in Sects. 2.10 and 9.1.

10. Porous medium equation (quasilinear, second order):

$$u_t = k \operatorname{div}(u^\gamma \nabla u)$$

where $k > 0$, $\gamma > 1$ are constant. This equation appears in the description of filtration phenomena, e.g. of the motion of water through the ground. We briefly meet the one-dimensional version of the porous medium equation in Sect. 2.10.

11. Minimal surface equation (quasilinear, second order):

$$\operatorname{div} \left(\frac{\nabla u}{\sqrt{1 + |\nabla u|^2}} \right) = 0 \quad (\mathbf{x} \in \mathbb{R}^2).$$

The graph of a solution u minimizes the area among all surfaces $z = v(x_1, x_2)$ whose boundary is a given curve. For instance, soap balls are minimal surfaces. We will not examine this equation (see e.g. *R. Mc Owen, 1996*).

12. Eikonal equation (fully nonlinear, first order):

$$|\nabla u| = c(\mathbf{x}).$$

It appears in geometrical optics: if u is a solution, its level surfaces $u(\mathbf{x}) = t$ describe the position of a light wave front at time t . A bidimensional version is examined at the end of Chap. 4.

Let us now give some examples of **systems**.

13. Navier's equation of linear elasticity: (three scalar equations of second order):

$$\rho \mathbf{u}_{tt} = \mu \Delta \mathbf{u} + (\mu + \lambda) \operatorname{grad} \operatorname{div} \mathbf{u}$$

where $\mathbf{u} = (u_1(\mathbf{x}, t), u_2(\mathbf{x}, t), u_3(\mathbf{x}, t))$, $\mathbf{x} \in \mathbb{R}^3$. The vector \mathbf{u} represents the displacement from equilibrium of a deformable continuum body of (constant) density ρ . We will examine the stationary version in Sect. 9.3.

14. Maxwell's equations in vacuum (six scalar linear equations of first order):

$$\begin{cases} \mathbf{E}_t - \operatorname{curl} \mathbf{B} = \mathbf{0}, & \mathbf{B}_t + \operatorname{curl} \mathbf{E} = \mathbf{0} \quad (\text{Ampère and Faraday laws}) \\ \operatorname{div} \mathbf{E} = 0, & \operatorname{div} \mathbf{B} = 0 \quad (\text{Gauss laws}), \end{cases}$$

where \mathbf{E} is the electric field and \mathbf{B} is the magnetic induction field. The unit measures are the "natural" ones, i.e. the light speed is $c = 1$ and the magnetic permeability is $\mu_0 = 1$. We will not examine this system (see e.g. *R. Dautray and J.L. Lions*, vol. 1, 1985).

15. Navier-Stokes equations (three quasilinear scalar equations of second order and one linear equation of first order):

$$\begin{cases} \mathbf{u}_t + (\mathbf{u} \cdot \nabla) \mathbf{u} = -\frac{1}{\rho} \nabla p + \nu \Delta \mathbf{u} \\ \operatorname{div} \mathbf{u} = 0, \end{cases}$$

where $\mathbf{u} = (u_1(\mathbf{x}, t), u_2(\mathbf{x}, t), u_3(\mathbf{x}, t))$, $p = p(\mathbf{x}, t)$, $\mathbf{x} \in \mathbb{R}^3$. This equation governs the motion of a viscous, homogeneous and incompressible fluid. Here \mathbf{u} is the fluid speed, p its pressure, ρ its density (constant) and ν is the kinematic viscosity, given by the ratio between the fluid viscosity and its density. The term $(\mathbf{u} \cdot \nabla) \mathbf{u}$ represents the inertial acceleration due to fluid transport. We will meet the stationary Navier-Stokes equation in Sect. 9.4.

1.3 Well Posed Problems

Usually, in the construction of a mathematical model, only some of the general laws of continuum mechanics are relevant, while the others are eliminated through the constitutive laws or suitably simplified according to the current situation. In

general, additional information are necessary to select or to predict the existence of a unique solution. These information are commonly supplied in the form of *initial and/or boundary data*, although other forms are possible. For instance, typical boundary conditions prescribe the value of the solution or of its normal derivative, or a combination of the two, at the boundary of the relevant domain. A main goal of a theory is to establish suitable conditions on the data in order to have a problem with the following features:

- a) *There exists at least one solution.*
- b) *There exists at most one solution.*
- c) *The solution depends continuously on the data.*

This last condition requires some explanation. Roughly speaking, property c) states that the correspondence

$$\text{data} \rightarrow \text{solution} \quad (1.6)$$

is *continuous* or, in other words, that a *small error on the data entails a small error on the solution*.

This property is extremely important and may be expressed as a **local stability of the solution with respect to the data**. Think for instance of using a computer to find an approximate solution: the insertion of the data and the computation algorithms entail approximation errors of various type. A significant sensitivity of the solution on small variations of the data would produce an unacceptable result.

The notion of continuity and the error measurements, both in the data and in the solution, are made precise by introducing a suitable notion of *distance*. In dealing with a numerical or a finite dimensional set of data, an appropriate distance may be the usual *euclidean distance*: if $\mathbf{x} = (x_1, x_2, \dots, x_n), \mathbf{y} = (y_1, y_2, \dots, y_n)$ then

$$\text{dist}(\mathbf{x}, \mathbf{y}) = |\mathbf{x} - \mathbf{y}| = \sqrt{\sum_{k=1}^n (x_k - y_k)^2}.$$

When dealing for instance with real functions, defined on a set A , common distances are:

$$\text{dist}(f, g) = \max_{\mathbf{x} \in A} |f(\mathbf{x}) - g(\mathbf{x})|,$$

which measures the maximum difference between f and g over A , or

$$\text{dist}(f, g) = \sqrt{\int_A (f - g)^2},$$

related to the so called *root-mean-square distance between f and g* .

Once the notion of distance has been chosen, the continuity of the correspondence (1.6) is easy to understand: *if the distance of the data tends to zero then the distance of the corresponding solutions tends to zero*.

When a problem possesses the properties a), b) c) above it is said to be **well posed**. When using a mathematical model, it is extremely useful, sometimes essential, to deal with well posed problems: existence of the solution indicates that the model is coherent, uniqueness and stability increase the possibility of providing accurate numerical approximations.

As one can imagine, complex models lead to complicated problems which require rather sophisticated techniques of theoretical analysis. Often, these problems become well posed and efficiently treatable by numerical methods if suitably reformulated in the abstract framework of Functional Analysis, as we will see in Chap. 6.

On the other hand, not only well posed problems are interesting for the applications. There are problems that are intrinsically *ill posed* because of the lack of uniqueness or of stability, but still of great interest for the modern technology. We only mention an important class of ill posed problems, given by the so called **inverse problems**, of which we provide a simple example in Sect. 5.10.

1.4 Basic Notations and Facts

We specify some of the symbols we will constantly use throughout the book and recall some basic notions about sets, topology and functions.

Sets and Topology. We denote by: \mathbb{N} , \mathbb{Z} , \mathbb{Q} , \mathbb{R} , \mathbb{C} the sets of natural numbers, integers, rational, real and complex numbers, respectively. \mathbb{R}^n is the n -dimensional vector space of the n -uples of real numbers. We denote by $\mathbf{e}^1, \dots, \mathbf{e}^n$ the unit vectors of the canonical base in \mathbb{R}^n . In \mathbb{R}^2 and \mathbb{R}^3 we may denote them by \mathbf{i} , \mathbf{j} and \mathbf{k} .

The symbol $B_r(\mathbf{x})$ denotes the *open ball* in \mathbb{R}^n , with radius r and center at \mathbf{x} , that is

$$B_r(\mathbf{x}) = \{\mathbf{y} \in \mathbb{R}^n; |\mathbf{x} - \mathbf{y}| < r\}.$$

If there is no need to specify the radius, we write simply $B(\mathbf{x})$. The volume of $B_r(\mathbf{x})$ and the area of $\partial B_r(\mathbf{x})$ are given by

$$|B_r| = \frac{\omega_n}{n} r^n \quad \text{and} \quad |\partial B_r| = \omega_n r^{n-1},$$

where ω_n is the surface area of the unit sphere¹ ∂B_1 in \mathbb{R}^n ; in particular $\omega_2 = 2\pi$ and $\omega_3 = 4\pi$.

Let $A \subseteq \mathbb{R}^n$. A point $\mathbf{x} \in \mathbb{R}^n$ is:

- An *interior point* if there exists a ball $B_r(\mathbf{x}) \subset A$; in particular $\mathbf{x} \in A$. The set of all the interior points of A is denoted by A° .

¹ In general, $\omega_n = n\pi^{n/2}/\Gamma(\frac{1}{2}n + 1)$ where $\Gamma(s) = \int_0^{+\infty} t^{s-1} e^{-t} dt$ is the *Euler gamma function*.

- A *boundary point* if any ball $B_r(\mathbf{x})$ contains points of A and of its complement $\mathbb{R}^n \setminus A$. The set of boundary points of A , the *boundary of A* , is denoted by ∂A ; observe that $\partial A = \partial(\mathbb{R}^n \setminus A)$.
- A *cluster point* of A if there exists a ball $B_r(\mathbf{x})$, $r > 0$, containing infinitely many points of A . Note that this is equivalent to asking that there exists a sequence $\{\mathbf{x}_m\} \subset A$ such that $\mathbf{x}_m \rightarrow \mathbf{x}$ as $m \rightarrow +\infty$. If $\mathbf{x} \in A$ and it is not a cluster point for A , we say that \mathbf{x} is an *isolated* point of A .

A set A is *open* if every point in A is an interior point; a *neighborhood* of a point \mathbf{x} is any open set A such that $\mathbf{x} \in A$.

A set C is *closed* if its complement $\mathbb{R}^n \setminus C$ is open. The set $\overline{A} = A \cup \partial A$ is the *closure of A* ; C is *closed* if and only if $C = \overline{C}$. Also, C is closed if and only if C is *sequentially closed*, that is if, for every sequence $\{\mathbf{x}_m\} \subset C$ such that $\mathbf{x}_m \rightarrow \mathbf{x}$, then $\mathbf{x} \in C$.

The unions of any family of open sets is open. The intersection of a *finite number* of open sets is open. The intersection of any family of closed sets is closed. The union of a *finite number* of closed sets is closed.

Since \mathbb{R}^n is simultaneously open and closed, its complement, that is the empty set \emptyset , is also open and closed. Only \mathbb{R}^n and \emptyset have this property among the subsets of \mathbb{R}^n .

By introducing the above notion of open set, \mathbb{R}^n becomes a *topological space*, equipped the so called *Euclidean topology*.

An open set A is *connected* if for every pair of points $\mathbf{x}, \mathbf{y} \in A$ there exists a regular curve joining them, entirely contained in A . Equivalently, A , open, is connected if it is not the union of two non-empty open subset. By a *domain* we mean an *open connected* set. Domains are usually denoted by the letter Ω .

A set A is *convex* if for every $\mathbf{x}, \mathbf{y} \in A$, the segment

$$[\mathbf{x}, \mathbf{y}] = \{\mathbf{x} + s(\mathbf{y} - \mathbf{x}) ; \forall s : 0 \leq s \leq 1\}$$

is contained in A . Clearly, any convex set is connected.

If $E \subset A$, we say that E is *dense in A* if $\overline{E} = \overline{A}$. This means that any point $\mathbf{x} \in A$ either it is an isolated point of E or it is a cluster point of E . For instance, \mathbb{Q} is dense in \mathbb{R} .

A is *bounded* if it is contained in some ball $B_r(\mathbf{0})$. The family of *compact* sets is particularly important. Let $K \subset \mathbb{R}^n$. First, we say that a family \mathcal{F} of open sets is an *open covering* of K if

$$K \subset \bigcup_{A \in \mathcal{F}} A.$$

K is *compact* if **every** open covering \mathcal{F} of K includes a finite covering of K . K is *sequentially compact* if, from every sequence $\{\mathbf{x}_m\} \subset K$, there exists a subsequence $\{\mathbf{x}_{m_k}\}$ such that $\mathbf{x}_{m_k} \rightarrow \mathbf{x} \in K$ as $k \rightarrow +\infty$.

If \overline{E} is compact and contained in A , we write $E \subset\subset A$ and we say that E is *compactly contained* in A .

In \mathbb{R}^n a subset K is compact if and only if it is closed and bounded, if and only if is sequentially compact.

Relative topology. In some situation it is convenient to consider a subset $E \subset \mathbb{R}^n$ as a topological space in itself, with an *induced* or *relative* Euclidean topology. In this topology we say that a set $A_0 \subset E$ is (*relatively*) *open in* E , if A_0 can be written as intersection between E and an open set in \mathbb{R}^n . That is if $E_0 = E \cap A$ for some A , open in \mathbb{R}^n .

Accordingly, a set $E_0 \subseteq E$ is relatively closed in E if $E \setminus E_0$ is relatively open. Clearly, a relatively open/closed set in E , could be neither open nor closed in the whole \mathbb{R}^n . For instance, the interval $[-1, 1/2)$ is relatively open in $E = [-1, 1]$, since, say,

$$[-1, 1/2) = E \cap (-2, 1/2),$$

but it is neither open nor closed in \mathbb{R} .

On the other hand, if $E = (-1, 0) \cup (0, 1)$ then, the two intervals $(-1, 0)$ and $(0, 1)$, both open in the topology of \mathbb{R} , are simultaneously open *and* closed non-empty subsets in the relative topology of E . Clearly this fact can occur only if E is disconnected. In fact, if E is *connected*, E and \emptyset are the only simultaneously relatively open and closed subsets in E .

Infimum and supremum of a set of real numbers. A set $A \subset \mathbb{R}$ is *bounded from below* if there exists a number l such that

$$l \leq x \text{ for every } x \in A. \quad (1.7)$$

The greatest among the numbers l with the property (1.7) is called the *infimum* or the *greatest lower bound* of A and denoted by $\inf A$.

More precisely, we say that $\lambda = \inf A$ if $\lambda \leq x$ for every $x \in A$ and if, for every $\varepsilon > 0$, we can find $\bar{x} \in A$ such that $x < \lambda + \varepsilon$. If $\inf A \in A$, then $\inf A$ is actually called the *minimum* of A , and may be denoted by $\min A$.

Similarly, $A \subset \mathbb{R}$ is *bounded from above* if there exists a number L such that

$$x \leq L \text{ for every } x \in A. \quad (1.8)$$

The smallest among the numbers L with the property (1.8) is called the *supremum* or the *least upper bound* of A and denoted by $\sup A$.

Precisely, we say that $\Lambda = \sup A$ if $\Lambda \geq x$ for every $x \in A$ and if, for every $\varepsilon > 0$, we can find $\bar{x} \in A$ such that $x > \Lambda - \varepsilon$. If $\sup A \in A$, then $\sup A$ is actually called the *maximum* of A , and may be denoted by $\max A$.

If A is unbounded from above or below, we set, respectively,

$$\sup A = +\infty \text{ or } \inf A = -\infty.$$

Upper and lower limits. Let $\{x_n\}$ be a sequence of real numbers. We say that $l \in \mathbb{R} \cup \{+\infty\} \cup \{-\infty\}$ is a limit point of $\{x_n\}$ if there exists a subsequence $\{x_{n_k}\}$ such that $x_{n_k} \rightarrow l$ as $k \rightarrow +\infty$.

Let E be the set of the limit points of $\{x_n\}$ and put

$$\lambda = \inf E \text{ and } \Lambda = \sup E.$$

The extended real numbers λ, Λ are called the *lower* and *upper limits* of $\{x_n\}$. We use the notations

$$\lambda = \liminf_{n \rightarrow \infty} x_n \text{ and } \Lambda = \limsup_{n \rightarrow \infty} x_n.$$

For instance, if $x_n = \cos n$, we have $E = [-1, 1]$, so that

$$\lambda = \liminf_{n \rightarrow \infty} \cos n = -1 \text{ and } \Lambda = \limsup_{n \rightarrow \infty} \cos n = 1.$$

Clearly, $\lim_{n \rightarrow \infty} x_n = l$ if and only if $E = \{l\}$; then

$$\liminf_{n \rightarrow \infty} x_n = \limsup_{n \rightarrow \infty} x_n = l.$$

Functions. Let $A \subseteq \mathbb{R}^n$ and $u : A \rightarrow \mathbb{R}$ be a real valued function defined in A . We say that u is *bounded from below* (resp. *above*) in A if the image

$$u(A) = \{y \in \mathbb{R}, y = u(\mathbf{x}) \text{ for some } \mathbf{x} \in A\}$$

is *bounded by below* (resp. *above*). The infimum (supremum) of $u(A)$ is called the *infimum (supremum) of u* and is denoted by

$$\inf_{\mathbf{x} \in A} u(\mathbf{x}) \text{ (resp. } \sup_{\mathbf{x} \in A} u(\mathbf{x})).$$

If the infimum (supremum) of $u(A)$ belongs to $u(A)$ then it is the *minimum (maximum)* of u .

We say that u is *continuous* at $\mathbf{x} \in A$ if \mathbf{x} is an isolated point of A or if $u(\mathbf{y}) \rightarrow u(\mathbf{x})$ as $\mathbf{y} \rightarrow \mathbf{x}$. If u is continuous at any point of A we say that u is continuous in A . The set of such functions is denoted by $C(A)$.

If K is compact and $u \in C(K)$ then u attains its maximum and minimum in K (*Weierstrass Theorem*).

The support of a function $u : A \rightarrow \mathbb{R}$ is the *closure in A of the set where it is different from zero*. In formulas:

$$\text{supp}(u) = \text{closure in } A \text{ of } \{\mathbf{x} \in A : u(\mathbf{x}) \neq 0\}.$$

We will denote by χ_A the *characteristic function of A* : $\chi_A = 1$ on A and $\chi_A = 0$ in $\mathbb{R}^n \setminus A$. The support of χ_A is \overline{A} .

A function is *compactly supported* in A if it vanishes outside a compact set contained in A . The symbol $C_0(A)$ denotes the subset of $C(A)$ of all functions that have a compact support in A .

We use one of the symbols u_{x_j} , $\partial_{x_j} u$, $\frac{\partial u}{\partial x_j}$ for the first partial derivatives of u , and ∇u or $\text{grad } u$ for the *gradient* of u :

$$\nabla u = \text{grad } u = (u_{x_1}, \dots, u_{x_n}).$$

Accordingly, we use the notations $u_{x_j x_k}$, $\partial_{x_j x_k} u$, $\frac{\partial^2 u}{\partial x_j \partial x_k}$ and so on for the higher order derivatives.

Given a unit vector ν , we use one the symbols $\nabla u \cdot \nu$, $\partial_\nu u$ or $\frac{\partial u}{\partial \nu}$ to denote the derivative of u in the direction ν .

Let Ω be a domain. We say that u is of class $C^k(\Omega)$, $k \geq 1$, or that it is a C^k -function in Ω , if u has continuous partials up to the order k (included) in Ω . The class of continuously differentiable functions of any order in Ω , is denoted by $C^\infty(\Omega)$. The symbol $C^k(\bar{\Omega})$ denotes the set of functions in $C^k(\Omega)$, whose derivatives, up to order k included, have a continuous extension up to $\partial\Omega$.

If $u \in C^1(\Omega)$ then u is differentiable in Ω and we can write, for $\mathbf{x} \in \Omega$ and $\mathbf{h} \in \mathbb{R}^n$, small:

$$u(\mathbf{x} + \mathbf{h}) - u(\mathbf{x}) = \nabla u(\mathbf{x}) \cdot \mathbf{h} + o(\mathbf{h})$$

where the symbol $o(\mathbf{h})$, “little o of \mathbf{h} ”, denotes a quantity such that $o(\mathbf{h}) / |\mathbf{h}| \rightarrow 0$ as $|\mathbf{h}| \rightarrow 0$.

Integrals. Up to Chap. 5 included, the integrals can be considered in the Riemann sense (proper or improper). A brief introduction to Lebesgue measure and integral is provided in Appendix B.

Let $1 \leq p < \infty$ and $q = p/(p-1)$, the *conjugate exponent* of p . The following Hölder's inequality holds

$$\left| \int_A uv \right| \leq \left(\int_A |u|^p \right)^{1/p} \left(\int_A |v|^q \right)^{1/q}. \quad (1.9)$$

The case $p = q = 2$ is known as the Schwarz inequality.

Uniform convergence of series. A series $\sum_{m=1}^{\infty} u_m$, where $u_m : A \subseteq \mathbb{R}^n \rightarrow \mathbb{R}$, is said to be *uniformly convergent in A* , with sum u if, setting $S_N = \sum_{m=1}^N u_m$, we have

$$\sup_{\mathbf{x} \in A} |S_N(\mathbf{x}) - u(\mathbf{x})| \rightarrow 0 \text{ as } N \rightarrow \infty.$$

- *Weierstrass test.* Let $|u_m(\mathbf{x})| \leq a_m$, for every $m \geq 1$ and $\mathbf{x} \in A$. If the numerical series $\sum_{m=1}^{\infty} a_m$ is convergent, then $\sum_{m=1}^{\infty} u_m$ converges absolutely and uniformly in A .
- *Limit and series.* Let $\sum_{m=1}^{\infty} u_m$ be uniformly convergent in A . If u_m is continuous at \mathbf{x}_0 for every $m \geq 1$, then u is continuous at \mathbf{x}_0 and

$$\lim_{\mathbf{x} \rightarrow \mathbf{x}_0} \sum_{m=1}^{\infty} u_m(\mathbf{x}) = \sum_{m=1}^{\infty} u_m(\mathbf{x}_0).$$

- *Term by term integration.* Let $\sum_{m=1}^{\infty} u_m$ be uniformly convergent in A . If A is bounded and u_m is integrable in A for every $m \geq 1$, then:

$$\int_A \sum_{m=1}^{\infty} u_m = \sum_{m=1}^{\infty} \int_A u_m.$$

- *Term by term differentiation.* Let A be a bounded open set and $u_m \in C^1(\overline{A})$, for every $m \geq 0$. If:

1. The series $\sum_{m=1}^{\infty} u_m(\mathbf{x})$ is convergent at some $\mathbf{x}_0 \in A$.
2. The series $\sum_{m=1}^{\infty} \partial_{x_j} u_m$ are uniformly convergent in \overline{A} , for every $j = 1, \dots, n$, then $\sum_{m=1}^{\infty} u_m$ converges uniformly in \overline{A} , its sum belongs to $C^1(\overline{A})$ and

$$\partial_{x_j} \sum_{m=1}^{\infty} u_m(\mathbf{x}) = \sum_{m=1}^{\infty} \partial_{x_j} u_m(\mathbf{x}) \quad (j = 1, \dots, n).$$

1.5 Smooth and Lipschitz Domains

We will need, especially in Chaps. 7–10, to distinguish the domains Ω in \mathbb{R}^n according to the degree of smoothness of their boundary.

Definition 1.1. We say that Ω is a C^1 –domain if for every point $\mathbf{p} \in \partial\Omega$, there exists a system of coordinates $(y_1, \dots, y_{n-1}, y_n) \equiv (\mathbf{y}', y_n)$, with origin at \mathbf{p} , a ball $B(\mathbf{p})$ and a function $\varphi_{\mathbf{p}}$ defined in a neighborhood $\mathcal{N}_{\mathbf{p}} \subset \mathbb{R}^{n-1}$ of $\mathbf{y}' = 0'$, such that

$$\varphi_{\mathbf{p}} \in C^1(\mathcal{N}_{\mathbf{p}}), \varphi_{\mathbf{p}}(\mathbf{0}') = 0$$

and

1. $\partial\Omega \cap B(\mathbf{p}) = \{(\mathbf{y}', y_n) : y_n = \varphi_{\mathbf{p}}(\mathbf{y}'), \mathbf{y}' \in \mathcal{N}_{\mathbf{p}}\}.$
2. $\Omega \cap B(\mathbf{p}) = \{(\mathbf{y}', y_n) : y_n > \varphi_{\mathbf{p}}(\mathbf{y}'), \mathbf{y}' \in \mathcal{N}_{\mathbf{p}}\}.$

The first condition expresses the fact that $\partial\Omega$ locally coincides with the graph of a C^1 –function. The second one requires that Ω is locally placed on one side of its boundary (see Fig. 1.1).

The boundary of a C^1 –domain does not have corners or edges and for every point $\mathbf{p} \in \partial\Omega$, a tangent straight line ($n = 2$) or plane ($n = 3$) or hyperplane ($n > 3$) is well defined, together with the *outward* and *inward* normal unit vectors. Moreover these vectors vary continuously on $\partial\Omega$.

The couples $(\varphi_{\mathbf{p}}, \mathcal{N}_{\mathbf{p}})$ appearing in the above definition are called *local charts*. If the functions $\varphi_{\mathbf{p}}$ are all C^k –functions, for some $k \geq 1$, Ω is said to be a C^k –domain. If Ω is a C^k –domain for every $k \geq 1$, it is said to be a C^∞ –domain. These are the domains we consider **smooth** domains.

If Ω is bounded then $\partial\Omega$ is compact and it is possible to cover $\partial\Omega$ by a finite numbers of balls $B_j = B(\mathbf{p}_j)$, $j = 1, \dots, N$, centered at $\mathbf{p}_j \in \partial\Omega$. Thus, the bound-

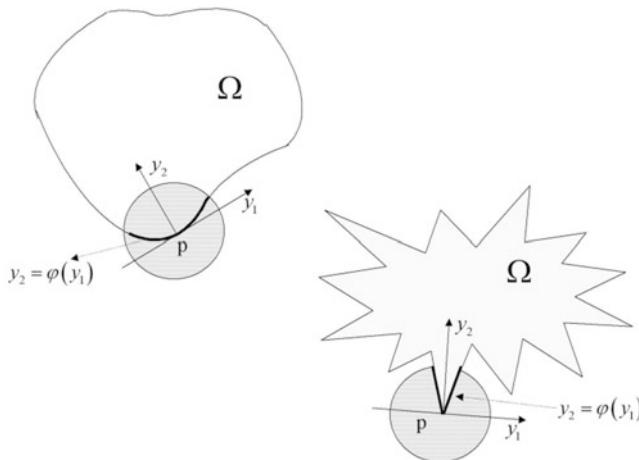


Fig. 1.1 A C^1 -domain and a Lipschitz domain

ary of Ω can be described by the finite family of local charts $(\varphi_j, \mathcal{N}_j)$, $j = 1, \dots, N$. Observe that the one-to-one transformation (*diffeomorphism*) $\mathbf{z} = \Phi_j(\mathbf{y})$, given by

$$\begin{cases} \mathbf{z}' = \mathbf{y}' \\ z_n = y_n - \varphi_j(\mathbf{y}') \end{cases} \quad \mathbf{y}' \in \mathcal{N}_j , \quad (1.10)$$

maps $\partial\Omega \cap B_j$ into a subset Γ_j of the hyperplane $z_n = 0$, so that $\partial\Omega \cap B_j$ *straightens*, as it is shown in Fig. 1.2.

In a great number of applications the relevant domains are rectangles, prisms, cones, cylinders or unions of them. Very important are polygonal domains obtained by *triangulation* procedures of smooth domains, for numerical approximations. These types of domains belong to the class of *Lipschitz domains*, whose boundary is locally described by the graph of a *Lipschitz function*.

Definition 1.2. We say that $u : \Omega \rightarrow \mathbb{R}$ is Lipschitz if there exists L such that

$$|u(\mathbf{x}) - u(\mathbf{y})| \leq L |\mathbf{x} - \mathbf{y}|$$

for every $\mathbf{x}, \mathbf{y} \in \Omega$. The number L is called the *Lipschitz constant* of u .

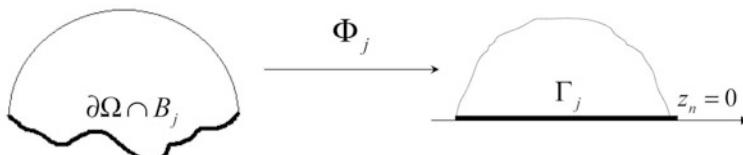


Fig. 1.2 Locally straightening $\partial\Omega$ through the diffeomorphism Φ_j

Roughly speaking, a function is Lipschitz in Ω if the increment quotients in every direction are bounded. In fact, Lipschitz functions are differentiable at all points of their domain with the exception of a negligible set of points. Precisely, we have (see e.g. [40], Ziemer, 1989):

Theorem 1.3 (Rademacher). *Let u be a Lipschitz function in $\Omega \subseteq \mathbb{R}^n$. Then u is differentiable at every point of Ω , except at a set of points of Lebesgue measure zero.*

Typical real Lipschitz functions in \mathbb{R}^n are $f(\mathbf{x}) = |\mathbf{x}|$ or, more generally, the *distance function from a closed set*, C , defined by

$$f(\mathbf{x}) = \text{dist}(\mathbf{x}, C) = \inf_{\mathbf{y} \in C} |\mathbf{x} - \mathbf{y}|.$$

Definition 1.4. *We say that a **bounded domain** Ω is **Lipschitz** if $\partial\Omega$ can be described by a family of local charts $(\varphi_j, \mathcal{N}_j)$, $j = 1, \dots, N$, where the functions φ_j , $j = 1, \dots, N$, are Lipschitz or, equivalently, if each map (1.10) is a bi-Lipschitz transformation, that is, both Φ_j and Φ_j^{-1} are Lipschitz.*

The so called mixed boundary value problems in dimension $n > 2$ require a splitting of the boundary of a domain Ω into subsets having some regularity. Typical examples of regular subsets in dimension $n = 2$ are unions of smooth arcs contained in $\partial\Omega$ and union of faces of a polyhedra for $n = 3$. Most applications deal with sets of this type.

For the sake of completeness, we introduce below a more general notion of regular subsets of $\partial\Omega$. Since the matter could quickly become very technical, we confine ourselves to a common case. Let Γ be a *relatively open* subset of $\partial\Omega$. Γ may have a boundary $\partial\Gamma$ with respect to the relative topology of $\partial\Omega$.

For instance, let Ω be a three dimensional spherical shell, bounded by two concentric spheres Γ_1 and Γ_2 . Both Γ_1 and Γ_2 are relatively open² subsets of $\partial\Omega$, with no boundary. However, if Γ is a half sphere contained, say, in Γ_1 , then its boundary is a circle. We denote by $\overline{\Gamma} = \Gamma \cup \partial\Gamma$ the *closure* of Γ in $\partial\Omega$.

Definition 1.5. *Let $\Omega \subset \mathbb{R}^n$, $n \geq 2$ be a bounded domain. We say that a relatively open subset $\Gamma \subset \partial\Omega$ is a *regular subset* of $\partial\Omega$ if:*

1. *$\overline{\Gamma}$ is locally Lipschitz; that is, for every $\mathbf{p} \in \overline{\Gamma}$ there exists a bi-Lipschitz map $\Psi_{\mathbf{p}}$ that flattens $\overline{\Gamma}$ near \mathbf{p} into a subset of the hyperplane $\{z_n = 0\}$ as in Fig. 1.2.*
2. *If $\partial\Gamma \neq \emptyset$, for every $\mathbf{p} \in \partial\Gamma$, $\Psi_{\mathbf{p}}$ also straightens $\partial\Gamma$, near \mathbf{p} , on the hyperplane $\{z_n = 0\}$.*

The second conditions implies that Γ lies on one side of $\partial\Gamma$, locally on $\partial\Omega$. The map $\Psi_{\mathbf{p}}$ may be obtained by composing the map $\Phi_{\mathbf{p}}$ that flattens $\partial\Omega$ near \mathbf{p} , defined as in (1.10), with another map that straightens $\Phi_{\mathbf{p}}(\partial\Gamma)$ near $\mathbf{z} = \mathbf{0}$, on the hyperplane $z_n = 0$.

² Also closed, in this case.

1.6 Integration by Parts Formulas

Let $\Omega \subset \mathbb{R}^n$, be a bounded C^1 - domain and

$$\mathbf{F} = (F_1, F_2, \dots, F_n) : \overline{\Omega} \rightarrow \mathbb{R}^n$$

be a vector field with components F_j , $j = 1, \dots, n$, of class $C^1(\overline{\Omega})$; we write $\mathbf{F} \in C^1(\overline{\Omega}; \mathbb{R}^n)$. The **Gauss divergence formula** holds:

$$\int_{\Omega} \operatorname{div} \mathbf{F} \, d\mathbf{x} = \int_{\partial\Omega} \mathbf{F} \cdot \boldsymbol{\nu} \, d\sigma, \quad (1.11)$$

where

$$\operatorname{div} \mathbf{F} = \sum_{j=1}^n \partial_{x_j} F_j$$

denotes the divergence of \mathbf{F} , $\boldsymbol{\nu}$ denotes the *outward normal* unit vector to $\partial\Omega$ and $d\sigma$ is the “surface” measure on $\partial\Omega$, locally given in terms of the local charts by³

$$d\sigma = \sqrt{1 + |\nabla \varphi_j(\mathbf{y}')|^2} d\mathbf{y}'.$$

A number of useful identities can be derived from (1.11). Applying (1.11) to $v\mathbf{F}$, with $v \in C^1(\overline{\Omega})$, and recalling the identity

$$\operatorname{div}(v\mathbf{F}) = v \operatorname{div} \mathbf{F} + \nabla v \cdot \mathbf{F},$$

we obtain the following **integration by parts** formula:

$$\int_{\Omega} v \operatorname{div} \mathbf{F} \, d\mathbf{x} = \int_{\partial\Omega} v \mathbf{F} \cdot \boldsymbol{\nu} \, d\sigma - \int_{\Omega} \nabla v \cdot \mathbf{F} \, d\mathbf{x}. \quad (1.12)$$

Choosing $\mathbf{F} = \nabla u$, $u \in C^2(\Omega) \cap C^1(\overline{\Omega})$, since

$$\operatorname{div} \nabla u = \Delta u \quad \text{and} \quad \nabla u \cdot \boldsymbol{\nu} = \partial_{\boldsymbol{\nu}} u,$$

the following **Green's identity** follows:

$$\int_{\Omega} v \Delta u \, d\mathbf{x} = \int_{\partial\Omega} v \partial_{\boldsymbol{\nu}} u \, d\sigma - \int_{\Omega} \nabla v \cdot \nabla u \, d\mathbf{x}. \quad (1.13)$$

In particular, the choice $v \equiv 1$ yields

$$\int_{\Omega} \Delta u \, d\mathbf{x} = \int_{\partial\Omega} \partial_{\boldsymbol{\nu}} u \, d\sigma. \quad (1.14)$$

³ The definition of the integral over the boundary $\partial\Omega$ of a C^1 -domain is given in Chap. 7, Sect. 9.

16 1 Introduction

If also $v \in C^2(\Omega) \cap C^1(\overline{\Omega})$, interchanging the roles of u and v in (1.13) and subtracting, we derive a second **Green's identity**:

$$\int_{\Omega} v \Delta u - u \Delta v \, d\mathbf{x} = \int_{\partial\Omega} (v \partial_{\boldsymbol{\nu}} u - u \partial_{\boldsymbol{\nu}} v) \, d\sigma. \quad (1.15)$$

Remark 1.6. All the above formulas hold for Lipschitz domains as well. In fact, the Rademacher theorem implies that at every point of the boundary of a Lipschitz domain, with the exception of a set of points of surface measure zero, there is a well defined tangent plane. This is enough for extending the formulas (1.12)–(1.15) to Lipchitz domains.

Chapter 2

Diffusion

2.1 The Diffusion Equation

2.1.1 Introduction

The one-dimensional *diffusion equation* is the linear second order partial differential equation

$$u_t - Du_{xx} = f,$$

where $u = u(x, t)$, x is a real space variable, t a time variable and D a positive constant, called *diffusion coefficient*. In space dimension $n > 1$, that is when $\mathbf{x} \in \mathbb{R}^n$, the diffusion equation reads

$$u_t - D\Delta u = f, \quad (2.1)$$

where Δ denotes the *Laplace operator*:

$$\Delta = \sum_{k=1}^n \frac{\partial^2}{\partial x_k^2}.$$

When $f \equiv 0$ the equation is said to be *homogeneous* and in this case the *superposition principle* holds: if u and v are solutions of (2.1) and a, b are real (or complex) numbers, $au + bv$ also is a solution of (2.1). More generally, if $u_k(\mathbf{x}, t)$ is a family of solutions depending on the parameter k (integer or real) and $g = g(k)$ is a function rapidly vanishing at infinity, then

$$\sum_{k=1}^{\infty} u_k(\mathbf{x}, t) g(k) \quad \text{and} \quad \int_{-\infty}^{+\infty} u_k(\mathbf{x}, t) g(k) dk$$

are still solutions.

A common example of diffusion is given by *heat conduction* in a solid body. Conduction comes from molecular collision, transferring heat by kinetic energy, without

macroscopic material movement. If the medium is homogeneous and isotropic with respect to the heat propagation, the evolution of the temperature is described by equation (2.1); f represents the intensity of an external distributed source. For this reason eq. (2.1) is also known as the **heat equation**.

On the other hand eq. (2.1) constitutes a much more general diffusion model, where by **diffusion** we mean, for instance, the *transport of a substance due to the molecular motion of the surrounding medium*. In this case, u could represent the concentration of a polluting material or of a solute in a liquid or a gas (dye in a liquid, smoke in the atmosphere) or even a probability density. We may say that the diffusion equation unifies at a macroscopic scale a variety of phenomena, that look quite different when observed at a microscopic scale.

Through equation (2.1) and some of its variants we will explore the deep connection between probabilistic and deterministic models, according (roughly) to the scheme

$$\text{diffusion processes} \leftrightarrow \text{probability density} \leftrightarrow \text{differential equations.}$$

The *star* in this field is *Brownian motion*, derived from the name of the botanist Brown, who observed in the middle of the 19th century, the apparently chaotic behavior of certain particles on a water surface, due to the molecular motion. This irregular motion is now modeled as a *stochastic process* under the terminology of *Wiener process or Brownian motion*. The operator

$$\frac{1}{2}\Delta$$

is strictly related to Brownian motion¹ and indeed it captures and synthesizes the microscopic features of that process.

Under equilibrium conditions, that is when there is no time evolution, the solution u depends only on the space variable and satisfies the *stationary* version of the diffusion equation (letting $D = 1$)

$$-\Delta u = f \quad (2.2)$$

($-u_{xx} = f$, in dimension $n = 1$). Equation (2.2) is known as the *Poisson equation*. When $f = 0$, it is called *Laplace's equation* and its solutions are so important in so many fields that they have deserved the special name of **harmonic functions**. This equations will be considered in the next chapter.

2.1.2 The conduction of heat

Heat is a form of energy which it is frequently convenient to consider as separated from other forms. For historical reasons, *calories* instead of Joules are used as units of measurement, each *calorie* corresponding to 4.182 Joules.

¹ In the theory of stochastic processes, $\frac{1}{2}\Delta$ represents the *infinitesimal generator of the Brownian motion*.

We want to derive a mathematical model for the heat conduction in a solid body. We assume that the body is homogeneous and isotropic, with constant *mass density* ρ , and that it can receive energy from an external source (for instance, from an electrical current or a chemical reaction or from external absorption/radiation). Denote by r the time rate per unit mass at which heat is supplied² by the external source.

Since heat is a form of energy, it is natural to use the law of conservation of energy, that we can formulate in the following way:

Let V be an arbitrary control volume inside the body. *The time rate of change of thermal energy in V equals the net flux of heat through the boundary ∂V of V , due to the conduction, plus the time rate at which heat is supplied by the external sources.*

If we denote by $e = e(\mathbf{x}, t)$ the thermal energy per unit mass, the total quantity of thermal energy inside V is given by

$$\int_V e \rho \, d\mathbf{x}$$

so that its time rate of change is³

$$\frac{d}{dt} \int_V e \rho \, d\mathbf{x} = \int_V e_t \rho \, d\mathbf{x}.$$

Denote by \mathbf{q} the *heat flux* vector⁴, which specifies the heat flow direction and the magnitude of the rate of flow across a unit area. More precisely, if $d\sigma$ is an area element contained in ∂V with *outer* unit normal $\boldsymbol{\nu}$, then

$$\mathbf{q} \cdot \boldsymbol{\nu} d\sigma$$

is the energy flow rate through $d\sigma$ and therefore the *total inner heat flux* through ∂V is given by

$$-\int_{\partial V} \mathbf{q} \cdot \boldsymbol{\nu} \, d\sigma \underset{(\text{divergence theorem})}{=} -\int_V \operatorname{div} \mathbf{q} \, d\mathbf{x}.$$

Finally, the contribution due to the external source is given by

$$\int_V r \rho \, d\mathbf{x}.$$

Thus, conservation of energy requires:

$$\int_V e_t \rho \, d\mathbf{x} = -\int_V \operatorname{div} \mathbf{q} \, d\mathbf{x} + \int_V r \rho \, d\mathbf{x}. \quad (2.3)$$

² Dimensions of r : $[r] = [\text{cal}] \times [\text{time}]^{-1} \times [\text{mass}]^{-1}$.

³ Assuming that the time derivative can be carried inside the integral.

⁴ $[\mathbf{q}] = [\text{cal}] \times [\text{length}]^{-2} \times [\text{time}]^{-1}$.

The arbitrariness of V allows us to convert the integral equation (2.3) into the pointwise relation

$$e_t \rho = -\operatorname{div} \mathbf{q} + r \rho \quad (2.4)$$

that constitutes a basic law of heat conduction. However, e and \mathbf{q} are unknown and we need additional information through *constitutive relations* for these quantities. We assume the following:

- **Fourier law** of heat conduction. Under “normal” conditions, for many solid materials, the heat flux is a linear function of the temperature gradient, that is:

$$\mathbf{q} = -\kappa \nabla u \quad (2.5)$$

where u is the absolute temperature and $\kappa > 0$, the *thermal conductivity*⁵, depends on the properties of the material. In general, κ may depend on u , \mathbf{x} and t , but often varies so little in cases of interest that it is reasonable to neglect its variation. Here we consider κ *constant* so that

$$\operatorname{div} \mathbf{q} = -\kappa \Delta u. \quad (2.6)$$

The minus sign in the law (2.5) reflects the tendency of heat to flow from hotter to cooler regions.

- The thermal energy is a linear function of the absolute temperature:

$$e = c_v u \quad (2.7)$$

where c_v denotes the *specific heat*⁶ (at constant volume) of the material. In many cases of interest c_v can be considered constant. The relation (2.7) is reasonably true over not too wide ranges of temperature.

Using (2.6) and (2.7), equation (2.4) becomes

$$u_t = \frac{\kappa}{c_v \rho} \Delta u + \frac{1}{c_v} r \quad (2.8)$$

which is the diffusion equation with $D = \kappa / (c_v \rho)$ and $f = r / c_v$. As we will see, the coefficient D , called *thermal diffusivity*, encodes the thermal response time of the material.

2.1.3 Well posed problems ($n = 1$)

As we have mentioned at the end of chapter one, the governing equations in a mathematical model have to be supplemented by additional information in or-

⁵ $[\kappa] = [\text{cal}] \times [\text{deg}]^{-1} \times [\text{time}]^{-1} \times [\text{length}]^{-1}$ (deg stays for Kelvin degree).

⁶ $[c_v] = [\text{cal}] \times [\text{deg}]^{-1} \times [\text{mass}]^{-1}$.

der to obtain a *well posed problem*, i.e. a problem that has exactly one solution, depending continuously on the data.

On physical grounds, it is not difficult to outline some typical well posed problems for the heat equation. Consider the evolution of the temperature u inside a cylindrical bar, whose lateral surface is *perfectly insulated* and whose length is much larger than its cross-sectional area A . Although the bar is three dimensional, we may assume that heat moves only down the length of the bar and that the heat transfer intensity is uniformly distributed in each section of the bar. Thus we may assume that $e = e(x, t)$, $r = r(x, t)$, with $0 \leq x \leq L$. Accordingly, the constitutive relations (2.5) and (2.7) read

$$e(x, t) = c_v u(x, t), \quad \mathbf{q} = -\kappa u_x \mathbf{i}.$$

By choosing $V = A \times (x, x + \Delta x)$ as the control volume in (2.3), the cross-sectional area A cancels out, and we obtain

$$\int_x^{x+\Delta x} c_v \rho u_t \, dx = \int_x^{x+\Delta x} \kappa u_{xx} \, dx + \int_x^{x+\Delta x} r \rho \, dx$$

that yields for u the one-dimensional heat equation

$$u_t - Du_{xx} = f.$$

We want to study the temperature evolution during an interval of time, say, from $t = 0$ until $t = T$. It is then reasonable to prescribe its initial distribution inside the bar: different initial configurations will correspond to different evolutions of the temperature along the bar. Thus we need to prescribe **the initial condition**

$$u(x, 0) = g(x),$$

where g models the initial temperature profile.

This is not enough to determine a unique evolution; it is necessary to know how the bar interacts with the surroundings. Indeed, starting with a given initial temperature distribution, we can change the evolution of u by controlling the temperature or the heat flux at the two ends of the bar⁷; for instance, we could keep the temperature at a certain fixed level or let it vary in a certain way, depending on time. This amounts to prescribing

$$u(0, t) = h_1(t), \quad u(L, t) = h_2(t) \tag{2.9}$$

at any time $t \in (0, T]$. The (2.9) are called **Dirichlet boundary conditions**.

We could also prescribe the heat flux at the end points. Since from Fourier law we have

$$\text{inward heat flow at } x = 0 : -\kappa u_x(0, t),$$

⁷ Remember that the bar has perfect lateral thermal insulation.

inward heat flow at $x = L : \kappa u_x(L, t)$,

the heat flux is assigned through the **Neumann boundary conditions**

$$-u_x(0, t) = h_1(t), u_x(L, t) = h_2(t)$$

at any time $t \in (0, T]$.

Another type of boundary condition is the **Robin or radiation condition**. Let the surroundings be kept at temperature U and assume that the inward heat flux from one end of the bar, say $x = L$, depends linearly on the difference $U - u$, that is⁸

$$\kappa u_x = \gamma(U - u) \quad (\gamma > 0). \quad (2.10)$$

Letting $\alpha = \gamma/\kappa > 0$ and $h = \gamma U/\kappa$, the Robin condition at $x = L$ reads

$$u_x + \alpha u = h.$$

Clearly, it is possible to assign **mixed conditions**: for instance, at one end a Dirichlet condition and at the other one a Neumann condition.

The problems associated with the above boundary conditions have a corresponding nomenclature. Summarizing, we can state the most common problems for the one dimensional heat equation as follows: *given $f = f(x, t)$ (external source) and $g = g(x)$ (initial or Cauchy data), determine $u = u(x, t)$ such that:*

$$\begin{cases} u_t - Du_{xx} = f & 0 < x < L, 0 < t < T \\ u(x, 0) = g(x) & 0 \leq x \leq L \\ + \text{boundary conditions} & 0 < t \leq T \end{cases}$$

where the boundary conditions may be:

- *Dirichlet:*

$$u(0, t) = h_1(t), u(L, t) = h_2(t).$$

- *Neumann:*

$$-u_x(0, t) = h_1(t), u_x(L, t) = h_2(t).$$

- *Robin or radiation:*

$$-u_x(0, t) + \alpha u(0, t) = h_1(t), u_x(L, t) + \alpha u(L, t) = h_2(t) \quad (\alpha > 0),$$

or *mixed* conditions. Accordingly, we have the initial-Dirichlet problem, the initial-Neumann problem and so on. When $h_1 = h_2 = 0$, we say that the boundary conditions are **homogeneous**.

⁸ Formula (3.31) is based on *Newton's law of cooling*: the heat loss from the surface of a body is a linear function of the temperature drop $U - u$ from the surroundings to the surface. It represents a good approximation to the radiative loss from a body when $|U - u|/u \ll 1$.

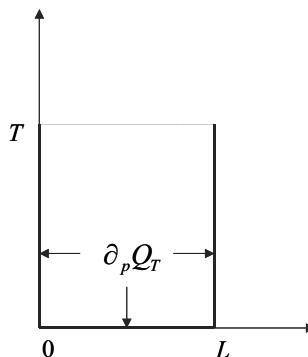


Fig. 2.1 The parabolic boundary of Q_T

Remark 2.1. Observe that only a special part of the boundary of the rectangle

$$Q_T = (0, L) \times (0, T),$$

called the **parabolic boundary** of Q_T , carries the data (see Fig. 2.1). No final condition (for $t = T, 0 < x < L$) is required.

In important applications, for instance in financial mathematics, x varies over unbounded intervals, typically $(0, \infty)$ or \mathbb{R} . In these cases one has to require that the solution does not grow too much at infinity. We will later consider the global Cauchy problem:

$$\begin{cases} u_t - Du_{xx} = f & x \in \mathbb{R}, 0 < t < T \\ u(x, 0) = g(x) & x \in \mathbb{R} \\ + \text{conditions as } x \rightarrow \pm\infty. \end{cases}$$

2.1.4 A solution by separation of variables

We will prove that, under reasonable hypotheses, the initial Dirichlet, Neumann or Robin and mixed problems are well posed. Sometimes this can be shown using elementary techniques like the *separation of variables method* that we describe below through a simple example of heat conduction. We will come back to this method from a more general point of view in Sect. 6.9.

As in the previous section, consider a bar (that we can consider one-dimensional) of length L , initially (at time $t = 0$) at constant temperature u_0 . Thereafter, the end point $x = 0$ is kept at the same temperature while the other end $x = L$ is kept at a constant temperature $u_1 > u_0$. We want to know how the temperature evolves inside the bar.

Before making any computations, let us try to conjecture what could happen. Given that $u_1 > u_0$, heat starts flowing from the hotter end, raising the temperature inside the bar and causing a heat outflow into the cold boundary. On the

other hand, the interior increase of temperature causes the hot inflow to decrease in time, while the outflow increases. We expect that sooner or later the two fluxes balance each other and that the temperature eventually reaches a steady state distribution. It would also be interesting to know how fast the steady state is reached.

We show that this is exactly the behavior predicted by our mathematical model, given by the heat equation

$$u_t - Du_{xx} = 0, \quad t > 0, 0 < x < L$$

with the initial-Dirichlet conditions

$$\begin{aligned} u(x, 0) &= u_0, & 0 \leq x \leq L \\ u(0, t) &= u_0, \quad u(L, t) = u_1, & t > 0. \end{aligned}$$

Since we are interested in the long term behavior of our solution, we leave t unlimited. Notice the *jump discontinuity* between the initial and the boundary data at $x = L$; we will take care of this little difficulty later.

- *Dimensionless variables.* First of all we introduce dimensionless variables, that is variables *independent of the units of measurement*. To do that we rescale space, time and temperature with respect to quantities that are characteristic of our problem. For the space variable we can use the length L of the bar as rescaling factor, setting

$$y = \frac{x}{L}$$

which is clearly dimensionless, being a ratio of lengths. Notice that $0 \leq y \leq 1$. How can we rescale time? Observe that the physical dimensions of the diffusion coefficient D are

$$[\text{length}]^2 \times [\text{time}]^{-1}.$$

Thus the constant $\tau = L^2/D$ gives a characteristic time scale for our diffusion problem. Therefore we introduce the dimensionless time

$$s = \frac{t}{\tau}. \tag{2.11}$$

Finally, we rescale the temperature by setting

$$z(y, s) = \frac{u(Ly, \tau s) - u_0}{u_1 - u_0}.$$

For the dimensionless temperature z we have:

$$\begin{aligned} z(y, 0) &= \frac{u(Ly, 0) - u_0}{u_1 - u_0} = 0, \quad 0 \leq y \leq 1 \\ z(0, s) &= \frac{u(0, \tau s) - u_0}{u_1 - u_0} = 0, \quad z(1, s) = \frac{u(L, \tau s) - u_0}{u_1 - u_0} = 1. \end{aligned}$$

Moreover

$$(u_1 - u_0)z_s = \frac{\partial t}{\partial s}u_t = \tau u_t = \frac{L^2}{D}u_t$$

$$(u_1 - u_0)z_{yy} = \left(\frac{\partial x}{\partial y}\right)^2 u_{xx} = L^2 u_{xx}.$$

Hence, since $u_t = Du_{xx}$,

$$(u_1 - u_0)(z_s - z_{yy}) = \frac{L^2}{D}u_t - L^2u_{xx} = \frac{L^2}{D}Du_{xx} - L^2u_{xx} = 0.$$

In conclusion, we find

$$z_s - z_{yy} = 0 \quad (2.12)$$

with the initial condition

$$z(y, 0) = 0 \quad (2.13)$$

and the boundary conditions

$$z(0, s) = 0, \quad z(1, s) = 1. \quad (2.14)$$

We see that, in the dimensionless formulation, the parameters L and D have disappeared, emphasizing the mathematical essence of the problem. On the other hand, we will later show the relevance of the dimensionless variables in test modelling.

- *The steady state solution.* We start solving problem (2.12), (2.13), (2.14) by first determining the steady state solution z^{St} , that satisfies the equation $z_{yy} = 0$ and the boundary conditions (2.14). An elementary computation gives

$$z^{St}(y) = y.$$

In terms of the original variables the steady state solution is

$$u^{St}(x) = u_0 + (u_1 - u_0) \frac{x}{L}$$

corresponding to a uniform heat flux along the bar given by the Fourier law:

$$\text{heat flux} = -\kappa u_x = -\kappa \frac{(u_1 - u_0)}{L}.$$

- *The transient regime.* Knowing the steady state solution, it is convenient to introduce the function

$$U(y, s) = z^{St}(y, s) - z(y, s) = y - z(y, s).$$

Since we expect our solution to eventually reach the steady state, U represents a *transient regime* that should converge to zero as $s \rightarrow \infty$. Furthermore, the rate of convergence to zero of U gives information on how fast the temperature reaches

its equilibrium distribution. U satisfies (2.12) with initial condition

$$U(y, 0) = y \quad (2.15)$$

and *homogeneous* boundary conditions

$$U(0, s) = 0 \quad \text{and} \quad U(1, s) = 0. \quad (2.16)$$

- *The method of separation of variables.* We are now in a position to find an explicit formula for U using the method of separation of variables. The main idea is to exploit the linear nature of the problem constructing the solution by superposition of simpler solutions of the form $w(s)v(y)$ in which the variables s and y appear in *separated form*. We emphasize that the reduction to **homogeneous boundary conditions** is crucial to carry on the computations.

Step 1. We look for non-trivial solutions to (2.12) of the form

$$U(y, s) = w(s)v(y)$$

with $v(0) = v(1) = 0$. By substitution into (2.12) we find

$$0 = U_s - U_{yy} = w'(s)v(y) - w(s)v''(y)$$

from which, separating the variables,

$$\frac{w'(s)}{w(s)} = \frac{v''(y)}{v(y)}. \quad (2.17)$$

Now, the left hand side in (2.17) is a function of s only, while the right hand side is a function of y only and the equality must hold for every $s > 0$ and every $y \in (0, L)$. This is possible only when both sides are equal to a common constant λ , say. Hence we have

$$v''(y) - \lambda v(y) = 0 \quad (2.18)$$

with

$$v(0) = v(1) = 0 \quad (2.19)$$

and

$$w'(s) - \lambda w(s) = 0. \quad (2.20)$$

Step 2. We first solve problem (2.18), (2.19). There are three different possibilities for the general solution of (2.18):

- If $\lambda = 0$,

$$v(y) = A + By \quad (A, B \text{ arbitrary constants})$$

and the conditions (2.19) imply $A = B = 0$.

b) If λ is a positive real number, say $\lambda = \mu^2 > 0$, then

$$v(y) = Ae^{-\mu y} + Be^{\mu y}$$

and again it is easy to check that the conditions (2.19) imply $A = B = 0$.

c) Finally, if $\lambda = -\mu^2 < 0$, then

$$v(y) = A \sin \mu y + B \cos \mu y.$$

From (2.19) we get

$$\begin{aligned} v(0) &= B = 0 \\ v(1) &= A \sin \mu + B \cos \mu = 0 \end{aligned}$$

from which

$$A \text{ arbitrary}, B = 0, \mu_m = m\pi, m = 1, 2, \dots .$$

Thus, only in case c) we find non-trivial solutions, given by

$$v_m(y) = A \sin m\pi y. \quad (2.21)$$

In this context, (2.18), (2.19) is called an *eigenvalue problem*; the special values μ_m are the *eigenvalues* and the solutions v_m are the corresponding *eigenfunctions*. With $\lambda = -\mu_m^2 = -m^2\pi^2$, the general solution of (2.20) is

$$w_m(s) = Ce^{-m^2\pi^2 s} \quad (C \text{ arbitrary constant}). \quad (2.22)$$

From (2.21) and (2.22) we obtain damped sinusoidal waves of the form

$$U_m(y, s) = A_m e^{-m^2\pi^2 s} \sin m\pi y.$$

Step 3. Although the solutions U_m satisfy the homogeneous Dirichlet conditions, they do not match the initial condition $U(y, 0) = y$. As we already mentioned, we try to construct the correct solution by superposing the U_m , that is, by setting

$$U(y, s) = \sum_{m=1}^{\infty} A_m e^{-m^2\pi^2 s} \sin m\pi y. \quad (2.23)$$

Some questions arise:

Q1. The initial condition requires

$$U(y, 0) = \sum_{m=1}^{\infty} A_m \sin m\pi y = y \quad \text{for } 0 \leq y \leq 1. \quad (2.24)$$

Is it possible to choose the coefficients A_m in order to satisfy (2.24)? In which sense does U attain the initial data? For instance, is it true, in some sense, that

$$U(z, s) \rightarrow y \quad \text{if} \quad (z, s) \rightarrow (y, 0)?$$

Q2. Any finite linear combination of the U_m is a solution of the heat equation; can we make sure that the same is true for U ? The answer is positive if we could differentiate term by term the infinite sum and get

$$(\partial_s - \partial_{yy})U(y, s) = \sum_{m=1}^{\infty} (\partial_s - \partial_{yy})U_m(y, s) = 0. \quad (2.25)$$

What about the boundary conditions?

Q3. Even if we have a positive answer to questions 1 and 2, are we confident that U is the unique solution of our problem and therefore that it describes the correct evolution of the temperature?

Q1. Question 1 is rather general and concerns the *Fourier series expansion*⁹ of a function, in particular of the initial data $f(y) = y$, in the interval $(0, 1)$. Due to the homogeneous Dirichlet conditions it is natural to expand $f(y) = y$ in a *sine Fourier series*, that is to expand the *2-periodic* and *odd* function that agrees with y in the interval $(-1, 1)$. The Fourier coefficients are given by the formulas

$$\begin{aligned} A_m &= 2 \int_0^1 y \sin m\pi y \, dy = -\frac{2}{m\pi} [y \cos m\pi y]_0^1 + \frac{2}{m\pi} \int_0^1 \cos m\pi y \, dy = \\ &= -2 \frac{\cos m\pi}{m\pi} = (-1)^{m+1} \frac{2}{m\pi}. \end{aligned}$$

The sine Fourier expansion of $f(y) = y$ on the interval $(0, 1)$ is therefore

$$y = \sum_{m=1}^{\infty} (-1)^{m+1} \frac{2}{m\pi} \sin m\pi y. \quad (2.26)$$

Where is the expansion (2.26) valid? It cannot be true at $y = 1$, since $\sin m\pi = 0$ for every m , and we would obtain $1 = 0$. This clearly reflects the presence of the jump discontinuity of the data at $y = 1$.

The theory of Fourier series implies that (2.26) is true at every point $y \in [0, 1]$ and that the series converges uniformly in every interval $[0, a]$, $a < 1$. Moreover, equality (2.26) holds **in the quadratic mean sense** (or $L^2(0, 1)$ sense), that is

$$\int_0^1 [y - \sum_{m=1}^N (-1)^{m+1} \frac{2}{m\pi} \sin m\pi y]^2 dy \rightarrow 0 \quad \text{as } N \rightarrow \infty.$$

⁹ See Appendix A.

From (2.23) and the expression of A_m , we obtain the *formal* solution

$$U(y, s) = \sum_{m=1}^{\infty} (-1)^{m+1} \frac{2}{m\pi} e^{-m^2\pi^2 s} \sin m\pi y \quad (2.27)$$

that attains the initial data in the least squares sense, i.e.

$$\lim_{s \rightarrow 0^+} \int_0^1 [U(y, s) - y]^2 dy = 0. \quad (2.28)$$

In fact, from Parseval's equality (A.8), we can write

$$\int_0^1 [U(y, s) - y]^2 dy = \frac{4}{\pi^2} \sum_{m=1}^{\infty} \frac{(e^{-m^2\pi^2 s} - 1)^2}{m^2}. \quad (2.29)$$

Since for $s \geq 0$

$$\frac{(e^{-m^2\pi^2 s} - 1)^2}{m^2} \leq \frac{1}{m^2}$$

and the series $\sum 1/m^2$ converges, then the series in (2.29) converges uniformly by Weierstrass test (see Sect. 1.4) in $[0, \infty)$ and we can take the limit under the sum, obtaining (2.28).

Q2. The analytical expression of U is rather reassuring: it is a superposition of sinusoids of increasing frequency m and of strongly damped amplitude, because of the negative exponential, at least when $s > 0$. Indeed, for $s > 0$, the rapid convergence to zero of each term and its derivatives in the series (2.27) allows us to differentiate term by term. Precisely, we have

$$\frac{\partial U_m}{\partial s} = \frac{\partial^2 U_m}{\partial y^2} = (-1)^m 2m\pi e^{-m^2\pi^2 s} \sin m\pi y,$$

so that, if $s \geq s_0 > 0$,

$$\left| \frac{\partial U_m}{\partial s} \right|, \left| \frac{\partial^2 U_m}{\partial y^2} \right| \leq 2m\pi e^{-m^2\pi^2 s_0}.$$

Since the numerical series

$$\sum_{m=1}^{\infty} m e^{-m^2\pi^2 s_0}$$

is convergent, we conclude by the Weierstrass test that the series

$$\sum_{m=1}^{\infty} \frac{\partial U_m}{\partial s} \quad \text{and} \quad \sum_{m=1}^{\infty} \frac{\partial^2 U_m}{\partial y^2}$$

converge uniformly in $[0, 1] \times [s_0, \infty)$. Thus (2.25) is true and U is a solution of (2.12).

It remains to check the Dirichlet conditions: if $s_0 > 0$,

$$U(z, s) \rightarrow 0 \quad \text{as } (z, s) \rightarrow (0, s_0) \text{ or } (z, s) \rightarrow (1, s_0).$$

This is true because we can take the two limits under the sum, due to the uniform convergence of the series (2.27) in any region $[0, 1] \times (b, +\infty)$ with $b > 0$. For the same reason, U has continuous derivatives of any order, up to the lateral boundary of the strip $[0, 1] \times (b, +\infty)$.

Note, in particular, that U *immediately* forgets the initial discontinuity and becomes smooth at any positive time.

Q3. To show that U is indeed the unique solution, we use the so-called *energy method*, that we will develop later in greater generality. Suppose W is another solution of problem (2.12), smooth in any region $[0, 1] \times (b, \infty)$ with $b > 0$ and satisfying (2.18). Then, by linearity,

$$v = U - W$$

satisfies

$$v_s - v_{yy} = 0 \tag{2.30}$$

and has zero initial-boundary data. Multiplying (2.30) by v , integrating in y over the interval $[0, 1]$ and keeping $s > 0$, fixed, we get

$$\int_0^1 vv_s \, dy - \int_0^1 vv_{yy} \, dy = 0. \tag{2.31}$$

Observe that

$$\int_0^1 vv_s \, dy = \frac{1}{2} \int_0^1 \partial_s(v^2) \, dy = \frac{1}{2} \frac{d}{ds} \int_0^1 v^2 \, dy. \tag{2.32}$$

Moreover, integrating by parts we can write

$$\begin{aligned} \int_0^1 vv_{yy} \, dy &= [v(1, s)v_y(1, s) - v(0, s)v_y(0, s)] - \int_0^1 (v_y)^2 \, dy \\ &= - \int_0^1 (v_y)^2 \, dy \end{aligned} \tag{2.33}$$

since $v(1, s) = v(0, s) = 0$. From (2.31), (2.32) and (2.33) we get

$$\frac{1}{2} \frac{d}{ds} \int_0^1 v^2 \, dy = - \int_0^1 (v_y)^2 \, dy \leq 0 \tag{2.34}$$

and therefore, the *nonnegative* function

$$E(s) = \int_0^1 v^2(y, s) \, dy$$

is non-increasing. On the other hand, using (2.28) for v instead of U , we get

$$E(s) \rightarrow 0 \quad \text{as } s \rightarrow 0$$

which forces $E(s) = 0$, for every $s > 0$. But v^2 is nonnegative and continuous in $[0, 1]$ if $s > 0$, so that it must be $v(y, s) = 0$ for $y \in [0, 1]$ and every $s > 0$. Then $U = W$.

- *Back to the original variables.* In terms of the original variables, our solution is expressed as

$$u(x, t) = u_0 + (u_1 - u_0) \frac{x}{L} - (u_1 - u_0) \sum_{m=1}^{\infty} (-1)^{m+1} \frac{2}{m\pi} e^{\frac{-m^2\pi^2 D}{L^2}t} \sin \frac{m\pi}{L} x.$$

This formula confirms our initial guess about the evolution of the temperature towards the steady state. Indeed, each term of the series converges to zero exponentially as $t \rightarrow +\infty$ and it is not difficult to show¹⁰ that

$$u(x, t) \rightarrow u_0 + (u_1 - u_0) \frac{x}{L} \quad \text{as } t \rightarrow +\infty.$$

Moreover, the first exponential ($m = 1$) in the series decays much more slowly than the others and very soon it determines the main deviation of u from the equilibrium, *independently of the initial condition*. Thus, as $t \rightarrow +\infty$, the leading term is the damped sinusoid

$$\frac{2}{\pi} e^{\frac{-\pi^2 D}{L^2}t} \sin \frac{\pi}{L} x.$$

In this mode there is a concentration of heat at $x = L/2$, where the temperature reaches its maximum amplitude $2 \exp(-\pi^2 Dt/L^2)/\pi$. At time $t = L^2/D$ the amplitude decays to $2 \exp(-\pi^2)/\pi \simeq 3.3 \times 10^{-5}$, about 0.005 per cent of its initial value. This simple calculation shows that to reach the steady state, a time of order L^2/D is required, a fundamental fact in heat diffusion.

Not surprisingly, the scaling factor in (2.11) was exactly $\tau = L^2/D$. The dimensionless formulation is extremely useful in experimental modelling tests. To achieve reliable results, these models must reproduce the same characteristics at different scales. For instance, if our bar were an experimental model of a much bigger beam of length L_0 and diffusion coefficient D_0 , to reproduce the same heat diffusion effects, we must choose material (D) and length (L) for our model bar such that

$$\frac{L^2}{D} = \frac{L_0^2}{D_0}.$$

Figure 2.2 shows the solution of the dimensionless problem (2.12), (2.15), (2.16) for $0 < t \leq 1$.

¹⁰ The Weierstrass test works here for $t \geq t_0 > 0$.

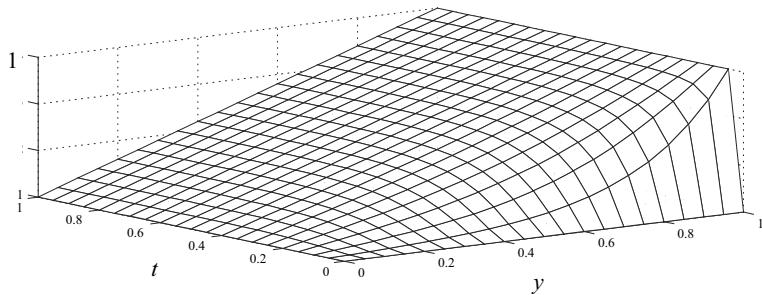


Fig. 2.2 The solution to the dimensionless problem (2.12), (2.13), (2.14)

2.1.5 Problems in dimension $n > 1$

The formulation of the well posed problems in Subsect. 2.1.3 can be easily generalized to any spatial dimension $n > 1$, in particular to $n = 2$ or $n = 3$. Suppose we want to determine the evolution of the temperature in a heat conducting body that occupies a bounded domain¹¹ $\Omega \subset \mathbb{R}^n$, during an interval of time $[0, T]$. Under the hypotheses of Subsect. 2.1.2, the temperature is a function $u = u(\mathbf{x}, t)$ that satisfies the heat equation $u_t - D\Delta u = f$ in the *space-time cylinder* (see Fig. 2.3)

$$Q_T = \Omega \times (0, T).$$

To select a unique solution we have to prescribe, first of all, the *initial distribution*

$$u(\mathbf{x}, 0) = g(\mathbf{x}), \quad \mathbf{x} \in \overline{\Omega},$$

where $\overline{\Omega} = \Omega \cup \partial\Omega$ denotes the *closure* of Ω .

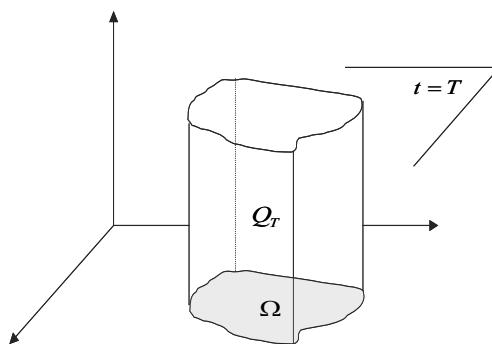


Fig. 2.3 The space-time cylinder Q_T

¹¹ Recall that by a *domain* we mean an *open connected set* in \mathbb{R}^n .

The control of the interaction of the body with the surroundings is modeled through *suitable conditions* on $\partial\Omega$. The most common ones are:

Dirichlet condition: the temperature is kept at a prescribed level on $\partial\Omega$; this amounts to assigning

$$u(\sigma, t) = h(\sigma, t), \quad \sigma \in \partial\Omega \text{ and } t \in (0, T].$$

Neumann condition: the heat flux through $\partial\Omega$ is assigned. To model this condition, we assume that the boundary $\partial\Omega$ is a smooth curve or surface, having a tangent line or plane at every point¹² with *outward* unit vector ν . From Fourier law we have

$$\mathbf{q} = \text{heat flux} = -\kappa \nabla u$$

so that the *inward heat flux* is

$$-\mathbf{q} \cdot \nu = \kappa \nabla u \cdot \nu = \kappa \partial_\nu u.$$

Thus the Neumann condition reads

$$\partial_\nu u(\sigma, t) = h(\sigma, t), \quad \sigma \in \partial\Omega \text{ and } t \in (0, T].$$

Radiation or Robin condition: the *inward* heat flux through $\partial\Omega$ depends linearly on the difference¹³ $U - u$:

$$-\mathbf{q} \cdot \nu = \gamma(U - u), \quad (\gamma > 0)$$

where U is the surroundings temperature. From the Fourier law we obtain

$$\partial_\nu u + \alpha u = h, \quad \text{on } \partial\Omega \times (0, T]$$

with $\alpha = \gamma/\kappa > 0$, $h = \gamma U/\kappa$.

Mixed conditions: the boundary of Ω is decomposed into various parts where different boundary conditions are prescribed. For instance, a formulation of a mixed Dirichlet-Neumann problem is obtained by writing

$$\partial\Omega = \overline{\Gamma}_D \cup \overline{\Gamma}_N \quad \text{with} \quad \Gamma_D \cap \Gamma_N = \emptyset,$$

where Γ_D and Γ_N are regular open subsets of $\partial\Omega$ in the sense of Definition 1.5, p. 14. Then we assign

$$\begin{aligned} u &= h_1 \text{ on } \Gamma_D \times (0, T] \\ \partial_\nu u &= h_2 \text{ on } \Gamma_N \times (0, T]. \end{aligned}$$

¹² We can also allow boundaries with corner points, like squares, cones, or edges, like cubes. It is enough that the set of points where the tangent plane does not exist has zero surface measure (zero length in two dimensions). Lipschitz domains have this property (see Sect. 1.5).

¹³ Linear Newton law of cooling.

Summarizing, we have the following typical problems: *given $f = f(\mathbf{x}, t)$ and $g = g(\mathbf{x})$, determine $u = u(\mathbf{x}, t)$ such that:*

$$\begin{cases} u_t - D\Delta u = f & \text{in } Q_T \\ u(\mathbf{x}, 0) = g(\mathbf{x}) & \text{in } \overline{\Omega} \\ + \text{boundary conditions on } \partial\Omega \times (0, T] \end{cases}$$

where the boundary conditions could be:

- *Dirichlet:*

$$u = h.$$

- *Neumann:*

$$\partial_{\mathbf{\nu}} u = h.$$

- *Radiation or Robin:*

$$\partial_{\mathbf{\nu}} u + \alpha u = h \quad (\alpha > 0).$$

- *Mixed, for instance:*

$$u = h_1 \text{ on } \Gamma_D, \quad \partial_{\mathbf{\nu}} u = h_2 \text{ on } \Gamma_N.$$

Also in dimension $n > 1$, the *global Cauchy problem* is important:

$$\begin{cases} u_t - D\Delta u = f & \mathbf{x} \in \mathbb{R}^n, 0 < t < T \\ u(\mathbf{x}, 0) = g(\mathbf{x}) & \mathbf{x} \in \mathbb{R}^n \\ + \text{condition as } |\mathbf{x}| \rightarrow \infty. \end{cases}$$

We again emphasize that no final condition (for $t = T, \mathbf{x} \in \Omega$) is required. The data is assigned on the *parabolic boundary* $\partial_p Q_T$ of Q_T , given by the union of the bottom points $\overline{\Omega} \times \{t = 0\}$ and the side points $S_T = \partial\Omega \times (0, T]$:

$$\partial_p Q_T = (\overline{\Omega} \times \{t = 0\}) \cup S_T.$$

2.2 Uniqueness and Maximum Principles

2.2.1 Integral method

Generalizing the energy method used in Subsect. 2.1.4, it is easy to show that all the problems we have formulated in the previous section have at most one solution under reasonable conditions on the data. Suppose u and v are solutions of one of those problems, sharing the same boundary conditions, and let $w = u - v$; we want to show that $w \equiv 0$. For the time being we do not worry about the precise hypotheses on u and v ; we assume they are sufficiently smooth in Q_T up to $\partial_p Q_T$ and observe that w satisfies the homogeneous equation

$$w_t - D\Delta w = 0 \tag{2.35}$$

in $Q_T = \Omega \times (0, T)$, with initial condition

$$w(\mathbf{x}, 0) = 0$$

in $\overline{\Omega}$, and one of the following conditions on $\partial\Omega \times (0, T]$:

$$w = 0 \quad (\text{Dirichlet}) \quad (2.36)$$

or

$$\partial_{\nu} w = 0 \quad (\text{Neumann}) \quad (2.37)$$

or

$$\partial_{\nu} w + \alpha w = 0 \quad \alpha > 0, \quad (\text{Robin}) \quad (2.38)$$

or

$$w = 0 \text{ on } \partial_D \Omega, \quad \partial_{\nu} w = 0 \text{ on } \partial_N \Omega \quad (\text{mixed}). \quad (2.39)$$

Multiply equation (2.35) by w and integrate on Ω ; we find

$$\int_{\Omega} w w_t d\mathbf{x} = D \int_{\Omega} w \Delta w d\mathbf{x}.$$

Now,

$$\int_{\Omega} w w_t d\mathbf{x} = \frac{1}{2} \frac{d}{dt} \int_{\Omega} w^2 d\mathbf{x} \quad (2.40)$$

and from Green's identity (1.13) with $u = v = w$,

$$\int_{\Omega} w \Delta w d\mathbf{x} = \int_{\partial\Omega} w \partial_{\nu} w d\sigma - \int_{\Omega} |\nabla w|^2 d\mathbf{x}. \quad (2.41)$$

Then, letting

$$E(t) = \int_{\Omega} w^2 d\mathbf{x},$$

(2.40) and (2.41) give

$$\frac{1}{2} E'(t) = D \int_{\partial\Omega} w \partial_{\nu} w d\sigma - D \int_{\Omega} |\nabla w|^2 d\mathbf{x}.$$

If Robin condition (2.38) holds,

$$\int_{\partial\Omega} w \partial_{\nu} w d\sigma = -\alpha \int_{\Omega} w^2 d\mathbf{x} \leq 0.$$

If one of the (2.36), (2.37), (2.39) holds, then

$$\int_{\partial\Omega} w \partial_{\nu} w d\sigma = 0.$$

In any case it follows that $E'(t) \leq 0$ and therefore E is a nonincreasing function. Since

$$E(0) = \int_{\Omega} w^2(\mathbf{x}, 0) d\mathbf{x} = 0,$$

we must have $E(t) = 0$ for every $t \geq 0$, and this implies $w(\mathbf{x}, t) \equiv 0$ in Ω for every $t > 0$. Thus $u = v$.

The above calculations are completely justified if Ω is a sufficiently smooth domain¹⁴ and, for instance, we require that u and v are continuous in $\overline{Q}_T = \overline{\Omega} \times [0, T]$, together with their first and second spatial derivatives and their first order time derivatives. We denote the set of these functions by the symbol $C^{2,1}(\overline{Q}_T)$ and summarize everything in the following statement.

Theorem 2.2. *The initial Dirichlet, Neumann, Robin and mixed problems have at most one solution belonging to $C^{2,1}(\overline{Q}_T)$.*

2.2.2 Maximum principles

The fact that heat flows from higher to lower temperature regions implies that a solution of the homogeneous heat equation attains its maximum and minimum values on $\partial_p Q_T$. This result is known as the *maximum principle* and reflects an aspect of the time irreversibility of the phenomena described by the heat equation, in the sense that the future cannot have an influence on the past (*causality principle*). In other words, the value of a solution u at time t is independent of any change of the data after t .

The following simple theorem translates these principles and we state it for functions in the class $C^{2,1}(Q_T) \cap C(\overline{Q}_T)$. These functions are continuous up to the boundary of Q_T , with derivatives continuous in the interior of Q_T . It is convenient to distinguish between subsolution and supersolution of the heat equation, according to the following definition.

Definition 2.3. *A function $w \in C^{2,1}(Q_T)$ such that $w_t - D\Delta w \leq 0$ (≥ 0) in Q_T is called a subsolution (supersolution) of the diffusion equation.*

We have:

Theorem 2.4. *Let $w \in C^{2,1}(Q_T) \cap C(\overline{Q}_T)$ such that*

$$w_t - D\Delta w = q(\mathbf{x}, t) \leq 0 \quad (\text{resp. } \geq 0) \text{ in } Q_T. \quad (2.42)$$

Then w attains its maximum (resp. minimum) on $\partial_p Q_T$:

$$\max_{\overline{Q}_T} w = \max_{\partial_p Q_T} w \quad (\text{resp. } \min_{\overline{Q}_T} w = \min_{\partial_p Q_T} w). \quad (2.43)$$

In particular, if w is negative (resp. positive) on $\partial_p Q_T$, then is negative (resp. positive) in all Q_T .

¹⁴ C^1 or even Lipschitz domains, for instance (see Sect. 1.5).

Proof. We recall that $\partial_p Q_T$ is the union of the base and the lateral boundary of Q_T . Let $q \leq 0$. The case $q \geq 0$ is analogous. We consider two cases.

Case 1. Assume first that $w \in C^2(\overline{Q}_T)$ and that $q(\mathbf{x}, t) < 0$ for all (\mathbf{x}, t) in \overline{Q}_T . We argue by contradiction. Suppose that w attains its maximum at a point $(\mathbf{x}_0, t_0) \notin \partial_p Q_T$. Then $\mathbf{x}_0 \in \Omega$ and $t_0 \in (0, T]$. From elementary calculus, we have

$$w_{x_j x_j}(\mathbf{x}_0, t_0) \leq 0$$

for every $j = 1, \dots, n$, so that

$$\Delta w(\mathbf{x}_0, t_0) \leq 0,$$

and either

$$w_t(\mathbf{x}_0, t_0) = 0 \quad \text{if } t_0 < T$$

or

$$w_t(\mathbf{x}_0, T) \geq 0.$$

In both cases, we obtain the contradiction

$$0 \leq w_t(\mathbf{x}_0, t_0) - \Delta w(\mathbf{x}_0, t_0) = q(\mathbf{x}_0, t_0) < 0.$$

Case 2. Consider now the general case $w \in C^2(Q_T) \cap C(\overline{Q}_T)$ and $q \leq 0$ in \overline{Q}_T . We reduce to Case 1. Let $\varepsilon \in (0, T)$ and set $u = w - \varepsilon t$. Then

$$u_t - D\Delta u = q - \varepsilon < 0 \tag{2.44}$$

and $u \in C^2(\overline{Q}_{T-\varepsilon})$. From Case 1 we deduce that

$$\max_{\overline{Q}_{T-\varepsilon}} u \leq \max_{\partial_p Q_{T-\varepsilon}} u \leq \max_{\partial_p Q_T} u. \tag{2.45}$$

Since $u \leq w \leq u + \varepsilon t$, we infer from (2.45),

$$\max_{\overline{Q}_{T-\varepsilon}} w \leq \max_{\overline{Q}_{T-\varepsilon}} u + \varepsilon T \leq \max_{\partial_p Q_T} u + \varepsilon T \leq \max_{\partial_p Q_T} w + \varepsilon T. \tag{2.46}$$

Since w is continuous in \overline{Q}_T , we deduce that

$$\max_{\overline{Q}_{T-\varepsilon}} w \rightarrow \max_{\overline{Q}_T} w \quad \text{as } \varepsilon \rightarrow 0.$$

Hence, letting $\varepsilon \rightarrow 0$ in (2.46) we find $\max_{\overline{Q}_T} w \leq \max_{\partial_p Q_T} w$, which concludes the proof. \square

As an immediate consequence of Theorem 2.4, we have that, if

$$w_t - D\Delta w = 0 \quad \text{in } Q_T,$$

then w attains its maximum and its minimum on $\partial_p Q_T$. In particular,

$$\min_{\partial_p Q_T} w \leq w(\mathbf{x}, t) \leq \max_{\partial_p Q_T} w \quad \text{for every } (\mathbf{x}, t) \in Q_T.$$

Moreover (for the proof see Problem 2.6):

Corollary 2.5 (Comparison and stability). *Let $v, w \in C^{2,1}(Q_T) \cap C(\overline{Q}_T)$ be solutions of*

$$v_t - D\Delta v = f_1 \quad \text{and} \quad w_t - D\Delta w = f_2$$

with f_1, f_2 bounded in Q_T . Then:

- a) *If $v \geq w$ on $\partial_p Q_T$ and $f_1 \geq f_2$ in Q_T then $v \geq w$ in all Q_T .*
- b) *The following stability estimate holds:*

$$\max_{\overline{Q}_T} |v - w| \leq \max_{\partial_p Q_T} |v - w| + T \sup_{\overline{Q}_T} |f_1 - f_2|. \quad (2.47)$$

Corollary 2.5 gives uniqueness for the initial-Dirichlet problem under much less restrictive hypotheses than those in Theorem 2.2: indeed it does not require the continuity of any derivatives of the solution up to $\partial_p Q_T$.

Inequality (2.47) is a uniform pointwise stability estimate, extremely useful in several applications. In fact if

$$v = g_1, \quad w = g_2 \quad \text{on } \partial_p Q_T$$

and

$$\max_{\partial_p Q_T} |g_1 - g_2| \leq \varepsilon \quad \text{and} \quad \sup_{\overline{Q}_T} |f_1 - f_2| \leq \varepsilon,$$

we deduce

$$\max_{\overline{Q}_T} |v - w| \leq \varepsilon(1 + T).$$

Thus, in finite time, a small uniform distance between the data implies small uniform distance between the corresponding solutions.

Theorem 2.4 is a version of the so called weak maximum principle, weak because this result says nothing about the possibility that a solution achieves its maximum or minimum at an interior point as well.

Actually a more precise result is the following, known as *strong maximum principle*¹⁵ (see Fig. 2.4).

Theorem 2.6. *Let $u \in C^{2,1}(Q_T) \cap C(\overline{Q}_T)$ be a subsolution (resp. supersolution) of the heat equation in Q_T . If u attains its maximum M (resp. minimum) at a point (x_1, t_1) with $x_1 \in \Omega$, $0 < t_1 \leq T$, then*

$$u = M \quad \text{in } \overline{Q}_{t_1}.$$

¹⁵ We omit the rather technical proof. See [14], M. Protter, H. Weinberger, 1984.

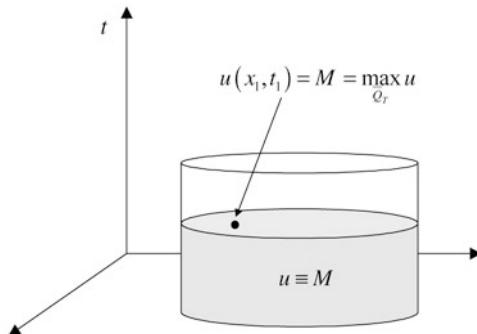


Fig. 2.4 The strong maximum principle

2.3 The Fundamental Solution

There are privileged solutions of the diffusion equation that can be used to construct many other ones. In this section we are going to discover one of these special building blocks, the most important one. First we point out some features of the heat equation.

2.3.1 Invariant transformations

The homogeneous diffusion equation has simple but important properties. Let $u = u(\mathbf{x}, t)$ be a solution of

$$u_t - D\Delta u = 0. \quad (2.48)$$

- *Time reversal.* The function

$$v(\mathbf{x}, t) = u(\mathbf{x}, -t),$$

obtained by the change of variable $t \mapsto -t$, is a solution of the **adjoint** or **backward** equation.

$$v_t + D\Delta v = 0.$$

Coherently, the (2.48) is sometimes called the **forward** equation. The non-invariance of (2.48) with respect to a change of sign in time is another aspect of time irreversibility.

- *Space and time translation invariance.* For \mathbf{y}, s fixed, the function

$$v(\mathbf{x}, t) = u(\mathbf{x} - \mathbf{y}, t - s),$$

is still a solution of (2.48). Clearly, for x, t fixed the function $u(\mathbf{x} - \mathbf{y}, t - s)$ is a solution of the backward equation with respect to y and s .

- *Parabolic dilations.* The transformation

$$\mathbf{x} \mapsto a\mathbf{x}, \quad t \mapsto bt, \quad u \mapsto cu \quad (a, b, c > 0)$$

represents a dilation (expansion or contraction) of the graph of u . Let us check for which values of a, b, c the function

$$u^*(\mathbf{x}, t) = cu(a\mathbf{x}, bt)$$

is still a solution of (2.48). We have:

$$u_t^*(\mathbf{x}, t) - D\Delta u^*(\mathbf{x}, t) = cbu_t(a\mathbf{x}, bt) - ca^2D\Delta u(a\mathbf{x}, bt)$$

and so u^* is a solution of (2.48) if

$$b = a^2. \quad (2.49)$$

The relation (2.49) suggests the name of *parabolic dilation* for the transformation

$$\mathbf{x} \mapsto a\mathbf{x} \quad t \mapsto a^2t \quad (a, b > 0).$$

Under this transformation the expressions

$$\frac{|\mathbf{x}|^2}{Dt} \quad \text{or} \quad \frac{\mathbf{x}}{\sqrt{Dt}}$$

are left unchanged. Moreover, we already observed that they are *dimensionless groups*. Thus it is not surprising that these combinations of the independent variables occur frequently in the study of diffusion phenomena.

- *Dilations and conservation of mass (or energy).* Let $u = u(\mathbf{x}, t)$ be a solution of (2.48) in the half-space $\mathbb{R}^n \times (0, +\infty)$. Then we just checked that the function

$$u^*(\mathbf{x}, t) = cu(a\mathbf{x}, a^2t) \quad (a > 0)$$

is also a solution in the same set. Suppose u satisfies the condition

$$\int_{\mathbb{R}^n} u(\mathbf{x}, t) d\mathbf{x} = q \quad \text{for every } t > 0. \quad (2.50)$$

If, for instance, u represents the concentration of a substance (density of mass), equation (2.50) states that the total mass is q at every time t . If u is a temperature, (2.50) says that the total internal energy is constant ($= q\rho c_v$). We ask for which a, c the solution u^* still satisfies (2.50). We have

$$\int_{\mathbb{R}^n} u^*(\mathbf{x}, t) d\mathbf{x} = c \int_{\mathbb{R}^n} u(a\mathbf{x}, a^2t) d\mathbf{x}.$$

Letting $\mathbf{y} = a\mathbf{x}$, so that $d\mathbf{y} = a^n d\mathbf{x}$, we find

$$\int_{\mathbb{R}^n} u^*(\mathbf{x}, t) d\mathbf{x} = ca^{-n} \int_{\mathbb{R}^n} u(\mathbf{y}, a^2 t) d\mathbf{y} = ca^{-n}$$

and for (2.50) to be satisfied we must have:

$$c = qa^n.$$

In conclusion, if $u = u(\mathbf{x}, t)$ is a solution of (2.48) in the half-space $\mathbb{R}^n \times (0, +\infty)$, satisfying (2.50), the same is true for

$$u^*(\mathbf{x}, t) = qa^n u(a\mathbf{x}, a^2 t). \quad (2.51)$$

2.3.2 The fundamental solution ($n = 1$)

We are now in position to construct our special solution, starting with dimension $n = 1$. To help intuition, think for instance of our solution as the concentration of a substance of total mass q and suppose we want to keep the total mass equal to q at any time.

We have seen that the combination of variables x/\sqrt{Dt} is not only invariant with respect to parabolic dilations but also dimensionless. It is then natural to check if there are solutions of (2.48) involving such dimensionless group. Since \sqrt{Dt} has the dimension of a length, the quantity q/\sqrt{Dt} is a typical order of magnitude for the concentration, so that it makes sense to look for solutions of the form

$$u^*(x, t) = \frac{q}{\sqrt{Dt}} U\left(\frac{x}{\sqrt{Dt}}\right) \quad (2.52)$$

where U is a (dimensionless) function of a single variable.

Here is the main question: is it possible to determine

$$U = U(\xi)$$

such that u^* is a solution of (2.48)? Solutions of the form (2.52) are called *similarity solutions*¹⁶.

Moreover, since we are interpreting u^* as a concentration, we require $U \geq 0$ and the total mass condition yields

$$1 = \frac{1}{\sqrt{Dt}} \int_{\mathbb{R}} U\left(\frac{x}{\sqrt{Dt}}\right) dx \underset{\xi=x/\sqrt{Dt}}{=} \int_{\mathbb{R}} U(\xi) d\xi,$$

¹⁶ A solution of a particular evolution problem is a *similarity* or *self-similar* solution if its spatial configuration (its graph at a fixed time) remains similar to itself at all times during the evolution. In one space dimension, *self-similar* solutions have the general form

$$u(x, t) = a(t) F(x/b(t))$$

where, preferably, u/a and x/b are dimensionless quantity.

so that we require that

$$\int_{\mathbb{R}} U(\xi) d\xi = 1. \quad (2.53)$$

Let us check if u^* is a solution to (2.48). We have ($\xi = x/\sqrt{Dt}$):

$$u_t^* = \frac{q}{\sqrt{D}} \left[-\frac{1}{2} t^{-\frac{3}{2}} U(\xi) - \frac{1}{2\sqrt{D}} xt^{-2} U'(\xi) \right] = -\frac{q}{2t\sqrt{Dt}} [U(\xi) + \xi U'(\xi)]$$

$$u_{xx}^* = \frac{q}{(Dt)^{3/2}} U''(\xi),$$

hence

$$u_t^* - Du_{xx}^* = -\frac{q}{t\sqrt{Dt}} \left\{ U''(\xi) + \frac{1}{2}\xi U'(\xi) + \frac{1}{2}U(\xi) \right\}.$$

We see that for u^* to be a solution of (2.48), U must be a solution in \mathbb{R} of the ordinary differential equation

$$U''(\xi) + \frac{1}{2}\xi U'(\xi) + \frac{1}{2}U(\xi) = 0. \quad (2.54)$$

Since $U \geq 0$, (2.53) implies¹⁷:

$$U(-\infty) = U(+\infty) = 0.$$

On the other hand, (2.54) is invariant with respect to the change of variables

$$\xi \mapsto -\xi$$

and therefore we look for *even solutions*:

$$U(-\xi) = U(\xi).$$

Then we can restrict ourselves to $\xi \geq 0$, asking

$$U'(0) = 0 \quad \text{and} \quad U(+\infty) = 0, \quad (2.55)$$

where $U'(0) = 0$ comes from the smoothness of U as a solution of (2.54). To solve (2.54) observe that it can be written in the form

$$\frac{d}{d\xi} \left\{ U'(\xi) + \frac{1}{2}\xi U(\xi) \right\} = 0$$

that yields

$$U'(\xi) + \frac{1}{2}\xi U(\xi) = C \quad (C \in \mathbb{R}). \quad (2.56)$$

¹⁷ Rigorously, the precise conditions are:

$$\liminf_{\xi \rightarrow \pm\infty} U(\xi) = 0.$$

Letting $\xi = 0$ in (2.56) and recalling (2.55), we deduce that $C = 0$ and therefore

$$U'(\xi) + \frac{1}{2}\xi U(\xi) = 0. \quad (2.57)$$

The general integral of (2.57) is

$$U(\xi) = c_0 e^{-\frac{\xi^2}{4}} \quad (c_0 \in \mathbb{R}).$$

This function is even, integrable and vanishes at infinity. It only remains to choose $c_0 > 0$ in order to ensure (2.53). Since¹⁸

$$\int_{\mathbb{R}} e^{-\frac{\xi^2}{4}} d\xi \Big|_{\xi=2z} = 2 \int_{\mathbb{R}} e^{-z^2} dz = 2\sqrt{\pi}$$

the choice is $c_0 = (4\pi)^{-1/2}$.

Going back to the original variables, we have found the following solution of (2.48)

$$u^*(x, t) = \frac{q}{\sqrt{4\pi D t}} e^{-\frac{x^2}{4Dt}}, \quad x \in \mathbb{R}, t > 0$$

positive, even in x , and such that

$$\int_{\mathbb{R}} u^*(x, t) dx = q \quad \text{for every } t > 0. \quad (2.58)$$

The choice $q = 1$ gives a family of *Gaussians*, parametrized with time, and it is natural to think of a *normal probability density*.

Definition 2.7. *The function*

$$\Gamma_D(x, t) = \frac{1}{\sqrt{4\pi D t}} e^{-\frac{x^2}{4Dt}}, \quad x \in \mathbb{R}, t > 0 \quad (2.59)$$

is called the **fundamental solution** of eq. (2.48).

2.3.3 The Dirac distribution

It is worthwhile to examine the behavior of the fundamental solution Γ_D . For every fixed $x \neq 0$,

$$\lim_{t \rightarrow 0^+} \Gamma_D(x, t) = \lim_{t \rightarrow 0^+} \frac{1}{\sqrt{4\pi D t}} e^{-\frac{x^2}{4Dt}} = 0 \quad (2.60)$$

¹⁸ Recall that

$$\int_{\mathbb{R}} e^{-z^2} dz = \sqrt{\pi}.$$

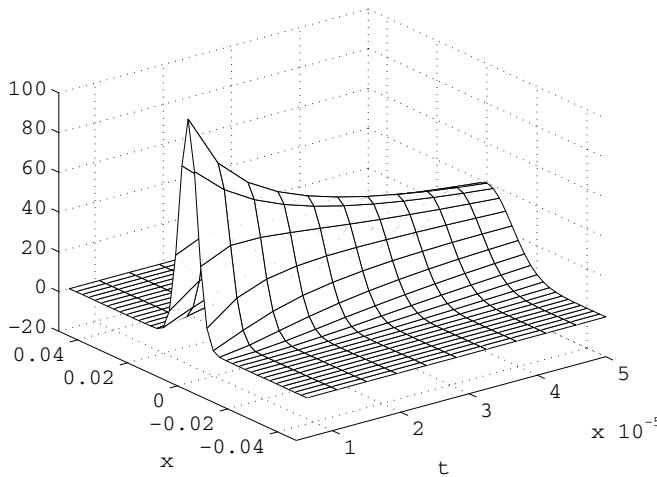


Fig. 2.5 The fundamental solution Γ_1

while

$$\lim_{t \rightarrow 0^+} \Gamma_D(0, t) = \lim_{t \rightarrow 0^+} \frac{1}{\sqrt{4\pi Dt}} = +\infty. \quad (2.61)$$

If we interpret Γ_D as a probability density, eqs. (2.60), (2.61) and (2.58) imply that when $t \rightarrow 0^+$ the fundamental solution tends to concentrate mass around the origin; eventually, the whole probability mass is concentrated at $x = 0$ (see Fig. 2.5).

The limiting density distribution can be mathematically modeled by the so called *Dirac distribution (or measure) at the origin*, denoted by the symbol δ . The Dirac distribution is not a function in the usual sense of Analysis; if it were, it should have the following properties:

- $\delta(0) = \infty$, $\delta(x) = 0$ for $x \neq 0$;
- $\int_{\mathbb{R}} \delta(x) dx = 1$,

clearly incompatible with any concept of classical function or integral. A rigorous definition of the Dirac measure requires the theory of *generalized functions* or *distributions* of L. Schwartz, that we will consider in Chap. 7. Here we restrict ourselves to some heuristic considerations. Let

$$\mathcal{H}(x) = \begin{cases} 1 & \text{if } x \geq 0 \\ 0 & \text{if } x < 0, \end{cases}$$

be the characteristic function of the interval $[0, \infty)$, known as the *Heaviside function*. Observe that

$$\frac{\mathcal{H}(x + \varepsilon) - \mathcal{H}(x - \varepsilon)}{2\varepsilon} = \begin{cases} \frac{1}{2\varepsilon} & \text{if } -\varepsilon \leq x < \varepsilon \\ 0 & \text{otherwise.} \end{cases} \quad (2.62)$$

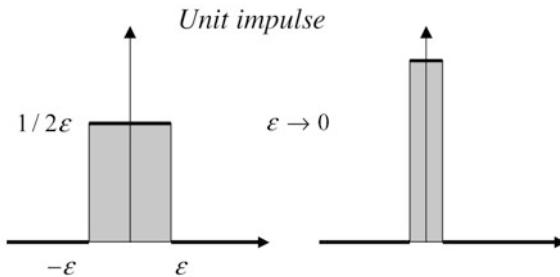


Fig. 2.6 Approximation of the Dirac measure

Denote by $I_\varepsilon(x)$ the quotient (2.62); the following properties hold:

- i) For every $\varepsilon > 0$,

$$\int_{\mathbb{R}} I_\varepsilon(x) dx = \frac{1}{2\varepsilon} 2\varepsilon = 1.$$

We can interpret I_ε as a *unit impulse of extent 2ε* (Fig. 2.6).

- ii)

$$\lim_{\varepsilon \downarrow 0} I_\varepsilon(x) = \begin{cases} 0 & \text{if } x \neq 0 \\ \infty & \text{if } x = 0. \end{cases}$$

- iii) If $\varphi = \varphi(x)$ is a smooth function, vanishing outside a bounded interval, (these kind of functions are called *test functions*), we have

$$\int_{\mathbb{R}} I_\varepsilon(x) \varphi(x) dx = \frac{1}{2\varepsilon} \int_{-\varepsilon}^{\varepsilon} \varphi(x) dx \xrightarrow{\varepsilon \rightarrow 0} \varphi(0).$$

Properties i) and ii) say that I_ε tends to a mathematical object that has precisely the formal features of the Dirac distribution at the origin. In particular iii) suggests how to identify this object, that is *through its action on test functions*.

Definition 2.8. We call *Dirac measure at the origin* the generalized function, denoted by δ , that acts on a test function φ as follows:

$$\delta[\varphi] = \varphi(0). \quad (2.63)$$

Equation (7.9) is often written in the form $\langle \delta, \varphi \rangle = \varphi(0)$ or even

$$\int \delta(x) \varphi(x) dx = \varphi(0)$$

where the integral symbol is purely formal. Observe that property ii) shows that

$$\mathcal{H}' = \delta$$

whose meaning is given in the following computations, where an integration by parts is used and φ is a test function:

$$\int_{\mathbb{R}} \varphi d\mathcal{H} = - \int_{\mathbb{R}} \mathcal{H}\varphi' = - \int_0^\infty \varphi' = \varphi(0), \quad (2.64)$$

since φ vanishes for large¹⁹ x .

With the notion of Dirac measure at hand, we can say that Γ_D satisfies the initial conditions

$$\Gamma_D(x, 0) = \delta(x).$$

If the unit mass is concentrated at a point $y \neq 0$, we denote by $\delta(x - y)$ the Dirac measure at y , defined through the formula

$$\int \delta(x - y) \varphi(x) dx = \varphi(y).$$

Then, by translation invariance, the fundamental solution $\Gamma_D(x - y, t)$ is a solution of the diffusion equation, that satisfies the initial condition

$$\Gamma_D(x - y, 0) = \delta(x - y).$$

Indeed, it is the unique solution satisfying the total mass condition (2.58), with $q = 1$.

As any solution u of (2.48) has several interpretations (concentration of a substance, probability density, temperature in a bar) hence also the fundamental solution can have several meanings.

We can think of it as a **unit source solution**: $\Gamma_D(x, t)$ gives the concentration at the point x , at time t , generated by the diffusion of a **unit mass initially ($t = 0$) concentrated at the origin**.

From another perspective, if we imagine a unit mass composed of a large number N of particles, $\Gamma_D(x, t)dx$ gives the probability that a single particle is placed between x and $x + dx$ at time t or, equivalently, the percentage of particles inside the interval $(x, x + dx)$ at time t .

Initially Γ_D is zero outside the origin. As soon as $t > 0$, Γ_D becomes positive everywhere: this amounts to saying that the unit mass diffuses instantaneously all over the x -axis and therefore *with infinite speed of propagation*. This could be a problem in using (2.48) as a realistic model, although (see Fig. 2.5) for $t > 0$, small, Γ_D is practically equal to zero outside an interval centered at the origin of length $4D$.

¹⁹ The first integral in (2.64) is a Riemann-Stieltjes integral, that formally can be written as

$$\int \varphi(x) \mathcal{H}'(x) dx$$

and interpreted as *the action of the generalized function \mathcal{H}' on the test function φ* .

2.3.4 The fundamental solution ($n > 1$)

In space dimension greater than 1, we can more or less repeat the same arguments. We look for positive, radial, self-similar solutions u^* to (2.48), with total mass equal to q at every time, that is

$$\int_{\mathbb{R}^n} u^*(\mathbf{x}, t) d\mathbf{x} = q \quad \text{for every } t > 0. \quad (2.65)$$

Since $q/(Dt)^{n/2}$ is a concentration per unit volume, we set

$$u^*(\mathbf{x}, t) = \frac{q}{(Dt)^{n/2}} U(\xi), \quad \xi = |\mathbf{x}|/\sqrt{Dt}.$$

We have, recalling the expression of the Laplace operator for radial functions (see Appendix C),

$$\begin{aligned} u_t^* &= -\frac{1}{2t(Dt)^{n/2}} [nU(\xi) + \xi U'(\xi)] \\ \Delta u^* &= \frac{1}{(Dt)^{1+n/2}} \left\{ U''(\xi) + \frac{n-1}{\xi} U'(\xi) \right\}. \end{aligned}$$

Therefore, for u^* to be a solution of (2.48), U must be a nonnegative solution in $(0, +\infty)$ of the ordinary differential equation

$$\xi U''(\xi) + (n-1)U'(\xi) + \frac{\xi^2}{2}U'(\xi) + \frac{n}{2}\xi U(\xi) = 0. \quad (2.66)$$

Multiplying by ξ^{n-2} , we can write (2.66) in the form

$$(\xi^{n-1}U')' + \frac{1}{2}(\xi^n U)' = 0$$

that gives

$$\xi^{n-1}U' + \frac{1}{2}\xi^n U = C \quad (C \in \mathbb{R}). \quad (2.67)$$

Assuming that $\lim_{\xi \rightarrow 0^+}$ of U and U' are finite, letting $\xi \rightarrow 0^+$ into (2.67), we deduce $C = 0$ and therefore

$$U' + \frac{1}{2}\xi U = 0.$$

Thus we obtain the family of solutions

$$U(\xi) = c_0 e^{-\frac{\xi^2}{4}}.$$

The total mass condition requires

$$\begin{aligned} 1 &= \frac{1}{(Dt)^{n/2}} \int_{\mathbb{R}^n} U\left(\frac{|\mathbf{x}|}{\sqrt{Dt}}\right) d\mathbf{x} = \frac{c_0}{(Dt)^{n/2}} \int_{\mathbb{R}^n} \exp\left(-\frac{|\mathbf{x}|^2}{4Dt}\right) d\mathbf{x} \\ &\stackrel{\mathbf{y}=\mathbf{x}/\sqrt{Dt}}{=} c_0 \int_{\mathbb{R}^n} e^{-|\mathbf{y}|^2} d\mathbf{y} = c_0 \left(\int_{\mathbb{R}} e^{-z^2} dz\right)^n = c_0 (4\pi)^{n/2} \end{aligned}$$

and therefore $c_0 = (4\pi)^{-n/2}$. Thus, we have obtained solutions of the form

$$u^*(\mathbf{x}, t) = \frac{q}{(4\pi Dt)^{n/2}} \exp\left(-\frac{|\mathbf{x}|^2}{4Dt}\right), \quad (t > 0).$$

Once more, the choice $q = 1$ is special.

Definition 2.9. *The function*

$$\Gamma_D(\mathbf{x}, t) = \frac{1}{(4\pi Dt)^{n/2}} \exp\left(-\frac{|\mathbf{x}|^2}{4Dt}\right) \quad (t > 0)$$

is called the **fundamental solution of the diffusion equation** (2.48).

The remarks after Definition 2.8, p. 45, can be easily generalized to the multi-dimensional case. In particular, it is possible to define the *n-dimensional Dirac measure* at a point \mathbf{y} , denoted by $\delta_n(\mathbf{x} - \mathbf{y})$, through the formula²⁰

$$\int \delta_n(\mathbf{x} - \mathbf{y}) \varphi(\mathbf{x}) dx = \varphi(\mathbf{y}) \quad (2.68)$$

that expresses the action of $\delta_n(\mathbf{x} - \mathbf{y})$ on the test function φ , smooth in \mathbb{R}^n and vanishing outside a compact set. For $n = 1$ we set $\delta_1 = \delta$.

For fixed \mathbf{y} , the fundamental solution $\Gamma_D(\mathbf{x} - \mathbf{y}, t)$ is the unique solution of the global Cauchy problem

$$\begin{cases} u_t - D\Delta u = 0 & \mathbf{x} \in \mathbb{R}^n, t > 0 \\ u(\mathbf{x}, 0) = \delta_n(\mathbf{x} - \mathbf{y}) & \mathbf{x} \in \mathbb{R}^n \end{cases}$$

that satisfies (2.65) with $q = 1$.

2.4 Symmetric Random Walk ($n = 1$)

In this section we start exploring the connection between probabilistic and deterministic models, in dimension $n = 1$. The main purpose is to construct a Brownian

²⁰ As in dimension $n = 1$, in (2.68) the integral has a symbolic meaning only.

motion, which is a **continuous model** (in both space and time), as a limit of a simple stochastic process, called *random walk*, which is instead a **discrete model** (in both space and time). During the realization of the limiting procedure we shall see how the diffusion equation can be approximated by a *difference equation*. Moreover, this new perspective will better clarify the nature of the diffusion coefficient.

2.4.1 Preliminary computations

Consider a unit mass particle²¹ that moves randomly along the x axis, according to the following rules: fix

- $h > 0$, space step.
- $\tau > 0$, time step.

1. During an interval of time τ , the particle takes one step of h unit length, starting from $x = 0$.
2. The particle moves to the left or to the right with probability $p = \frac{1}{2}$, independently of the previous steps (Fig. 2.7).

At time $t = N\tau$, that is after N steps, the particle will be located at a point $x = mh$, where $N \geq 0$ and m are integers, with $-N \leq m \leq N$.

Our task is: *Compute the probability $p(x, t)$ to find the particle at x , at time t .*

Random walks can occur in a wide variety of situations. To give an example, think of a gambling game in which a *fair coin* is thrown. If heads comes out, the particle moves to the right and the player gains 1 *dollar*; if tails comes out it moves to the left and the player loses 1 *dollar*: $p(x, t)$ is the probability to gain m dollars after N throws.

- *Computation of $p(x, t)$.*

Let $x = mh$ be the position of the particle after N steps. To reach x , the particle takes some number of steps to the right, say k , and $N - k$ steps to the left. Clearly, $0 \leq k \leq N$ and

$$m = k - (N - k) = 2k - N \quad (2.69)$$

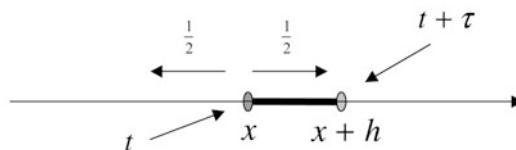


Fig. 2.7 Symmetric random walk

²¹ One can also think of a large number of particles of total mass one.

so that N and m are both even or both odd integers and

$$k = \frac{1}{2} (N + m).$$

Thus, $p(x, t) = p_k$ where

$$p_k = \frac{\text{number of walks with } k \text{ steps to the right after } N \text{ steps}}{\text{number of possible walks after } N \text{ steps}}. \quad (2.70)$$

Now, the number of possible walks with k steps to the right and $N - k$ to the left is given by the binomial coefficient²²

$$C_{N,k} = \binom{N}{k} = \frac{N!}{k!(N-k)!}.$$

On the other hand, the number of possible walks after N steps is 2^N (why?); hence, from (2.70):

$$p_k = \frac{C_{N,k}}{2^N} \quad x = mh, \quad t = N\tau, \quad k = \frac{1}{2} (N + m). \quad (2.71)$$

- *Mean displacement and standard deviation of x .*

Our ultimate goal is to let h and τ go to zero, keeping x and t fixed, in order to get a continuous walk, which incorporates the main features of the discrete random walk. This is a delicate point, since, if we want to eventually obtain a continuous faithful copy of the random walk, we need to isolate some quantitative parameters, able to capture the essential features of the walk and maintain them unchanged. In our case there are two key parameters²³

- (a) The **mean displacement** of x after N steps = $\langle x \rangle = \langle m \rangle h$.
- (b) The **second moment** of x after N steps = $\langle x^2 \rangle = \langle m^2 \rangle h^2$.

The quantity $\sqrt{\langle x^2 \rangle} = \sqrt{\langle m^2 \rangle}h$ measures the average distance from the origin after N steps.

²² The set of walks with k steps to the right and $N - k$ to the left is in one to one correspondence with the set of *sequences of N binary digits*, where 1 means *right* and 0 means *left*. There are exactly $C_{N,k}$ of these sequences.

²³ If a random variable x takes N possible outcomes x_1, \dots, x_N with probability p_1, \dots, p_N , its *moments of (integer) order $q \geq 1$* are given by

$$E(x^q) = \langle x^q \rangle = \sum_{j=1}^N x_j^q p_j.$$

The first moment ($q = 1$) is the *mean or expected value of x* , while

$$\text{var}(x) = \langle x^2 \rangle - \langle x \rangle^2$$

is the *variance of x* . The square root of the variance is called *standard deviation*.

First observe that, from (2.69), we have

$$\langle m \rangle = 2 \langle k \rangle - N \quad (2.72)$$

and

$$\langle m^2 \rangle = 4 \langle k^2 \rangle - 4 \langle k \rangle N + N^2. \quad (2.73)$$

Thus, to compute $\langle m \rangle$ and $\langle m^2 \rangle$ it is enough to compute $\langle k \rangle$ and $\langle k^2 \rangle$. We have, by definition and from (2.71),

$$\langle k \rangle = \sum_{k=1}^N kp_k = \frac{1}{2^N} \sum_{k=1}^N k C_{N,k}, \quad \langle k^2 \rangle = \sum_{k=1}^N k^2 p_k = \frac{1}{2^N} \sum_{k=1}^N k^2 C_{N,k}. \quad (2.74)$$

Although it is possible to make the calculations directly from (2.74), it is easier to use the probability generating function, defined by

$$G(s) = \sum_{k=0}^N p_k s^k = \frac{1}{2^N} \sum_{k=0}^N C_{N,k} s^k.$$

The function G contains in compact form all the information on the moments of k and works for all the discrete random variables taking integer values. In particular, we have

$$G'(s) = \frac{1}{2^N} \sum_{k=1}^N k C_{N,k} s^{k-1}, \quad G''(s) = \frac{1}{2^N} \sum_{k=2}^N k(k-1) C_{N,k} s^{k-2}. \quad (2.75)$$

Letting $s = 1$ and using (2.74), we get

$$G'(1) = \frac{1}{2^N} \sum_{k=1}^N k C_{N,k} = \langle k \rangle \quad (2.76)$$

and

$$G''(1) = \frac{1}{2^N} \sum_{k=2}^N k(k-1) C_{N,k} = \langle k(k-1) \rangle = \langle k^2 \rangle - \langle k \rangle. \quad (2.77)$$

On the other hand, letting $a = 1$ and $b = s$ in the elementary formula

$$(a+b)^N = \sum_{k=0}^N C_{N,k} a^{N-k} b^k,$$

we deduce

$$G(s) = \frac{1}{2^N} (1+s)^N$$

and therefore

$$G'(1) = \frac{N}{2} \quad \text{and} \quad G''(1) = \frac{N(N-1)}{4}. \quad (2.78)$$

From (2.78), (2.76) and (2.77) we easily find

$$\langle k \rangle = \frac{N}{2} \quad \text{and} \quad \langle k^2 \rangle = \frac{N(N+1)}{4}.$$

Finally, since $m = 2k - N$, we have

$$\langle m \rangle = 2\langle k \rangle - N = 2\frac{N}{2} - N = 0$$

and also $\langle x \rangle = \langle m \rangle h = 0$, which is not surprising, given the symmetry of the walk. Furthermore

$$\langle m^2 \rangle = 4\langle k^2 \rangle - 4N\langle k \rangle + N^2 = N^2 + N - 2N^2 + N^2 = N$$

from which

$$\sqrt{\langle x^2 \rangle} = \sqrt{Nh}, \quad (2.79)$$

which is the *standard deviation of x*, since $\langle x \rangle = 0$. Formula (2.79) contains a key information: at time $N\tau$, the distance from the origin is of order \sqrt{Nh} , that is **the order of the time scale is the square of the space scale**. In other words, if we want to leave the standard deviation unchanged in the limit process, we must rescale the time as the square of the space, that is we must use a *space-time parabolic dilation!*

But let us proceed step by step. The next one is to deduce a *difference equation* for the transition probability $p = p(x, t)$. It is on this equation that we will carry out the limit procedure.

2.4.2 The limit transition probability

The particle motion has no memory since each move is independent from the previous ones. If the particle location at time $t + \tau$ is x , this means that at time t its location was at $x - h$ or at $x + h$, with equal probability. The total probability formula then gives

$$p(x, t + \tau) = \frac{1}{2}p(x - h, t) + \frac{1}{2}p(x + h, t) \quad (2.80)$$

with the initial conditions

$$p(0, 0) = 1 \quad \text{and} \quad p(x, 0) = 0 \quad \text{if } x \neq 0.$$

Keeping x and t fixed, let us examine what happens when $h \rightarrow 0, \tau \rightarrow 0$. It is convenient to think of p as a smooth function, defined in the whole half plane $\mathbb{R} \times (0, +\infty)$ and not only at the discrete set of points $(mh, N\tau)$. By passing to

the limit, we will find a continuous probability distribution and $p(x, t)$ has to be interpreted as *probability density*. Using Taylor's formula we can write²⁴

$$\begin{aligned} p(x, t + \tau) &= p(x, t) + p_t(x, t)\tau + o(\tau), \\ p(x \pm h, t) &= p(x, t) \pm p_x(x, t)h + \frac{1}{2}p_{xx}(x, t)h^2 + o(h^2). \end{aligned}$$

Substituting into (2.80), after some simplifications, we find

$$p_t\tau + o(\tau) = \frac{1}{2}p_{xx}h^2 + o(h^2).$$

Dividing by τ , we get

$$p_t + o(1) = \frac{1}{2}\frac{h^2}{\tau}p_{xx} + o\left(\frac{h^2}{\tau}\right). \quad (2.81)$$

This is the crucial point; in the last equation we meet again the combination $\frac{h^2}{\tau}$!!

If we want to obtain something nontrivial when $h, \tau \rightarrow 0$, **we must require that h^2/τ has a finite and positive limit**; the simplest choice is to keep

$$\frac{h^2}{\tau} = 2D \quad (2.82)$$

for some number $D > 0$ (the number 2 is there for aesthetic reasons only) whose physical dimensions are clearly $[D] = [\text{length}]^2 \times [\text{time}]^{-1}$.

Passing to the limit in (2.81), we get for p the equation

$$p_t = Dp_{xx} \quad (2.83)$$

while the initial condition becomes

$$\lim_{t \rightarrow 0^+} p(x, t) = \delta(x). \quad (2.84)$$

We have already seen that the unique solution of (2.83), (2.84) is

$$p(x, t) = T_D(x, t)$$

since, due to the meaning of p ,

$$\int_{\mathbb{R}} p(x, t) dx = 1.$$

²⁴ Recall that the symbol $o(z)$, ("little o of z ") denotes a quantity of lower order with respect to z ; precisely

$$\frac{o(z)}{z} \rightarrow 0 \quad \text{when } z \rightarrow 0.$$

Thus, the constant D in (2.82) is precisely the *diffusion coefficient*. Recalling that

$$h^2 = \frac{\langle x^2 \rangle}{N}, \quad \tau = \frac{t}{N},$$

we have

$$\frac{h^2}{\tau} = \frac{\langle x^2 \rangle}{t} = 2D$$

that means: *in unit time, the particle diffuses up to an average distance of $\sqrt{2D}$* . Also, from (2.82) we deduce

$$\frac{h}{\tau} = \frac{2D}{h} \rightarrow +\infty. \quad (2.85)$$

This shows that the average speed h/τ of the particle at each step becomes unbounded. Therefore, the fact that the particle diffuses in unit time up to a finite average distance is purely due to the rapid fluctuations of its motion.

2.4.3 From random walk to Brownian motion

What happened in the limit to the random walk? What kind of motion did it become?

We can answer using some more tools from probability theory. Let $x_j = x(j\tau)$ be the position of our particle after j steps and let, for $j \geq 1$,

$$h\xi_j = x_j - x_{j-1}.$$

The ξ_j are *independent, identically distributed random variables*: each one takes on the value 1 or -1 with probability $\frac{1}{2}$. They have expectation $\langle \xi_j \rangle = 0$ and variance $\langle \xi_j^2 \rangle = 1$.

The displacement of the particle after N steps is

$$x_N = h \sum_{j=1}^N \xi_j.$$

If we choose

$$h = \sqrt{\frac{2Dt}{N}},$$

that is $\frac{h^2}{\tau} = 2D$, and let $N \rightarrow \infty$, the *Central Limit Theorem* insures that x_N converges in law²⁵ to a random variable $X = X(t)$, normally distributed with mean 0 and variance $2Dt$, whose density is $\Gamma_D(x, t)$.

²⁵ That is, if $N \rightarrow +\infty$, for every, $a, b \in \mathbb{R}$, $a < b$,

$$\text{Prob}\{a < x_N < b\} \rightarrow \int_a^b \Gamma_D(x, t) dx.$$

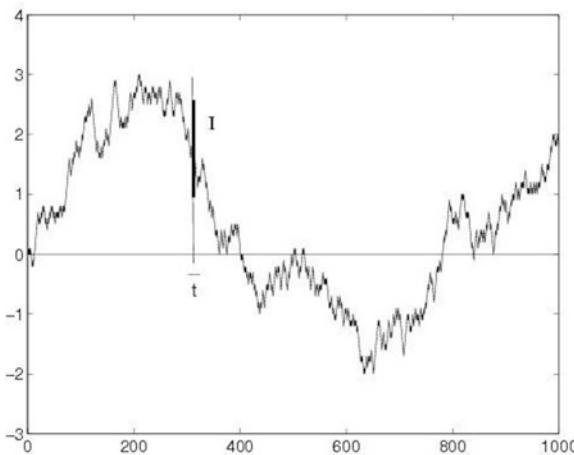


Fig. 2.8 A Brownian path

The random walk has become a continuous walk; if $D = 1/2$, it is called (1-dimensional) **Brownian motion or Wiener process**, that we will characterize later through its essential features.

Usually the symbol $B = B(t)$ is used to indicate the random position of a Brownian particle. The family of random variables $B(t)$, where t plays the role of a parameter, is defined on a common probability space (Ω, \mathcal{F}, P) , where Ω is the set of elementary events, \mathcal{F} a σ -algebra in Ω of measurable events, and P a suitable probability measure²⁶ in \mathcal{F} ; therefore the right notation should be

$$B = B(t, \omega),$$

with $\omega \in \Omega$, but the dependence on ω is usually omitted and understood (for simplicity or laziness).

The family of random variables $B(t, \omega)$, with time t as a real parameter, is a **continuous stochastic process**. Keeping $\omega \in \Omega$ fixed, we get the real function

$$t \longmapsto B(t, \omega)$$

whose graph describes a Brownian path (see Fig. 2.8).

Keeping t fixed, we get the random variable

$$\omega \longmapsto B(t, \omega).$$

Without caring too much of what Ω really is, it is important to be able to compute the probability

$$P\{B(t) \in I\}$$

²⁶ See Appendix B.

where $I \subseteq R$ is a reasonable subset of R , (*a so called Borel set*²⁷). Figure 2.8 shows the meaning of this computation: fixing t amounts to fixing a vertical straight line, say $t = \bar{t}$. Let I be a subset of this line; in the picture I is an interval. $P\{B(\bar{t}) \in I\}$ is the probability that the particle hits I at time \bar{t} .

The main properties of Brownian motion are listed below. To be minimalist we could synthesize everything in the formula²⁸

$$dB \sim \sqrt{dt}N(0, 1) = N(0, dt) \quad (2.86)$$

where $N(0, 1)$ is a normal random variable, with zero mean and variance equal to one.

- *Path continuity.* With probability 1, the possible paths of a Brownian particle are continuous functions

$$t \longmapsto B(t), \quad t \geq 0.$$

Since from (2.85) the instantaneous speed of the particle is infinite, their graphs are nowhere differentiable!

- *Gaussian law for increments.* We can allow the particle to start from a point $x \neq 0$, by considering the process

$$B^x(t) = x + B(t).$$

With every point x is associated a probability P^x , with the following properties (if $x = 0$, $P^0 = P$):

- $P^x\{B^x(0) = x\} = P\{B(0) = 0\} = 1$.
- For every $s \geq 0, t \geq 0$, the increment

$$B^x(t+s) - B^x(s) = B(t+s) - B(s) \quad (2.87)$$

has normal law, with *zero mean and variance* t , with density given by

$$\Gamma(x, t) \equiv \Gamma_{\frac{1}{2}}(x, t) = \frac{1}{\sqrt{2\pi t}} e^{-\frac{x^2}{2t}}.$$

Moreover, the increment (2.87) is independent of any event occurred at a time $\leq s$. For instance, the two events

$$\{B^x(t_2) - B^x(t_1) \in I_2\} \quad \{B^x(t_1) - B^x(t_0) \in I_1\},$$

with $t_0 < t_1 < t_2$, are independent.

²⁷ An interval or a set obtained by countable unions and intersections of intervals, for instance. See Appendix B.

²⁸ If X is a random variable, we write $X \sim N(\mu, \sigma^2)$ if X has normal distribution with mean μ and variance σ^2 .

- *Transition probability.* For each Borel set $I \subseteq \mathbb{R}$, is defined a *transition function*

$$P(x, t, I) = P^x \{B^x(t) \in I\}$$

assigning the probability that the particle, initially at x , belongs to I at time t . We can write:

$$P(x, t, I) = P\{B(t) \in I - x\} = \int_{I-x} \Gamma(y, t) dy = \int_I \Gamma(y - x, t) dy.$$

- *Invariance.* The motion is invariant with respect to translations.
- *Markov and strong Markov properties.* Let μ be a probability measure²⁹ on \mathbb{R} . If the initial position of the particle is random with a probability distribution μ , we can consider a *Brownian motion with initial distribution* μ , and for it we use the symbol B^μ . With this motion is associated a probability distribution P^μ such that, for every Borel set $F \subseteq \mathbb{R}$,

$$P^\mu \{B^\mu(0) \in F\} = \mu(F).$$

The probability that the particle belongs to I at time t can be computed through the formula

$$\begin{aligned} P^\mu \{B^\mu(t) \in I\} &= \int_{\mathbb{R}} P^x \{B^x(t) \in I\} \mu(dx) \\ &= \int_{\mathbb{R}} P(x, t, I) \mu(dx). \end{aligned}$$

The *Markov property* can be stated as follows: given any condition H , related to the behavior of the particle before time $s \geq 0$, the process

$$Y(t) = B^x(t + s)$$

is a Brownian motion with initial distribution³⁰

$$\mu(I) = P^x \{B^x(s) \in I | H\}.$$

This property establishes the independence of the *future* process $B^x(t + s)$ from the *past* (absence of memory) when the *present* $B^x(s)$ is known and reflects the *absence of memory* of the random walk.

In the strong *Markov property*, s is substituted by a random time τ , which depends only on the behavior of the particle in the interval $[0, \tau]$. In other words, to decide whether or not the event $\{\tau \leq t\}$ is true, it is enough to know the behavior of the particle up to time t . These kinds of random times are called *stopping times*. An important example is the *first exit time from a domain*, that we will consider

²⁹ See Appendix B for the definition of a probability measure μ and of the integral with respect to the measure μ .

³⁰ $P(A | H)$ denotes the conditional probability of A , given H .

in the next chapter. Instead, the random time τ defined by

$$\tau = \inf \{t : B(t) > 10 \text{ and } B(t+1) < 10\}$$

is *not* a stopping time. Indeed (measuring time in *seconds*), τ is “the smallest” among the times t such that the Brownian path is above level 10 at time t , and after one second is below 10. Clearly, to decide whether $\tau \leq 3$, say, it is not enough to know the path up to time $t = 3$, since τ involves the behavior of the path up to the *future* time $t = 4$.

- *Expectation.* Given a sufficiently smooth function $g = g(y)$, $y \in \mathbb{R}$, we can define the random variable

$$Z(t) = (g \circ B^x)(t) = g(B^x(t)).$$

Its expected value is given by the formula

$$E^x[Z(t)] = \int_{\mathbb{R}} g(y) P(x, t, dy) = \int_{\mathbb{R}} g(y) \Gamma(y - x, t) dy.$$

We will meet this formula in a completely different situation later on.

2.5 Diffusion, Drift and Reaction

2.5.1 Random walk with drift

The hypothesis of symmetry of our random walk can be removed. Suppose our unit mass particle moves along the x axis with space step $h > 0$, every time interval of duration $\tau > 0$, according to the following rules (Fig. 2.9).

1. The particle starts from $x = 0$.
2. It moves to the right with probability $p_0 \neq \frac{1}{2}$ and to the left with probability $q_0 = 1 - p_0$, independently of the previous steps.

Rule 2 breaks the symmetry of the walk and models a particle tendency to move to the right or to the left, according to the sign of $p_0 - q_0$ being positive or negative, respectively. Again we denote by $p = p(x, t)$ the probability that the

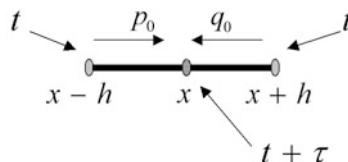


Fig. 2.9 Random walk with drift

particle location is $x = mh$ at time $t = N\tau$. From the total probability formula we have:

$$p(x, t + \tau) = p_0 p(x - h, t) + q_0 p(x + h, t), \quad (2.88)$$

with the usual initial conditions

$$p(0, 0) = 1 \quad \text{and} \quad p(x, 0) = 0 \quad \text{if } x \neq 0.$$

As in the symmetric case, we want to examine what happens when we pass to the limit for $h \rightarrow 0, \tau \rightarrow 0$. From Taylor formula, we have

$$\begin{aligned} p(x, t + \tau) &= p(x, t) + p_t(x, t)\tau + o(\tau), \\ p(x \pm h, t) &= p(x, t) \pm p_x(x, t)h + \frac{1}{2}p_{xx}(x, t)h^2 + o(h^2). \end{aligned}$$

Substituting into (2.88), we get

$$p_t\tau + o(\tau) = \frac{1}{2}p_{xx}h^2 + (q_0 - p_0)hp_x + o(h^2). \quad (2.89)$$

A new term appears: $(q_0 - p_0)hp_x$. Dividing by τ , we obtain

$$p_t + o(1) = \frac{1}{2}\frac{h^2}{\tau}p_{xx} + \boxed{\frac{(q_0 - p_0)h}{\tau}p_x} + o\left(\frac{h^2}{\tau}\right). \quad (2.90)$$

Again, here is the crucial point. If we let $h, \tau \rightarrow 0$, we realize that the assumption

$$\frac{h^2}{\tau} = 2D \quad (2.91)$$

alone is not sufficient anymore to get something nontrivial from (2.90): indeed, if we keep p_0 and q_0 constant, we have

$$\frac{(q_0 - p_0)h}{\tau} \rightarrow \infty$$

and from (2.90) we get a contradiction. What else do we have to require? Writing

$$\frac{(q_0 - p_0)h}{\tau} = \frac{(q_0 - p_0)h^2}{h\tau}$$

we see we must require, in addition to (2.91), that

$$\frac{q_0 - p_0}{h} \rightarrow \beta \quad (2.92)$$

with β finite. Notice that, since $q_0 + p_0 = 1$, (2.92) is equivalent to

$$p_0 = \frac{1}{2} - \frac{\beta}{2}h + o(h) \quad \text{and} \quad q_0 = \frac{1}{2} + \frac{\beta}{2}h + o(h), \quad (2.93)$$

that could be interpreted as a *symmetry of the motion at a microscopic scale*.

With (2.92) at hand, we have

$$\frac{(q_0 - p_0)}{h} \frac{h^2}{\tau} \rightarrow 2D\beta \equiv b$$

and (2.90) becomes in the limit,

$$p_t = Dp_{xx} + bp_x. \quad (2.94)$$

We already know that Dp_{xx} models a diffusion phenomenon. Let us *unmask* the term bp_x , by first examining the dimensions of b . Since $q_0 - p_0$ is dimensionless, being a difference of probabilities, the dimensions of b are those of h/τ , namely of a **velocity**.

Thus the coefficient b codifies the tendency of the limiting continuous motion, to move towards a privileged direction with speed $|b|$: to the right if $b < 0$, to the left if $b > 0$. In other words, there exists a *current of intensity* $|b|$, driving the particle. *The random walk has become a diffusion process with drift.*

The last point of view calls for an analogy with the diffusion of a substance transported along a channel.

2.5.2 Pollution in a channel

In this section we examine a simple convection-diffusion model of a pollutant on the surface of a narrow channel. A water stream of constant speed v transports the pollutant along the positive direction of the x axis. We can neglect the depth of the water (thinking to a floating pollutant) and the transverse dimension (thinking of a very narrow channel).

Our purpose is to derive a mathematical model capable of describing the evolution of the concentration³¹ $c = c(x, t)$ of the pollutant. Accordingly, the integral

$$\int_x^{x+\Delta x} c(y, t) dy \quad (2.95)$$

gives the mass inside the interval $(x, x + \Delta x)$ at time t (Fig. 2.10). In the present case there are neither sources nor sinks of pollutant, therefore to construct a model we use the **law of mass conservation**: *the growth rate of the mass contained in an interval $(x, x + \Delta x)$ equals the net mass flux into $(x, x + \Delta x)$ through the end points.*

From (2.95), the growth rate of the mass contained in an interval $(x, x + \Delta x)$ is given by³²

$$\frac{d}{dt} \int_x^{x+\Delta x} c(y, t) dy = \int_x^{x+\Delta x} c_t(y, t) dy. \quad (2.96)$$

³¹ $[c] = [\text{mass}] \times [\text{length}]^{-1}$.

³² Assuming we can take the derivative inside the integral.

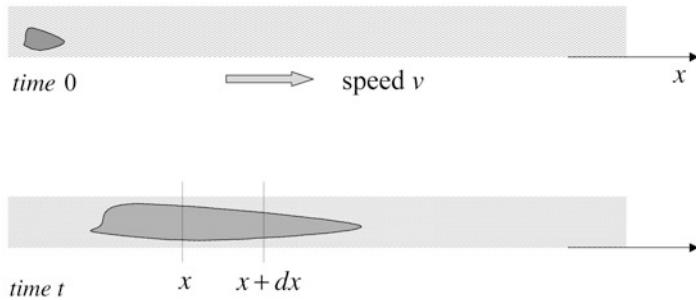


Fig. 2.10 Pollution in a narrow channel

Denote by $q = q(x, t)$ the mass flux³³ entering the interval $(x, x + \Delta x)$, through the point x at time t . The net mass flux into $(x, x + \Delta x)$ through the end points is

$$q(x, t) - q(x + \Delta x, t). \quad (2.97)$$

Equating (2.96) and (2.97), the law of mass conservation reads

$$\int_x^{x+\Delta x} c_t(y, t) dy = q(x, t) - q(x + \Delta x, t).$$

Dividing by Δx and letting $\Delta x \rightarrow 0$, we find the basic law

$$c_t = -q_x. \quad (2.98)$$

At this point we have to decide which kind of mass flux we are dealing with. In other words, we need a *constitutive relation* for q . There are several possibilities, for instance:

- a) *Convection*. The flux is determined by the water stream only. This case corresponds to a bulk of pollutant that is driven by the stream, without deformation or expansion. Translating into mathematical terms we find

$$q(x, t) = vc(x, t)$$

where, we recall, v denotes the stream speed.

- b) *Diffusion*. The pollutant expands from higher to lower concentration regions. We have seen something like that in heat conduction, where, according to the Fourier law, the heat flux is proportional and opposite to the temperature gradient. Here we can adopt a similar law, that, in this setting, is known as the *Fick's law* and reads

$$q(x, t) = -Dc_x(x, t), \quad (2.99)$$

³³ $[q] = [\text{mass}] \times [\text{time}]^{-1}$.

where the constant D depends on the polluting and has the usual dimensions ($[D] = [length]^2 \times [time]^{-1}$).

In our case, convection and diffusion are both present and therefore we superpose the two effects, by writing

$$q(x, t) = vc(x, t) - Dc_x(x, t).$$

From (2.98) we deduce

$$c_t = Dc_{xx} - vc_x \quad (2.100)$$

which constitutes our mathematical model and turns out to be identical to (10.22).

Since D and v are constant, it is easy to determine the evolution of a mass Q of pollutant, initially located at the origin (say). Its concentration is the solution of (2.100) with initial condition

$$c(x, 0) = Q\delta(x)$$

where δ is the Dirac measure at the origin. To find an explicit formula, we can get rid of the drift term $-vc_x$ by setting

$$w(x, t) = c(x, t) e^{hx+kt}$$

with h, k to be chosen suitably. We have:

$$\begin{aligned} w_t &= (c_t + kc) e^{hx+kt} \\ w_x &= (c_x + hc) e^{hx+kt}, \quad w_{xx} = (c_{xx} + 2hc_x + h^2 c) e^{hx+kt}. \end{aligned}$$

Using the equation $c_t = Du_{xx} - vc_x$, we can write

$$\begin{aligned} w_t - Dw_{xx} &= e^{hx+kt}[c_t - Dc_{xx} - 2Dhc_x + (k - Dh^2)c] = \\ &= e^{hx+kt}[(-v - 2Dh)c_x + (k - Dh^2)c]. \end{aligned}$$

Thus if we choose

$$h = -\frac{v}{2D} \quad \text{and} \quad k = Dh^2 = \frac{v^2}{4D},$$

w is a solution of the diffusion equation $w_t - Dw_{xx} = 0$, with the initial condition

$$w(x, 0) = c(x, 0) e^{-\frac{v}{2D}x} = Q\delta(x) e^{-\frac{v}{2D}x}.$$

In Chap. 7 we show that, in a suitable sense,

$$\delta(x) e^{-\frac{v}{2D}x} = \delta(x),$$

so that $w(x, t) = Q\Gamma_D(x, t)$ and finally

$$c(x, t) = Qe^{\frac{v}{2D}(x - \frac{v}{2}t)} \Gamma_D(x, t). \quad (2.101)$$

The concentration c is thus given by the fundamental solution Γ_D , “carried” by the travelling wave $\exp\left\{\frac{v}{2D}(x - \frac{v}{2}t)\right\}$, in motion to the right with speed $v/2$.

In realistic situations, the pollutant undergoes some sort of decay, for instance by biological decomposition. The resulting equation for the concentration becomes

$$c_t = Dc_{xx} - vc_x - \gamma c$$

where γ is a rate of decay³⁴. We deal with this case in the next section via a suitable variant of our random walk.

2.5.3 Random walk with drift and reaction

We go back to our 1– dimensional random walk, assuming that the particle loses mass at the constant rate $\gamma > 0$. This means that in an interval of time from t to $t + \tau$ a percentage of mass

$$Q(x, t) = \tau\gamma p(x, t)$$

disappears. The difference equation (2.88) for p becomes

$$p(x, t + \tau) = p_0[p(x - h, t) - Q(x - h, t)] + q_0[p(x + h, t) - Q(x + h, t)].$$

We have:

$$\begin{aligned} p_0Q(x - h, t) + q_0Q(x + h, t) &= Q(x, t) + (q_0 - p_0)hQ_x(x, t) + \dots \\ &= \tau\gamma p(x, t) + O(\tau h), \end{aligned}$$

where the symbol “ $O(k)$ ” (“big O of k ”) denotes a quantity such that $O(k)/k$ remains bounded as $k \rightarrow 0$.

Thus, eq. (2.89) modifies into

$$p_t\tau + o(\tau) = \frac{1}{2}p_{xx}h^2 + (q_0 - p_0)hp_x - \tau\gamma p + O(\tau h) + o(h^2).$$

Dividing by τ , letting $h, \tau \rightarrow 0$ and assuming

$$\frac{h^2}{\tau} = 2D, \quad \frac{q_0 - p_0}{h} \rightarrow \beta,$$

we get

$$p_t = Dp_{xx} + bp_x - \gamma p \quad (b = 2D\beta). \quad (2.102)$$

The term $-\gamma p$ appears in (2.102) as a decaying term. On the other hand, as we will see in the next subsection, γ could be *negative*, meaning that this time we have a *creation of mass* at the rate $|\gamma|$. For this reason the last term is generically called a *reaction term* and (2.102) is a *diffusion equation with drift and reaction*.

³⁴ $[\gamma] = [\text{time}]^{-1}$.

Going back to equation (2.102), it is useful to look separately at the effect of the three terms in its right hand side.

- $p_t = Dp_{xx}$ models pure diffusion. The typical effects are spreading and smoothing, as shown by the typical behavior of the fundamental solution Γ_D .
- $p_t = bp_x$ is a pure transport equation, that we will consider in detail in Chap. 4. The solutions are travelling waves of the form $g(x + bt)$.
- $p_t = -\gamma p$ models pure reaction. The solutions are multiples of $e^{-\gamma t}$, exponentially decaying (increasing) if $\gamma > 0$ ($\gamma < 0$).

So far we have given a probabilistic interpretation for a motion in all \mathbb{R} , where no boundary condition is present. The Problems 2.11 and 2.12 give a probabilistic interpretation of the Neumann and Dirichlet condition in terms of *reflecting absorbing* boundaries, respectively.

2.5.4 Critical dimension in a simple population dynamics

When $-\gamma = a > 0$ in (2.102), a competition between reaction and diffusion occurs. We examine this effect on the following simple population dynamics problem:

$$\begin{cases} u_t - Du_{xx} = au & 0 < x < L, t > 0 \\ u(0, t) = u(L, t) = 0 & t > 0 \\ u(x, 0) = g(x) & 0 < x < L, \end{cases} \quad (2.103)$$

where u represents the density of a population of individuals. In this case, the homogeneous Dirichlet conditions model an hostile external environment³⁵. Given this kind of boundary condition, the population decays by diffusion while tends to increase by reaction. Thus the two effects compete and we want to explore which factors determine the overwhelming one.

First of all, since a is constant, we can get rid of the term au by setting

$$u(x, t) = e^{at} w(x, t).$$

We have:

$$u_t = e^{at}(aw + w_t), \quad u_x = e^{at}w_x, \quad u_{xx} = e^{at}w_{xx}$$

and substituting into the differential equation, after some simple algebra, we find for w the equation

$$w_t - Dw_{xx} = 0,$$

with the same boundary and initial conditions:

$$\begin{aligned} w(0, t) &= w(L, t) = 0 \\ w(x, 0) &= g(x). \end{aligned}$$

³⁵ A homogeneous Neumann condition would represent the evolution of an isolated population, without external exchange.

Then, we can easily exhibit an explicit formula for the solution, using the separation of variables³⁶:

$$w(x, t) = \sum_{k=1}^{\infty} b_k \exp\left(-D \frac{k^2 \pi^2}{L^2} t\right) \sin \frac{k \pi x}{L} \quad (2.104)$$

with

$$b_k = \frac{1}{L} \int_0^L g(x) \sin(k \pi x / L) dx.$$

If $g \in C^2([0, L])$ and $g(0) = g(L) = 0$, the series (2.104) converges uniformly for $t > 0$ and $0 \leq x \leq L$. Going back to u , we get for the solution to problem (2.103) the following expression:

$$u(x, t) = \sum_{k=1}^{\infty} b_k \exp\left\{(a - D \frac{k^2 \pi^2}{L^2})t\right\} \sin \frac{k \pi x}{L}. \quad (2.105)$$

Formula (2.105) displays an important difference from the pure diffusion case $a = 0$, as far as the asymptotic behavior for $t \rightarrow +\infty$ is concerned. Assuming $b_1 \neq 0$, the population evolution is determined by the largest exponential in the series (2.105), corresponding to $k = 1$. It is now an easy matter to draw the following conclusions.

1. If

$$a - D \frac{\pi^2}{L^2} < 0, \quad \text{then} \quad \lim_{t \rightarrow +\infty} u(x, t) = 0 \quad (2.106)$$

uniformly in $[0, L]$, since $a - D \frac{k^2 \pi^2}{L^2} < a - D \frac{\pi^2}{L^2}$ for every $k > 1$.

2. If

$$a - D \frac{\pi^2}{L^2} > 0, \quad \text{then} \quad \lim_{t \rightarrow +\infty} u(x, t) = \infty$$

for $x \neq 0, L$, since the first exponential blows up exponentially and the other terms are either of lower order or vanish exponentially.

3. If $a - D \frac{\pi^2}{L^2} = 0$, (2.105) becomes

$$u(x, t) = b_1 \sin \frac{\pi x}{L} + \sum_{k=2}^{\infty} b_k \exp\left\{(a - D \frac{k^2 \pi^2}{L^2})t\right\} \sin \frac{k \pi x}{L}.$$

Since $a - D k^2 \pi^2 / L^2 < 0$ if $k > 1$, we deduce that

$$u(x, t) \rightarrow b_1 \sin \frac{\pi x}{L} \quad \text{as } t \rightarrow +\infty,$$

uniformly in $[0, L]$.

³⁶ See Problem 2.1.

Now, the coefficients a and D are intrinsic parameters, encoding the features of the population and of the environment. When these parameters are fixed, the habitat size plays a major role. In fact, the value

$$L_0 = \pi \sqrt{\frac{D}{a}}$$

represents a critical value for the population survival. If $L < L_0$ the habitat is too small to avoid the extinction of the population; on the contrary, if $L > L_0$, one observes exponential growth. If $L = L_0$, diffusion and reaction *balance* and the population evolves towards the solution $b_1 \sin(\pi x/L)$ of the stationary problem

$$\begin{cases} -Du_{xx} = au & \text{in } (0, L) \\ u(0) = u(L) = 0, \end{cases}$$

called *steady state solution*.

Finally note that if for some $\bar{k} \geq 1$ we have $a - D\bar{k}\pi^2/L^2 > 0$, then for every k , $1 \leq k < \bar{k}$, all the values $a - Dk\pi^2/L^2$ are positive and the corresponding terms in the series (2.105) contributes to the exponential growth of the solution. In terms of population dynamics this means that the *vibration modes*

$$b_k \exp \left\{ \left(a - D \frac{k^2 \pi^2}{L^2} \right) t \right\} \sin \frac{k\pi x}{L}$$

for $k = 1, 2, \dots, \bar{k}$ are activated.

2.6 Multidimensional Random Walk

2.6.1 The symmetric case

What we have done in dimension $n = 1$ can be extended without much effort to dimension $n > 1$, in particular $n = 2, 3$. To define a symmetric random walk, we introduce the *lattice* \mathbb{Z}^n given by the set of points $\mathbf{x} \in \mathbb{R}^n$, whose coordinates are signed integers. Given the *space step* $h > 0$, the symbol $h\mathbb{Z}^n$ denotes the lattice of points whose coordinates are signed integers *multiplied by h*.

Every point $\mathbf{x} \in h\mathbb{Z}^n$, has a “discrete neighborhood” of $2n$ points at distance h , given by

$$\mathbf{x} + h\mathbf{e}_j \quad \text{and} \quad \mathbf{x} - h\mathbf{e}_j \quad (j = 1, \dots, n),$$

where $\{\mathbf{e}_1, \dots, \mathbf{e}_n\}$ is the canonical basis in \mathbb{R}^n . Our particle moves in $h\mathbb{Z}^n$ according to the following rules (Fig. 2.11).

1. It starts from $\mathbf{x} = \mathbf{0}$.
2. If it is located in \mathbf{x} at time t , at time $t + \tau$ the particle location is at one of the $2n$ points $\mathbf{x} \pm h\mathbf{e}_j$, with probability $p = \frac{1}{2n}$.
3. Each step is independent of the previous ones.

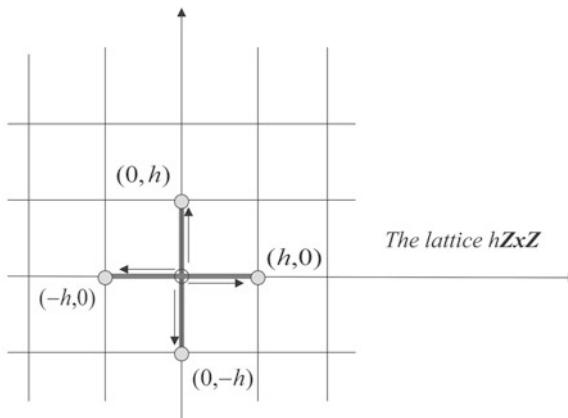


Fig. 2.11 Bidimensional random walk

As in the 1-dimensional case, our task is to compute the probability $p(\mathbf{x}, t)$ of finding the particle at \mathbf{x} at time t .

Clearly the initial conditions for p are

$$p(\mathbf{0}, 0) = 1 \quad \text{and} \quad p(\mathbf{x}, 0) = 0 \quad \text{if } \mathbf{x} \neq \mathbf{0}.$$

The total probability formula gives

$$p(\mathbf{x}, t + \tau) = \frac{1}{2n} \sum_{j=1}^n \{p(\mathbf{x} + h\mathbf{e}_j, t) + p(\mathbf{x} - h\mathbf{e}_j, t)\}. \quad (2.107)$$

Indeed, to reach the point \mathbf{x} at time $t + \tau$, at time t the particle must have been located at one of the points in the discrete neighborhood of \mathbf{x} and moved from there towards \mathbf{x} with probability $1/2n$. Keeping \mathbf{x} and t fixed, we want to examine what happens when we let $h \rightarrow 0, \tau \rightarrow 0$. Assuming that p is defined and smooth in all of $\mathbb{R}^n \times (0, +\infty)$, we use Taylor's formula to write

$$\begin{aligned} p(\mathbf{x}, t + \tau) &= p(\mathbf{x}, t) + p_t(\mathbf{x}, t) \tau + o(\tau) \\ p(\mathbf{x} \pm h\mathbf{e}_j, t) &= p(\mathbf{x}, t) \pm p_{x_j}(\mathbf{x}, t) h + \frac{1}{2} p_{x_j x_j}(\mathbf{x}, t) h^2 + o(h^2). \end{aligned}$$

Substituting into (2.107), after some simplifications, we get

$$p_t \tau + o(\tau) = \frac{h^2}{2n} \Delta p + o(h^2).$$

Dividing by τ we obtain the equation

$$p_t + o(1) = \frac{1}{2n} \frac{h^2}{\tau} \Delta p + o\left(\frac{h^2}{\tau}\right). \quad (2.108)$$

The situation is quite similar to the 1–dimensional case: still, to obtain eventually something nontrivial, we must require that the ratio h^2/τ has a finite and positive limit. The simplest choice is

$$\frac{h^2}{\tau} = 2nD \quad (2.109)$$

with $D > 0$. From (2.109), we deduce that *in unit time, the particle diffuses up to an average distance $\sqrt{2nD}$.* The physical dimensions of D have not changed. Letting $h \rightarrow 0, \tau \rightarrow 0$ in (2.108), we find for p the diffusion equation

$$p_t = D\Delta p, \quad (2.110)$$

with the initial conditions

$$\lim_{t \rightarrow 0^+} p(\mathbf{x}, t) = \delta_3(\mathbf{x}). \quad (2.111)$$

Since $\int_{\mathbb{R}^n} p(\mathbf{x}, t) d\mathbf{x} = 1$ for every t , the unique solution is given by

$$p(\mathbf{x}, t) = \Gamma_D(\mathbf{x}, t) = \frac{1}{(4\pi Dt)^{n/2}} e^{-\frac{|\mathbf{x}|^2}{4Dt}}, \quad t > 0.$$

The n –dimensional random walk has become a continuous walk; when $D = \frac{1}{2}$, it is called *n –dimensional Brownian motion*. Denote by $\mathbf{B}(t) = \mathbf{B}(t, \omega)$ the random position of a Brownian particle, defined for every $t > 0$ on a probability space (Ω, \mathcal{F}, P) ³⁷.

The family of random variables $\mathbf{B}(t, \omega)$, with time t as a real parameter, is a **vector valued continuous stochastic process**. For $\omega \in \Omega$ fixed, the *vector* function

$$t \mapsto \mathbf{B}(t, \omega)$$

describes an n –dimensional Brownian path, whose main features are listed below.

- *Path continuity.* With probability 1, the Brownian paths are continuous for $t \geq 0$.
- *Gaussian law for increments.* The process

$$\mathbf{B}^{\mathbf{x}}(t) = \mathbf{x} + \mathbf{B}(t)$$

defines a Brownian motion with start at \mathbf{x} . With every point \mathbf{x} is associated a probability $P^{\mathbf{x}}$, with the following properties (if $\mathbf{x} = 0$, $P^0 = P$).

- a) $P^{\mathbf{x}}\{\mathbf{B}^{\mathbf{x}}(0) = \mathbf{x}\} = P\{\mathbf{B}(0) = \mathbf{0}\} = 1$.
- b) For every $s \geq 0, t \geq 0$, the increment

$$\mathbf{B}^{\mathbf{x}}(t+s) - \mathbf{B}^{\mathbf{x}}(s) = \mathbf{B}(t+s) - \mathbf{B}(s) \quad (2.112)$$

follows a *normal law with zero mean value and covariance matrix equal to tI_n ,*

³⁷ See Appendix B.

whose density is

$$\Gamma(\mathbf{x}, t) = \Gamma_{\frac{1}{2}}(\mathbf{x}, t) = \frac{1}{(2\pi t)^{n/2}} e^{-\frac{|\mathbf{x}|^2}{2t}}.$$

Moreover, (2.112) is independent of any event occurred at any time less than s . For instance, the two events

$$\{\mathbf{B}(t_2) - \mathbf{B}(t_1) \in A_1\} \quad \{\mathbf{B}(t_1) - \mathbf{B}(t_0) \in A_2\}$$

are independent if $t_0 < t_1 < t_2$.

- *Transition function.* For each Borel set $A \subseteq \mathbb{R}^n$ a *transition function*

$$P(\mathbf{x}, t, A) = P^{\mathbf{x}} \{ \mathbf{B}^{\mathbf{x}}(t) \in A \}$$

is defined, representing the probability that the particle, initially located at \mathbf{x} , belongs to A at time t . We have:

$$P(\mathbf{x}, t, A) = P \{ \mathbf{B}(t) \in A - \mathbf{x} \} = \int_{A - \mathbf{x}} \Gamma(\mathbf{y}, t) d\mathbf{y} = \int_A \Gamma(\mathbf{y} - \mathbf{x}, t) d\mathbf{y}.$$

- *Invariance.* The motion is invariant with respect to rotations and translations.
- *Markov and strong Markov properties.* Let μ be a probability measure³⁸ on \mathbb{R}^n . If the particle has a random initial position with probability distribution μ , we can consider a *Brownian motion with initial distribution* μ , and for it we use the symbol \mathbf{B}^μ . To \mathbf{B}^μ is associated a probability distribution P^μ such that

$$P^\mu \{ \mathbf{B}^\mu(0) \in A \} = \mu(A).$$

The probability that the particle belongs to A at time t can be computed through the formula

$$P^\mu \{ \mathbf{B}^\mu(t) \in A \} = \int_{\mathbb{R}^n} P(\mathbf{x}, t, A) \mu(d\mathbf{x}). \quad (2.113)$$

The *Markov property* can be stated as follows: given any condition H , related to the behavior of the particle before time $s \geq 0$, the process $\mathbf{Y}(t) = \mathbf{B}^{\mathbf{x}}(t+s)$ is a Brownian motion with initial distribution

$$\mu(A) = P^{\mathbf{x}} \{ \mathbf{B}^{\mathbf{x}}(s) \in A | H \}.$$

Again, this property establishes the independence of the *future* process $\mathbf{B}^{\mathbf{x}}(t+s)$ from the *past*, when the *present* $\mathbf{B}^{\mathbf{x}}(s)$ is known and encodes the *absence of memory* of the process. In the strong Markov property, a stopping time τ takes the place of s .

³⁸ See Appendix B for the definition of a probability measure μ and of the integral with respect to the measure μ .

- *Expectation.* Given any sufficiently smooth real function $g = g(\mathbf{y})$, $\mathbf{y} \in \mathbb{R}^n$, we can define the real random variable

$$Z(t) = (g \circ \mathbf{B}^{\mathbf{x}})(t) = g(\mathbf{B}^{\mathbf{x}}(t)).$$

Its *expectation* is given by the formula

$$E[Z(t)] = \int_{\mathbb{R}^n} g(\mathbf{y}) P(\mathbf{x}, t, d\mathbf{y}) = \int_{\mathbb{R}^n} g(\mathbf{y}) \Gamma(\mathbf{y} - \mathbf{x}, t) d\mathbf{y}.$$

2.6.2 Walks with drift and reaction

As in the 1-dimensional case, we can construct several variants of the symmetric random walk. For instance, we can allow a different behavior along each direction, by choosing the spatial step h_j depending on \mathbf{e}_j . As a consequence the limit process models an anisotropic motion, codified in the matrix

$$\mathbf{D} = \begin{pmatrix} D_1 & 0 & \cdots & 0 \\ 0 & D_2 & & 0 \\ \vdots & & \ddots & \vdots \\ 0 & 0 & \cdots & D_n \end{pmatrix}$$

where $D_j = h_j^2/2n\tau$ is the diffusion coefficient in the direction \mathbf{e}_j . The resulting equation for the transition probability $p(\mathbf{x}, t)$ is

$$p_t = \sum_{j=1}^n D_j p_{x_j x_j}. \quad (2.114)$$

We may also break the symmetry by asking that along the direction \mathbf{e}_j the probability to go to the left (right) is q_j (resp. p_j) with $\sum_{j=1}^n (p_j + q_j) = 1$. If

$$\frac{q_j - p_j}{h_j} \rightarrow \beta_j \quad \text{and} \quad b_j = 2D_j \beta_j,$$

the vector $\mathbf{b} = (b_1, \dots, b_n)$ plays a role of a *drift vector*, reflecting the tendency of motion to move asymmetrically along each coordinate axis. Adding a reaction term of the form cp , the resulting *drift-reaction-diffusion* equation is

$$p_t = \sum_{j=1}^n D_j p_{x_j x_j} + \sum_{j=1}^n b_j u_{x_j} + cp. \quad (2.115)$$

We let the reader to fill in all the details in the argument leading to equations (2.114) and (2.115). We will deal with general equations of these type in Chap. 10.

2.7 An Example of Reaction–Diffusion in Dimension $n = 3$

In this section we examine a model of reaction-diffusion in a fissionable material. Although we deal with a greatly simplified model, some interesting implications can be drawn.

By shooting neutrons into an uranium nucleus it may happen that the nucleus breaks into two parts, releasing other neutrons already present in the nucleus and causing a chain reaction. Some macroscopic aspects of this phenomenon can be described by means of an elementary model.

Suppose that a cylinder with height h and radius R is made of a fissionable material of constant density ρ , with total mass $M = \pi\rho R^2 h$. At a macroscopic level, the free neutrons diffuse like a chemical in a porous medium, with a flux proportional and opposite to its density gradient. In other terms, if $N = N(x, y, z, t)$ is the *neutron density* and no fission occurs, the *flux of neutrons is equal to $-\kappa\nabla N$* , where κ is a positive constant depending on the material. The mass conservation law then gives $N_t = \kappa\Delta N$. When fission occurs at a constant rate $\gamma > 0$, we get the equation

$$N_t = \kappa\Delta N + \gamma N, \quad (2.116)$$

where reaction and diffusion are competing: diffusion tends to slow down N , while, clearly, the reaction term tends to exponentially increase N . A crucial question is to examine the behavior of N in the long run (i.e. as $t \rightarrow +\infty$).

We look for *bounded* solutions satisfying a homogeneous Dirichlet condition on the boundary of the cylinder, with the idea that the density is higher at the center of the cylinder and very low near the boundary. Then, it is reasonable to assume that N has a radial distribution with respect to the axis of the cylinder. More precisely, using the cylindrical coordinates (r, θ, z) , with z along the axis of the cylinder, and

$$x = r \cos \theta, \quad y = r \sin \theta,$$

we can write $N = N(r, z, t)$ and the homogeneous Dirichlet condition on the boundary of the cylinder translates into

$$\begin{aligned} N(R, z, t) &= 0 & 0 < z < h \\ N(r, 0, t) &= N(r, h, t) = 0 & 0 < r < R, \end{aligned} \quad (2.117)$$

for every $t > 0$. Accordingly we prescribe an initial condition

$$N(r, z, 0) = N_0(r, z) \quad (2.118)$$

such that

$$N_0(R, z) = 0 \text{ for } 0 < z < h, \text{ and } N_0(r, 0) = N_0(r, h) = 0. \quad (2.119)$$

To solve problem (2.116), (2.117), (2.118), first we get rid of the reaction term by setting

$$N(r, z, t) = \mathcal{N}(r, z, t) e^{\gamma t}. \quad (2.120)$$

Then, writing the Laplace operator in cylindrical coordinates³⁹, \mathcal{N} solves the equation

$$\mathcal{N}_t = \kappa \left[\mathcal{N}_{rr} + \frac{1}{r} \mathcal{N}_r + \mathcal{N}_{zz} \right] \quad (2.121)$$

with the same initial and boundary conditions of N . By maximum principle, we know that there exists only one solution, continuous up to the boundary of the cylinder. To find an explicit formula for the solution, we use the method of separation of variables, first searching for bounded solutions of the form

$$\mathcal{N}(r, z, t) = u(r)v(z)w(t), \quad (2.122)$$

satisfying the homogeneous Dirichlet conditions $u(R) = 0$ and $v(0) = v(h) = 0$.

Substituting (2.122) into (2.121), we find

$$u(r)v(z)w'(t) = \kappa[u''(r)v(z)w(t) + \frac{1}{r}u'(r)v(z)w(t) + u(r)v''(z)w(t)].$$

Dividing by \mathcal{N} and rearranging the terms, we get,

$$\frac{w'(t)}{Dw(t)} - \left[\frac{u''(r)}{u(r)} + \frac{1}{r} \frac{u'(r)}{u(r)} \right] = \frac{v''(z)}{v(z)}. \quad (2.123)$$

The two sides of (2.123) depend on different variables so that they must be equal to a common constant b . Then for v we have the eigenvalue problem

$$\begin{aligned} v''(z) - bv(z) &= 0 \\ v(0) = v(h) &= 0. \end{aligned}$$

The eigenvalues are $b_m \equiv -\nu_m^2 = -\frac{m^2\pi^2}{h^2}$, $m \geq 1$ integer, with corresponding eigenfunctions

$$v_m(z) = \sin \nu_m z.$$

The equation for w and u can be written in the form:

$$\frac{w'(t)}{Dw(t)} + \nu_m^2 = \frac{u''(r)}{u(r)} + \frac{1}{r} \frac{u'(r)}{u(r)} \quad (2.124)$$

where the variables r and t are again separated. This forces the two sides of (2.124) to be equal to a common constant μ . Then the equation for u is

$$u''(r) + \frac{1}{r}u'(r) - \mu u(r) = 0 \quad (2.125)$$

with

$$u(R) = 0 \quad \text{and} \quad u \text{ bounded in } [0, R]. \quad (2.126)$$

³⁹ See Appendix C.

The (2.125) is a *Bessel equation of order zero with parameter $-\mu$* (see Sect. 6.4.2); conditions (2.126) force⁴⁰ $\mu = -\lambda^2 < 0$. The only bounded solution of (2.125), (2.126) is $J_0(\lambda r)$, where

$$J_0(x) = \sum_{k=0}^{\infty} \frac{(-1)^k}{(k!)^2} \left(\frac{x}{2}\right)^{2k}$$

is the *Bessel function of first kind and order zero*. To match the boundary condition $u(R) = 0$ we require $J_0(\lambda R) = 0$. Now, J_0 has an infinite number of positive simple zeros⁴¹ λ_n , $n \geq 1$:

$$0 < \lambda_1 < \lambda_2 < \dots < \lambda_n < \dots$$

Thus, if $\lambda R = \lambda_n$, we find infinitely many solutions of (2.125), given by

$$u_n(r) = J_0\left(\frac{\lambda_n r}{R}\right)$$

and $\mu = \mu_n = -\lambda_n^2/R^2$. Finally, for w we have the equation

$$w'(t) = D(\mu_n - \nu_m^2)w(t)$$

that gives

$$w_{mn}(t) = \exp[D(\mu_n - \nu_m^2)t], \quad c \in \mathbb{R}.$$

To summarize, we have determined so far a countable number of solutions of the differential equation, namely

$$\begin{aligned} \mathcal{N}_{mn}(r, z, t) &= u_n(r)v_m(z)w_{mn}(t) = \\ &= J_0\left(\frac{\lambda_n r}{R}\right)\sin\nu_m z \exp\left[-\kappa\left(\nu_m^2 + \frac{\lambda_n^2}{R^2}\right)t\right], \end{aligned}$$

all satisfying the homogeneous Dirichlet conditions. It remains to satisfy the initial condition. Due to the linearity of the problem, we look for a solution obtained by

⁴⁰ In fact, write Bessel's equation (2.125) in the form $(ru')' - \mu ru = 0$. Multiplying by u and integrating over $(0, R)$, we have

$$\int_0^R (ru')' u dr = \mu \int_0^R u^2 dr. \quad (2.127)$$

Integrating by parts and using (2.126), we get

$$\int_0^R (ru')' u dr = [(ru')u]_0^R - \int_0^R (u')^2 dr = - \int_0^R (u')^2 dr < 0$$

and from (2.127) we get $\mu < 0$.

⁴¹ The zeros of the Bessel functions are known with a considerable degree of accuracy. The first five zeros of J_0 are: 2.4048..., 5.5201..., 8.6537..., 11.7915..., 14.9309....

superposition of the \mathcal{N}_{mn} , that is

$$\mathcal{N}(r, z, t) = \sum_{n,m=1}^{\infty} c_{mn} \mathcal{N}_{mn}(r, z, t).$$

Then, we choose the coefficients c_{mn} in order to have

$$\sum_{n,m=1}^{\infty} c_{mn} \mathcal{N}_{mn}(r, z, 0) = \sum_{n,m=1}^{\infty} c_{mn} J_0\left(\frac{\lambda_n r}{R}\right) \sin \frac{m\pi}{h} z = N_0(r, z). \quad (2.128)$$

Since $N_0(r, 0) = N_0(r, h) = 0$, formula (2.128) suggests an expansion of N_0 in sine Fourier series with respect to z . Let

$$c_m(r) = \frac{2}{h} \int_0^h N_0(r, z) \sin \frac{m\pi}{h} z \, dz, \quad m \geq 1,$$

and

$$N_0(r, z) = \sum_{m=1}^{\infty} c_m(r) \sin \frac{m\pi}{h} z.$$

Then (2.128) shows that, for fixed $m \geq 1$, the c_{mn} are the coefficients of the expansion of $c_m(r)$ in the *Fourier-Bessel series*

$$\sum_{n=1}^{\infty} c_{mn} J_0\left(\frac{\lambda_n r}{R}\right) = c_m(r).$$

We are not really interested in the exact formula for the c_{mn} , however we will come back to this point in Remark 2.10 below.

In conclusion, recalling (2.120), the analytic expression of the solution of our original problem is the following:

$$N(r, z, t) = \sum_{n,m=1}^{\infty} c_{mn} J_0\left(\frac{\lambda_n r}{R}\right) \exp\left\{\left(\gamma - \kappa\nu_m^2 - \kappa \frac{\lambda_n^2}{R^2}\right)t\right\} \sin \nu_m z. \quad (2.129)$$

Of course, (2.129) is only a formal solution, since we should check in which sense the boundary and initial condition are attained and that term by term differentiation can be performed. This can be done under reasonable smoothness properties of N_0 and we do not pursue the calculations here.

Rather, we notice that from (2.129) we can draw an interesting conclusion on the long range behavior of N . Consider for instance the value of N at the center of the cylinder, that is at the point $r = 0$ and $z = h/2$; we have, since $J_0(0) = 1$

2.7 An Example of Reaction–Diffusion in Dimension $n = 3$

75

and $\nu_m^2 = \frac{m^2\pi^2}{h^2}$,

$$N\left(0, \frac{h}{2}, t\right) = \sum_{n,m=1}^{\infty} c_{mn} \exp\left\{\left(\gamma - \kappa \frac{m^2\pi^2}{h^2} - \kappa \frac{\lambda_n^2}{R^2}\right)t\right\} \sin \frac{m\pi}{2}.$$

The exponential factor is maximized for $m = n = 1$, so, assuming $c_{11} \neq 0$, the leading term in the sum is

$$c_{11} \exp\left\{\left(\gamma - \kappa \frac{\pi^2}{h^2} - \kappa \frac{\lambda_1^2}{R^2}\right)t\right\}.$$

If now $\gamma - \kappa \left(\frac{\pi^2}{h^2} + \frac{\lambda_1^2}{R^2}\right) < 0$, each term in the series goes to zero as $t \rightarrow +\infty$ and the reaction dies out. On the opposite, if

$$\gamma - \kappa \left(\frac{\pi^2}{h^2} + \frac{\lambda_1^2}{R^2}\right) > 0,$$

that is

$$\frac{\gamma}{\kappa} > \frac{\pi^2}{h^2} + \frac{\lambda_1^2}{R^2}, \quad (2.130)$$

the leading term increases exponentially with time. To be true, (2.130) requires that the following relations be both satisfied:

$$h^2 > \frac{\kappa\pi^2}{\gamma} \quad \text{and} \quad R^2 > \frac{\kappa\lambda_1^2}{\gamma}. \quad (2.131)$$

Equations (2.131) give a lower bound for the height and the radius of the cylinder. Thus we deduce that *there exists a critical mass of material, below which the reaction cannot be sustained*.

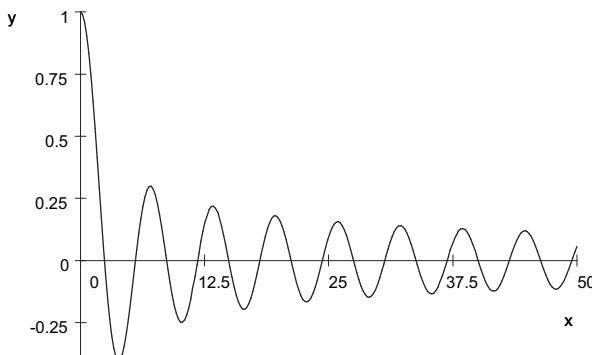


Fig. 2.12 The Bessel function J_0

Remark 2.10. A sufficiently smooth function f , for instance of class $C^1([0, R])$, can be expanded in a Fourier-Bessel series, where the Bessel functions $J_0\left(\frac{\lambda_n r}{R}\right)$, $n \geq 1$, play the same role of the trigonometric functions. More precisely⁴², setting $R = 1$ for simplicity, the functions $J_0(\lambda_n r)$ satisfy the following orthogonality relations:

$$\int_0^1 x J_0(\lambda_m x) J_0(\lambda_n x) dx = \begin{cases} 0 & m \neq n \\ \frac{1}{2} c_n^2 & m = n \end{cases}$$

where

$$c_n = \sum_{k=0}^{\infty} \frac{(-1)^k}{k! (k+1)!} \left(\frac{\lambda_n}{2}\right)^{2k+1}.$$

Then

$$f(x) = \sum_{n=0}^{\infty} f_n J_0(\lambda_n x) \quad (2.132)$$

with the coefficients f_n assigned by the formula

$$f_n = \frac{2}{c_n^2} \int_0^1 x f(x) J_0(\lambda_n x) dx.$$

The series (2.132) converges in the following least square sense: if

$$S_N(x) = \sum_{n=0}^N f_n J_0(\lambda_n x)$$

then

$$\lim_{N \rightarrow +\infty} \int_0^1 [f(x) - S_N(x)]^2 x dx = 0. \quad (2.133)$$

In Chap. 6, we will interpret (2.133) from the point of view of Hilbert space theory.

2.8 The Global Cauchy Problem ($n = 1$)

2.8.1 The homogeneous case

In this section we consider the global Cauchy problem

$$\begin{cases} u_t - Du_{xx} = 0 & \text{in } \mathbb{R} \times (0, \infty) \\ u(x, 0) = g(x) & \text{in } \mathbb{R}, \end{cases} \quad (2.134)$$

where g , the *initial datum*, is given. We confine ourselves to the one dimensional case; techniques, ideas and formulas can be extended without too much effort to the n -dimensional case.

⁴² See e.g. [23], *R. Courant, D. Hilbert*, vol. 1, 1953.

The problem (2.134) models the evolution of the temperature or of the concentration of a substance along a very long (infinite) bar or channel, respectively, given the initial ($t = 0$) distribution.

By heuristic considerations, we can guess what could be a candidate solution. Interpret u as a linear mass concentration. Then, $u(x, t) dx$ gives the mass inside the interval $(x, x + dx)$ at time t .

We want to determine the evolution of $u(x, t)$, due to the diffusion of a mass whose initial concentration is given by g .

Thus, the quantity $g(y) dy$ represents the mass concentrated in the interval $(y, y + dy)$ at time $t = 0$. As we have seen, $\Gamma(x - y, t)$ is a *unit source solution*, representing the concentration at x at time t , due to the diffusion of a unit mass, initially concentrated in the same interval. Accordingly,

$$\Gamma_D(x - y, t) g(y) dy$$

gives the concentration at x at time t , due to the diffusion of the mass $g(y) dy$.

Thanks to the linearity of the diffusion equation, we can use the *superposition principle* and compute the solution as the sum of all contributions. In this way, we get the formula

$$u(x, t) = \int_{\mathbb{R}} \Gamma_D(x - y, t) g(y) dy = \frac{1}{\sqrt{4\pi D t}} \int_{\mathbb{R}} e^{-\frac{(x-y)^2}{4Dt}} g(y) dy. \quad (2.135)$$

Clearly, one has to check rigorously that, under reasonable hypotheses on the initial datum g , formula (2.135) really gives the unique solution of the Cauchy problem. This is not a negligible question. First of all, if g grows too much at infinity, more than an exponential of the type e^{ax^2} , $a > 0$, in spite of the rapid convergence to zero of the Gaussian, the integral in (2.135) could be divergent and formula (2.135) loses any meaning.

Even more delicate is the question of the uniqueness of the solution, as we will see later.

Remark 2.11. Formula (2.135) has a probabilistic interpretation. Let $D = \frac{1}{2}$ and let $B^x(t)$ be the position of a Brownian particle, started at x . Let $g(y)$ be the gain obtained when the particle crosses y . Then, we can write:

$$u(x, t) = E^x [g(B^x(t))]$$

where E^x denotes the *expected value* with respect to the probability P^x , with density $\Gamma(x - y, t)$ ⁴³.

In other words: *to compute u at the point (x, t) , consider a Brownian particle starting at x , compute its position $B^x(t)$ at time t , and finally compute the expected value of $g(B^x(t))$.*

⁴³ See Appendix B.

2.8.2 Existence of a solution

The following theorem states that (2.135) is indeed a solution of the global Cauchy problem under rather general hypotheses on g , satisfied in most of the interesting applications. Although it is a little bit technical, the proof can be found at the end of this section.

Theorem 2.12. *Assume that there exist positive numbers a and c such that*

$$|g(x)| \leq ce^{ax^2} \quad \text{for all } x \in \mathbb{R}. \quad (2.136)$$

Let u be given by formula (2.135) and $T < \frac{1}{4aD}$. Then, the following properties hold.

i) *There are positive numbers c_1 and A such that*

$$|u(x, t)| \leq Ce^{Ax^2} \quad \text{for all } (x, t) \in \mathbb{R} \times (0, T].$$

ii) *$u \in C^\infty(\mathbb{R} \times (0, T])$ and in the strip $\mathbb{R} \times (0, T]$*

$$u_t - Du_{xx} = 0.$$

iii) *Let $(x, t) \rightarrow (x_0, 0^+)$. If g is continuous at x_0 then $u(x, t) \rightarrow g(x_0)$.*

Remark 2.13. The theorem says that, if we allow an initial data with a controlled exponential growth at infinity expressed by (2.136), then (2.135) is a solution in the strip $\mathbb{R} \times (0, T)$. We will see that, under the stated conditions, (2.135) is actually the unique solution.

In some applications (e.g. to Finance), the initial datum grows at infinity no more than $c_1 e^{a_1 |x|}$. In this case (2.136) is satisfied by choosing any positive number a and a suitable c . This means that there is really no limitation on T , since

$$T < \frac{1}{4Da}$$

and a can be chosen as small as we like.

Remark 2.14. The property ii) shows a typical and important phenomenon connected with the diffusion equation: even if the initial data is discontinuous at some point, immediately after the solution is smooth. The diffusion is therefore a **smoothing process**.

In Fig. 2.13, this phenomenon is shown for the initial data $g(x) = \chi_{(-2,0)}(x) + \chi_{(1,4)}(x)$, where $\chi_{(a,b)}$ denotes the characteristic function of the interval (a, b) . By iii), if the initial data g is continuous in all of \mathbb{R} , then the solution is continuous up to $t = 0$, that is in $\mathbb{R} \times [0, T)$.

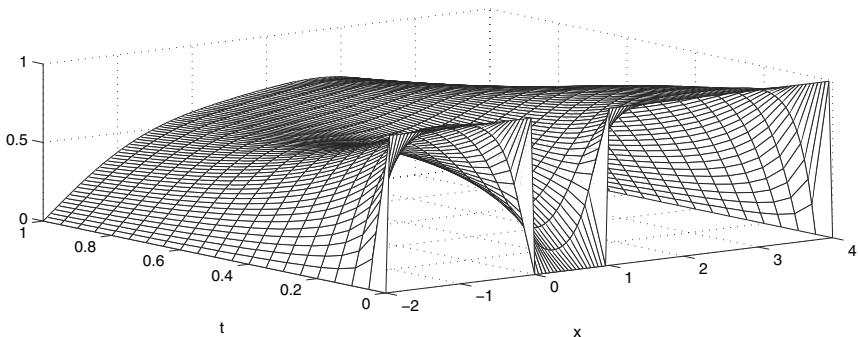


Fig. 2.13 Smoothing effect of the diffusion equation

2.8.3 The nonhomogeneous case. Duhamel's method

The difference equation (or the total probability formula)

$$p(x, t + \tau) = \frac{1}{2}p(x - h, t) + \frac{1}{2}p(x + h, t),$$

that we found in Sect. 4.2 during the analysis of the symmetric random walk, could be considered a probabilistic version of the *mass conservation principle*: the mass located at x at time $t + \tau$ is the sum of the masses diffused from $x + h$ and $x - h$ at time t ; no mass has been lost or added over the time interval $[t, t + \tau]$. Accordingly, the expression

$$p(x, t + \tau) - [\frac{1}{2}p(x - h, t) + \frac{1}{2}p(x + h, t)] \quad (2.137)$$

could be considered as a measure of the lost/added mass over the time interval from t to $t + \tau$. Expanding with Taylor's formula as we did in Sect. 4.2, keeping $h^2/\tau = 2D$, dividing by τ and letting $h, \tau \rightarrow 0$ in (2.137), we find

$$p_t - Dp_{xx}.$$

Thus the differential operator $\partial_t - D\partial_{xx}$ measures the instantaneous density production rate.

Suppose now that, from time $t = 0$ until a certain time $t = s > 0$, no mass is present and that at time s a unit mass at the point y (infinite density) appears. We know that we can model this kind of source by means of a Dirac measure at y , that has to be time dependent since the mass appears only at time s . We can write it in the form

$$\delta_2(x - y, t - s).$$

Thus, we are led to the nonhomogeneous equation

$$p_t - Dp_{xx} = \delta_2(x - y, t - s)$$

with $p(x, 0) = 0$ as initial condition. What could be the solution? Until $t = s$ nothing happens and *after* s we have $\delta_2(x - y, t - s) = 0$. Therefore it is like starting from time $t = s$ and solving the problem

$$p_t - Dp_{xx} = 0, \quad x \in \mathbb{R}, t > s$$

with initial condition

$$p(x, s) = \delta_2(x - y, t - s).$$

We have solved this problem when $s = 0$; the solution is $\Gamma_D(x - y, t)$. By the time translation invariance of the diffusion equation, we deduce that the solution for any $s > 0$ is given by

$$p(x, t) = \Gamma_D(x - y, t - s). \quad (2.138)$$

Consider now a distributed source on the half-plane $t > 0$, capable to produce mass density at the time rate $f(x, t)$. Precisely, $f(x, t) dx dt$ is the mass produced⁴⁴ between x and $x+dx$, over the time interval $(t, t+dt)$. If initially no mass is present, we are lead to the *nonhomogeneous Cauchy problem*

$$\begin{cases} v_t - Dv_{xx} = f(x, t) & \text{in } \mathbb{R} \times (0, T) \\ v(x, 0) = 0 & \text{in } \mathbb{R}. \end{cases} \quad (2.139)$$

As in Sect. 2.8.1, we motivate the form of the solution at the point (x, t) using heuristic considerations.

Let us compute the contribution dv to $v(x, t)$ due to a mass $f(y, s) dy ds$. It is like having a source term of the form

$$f^*(x, t) = f(x, t) \delta_2(x - y, t - s)$$

and therefore, recalling (2.138), we have

$$dv(x, t) = \Gamma_D(x - y, t - s) f(y, s) dy ds. \quad (2.140)$$

We obtain the solution $v(x, t)$ by superposition, summing all the contributions (2.140). We split it into the following two steps:

- We sum over y the contributions for fixed s , to get the total density at (x, t) , due to the diffusion of mass produced in the time interval $(s, s+ds)$. The result is $w(x, t; s) ds$, where

$$w(x, t; s) = \int_{\mathbb{R}} \Gamma_D(x - y, t - s) f(y, s) dy. \quad (2.141)$$

⁴⁴ Negative production ($f < 0$) means removal.

- We sum the above contributions for s ranging from 0 to t :

$$v(x, t) = \int_0^t \int_{\mathbb{R}} \Gamma_D(x - y, t - s) f(y, s) dy ds.$$

The above construction is an example of application of the *Duhamel's method*, that we state below:

Duhamel's method. The procedure to compute the solution u of problem (2.139) at the point (x, t) consists in the following two steps:

1. Construct a family of solutions of homogeneous Cauchy problems, with variable initial time s , $0 \leq s \leq t$, and initial data $f(x, s)$.
2. Integrate the above family with respect to s , over $(0, t)$.

Indeed, let us examine the two steps.

1. Consider the family of homogeneous Cauchy problems

$$\begin{cases} w_t - Dw_{xx} = 0 & x \in \mathbb{R}, t > s \\ w(x, s; s) = f(x, s) & x \in \mathbb{R}, \end{cases} \quad (2.142)$$

where the initial time s plays the role of a parameter.

The function $\Gamma^{y,s}(x, t) = \Gamma_D(x - y, t - s)$ is the fundamental solution of the diffusion equation that satisfies for $t = s$ the initial condition

$$\Gamma^{y,s}(x, s) = \delta(x - y).$$

Hence, the solution of (2.142) is given by the function (2.141):

$$w(x, t; s) = \int_{\mathbb{R}} \Gamma_D(x - y, t - s) f(y, s) dy.$$

Thus, $w(x, t; s)$ is the required family.

2. Integrating w over $(0, t)$ with respect to s , we find

$$v(x, t) = \int_0^t w(x, t; s) ds = \int_0^t \int_{\mathbb{R}} \Gamma_D(x - y, t - s) f(y, s) dy ds. \quad (2.143)$$

Using (2.142) we have

$$v_t - Dv_{xx} = w(x, t; t) + \int_0^t [w_t(x, t; s) - Dw_{xx}(x, t; s)] ds = f(x, t).$$

Moreover, $v(x, 0) = 0$ and therefore v is a solution to (2.139).

Everything works under rather mild hypotheses on f . More precisely⁴⁵:

⁴⁵ For a proof, see [12], *McOwen*, 1996.

Theorem 2.15. If f and its derivatives f_t, f_x, f_{xx} are continuous and bounded in $\mathbb{R} \times [0, T]$, then (2.143) gives a solution v of problem (2.139) in $\mathbb{R} \times [0, T]$, continuous up to $t = 0$, with derivatives v_t, v_x, v_{xx} continuous in $\mathbb{R} \times (0, T)$.

The formula for the general Cauchy problem

$$\begin{cases} u_t - Du_{xx} = f(x, t) & \text{in } \mathbb{R} \times (0, T) \\ u(x, 0) = g(x) & \text{in } \mathbb{R} \end{cases} \quad (2.144)$$

is obtained by superposition of (2.135) and (2.139). Under the hypotheses on f and g stated in Theorems 2.12 and 2.15, the function

$$u(x, t) = \int_{\mathbb{R}} \Gamma_D(x - y, t) g(y) dy + \int_0^t \int_{\mathbb{R}} \Gamma(x - y, t - s) f(y, s) dy ds$$

is a solution of (2.144) in $\mathbb{R} \times (0, T)$,

$$T < \frac{1}{4Da},$$

continuous with its derivatives u_t, u_x, u_{xx} .

The initial condition means that

$$u(x, t) \rightarrow g(x_0) \quad \text{as} \quad (x, t) \rightarrow (x_0, 0)$$

at any point x_0 of continuity of g . In particular, if g is continuous in \mathbb{R} then u is continuous in $\mathbb{R} \times [0, T]$.

2.8.4 Global maximum principles and uniqueness.

The uniqueness of the solution to the global Cauchy problem is still to be discussed. This is not a trivial question since the following counterexample of Tychonov shows that there could be several solutions of the homogeneous problem. Let

$$h(t) = \begin{cases} e^{-t^{-2}} & \text{for } t > 0 \\ 0 & \text{for } t \leq 0. \end{cases}$$

It can be checked⁴⁶ that the function

$$T(x, t) = \sum_{k=0}^{\infty} \frac{x^{2k}}{(2k)!} \frac{d^k}{dt^k} h(t)$$

is a solution of

$$u_t - u_{xx} = 0 \quad \text{in } \mathbb{R} \times (0, +\infty)$$

with

$$u(x, 0) = 0 \quad \text{in } \mathbb{R}.$$

⁴⁶ Not an easy task! See [7], F. John, 1982.

Since also

$$u(x, t) \equiv 0$$

is a solution of the same problem, we conclude that, in general, the Cauchy problem is *not well posed*.

What is wrong with \mathcal{T} ? It grows too much at infinity for small times. Indeed the best estimate available for \mathcal{T} is the following:

$$|\mathcal{T}(x, t)| \leq C \exp \left\{ \frac{x^2}{\theta t} \right\} \quad (\theta > 0)$$

that quickly deteriorates when $t \rightarrow 0^+$, due to the factor $1/\theta t$.

If instead of $1/\theta t$ we had a constant a , as in condition *i*) of Theorem 2.12, p. 78, then we can assure uniqueness.

In other words, among the class of functions with growth at infinity controlled by an exponential of the type Ce^{Ax^2} for any $t \geq 0$ (the so called *Tychonov class*), the solution to the homogeneous Cauchy problem is unique.

This is a consequence of the following maximum principle.

Theorem 2.16. *Global maximum principle.* Let z be continuous in $\mathbb{R} \times [0, T]$, with derivatives z_x, z_{xx}, z_t continuous in $\mathbb{R} \times (0, T)$, such that, in $\mathbb{R} \times (0, T)$:

$$z_t - Dz_{xx} \leq 0 \quad (\text{resp. } \geq 0)$$

and

$$z(x, t) \leq Ce^{Ax^2}, \quad (\text{resp. } \geq -Ce^{Ax^2}) \quad (2.145)$$

where $C > 0$. Then

$$\sup_{\mathbb{R} \times [0, T]} z(x, t) \leq \sup_{\mathbb{R}} z(x, 0) \quad \left(\text{resp. } \inf_{\mathbb{R} \times [0, T]} z(x, t) \geq \inf_{\mathbb{R}} z(x, 0) \right).$$

The proof is rather difficult⁴⁷, but if we assume that z is bounded from above or below ($A = 0$ in (2.145)), then the proof relies on a simple application of the weak maximum principle, Theorem 2.4, p. 36. In Problem 2.15 we ask the reader to fill in the details of the proof.

We now are in position to prove the following uniqueness result.

Corollary 2.17. *Uniqueness I.* Suppose u is a solution of

$$\begin{cases} u_t - Du_{xx} = 0 & \text{in } \mathbb{R} \times (0, T) \\ u(x, 0) = 0 & \text{in } \mathbb{R}, \end{cases}$$

continuous in $\mathbb{R} \times [0, T]$, with derivatives u_x, u_{xx}, u_t continuous in $\mathbb{R} \times (0, T)$. If $|u|$ satisfies (2.145) then $u \equiv 0$.

⁴⁷ See [7], F. John, 1982.

Proof. From Theorem 2.16 we have

$$0 = \inf_{\mathbb{R}} u(x, 0) \leq \inf_{\mathbb{R} \times [0, T]} u(x, t) \leq \sup_{\mathbb{R} \times [0, T]} u(x, t) \leq \sup_{\mathbb{R}} u(x, 0) = 0$$

so that $u \equiv 0$. \square

Notice that if

$$|g(x)| \leq ce^{ax^2} \quad \text{for every } x \in \mathbb{R} \quad (c, a \text{ positive}), \quad (2.146)$$

we know from Theorem 2.12 that

$$u(x, t) = \int_{\mathbb{R}} \Gamma_D(x - y, t) g(y) dy$$

satisfies the estimate

$$|u(x, t)| \leq Ce^{Ax^2} \quad \text{in } \mathbb{R} \times (0, T] \quad (2.147)$$

and therefore it belongs to the Tychonov class in $\mathbb{R} \times (0, T]$, for $T < 1/(4Da)$. Moreover, if f is as in Theorem 2.15 and

$$v(x, t) = \int_0^t \int_{\mathbb{R}} \Gamma_D(x - y, t - s) f(y, s) dy ds,$$

we easily get the estimate

$$t \inf_{\mathbb{R} \times [0, T)} f \leq v(x, t) \leq t \sup_{\mathbb{R} \times [0, T)} f, \quad (2.148)$$

for every $x \in \mathbb{R}$, $0 \leq t \leq T$. In fact:

$$v(x, t) \leq \sup_{\mathbb{R} \times [0, T)} f \int_0^t \int_{\mathbb{R}} \Gamma_D(x - y, t - s) dy ds = t \sup_{\mathbb{R} \times [0, T)} f$$

since

$$\int_{\mathbb{R}} \Gamma_D(x - y, t - s) dy = 1$$

for every x, t, s , $t > s$. In the same way it can be shown that $v(x, t) \geq t \inf_{\mathbb{R} \times [0, T)} f$. As a consequence, we have:

Corollary 2.18. Uniqueness II. *Let g be continuous in \mathbb{R} , satisfying (2.147), and let f be as in Theorem 2.15. Then the Cauchy problem (2.144) has a unique solution u in $\mathbb{R} \times [0, T]$, for $T < 1/(4Da)$, belonging to the Tychonov class. This solution is given by (2.8.3) and moreover*

$$\inf_{\mathbb{R}} g + t \inf_{\mathbb{R} \times [0, T)} f \leq u(x, t) \leq \sup_{\mathbb{R}} g + t \sup_{\mathbb{R} \times [0, T)} f. \quad (2.149)$$

Proof. If u and v are solutions of the same Cauchy problem (2.144), then $w = u - v$ is a solution of (2.144) with $f = g = 0$ and satisfies the hypotheses of Corollary 2.17. It follows that $w(x, t) \equiv 0$. \square

- *Stability and comparison.* As in Corollary 2.5, p. 38, the inequality (2.149) is a stability estimate for the correspondence

$$\text{data} \longmapsto \text{solution.}$$

Indeed, let u_1 and u_2 be solutions of (2.144) with data g_1, f_1 and g_2, f_2 , respectively. Under the hypotheses of Corollary 2.18, by (2.149) we can write

$$\sup_{\mathbb{R} \times [0, T]} |u_1 - u_2| \leq \sup_{\mathbb{R}} |g_1 - g_2| + T \sup_{\mathbb{R} \times [0, T]} |f_1 - f_2|.$$

Therefore if

$$\sup_{\mathbb{R} \times [0, T]} |f_1 - f_2| \leq \varepsilon, \quad \sup_{\mathbb{R}} |g_1 - g_2| \leq \varepsilon$$

also

$$\sup_{\mathbb{R} \times [0, T]} |u_1 - u_2| \leq \varepsilon (1 + T)$$

that means *uniform pointwise stability*.

This is not the only consequence of (2.149). We can use it to compare two solutions. For instance, from the left inequality we immediately deduce that if $f \geq 0$ and $g \geq 0$, also $u \geq 0$.

Similarly, if $f_1 \geq f_2$ and $g_1 \geq g_2$, then

$$u_1 \geq u_2.$$

- *Backward equations* arise in several applied contexts, from *control theory* and *dynamic programming* to *probability* and *finance*. An example is the celebrated *Black–Scholes equation*, that we will present in the next section.

Due to the time irreversibility, to have a well posed problem for the backward equation in the time interval $[0, T]$ we must prescribe a *final condition*, that is for $t = T$, rather than an initial one. On the other hand, the change of variable $t \mapsto T - t$ transforms the backward into the forward equation, so that, from the mathematical point of view, the two equations are equivalent. Except for this remark, the theory we have developed so far remains valid.

2.8.5 The proof of the existence theorem 2.12

Proof of i). We want to show that u is well defined in the strip $(0, T]$ with $T < \frac{1}{4D\alpha}$. Since $|g(y)| \leq ce^{ay^2}$, we have, for any positive $t \leq T$:

$$|u(x, t)| \leq \frac{c}{\sqrt{4\pi Dt}} \int_{\mathbb{R}} e^{-\frac{(x-y)^2}{4Dt}} e^{ay^2} dy \underset{z=x-y}{=} \frac{c}{\sqrt{4\pi Dt}} \int_{\mathbb{R}} e^{-\frac{z^2}{4Dt}} e^{a(x-z)^2} dz. \quad (2.150)$$

Since $\frac{1}{4Dt} > a$, we can find a small positive ε such that

$$\frac{1-\varepsilon}{4Dt} > a + \varepsilon, \quad \forall t \in (0, T]. \quad (2.151)$$

Then

$$e^{-\frac{z^2}{4Dt}} = e^{-\frac{\varepsilon z^2}{4Dt}} e^{-\frac{(1-\varepsilon)z^2}{4Dt}} \leq e^{-\frac{\varepsilon z^2}{4Dt}} e^{-(a+\varepsilon)z^2}$$

and

$$e^{-\frac{z^2}{4Dt}} e^{a(x-z)^2} \leq e^{-\frac{\varepsilon z^2}{4Dt}} e^{-(a+\varepsilon)z^2 + a(x-z)^2} = e^{-\frac{\varepsilon z^2}{4Dt}} e^{-\varepsilon z^2 - 2xz + ax^2}. \quad (2.152)$$

Now, we have

$$-\varepsilon z^2 - 2xz + ax^2 = -\varepsilon z^2 - 2axz + (a-A)x^2 + Ax^2 \leq Ax^2 \quad (2.153)$$

if $a^2 + \varepsilon(a-A) \leq 0$, that is if $A \geq \frac{a^2 + \varepsilon a}{\varepsilon}$.

With this choice of A , from (2.150), (2.152) and (2.153), we can write

$$|u(x, t)| \leq \frac{ce^{Ax^2}}{\sqrt{4\pi Dt}} \int_{\mathbb{R}} e^{-\frac{\varepsilon z^2}{4Dt}} dz \underset{z=\sqrt{4Dt/\varepsilon}y}{=} \frac{ce^{Ax^2}}{\sqrt{\varepsilon\pi}} \int_{\mathbb{R}} e^{-y^2} dy = \frac{c}{\sqrt{\varepsilon}} e^{Ax^2}.$$

Thus u is well defined in the strip $R \times (0, T]$, $T < \frac{1}{4D^2a}$, and moreover i) holds with $C = c/\sqrt{\varepsilon}$.

Proof of ii). We need to differentiate under the integral sign. Observe that $\Gamma_D \in C^\infty(\mathbb{R} \times (0, \infty))$. Given a point (x, t) , we need to bound, in a neighborhood of (x, t) , each derivative $\partial_t^h \partial_x^k \Gamma_D(x-y, t) g(y)$ by a integrable function $G = G(y)$, in general depending on the derivative itself.

Now, the analytic expression of a derivative $\partial_t^h \partial_x^k \Gamma_D(x-y, t)$ is a sum of terms of the form

$$t^{-r} (x-y)^s \exp\left\{-\frac{(x-y)^2}{4Dt}\right\} \quad (r, s \geq 0)$$

up to multiplicative constants. Let $x \in [-R, R]$, $t_0 \leq t \leq T$. Since for any $b > 0$ and any $xy \in \mathbb{R}$ we have $2xy \leq b^{-1}x^2 + by^2$, we can write

$$e^{-\frac{(x-y)^2}{4Dt}} \leq e^{-\frac{x^2 - 2xy + y^2}{4DT}} \leq e^{\frac{R^2}{4bDT} - \frac{(1-b)y^2}{4DT}}.$$

On the other hand,

$$t^{-r} |x-y|^s \leq t_0^{-r} (R+|y|)^s.$$

Therefore, we can write, for $x \in [-R, R]$, $t_0 \leq t \leq T$,

$$t^{-r} |x-y|^s \exp\left\{-\frac{(x-y)^2}{4Dt}\right\} |g(y)| \leq t_0^{-r} e^{\frac{R^2}{4bDT}} (R+|y|^s) e^{-\left(\frac{(1-b)}{4DT} - a\right)y^2} \equiv G_{R, t_0, s, r, b}(y).$$

If b is chosen so small to have $\frac{(1-b)}{4DT} - a > 0$ then, $G_{R, t_0, s, r, b}$ is integrable over \mathbb{R} and therefore differentiation under the integral sign is possible for $x \in [-R, R]$, $t_0 \leq t \leq T$. On the other hand, R is an arbitrary positive number and t_0 is an arbitrary number in $[t_0, T]$ so that $u \in C^\infty(\mathbb{R} \times (0, T])$ and, in particular,

$$u_t - \Delta u = \int_{\mathbb{R}} \{\partial_t \Gamma_D(x-y, t) - \Delta_x \Gamma_D(x-y, t)\} g(y) dy = 0.$$

Proof of iii). Assume g is continuous at x_0 . We want to show that, given $\varepsilon > 0$, if $|x - x_0|$ and t are sufficiently small, depending on x_0 and ε , then

$$|u(x, t) - g(x_0)| \leq \varepsilon.$$

Let $\varepsilon > 0$ and $\delta > 0$, $\delta = \delta(\varepsilon, x_0)$, be such that, if $|y - x_0| < \delta$ then

$$|g(y) - g(x_0)| < \varepsilon/2. \quad (2.154)$$

Since $\int_{\mathbb{R}} \Gamma_D(x - y) dy = 1$ for every $t > 0$, we can write

$$\begin{aligned} u(x, t) - g(x_0) &= \int_{\mathbb{R}} \Gamma_D(x - y, t) [g(y) - g(x_0)] dy = \\ &= \int_{\{y: |y - x_0| < \delta\}} (\dots) dy + \int_{\{y: |y - x_0| > \delta\}} (\dots) dy \equiv I + II. \end{aligned}$$

From (2.154) we have

$$|I| \leq \int_{\{y: |y - x_0| < \delta\}} \Gamma_D(x - y, t) |g(y) - g(x_0)| dy < \varepsilon/2. \quad (2.155)$$

Moreover, since

$$|g(y) - g(x_0)| \leq ce^{ay^2} + ce^{ax_0^2}$$

we get

$$|II| \leq ce^{ax_0^2} \int_{\{y: |y - x_0| > \delta\}} \Gamma_D(x - y, t) dy + c \int_{\{y: |y - x_0| > \delta\}} \Gamma_D(x - y, t) e^{ay^2} dy. \quad (2.156)$$

Setting $z = x - y$, we find

$$\int_{\{y: |y - x_0| > \delta\}} \Gamma_D(x - y, t) dy = \int_{\{z: |z - (x - x_0)| > \delta\}} \frac{1}{\sqrt{4\pi Dt}} e^{-\frac{z^2}{4Dt}} dz.$$

Observe now that, if $|z - (x - x_0)| > \delta$ and $|x - x_0| < \delta/2$, we deduce

$$\delta < |z - (x - x_0)| \leq |z| + |x - x_0| \leq |z| + \frac{\delta}{2}$$

whence

$$\{z : |z - (x - x_0)| > \delta\} \subset \left\{ z : |z| > \frac{\delta}{2} \right\}.$$

Therefore, if $|x - x_0| < \delta/2$,

$$\begin{aligned} \int_{\{z: |z - (x - x_0)| > \delta\}} \frac{1}{\sqrt{4\pi Dt}} e^{-\frac{z^2}{4Dt}} dz &\leq \int_{\{z: |z| > \delta/2\}} \frac{1}{\sqrt{4\pi Dt}} e^{-\frac{z^2}{4Dt}} dz \\ &= \frac{1}{\sqrt{\pi}} \int_{\{s: |s| > \frac{\delta}{4\sqrt{Dt}}\}} e^{-s^2} ds \end{aligned}$$

and the last integral tends to zero as $t \rightarrow 0^+$. Via a similar argument⁴⁸, arguing as in the proof of i), we can show that also the second integral in (2.156) tends a zero as $t \rightarrow 0^+$.

⁴⁸ We leave the details to the reader.

Thus, if if $|x - x_0| < \delta/2$ and t is sufficiently small, $|II| < \varepsilon/2$ and from (2.155) we get

$$|u(x, t) - g(x_0)| \leq |I| + |II| < \varepsilon$$

from which *iii*) follows. \square

2.9 An Application to Finance

2.9.1 European options

In this section we apply the above theory to determine the price of some financial products, in particular of some *derivative* products, called *European options*. A financial product is a *derivative* if its payoff depends on the price behavior of an asset, in jargon *the underlying*, for instance a *stock*, a *currency* or a *commodity*. Among the simplest derivatives are the **European call** and **put options**, that are contracts on a prescribed asset between a *holder* and a *subscriber*, with the following rules. At the drawing up time of the contract (say at time $t = 0$) an **exercise** or **strike price** E is fixed.

At an **expiry date** T , fixed in the future:

- The holder of a call option **can (but is not obliged to)** exercise the option by **purchasing** the asset at the price E . If the holder decides to buy the asset, the subscriber **must** sell it.
- The holder of a put option **can (but is not obliged to)** exercise the option by **selling** at the price E . If the holder decides to sell the asset, the subscriber **must** buy it.

Since an option gives to the holder a right without any obligation, the option has a price and the basic question is: **what is the “right” price that must be paid** at $t = 0$?

This price certainly depends on the evolution of the price S of the underlying, on the strike price E , on the expiring time T and on the current riskless interest rate $r > 0$.

For instance, for a call, to a lower E corresponds a greater price; the opposite holds for a put. The price fluctuations of the underlying affect in crucial way the value of an option, since they incorporate the amount of risk.

To answer our basic question, we introduce the **value function** $V = V(S, t)$, giving the proper price of the option if at time t the price of the underlying is S . What we need to know is $V(S(0), 0)$. When we like to distinguish between **call** and **put**, we use the notations $C(S, t)$ and $P(S, t)$, respectively.

The problem is then to determine V in agreement with the financial market, where both the underlying and the option are exchanged. We shall use the Black-Scholes method, based on the assumption of a reasonable evolution model for S and on the fundamental principle of *no arbitrage possibilities*.

2.9.2 An evolution model for the price S

Since S depends on more or less foreseeable factors, it is clear that we cannot expect a deterministic model for the evolution of S . To construct it we assume a *market efficiency* in the following sense:

- a) The market responds instantaneously to new information on the asset.
- b) The price has no memory: its past history is fully stored in the present price, without further information.

Condition a) implies the adoption of a continuous model. Condition b) basically requires that a change dS of the underlying price has the Markov property, like Brownian motion.

Consider now a time interval from t to $t + dt$, during which S undergoes a change from S to $S + dS$. One of the most common models assumes that the **return** dS/S is given by the sum of two terms.

One is a deterministic term, which gives a contribution μdt due to a constant *drift* μ , representing the average growth rate of S . With this term alone, we would have

$$\frac{dS}{S} = \mu dt$$

and therefore $d \log S = \mu dt$, that gives the exponential growth $S(t) = S(0) e^{\mu t}$.

The other term is stochastic and takes into account the random aspects of the evolution. It gives the contribution

$$\sigma dB$$

where dB is an increment of a Brownian motion and has zero mean and variance dt . The coefficient σ , that we assume to be constant, is called the **volatility** and measures the standard deviation of the return.

Summing the contributions we have

$$\frac{dS}{S} = \mu dt + \sigma dB. \quad (2.157)$$

Note the physical dimensions of μ and σ : $[\mu] = [\text{time}]^{-1}$, $[\sigma] = [\text{time}]^{-\frac{1}{2}}$.

The (2.157) is a **stochastic differential equation** (*s.d.e.*). To solve it one is tempted to write

$$d \log S = \mu dt + \sigma dB,$$

to integrate between 0 and t , and to obtain

$$\log \frac{S(t)}{S(0)} = \mu t + \sigma (B(t) - B(0)) = \mu t + \sigma B(t)$$

since $B(0) = 0$. However, this is not correct. The diffusion term σdB requires the use of the **Itô formula**, a stochastic version of the chain rule. Let us make a few intuitive remarks on this important formula.

Digression on Itô's formula. Let $B = B(t)$ the usual Brownian motion. An Itô process $X = X(t)$ is a solution of a *s.d.e.* of the type

$$dX = a(X, t) dt + \sigma(X, t) dB \quad (2.158)$$

where a is the *drift term* and σ is the *volatility coefficient*. When $\sigma = 0$, the equation is deterministic and the trajectories can be computed with the usual analytic methods. Moreover, given a smooth function $F = F(x, t)$, we can easily compute the variation of F along those trajectories. It is enough to compute

$$dF = F_t dt + F_x dX = \{F_t + aF_x\} dt.$$

Let now be σ nonzero; the preceding computation would give

$$dF = F_t dt + F_x dX = \{F_t + aF_x\} dt + \sigma F_x dB,$$

but **this formula does not give the complete differential of F** . Indeed, using Taylor's formula, one has, letting $X(0) = X_0$:

$$F(X, t) = F(X_0, 0) + F_t dt + F_x dX + \frac{1}{2} \left\{ F_{xx}(dX)^2 + 2F_{xt} dX dt + F_{tt}(dt)^2 \right\} + \dots$$

The differential of F along the trajectories of (2.158) is obtained by selecting in the right hand side of the preceding formula the terms which are **linear** with respect to dt or dX . We first find the terms

$$F_t dt + F_x dX = \{F_t + aF_x\} dt + \sigma F_x dB.$$

The terms $2F_{xt} dX dt$ and $F_{tt}(dt)^2$ are nonlinear with respect to dt and dX and therefore they are not included in the differential. Let us now check the term $(dX)^2$. We have

$$(dX)^2 = (adt + \sigma dB)^2 = a^2(dt)^2 + 2a\sigma dB dt + \boxed{\sigma^2(dB)^2}.$$

While $a^2(dt)^2$ and $2a\sigma dB dt$ are nonlinear with respect to dt and dX , the framed term **turns out to be exactly**

$$\sigma^2 dt.$$

Formally, this is a consequence of the basic formula⁴⁹

$$dB \sim \sqrt{dt} N(0, 1)$$

that assigns \sqrt{dt} for the standard deviation of dB .

⁴⁹ See (2.86), Sect. 4.

Thus the differential of F along the trajectories of (2.158) is given by the following **Itô formula**:

$$dF = \left\{ F_t + aF_x + \frac{1}{2}\sigma^2 F_{xx} \right\} dt + \sigma F_x dB. \quad (2.159)$$

We are now ready to solve (2.157), that we write in the form

$$dS = \mu S dt + \sigma S dB.$$

Let $F(S) = \log S$. Since

$$F_t = 0, \quad F_S = \frac{1}{S}, \quad F_{ss} = -\frac{1}{S^2}$$

Itô's formula gives, with $X = S$, $a(S, t) = \mu S$, $\sigma(S, t) = \sigma S$,

$$d \log S = \left(\mu - \frac{1}{2}\sigma^2 \right) dt + \sigma dB.$$

We can now integrate between 0 and t , obtaining

$$\log S(t) = \log S_0 + \left(\mu - \frac{1}{2}\sigma^2 \right) t + \sigma B(t). \quad (2.160)$$

The (2.160) shows that the random variable $Y = \log S$ has a normal distribution, with mean $\log S_0 + (\mu - \frac{1}{2}\sigma^2)t$ and variance $\sigma^2 t$. Its probability density is therefore

$$f(y) = \frac{1}{\sqrt{2\pi\sigma^2 t}} \exp \left\{ -\frac{(y - \log S_0 - (\mu - \frac{1}{2}\sigma^2)t)^2}{2\sigma^2 t} \right\}$$

and the density of S is given by

$$p(s) = \frac{1}{s} f(\log s) = \frac{1}{s\sqrt{2\pi\sigma^2 t}} \left\{ -\frac{(\log s - \log S_0 - (\mu - \frac{1}{2}\sigma^2)t)^2}{2\sigma^2 t} \right\}$$

which is called a **lognormal density**.

2.9.3 The Black-Scholes equation

We now construct a differential equation able to describe the evolution of $V(S, t)$. We work under the following hypotheses:

- S follows a **lognormal law**.
- The volatility σ is constant and known.
- There are no transaction costs or dividends.

- It is possible to buy or sell any number of the underlying asset.
- There is an interest rate $r > 0$, for a riskless investment. This means that 1 dollar in a bank at time $t = 0$ becomes e^{rT} dollars at time T .
- The market is **arbitrage free**.

The last hypothesis is crucial in the construction of the model and means that *there is no opportunity for instantaneous risk-free profit*. It could be considered as a sort of conservation law for money!

The translation of this principle into mathematical terms is linked with the notion of *hedging* and the existence of *self-financing portfolios*⁵⁰. The basic idea is first to compute the return of V through Itô's formula and then to construct a riskless portfolio Π , consisting of shares of S and the option. By the arbitrage free hypothesis, Π must grow at the current interest rate r , i.e. $d\Pi = r\Pi dt$, which turns out to coincide with the fundamental Black-Scholes equation.

Let us then use the Itô's formula to compute the differential of V . Since

$$dS = \mu S dt + \sigma S dB,$$

we find

$$dV = \left\{ V_t + \mu S V_S + \frac{1}{2} \sigma^2 S^2 V_{SS} \right\} dt + \sigma S V_S dB. \quad (2.161)$$

Now we try to get rid of the risk term $\sigma S V_S dB$ by constructing a portfolio Π , consisting of the option and a quantity⁵¹ $-\Delta$ of underlying:

$$\Pi = V - S\Delta.$$

This is an important financial operation called *hedging*. Consider now the interval of time $(t, t+dt)$ during which Π undergoes a variation $d\Pi$. If we manage to keep Δ equal to its value at t during the interval $(t, t+dt)$, the variation of Π is given by

$$d\Pi = dV - \Delta dS.$$

This is a key point in the whole construction, that needs to be carefully justified⁵². Although we content ourselves with an intuitive level, we will come back to this question in the last section of the chapter.

Using (2.161) we find

$$\begin{aligned} d\Pi &= dV - \Delta dS = \\ &= \left\{ V_t + \mu S V_S + \frac{1}{2} \sigma^2 S^2 V_{SS} - \mu S \Delta \right\} dt + \sigma S (V_S - \Delta) dB. \end{aligned} \quad (2.162)$$

⁵⁰ A *portfolio* is a collection of *securities* (e.g. stocks) holdings.

⁵¹ We borrow from *finance* the use of the greek letter Δ in this context. Clearly here it has nothing to do with the Laplace operator.

⁵² In fact, saying that we keep Δ constant for an infinitesimal time interval so that we can cancel $Sd\Delta$ from the differential $d\Pi$ requires a certain amount of impudence. . . .

Thus, if we choose

$$\Delta = V_S, \quad (2.163)$$

meaning that Δ is the value of V_S at time t , we eliminate the stochastic component in (2.162). The evolution of the portfolio Π is now entirely deterministic and its dynamics is given by the following equation:

$$d\Pi = \left\{ V_t + \frac{1}{2}\sigma^2 S^2 V_{SS} \right\} dt. \quad (2.164)$$

The choice (2.163) appears almost miraculous, but it is partly justified by the fact that V and S are dependent and the random component in their dynamics is proportional to S . Thus, in a suitable linear combination of V and S such component should disappear.

It is the moment to use the no-arbitrage principle. Investing Π at the riskless rate r , after a time dt we have an increment $r\Pi dt$. Compare $r\Pi dt$ with $d\Pi$ given by (2.164).

- If $d\Pi > r\Pi dt$, we borrow an amount Π to invest in the portfolio. The return $d\Pi$ would be greater of the cost $r\Pi dt$, so that we make an instantaneous riskless profit

$$d\Pi - r\Pi dt.$$

- If $d\Pi < r\Pi dt$, we sell the portfolio Π investing it in a bank at the rate r . This time we would make an instantaneous risk free profit

$$r\Pi dt - d\Pi.$$

Therefore, the arbitrage free hypothesis forces

$$d\Pi = \left\{ V_t + \frac{1}{2}\sigma^2 S^2 V_{SS} \right\} dt = r\Pi dt. \quad (2.165)$$

Substituting

$$\Pi = V - S\Delta = V - V_S S$$

into (2.165), we obtain the celebrated **Black-Scholes equation**:

$$\mathcal{L}V = V_t + \frac{1}{2}\sigma^2 S^2 V_{SS} + rSV_S - rV = 0. \quad (2.166)$$

Note that the coefficient μ , the drift of S , does not appear in (2.166). This fact is apparently counter-intuitive and shows an interesting aspect of the model. The financial meaning of the Black-Scholes equation is emphasized from the following

decomposition of its right hand side:

$$\mathcal{L}V = V_t + \underbrace{\frac{1}{2}\sigma^2 S^2 V_{SS}}_{\text{portfolio return}} - \underbrace{r(V - SV_S)}_{\text{bank investment}}.$$

The Black-Scholes equation is a little more general than the equations we have seen so far. Indeed, the diffusion and the drift coefficients are both depending on S . However, as we shall see below, we can transform it into the diffusion equation $u_t = u_{xx}$.

Observe that the coefficient of V_{SS} is positive, so that (2.166) is a **backward equation**. To get a well posed problem, we need a **final condition** (at $t = T$), a side condition at $S = 0$ and one condition for $S \rightarrow +\infty$.

- *Final conditions.* We examine what conditions we have to impose at $t = T$.

Call. If at time T we have $S > E$ then we exercise the option, with a profit $S - E$. If $S \leq E$, we do not exercise the option with no profit. The *final payoff* of the option is therefore

$$C(S, T) = \max\{S - E, 0\} = (S - E)^+, \quad S > 0.$$

Put. If at time T we have $S \geq E$, we do not exercise the option, while we exercise the option if $S < E$. The *final payoff* of the option is therefore

$$P(S, T) = \max\{E - S, 0\} = (E - S)^+, \quad S > 0.$$

- *Boundary conditions.* We now examine the conditions to be imposed at $S = 0$ and for $S \rightarrow +\infty$.

Call. If $S = 0$ at a time t , (2.157) implies $S = 0$ thereafter, and the option has no value; therefore

$$C(0, t) = 0, \quad t \geq 0.$$

As $S \rightarrow +\infty$, at time t , the option will be exercised and its value becomes practically equal to S minus the discounted exercise price, that is

$$C(S, t) - (S - e^{-r(T-t)}E) \rightarrow 0 \quad \text{as } S \rightarrow \infty.$$

Put. If at a certain time is $S = 0$, so that $S = 0$ thereafter, the final profit is E . Thus, to determine $P(0, t)$ we need to determine the present value of E at time T , that is

$$P(0, t) = Ee^{-r(T-t)}.$$

If $S \rightarrow +\infty$, we do not exercise the option, hence

$$P(S, t) = 0 \quad \text{as } S \rightarrow +\infty.$$

2.9.4 The solutions

Let us summarize our model in the two cases.

- *Black-Scholes equation*

$$V_t + \frac{1}{2}\sigma^2 S^2 V_{SS} + rSV_S - rV = 0. \quad (2.167)$$

- *Final payoffs*

$$\begin{aligned} C(S, T) &= (S - E)^+ && \text{(call)} \\ P(S, T) &= (E - S)^+ && \text{(put).} \end{aligned}$$

- *Boundary conditions*

$$\begin{aligned} C(0, t) &= 0 \quad \text{and} \quad C(S, t) - (S - e^{-r(T-t)}E) \rightarrow 0 \text{ as } S \rightarrow \infty && \text{(call)} \\ P(0, t) &= Ee^{-r(T-t)} \quad \text{and} \quad P(S, T) = 0 \text{ as } S \rightarrow \infty && \text{(put).} \end{aligned}$$

It turns out that the above problems can be reduced to a global Cauchy problem for the heat equation. In this way it is possible to find explicit formulas for the solutions. First of all we make a change of variables to reduce the Black-Scholes equation to constant coefficients and to pass from backward to forward in time. Also note that $1/\sigma^2$ can be considered an intrinsic reference time while the exercise price E gives a characteristic order of magnitude for S and V . Thus, $1/\sigma^2$ and E can be used as rescaling factors to introduce dimensionless variables.

Let us set

$$x = \log \frac{S}{E}, \quad \tau = \frac{1}{2}\sigma^2(T - t), \quad w(x, \tau) = \frac{1}{E}V\left(Ee^x, T - \frac{2\tau}{\sigma^2}\right).$$

When S goes from 0 to $+\infty$, x varies from $-\infty$ to $+\infty$. When $t = T$ we have $\tau = 0$. Moreover:

$$\begin{aligned} V_t &= -\frac{1}{2}\sigma^2 E w_\tau \\ V_S &= \frac{E}{S}w_x, \quad V_{SS} = -\frac{E}{S^2}w_x + \frac{E}{S^2}w_{xx}. \end{aligned}$$

Substituting into (2.167), after some simplifications, we get

$$-\frac{1}{2}\sigma^2 w_\tau + \frac{1}{2}\sigma^2(-w_x + w_{xx}) + rw_x - rw = 0$$

or

$$w_\tau = w_{xx} + (k - 1)w_x - kw$$

where $k = \frac{2r}{\sigma^2}$ is a dimensionless parameter. By further setting⁵³

$$w(x, \tau) = e^{-\frac{k-1}{2}x - \frac{(k+1)^2}{4}\tau} v(x, \tau)$$

we find that v satisfies

$$v_\tau - v_{xx} = 0, \quad -\infty < x < +\infty, \quad 0 \leq \tau \leq T.$$

The final condition for V becomes an initial condition for v . Precisely, after some manipulations, we have

$$v(x, 0) = g(x) = \begin{cases} e^{\frac{1}{2}(k+1)x} - e^{\frac{1}{2}(k-1)x} & x > 0 \\ 0 & x \leq 0 \end{cases}$$

for the call option, and

$$v(x, 0) = g(x) = \begin{cases} e^{\frac{1}{2}(k-1)x} - e^{\frac{1}{2}(k+1)x} & x < 0 \\ 0 & x \geq 0 \end{cases}$$

for the put option.

Now we can use the preceding theory and in particular Theorem 2.12, p. 78 and Corollary 2.17, p. 83. The solution is unique and it is given by

$$v(x, \tau) = \frac{1}{\sqrt{4\pi\tau}} \int_{\mathbb{R}} g(y) e^{-\frac{(x-y)^2}{4\tau}} dy.$$

To have a more significant formula, let $y = \sqrt{2\tau}z + x$; then, focusing on the **call** option:

$$\begin{aligned} v(x, \tau) &= \frac{1}{\sqrt{2\pi}} \int_{\mathbb{R}} g(\sqrt{2\tau}z + x) e^{-\frac{z^2}{2}} dz = \\ &= \frac{1}{\sqrt{2\pi}} \left\{ \int_{-x/\sqrt{2\tau}}^{\infty} e^{\frac{1}{2}(k+1)(\sqrt{2\tau}z+x)-\frac{1}{2}z^2} dz - \int_{-x/\sqrt{2\tau}}^{\infty} e^{\frac{1}{2}(k-1)(\sqrt{2\tau}z+x)-\frac{1}{2}z^2} dz \right\}. \end{aligned}$$

After some manipulations in the two integrals⁵⁴, we obtain

$$v(x, \tau) = e^{\frac{1}{2}(k+1)x + \frac{1}{4}(k+1)^2\tau} N(d_+) - e^{\frac{1}{2}(k-1)x + \frac{1}{4}(k-1)^2\tau} N(d_-)$$

⁵³ See Problem 2.17.

⁵⁴ For instance, to evaluate the first integral, complete the square at the exponent, writing

$$\frac{1}{2}(k+1)\left(\sqrt{2\tau}z + x\right) - \frac{1}{2}z^2 = \frac{1}{2}(k+1)x + \frac{1}{4}(k+1)^2\tau - \frac{1}{2}\left[z - \frac{1}{2}(k+1)\sqrt{2\tau}\right]^2.$$

Then, setting $y = z - \frac{1}{2}(k+1)\sqrt{2\tau}$,

$$\int_{-x/\sqrt{2\tau}}^{\infty} e^{\frac{1}{2}(k+1)(\sqrt{2\tau}z+x)-\frac{1}{2}z^2} dz = e^{\frac{1}{2}(k+1)x + \frac{1}{4}(k+1)^2\tau} \int_{-x/\sqrt{2\tau} - (k+1)\sqrt{\tau}/\sqrt{2}}^{\infty} e^{-\frac{1}{2}y^2} dy.$$

where

$$N(z) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^z e^{-\frac{1}{2}y^2} dy$$

is the distribution of a standard normal random variable and

$$d_{\pm} = \frac{x}{\sqrt{2\tau}} + \frac{1}{2}(k \pm 1)\sqrt{2\tau}.$$

Going back to the original variables we have, for the **call**:

$$C(S, t) = SN(d_+) - Ee^{-r(T-t)}N(d_-)$$

with

$$d_{\pm} = \frac{\log(S/E) + (r \pm \frac{1}{2}\sigma^2)(T-t)}{\sigma\sqrt{T-t}}.$$

The formula for the **put** is

$$P(S, t) = Ee^{-r(T-t)}N(-d_-) - SN(-d_+).$$

It can be shown that⁵⁵

$$\begin{aligned} \Delta = C_S &= N(d_+) > 0 && \text{for the call} \\ \Delta = P_S &= N(d_+) - 1 < 0 && \text{for the put.} \end{aligned}$$

Note that C_S and P_S are strictly increasing with respect to S , since N is a strictly increasing function and d_+ is strictly increasing with S . The functions C , P are therefore *strictly convex functions* of S , for every t , namely

$$C_{ss} > 0 \quad \text{and} \quad P_{ss} > 0.$$

- *Put-call parity.* Put and call options, with the same exercise price and expiry time, can be connected by forming the following portfolio:

$$\Pi = S + P - C$$

where the minus in front of C shows a so called *short position* (negative holding). For this portfolio, the final payoff is

$$\Pi(S, T) = S + (E - S)^+ - (S - E)^+.$$

If $E \geq S$, we have

$$\Pi(S, T) = S + (E - S) - 0 = E$$

while if $E \leq S$,

$$\Pi(S, T) = S + 0 - (S - E) = E.$$

⁵⁵ The calculations are rather . . . painful.

Thus at expiry the payoff is always equal to E and it constitutes a riskless profit, whose value at t must be equal to the discounted value of E , because of the no arbitrage condition. Hence we find the following relation (*put-call parity*)

$$S + P - C = Ee^{-r(T-t)}. \quad (2.168)$$

Formula (2.168) also shows that, given the value of C (or P), we can find the value of P (or C). From (2.168), since

$$Ee^{-r(T-t)} \leq E$$

and $P \geq 0$, we get

$$C(S, t) = S + P - Ee^{-r(T-t)} \geq S - E$$

and therefore, since $C \geq 0$,

$$C(S, t) \geq (S - E)^+.$$

It follows that the value of C is always greater than the final payoff. It is not so for a put. In fact

$$P(0, t) = Ee^{-r(T-t)} \leq E$$

so that the value of P is below the final payoff when S is near 0, while it is above just before expiring. Figures 2.14 and 2.15 show the behavior of C and P versus S , for some values of $T - t$ up to expiry.

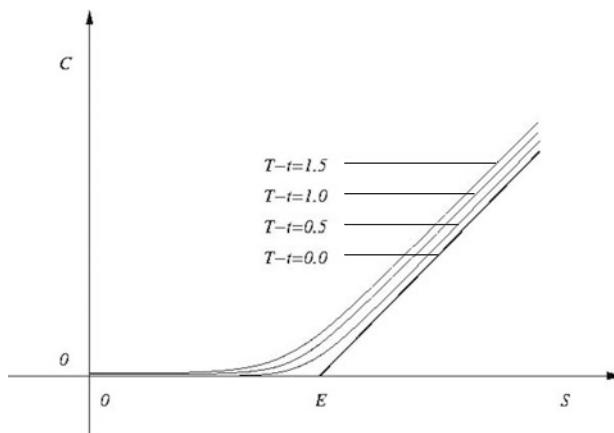


Fig. 2.14 The value function for an European call option

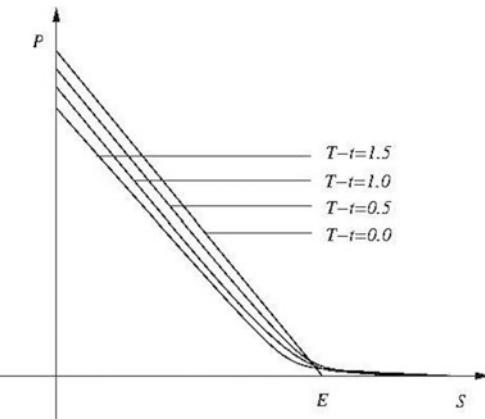


Fig. 2.15 The value function of an European put option

- *Different volatilities.* The maximum principle arguments in Sect. 2.8 can be used to compare the value of two options with different volatilities σ_1 and σ_2 , having the same exercise price E and the same strike time T . Assume that $\sigma_1 > \sigma_2$ and denote by $C^{(1)}$, $C^{(2)}$ the value of the corresponding call options. Diminishing the amount of risk the value of the option should decrease and indeed we want to confirm that

$$C^{(1)} > C^{(2)}, \quad S > 0, 0 \leq t < T.$$

Let $W = C^{(1)} - C^{(2)}$. Then

$$W_t + \frac{1}{2}\sigma_2^2 S^2 W_{SS} + rSW_S - rW = \frac{1}{2}(\sigma_2^2 - \sigma_1^2)S^2 C_{SS}^{(1)} \quad (2.169)$$

with $W(S, T) = 0$, $W(0, t) = 0$ and $W \rightarrow 0$ as $S \rightarrow +\infty$.

The (2.169) is a non-homogeneous equation, whose right hand side is *negative* for $S > 0$, because $C_{SS}^{(1)} > 0$. Since W is continuous in the half strip $[0, +\infty) \times [0, T]$ and vanishes at infinity, it attains its global minimum at a point (S_0, t_0) .

We claim that the minimum is zero and cannot be attained at a point in $(0, +\infty) \times [0, T]$. Since the equation is backward, $t_0 = 0$ is excluded. Suppose $W(S_0, t_0) \leq 0$ with $S_0 > 0$ and $0 < t_0 < T$. We have

$$W_t(S_0, t_0) = 0$$

and

$$W_S(S_0, t_0) = 0, \quad W_{SS}(S_0, t_0) \geq 0.$$

Substituting $S = S_0, t = t_0$ into (2.169) we get a contradiction. Therefore $W = C^{(1)} - C^{(2)} > 0$ for $S > 0, 0 < t < T$.

2.9.5 Hedging and self-financing strategy

The mathematical translation of the *no arbitrage* principle can be made more rigorously, with respect to what we did in Sect. 2.9.2, by introducing the concept of *self-financing* portfolio. The idea is to “duplicate” V by means of a portfolio consisting of a number of shares of S and a bond Z , a free risk investment growing at the rate r , e.g. $Z(t) = e^{rt}$.

To this purpose let us try to determine two processes $\phi = \phi(t)$ and $\psi = \psi(t)$ such that

$$V = \phi S + \psi Z \quad (0 \leq t \leq T) \quad (2.170)$$

in order to eliminate any risk factor. In fact, playing the part of the subscriber of a *call* (that has to sell), the risk is that at time T the price $S(T)$ is greater than E , so that the holder will exercise the option. If in the meantime the subscriber has constructed the portfolio (2.170), the profit from it exactly meets the funds necessary to pay the holder. On the other hand, if the option has zero value at time T , the portfolio has no value as well.

For the operation to make sense, it is necessary that the subscriber *does not put extra money in this strategy (hedging)*. This can be assured by requiring that the portfolio (2.170) be *self-financing* that is, **its changes in value be dependent from variations of S and Z alone**.

In formulas, this amounts to requiring

$$dV = \phi dS + \psi dZ \quad (0 \leq t \leq T). \quad (2.171)$$

Actually, we have already met something like (2.171), when we have constructed the portfolio $\Pi = V - S\Delta$ or

$$V = \Pi + S\Delta,$$

asking that $dV = d\Pi + \Delta dS$. This construction is nothing else than a duplication of V by means of a *self-financing portfolio*, with Π playing the role of Z and choosing $\psi = 1$.

But, what is the real meaning of (2.171)? We see it better in a discrete setting. Consider a sequence of times

$$t_0 < t_1 < \dots < t_N$$

and suppose that the intervals $(t_j - t_{j-1})$ are very small. Denote by S_j and Z_j the values at t_j of S and Z . Consequently, look for two sequences

$$\phi_j \text{ and } \psi_j$$

corresponding to the quantity of S and Z to be used in the construction of the portfolio (2.170) from t_{j-1} to t_j . Notice that ϕ_j and ψ_j are chosen at time t_{j-1} .

Thus, given the interval (t_{j-1}, t_j) ,

$$V_j = \phi_j S_j + \psi_j Z_j$$

represents the closing value of the portfolio while

$$\phi_{j+1} S_j + \psi_{j+1} Z_j$$

is the opening value, the amount of money necessary to buy the new one. The **self-financing condition means** that the value V_j of the portfolio at time t_j , determined by the couple (ϕ_j, ψ_j) , exactly meets the purchasing cost of the portfolio in the interval (t_j, t_{j+1}) , determined by (ϕ_{j+1}, ψ_{j+1}) . This means

$$\phi_{j+1} S_j + \psi_{j+1} Z_j = \phi_j S_j + \psi_j Z_j \quad (2.172)$$

or that **the financial gap**

$$D_j = \phi_{j+1} S_j + \psi_{j+1} Z_j - V_j$$

must be zero, otherwise an amount of cash D_j has to be injected to sustain the strategy ($D_j > 0$) or the same amount of money can be drawn from it ($D_j < 0$). From (2.172) we deduce that

$$\begin{aligned} V_{j+1} - V_j &= (\phi_{j+1} S_{j+1} + \psi_{j+1} Z_{j+1}) - (\phi_j S_j + \psi_j Z_j) \\ &= (\phi_{j+1} S_{j+1} + \psi_{j+1} Z_{j+1}) - (\phi_{j+1} S_j + \psi_{j+1} Z_j) \\ &= \phi_{j+1} (S_{j+1} - S_j) + \psi_{j+1} (Z_{j+1} - Z_j) \end{aligned}$$

or

$$\Delta V_j = \phi_{j+1} \Delta S_j + \psi_{j+1} \Delta Z_j$$

whose continuous version is exactly (2.171).

Going back to the continuous case, by combining formulas (2.161) and (2.171) for dV , we get

$$\left\{ V_t + \mu S V_S + \frac{1}{2} \sigma^2 S^2 V_{SS} \right\} dt + \sigma S V_S dB = \phi (\mu S dt + \sigma S dB) + \psi r Z dt.$$

Choosing $\phi = V_S$, we rediscover the Black and Scholes equation

$$V_t + \frac{1}{2} \sigma^2 S^2 V_{SS} + r S V_S - r V = 0. \quad (2.173)$$

On the other hand, if V satisfies (2.173) and

$$\phi = V_S, \quad \psi = Z^{-1} (V - V_S S) = e^{-rt} (V - V_S S),$$

it can be proved that the self financing condition (2.171) is satisfied for the portfolio $\phi S + \psi Z$.

2.10 Some Nonlinear Aspects

All the mathematical models we have examined so far are *linear*. On the other hand, the nature of most real problems is nonlinear. For example, *nonlinear diffusion* has to be taken into account in filtration problems, *nonlinear drift* terms are quite important in fluid dynamics while *nonlinear reaction* terms occur frequently in population dynamics and kinetics chemistry.

The presence of a nonlinearity in a mathematical model gives rise to many interesting phenomena that cannot occur in the linear case; typical instances are finite speed of diffusion, finite time blow-up or existence of travelling wave solutions of certain special profiles, each one with its own characteristic velocity.

In this section we try to convey some intuition of what could happen in two typical and important examples from filtration through a porous medium and population dynamics. In Chap. 4, we shall deal with nonlinear transport models.

2.10.1 Nonlinear diffusion. The porous medium equation

Consider a gas of density $\rho = \rho(\mathbf{x}, t)$ flowing through a porous medium. Denote by $\mathbf{v} = \mathbf{v}(\mathbf{x}, t)$ the velocity of the gas and by κ the *porosity* of the medium, representing the volume fraction filled with gas. Conservation of mass reads, in this case:

$$\kappa \rho_t + \operatorname{div}(\rho \mathbf{v}) = 0. \quad (2.174)$$

Besides (2.174), the flow is governed by the two following constitutive (empirical) laws.

- **Darcy's law:**

$$\mathbf{v} = -\frac{\mu}{\nu} \nabla p \quad (2.175)$$

where $p = p(\mathbf{x}, t)$ is the pressure, μ is the *permeability* of the medium and ν is the *viscosity* of the gas. We assume μ and ν are positive constants.

- **Equation of state:**

$$p = p_0 \rho^a, \quad p_0 > 0, a > 0. \quad (2.176)$$

From (2.175) and (2.176) we have

$$\rho \mathbf{v} = -\frac{\mu p_0 a}{\nu} \rho^a \nabla \rho = -\frac{\mu p_0 a}{\nu (a+1)} \nabla \rho^{a+1}$$

so that

$$\operatorname{div}(\rho \mathbf{v}) = -\frac{\mu p_0 a}{\nu (a+1)} \Delta \rho^{a+1}.$$

Setting $m = 1 + a > 1$, from (2.174) we obtain

$$\rho_t = \frac{(m-1)\mu p_0}{\kappa m\nu} \Delta(\rho^m).$$

Rescaling time ($t \mapsto \frac{(m-1)\mu p_0}{\kappa m\nu} t$) we finally get the **porous medium equation**

$$\rho_t = \Delta(\rho^m). \quad (2.177)$$

Since

$$\Delta(\rho^m) = \operatorname{div}(m\rho^{m-1}\nabla\rho)$$

we see that the diffusion coefficient is $D(\rho) = m\rho^{m-1}$, showing that the diffusive effect increases with the density.

The porous medium equation can be written in terms of the pressure variable

$$u = p/p_0 = \rho^{m-1}.$$

It is not difficult to check that the equation for u is given by

$$u_t = u\Delta u + \frac{m}{m-1} |\nabla u|^2 \quad (2.178)$$

showing once more the dependence on u of the diffusion coefficient.

One of the basic questions related to the equation (2.177) or (2.178) is to understand how an initial data ρ_0 , confined in a small region Ω , evolves with time. The key object to examine is therefore the unknown *boundary* of the gas $\partial\Omega$, whose speed of expansion we expect to be proportional to $|\nabla u|$ (from (2.175)). This means that we expect a *finite speed of propagation*, in contrast with the classical case $m = 1$.

The porous media equation cannot be treated by elementary means, since at very low density the diffusion has a very low effect and the equation degenerates. However we can get some clue of what happens by examining a sort of fundamental solutions, the so called *Barenblatt solutions*, in spatial dimension 1.

The equation is

$$\rho_t = (\rho^m)_{xx}. \quad (2.179)$$

We look for *nonnegative self-similar* solutions of the form

$$\rho(x, t) = t^{-\alpha} U(xt^{-\beta}) \equiv t^{-\alpha} U(\xi)$$

which U even, $U'(0) = 0$, satisfying

$$\int_{-\infty}^{+\infty} \rho(x, t) dx = 1.$$

This condition requires

$$1 = \int_{-\infty}^{+\infty} t^{-\alpha} U(xt^{-\beta}) dx = t^{\beta-\alpha} \int_{-\infty}^{+\infty} U(\xi) d\xi$$

so that we must have $\alpha = \beta$ and $\int_{-\infty}^{+\infty} U(\xi) d\xi = 1$. Substituting into (2.179), we find

$$\alpha t^{-\alpha-1}(-U - \xi U') = t^{-m\alpha-2\alpha}(U^m)''.$$

Thus, if we choose $\alpha = 1/(m+1)$, we get for U the differential equation

$$(m+1)(U^m)'' + \xi U' + U = 0$$

that can be written in the form

$$\frac{d}{d\xi} [(m+1)(U^m)' + \xi U] = 0.$$

Thus, we have

$$(m+1)(U^m)' + \xi U = \text{constant}.$$

Computing at $\xi = 0$, since $U'(0) = 0$, we deduce that the constant is equal to zero, so that

$$(m+1)(U^m)' = (m+1)mU^{m-1}U' = -\xi U$$

or

$$(m+1)mU^{m-2}U' = -\xi.$$

This in turn is equivalent to

$$\frac{(m+1)m}{m-1}(U^{m-1})' = -\xi$$

whose solution is

$$U(\xi) = [A - B_m \xi^2]^{1/(m-1)}$$

where A is an arbitrary constant and $B_m = \frac{(m-1)}{2m(m+1)}$. Clearly, to have a physical meaning, we must have

$$A - B_m \xi^2 \geq 0.$$

In conclusion we have found solutions of the porous medium equation of the form

$$\rho(x, t) = \begin{cases} \frac{1}{t^\alpha} \left[A - B_m \frac{x^2}{t^{2\alpha}} \right]^{1/(m-1)} & \text{if } x^2 \leq At^{2\alpha}/B_m \\ 0 & \text{if } x^2 > At^{2\alpha}/B_m \end{cases} \quad (\alpha = 1/(m+1)),$$

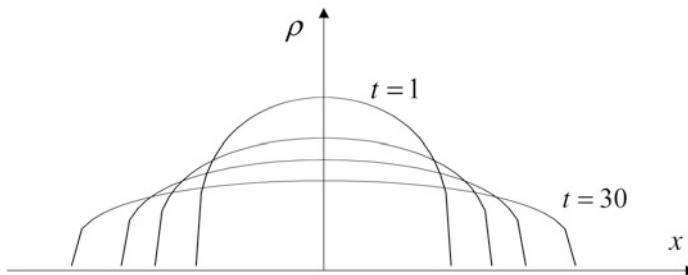


Fig. 2.16 The Barenblatt solution $\rho(x, t) = t^{-1/5}[1 - x^2t^{-2/5}]_+^{1/3}$ for $t = 1, 4, 10, 30$

known as *Barenblatt solutions* (see Fig. 2.16). The points

$$x = \pm \sqrt{A/B_m} t^\alpha \equiv \pm r(t)$$

represent the interface between the part filled by gas and the empty part. Its speed of propagation is therefore

$$\dot{r}(t) = \alpha \sqrt{A/B_m} t^{\alpha-1}.$$

2.10.2 Nonlinear reaction. Fischer's equation

In 1937 Fisher⁵⁶ introduced a model for the spatial spread of a so-called *favoured*⁵⁷ (or *advantageous*) *gene in a population*, over an infinitely long one dimensional habitat. Denoting by v the gene concentration, Fisher's equation reads

$$v_\tau = Dv_{yy} + rv \left(1 - \frac{v}{M}\right) \quad \tau > 0, y \in \mathbb{R}, \quad (2.180)$$

where D , r , and M are positive parameters. An important question is to determine whether the gene has a typical speed of propagation. Accordingly to the terminology in the introduction, (2.180) is a *semilinear equation* where diffusion is coupled with *logistic growth* through the reaction term

$$f(v) = rv \left(1 - \frac{v}{M}\right).$$

The parameter r represents a *biological potential* (net birth-death rate, with dimension $[time]^{-1}$), while M is the *carrying capacity* of the habitat. If we rescale time, space and concentration in the following way

$$t = r\tau, \quad x = \sqrt{r/D}y, \quad u = v/M,$$

⁵⁶ Fisher, R.A., The wave of advance of advantageous gene. Ann. Eugenics, **7**, 355–369, 1937.

⁵⁷ That is a *gene* that has an advantage in the struggle for life.

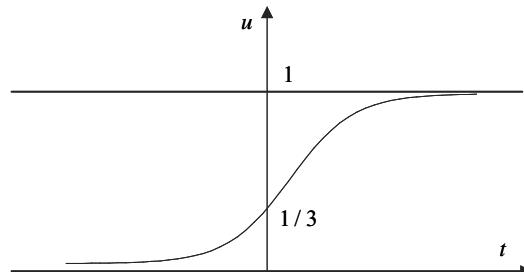


Fig. 2.17 Logistic curve ($r = 0.1, u_0 = 1/3$)

(2.180) takes the dimensionless form

$$u_t = u_{xx} + u(1-u), \quad t > 0. \quad (2.181)$$

Note the two equilibria $u \equiv 0$ and $u \equiv 1$. In absence of diffusion, 0 is unstable, and 1 is asymptotically stable. A trajectory with initial data $u(0) = u_0$ between 0 and 1 has the typical behavior shown in Fig. 2.17. Therefore, if

$$u(x, 0) = u_0(x), \quad x \in \mathbb{R}, \quad (2.182)$$

is an initial data for the equation (2.180), with $0 < u_0(x) < 1$, we expect a competitive action between diffusion and reaction, with diffusion trying to spread and lower u_0 against the reaction tendency to increase u towards the equilibrium solution 1. What we intend to show here is the existence of permanent travelling waves solutions connecting the two equilibrium states, that is solutions of the form

$$u(x, t) = U(z), \quad z = x - ct,$$

with c denoting the propagation speed, satisfying the conditions

$$0 < u < 1, \quad t > 0, x \in \mathbb{R}$$

and

$$\lim_{x \rightarrow -\infty} u(x, t) = 1 \quad \text{and} \quad \lim_{x \rightarrow +\infty} u(x, t) = 0. \quad (2.183)$$

The first condition in (2.183), states that the gene concentration is saturated at the far left end while the second condition denotes zero concentration at the far right end. Clearly, this kind of solutions realizes a balance between diffusion and reaction. Since the equation (2.180) is invariant under the transformation $x \mapsto -x$, it suffices to consider $c > 0$, that is right-moving waves only.

Since

$$u_t = -cU', \quad u_x = U', \quad u_{xx} = U'', \quad (' = d/dz)$$

substituting $u(x, t) = U(z)$ into (2.181), we find for U the ordinary differential equation

$$U'' + cU' + U - U^2 = 0 \quad (2.184)$$

with

$$\lim_{z \rightarrow -\infty} U(z) = 1 \quad \text{and} \quad \lim_{z \rightarrow +\infty} U(z) = 0. \quad (2.185)$$

Letting $U' = V$, the equation (2.184) is equivalent to the system

$$\frac{dU}{dz} = V, \quad \frac{dV}{dz} = -cV - U + U^2 \quad (2.186)$$

in the phase plane (U, V) . This system has two equilibrium points $(0, 0)$ and $(1, 0)$, corresponding to two steady states. Our travelling wave solution corresponds to an orbit connecting $(1, 0)$ to $(0, 0)$, with $0 < U < 1$.

We first examine the local behavior of the orbits near the equilibrium points. The coefficients matrices of the linearized systems at $(0, 0)$ and $(1, 0)$ are, respectively,

$$J(0, 0) = \begin{pmatrix} 0 & 1 \\ -1 & -c \end{pmatrix} \quad \text{and} \quad J(1, 0) = \begin{pmatrix} 0 & 1 \\ 1 & -c \end{pmatrix}.$$

The eigenvalues of $J(0, 0)$ are

$$\lambda_{\pm} = \frac{1}{2} \left[-c \pm \sqrt{c^2 - 4} \right],$$

with corresponding eigenvectors

$$\mathbf{h}_{\pm} = \begin{pmatrix} -c \mp \sqrt{c^2 - 4} \\ 2 \end{pmatrix}.$$

If $c \geq 2$ the eigenvalues are both negative while if $c < 2$ they are complex. Therefore

$$(0, 0) \text{ is a } \begin{cases} \text{stable node if } c \geq 2 \\ \text{stable focus if } c < 2. \end{cases}$$

The eigenvalues of $J(1, 0)$ are

$$\mu_{\pm} = \frac{1}{2} \left[-c \pm \sqrt{c^2 + 4} \right],$$

of opposite sign, hence $(1, 0)$ is a saddle point. The unstable and stable separatrices leave $(1, 0)$ along the directions of the two eigenvectors

$$\mathbf{k}_+ = \begin{pmatrix} c + \sqrt{c^2 + 4} \\ 2 \end{pmatrix} \quad \text{and} \quad \mathbf{k}_- = \begin{pmatrix} c - \sqrt{c^2 + 4} \\ 2 \end{pmatrix},$$

respectively.

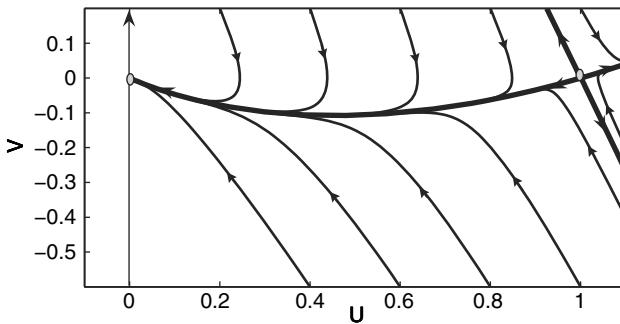


Fig. 2.18 Orbits of the system (2.186)

Now, the constraint $0 < U < 1$ rules out the case $c < 2$, since in this case U changes sign along the orbit approaching $(0, 0)$. For $c \geq 2$, all orbits⁵⁸ in a neighborhood of the origin approach $(0, 0)$ for $z \rightarrow +\infty$ asymptotically with slope λ_+ . On the other hand, the only orbit going to $(1, 0)$ as $z \rightarrow -\infty$ and remaining in the region $0 < U < 1$ is the unstable separatrix γ of the saddle point.

Figure 2.18 shows the orbits configuration in the region of interest (see Problem 2.27). The conclusion is that *for each $c \geq 2$ there exists a unique travelling wave solution of equation (2.180) with speed c . Moreover U is strictly decreasing.*

In terms of original variables, there is a unique travelling wave solution for every speed c satisfying the inequality $c \geq c_{\min} = 2\sqrt{rD}$.

Thus, we have a continuous “spectrum” of possible speeds of propagation. It turns out that the minimum speed $c = c_{\min}$ is particularly important.

Indeed, having found a travelling solution is only the beginning of the story. There is a number of questions that arise naturally. Among them, the study of the *stability* of the travelling waves or of the asymptotic behavior (as $t \rightarrow +\infty$) of a solution with an initial data u_0 of *transitional* type, that is

$$u_0(x) = \begin{cases} 1 & x \leq a \\ 0 < u_0 < 1 & a < x < b \\ 0 & x \geq b. \end{cases} \quad (2.187)$$

Should we expect that the travelling wave is insensitive to small perturbations? Does the solution with initial condition (2.187) evolve towards one of the travelling waves we have just found?

The interested reader can find the answers in the many specialized texts or papers on the subject⁵⁹. Here we only mention that among the travelling wave solutions we have found, *only the minimum speed one* can be the asymptotic representation of solutions with transitional type initial condition. The biological implication of this result is that c_{\min} determines the required speed of propagation of an advantageous gene.

⁵⁸ Except for two orbits on the stable manifold tangent to \mathbf{h}_- at $(0, 0)$, in the case $c > 2$.

⁵⁹ See for instance [26], Murray, vol I, 2001

Problems

2.1. Use the method of separation of variables to solve the following initial-Dirichlet problem:

$$\begin{cases} u_t - Du_{xx} = au & 0 < x < L, t > 0 \\ u(0, t) = u(L, t) = 0 & t > 0 \\ u(x, 0) = g(x) & 0 \leq x \leq L. \end{cases}$$

Assume $g(x) \in C^2([0, L])$, $g(0) = g(L) = 0$ and show that the solution is continuous in $[0, L] \times [0, +\infty)$ and of class C^∞ in $(0, L) \times (0, +\infty)$.

2.2. Use the method of separation of variables to solve the following initial-Neumann problem:

$$\begin{cases} u_t - u_{xx} = 0 & 0 < x < L, t > 0 \\ u_x(0, t) = u_x(L, t) = 0 & 0 < x < L \\ u(x, 0) = g(x) & t > 0. \end{cases}$$

Show that $u(x, t) \rightarrow \frac{1}{L} \int_0^L g(y) dy$ as $t \rightarrow +\infty$, for every $x \in (0, L)$.

2.3. Use the method of separation of variables to solve the following nonhomogeneous initial-Neumann problem:

$$\begin{cases} u_t - u_{xx} = tx & 0 < x < L, t > 0 \\ u(x, 0) = 1 & 0 \leq x \leq L \\ u_x(0, t) = u_x(L, t) = 0 & t > 0. \end{cases}$$

[Hint: Write the candidate solution as $u(x, t) = \sum_{k \geq 0} c_k(t) v_k(x)$ where v_k are the eigenfunctions of the eigenvalue problem associated with the homogeneous equation].

2.4. Use the method of separation of variables to solve (at least formally) the following mixed problem:

$$\begin{cases} u_t - Du_{xx} = 0 & 0 < x < L, t > 0 \\ u(x, 0) = g(x) & 0 \leq x \leq L \\ u_x(0, t) = 0 & t > 0. \\ u_x(L, t) + u(L, t) = U & \end{cases}$$

[Answer: $u(x, t) = U + \sum_{k \geq 1} c_k e^{-D\mu_k^2 t} \cos(\mu_k x)$, where the numbers μ_k are the positive solutions of the equation $\mu \tan(\mu L) = 1$, and $c_k = \frac{1}{\alpha_k} \int_0^L (g(x) - U) \cos(\mu_k x) dx$, $\alpha_k = \int_0^L (\cos \mu_k x)^2 dx$, $k \geq 1$].

2.5. *Evolution of a chemical solution.* Consider a tube of length L and constant cross section A , with axis of symmetry along the x axis. The tube contains a fluid with a saline volume concentration of c grams/cm³. Assume that:

- a) $A \ll L$, so that we can assume $c = c(x, t)$.
- b) The salt diffuses along the x axis.
- c) The velocity of the fluid is negligible.
- d) From the left boundary of the pipe, at $x = 0$, a solution of constant concentration C_0 enter the tube at a speed of R_0 cm³ per second, while at the other hand the solution is removed at the same speed.

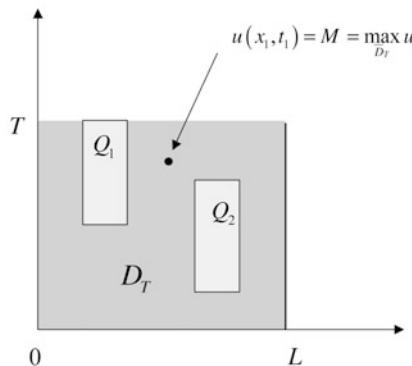


Fig. 2.19 At which points (x, t) , $u(x, t) = M$?

1. Using Fick's law (see (2.99), p. 61) to show that c satisfy the diffusion equation with mixed Neumann/Robin at the and of the tube.
2. Solve explicitly the problem and show that $c(x, t)$ tends to an equilibrium concentration $c_\infty(x)$ as $t \rightarrow +\infty$.

[*Partial answer:* 1. Let k be the diffusion coefficient in the Fick's law. The boundary conditions are:

$$c_x(0, t) = -C_0 R_0 / kA, \quad c_x(L, t) + (R_0 / kA)c(L, t) = 0.$$

$$2. c_\infty = C_0 + (R_0 / kA)(L - x).$$

2.6. Prove Corollary 2.5, p. 38.

[*Hint:* b) Let $u = v - w$, $M = \max_{\bar{Q}_T} |f_1 - f_2|$ and apply Theorem 2.4, p. 36, to $z_\pm = \pm u - Mt$.]

2.7. Let $g(t) = M$ for $0 \leq t \leq 1$ and $g(t) = M - (1-t)^4$ for $1 < t \leq 2$. Let u be the solution of $u_t - u_{xx} = 0$ in $Q_2 = (0, 2) \times (0, 2)$, $u = g$ on $\partial_p Q_2$. Compute $u(1, 1)$ and check that it is the maximum of u . Is this in contrast with the maximum principle of Theorem 2.4, p. 36?

2.8. Suppose $u = u(x, t)$ is a solution of the heat equation in a plane domain $D_T = Q_T \setminus (\bar{Q}_1 \cup \bar{Q}_2)$ where Q_1 and Q_2 are the rectangles in Fig. 2.19. Assume that u attains its maximum M at the interior point (x_1, t_1) . Where else $u = M$?

2.9. a) Find the similarity solutions of the equation $u_t - u_{xx} = 0$ of the form $u(x, t) = U(x/\sqrt{t})$ and express the result in term of the *error function*

$$\operatorname{erf}(x) = \frac{2}{\sqrt{\pi}} \int_0^x e^{-z^2} dz.$$

b) Find the solution of $u_t - u_{xx} = 0$ in $x > 0, t > 0$ satisfying the conditions $u(0, t) = 1$, $u(x, t) \rightarrow 0$ as $x \rightarrow +\infty$, $t > 0$, and $u(x, 0) = 0$, $x > 0$.

2.10. Determine for which α and β there exist similarity solutions to $u_t - u_{xx} = f(x)$ of the form $t^\alpha U(x/t^\beta)$ in each one of the following cases:

$$(a) \quad f(x) = 0, \quad (b) \quad f(x) = 1, \quad (c) \quad f(x) = x.$$

[Answer: (a) α arbitrary, $\beta = 1/2$. (b) $\alpha = 1$, $\beta = 1/2$. (c) $\alpha = 3/2$, $\beta = 1/2$].

2.11. *Reflecting barriers and Neumann condition.* Consider the symmetric random walk of Sect. 2.4. Suppose that a perfectly *reflecting* barrier is located at the point $L = \bar{m}h + \frac{h}{2} > 0$. By this we mean that if the particle hits the point $L - \frac{h}{2}$ at time t and moves to the right, then it is reflected and it comes back to $L - \frac{h}{2}$ at time $t + \tau$. Show that when $h, \tau \rightarrow 0$ and $h^2/\tau = 2D$, $p = p(x, t)$ is a solution of the problem

$$\begin{cases} p_t - Dp_{xx} = 0 & x < L, t > 0 \\ p(x, 0) = \delta(x) & x < L \\ p_x(L, t) = 0 & t > 0 \end{cases}$$

and moreover $\int_{-\infty}^L p(x, t) dx = 1$. Compute explicitly the solution.

[Answer: $p(x, t) = \Gamma_D(x, t) + \Gamma_D(x - 2L, t)$].

2.12. *Absorbing barriers and Dirichlet condition.* Consider the symmetric random walk of Sect. 2.4. Suppose that a perfectly *absorbing* barrier is located at the point $L = \bar{m}h > 0$. By this we mean that if the particle hits the point $L - h$ at time t and moves to the right then it is absorbed and stops at L . Show that when $h, \tau \rightarrow 0$ and $h^2/\tau = 2D$, $p = p(x, t)$ is a solution of the problem

$$\begin{cases} p_t - Dp_{xx} = 0 & x < L, t > 0 \\ p(x, 0) = \delta(x) & x < L \\ p(L, t) = 0 & t > 0. \end{cases}$$

Compute explicitly the solution.

[Answer: $p(x, t) = \Gamma_D(x, t) - \Gamma_D(x - 2L, t)$].

2.13. *Elastic restoring force.* The one-dimensional motion of a particle follows the following rules, where N is a natural integer and m is an integer:

1. During an interval of time τ the particle takes one step of h unit length, starting from $x = 0$ at time $t = 0$.

2. If at some step the position of the particle is at the point mh , with $-N \leq m \leq N$, it moves to the right or to the left with probability given, respectively, by:

$$p_0 = \frac{1}{2}(1 - \frac{m}{N}), \quad q_0 = \frac{1}{2}(1 + \frac{m}{N})$$

independently from the previous steps. Explain why this rules model an elastic restoring force on the particle motion. Show that, if $h^2/\tau = 2D$ and $N\tau = \gamma > 0$, when $h, \tau \rightarrow 0$ and $N \rightarrow +\infty$, the limit transition probability is a solution of the following equation:

$$p_t = Dp_{xx} + \frac{1}{\gamma}(xp_x)_x.$$

2.14. Use the partial Fourier transform $\hat{u}(\xi, t) = \int_{\mathbb{R}} e^{-ix\xi} u(x, t) dx$ to solve the global Cauchy problem (2.134) and rediscover formula (2.135).

2.15. Prove Theorem 2.16, p. 83, under the condition

$$z(x, t) \leq C, \quad x \in \mathbb{R}, \quad 0 \leq t \leq T,$$

by filling the details in the following steps.

a) Let $\sup_{\mathbb{R}} z(x, 0) = M_0$ and define

$$w(x, t) = \frac{2C}{L^2} \left(\frac{x^2}{2} + Dt \right) + M_0.$$

Check that $w_t - Dw_{xx} = 0$ and use the maximum principle to show that $w \geq z$ in the rectangle $R_L = [-L, L] \times [0, T]$.

b) Fix an arbitrary point (x_0, t_0) and choose L large enough to have $(x_0, t_0) \in R_L$. Using a) deduce that $z(x_0, t_0) \leq M_0$.

2.16. Show that, if $g(x) = 0$ for $x < 0$ and $g(x) = 1$ for $x > 0$, then

$$\lim_{(x,t) \rightarrow (0,0)} \int_{\mathbb{R}} \Gamma_D(x-y) g(y) dy$$

does not exist.

2.17. Find an explicit formula for the solution of the global Cauchy problem

$$\begin{cases} u_t = Du_{xx} + bu_x + cu & x \in \mathbb{R}, t > 0 \\ u(x, 0) = g(x) & x \in \mathbb{R} \end{cases}$$

where D, b, c are constant coefficients. Show that, if $c < 0$ and g is bounded, $u(x, t) \rightarrow 0$ as $t \rightarrow +\infty$.

2.18. Find an explicit formula for the solution of the Cauchy problem

$$\begin{cases} u_t = u_{xx} & x > 0, t > 0 \\ u(x, 0) = g(x) & x \geq 0 \\ u(0, t) = 0 & t > 0 \end{cases}$$

with g continuous and $g(0) = 0$.

2.19. Let $Q_T = \Omega \times (0, T)$, with Ω bounded domain in \mathbb{R}^n . Let $u \in C^{2,1}(Q_T) \cap C(\overline{Q}_T)$ satisfy the equation

$$u_t = D\Delta u + \mathbf{b}(\mathbf{x}, t) \cdot \nabla u + c(\mathbf{x}, t) u \quad \text{in } Q_T,$$

where \mathbf{b} and c are continuous in \overline{Q}_T . Show that if $u \geq 0$ (resp. $u \leq 0$) on $\partial_p Q_T$ then $u \geq 0$ (resp. $u \leq 0$) in Q_T .

[Hint: Assume first that $c(\mathbf{x}, t) \leq a < 0$. Then reduce to this case by setting $u = ve^{kt}$ with a suitable $k > 0$].

2.20. Fill in the details in the arguments of Sect. 6.2, leading to formulas (2.114) and (2.115).

2.21. An invasion problem. A population of density $P = P(x, y, t)$ and total mass $M(t)$ is initially (at time $t = 0$) concentrated at an isolated point, say, the origin $(0, 0)$, it grows at a constant rate $a > 0$ and it diffuses with constant diffusion coefficient D .

- Write the problem governing the evolution of P and solve it explicitly.
- Compute the evolution of the mass

$$M(t) = \int_{\mathbb{R}^2} P(x, y, t) dx dy.$$

- Let B_R be the circle centered at $(0, 0)$ and radius R . Determine $R = R(t)$ such that

$$\int_{\mathbb{R}^2 \setminus B_{R(t)}} P(x, y, t) dx dy = M(0).$$

Call *metropolitan area* the region $B_{R(t)}$ and *rural area* the region $\mathbb{R}^2 \setminus B_{R(t)}$. Compute the speed of the *metropolitan propagation front* $\partial B_{R(t)}$.

[Hint: c) We find:

$$\int_{\mathbb{R}^2 \setminus B_{R(t)}} P(x, y, t) dx dy = M(0) \exp \left\{ at - \frac{R^2(t)}{4Dt} \right\}$$

from which $R(t) = 2t\sqrt{aD}$. The speed of the metropolitan front is therefore equal to $2\sqrt{aD}$.

2.22. Pollution in dimension $n \geq 2$. Using conservation of mass, write a model of drift-diffusion for the concentration c of a pollutant in dimension $n = 2, 3$, assuming the following constitutive law for the flux vector

$$\mathbf{q}(\mathbf{x}, t) = \underbrace{\mathbf{v}(\mathbf{x}, t)c(\mathbf{x}, t)}_{\text{drift}} - \underbrace{\kappa(\mathbf{x}, t)\nabla c(\mathbf{x}, t)}_{\text{diffusion}}.$$

[Answer: $c_t + \operatorname{div}(\mathbf{v}c - \kappa(\mathbf{x}, t)\nabla c) = 0$].

2.23. Solve the following initial-Dirichlet problem in $B_1 = \{\mathbf{x} \in \mathbb{R}^3 : |\mathbf{x}| < 1\}$:

$$\begin{cases} u_t = \Delta u & \mathbf{x} \in B_1, t > 0 \\ u(\mathbf{x}, 0) = 0 & \mathbf{x} \in B_1 \\ u(\boldsymbol{\sigma}, t) = 1 & \boldsymbol{\sigma} \in \partial B_1, t > 0. \end{cases}$$

Compute $\lim_{t \rightarrow +\infty} u$.

[Hint: The solution is radial so that $u = u(r, t)$, $r = |\mathbf{x}|$. Observe that $\Delta u = u_{rr} + \frac{2}{r}u_r = \frac{1}{r}(ru)_{rr}$. Let $v = ru$, reduce to homogeneous Dirichlet condition and use separation of variables].

2.24. Solve the following initial-Neumann problem in $B_1 = \{\mathbf{x} \in \mathbb{R}^3 : |\mathbf{x}| < 1\}$:

$$\begin{cases} u_t = \Delta u & \mathbf{x} \in B_1, t > 0 \\ u(\mathbf{x}, 0) = |\mathbf{x}| & \mathbf{x} \in B_1 \\ u_\nu(\boldsymbol{\sigma}, t) = 1 & \boldsymbol{\sigma} \in \partial B_1, t > 0. \end{cases}$$

2.25. Solve the following nonhomogeneous initial-Dirichlet problem in the unit sphere B_1 ($u = u(r, t)$, $r = |\mathbf{x}|$):

$$\begin{cases} u_t - (u_{rr} + \frac{2}{r}u_r) = qe^{-t} & 0 < r < 1, t > 0 \\ u(r, 0) = U & 0 \leq r \leq 1 \\ u(1, t) = 0 & t > 0. \end{cases}$$

[Answer: The solution is

$$u(r, t) = \frac{2}{r} \sum_{n=1}^{\infty} \frac{(-1)^n}{\lambda_n} \sin(\lambda_n r) \left\{ \frac{q}{1 - \lambda_n^2} (e^{-t} - e^{-\lambda_n^2 t}) - U e^{-\lambda_n^2 t} \right\}, \quad \lambda_n = n\pi.$$

2.26. Using the maximum principle, compare the values of two call options $C^{(1)}$ and $C^{(2)}$ in the following cases:

- (a) Same exercise price and $T_1 > T_2$. (b) Same expiry time and $E_1 > E_2$.

2.27. Consider system (2.186). Justify carefully the orbit configuration of Fig. 2.18 and in particular that the unstable separatrix γ of the saddle point $(1, 0)$ connects the two equilibrium points $(0, 0)$ and $(1, 0)$, by filling in the details in the following steps:

- Let $\mathbf{F} = V\mathbf{i} + (-cV + U^2 - U)\mathbf{j}$ and \mathbf{n} be the interior normal to the boundary of the triangle Ω in Fig. 2.20. Show that, if β is suitably chosen, $\mathbf{F} \cdot \mathbf{n} > 0$ along $\partial\Omega$.
- Deduce that all the orbits of system (2.186) starting at a point in Ω cannot leave Ω (i.e. Ω is a *positively invariant region*) and converge to the origin as $z \rightarrow +\infty$.
- Finally, deduce that the unstable separatrix γ of the saddle point $(1, 0)$ approaches $(0, 0)$ as $z \rightarrow +\infty$.

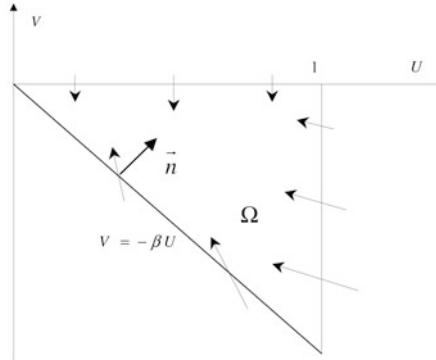


Fig. 2.20 Trapping region for the orbits of the vector field $\mathbf{F} = V\mathbf{i} + (-cV + U^2 - U)\mathbf{j}$

Chapter 3

The Laplace Equation

3.1 Introduction

The Laplace equation $\Delta u = 0$ occurs frequently in the applied sciences, in particular in the study of the *steady state phenomena*. Its solutions are called *harmonic* functions. For instance, the equilibrium position of a perfectly elastic membrane is a harmonic function as it is the velocity potential of a homogeneous fluid. Also, the steady state temperature of a homogeneous and isotropic body is a harmonic function and in this case Laplace equation constitutes the stationary counterpart (time independent) of the diffusion equation.

Slightly more generally, Poisson's equation $\Delta u = f$ plays an important role in the theory of *conservative fields* (electrical, magnetic, gravitational, ...), where the vector field is derived from the gradient of a potential.

For example, let \mathbf{E} be a force field due to a distribution of electric charges in a domain $\Omega \subset \mathbb{R}^3$. Then, in standard units, $\operatorname{div} \mathbf{E} = 4\pi\rho$, where ρ represents the density of the charge distribution. When a *potential* u exists such that $\nabla u = -\mathbf{E}$, then $\Delta u = \operatorname{div} \nabla u = -4\pi\rho$, which is Poisson's equation. If the electric field is created by charges located outside Ω , then $\rho = 0$ in Ω and u is harmonic therein. Analogously, the potential of a gravitational field due to a mass distribution is a harmonic function in a region free from mass.

In dimension two, the theories of harmonic and holomorphic functions are strictly connected¹. Indeed, the real and the imaginary part of a holomorphic function are harmonic. For instance, since the functions

$$z^m = r^m (\cos m\theta + i \sin m\theta), \quad m \in \mathbb{N},$$

¹ A complex function $f = f(z)$ is *holomorphic* in an open subset Ω of the complex plane if for every $z_0 \in \Omega$, the limit

$$\lim_{z \rightarrow z_0} \frac{f(z) - f(z_0)}{z - z_0} = f'(z_0)$$

exists and it is finite.

(r, θ) polar coordinates) are holomorphic in the whole plane \mathbb{C} , the functions

$$u(r, \theta) = r^m \cos m\theta \quad \text{and} \quad v(r, \theta) = r^m \sin m\theta, \quad m \in \mathbb{N},$$

are harmonic in \mathbb{R}^2 (called *elementary harmonics*). In Cartesian coordinates, they are harmonic polynomials; for $m = 1, 2, 3$ we find

$$x, y, xy, x^2 - y^2, x^3 - 3xy^2, 3x^2y - y^3.$$

Other examples are

$$u(x, y) = e^{\alpha x} \cos \alpha y, \quad v(x, y) = e^{\alpha x} \sin \alpha y \quad (\alpha \in \mathbb{R}),$$

the real and imaginary parts of $f(z) = e^{i\alpha z}$, both harmonic in \mathbb{R}^2 , and

$$u(r, \theta) = \log r, \quad v(r, \theta) = \theta,$$

the real and imaginary parts of $f(z) = \log_0 z = \log r + i\theta$, harmonic in $\mathbb{R}^2 \setminus (0, 0)$ and $\mathbb{R}^2 \setminus \{\theta = 0\}$, respectively.

In this chapter we present the formulation of the most important well posed problems and the classical properties of harmonic functions, focusing mainly on dimensions two and three. As in Chap. 2, we emphasize some probabilistic aspects, exploiting the connection among random walks, Brownian motion and the Laplace operator. A central notion is the concept of *fundamental solution*, that we develop in conjunction with the very basic elements of the so called *potential theory*.

3.2 Well Posed Problems. Uniqueness

Consider the Poisson equation

$$\Delta u = f \quad \text{in } \Omega \tag{3.1}$$

where $\Omega \subset \mathbb{R}^n$ is a **bounded domain**. The well posed problems associated with equation (3.1) are the stationary counterparts of the corresponding problems for the diffusion equation. Clearly here there is no initial condition. On the boundary $\partial\Omega$ we may assign:

- *Dirichlet data*

$$u = g. \tag{3.2}$$

- *Neumann data*

$$\partial_{\nu} u = h, \tag{3.3}$$

where ν is the outward normal unit vector to $\partial\Omega$.

- A *Robin (radiation) condition*

$$\partial_\nu u + \alpha u = h \quad (\alpha > 0). \quad (3.4)$$

- A *mixed condition*; for instance,

$$\begin{aligned} u &= g && \text{on } \Gamma_D \\ \partial_\nu u &= h && \text{on } \Gamma_N, \end{aligned} \quad (3.5)$$

where $\overline{\Gamma_D} \cup \overline{\Gamma_N} = \partial\Omega$, $\Gamma_D \cap \Gamma_N = \emptyset$, and Γ_D, Γ_N are relatively open regular² subsets of $\partial\Omega$.

When $g = h = 0$ we say that the above boundary conditions are *homogeneous*.

We give some interpretations. If u is the position of a perfectly flexible membrane and f is an external distributed load (vertical force per unit surface), then (3.1) models a steady state.

The Dirichlet condition corresponds to fixing the position of the membrane at its boundary. Robin condition describes an elastic attachment at the boundary while a homogeneous Neumann condition corresponds to a free vertical motion of the boundary of the membrane.

If u is the steady state concentration of a substance, the Dirichlet condition prescribes the level of u at the boundary, while the Neumann condition assigns the flux of u through the boundary.

Using Green's identity (1.13) we can prove the following uniqueness result.

Theorem 3.1. *Let $\Omega \subset \mathbb{R}^n$ be a smooth, bounded domain. Then there exists at most one solution $u \in C^2(\Omega) \cap C^1(\overline{\Omega})$ of (3.1), satisfying on $\partial\Omega$ one of the conditions (3.2), (3.4) or (3.5).*

In the case of the Neumann condition, that is when

$$\partial_\nu u = h \quad \text{on } \partial\Omega,$$

two solutions differ by a constant.

Proof. Let u and v be solutions of the same problem, sharing the same boundary data, and let $w = u - v$. Then w is harmonic and satisfies homogeneous boundary conditions (one among (3.2)-(3.5)). Substituting $u = v = w$ into (1.13) we find

$$\int_{\Omega} |\nabla w|^2 d\mathbf{x} = \int_{\partial\Omega} w \partial_\nu w d\sigma.$$

If Dirichlet, Neumann or mixed conditions hold, we have

$$\int_{\partial\Omega} w \partial_\nu w d\sigma = 0.$$

² See Definition 1.5, p. 14.

When a Robin condition holds

$$\int_{\partial\Omega} w \partial_\nu w \, d\sigma = - \int_{\partial\Omega} \alpha w^2 \, d\sigma \leq 0.$$

In any case we obtain that

$$\int_{\Omega} |\nabla w|^2 \, d\mathbf{x} \leq 0. \quad (3.6)$$

From (3.6) we infer $\nabla w = \mathbf{0}$ and therefore $w = u - v = \text{constant}$. This concludes the proof in the case of Neumann condition. In the other cases, the constant must be zero (why?), hence $u = v$. \square

Remark 3.2. Consider the Neumann problem

$$\Delta u = f \text{ in } \Omega, \quad \partial_\nu u = h \text{ on } \partial\Omega.$$

Integrating the equation on Ω and using Gauss' formula, we find

$$\int_{\Omega} f \, d\mathbf{x} = \int_{\partial\Omega} h \, d\sigma. \quad (3.7)$$

The relation (3.7) appears as a *compatibility* condition on the data f and h , that has *necessarily* to be satisfied in order for the Neumann problem to admit a solution. Thus, when having to solve a Neumann problem, the first thing to do is to check the validity of (3.7). If it does not hold, the problem does not have any solution. We will examine later the physical meaning of (3.7).

3.3 Harmonic Functions

3.3.1 Discrete harmonic functions

In Chap. 2 we have examined the connection between Brownian motion and diffusion equation. We now go back to the multidimensional symmetric random walk considered in Sect. 2.6, analyzing its relation with the Laplace operator Δ . For simplicity we will work in dimension $n = 2$ but both arguments and conclusions may be easily extended to any dimension $n > 2$. We fix a time step $\tau > 0$, a space step $h > 0$ and denote by $h\mathbb{Z}^2$ the *lattice* of points $\mathbf{x} = (x_1, x_2)$ whose coordinates are integer multiples of h . Let $p(\mathbf{x}, t) = p(x_1, x_2, t)$ be the transition probability function, giving the probability to find our random particle at \mathbf{x} , at time t . From the total probability formula, we found a difference equation for p , that we rewrite in dimension two:

$$p(\mathbf{x}, t + \tau) = \frac{1}{4} \{p(\mathbf{x} + h\mathbf{e}_1, t) + p(\mathbf{x} - h\mathbf{e}_1, t) + p(\mathbf{x} + h\mathbf{e}_2, t) + p(\mathbf{x} - h\mathbf{e}_2, t)\}.$$

We can write this formula in a more significant way by introducing the *mean value operator* M_h , whose action on a function $u = u(\mathbf{x})$ is defined by the following

formula:

$$\begin{aligned} M_h u(\mathbf{x}) &= \frac{1}{4} \{u(\mathbf{x}+h\mathbf{e}_1) + u(\mathbf{x}-h\mathbf{e}_1) + u(\mathbf{x}+h\mathbf{e}_2) + u(\mathbf{x}-h\mathbf{e}_2)\} \\ &= \frac{1}{4} \sum_{|\mathbf{x}-\mathbf{y}|=h} u(\mathbf{y}). \end{aligned}$$

Note that $M_h u(\mathbf{x})$ gives the average of u over the points of the lattice $h\mathbb{Z}^2$ at distance h from \mathbf{x} . We say that these points constitute the *discrete neighborhood of \mathbf{x} of radius h* .

Thus, we can write

$$p(\mathbf{x}, t + \tau) = M_h p(\mathbf{x}, t). \quad (3.8)$$

In (3.8), the probability p at time $t + \tau$ is determined by the action of M_h at the previous time, and then it is natural to interpret the mean value operator as *the generator of the random walk*.

Now we come to the Laplacian. If u is twice continuously differentiable, it is not difficult to show that³

$$\lim_{h \rightarrow 0} \frac{M_h u(\mathbf{x}) - u(\mathbf{x})}{h^2} \rightarrow \frac{1}{4} \Delta u(\mathbf{x}). \quad (3.9)$$

The formula (3.9) induces to define, for any fixed $h > 0$, a *discrete Laplace operator* through the formula

$$\Delta_h^* = \frac{4}{h^2} (M_h - I)$$

where I denotes the *identity* operator (i.e. $Iu = u$). The operator Δ_h^* acts on functions u defined in the whole lattice $h\mathbb{Z}^2$ and, coherently, we say that u is *d-harmonic* (d for *discrete*) if $\Delta_h^* u = 0$.

Thus, the value of a *d-harmonic* function at any point \mathbf{x} is given by the average of the values at the points in the discrete neighborhood of \mathbf{x} of radius h .

We can proceed further and define a discrete Dirichlet problem. Let A be a subset of $h\mathbb{Z}^2$. We say that $\mathbf{x} \in A$ is:

- An *interior* point of A if its h -neighborhood is contained in A .
- A *boundary points* (Fig. 3.1) if it is not an interior point but its h -neighborhood contains at least an interior point. The set of the boundary points of A , the *boundary* of A , is denoted by ∂A .

The points of A whose h -neighborhood does not contain interior points of A are considered *isolated points*.

³ Using a second order Taylor's polynomial, after some simplifications, we get:

$$M_h u(\mathbf{x}) = u(\mathbf{x}) + \frac{h^2}{4} \{u_{x_1 x_1}(\mathbf{x}) + u_{x_2 x_2}(\mathbf{x})\} + o(h^2)$$

from which formula (3.9) comes easily.

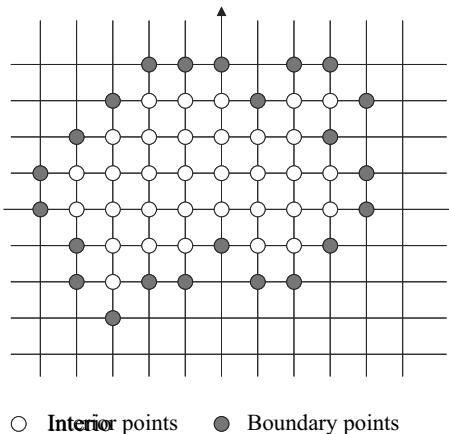


Fig. 3.1 A domain for the discrete Dirichlet problem

We say that A is *connected* if A does not have isolated points and, given any couple of points $\mathbf{x}_0, \mathbf{x}_1$ in A , it is possible to connect them by a walk⁴ on $h\mathbb{Z}^2$ entirely contained in A .

Discrete Dirichlet problem. Let A be a *bounded connected* subset of $h\mathbb{Z}^2$ and g be a function defined on the boundary ∂A of A . We want to determine u , defined on A , such that

$$\begin{cases} \Delta_h^* u = 0 \text{ at the interior points of } A \\ u = g \quad \text{on } \partial A. \end{cases} \quad (3.10)$$

We deduce immediately three important properties of a solution u :

1. Maximum principle: *If u attains its maximum or its minimum at an interior point then u is constant.* Indeed, suppose that $\mathbf{x} \in A$ is an interior point and that $u(\mathbf{x}) = M \geq u(\mathbf{y})$, for every $\mathbf{y} \in A$. Since $u(\mathbf{x})$ is the average of the four values of u at the points at distance h from \mathbf{x} , at all these points u must be equal to M . Let $\mathbf{x}_1 \neq \mathbf{x}$ be one of these neighboring points. By the same argument, $u(\mathbf{y}) = M$ for every \mathbf{y} in the h -neighborhood of \mathbf{x}_1 . Since A is connected, proceeding in this way we prove that $u(\mathbf{y}) = M$ at every point of A .
2. *u attains its maximum and its minimum on ∂A .* This is an immediate consequence of 1.
3. *The solution of the discrete Dirichlet problem is unique.*

The discrete Dirichlet problem (3.10) has a remarkable probabilistic interpretation that can be used to construct its solution. Let us go back to our random particle.

⁴ Recall that consecutive points in a walk have distance h .

First of all, we want to show that whatever its starting point $\mathbf{x} \in A$ is, the particle hits the boundary ∂A with probability one.

For every $\Gamma \subseteq \partial A$, we denote by $P(\mathbf{x}, \Gamma)$ the probability that the particle starting from $\mathbf{x} \in A$ hits ∂A for the first time at a point $\mathbf{y} \in \Gamma$. We have to prove that $P(\mathbf{x}, \partial A) = 1$ for every $\mathbf{x} \in A$.

Clearly, if $\mathbf{x} \in \Gamma$ we have $P(\mathbf{x}, \Gamma) = 1$, while if $\mathbf{x} \in \partial A \setminus \Gamma$, $P(\mathbf{x}, \Gamma) = 0$. It turns out that, for fixed Γ , the function

$$\mathbf{x} \mapsto w_\Gamma(\mathbf{x}) = P(\mathbf{x}, \Gamma)$$

is *d-harmonic* in the interior of A , that is $\Delta_h^* w_\Gamma = 0$. To see this, denote by $p(1, \mathbf{x}, \mathbf{y})$ the *one step transition probability*, i.e. the probability to go from \mathbf{x} to \mathbf{y} in one step. Given the symmetry of the walk, we have $p(1, \mathbf{x}, \mathbf{y}) = 1/4$ if $|\mathbf{x} - \mathbf{y}| = h$ and $p(1, \mathbf{x}, \mathbf{y}) = 0$ otherwise.

Now, to hit Γ starting from \mathbf{x} , the particle first hits a point \mathbf{y} in its h -neighborhood and from there it reaches Γ , independently of the first step. Then, by the total probability formula we can write

$$w_\Gamma(\mathbf{x}) = P(\mathbf{x}, \Gamma) = \sum_{\mathbf{y} \in h\mathbb{Z}^2} p(1, \mathbf{x}, \mathbf{y}) P(\mathbf{y}, \Gamma) = M_h P(\mathbf{x}, \Gamma) = M_h w_\Gamma(\mathbf{x}),$$

which entails

$$0 = (I - M_h)w_\Gamma = \frac{h^2}{4} \Delta_h^* w_\Gamma.$$

Thus, w_Γ is *d-harmonic* in A . In particular, $w_{\partial A}(\mathbf{x}) = P(\mathbf{x}, \partial A)$ is *d-harmonic* in A and $w_{\partial A} = 1$ on ∂A . On the other hand, the function $z(\mathbf{x}) \equiv 1$ satisfies the same discrete Dirichlet problem, so that, by the uniqueness property 3 above,

$$w_{\partial A}(\mathbf{x}) = P(\mathbf{x}, \partial A) \equiv 1 \text{ in } A. \quad (3.11)$$

This means that the particle hits the boundary ∂A with probability one. On the other hand, by linearity, if $\Gamma_1, \Gamma_2 \subset \partial A$ and $\Gamma_1 \cap \Gamma_2 = \emptyset$, then: $P(\mathbf{x}, \Gamma_1 \cup \Gamma_2) = P(\mathbf{x}, \Gamma_1) + P(\mathbf{x}, \Gamma_2)$. Thus the set function

$$\Gamma \mapsto P(\mathbf{x}, \Gamma)$$

defines a probability measure on ∂A , for any fixed $\mathbf{x} \in A$.

We now construct the solution u to (3.10). Interpret the boundary data g as a *payoff*: if the particle starts from \mathbf{x} and hits the boundary for the first time at \mathbf{y} , it gains $g(\mathbf{y})$. We have:

Proposition 3.3. *The value $u(\mathbf{x})$ is given by the expected value of the winnings $g(\cdot)$ with respect to the probability $P(\mathbf{x}, \cdot)$. That is*

$$u(\mathbf{x}) = \sum_{\mathbf{y} \in \partial A} g(\mathbf{y}) P(\mathbf{x}, \{\mathbf{y}\}). \quad (3.12)$$

Proof. Each term

$$g(\mathbf{y}) P(\mathbf{x}, \{\mathbf{y}\}) = g(\mathbf{y}) w_{\{\mathbf{y}\}}(\mathbf{x})$$

is d -harmonic in A and therefore u is d -harmonic in A as well. Moreover, if $\mathbf{x} \in \partial A$ then $u(\mathbf{x}) = g(\mathbf{x})$, since each term in the sum is equal to $g(\mathbf{x})$ if $\mathbf{y} = \mathbf{x}$ or to zero if $\mathbf{y} \neq \mathbf{x}$. \square

As $h \rightarrow 0$, formula (3.9) shows that, formally, d -harmonic functions “become” harmonic. Thus, it seems reasonable that appropriate versions of the above properties and results should hold in the continuous case. We start with the mean value properties.

3.3.2 Mean value properties

Guided by their discrete characterization, we want to establish some fundamental properties of the harmonic functions. To be precise, we say that a function u is *harmonic* in a domain $\Omega \subseteq \mathbb{R}^n$ if $u \in C^2(\Omega)$ and $\Delta u = 0$ in Ω .

Since d -harmonic functions are defined through a mean value property, we expect that harmonic functions inherit a mean value property of the following kind: the value at the center of any ball $B \subset\subset \Omega$, i.e. compactly contained in Ω , equals the average of the values on the boundary ∂B . Actually, something more is true.

Theorem 3.4. *Let u be harmonic in $\Omega \subseteq \mathbb{R}^n$. Then, for any ball $B_R(\mathbf{x}) \subset\subset \Omega$ the following mean value formulas hold:*

$$u(\mathbf{x}) = \frac{n}{\omega_n R^n} \int_{B_R(\mathbf{x})} u(\mathbf{y}) d\mathbf{y} \quad (3.13)$$

$$u(\mathbf{x}) = \frac{1}{\omega_n R^{n-1}} \int_{\partial B_R(\mathbf{x})} u(\boldsymbol{\sigma}) d\boldsymbol{\sigma} \quad (3.14)$$

where ω_n is the surface measure⁵ of ∂B_1 .

Proof. Let us start from the second formula. For $r < R$ define

$$g(r) = \frac{1}{\omega_n r^{n-1}} \int_{\partial B_r(\mathbf{x})} u(\boldsymbol{\sigma}) d\boldsymbol{\sigma}.$$

Perform the change of variables $\boldsymbol{\sigma} = \mathbf{x} + r\boldsymbol{\sigma}'$. Then

$$\boldsymbol{\sigma}' \in \partial B_1(\mathbf{0}), \quad d\boldsymbol{\sigma} = r^{n-1} d\boldsymbol{\sigma}'$$

and

$$g(r) = \frac{1}{\omega_n} \int_{\partial B_1(\mathbf{0})} u(\mathbf{x} + r\boldsymbol{\sigma}') d\boldsymbol{\sigma}'.$$

Let $v(\mathbf{y}) = u(\mathbf{x} + r\mathbf{y})$ and observe that

$$\begin{aligned} \nabla v(\mathbf{y}) &= r \nabla u(\mathbf{x} + r\mathbf{y}) \\ \Delta v(\mathbf{y}) &= r^2 \Delta u(\mathbf{x} + r\mathbf{y}). \end{aligned}$$

⁵ See footnote 1, pag. 7.

Then we have

$$\begin{aligned} g'(r) &= \frac{1}{\omega_n} \int_{\partial B_1(\mathbf{0})} \frac{d}{dr} u(\mathbf{x} + r\sigma') d\sigma' = \frac{1}{\omega_n} \int_{\partial B_1(\mathbf{0})} \nabla u(\mathbf{x} + r\sigma') \cdot \sigma' d\sigma' \\ &= \frac{1}{\omega_n r} \int_{\partial B_1(\mathbf{0})} \nabla v(\sigma') \cdot \sigma' d\sigma' = (\text{divergence theorem}) \\ &= \frac{1}{\omega_n r} \int_{B_1(\mathbf{0})} \Delta v(\mathbf{y}) d\mathbf{y} = \frac{r}{\omega_n} \int_{B_1(\mathbf{0})} \Delta u(\mathbf{x} + r\mathbf{y}) d\mathbf{y} = 0. \end{aligned}$$

Thus, g is constant and since $g(r) \rightarrow u(\mathbf{x})$ for $r \rightarrow 0$, we get (3.14).

To obtain (3.13), let $R = r$ in (3.14), multiply by r and integrate both sides between 0 and R . We find

$$\frac{R^n}{n} u(\mathbf{x}) = \frac{1}{\omega_n} \int_0^R dr \int_{\partial B_r(\mathbf{x})} u(\sigma) d\sigma = \frac{1}{\omega_n} \int_{B_R(\mathbf{x})} u(\mathbf{y}) d\mathbf{y}$$

from which (3.13) follows. \square

Even more significant is a converse of Theorem 3.4. We say that a **continuous function u satisfies the mean value property in Ω , if (3.13) or (3.14) holds for any ball $B_R(\mathbf{x}) \subset\subset \Omega$.** It turns out that if u is continuous and possesses the mean value property in a domain Ω , then u is harmonic in Ω . Thus we obtain a characterization of harmonic functions through a mean value property, as in the discrete case. As a by product, we deduce that every harmonic function in a domain Ω is continuously differentiable of any order in Ω , that is, it belongs to $C^\infty(\Omega)$. Notice that this is not a trivial fact since it involves derivatives not appearing in the expression of the Laplace operator. For instance, $u(x, y) = x + y|y|$ is a solution of $u_{xx} + u_{xy} = 0$ in all \mathbb{R}^2 but it is not twice differentiable with respect to y at $(0, 0)$.

Theorem 3.5. *Let $u \in C(\Omega)$. If u satisfies the mean value property, then $u \in C^\infty(\Omega)$ and it is harmonic in Ω .*

We postpone the proof to the end of the Sect. 3.4. Here we point out the following simple consequence.

Theorem 3.6. *Let $\Omega \subset \mathbb{R}^n$ be a bounded domain and $\{u_k\}_{k \geq 1}$ be a sequence of harmonic functions in Ω . If u_k converges uniformly to u in $\overline{\Omega}$, then $u \in C^\infty(\Omega)$ and is harmonic in Ω .*

Proof. Let $\mathbf{x} \in \Omega$ and $B_r(\mathbf{x}) \subset\subset \Omega$. By Theorem 3.4, for every $k \geq 1$ we can write

$$u_k(\mathbf{x}) = \frac{1}{\omega_n R^n} \int_{B_R(\mathbf{x})} u_k(\mathbf{y}) d\mathbf{y}. \quad (3.15)$$

Since $u_k \rightarrow u$ uniformly in Ω , $\int_{B_R(\mathbf{x})} u_k(\mathbf{y}) d\mathbf{y} \rightarrow \int_{B_R(\mathbf{x})} u(\mathbf{y}) d\mathbf{y}$ so that, letting $k \rightarrow +\infty$ in (3.15) we get

$$u(\mathbf{x}) = \frac{1}{\omega_n R^n} \int_{B_R(\mathbf{x})} u(\mathbf{y}) d\mathbf{y}.$$

Since \mathbf{x} is an arbitrary point in Ω , we deduce that u has the mean value property and therefore is harmonic. \square

3.3.3 Maximum principles

As in the discrete case, a function satisfying the mean value property in a domain⁶ Ω cannot attain its maximum or minimum at an *interior point of* Ω , unless it is constant. In case Ω is bounded and u (nonconstant) is continuous up to the boundary of Ω , it follows that u attains both its maximum and minimum **only on** $\partial\Omega$. This result expresses a maximum principle that we state precisely in the following theorem.

Theorem 3.7. *Let $\Omega \subseteq \mathbb{R}^n$ be a domain and $u \in C(\bar{\Omega})$. If u has the mean value property and attains its maximum or minimum at $p \in \Omega$, then u is constant. In particular, if Ω is bounded and $u \in C(\bar{\Omega})$ is not constant, then, for every $\mathbf{x} \in \Omega$,*

$$u(\mathbf{x}) < \max_{\partial\Omega} u \quad \text{and} \quad u(\mathbf{x}) > \min_{\partial\Omega} u.$$

Proof. Let \mathbf{p} be a minimum point⁷ for u :

$$m = u(\mathbf{p}) \leq u(\mathbf{y}), \quad \forall \mathbf{y} \in \Omega.$$

We want to show that $u \equiv m$ in Ω . Let \mathbf{q} be another arbitrary point in Ω . Since Ω is connected, it is possible to find a finite sequence of balls $B(\mathbf{x}_j) \subset\subset \Omega$, $j = 0, \dots, N$, such that (Fig. 3.2):

- $\mathbf{x}_j \in B(\mathbf{x}_{j-1})$, $j = 1, \dots, N$.
- $x_0 = \mathbf{p}$, $x_N = \mathbf{q}$.

The mean value property gives

$$m = u(\mathbf{p}) = \frac{1}{|B(\mathbf{p})|} \int_{B(\mathbf{p})} u(\mathbf{y}) d\mathbf{y}.$$

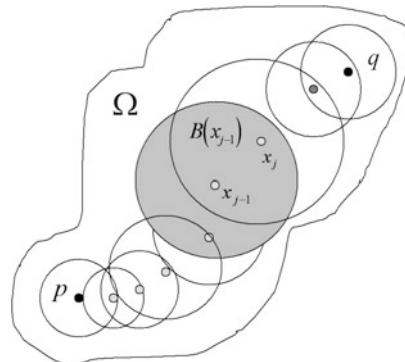


Fig. 3.2 A sequence of overlapping circles connecting the points \mathbf{p} and \mathbf{q}

⁶ Recall that a *domain* is an open *connected* set.

⁷ The argument for the maximum is the same.

Suppose there exists $\mathbf{z} \in B(\mathbf{p})$ such that $u(\mathbf{z}) > m$. Then, given a circle $B_r(\mathbf{z}) \subset B(\mathbf{p})$, we can write:

$$\begin{aligned} m &= \frac{1}{|B(\mathbf{p})|} \int_{B(\mathbf{p})} u(\mathbf{y}) d\mathbf{y} \\ &= \frac{1}{|B(\mathbf{p})|} \left\{ \int_{B(\mathbf{p}) \setminus B_r(\mathbf{z})} u(\mathbf{y}) d\mathbf{y} + \int_{B_r(\mathbf{z})} u(\mathbf{y}) d\mathbf{y} \right\}. \end{aligned} \quad (3.16)$$

Since $u(\mathbf{y}) \geq m$ for every \mathbf{y} and, by the mean value again,

$$\int_{B_r(\mathbf{z})} u(\mathbf{y}) d\mathbf{y} = u(\mathbf{z}) |B_r(\mathbf{z})| > m |B_r(\mathbf{z})|,$$

continuing from (3.16) we obtain

$$> \frac{1}{|B(\mathbf{p})|} \{m |B(\mathbf{p}) \setminus B_r(\mathbf{z})| + m |B_r(\mathbf{z})|\} = m,$$

and therefore the contradiction $m > m$.

Thus it must be that $u \equiv m$ in $B(\mathbf{p})$ and in particular $u(\mathbf{x}_1) = m$. We repeat now the same argument with \mathbf{x}_1 in place of \mathbf{p} to show that $u \equiv m$ in $B(\mathbf{x}_1)$ and in particular $u(\mathbf{x}_2) = m$. Iterating the procedure we eventually deduce that $u(\mathbf{x}_N) = u(\mathbf{q}) = m$. Since \mathbf{q} is an arbitrary point of Ω , we conclude that $u \equiv m$ in Ω . \square

An important consequence of the maximum principle is the following corollary.

Corollary 3.8. *Let $\Omega \subset \mathbb{R}^n$ be a bounded domain and $g \in C(\partial\Omega)$. The problem*

$$\begin{cases} \Delta u = 0 & \text{in } \Omega \\ u = g & \text{on } \partial\Omega \end{cases} \quad (3.17)$$

has at most a solution $u_g \in C^2(\Omega) \cap C(\overline{\Omega})$. Moreover, let u_{g_1} and u_{g_2} be the solutions corresponding to the data $g_1, g_2 \in C(\partial\Omega)$. Then:

(a) (*Comparison*). If $g_1 \geq g_2$ on $\partial\Omega$ then

$$u_{g_1} \geq u_{g_2} \quad \text{in } \Omega. \quad (3.18)$$

(b) (*Stability*).

$$|u_{g_1}(\mathbf{x}) - u_{g_2}(\mathbf{x})| \leq \max_{\partial\Omega} |g_1 - g_2| \quad \text{for every } \mathbf{x} \in \Omega. \quad (3.19)$$

Proof. We first show (a) and (b). Let $w = u_{g_1} - u_{g_2}$. Then w is harmonic and $w = g_1 - g_2 \geq 0$ on $\partial\Omega$. From Theorem 3.7

$$w(\mathbf{x}) \geq \min_{\partial\Omega} (g_1 - g_2) \geq 0 \quad \text{for every } \mathbf{x} \in \Omega.$$

This is (3.18). To prove (b), apply Theorem 3.7 to w and $-w$ to find

$$\pm w(\mathbf{x}) \leq \max_{\partial\Omega} |g_1 - g_2| \quad \text{for every } \mathbf{x} \in \Omega$$

which is equivalent to (3.19).

Now if $g_1 = g_2$, (3.19) implies $w = u_{g_1} - u_{g_2} \equiv 0$, so that the Dirichlet problem (3.17) has at most one solution. \square

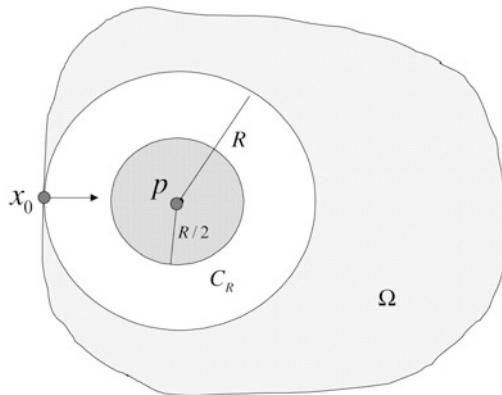


Fig. 3.3 Interior sphere condition at \mathbf{x}_0

Remark 3.9. Inequality (3.19) is a *stability estimate*. Indeed, suppose g is known within an absolute error less than ε , or, in other words, suppose g_1 is an approximation of g and $\max_{\partial\Omega} |g - g_1| < \varepsilon$; then (3.19) gives

$$\max_{\bar{\Omega}} |u_{g_1} - u_g| < \varepsilon$$

so that the approximate solution is known within the same absolute error.

Theorem 3.7 states that if $u \in C^2(\Omega) \cap C(\bar{\Omega})$ is a nonconstant harmonic function, then it attains its maximum and minimum *only* at the boundary. Maximum principles of various types hold for more general operators, as we shall see in Chap. 8.

3.3.4 The Hopf principle

Another important result concerns the slope of a harmonic function u at a boundary point satisfying a so-called *interior sphere condition*, where u attains an extremum.

Definition 3.10. We say that $\mathbf{x}_0 \in \partial\Omega$ satisfies an *interior sphere condition* if there exists a ball $B_R(\mathbf{p}_0) \subset \Omega$ such that $B_R(\mathbf{p}_0) \cap \partial\Omega = \{\mathbf{x}_0\}$ (Fig. 3.3).

The following *Hopf's principle* holds.

Theorem 3.11. Let Ω be a domain in \mathbb{R}^n and $u \in C^2(\Omega) \cap (\bar{\Omega})$ be harmonic in Ω . Assume that:

- i) At $\mathbf{x}_0 \in \partial\Omega$ an interior sphere condition holds.
- ii) $u(\mathbf{x}_0) = m = \min_{\bar{\Omega}} u$.

If the exterior normal derivative of u at \mathbf{x}_0 exists, then

$$\partial_{\nu} u(\mathbf{x}_0) < 0 \quad (3.20)$$

unless $u \equiv m$ in Ω .

Proof. Replacing u by $u - m$, we can assume that $m = 0$. Referring to Fig. 3.3, the idea is to construct a radial function z , in the spherical shell

$$C_R = B_R(\mathbf{p}) \setminus \overline{B}_{R/2}(\mathbf{p}),$$

with the following properties:

- a) $z = m_1 = \min_{\partial B_{R/2}(\mathbf{p})} u$ on $\partial B_{R/2}(\mathbf{p})$, $z = 0$ on $\partial B_R(\mathbf{p})$ and, in particular, $z(\mathbf{x}_0) = 0$.
- b) $\partial_{\nu} z(\mathbf{x}_0) < 0$.
- c) $\Delta z > 0$ in C_R .

Let u be nonconstant. Then $m_1 > 0$, by Theorem 3.7. Define

$$z(\mathbf{x}) = \frac{m_1}{\gamma} \left\{ e^{-\alpha|\mathbf{x}-\mathbf{p}|^2} - e^{-\alpha R^2} \right\}$$

where $\gamma = e^{-\alpha R^2/4} - e^{\alpha R^2}$ and $\alpha > 0$ is to be suitably chosen.

Then a) is satisfied. Observing that the exterior normal at \mathbf{x}_0 is given by $\nu = (\mathbf{x}_0 - \mathbf{p})/R$, we compute

$$\partial_{\nu} z(\mathbf{x}_0) = \nabla z(\mathbf{x}_0) \cdot \nu = -\frac{2\alpha m_1}{\gamma} e^{-\alpha R^2} (\mathbf{x}_0 - \mathbf{p}) \cdot \frac{(\mathbf{x}_0 - \mathbf{p})}{R} = -\frac{2\alpha R m_1}{\gamma} e^{-\alpha R^2} < 0$$

so that b) holds. Finally, we have, in C_R :

$$\Delta z(\mathbf{x}) = \frac{m_1}{\gamma} e^{-\alpha|\mathbf{x}-\mathbf{p}|^2} \{4\alpha^2 |\mathbf{x}-\mathbf{p}|^2 - 2n\alpha\} \geq \frac{m_1}{\gamma} e^{-\alpha R^2} \{\alpha^2 R^2 - 2n\alpha\}$$

and therefore c) is satisfied if $\alpha > 2n/R$.

Once z is constructed, we set $w = u - z$. Then $w \geq 0$ on ∂C_R and $\Delta w < 0$ in C_R . Hence $w > 0$ in C_R . In fact if $\mathbf{x}_1 \in C_R$ and $w(\mathbf{x}_1) = \min_{C_R} w \leq 0$, then we have the contradiction $\Delta w(\mathbf{x}_1) \geq 0$. Therefore, since $w(\mathbf{x}_0) = 0$, we have

$$\partial_{\nu} u(\mathbf{x}_0) - \partial_{\nu} z(\mathbf{x}_0) = \partial_{\nu} w(\mathbf{x}_0) \leq 0$$

and from property b), $\partial_{\nu} u(\mathbf{x}_0) \leq \partial_{\nu} z(\mathbf{x}_0) < 0$. The proof is complete. \square

The key point in Theorem 3.11 is that the normal derivative of u at \mathbf{x}_0 cannot be zero. This has a number of consequences, for instance a maximum principle for the Robin problem (see Problem 3.4).

3.3.5 The Dirichlet problem in a disc. Poisson's formula

To prove the existence of a solution to one of the boundary value problems we considered in Sect. 3.2 is, in general, not an elementary task. In Chap. 8, we solve this question in a wider context, using the more advanced tools of Functional Analysis. However, in special cases, elementary methods, like separation of variables, work. We use it to compute the solution of the Dirichlet problem in a disc. Precisely, let

$B_R = B_R(\mathbf{p})$ be the disc of radius R centered at $\mathbf{p} = (p_1, p_2)$ and $g \in C(\partial B_R)$. We want to prove the following theorem.

Theorem 3.12. *The unique solution $u \in C^2(B_R) \cap C(\overline{B}_R)$ of the problem*

$$\begin{cases} \Delta u = 0 & \text{in } B_R \\ u = g & \text{on } \partial B_R \end{cases} \quad (3.21)$$

is given by **Poisson's formula**

$$u(\mathbf{x}) = \frac{R^2 - |\mathbf{x} - \mathbf{p}|^2}{2\pi R} \int_{\partial B_R(\mathbf{p})} \frac{g(\boldsymbol{\sigma})}{|\mathbf{x} - \boldsymbol{\sigma}|^2} d\sigma. \quad (3.22)$$

In particular, $u \in C^\infty(B_R)$.

Proof. The symmetry of the domain suggests the use of polar coordinates

$$x_1 = p_1 + r \cos \theta \quad x_2 = p_2 + r \sin \theta.$$

Accordingly, let

$$U(r, \theta) = u(p_1 + r \cos \theta, p_2 + r \sin \theta), \quad G(\theta) = g(p_1 + R \cos \theta, p_2 + R \sin \theta).$$

The Laplace equation becomes⁸

$$U_{rr} + \frac{1}{r} U_r + \frac{1}{r^2} U_{\theta\theta} = 0, \quad 0 < r < R, \quad 0 \leq \theta \leq 2\pi, \quad (3.23)$$

with the Dirichlet condition

$$U(R, \theta) = G(\theta), \quad 0 \leq \theta \leq 2\pi.$$

Since we ask that u be continuous in \overline{B}_R , then U and G have to be continuous in $[0, R] \times [0, 2\pi]$ and $[0, 2\pi]$, respectively; moreover both have to be 2π -periodic with respect to θ .

We now use the method of separation of variables, by looking first for solutions of the form

$$U(r, \theta) = v(r) w(\theta),$$

with v, w bounded and w 2π -periodic. Substitution into (3.23) gives

$$v''(r) w(\theta) + \frac{1}{r} v'(r) w(\theta) + \frac{1}{r^2} v(r) w''(\theta) = 0$$

or, separating the variables,

$$-\frac{r^2 v''(r) + r v'(r)}{v(r)} = \frac{w''(\theta)}{w(\theta)}.$$

This identity is possible only when the two quotients have a common constant value λ .

⁸ See Appendix B.

Thus we are lead to the ordinary differential equation

$$r^2 v''(r) + rv'(r) - \lambda v(r) = 0 \quad (3.24)$$

and to the eigenvalue problem

$$\begin{cases} w''(\theta) - \lambda w(\theta) = 0 \\ w(0) = w(2\pi). \end{cases} \quad (3.25)$$

We leave to the reader to check that problem (3.25) has only the zero solution for $\lambda \geq 0$. If

$$\lambda = -\mu^2, \quad \mu > 0,$$

the differential equation in (3.25) has the general integral

$$w(\theta) = a \cos \mu\theta + b \sin \mu\theta, \quad (a, b \in \mathbb{R}).$$

The 2π -periodicity forces $\mu = m$, a nonnegative integer.

Equation (3.24), with $\lambda = -m^2$, has the general solution⁹

$$v(r) = d_1 r^{-m} + d_2 r^m, \quad (d_1, d_2 \in \mathbb{R}).$$

Since v has to be bounded, we exclude r^{-m} , $m > 0$, and hence $d_1 = 0$.

We have found a countable number of 2π -periodic harmonic functions

$$r^m \{a_m \cos m\theta + b_m \sin m\theta\}, \quad m = 0, 1, 2, \dots \quad (3.26)$$

We superpose now the functions in (3.26) by writing

$$U(r, \theta) = a_0 + \sum_{m=1}^{\infty} r^m \{a_m \cos m\theta + b_m \sin m\theta\}, \quad (3.27)$$

where the coefficients a_m and b_m are still to be chosen in order to satisfy the boundary condition

$$\lim_{(r, \theta) \rightarrow (R, \xi)} U(r, \theta) = G(\xi), \quad \forall \xi \in [0, 2\pi]. \quad (3.28)$$

Case $G \in C^1([0, 2\pi])$. In this case, G can be expanded in a uniformly convergent Fourier series

$$G(\xi) = \frac{\alpha_0}{2} + \sum_{m=1}^{\infty} \{\alpha_m \cos m\xi + \beta_m \sin m\xi\},$$

where

$$\alpha_m = \frac{1}{\pi} \int_0^{2\pi} G(\varphi) \cos m\varphi \, d\varphi, \quad \beta_m = \frac{1}{\pi} \int_0^{2\pi} G(\varphi) \sin m\varphi \, d\varphi.$$

Then, the boundary condition (3.28) is satisfied if we choose

$$a_0 = \frac{\alpha_0}{2}, \quad a_m = R^{-m} \alpha_m, \quad b_m = R^{-m} \beta_m.$$

⁹ It is an Euler equation. The change of variables $s = \log r$ reduces it to the equation

$$v''(s) - m^2 v(s) = 0.$$

Substitution of these values of a_0, a_m, b_m into (3.27) gives, for $r \leq R$,

$$\begin{aligned} U(r, \theta) &= \frac{\alpha_0}{2} + \frac{1}{\pi} \sum_{m=1}^{\infty} \left(\frac{r}{R}\right)^m \int_0^{2\pi} G(\varphi) \{ \cos m\varphi \cos m\theta + \sin m\varphi \sin m\theta \} d\varphi \\ &= \frac{1}{\pi} \int_0^{2\pi} G(\varphi) \left[\frac{1}{2} + \sum_{m=1}^{\infty} \left(\frac{r}{R}\right)^m \cos m(\varphi - \theta) \right] d\varphi \\ &= \frac{1}{\pi} \int_0^{2\pi} G(\varphi) \left[-\frac{1}{2} + \sum_{m=0}^{\infty} \left(\frac{r}{R}\right)^m \cos m(\varphi - \theta) \right] d\varphi. \end{aligned}$$

Note that in the second equality above, the exchange of sum and integration is possible because of the uniform convergence of the series. Moreover, for $r < R$, we can differentiate under the integral sign and then term by term as many times as we want. Therefore, since for every $m \geq 1$ the functions

$$\left(\frac{r}{R}\right)^m \cos m(\varphi - \theta)$$

are smooth and harmonic, also $U \in C^\infty(B_R)$ and is harmonic for $r < R$. To obtain a better formula, observe that

$$\sum_{m=0}^{\infty} \left(\frac{r}{R}\right)^m \cos m(\varphi - \theta) = \operatorname{Re} \left[\sum_{m=0}^{\infty} \left(e^{i(\varphi-\theta)} \frac{r}{R}\right)^m \right].$$

Since

$$\operatorname{Re} \sum_{m=0}^{\infty} \left(e^{i(\varphi-\theta)} \frac{r}{R}\right)^m = \operatorname{Re} \frac{1}{1 - e^{i(\varphi-\theta)} \frac{r}{R}} = \frac{R^2 - rR \cos(\varphi - \theta)}{R^2 + r^2 - 2rR \cos(\varphi - \theta)}$$

we find

$$-\frac{1}{2} + \sum_{m=0}^{\infty} \left(\frac{r}{R}\right)^m \cos m(\varphi - \theta) = \frac{1}{2} \frac{R^2 - r^2}{R^2 + r^2 - 2rR \cos(\varphi - \theta)}. \quad (3.29)$$

Inserting (3.29) into the formula for U , we get **Poisson's formula** in polar coordinates:

$$U(r, \theta) = \frac{R^2 - r^2}{2\pi} \int_0^{2\pi} \frac{G(\varphi)}{R^2 + r^2 - 2rR \cos(\theta - \varphi)} d\varphi. \quad (3.30)$$

Going back to Cartesian coordinates¹⁰, we obtain Poisson's formula (3.22). Corollary 3.8 p. 125, assures that (3.30) is indeed the unique solution of the Dirichlet problem (3.21).

Case $G \in C([0, 2\pi])$. Even with G only continuous, formula (3.30) makes perfect sense and defines a harmonic function in B_R . It can be shown that¹¹

$$\lim_{(r, \theta) \rightarrow (R, \xi)} U(r, \theta) = G(\xi), \quad \forall \xi \in [0, 2\pi].$$

Therefore (3.30) is the unique solution to (3.21), once more by Corollary 3.8. \square

¹⁰ With $\boldsymbol{\sigma} = R(\cos \varphi, \sin \varphi)$, $d\sigma = R d\varphi$ and

$$\begin{aligned} |\mathbf{x} - \boldsymbol{\sigma}|^2 &= (r \cos \theta - R \cos \varphi)^2 + (r \sin \theta - R \sin \varphi)^2 \\ &= R^2 + r^2 - 2Rr (\cos \varphi \cos \theta + \sin \varphi \sin \theta) \\ &= R^2 + r^2 - 2Rr \cos(\theta - \varphi). \end{aligned}$$

¹¹ See Problem 3.20.

- *Poisson's formula in dimension $n > 2$.* Theorem 3.12 has an appropriate extension in any number of dimensions. When $B_R = B_R(\mathbf{p})$ is an n -dimensional ball, the solution of the Dirichlet problem (3.21) is given by (see Sect. 3.7.3)

$$u(\mathbf{x}) = \frac{R^2 - |\mathbf{x} - \mathbf{p}|^2}{\omega_n R} \int_{\partial B_R(\mathbf{p})} \frac{g(\boldsymbol{\sigma})}{|\mathbf{x} - \boldsymbol{\sigma}|^n} d\sigma. \quad (3.31)$$

We are now in position to prove Theorem 3.5, p. 123, the converse of the mean value property (*m.v.p.*).

Proof of Theorem 3.5. First observe that if two functions satisfy the *mean value property*, (*m.v.p.*) in a domain Ω , their difference satisfies this property as well. Assume now that $u \in C(\Omega)$ satisfies the *m.v.p.* and consider a ball $B \subset\subset \Omega$. We want to show that u is harmonic and infinitely differentiable in Ω . Denote by v the solution of the Dirichlet problem

$$\begin{cases} \Delta v = 0 & \text{in } B \\ v = u & \text{on } \partial B. \end{cases}$$

We know that

$$v \in C^\infty(B) \cap C(\overline{B})$$

and, being harmonic, it satisfies the *m.v.p.* in B . Then, also $w = v - u$ satisfies the *m.v.p.* in B and therefore by Theorem 3.7 it attains its maximum and minimum on ∂B . Since $w = 0$ on ∂B , we conclude that $u = v$ in B . Since B is arbitrary, $u \in C^\infty(\Omega)$ and it is harmonic in Ω . \square

3.3.6 Harnack's inequality and Liouville's theorem

From the mean value and Poisson's formulas we deduce another maximum principle, known as *Harnack's inequality*:

Theorem 3.13. *Let u be harmonic and nonnegative in a domain $\Omega \subset \mathbb{R}^n$. Assume that $B_R(\mathbf{z}) \subset\subset \Omega$. Then, for any $\mathbf{x} \in \overline{B}_R(\mathbf{z})$,*

$$\frac{R^{n-2}(R-r)}{(R+r)^{n-1}} u(\mathbf{z}) \leq u(\mathbf{x}) \leq \frac{R^{n-2}(R+r)}{(R-r)^{n-1}} u(\mathbf{z}), \quad (3.32)$$

where $r = |\mathbf{x} - \mathbf{z}|$. As a consequence, for every compact set $K \subset \Omega$, there exists a constant C , depending only on n and the distance of K from $\partial\Omega$, such that

$$\max_K u \leq C \min_K u. \quad (3.33)$$

Proof. We can assume that $\mathbf{z} = \mathbf{0}$. From Poisson's formula (3.31):

$$u(\mathbf{x}) = \frac{R^2 - |\mathbf{x}|^2}{\omega_n R} \int_{\partial B_R} \frac{u(\boldsymbol{\sigma})}{|\boldsymbol{\sigma} - \mathbf{x}|^n} d\sigma.$$

Observe that $R - |\mathbf{x}| \leq |\boldsymbol{\sigma} - \mathbf{x}| \leq R + |\mathbf{x}|$ and $R^2 - |\mathbf{x}|^2 = (R - |\mathbf{x}|)(R + |\mathbf{x}|)$. Then, by the mean value property,

$$u(\mathbf{x}) \leq \frac{(R + |\mathbf{x}|)}{(R - |\mathbf{x}|)^{n-1}} \frac{1}{\omega_n R} \int_{\partial B_R} u(\boldsymbol{\sigma}) d\sigma = \frac{R^{n-2}(R + |\mathbf{x}|)}{(R - |\mathbf{x}|)^{n-1}} u(\mathbf{0}).$$

Analogously,

$$u(\mathbf{x}) \geq \frac{R(R - |\mathbf{x}|)}{(R + |\mathbf{x}|)^{n-1}} \frac{1}{\omega_n R^2} \int_{\partial B_R} u(\boldsymbol{\sigma}) d\sigma = \frac{R^{n-2} (R - |\mathbf{x}|)}{(R + |\mathbf{x}|)^{n-1}} u(\mathbf{0}).$$

b) Let $d = \text{dist}(K, \partial\Omega)$. Let \mathbf{z}_m and \mathbf{z}_M be a minimum and a maximum point for u in K , respectively. Then, we can construct a sequence of balls $B_{d/3}(\mathbf{x}_j)$, $j = 0, \dots, N$, where $N = N(d, n)$, such that (Fig. 3.2):

- $\mathbf{x}_j \in B(\mathbf{x}_{j-1}), j = 1, \dots, N.$
- $x_0 = \mathbf{z}_m, x_N = \mathbf{z}_M.$

A repeated use of inequality (3.32) in each of the balls gives (3.33). We leave the details to the reader. \square

Harnack's inequality has important consequences. We present two of them. The first one says that the only harmonic functions in \mathbb{R}^n bounded from below or above are the constant functions.

Corollary 3.14 (Liouville's Theorem). *If u is harmonic in \mathbb{R}^n and $u(\mathbf{x}) \geq M$, then u is constant.*

Proof. The function $w = u - M$ is harmonic in \mathbb{R}^n and nonnegative. Fix $\mathbf{x} \in \mathbb{R}^n$ and choose $R > |\mathbf{x}|$; Harnack's inequality gives

$$\frac{R^{n-2} (R - |\mathbf{x}|)}{(R + |\mathbf{x}|)^{n-1}} w(\mathbf{0}) \leq w(\mathbf{x}) \leq \frac{R^{n-2} (R + |\mathbf{x}|)}{(R - |\mathbf{x}|)^{n-1}} w(\mathbf{0}). \quad (3.34)$$

Letting $R \rightarrow \infty$ in (3.34) we get $w(\mathbf{0}) \leq w(\mathbf{x}) \leq w(\mathbf{0})$, whence $w(\mathbf{0}) = w(\mathbf{x})$. Since \mathbf{x} is arbitrary we conclude that w , and therefore also u , is constant. \square

The second consequence concerns monotone sequences of harmonic functions.

Corollary 3.15. *Let $\Omega \subset \mathbb{R}^n$ be a bounded domain and $\{u_k\}_{k \geq 1}$ be a sequence of harmonic functions in Ω . Assume that:*

- i) *For every $k \geq 1$, $u_k \leq u_{k+1}$ in Ω .*
- ii) *$u_k(\mathbf{x})$ converges at some point $\mathbf{x}_0 \in \Omega$.*

Then $\{u_k\}_{k \geq 1}$ converges uniformly in every compact set $K \subset \Omega$ to a harmonic function u .

Proof. Let $u_k - u_j$, $k > j$. Then $u_k - u_j$ is harmonic and nonnegative in Ω . Let $K \subset \Omega$ be a compact set. We may always assume that $\mathbf{x}_0 \in K$. By (3.33), we can write

$$u_k(\mathbf{x}) - u_j(\mathbf{x}) \leq C(u_k(\mathbf{x}_0) - u_j(\mathbf{x}_0)), \quad \forall \mathbf{x} \in K.$$

Hence $\{u_k\}$ converges uniformly in K to a function u , which is harmonic, by Theorem 3.6, p. 123. \square

3.3.7 Analyticity of harmonic functions

Another important consequence of Poisson's formula is the possibility to control the derivatives of any order of a harmonic function u at a point \mathbf{x} , by the maximum of u in a small ball centered at \mathbf{x} . Let u be a harmonic function in a domain Ω and $B_r(\mathbf{x}) \subset\subset \Omega$. Since $u \in C^\infty(\Omega)$, we can differentiate the equation $\Delta u = 0$ any number of times and conclude that any derivative of u is still a harmonic function in Ω .

Let us start with an estimate of a first order derivative u_{x_j} . By the Mean Value Theorem, we can write

$$u_{x_j}(\mathbf{x}) = \frac{n}{\omega_n r^n} \int_{B_r(\mathbf{x})} u_{x_j}(\mathbf{y}) d\mathbf{y} = \frac{n}{\omega_n r^n} \int_{\partial B_r(\mathbf{x})} u(\boldsymbol{\sigma}) \nu_j d\sigma$$

where $\boldsymbol{\nu} = (\nu_1, \dots, \nu_n) = \frac{\boldsymbol{\sigma}-\mathbf{p}}{r}$, is the exterior normal to $B_r(\mathbf{p})$. Hence

$$|u_{x_j}(\mathbf{x})| \leq \frac{n}{r} \max_{\partial B_r(\mathbf{x})} |u|. \quad (3.35)$$

To deal with derivatives of any order, it is convenient to introduce a flexible notation. We call *multi-index* an *n-tuple* $\alpha = (\alpha_1, \dots, \alpha_n)$, where each α_j is a non-negative integer. Define *length* and *factorial* of α by

$$|\alpha| = \alpha_1 + \cdots + \alpha_n \quad \text{and} \quad \alpha! = \alpha_1! \alpha_2! \cdots \alpha_n!$$

respectively. Finally, we set

$$\mathbf{x}^\alpha = x_1^{\alpha_1} x_2^{\alpha_2} \cdots x_n^{\alpha_n} \quad \text{and} \quad D^\alpha = \frac{\partial^{\alpha_1}}{\partial x_1^{\alpha_1}} \frac{\partial^{\alpha_2}}{\partial x_2^{\alpha_2}} \cdots \frac{\partial^{\alpha_n}}{\partial x_n^{\alpha_n}}.$$

If $|\alpha| = 0$, we set $D^\alpha u = u$.

Lemma 3.16. *Let u be harmonic in $\Omega \subseteq \mathbb{R}^n$. Then for any ball $B_r(\mathbf{x}) \subset\subset \Omega$ and any multi-index α ,*

$$|D^\alpha u(\mathbf{x})| \leq \frac{n^k k! e^{k-1}}{r^k} \max_{\partial B_r(\mathbf{x})} |u|, \quad k = |\alpha|. \quad (3.36)$$

Proof. Let $M = \max_{\partial B_r(\mathbf{x})} |u|$. We proceed by induction on k . The step $k = 1$ is provided by (3.35). Assume now that (3.36) holds for $|\alpha| = k$ and any ball $B_\rho(\mathbf{x}) \subset\subset \Omega$. Let $D^\beta u = \partial_{x_j} D^\alpha u$ be a derivative of order $k+1$. Applying step $k=1$, we write

$$|D^\beta u(\mathbf{x})| \leq \frac{n}{\rho} \max_{\partial B_\rho(\mathbf{x})} |D^\alpha u|, \quad k = |\alpha|, \quad (3.37)$$

as long as $B_\rho(\mathbf{x}) \subset\subset \Omega$. On the other hand, the induction hypothesis gives, for any point $\mathbf{y} \in \partial B_\rho(\mathbf{x})$,

$$|D^\alpha u(\mathbf{y})| \leq \frac{n^k k! e^{k-1}}{\delta^k} \max_{\partial B_\delta(\mathbf{y})} |u|, \quad k = |\alpha|, \quad (3.38)$$

as long as $\partial B_\delta(\mathbf{y}) \subset \subset \Omega$. To get an estimate in terms of M , we have to choose ρ and δ such that $\partial B_\delta(\mathbf{y}) \subset \overline{B_r}(\mathbf{x})$ for any $\mathbf{y} \in \partial B_\delta(\mathbf{x})$, that is $\delta + \rho \leq r$. The most convenient choice is

$$\rho = \frac{r}{k+1}, \quad \delta = r - \frac{r}{k+1} = \frac{kr}{k+1}.$$

Inserting into (3.37) and (3.38) we get, using the elementary inequality $(1 + \frac{1}{k})^k < e$,

$$\begin{aligned} |D^\beta u(\mathbf{x})| &\leq \frac{(k+1)n}{r} \frac{n^k k! e^{k-1}}{r^k \left(\frac{k}{k+1}\right)^k} M \\ &= \frac{n^{k+1} (k+1)! e^{k-1} \left(1 + \frac{1}{k}\right)^k}{r^{k+1}} < \frac{n^{k+1} (k+1)! e^k}{r^{k+1}} \end{aligned}$$

which is (3.36) for $k+1$. \square

We have seen that if u is harmonic in a domain Ω , then $u \in C^\infty(\Omega)$. On the other hand, in dimension $n = 2$, we know that u is analytic, being the real part of an holomorphic function. Thanks to the estimates (3.36), the same property holds in any dimension. We have:

Theorem 3.17. *Let u be harmonic in Ω . Then u is (real) analytic, that is, for any $\mathbf{x} \in \Omega$ there exists $\rho = \rho(n, \text{dist}(\mathbf{x}, \partial\Omega))$ such that*

$$u(\mathbf{z}) = \sum_{k=0}^{\infty} \frac{D^\alpha u(\mathbf{x})}{\alpha!} (\mathbf{z} - \mathbf{x})^\alpha$$

in $\overline{B_\rho}(\mathbf{x})$.

Proof. Let $\mathbf{z} - \mathbf{x} = \rho \mathbf{h}$, where \mathbf{h} is a unit vector. Consider the Taylor series for u at \mathbf{x} , given by

$$\sum_{|\alpha|=0}^{\infty} \frac{D^\alpha u(\mathbf{x})}{\alpha!} (\mathbf{z} - \mathbf{x})^\alpha = \sum_{k=0}^{\infty} \rho^k \sum_{|\alpha|=k} \frac{D^\alpha u(\mathbf{x})}{\alpha!} \mathbf{h}^\alpha. \quad (3.39)$$

From (3.36) we have, if $|\alpha| = k$,

$$|D^\alpha u(\mathbf{x})| \leq k! \left(\frac{ne}{r}\right)^k \max_{\partial B_r(\mathbf{x})} |u|$$

as long as $r < \text{dist}(\mathbf{x}, \partial\Omega)$. Thus, we deduce ($|\mathbf{h}| = 1$) that

$$\rho^k \sum_{|\alpha|=k} \frac{|D^\alpha u(\mathbf{x})|}{\alpha!} \leq \max_{\partial B_r(\mathbf{x})} |u| \sum_{|\alpha|=k} \frac{k!}{\alpha!} \left(\frac{ne}{r}\rho\right)^k.$$

Choosing, say, $\rho = r/(2n^2e)$ and recalling the formula

$$n^k = \sum_{|\alpha|=k} \frac{k!}{\alpha!},$$

we get

$$\sum_{|\alpha|=k} \frac{k!}{\alpha!} \left(\frac{ne}{r}\rho\right)^k = \frac{1}{2^k}.$$

By the Weierstrass test, the series (3.39) converges uniformly in $\overline{B_\rho}(\mathbf{x})$ and therefore

$$u(\mathbf{x}) = \sum_{|\alpha|=0}^{\infty} \frac{D^\alpha u(\mathbf{x})}{\alpha!} (\mathbf{z} - \mathbf{x})^\alpha$$

in $\overline{B_\rho}(\mathbf{x})$. □

3.4 A probabilistic solution of the Dirichlet problem

In Sect. 3.1 we solved the discrete Dirichlet problem via a probabilistic method. The key ingredients in the construction of the solution, leading to formula (3.12), were the mean value property and the absence of memory of the random walk (each step is independent of the preceding ones). In the continuous case the appropriate versions of those tools are available, with the Markov property encoding the absence of memory of Brownian motion¹². Thus it is reasonable that a suitable continuous version of formula (3.12) should give the solution of the Dirichlet problem for the Laplace operator. As before, we work in dimension $n = 2$, but methods and conclusions can be extended without much effort to any number of dimensions.

Let $\Omega \subset \mathbb{R}^2$ be a bounded domain and $g \in C(\partial\Omega)$. We want to derive a representation formula for the unique solution $u \in C^2(\Omega) \cap C(\overline{\Omega})$ of the problem

$$\begin{cases} \Delta u = 0 & \text{in } \Omega \\ u = g & \text{on } \partial\Omega. \end{cases} \quad (3.40)$$

Let $\mathbf{X}(t)$ be the position of a Brownian particle started at $\mathbf{x} \in \Omega$ and define *the first exit time from Ω* , $\tau = \tau(\mathbf{x})$, as follows (Fig. 3.4):

$$\tau(\mathbf{x}) = \left\{ \inf_{t \geq 0} t : \mathbf{X}(t) \in \mathbb{R}^2 \setminus \Omega \right\}.$$

The time τ is a *stopping time*: to decide whether the event $\{\tau \leq t\}$ occurs or not, *it suffices to observe the process until time t* . In fact, for fixed $t \geq 0$, to decide whether or not $\tau \leq t$ is true, it is enough to consider the event

$$E = \{\mathbf{X}(s) \in \Omega, \text{ for all times } s \text{ from 0 until } t, t \text{ included}\}.$$

If this event occurs, then it must be that $\tau > t$. If E does not occur, it means that there are points $\mathbf{X}(s)$ outside Ω for some $s \leq t$ and therefore it must be that $\tau \leq t$.

The first thing we have to check is that the particle leaves Ω in finite time, almost surely. Precisely:

¹² See Sect. 2.6.

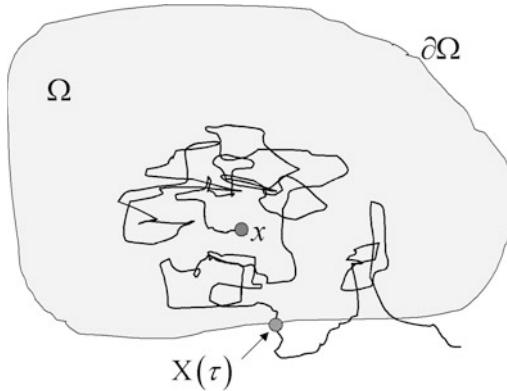


Fig. 3.4 First exit point from Ω

Lemma 3.18. For every $\mathbf{x} \in \Omega$, $\tau(\mathbf{x})$ is finite with probability 1, that is:

$$P\{\tau(\mathbf{x}) < \infty\} = 1.$$

Proof. It is enough to show that our particle remains inside any circle $B_r = B_r(\mathbf{x}) \subset \Omega$, with zero probability. If we denote by τ_r the first exit time from B_r , we have to prove that $P\{\tau_r = \infty\} = 0$.

Suppose $\mathbf{X}(t) \in B_r$ until $t = k$ (k integer). Then, for $j = 1, 2, \dots, k$, it must be that

$$|\mathbf{X}(j) - \mathbf{X}(j-1)| < 2r.$$

Thus (the occurrence of) the event $\{\tau_r > k\}$ implies (the occurrence of) all the events

$$E_j = \{|\mathbf{X}(j) - \mathbf{X}(j-1)| < 2r\}, \quad j = 1, 2, \dots, k,$$

and therefore also of their intersection. As a consequence

$$P\{\tau_r > k\} \leq P\{\cap_{j=1}^k E_j\}. \quad (3.41)$$

On the other hand, the increments $\mathbf{X}(j) - \mathbf{X}(j-1)$ are mutually independent and equidistributed according to a standard normal law, hence we can write

$$P\{E_j\} = \frac{1}{2\pi} \int_{\{|\mathbf{z}| < 2r\}} \exp\left(-\frac{|\mathbf{z}|^2}{2}\right) d\mathbf{z} \equiv \gamma < 1$$

and

$$P\{\cap_{j=1}^k E_j\} = \prod_{j=1}^k P\{E_j\} = \gamma^k. \quad (3.42)$$

Since $\{\tau_r = \infty\}$ implies $\{\tau_r > k\}$, from (3.41) and (3.42) we have

$$P\{\tau_r = \infty\} \leq P\{\tau_r > k\} \leq \gamma^k.$$

Letting $k \rightarrow +\infty$ we get $P\{\tau_r = \infty\} = 0$. □

Lemma 3.18 implies that $\mathbf{X}(t)$ hits the boundary $\partial\Omega$ in finite time $\tau = \tau(\mathbf{x})$, with probability 1. We can therefore introduce on $\partial\Omega$ a probability distribution associated with the random variable $\mathbf{X}(\tau)$ by setting

$$P(\mathbf{x}, \tau(\mathbf{x}), F) = P\{\mathbf{X}(\tau) \in F\} \quad (\tau = \tau(\mathbf{x})),$$

for every “reasonable” subset $F \subset \partial\Omega$. For fixed \mathbf{x} in Ω , the set function

$$F \longmapsto P(\mathbf{x}, \tau(\mathbf{x}), F)$$

defines a probability measure on $\partial\Omega$, since $P(\mathbf{x}, \tau(\mathbf{x}), \partial\Omega) = 1$, according to Lemma 3.18. $P(\mathbf{x}, \tau(\mathbf{x}), F)$ is called the *escape probability from Ω , through F* ¹³.

By analogy with formula (3.12), we can now guess the type of formula we expect for the solution u of problem (3.40). To get the value $u(\mathbf{x})$, let a Brownian particle start from \mathbf{x} , and let $\mathbf{X}(\tau) \in \partial\Omega$ its first exit point from Ω . Then, compute the random “gain” $g(\mathbf{X}(\tau))$ and take its expected value with respect to the distribution $P(\mathbf{x}, \tau(\mathbf{x}), \cdot)$. This is $u(\mathbf{x})$. Everything works if $\partial\Omega$ is not too bad. Precisely, we have:

Theorem 3.19. *Let Ω be a bounded Lipschitz domain and $g \in C(\partial\Omega)$. The unique solution $u \in C^2(\Omega) \cap C(\overline{\Omega})$ of problem (3.40) is given by*

$$u(\mathbf{x}) = E^{\mathbf{x}}[g(\mathbf{X}(\tau))] = \int_{\partial\Omega} g(\sigma) P(\mathbf{x}, \tau(\mathbf{x}), d\sigma). \quad (3.43)$$

Proof (sketch). For fixed $F \subseteq \partial\Omega$, consider the function

$$u_F: \mathbf{x} \longmapsto P(\mathbf{x}, \tau(\mathbf{x}), F).$$

We claim that u_F harmonic in Ω . Assuming that u_F is continuous¹⁴ in Ω , from Theorem 3.7, p. 124, it is enough to show that $P(\mathbf{x}, \tau(\mathbf{x}), F)$ satisfies the mean value property. Let

$$B_R = B_R(\mathbf{x}) \subset \subset \Omega.$$

If $\tau_R = \tau_R(\mathbf{x})$ is the first exit time from B_R , then $\mathbf{X}(\tau_R)$ has a uniform distribution on ∂B_R , due to the invariance by rotation of the Brownian motion.

This means that, starting from the center, the escape probability from B_R through any arc $K \subset \partial B_R$ is given by

$$\frac{\text{length of } K}{2\pi R}.$$

Now, before reaching F , the particle must hit ∂B_R . Since τ_R is a stopping time, we may use the strong Markov property. Thus, after τ_R , $\mathbf{X}(t)$ can be considered as a Brownian motion with uniform initial distribution on ∂B_R , expressed by the formula (Fig. 3.5)

$$\mu(ds) = \frac{ds}{2\pi R},$$

¹³ More precisely, $P(\mathbf{x}, \tau(\mathbf{x}), F)$ is well defined on the σ -algebra of the *Borel subsets set of $\partial\Omega$* and it can be shown that it is σ -additive (see Appendix B).

¹⁴ Which should be at least intuitively clear.

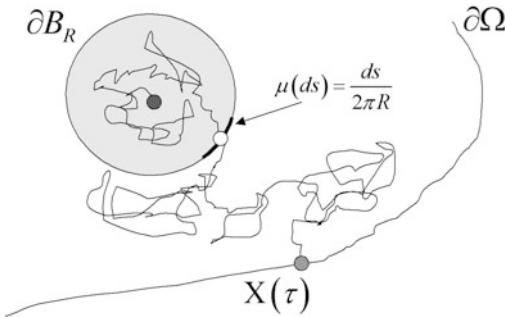


Fig. 3.5 Strong Markov property of a Brownian particle

where ds is the length element on ∂B_R . Therefore, the particle escapes ∂B_R through some arc of length ds , centered at a point s , and from there it reaches F with probability $P(s, \tau(s), F)\mu(ds)$. By integrating this probability on ∂B_R , we obtain $P(x, \tau(x), F)$, namely:

$$P(x, \tau(x), F) = \int_{\partial B_R(x)} P(s, \tau(s), F)\mu(ds) = \frac{1}{2\pi R} \int_{\partial B_R(x)} P(s, \tau(s), F)ds$$

which is the mean value property for $P(x, \tau(x), F)$.

Observe now that if $\sigma \in \partial\Omega$ then $\tau(\sigma) = 0$ and hence

$$P(\sigma, \tau(\sigma), F) = \begin{cases} 1 & \text{if } \sigma \in F \\ 0 & \text{if } \sigma \in \partial\Omega \setminus F. \end{cases}$$

Therefore, if $\sigma \in \partial\Omega$, $P(\sigma, \tau(\sigma), F)$ coincides with the characteristic function of F . Thus, if $d\sigma$ is an arc element on $\partial\Omega$ centered in σ , the function

$$x \mapsto g(\sigma) P(x, \tau(x), d\sigma) \tag{3.44}$$

is harmonic in Ω , it attains the value $g(\sigma)$ on $d\sigma$ and it is zero on $\partial\Omega \setminus d\sigma$. To obtain the harmonic function equal to g on $\partial\Omega$, we integrate over $\partial\Omega$ all the contributions from (3.44). This gives the representation (3.43).

Rigorously, to assert that (3.43) is indeed the required solution, we should check that $u(x) \rightarrow g(\sigma)$ when $x \rightarrow \sigma$. It can be proved¹⁵ that this is true if Ω is, for instance, a Lipschitz domain¹⁶. \square

Remark 3.20. The measure

$$F \mapsto P(x, \tau(x), F)$$

is called the *harmonic measure at x of the domain Ω* and in general it can not be expressed by an explicit formula. In the particular case $\Omega = B_R(p)$, Poisson's formula (3.22) indicates that the harmonic measure for the disc $B_R(p)$ is given

¹⁵ The proof is rather delicate. See [47], Øksendal, 1995.

¹⁶ We will be back on the boundary behavior of harmonic functions in the next section.

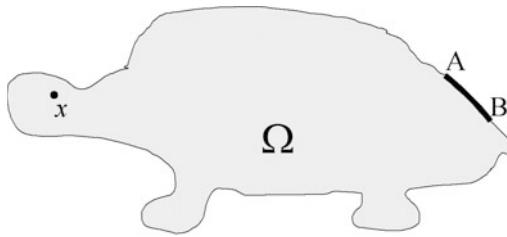


Fig. 3.6 A modification of the Dirichlet data on the arc AB affects the value of the solution at \mathbf{x}

by

$$P(\mathbf{x}, \tau(\mathbf{x}), d\sigma) = \frac{1}{2\pi R} \frac{R^2 - |\mathbf{x} - \mathbf{p}|^2}{|\sigma - \mathbf{x}|^2} d\sigma.$$

Remark 3.21. Formula (3.43) shows that the value of the solution at a point \mathbf{x} depends on the boundary data on all $\partial\Omega$ (except for sets of length zero). In the case of Fig. 3.6, a change of the datum g on the arc AB affects the value of the solution at \mathbf{x} , even if this point is far from AB and near $\partial\Omega$.

Recurrence and Brownian motion

We have seen that the solution of a Dirichlet problem can be constructed by using the general properties of a Brownian motion. On the other hand, the deterministic solution of some Dirichlet problems can be used to deduce interesting properties of the Brownian motion.

We examine two simple examples. Recall from Sect. 3.1 that $\ln|\mathbf{x}|$ is harmonic in the plane except $\mathbf{x} = \mathbf{0}$.

Let a, R be real numbers, $R > a > 0$. It is easy to check that the function

$$u_R(\mathbf{x}) = \frac{\ln|R| - \ln|\mathbf{x}|}{\ln R - \ln a}$$

is harmonic in the ring $C_{a,R} = \{\mathbf{x} \in \mathbb{R}^2; a < |\mathbf{x}| < R\}$ and moreover

$$u_R(\mathbf{x}) = 1 \text{ on } \partial B_a(\mathbf{0}), \quad u_R(\mathbf{x}) = 0 \text{ on } \partial B_R(\mathbf{0}).$$

Thus $u_R(\mathbf{x})$ represents the escape probability from the ring through $\partial B_a(\mathbf{0})$, starting at \mathbf{x} :

$$u_R(\mathbf{x}) = P_R(\mathbf{x}, \tau(\mathbf{x}), \partial B_a(\mathbf{0})).$$

Letting $R \rightarrow +\infty$, we get

$$P_R(\mathbf{x}, \tau(\mathbf{x}), \partial B_a(\mathbf{0})) = \frac{\ln|R| - \ln|\mathbf{x}|}{\ln R - \ln a} \rightarrow 1 = P_\infty(\mathbf{x}, \tau(\mathbf{x}), \partial B_a(\mathbf{0})).$$

This means that, starting at \mathbf{x} , the probability to *enter (sooner or later)* the disc $B_a(\mathbf{0})$ is 1. Due to the invariance by translations of the Brownian motion, the origin can be replaced by any other point without changing the conclusions. Moreover, since we have proved in Lemma 3.18 that the exit probability from any disc is also 1, we can state the following result: *given any point \mathbf{x} and any disc in the plane, a Brownian particle started at \mathbf{x} enters the disc and exits from it an infinite number of times, with probability 1.* We say that a bidimensional Brownian motion is **recurrent**.

In three dimensions a Brownian motion is *not* recurrent. In fact (see the next section), the function

$$u_R(\mathbf{x}) = \frac{\frac{1}{|\mathbf{x}|} - \frac{1}{R}}{\frac{1}{a} - \frac{1}{R}}$$

is harmonic in the spherical shell

$$S_{a,R} = \{\mathbf{x} \in \mathbb{R}^3; a < |\mathbf{x}| < R\}$$

and

$$\begin{aligned} u_R(\mathbf{x}) &= 1 \text{ on } \partial B_a(\mathbf{0}), \\ u_R(\mathbf{x}) &= 0 \text{ on } \partial B_R(\mathbf{0}). \end{aligned}$$

Then $u_R(\mathbf{x})$ represents the escape probability from the shell through $\partial B_a(\mathbf{0})$, starting at \mathbf{x} :

$$u_R(\mathbf{x}) = P_R(\mathbf{x}, \tau(\mathbf{x}), \partial B_a(\mathbf{0})).$$

This time, letting $R \rightarrow +\infty$, we find

$$P_R(\mathbf{x}, \tau(\mathbf{x}), \partial B_a(\mathbf{0})) = \frac{\frac{1}{|\mathbf{x}|} - \frac{1}{R}}{\frac{1}{a} - \frac{1}{R}} \rightarrow \frac{a}{|\mathbf{x}|} = P_\infty(\mathbf{x}, \tau(\mathbf{x}), \partial B_a(\mathbf{0})).$$

Thus, the probability to enter, sooner or later, the sphere $B_a(\mathbf{0})$ is not 1 and it becomes smaller and smaller as the distance of \mathbf{x} from the origin increases.

3.5 Sub/Superharmonic Functions. The Perron Method

3.5.1 Sub/superharmonic functions

We present in this section a method due to O. Perron for constructing the solution of the Dirichlet problem

$$\begin{cases} \Delta u = 0 & \text{in } \Omega \\ u = g & \text{on } \partial\Omega \end{cases} \quad (3.45)$$

where $\Omega \subset \mathbb{R}^n$ is a **bounded** domain and $g \in C(\partial\Omega)$.

One of the peculiarity of Perron's method lies in its flexibility, and, indeed, it works with much more general operators than the Laplacian and, in particular, with fully nonlinear elliptic operators, of great importance in several applied areas.

We need to introduce the class of sub/super harmonic functions and establish some of their properties. For a function $u \in C(\Omega)$, $\Omega \subseteq \mathbb{R}^n$, denote its volume average over a ball $B_r(\mathbf{x}) \subset\subset \Omega$ by

$$A(u; \mathbf{x}, r) = \frac{n}{\omega_n r^n} \int_{B_r(\mathbf{x})} u(\mathbf{y}) d\mathbf{y}$$

and its spherical average by

$$S(u; \mathbf{x}, r) = \frac{1}{\omega_n r^{n-1}} \int_{\partial B_r(\mathbf{x})} u(\boldsymbol{\sigma}) d\sigma.$$

Definition 3.22. A function $u \in C(\Omega)$ is subharmonic in Ω if

$$u(\mathbf{x}) \leq S(u; \mathbf{x}, r) \quad (3.46)$$

for every $\mathbf{x} \in \Omega$ and every $B_r(\mathbf{x}) \subset\subset \Omega$.

In (3.46), it is equivalent to use volume averages (see Problem 3.8 a)). Superharmonic functions are defined reversing the inequality in (3.46).

Typical subharmonic functions are convex cones or paraboloids (e.g. $u(\mathbf{x}) = |\mathbf{x} - \mathbf{p}|^\alpha$, $\alpha \geq 1$). If $u \in C^2(\Omega)$, then u is subharmonic in Ω if and only if $\Delta u \geq 0$ (see Problem 3.8 d)). The properties we shall need are the following ones.

1. Maximum principle. If $u \in C(\overline{\Omega})$ is subharmonic (superharmonic) and non-constant, then

$$u(\mathbf{x}) < \max_{\partial\Omega} u, \quad (u(\mathbf{x}) > \min_{\partial\Omega} u) \quad \text{for every } \mathbf{x} \in \Omega.$$

2. If u_1, u_2, \dots, u_N are subharmonic (superharmonic) in Ω , then

$$u = \max \{u_1, u_2, \dots, u_N\} \quad (u = \min \{u_1, u_2, \dots, u_N\})$$

is subharmonic (superharmonic) in Ω .

3. Let u be subharmonic (superharmonic) in Ω . For a ball $B \subset\subset \Omega$, let $P(u; B)$ be the harmonic function in B , with $P(u; B) = u$ on ∂B . $P(u; B)$ is called the harmonic lifting of u in B . Then, the function

$$u^B = \begin{cases} P(u, B) & \text{in } B \\ u & \text{in } \Omega \setminus B \end{cases}$$

is subharmonic (superharmonic) in Ω .

Proof. 1. It follows repeating verbatim the proof of Theorem 3.7, p. 124.

2. For each $j = 1, \dots, N$, and every ball $B_r(\mathbf{x}) \subset\subset \Omega$, we have

$$u_j(\mathbf{x}) \leq S(u_j; \mathbf{x}, r) \leq S(u; \mathbf{x}, r).$$

Thus

$$u(\mathbf{x}) \leq S(u; \mathbf{x}, r)$$

and u is subharmonic.

3. By the maximum principle, $u \leq P(u, B)$. Let $B_r(\mathbf{x}) \subset\subset \Omega$. If $B_r(\mathbf{x}) \cap B = \emptyset$ or $B_r(\mathbf{x}) \subset B$ there is nothing to prove. In the other case, let w be the harmonic lifting of u^B in $B_r(\mathbf{x})$. Then, by maximum principle,

$$u^B(\mathbf{x}) \leq w(\mathbf{x}) = S(w; \mathbf{x}, r) = S(u^B; \mathbf{x}, r)$$

and therefore u^B is subharmonic. □

3.5.2 The method

We now go back to the Dirichlet problem. The idea is to construct the solution by looking at the class of subharmonic functions in Ω , smaller than g on the boundary, and to take their supremum. It is like constructing a line segment l by taking the supremum of all convex parabolas below l . Thus, let

$$S_g = \{v \in C(\overline{\Omega}) : v \text{ subharmonic in } \Omega, v \leq g \text{ on } \partial\Omega\}.$$

Our candidate solution is:

$$u_g(\mathbf{x}) = \sup\{v(\mathbf{x}) : v \in S_g\}, \quad \mathbf{x} \in \overline{\Omega}.$$

Note that:

- S_g is nonempty, since

$$v(\mathbf{x}) \equiv \min_{\partial\Omega} g \in S_g.$$

- The function u_g is well defined since, by maximum principle, $u_g \leq \max_{\partial\Omega} g$.

We have to check two things: first that u_g is harmonic in Ω , second that u_g assumes continuously the boundary data. Let us show that u_g is harmonic.

Theorem 3.23. u_g is harmonic in Ω .

Proof. Given $\mathbf{x}_0 \in \Omega$, by the definition of u_g , there exists $\{u_k\} \subset S_g$ such that $u_k(\mathbf{x}_0) \rightarrow u_g(\mathbf{x}_0)$. The functions

$$w_k = \max\{u_1, u_2, \dots, u_k\}, \quad k \geq 1,$$

belong to S_g (by property 2 above) and $w_k \leq w_{k+1}$. Moreover, since $u_k(\mathbf{x}_0) \leq w_k(\mathbf{x}_0) \leq w_g(\mathbf{x}_0)$, we infer that

$$\lim_{k \rightarrow +\infty} w_k(\mathbf{x}_0) = u_g(\mathbf{x}_0).$$

Let B be a ball such that $\mathbf{x}_0 \in B$ and $\overline{B} \subset \Omega$. For each $k \geq 1$, we have $w_k \leq w_k^B \leq w_{k+1}^B$ and hence,

$$\lim_{k \rightarrow +\infty} w_k^B(\mathbf{x}_0) = u_g(\mathbf{x}_0).$$

By Corollary 3.15, p. 132,

$$\lim_{k \rightarrow +\infty} w_k^B(\mathbf{x}) = w(\mathbf{x})$$

uniformly in \overline{B} , where w is harmonic in B , and clearly $w(\mathbf{x}_0) = u_g(\mathbf{x}_0)$.

We show that $w = u_g$ in B . By the definition of u_g we have $w \leq u_g$. Suppose that there exists $\mathbf{x}_1 \in B$ such that

$$w(\mathbf{x}_1) < u_g(\mathbf{x}_1).$$

Let $\{v_k\} \subset S_g$ such that $v_k(\mathbf{x}_1) \rightarrow u_g(\mathbf{x}_1)$. Define, for $k \geq 1$,

$$z_k = \max \{v_1, v_2, \dots, v_k, w_k\}.$$

Then z_k^B belongs to S_g and $w_k \leq z_k^B \leq u_g$, $v_k \leq z_k^B \leq u_g$ in $\overline{\Omega}$. Thus

$$\lim_{k \rightarrow +\infty} z_k^B(\mathbf{x}_1) = u_g(\mathbf{x}_1) \text{ and } \lim_{k \rightarrow +\infty} z_k^B(\mathbf{x}_0) = u_g(\mathbf{x}_0).$$

By Corollary 3.15, the sequence $\{z_k^B\}$ converges in B to a harmonic function z with $z(\mathbf{x}_1) = u_g(\mathbf{x}_1)$. By construction $w \leq z$ in B and $w(\mathbf{x}_0) = z(\mathbf{x}_0) = u_g(\mathbf{x}_0)$.

Then the function $z - w$ is harmonic, nonnegative with an interior minimum equal to zero at \mathbf{x}_0 . The maximum principle yields $z - w \equiv 0$ in B that leads to the contradiction

$$w(\mathbf{x}_1) = z(\mathbf{x}_1) = u_g(\mathbf{x}_1) > w(\mathbf{x}_1).$$

Thus, $u_g = w$ in B and u_g is harmonic in B . Since \mathbf{x}_0 is arbitrary, we conclude that u_g is harmonic in Ω . \square

3.5.3 Boundary behavior

Through Theorem 3.23, we can associate to each $g \in C(\partial\Omega)$ the harmonic function u_g . We have to check if

$$u_g(\mathbf{x}) \rightarrow g(\mathbf{p})$$

as $\mathbf{x} \rightarrow \mathbf{p}$, for every $\mathbf{p} \in \partial\Omega$. This is not always true, as we shall see later on, and it depends on a particular smoothness condition on $\partial\Omega$. In order to control the behavior of u at a boundary point \mathbf{p} , we introduce the key notion of *barrier*.

Definition 3.24. Let $\mathbf{p} \in \partial\Omega$. We say that $h \in C(\overline{\Omega})$ is a barrier in Ω at \mathbf{p} if:

- i) h is superharmonic in Ω .
- ii) $h > 0$ in $\overline{\Omega} \setminus \{\mathbf{p}\}$ and $h(\mathbf{p}) = 0$.

Remark 3.25. The notion of barrier has a local nature. Indeed, let h_{loc} be a barrier in $B_r(\mathbf{p}) \cap \Omega$ at \mathbf{p} , for some ball $B_r(\mathbf{p})$. Set $C_r = \Omega \cap (B_r(\mathbf{p}) \setminus B_{r/2}(\mathbf{p}))$ and $m = \min_{\overline{C_r}} h_{loc}$. Then

$$h(\mathbf{x}) = \begin{cases} \min \{m, h_{loc}(\mathbf{x})\} & \text{in } \overline{B}_r(\mathbf{p}) \cap \overline{\Omega} \\ m & \text{in } \overline{\Omega} \setminus B_r(\mathbf{p}) \end{cases}$$

is a barrier in Ω at \mathbf{p} .

We say that $\mathbf{p} \in \partial\Omega$ is a *regular point* if there exists a barrier at \mathbf{p} . If every $\mathbf{p} \in \partial\Omega$ is regular, we say that Ω is regular (for the Dirichlet problem). The following theorem clarifies the role of a barrier.

Theorem 3.26. *Let $\mathbf{p} \in \partial\Omega$ be a regular point. Then, for every $g \in C(\partial\Omega)$, $u_g(\mathbf{x}) \rightarrow g(\mathbf{p})$ as $\mathbf{x} \rightarrow \mathbf{p}$.*

Proof. Let h be a barrier at \mathbf{p} and $g \in C(\partial\Omega)$. To show that $u_g(\mathbf{x}) \rightarrow g(\mathbf{p})$ as $\mathbf{x} \rightarrow \mathbf{p}$ it is enough to prove that, for every $\varepsilon > 0$, there exists $k_\varepsilon > 0$ such that

$$g(\mathbf{p}) - \varepsilon - k_\varepsilon h(\mathbf{x}) \leq u_g(\mathbf{x}) \leq g(\mathbf{p}) + \varepsilon + k_\varepsilon h(\mathbf{x}), \quad \forall \mathbf{x} \in \overline{\Omega}.$$

Put

$$z(\mathbf{x}) = g(\mathbf{p}) + \varepsilon + k_\varepsilon h(\mathbf{x}).$$

Then $z \in C(\overline{\Omega})$ and it is superharmonic in Ω . We want to choose $k_\varepsilon > 0$ so that $z \geq g$ on $\partial\Omega$.

By the continuity of g , there exists δ_ε such that, if $\mathbf{y} \in \partial\Omega$ and $|\mathbf{y} - \mathbf{p}| \leq \delta_\varepsilon$, then $|g(\mathbf{y}) - g(\mathbf{p})| \leq \varepsilon$. For these points \mathbf{y} , we have

$$z(\mathbf{y}) = g(\mathbf{p}) - g(\mathbf{y}) + \varepsilon + k_\varepsilon h(\mathbf{y}) + g(\mathbf{y}) \geq g(\mathbf{y})$$

since $h(\mathbf{y}) \geq 0$ in $\overline{\Omega}$. If $|\mathbf{y} - \mathbf{p}| > \delta_\varepsilon$, from ii) we have $h(\mathbf{y}) \geq m_{\delta_\varepsilon} > 0$ so that

$$z(\mathbf{y}) \geq g(\mathbf{y})$$

if $k_\varepsilon > \max |g| / m_{\delta_\varepsilon}$. The maximum principle gives

$$z \geq v, \text{ in } \overline{\Omega}, \quad \forall v \in S_g,$$

from which $z \geq u_g$ in $\overline{\Omega}$. A similar argument gives that $g(\mathbf{p}) - \varepsilon - k_\varepsilon h(\mathbf{x}) \leq u_g(\mathbf{x})$ in $\overline{\Omega}$. \square

The following consequence is immediate.

Corollary 3.27. *If Ω is regular for the Dirichlet problem, u_g is the unique solution of problem (3.45).*

We list below some sufficient conditions for a boundary point \mathbf{p} to be regular.

a) *Exterior sphere condition* (Fig. 3.7). There exists $B_r(\mathbf{x}_0) \subset \mathbb{R}^n \setminus \overline{\Omega}$ such that $\overline{B_r(\mathbf{x}_0)} \cap \partial\Omega = \{\mathbf{p}\}$. Then, a barrier at \mathbf{p} is given by

$$h(\mathbf{x}) = 1 - \frac{r^{n-2}}{|\mathbf{x} - \mathbf{x}_0|^{n-2}} \quad \text{for } n > 2$$

and

$$h(\mathbf{x}) = 1 - \log \frac{r}{|\mathbf{x} - \mathbf{x}_0|} \quad \text{for } n = 2.$$

b) Ω is *convex*. At every point $\mathbf{p} \in \partial\Omega$, there exists a so called *support hyperplane* $\pi(\mathbf{p})$ of equation $(\mathbf{x} - \mathbf{p}) \cdot \boldsymbol{\nu} = 0$, such that Ω is contained in the half

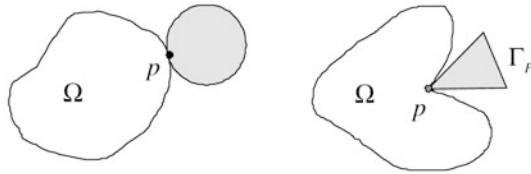


Fig. 3.7 Exterior sphere and exterior cone conditions at \mathbf{p}

space

$$S^+ = \{\mathbf{x} : (\mathbf{x} - \mathbf{p}) \cdot \boldsymbol{\nu} > 0\}$$

and $\pi(\mathbf{p}) \cap \partial\Omega = \{\mathbf{p}\}$. A barrier at \mathbf{p} is given by the affine linear function $h(\mathbf{x}) = (\mathbf{x} - \mathbf{p}) \cdot \boldsymbol{\nu}$.

c) *Exterior cone condition* (Fig. 3.7). There exists a closed cone $\Gamma_{\mathbf{p}}$ with vertex at \mathbf{p} and a ball $B_r(\mathbf{p})$ such that $\overline{B_r}(\mathbf{p}) \cap \Gamma_{\mathbf{p}} \subset \mathbb{R}^n \setminus \Omega$. A barrier at \mathbf{p} in $B_r(\mathbf{p}) \setminus \Omega$ is given by¹⁷ u_g where $g(\mathbf{x}) = |\mathbf{x} - \mathbf{p}|$ and it works also for $B_r(\mathbf{p}) \cap \Omega$, since $B_r(\mathbf{p}) \cap \Omega \subset B_r(\mathbf{p}) \setminus \Omega$. In particular, a bounded Lipschitz domain is regular.

Remark 3.28. In the construction of the solution, we could have started with the class

$$S^g = \{v \in C(\overline{\Omega}) : v \text{ superharmonic in } \Omega, v \geq g \text{ on } \partial\Omega\}$$

and define:

$$u^g(\mathbf{x}) = \inf\{v(\mathbf{x}) : v \in S^g\}, \quad \mathbf{x} \in \overline{\Omega}.$$

The same proof, with obvious changes, gives that u^g is harmonic in Ω . Clearly $u_g \leq u^g$ and they coincide if Ω is regular for the Dirichlet problem.

A natural question is to ask if nonregular points exist. As we have anticipated, the answer is yes. Here is a celebrated example due to Lebesgue.

Example 3.29. The Lebesgue spine ($n = 3$). Let $S_c = \{\rho < e^{-c/x_1}, x_1 > 0\}$, $0 < c \leq 1$, $\rho^2 = x_2^2 + x_3^2$ and $\Omega = B_1 \setminus \overline{S}_1$ (see Fig. 3.8). Then the origin is an inward exponential cusp and it is not a regular point.

Proof. Define

$$w(\mathbf{x}) = \int_0^1 \frac{t}{\sqrt{(x_1 - t)^2 + \rho^2}} dt.$$

We have:

1. w is harmonic¹⁸ in $\mathbb{R}^3 \setminus \{(x_1, 0, 0) : 0 \leq x_1 \leq 1\}$.
2. w is bounded in Ω .

¹⁷ It is not elementary to show it. This result is known as *Zaremba's Theorem*. We refer to [6], *Helms, 1969*.

¹⁸ We shall see in the next section that w is the Newtonian potential of density t on the segment $l = \{(x_1, 0, 0) : 0 \leq x_1 \leq 1\}$, which is harmonic outside l .

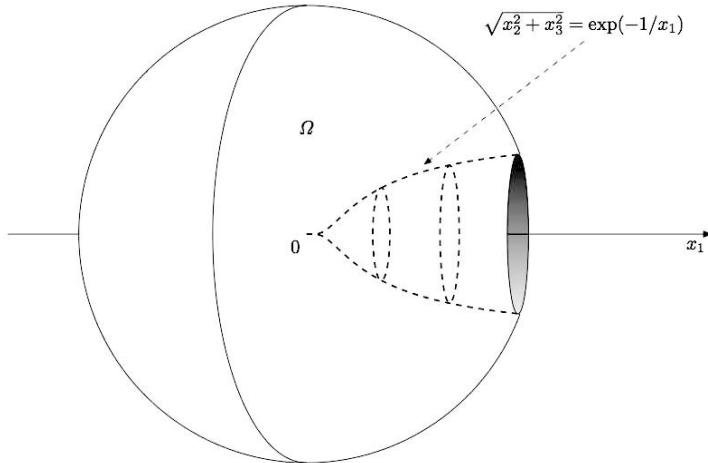


Fig. 3.8 The Lebesgue spine

To prove it, consider first $x_1 \leq 0$. Then $t/\sqrt{(x_1 - t)^2 + \rho^2} \leq t/\sqrt{t^2 + x_1^2 + \rho^2} \leq 1$ so that $|w| \leq 1$. If $0 < x_1 < 1$, write

$$w(\mathbf{x}) = \int_0^1 \frac{t - x_1}{\sqrt{(x_1 - t)^2 + \rho^2}} dt + x_1 \int_0^1 \frac{1}{\sqrt{(x_1 - t)^2 + \rho^2}} dt = A(x_1, \rho) + B(x_1, \rho).$$

Then, clearly, $|A(x_1, \rho)| \leq 1$. Moreover, since in Ω $\rho > e^{-1/x_1}$, we have

$$|B(x_1, \rho)| \leq x_1 \int_{\rho \leq |x_1 - t| \leq 1} \frac{dt}{|x_1 - t|} + \frac{x_1}{\rho} \int_{|x_1 - t| \leq \rho} dt \leq -x_1 2 \log \rho + \frac{x_1}{\rho} 2\rho \leq 4.$$

3. If $0 < c < 1$, the surface $\partial S_c \cap B_1$ is contained in Ω and we have

$$\lim_{\mathbf{x} \in \partial S_c, \mathbf{x} \rightarrow 0} w(\mathbf{x}) = 1 + 2c. \quad (3.47)$$

Indeed, the two terms $A(x_1, \rho)$ and $B(x_1, \rho)$ can be explicitly computed. Namely:

$$A(x_1, \rho) = \sqrt{(x_1 - 1)^2 + \rho^2} - \sqrt{x_1^2 + \rho^2}$$

$$B(x_1, \rho) = x_1 \log \left| \left(1 - x_1 + \sqrt{(x_1 - 1)^2 + \rho^2} \right) \left(x_1 + \sqrt{(x_1 - 1)^2 + \rho^2} \right) \right| - 2x_1 \log \rho.$$

Then (3.47) follows easily.

Now, for $c = 1$, the restriction $g = w|_{\partial\Omega}$ is well defined and continuous up to $\mathbf{x} = \mathbf{0}$. Consider u_g . Every point of $\partial\Omega \setminus \{\mathbf{p}\}$ is regular and therefore, if $\mathbf{p} \neq \mathbf{0}$,

$$\lim_{\mathbf{x} \rightarrow \mathbf{p}} (w(\mathbf{x}) - u_g(\mathbf{x})) = 0.$$

Being $u - u_g$ bounded, we conclude that $w - u_g \equiv 0$ (see Problem 3.10 b). However, we have seen that along each surface ∂S_c , $0 < c < 1$, w has a different limit as $\mathbf{x} \rightarrow \mathbf{0}$. Therefore the limit of u_g as $\mathbf{x} \rightarrow \mathbf{0}$ does not exist and $\mathbf{0}$ is not regular. \square

3.6 Fundamental Solution and Newtonian Potential

3.6.1 The fundamental solution

Equation (3.43), p. 137, is not the only representation formula for the solution of the Dirichlet problem. We shall derive deterministic formulas involving various types of *potentials*, constructed using a special function, called the *fundamental solution* of the Laplace operator.

As we did for the diffusion equation, let us look at the invariance properties characterizing the operator Δ : the invariances by *translations* and by *rotations*.

Let $u = u(\mathbf{x})$ be harmonic in \mathbb{R}^n . Invariance by translations means that the function $v(\mathbf{x}) = u(\mathbf{x} - \mathbf{y})$, for each fixed \mathbf{y} , is also harmonic, as it is immediate to check.

Invariance by rotations means that, given a rotation in \mathbb{R}^n , represented by an orthogonal matrix \mathbf{M} (i.e. $\mathbf{M}^\top = \mathbf{M}^{-1}$), also $v(\mathbf{x}) = u(\mathbf{M}\mathbf{x})$ is harmonic in \mathbb{R}^n . To check it, observe that, if we denote by D^2u the Hessian of u , we have

$$\Delta u = \text{Tr}D^2u = \text{trace of the Hessian of } u.$$

Since

$$D^2v(\mathbf{x}) = \mathbf{M}^\top D^2u(\mathbf{M}\mathbf{x}) \mathbf{M}$$

and \mathbf{M} is orthogonal, we have

$$\Delta v(\mathbf{x}) = \text{Tr}[\mathbf{M}^\top D^2u(\mathbf{M}\mathbf{x}) \mathbf{M}] = \text{Tr}D^2u(\mathbf{M}\mathbf{x}) = \Delta u(\mathbf{M}\mathbf{x}) = 0$$

and therefore v is harmonic.

Now, a typical rotationally invariant quantity is *the distance function from a point*, for instance from the origin, that is $r = |\mathbf{x}|$. Thus, let us look for *radially symmetric* harmonic functions $u = u(r)$.

Consider first $n = 2$; using polar coordinates and recalling (3.23), we find

$$\frac{\partial^2 u}{\partial r^2} + \frac{1}{r} \frac{\partial u}{\partial r} = 0,$$

so that

$$u(r) = C \log r + C_1.$$

In dimension $n = 3$, using spherical coordinates (r, ψ, θ) , $r > 0$, $0 < \psi < \pi$, $0 < \theta < 2\pi$, the operator Δ has the following expression¹⁹:

$$\Delta = \underbrace{\frac{\partial^2}{\partial r^2} + \frac{2}{r} \frac{\partial}{\partial r}}_{\text{radial part}} + \underbrace{\frac{1}{r^2} \left\{ \frac{1}{(\sin \psi)^2} \frac{\partial^2}{\partial \theta^2} + \frac{\partial^2}{\partial \psi^2} + \cot \psi \frac{\partial}{\partial \psi} \right\}}_{\text{spherical part (Laplace-Beltrami operator)}}.$$

¹⁹ See Appendix D.

The Laplace equation for $u = u(r)$ becomes

$$\frac{\partial^2 u}{\partial r^2} + \frac{2}{r} \frac{\partial u}{\partial r} = 0,$$

whose general integral is

$$u(r) = \frac{C}{r} + C_1, \quad C, C_1 \text{ arbitrary constants.}$$

Choose $C_1 = 0$ and $C = \frac{1}{4\pi}$ if $n = 3$, $C = -\frac{1}{2\pi}$ if $n = 2$. The function

$$\Phi(\mathbf{x}) = \begin{cases} -\frac{1}{2\pi} \log |\mathbf{x}| & n = 2 \\ \frac{1}{4\pi |\mathbf{x}|} & n = 3 \end{cases} \quad (3.48)$$

is called the **fundamental solution** for the Laplace operator Δ . As we shall prove in Chap. 7, the above choice of the constant C is made in order to have

$$\Delta \Phi(\mathbf{x}) = -\delta_n(\mathbf{x})$$

where $\delta_n(\mathbf{x})$ denotes the n -dimensional Dirac measure at $\mathbf{x} = \mathbf{0}$.

The physical meaning of Φ is remarkable: if $n = 3$, in standard units, $4\pi\Phi$ represents the electrostatic potential due to a unitary charge located at the origin and vanishing at infinity²⁰.

Clearly, if the origin is replaced by a point \mathbf{y} , the corresponding potential is $4\pi\Phi(\mathbf{x} - \mathbf{y})$ and

$$\Delta_{\mathbf{x}} \Phi(\mathbf{x} - \mathbf{y}) = -\delta_3(\mathbf{x} - \mathbf{y}).$$

By symmetry, we also have $\Delta_{\mathbf{y}} \Phi(\mathbf{x} - \mathbf{y}) = -\delta_3(\mathbf{x} - \mathbf{y})$.

Remark 3.30. In dimension $n > 3$, the fundamental solution of the Laplace operator is

$$\Phi(\mathbf{x}) = \omega_n^{-1} |\mathbf{x}|^{2-n},$$

where, we recall, ω_n is the surface area of the unit sphere in \mathbb{R}^n .

3.6.2 The Newtonian potential

Suppose that $(4\pi)^{-1} f(\mathbf{x})$ gives the density of a charge located inside a compact set in \mathbb{R}^3 . Then $\Phi(\mathbf{x} - \mathbf{y}) f(\mathbf{y}) d\mathbf{y}$ represents the potential at \mathbf{x} due to the charge $(4\pi)^{-1} f(\mathbf{y}) d\mathbf{y}$ inside a small region of volume $d\mathbf{y}$ around \mathbf{y} . The full potential is

²⁰ In dimension 2,

$$2\pi\Phi(x_1, x_2) = -\log \sqrt{x_1^2 + x_2^2}$$

represents the potential due to a charge of density 1, distributed along the x_3 axis.

given by the sum of all the contributions; we get

$$u(\mathbf{x}) = \int_{\mathbb{R}^3} \Phi(\mathbf{x} - \mathbf{y}) f(\mathbf{y}) d\mathbf{y} = \frac{1}{4\pi} \int_{\mathbb{R}^3} \frac{f(\mathbf{y})}{|\mathbf{x} - \mathbf{y}|} d\mathbf{y}, \quad (3.49)$$

which is the *convolution between f and Φ* and it is called the **Newtonian potential** of f . Formally, we have

$$\Delta u(\mathbf{x}) = \int_{\mathbb{R}^3} \Delta_{\mathbf{x}} \Phi(\mathbf{x} - \mathbf{y}) f(\mathbf{y}) d\mathbf{y} = - \int_{\mathbb{R}^3} \delta_3(\mathbf{x} - \mathbf{y}) f(\mathbf{y}) d\mathbf{y} = -f(\mathbf{x}). \quad (3.50)$$

Under suitable hypotheses on f , (3.50) is indeed true (see Theorem 3.31 below). Clearly, $u(\mathbf{x})$ is not the only solution of $\Delta u = -f$, since $u + c$, c constant, is a solution as well. However, the Newtonian potential is the only solution vanishing at infinity. All this is stated precisely in the theorem below, where, for simplicity, we assume $f \in C^2(\mathbb{R}^3)$ with compact support²¹. We have:

Theorem 3.31. *Let $f \in C^2(\mathbb{R}^3)$ with **compact support**. Let u be the Newtonian potential of f , defined by (3.49). Then, u is the only solution in \mathbb{R}^3 of*

$$\Delta u = -f \quad (3.51)$$

belonging to $C^2(\mathbb{R}^3)$ and vanishing at infinity.

Proof. Let $v \in C^2(\mathbb{R}^3)$ be another solution to (3.51), vanishing at infinity. Then $u - v$ is a *bounded* harmonic function in all \mathbb{R}^3 and therefore is constant by Corollary 3.14, p. 132. Since it vanishes at infinity it must be zero; thus $u = v$.

To show that (3.49) belongs to $C^2(\mathbb{R}^3)$ and satisfies (3.51), observe that we can write (3.49) in the alternative form

$$u(\mathbf{x}) = \int_{\mathbb{R}^3} \Phi(\mathbf{y}) f(\mathbf{x} - \mathbf{y}) d\mathbf{y} = \frac{1}{4\pi} \int_{\mathbb{R}^3} \frac{f(\mathbf{x} - \mathbf{y})}{|\mathbf{y}|} d\mathbf{y}.$$

Since $1/|\mathbf{y}|$ is integrable near the origin and f is zero outside a compact set, we can take first and second order derivatives under the integral sign to get

$$u_{x_i x_j}(\mathbf{x}; f) = \frac{1}{4\pi} \int_{\mathbb{R}^3} \Phi(\mathbf{y}) f_{x_i x_j}(\mathbf{x} - \mathbf{y}) d\mathbf{y}. \quad (3.52)$$

Since $f_{x_i x_j} \in C(\mathbb{R}^3)$, formula (3.52) shows that also $u_{x_i x_j}(\mathbf{x})$ is continuous and therefore $u(\mathbf{x}) \in C^2(\mathbb{R}^3)$.

It remains to prove (3.51). Since $\Delta_{\mathbf{x}} f(\mathbf{x} - \mathbf{y}) = \Delta_{\mathbf{y}} f(\mathbf{x} - \mathbf{y})$, from (3.52), we have

$$\Delta u(\mathbf{x}) = \int_{\mathbb{R}^3} \Phi(\mathbf{y}) \Delta_{\mathbf{x}} f(\mathbf{x} - \mathbf{y}) d\mathbf{y} = \int_{\mathbb{R}^3} \Phi(\mathbf{y}) \Delta_{\mathbf{y}} f(\mathbf{x} - \mathbf{y}) d\mathbf{y}.$$

We want to integrate by parts using formula (1.13), but, since Φ has a singularity at $\mathbf{y} = \mathbf{0}$, we have first to isolate the origin, by choosing a small ball $B_r = B_r(\mathbf{0})$ and

²¹ Recall that the support of a continuous function f is the closure of the set where f is not zero.

writing

$$\Delta u(\mathbf{x}) = \int_{B_r(\mathbf{0})} \cdots d\mathbf{y} + \int_{\mathbb{R}^3 \setminus B_r(\mathbf{0})} \cdots d\mathbf{y} \equiv I_r + J_r. \quad (3.53)$$

We have, using spherical coordinates,

$$|I_r| \leq \frac{\max |\Delta f|}{4\pi} \int_{B_r(\mathbf{0})} \frac{1}{|\mathbf{y}|} d\mathbf{y} = \max |\Delta f| \int_0^r \rho d\rho = \frac{\max |\Delta f|}{2} r^2$$

so that

$$I_r \rightarrow 0 \quad \text{if } r \rightarrow 0.$$

Keeping in mind that f vanishes outside a compact set, we can integrate J_r by parts (twice). Denoting by $\nu(\sigma)$ the outward²² unit normal to ∂B_r at σ , we obtain:

$$\begin{aligned} J_r &= \frac{1}{4\pi r} \int_{\partial B_r} \nabla_{\mathbf{y}} f(\mathbf{x} - \sigma) \cdot \nu(\sigma) d\sigma - \int_{\mathbb{R}^3 \setminus B_r(\mathbf{0})} \nabla \Phi(\mathbf{y}) \cdot \nabla_{\mathbf{y}} f(\mathbf{x} - \mathbf{y}) d\mathbf{y} \\ &= \frac{1}{4\pi r} \int_{\partial B_r} \nabla_{\mathbf{y}} f(\mathbf{x} - \sigma) \cdot \nu(\sigma) d\sigma - \int_{\partial B_r} f(\mathbf{x} - \sigma) \nabla \Phi(\sigma) \cdot \nu(\sigma) d\sigma \end{aligned}$$

since $\Delta \Phi = 0$ in $\mathbb{R}^3 \setminus B_r(\mathbf{0})$. We have:

$$\frac{1}{4\pi r} \left| \int_{\partial B_r} \nabla_{\mathbf{y}} f(\mathbf{x} - \sigma) \cdot \nu(\sigma) d\sigma \right| \leq r \max |\nabla f| \rightarrow 0 \quad \text{as } r \rightarrow 0.$$

On the other hand, $\nabla \Phi(\mathbf{y}) = -\mathbf{y} |\mathbf{y}|^{-3} / 4\pi$ and the outward unit normal on ∂B_r is $\nu(\sigma) = -\sigma/r$, so that

$$\int_{\partial B_r} f(\mathbf{x} - \sigma) \nabla \Phi(\sigma) \cdot \nu(\sigma) d\sigma = \frac{1}{4\pi r^2} \int_{\partial B_r} f(\mathbf{x} - \sigma) d\sigma \rightarrow f(\mathbf{x}) \quad \text{as } r \rightarrow 0.$$

Thus, $J_r \rightarrow -f(\mathbf{x})$ as $r \rightarrow 0$. Passing to the limit as $r \rightarrow 0$ in (3.53) we get

$$\Delta u(\mathbf{x}) = -f(\mathbf{x}). \quad \square$$

Remark 3.32. Theorem 3.31 actually holds under much less restrictive hypotheses on f . For instance, it is enough that $f \in C^1(\mathbb{R}^3)$ and $|f(\mathbf{x})| \leq C |\mathbf{x}|^{-3-\varepsilon}$, $\varepsilon > 0$.

Remark 3.33. An appropriate version of Theorem 3.31 holds in dimension $n = 2$, with the Newtonian potential replaced by the *logarithmic potential*

$$u(\mathbf{x}) = \int_{\mathbb{R}^2} \Phi(\mathbf{x} - \mathbf{y}) f(\mathbf{y}) d\mathbf{y} = -\frac{1}{2\pi} \int_{\mathbb{R}^2} \log |\mathbf{x} - \mathbf{y}| f(\mathbf{y}) d\mathbf{y}. \quad (3.54)$$

The logarithmic potential does not vanish at infinity; its asymptotic behavior is (see Problem 3.11)

$$u(\mathbf{x}) = -\frac{M}{2\pi} \log |\mathbf{x}| + O\left(\frac{1}{|\mathbf{x}|}\right) \quad \text{as } |\mathbf{x}| \rightarrow +\infty \quad (3.55)$$

²² With respect to $\mathbb{R}^3 \setminus B_r$.

where

$$M = \int_{\mathbb{R}^2} f(\mathbf{y}) d\mathbf{y}.$$

The logarithmic potential (3.54) is the only solution of $\Delta u = -f$ in \mathbb{R}^2 satisfying (3.55).

3.6.3 A divergence-curl system. Helmholtz decomposition formula

Using the properties of the Newtonian potential we can solve the following two problems, that appear in several applications e.g. to linear elasticity, fluid dynamics or electrostatics.

1) *Reconstruction of a vector field $\mathbf{u} \in \mathbb{R}^3$, from the knowledge of its curl and divergence.* Precisely, given a scalar f and a vector field $\boldsymbol{\omega}$, we want to find a vector field \mathbf{u} such that

$$\begin{cases} \operatorname{div} \mathbf{u} = f \\ \operatorname{curl} \mathbf{u} = \boldsymbol{\omega} \end{cases} \quad \text{in } \mathbb{R}^3.$$

We assume that \mathbf{u} has continuous second derivatives and vanishes at infinity, as it is required in most applications.

2) *Decomposition of a vector field $\mathbf{u} \in \mathbb{R}^3$ into the sum of a curl-free vector field and a divergence-free vector field.* Precisely, given \mathbf{u} , we want to find a scalar φ and a vector field \mathbf{w} such that the following *Helmholtz decomposition formula* holds

$$\mathbf{u} = \nabla \varphi + \operatorname{curl} \mathbf{w}. \quad (3.56)$$

Consider problem 1). First of all observe that, since $\operatorname{div} \operatorname{curl} \mathbf{u} = 0$, a necessary condition for the existence of a solution is $\operatorname{div} \boldsymbol{\omega} = 0$.

Let us check uniqueness. If \mathbf{u}_1 and \mathbf{u}_2 are solutions sharing the same data f and $\boldsymbol{\omega}$, their difference $\mathbf{w} = \mathbf{u}_1 - \mathbf{u}_2$ vanishes at infinity and satisfies

$$\operatorname{div} \mathbf{w} = 0 \quad \text{and} \quad \operatorname{curl} \mathbf{w} = \mathbf{0}, \quad \text{in } \mathbb{R}^3.$$

From $\operatorname{curl} \mathbf{w} = \mathbf{0}$ we infer the existence of a scalar function U such that $\nabla U = \mathbf{w}$. From $\operatorname{div} \mathbf{w} = 0$, we deduce

$$\operatorname{div} \nabla U = \Delta U = 0.$$

Thus U is harmonic. Hence its derivatives, that is the components w_j of \mathbf{w} , are bounded harmonic functions in \mathbb{R}^3 . Liouville's Theorem 3.14, p. 132, implies that each w_j is constant and therefore identically zero, since it vanishes at infinity. We conclude that, under the stated assumptions, the *solution of problem 1) is unique*.

To find \mathbf{u} , split it into $\mathbf{u} = \mathbf{v} + \mathbf{z}$ and look for \mathbf{v} and \mathbf{z} such that

$$\operatorname{div} \mathbf{z} = 0, \quad \operatorname{curl} \mathbf{z} = \boldsymbol{\omega}$$

$$\operatorname{div} \mathbf{v} = f, \quad \operatorname{curl} \mathbf{v} = \mathbf{0}.$$

As before, from $\operatorname{curl} \mathbf{v} = \mathbf{0}$, we infer the existence of a scalar function φ such that $\nabla \varphi = \mathbf{v}$, while $\operatorname{div} \mathbf{v} = f$ implies $\Delta \varphi = f$. We have seen that, under suitable hypotheses on f , φ is given by the Newtonian potential of f , that is:

$$\varphi(\mathbf{x}) = -\frac{1}{4\pi} \int_{\mathbb{R}^3} \frac{1}{|\mathbf{x} - \mathbf{y}|} f(\mathbf{y}) d\mathbf{y}$$

and $\mathbf{v} = \nabla \varphi$. To find \mathbf{z} , recall the identity

$$\operatorname{curl} \operatorname{curl} \mathbf{z} = \nabla(\operatorname{div} \mathbf{z}) - \Delta \mathbf{z}.$$

Since $\operatorname{div} \mathbf{z} = 0$, we get

$$\Delta \mathbf{z} = -\operatorname{curl} \operatorname{curl} \mathbf{z} = -\operatorname{curl} \boldsymbol{\omega}$$

so that

$$\mathbf{z}(\mathbf{x}) = \frac{1}{4\pi} \int_{\mathbb{R}^3} \frac{1}{|\mathbf{x} - \mathbf{y}|} \operatorname{curl} \boldsymbol{\omega}(\mathbf{y}) d\mathbf{y}.$$

Let us summarize the conclusions in the next theorem, also specifying the hypotheses²³ on f and $\boldsymbol{\omega}$.

Theorem 3.34. *Let $f \in C^1(\mathbb{R}^3)$, $\boldsymbol{\omega} \in C^2(\mathbb{R}^3; \mathbb{R}^3)$ such that $\operatorname{div} \boldsymbol{\omega} = 0$ and, for $|\mathbf{x}|$ large,*

$$|f(\mathbf{x})| \leq \frac{M}{|\mathbf{x}|^{3+\varepsilon}}, \quad |\operatorname{curl} \boldsymbol{\omega}(\mathbf{x})| \leq \frac{M}{|\mathbf{x}|^{3+\varepsilon}} \quad (\varepsilon > 0).$$

Then, the unique solution vanishing at infinity of the system

$$\begin{cases} \operatorname{div} \mathbf{u} = f \\ \operatorname{curl} \mathbf{u} = \boldsymbol{\omega} \end{cases} \quad \text{in } \mathbb{R}^3$$

is given by the vector field

$$\mathbf{u}(\mathbf{x}) = \int_{\mathbb{R}^3} \frac{1}{4\pi |\mathbf{x} - \mathbf{y}|} \operatorname{curl} \boldsymbol{\omega}(\mathbf{y}) d\mathbf{y} - \nabla \int_{\mathbb{R}^3} \frac{1}{4\pi |\mathbf{x} - \mathbf{y}|} f(\mathbf{y}) d\mathbf{y}. \quad (3.57)$$

²³ See Remark 3.32, p. 150.

Consider now problem 2). If \mathbf{u} , $f = \operatorname{div} \mathbf{u}$ and $\boldsymbol{\omega} = \operatorname{curl} \mathbf{u}$ satisfy the assumptions of Theorem 3.34, we can write

$$\mathbf{u}(\mathbf{x}) = \int_{\mathbb{R}^3} \frac{1}{4\pi |\mathbf{x} - \mathbf{y}|} \operatorname{curl} \operatorname{curl} \mathbf{u}(\mathbf{y}) d\mathbf{y} - \nabla \int_{\mathbb{R}^3} \frac{1}{4\pi |\mathbf{x} - \mathbf{y}|} \operatorname{div} \mathbf{u}(\mathbf{y}) d\mathbf{y}.$$

Since \mathbf{u} is rapidly vanishing at infinity, we have²⁴

$$\int_{\mathbb{R}^3} \frac{1}{|\mathbf{x} - \mathbf{y}|} \operatorname{curl} \operatorname{curl} \mathbf{u}(\mathbf{y}) d\mathbf{y} = \operatorname{curl} \int_{\mathbb{R}^3} \frac{1}{|\mathbf{x} - \mathbf{y}|} \operatorname{curl} \mathbf{u}(\mathbf{y}) d\mathbf{y}. \quad (3.58)$$

We conclude that

$$\mathbf{u} = \nabla \varphi + \operatorname{curl} \mathbf{w}$$

where

$$\varphi(\mathbf{x}) = - \int_{\mathbb{R}^3} \frac{1}{4\pi |\mathbf{x} - \mathbf{y}|} \operatorname{div} \mathbf{u}(\mathbf{y}) d\mathbf{y}$$

and

$$\mathbf{w}(\mathbf{x}) = \int_{\mathbb{R}^3} \frac{1}{4\pi |\mathbf{x} - \mathbf{y}|} \operatorname{curl} \mathbf{u}(\mathbf{y}) d\mathbf{y}.$$

- *An application to fluid dynamics.* Consider the three dimensional flow of an incompressible fluid, of both constant density ρ and viscosity μ , subject to a conservative external force²⁵ $\mathbf{F} = \nabla f$.

Let $\mathbf{u} = \mathbf{u}(\mathbf{x}, t)$ denote the velocity field, $p = p(\mathbf{x}, t)$ the pressure and $\mathbf{T} = (T_{ij})_{i,j=1,2,3}$ the stress tensor. The laws of conservation of mass and linear momentum give the following equations:

$$\rho_t + \operatorname{div}(\rho \mathbf{u}) = 0$$

and

$$\rho \frac{D\mathbf{u}}{Dt} = \rho (\mathbf{u}_t + (\mathbf{u} \cdot \nabla) \mathbf{u}) = \mathbf{F} + \operatorname{div} \mathbf{T}$$

where $\operatorname{div} \mathbf{T}$ is the vector of components $\sum_{j=1,2,3} \partial_{x_j} T_{ij}$, $i = 1, 2, 3$.

The so called Newtonian fluids (like water) are characterized by the constitutive law

$$\mathbf{T} = -p \mathbf{I} + \mu \nabla \mathbf{u}$$

where \mathbf{I} is the identity matrix. Being ρ constant, we get for \mathbf{u} and p the celebrated *Navier-Stokes equations*:

$$\operatorname{div} \mathbf{u} = 0 \quad (3.59)$$

²⁴ In fact, if $g \in C^1(\mathbb{R}^3)$ and $|g(\mathbf{x})| \leq \frac{M}{|\mathbf{x}|^{3+\varepsilon}}$, one can show that

$$\frac{\partial}{\partial x_j} \int_{\mathbb{R}^3} \frac{1}{|\mathbf{x} - \mathbf{y}|} g(\mathbf{y}) d\mathbf{y} = \int_{\mathbb{R}^3} \frac{1}{|\mathbf{x} - \mathbf{y}|} \frac{\partial g}{\partial y_j} d\mathbf{y}.$$

²⁵ Gravity, for instance.

and

$$\frac{D\mathbf{u}}{Dt} = \mathbf{u}_t + (\mathbf{u} \cdot \nabla) \mathbf{u} = -\frac{1}{\rho} \nabla p + \nu \Delta \mathbf{u} + \frac{1}{\rho} \nabla f \quad (\nu = \mu/\rho). \quad (3.60)$$

We look for solution of (3.59), (3.60) subject to a given initial condition

$$\mathbf{u}(\mathbf{x}, 0) = \mathbf{g}(\mathbf{x}) \quad \mathbf{x} \in \mathbb{R}^3, \quad (3.61)$$

where \mathbf{g} is also divergence-free:

$$\operatorname{div} \mathbf{g} = 0.$$

The quantity $\frac{D\mathbf{u}}{Dt}$ is called the *material derivative of \mathbf{u}* , given by the sum of \mathbf{u}_t , the fluid acceleration due to the non-stationary character of the motion, and of $(\mathbf{v} \cdot \nabla) \mathbf{v}$, the inertial acceleration due to fluid transport²⁶.

In general, the system (3.59), (3.60) is extremely difficult to solve. In the case of slow flow, for instance due to high viscosity, the inertial term becomes negligible, compared for instance to $\nu \Delta \mathbf{u}$, and (3.60) simplifies to the linearized equation

$$\mathbf{u}_t = -\frac{1}{\rho} \nabla p + \nu \Delta \mathbf{u} + \nabla f. \quad (3.62)$$

It is possible to find an explicit formula for the solution of (3.59), (3.61), (3.62) by writing everything in terms of $\boldsymbol{\omega} = \operatorname{curl} \mathbf{u}$. In fact, taking the curl of (3.62) and (3.61), we obtain, since

$$\operatorname{curl}(\nabla p + \nu \Delta \mathbf{u} + \nabla f) = \nu \Delta \boldsymbol{\omega},$$

$$\begin{cases} \boldsymbol{\omega}_t = \nu \Delta \boldsymbol{\omega} & \mathbf{x} \in \mathbb{R}^3, t > 0 \\ \boldsymbol{\omega}(\mathbf{x}, 0) = \operatorname{curl} \mathbf{g}(\mathbf{x}) & \mathbf{x} \in \mathbb{R}^3. \end{cases}$$

This is a global Cauchy problem for the heat equation. If $\mathbf{g} \in C^2(\mathbb{R}^3; \mathbb{R}^3)$ and $\operatorname{curl} \mathbf{g}$ is bounded, we have

$$\boldsymbol{\omega}(\mathbf{x}, t) = \frac{1}{(4\pi\nu t)^{3/2}} \int_{\mathbb{R}^3} \exp\left(-\frac{|\mathbf{y}|^2}{4\nu t}\right) \operatorname{curl} \mathbf{g}(\mathbf{x} - \mathbf{y}) d\mathbf{y}. \quad (3.63)$$

Moreover, for $t > 0$, we can take the divergence operator under the integral in (3.63) and deduce that $\operatorname{div} \boldsymbol{\omega} = 0$. Therefore, if $\operatorname{curl} \mathbf{g}(\mathbf{x})$ vanishes rapidly at

²⁶ The i -th component of $(\mathbf{v} \cdot \nabla) \mathbf{v}$ is given by $\sum_{j=1}^3 v_j \frac{\partial v_i}{\partial x_j}$. Let us compute $\frac{D\mathbf{v}}{Dt}$, for example, for a plane fluid, uniformly rotating with angular speed ω . Then $\mathbf{v}(x, y) = -\omega y \mathbf{i} + \omega x \mathbf{j}$. Since $\mathbf{v}_t = \mathbf{0}$, the motion is stationary and

$$\frac{D\mathbf{v}}{Dt} = (\mathbf{v} \cdot \nabla) \mathbf{v} = \left(-\omega y \frac{\partial}{\partial x} + \omega x \frac{\partial}{\partial y} \right) (-\omega y \mathbf{i} + \omega x \mathbf{j}) = -\omega^2 (-x \mathbf{i} + y \mathbf{j}),$$

which is the centrifugal acceleration.

infinity²⁷, we can recover \mathbf{u} by solving the system

$$\operatorname{curl} \mathbf{u} = \boldsymbol{\omega}, \quad \operatorname{div} \mathbf{u} = 0,$$

according to formula (3.57) with $f = 0$.

Finally to find the pressure, from (3.62) we have

$$\nabla p = -\rho \mathbf{u}_t + \mu \Delta \mathbf{u} - \nabla f. \quad (3.64)$$

Since

$$\boldsymbol{\omega}_t = \nu \Delta \boldsymbol{\omega},$$

the right hand side curl-free; hence (3.64) can be solved and determines p up to an additive constant (as it should be).

In conclusion:

Proposition 3.35. *Let, $f \in C^1(\mathbb{R}^3)$, $\mathbf{g} \in C^2(\mathbb{R}^3; \mathbb{R}^3)$, with $\operatorname{div} \mathbf{g} = 0$ and $\operatorname{curl} \mathbf{g}$ rapidly vanishing at infinity. There exist a unique $\mathbf{u} \in C^2(\mathbb{R}^3; \mathbb{R}^3)$, with $\operatorname{curl} \mathbf{u}$ vanishing at infinity, and $p \in C^1(\mathbb{R}^3)$, unique up to an additive constant, satisfying the system (3.59), (3.61), (3.62).*

3.7 The Green Function

3.7.1 An integral identity

Formula (3.49) gives a representation of the solution to Poisson's equation in all \mathbb{R}^3 . In bounded domains Ω , any representation formula has to take into account the boundary values, as indicated in the following theorem.

To avoid any confusion we clarify some notations. We write $r_{\mathbf{xy}} = |\mathbf{x} - \mathbf{y}|$. If \mathbf{x} is fixed, the symbol $\Phi(\mathbf{x} - \cdot)$ denotes the function

$$\mathbf{y} \mapsto \Phi(\mathbf{x} - \mathbf{y}).$$

If $\sigma \in \partial\Omega$, $\nu = \nu(\sigma)$ denotes the outward unit normal to $\partial\Omega$ at σ . The (outward) normal derivative of u at σ is denoted by one of the symbols

$$\partial_\nu u = \frac{\partial u}{\partial \nu} = \nabla u(\sigma) \cdot \nu(\sigma).$$

Accordingly,

$$\partial_\nu \Phi(\mathbf{x} - \sigma) = \nabla_y \Phi(\mathbf{x} - \sigma) \cdot \nu(\sigma) = -\nabla_x \Phi(\mathbf{x} - \sigma) \cdot \nu(\sigma).$$

²⁷ $|\operatorname{curl} \mathbf{g}(\mathbf{x})| \leq M / |\mathbf{x}|^{3+\varepsilon}$, $\varepsilon > 0$, it is enough.

Theorem 3.36. Let $\Omega \subset \mathbb{R}^n$ be a bounded, smooth domain and $u \in C^2(\overline{\Omega})$. Then, for every $\mathbf{x} \in \Omega$,

$$\begin{aligned} u(\mathbf{x}) = & - \int_{\Omega} \Phi(\mathbf{x} - \mathbf{y}) \Delta u(\mathbf{y}) d\mathbf{y} + \\ & + \int_{\partial\Omega} \Phi(\mathbf{x} - \boldsymbol{\sigma}) \partial_{\boldsymbol{\nu}} u(\boldsymbol{\sigma}) d\sigma - \int_{\partial\Omega} u(\boldsymbol{\sigma}) \partial_{\boldsymbol{\nu}} \Phi(\mathbf{x} - \boldsymbol{\sigma}) d\sigma. \end{aligned} \quad (3.65)$$

The last two terms in the right hand side of (3.65) are called *single* and *double layer potentials* of $\partial_{\boldsymbol{\nu}} u$ and $-u$, respectively. We are going to examine these surface potentials later. The first one is the Newtonian potential of $-\Delta u$ in Ω .

Proof. We do it for $n = 3$. The proof for $n = 2$ is similar. For fixed $\mathbf{x} \in \Omega$, we would like to apply *Green's identity* (1.15), p. 16,

$$\int_{\Omega} (v \Delta u - u \Delta v) d\mathbf{x} = \int_{\partial\Omega} (v \partial_{\boldsymbol{\nu}} u - u \partial_{\boldsymbol{\nu}} v) d\sigma, \quad (3.66)$$

to u and $\Phi(\mathbf{x} - \cdot)$. However, $\Phi(\mathbf{x} - \cdot)$ has a singularity at \mathbf{x} , so that it cannot be inserted directly into (3.66). Let us isolate the singularity inside a ball $B_{\varepsilon}(\mathbf{x})$, with ε small. In the domain $\Omega_{\varepsilon} = \Omega \setminus \overline{B}_{\varepsilon}(\mathbf{x})$, $\Phi(\mathbf{x} - \cdot)$ is smooth and harmonic.

Thus, replacing Ω with Ω_{ε} , we can apply (3.66) to u and $\Phi(\mathbf{x} - \cdot)$. Since

$$\partial\Omega_{\varepsilon} = \partial\Omega \cup \partial B_{\varepsilon}(\mathbf{x}),$$

and $\Delta_{\mathbf{y}} \Phi(\mathbf{x} - \mathbf{y}) = 0$, we find:

$$\begin{aligned} \int_{\Omega_{\varepsilon}} \frac{1}{r_{\mathbf{x}\mathbf{y}}} \Delta u d\mathbf{y} &= \int_{\partial\Omega_{\varepsilon}} \left(\frac{1}{r_{\mathbf{x}\boldsymbol{\sigma}}} \frac{\partial u}{\partial \boldsymbol{\nu}} - u \frac{\partial}{\partial \boldsymbol{\nu}} \frac{1}{r_{\mathbf{x}\boldsymbol{\sigma}}} \right) d\sigma \\ &= \int_{\partial\Omega} (\dots) d\sigma + \int_{\partial B_{\varepsilon}(\mathbf{x})} \frac{1}{r_{\mathbf{x}\boldsymbol{\sigma}}} \frac{\partial u}{\partial \boldsymbol{\nu}} d\sigma + \int_{\partial B_{\varepsilon}(\mathbf{x})} u \frac{\partial}{\partial \boldsymbol{\nu}} \frac{1}{r_{\mathbf{x}\boldsymbol{\sigma}}} d\sigma. \end{aligned} \quad (3.67)$$

We let now $\varepsilon \rightarrow 0$ in (3.67). We have:

$$\int_{\Omega_{\varepsilon}} \frac{1}{r_{\mathbf{x}\boldsymbol{\sigma}}} \Delta u d\mathbf{y} \rightarrow \int_{\Omega} \frac{1}{r_{\mathbf{x}\boldsymbol{\sigma}}} \Delta u d\mathbf{y}, \quad \text{as } \varepsilon \rightarrow 0 \quad (3.68)$$

since $\Delta u \in C(\overline{\Omega})$ and $r_{\mathbf{x}\boldsymbol{\sigma}}^{-1}$ is positive and integrable in Ω .

On $\partial B_{\varepsilon}(\mathbf{x})$, we have $r_{\mathbf{x}\boldsymbol{\sigma}} = \varepsilon$ and $|\partial_{\boldsymbol{\nu}} u| \leq M$, since $|\nabla u|$ is bounded; then

$$\left| \int_{\partial B_{\varepsilon}(\mathbf{x})} \frac{1}{r_{\mathbf{x}\boldsymbol{\sigma}}} \frac{\partial u}{\partial \boldsymbol{\nu}} d\sigma \right| \leq 4\pi\varepsilon M \rightarrow 0, \quad \text{as } \varepsilon \rightarrow 0. \quad (3.69)$$

The most delicate term is

$$\int_{\partial B_{\varepsilon}(\mathbf{x})} u \frac{\partial}{\partial \boldsymbol{\nu}} \frac{1}{r_{\mathbf{x}\boldsymbol{\sigma}}} d\sigma.$$

On $\partial B_{\varepsilon}(\mathbf{x})$, the outward (with respect to Ω_{ε}) unit normal at $\boldsymbol{\sigma}$ is $\boldsymbol{\nu}(\boldsymbol{\sigma}) = \frac{\mathbf{x} - \boldsymbol{\sigma}}{\varepsilon}$, so that

$$\frac{\partial}{\partial \boldsymbol{\nu}} \frac{1}{r_{\mathbf{x}\boldsymbol{\sigma}}} = \nabla_{\mathbf{y}} \frac{1}{r_{\mathbf{x}\boldsymbol{\sigma}}} \cdot \boldsymbol{\nu}(\boldsymbol{\sigma}) = \frac{\mathbf{x} - \boldsymbol{\sigma}}{\varepsilon^3} \frac{\mathbf{x} - \boldsymbol{\sigma}}{\varepsilon} = \frac{1}{\varepsilon^2}.$$

As a consequence,

$$\int_{\partial B_\varepsilon(\mathbf{x})} u \frac{\partial}{\partial \boldsymbol{\nu}} \frac{1}{r_{\mathbf{x}\sigma}} d\sigma = \frac{1}{\varepsilon^2} \int_{\partial B_\varepsilon(\mathbf{x})} u d\sigma \rightarrow 4\pi u(\mathbf{x}) \quad (3.70)$$

as $\varepsilon \rightarrow 0$, by the continuity of u .

Letting $\varepsilon \rightarrow 0$ in (3.67), from (3.68), (3.69), (3.70) we obtain (3.65). \square

3.7.2 Green's function

The function Φ defined in (3.48) is the fundamental solution for the Laplace operator Δ in all \mathbb{R}^n , for $n = 2, 3$. We can also define a fundamental solution for the Laplace operator in any open set and in particular in any *bounded* domain $\Omega \subset \mathbb{R}^3$, representing, up to the factor 4π , the potential due to a unit charge, placed at a point $\mathbf{x} \in \Omega$ and equal to zero on $\partial\Omega$.

This function, that we denote by $G(\mathbf{x}, \mathbf{y})$, is called the *Green function in Ω* , for the operator Δ ; for fixed $\mathbf{x} \in \Omega$, G satisfies

$$\Delta_{\mathbf{y}} G(\mathbf{x}, \mathbf{y}) = -\delta_3(\mathbf{x} - \mathbf{y}) \quad \text{in } \Omega$$

and

$$G(\mathbf{x}, \boldsymbol{\sigma}) = 0, \quad \text{for } \boldsymbol{\sigma} \in \partial\Omega.$$

More explicitly, the Green function can be written in the form

$$G(\mathbf{x}, \mathbf{y}) = \Phi(\mathbf{x} - \mathbf{y}) - \varphi(\mathbf{x}, \mathbf{y}) \quad (3.71)$$

where φ , for fixed $\mathbf{x} \in \Omega$, solves the Dirichlet problem

$$\begin{cases} \Delta_{\mathbf{y}} \varphi = 0 & \text{in } \Omega \\ \varphi(\mathbf{x}, \boldsymbol{\sigma}) = \Phi(\mathbf{x} - \boldsymbol{\sigma}) & \text{on } \partial\Omega. \end{cases} \quad (3.72)$$

The Green function has the following important properties (see Problem 3.14):

- (a) *Positivity:* $G(\mathbf{x}, \mathbf{y}) > 0$ for every $\mathbf{x}, \mathbf{y} \in \Omega$, with $G(\mathbf{x}, \mathbf{y}) \rightarrow +\infty$ when $\mathbf{x} - \mathbf{y} \rightarrow 0$.
- (b) *Symmetry:* $G(\mathbf{x}, \mathbf{y}) = G(\mathbf{y}, \mathbf{x})$.

The existence of the Green function for a particular domain depends on the solvability of the Dirichlet problem (3.72). According to Corollary 3.27, p. 144, we know that this is the case if Ω is regular for the Dirichlet problem, for instance, if Ω is a bounded Lipschitz domain.

Even if we know that the Green function exists, explicit formulas are available only for special domains. Sometimes, a technique known as *method of electrostatic images* works. In this method $4\pi\varphi(\mathbf{x}, \cdot)$ is considered as the potential due to an

imaginary charge q , placed at a suitable point \mathbf{x}^* , the *image of \mathbf{x}* , in the complement of Ω . The charge q and the point \mathbf{x}^* have to be chosen so that $4\pi\varphi(\mathbf{x}, \cdot)$, on $\partial\Omega$, is equal to the potential created by the unit charge in \mathbf{x} .

The simplest way to illustrate the method is to determine the Green's function for the upper half-space, although this is an unbounded domain. Clearly, we require that G vanishes at infinity.

- *Green's function for the upper half space in \mathbb{R}^3 .* Let \mathbb{R}_+^3 be the upper half space:

$$\mathbb{R}_+^3 = \{(x_1, x_2, x_3) : x_3 > 0\}.$$

Fix $\mathbf{x} = (x_1, x_2, x_3)$ and observe that, if we choose $\mathbf{x}^* = (x_1, x_2, -x_3)$, then, on $y_3 = 0$, we have

$$|\mathbf{x}^* - \mathbf{y}| = |\mathbf{x} - \mathbf{y}|.$$

Thus, if $\mathbf{x} \in \mathbb{R}_+^3$, \mathbf{x}^* belongs to the complement of \mathbb{R}_+^3 , the function

$$\varphi(\mathbf{x}, \mathbf{y}) = \Phi(\mathbf{x}^* - \mathbf{y}) = \frac{1}{4\pi |\mathbf{x}^* - \mathbf{y}|}$$

is harmonic in \mathbb{R}_+^3 and $\varphi(\mathbf{x}, \mathbf{y}) = \Phi(\mathbf{x} - \mathbf{y})$ on the plane $y_3 = 0$. In conclusion,

$$G(\mathbf{x}, \mathbf{y}) = \frac{1}{4\pi |\mathbf{x} - \mathbf{y}|} - \frac{1}{4\pi |\mathbf{x}^* - \mathbf{y}|} \quad (3.73)$$

is the Green's function for the upper half space.

- *Green's function for sphere.* Let

$$\Omega = B_R = B_R(\mathbf{0}) \subset \mathbb{R}^3.$$

To determine the Green's function for B_R , set

$$\varphi(\mathbf{x}, \mathbf{y}) = \frac{q}{4\pi |\mathbf{x}^* - \mathbf{y}|},$$

\mathbf{x} fixed in B_R , and try to determine \mathbf{x}^* , outside B_R , and q , so that

$$\frac{q}{|\mathbf{x}^* - \mathbf{y}|} = \frac{1}{|\mathbf{x} - \mathbf{y}|}, \quad (3.74)$$

when $|\mathbf{y}| = R$. The (3.74) for $|\mathbf{y}| = R$, gives

$$|\mathbf{x}^* - \mathbf{y}|^2 = q^2 |\mathbf{x} - \mathbf{y}|^2 \quad (3.75)$$

or

$$|\mathbf{x}^*|^2 - 2\mathbf{x}^* \cdot \mathbf{y} + R^2 = q^2 (|\mathbf{x}|^2 - 2\mathbf{x} \cdot \mathbf{y} + R^2).$$

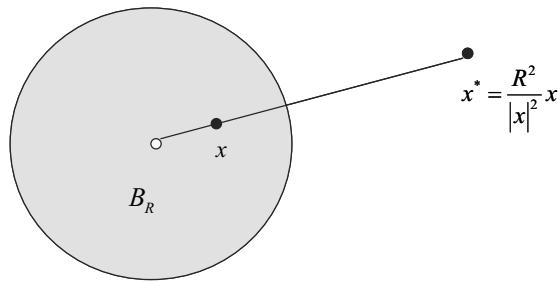


Fig. 3.9 The image \mathbf{x}^* of \mathbf{x} in the construction of the Green's function for the sphere

Rearranging the terms we have

$$|\mathbf{x}^*|^2 + R^2 - q^2(R^2 + |\mathbf{x}|^2) = 2\mathbf{y} \cdot (\mathbf{x}^* - q^2\mathbf{x}). \quad (3.76)$$

Since the left hand side does not depend on \mathbf{y} , it must be that

$$\mathbf{x}^* = q^2\mathbf{x}$$

and

$$q^4 |\mathbf{x}|^2 - q^2(R^2 + |\mathbf{x}|^2) + R^2 = 0$$

from which

$$q = R/|\mathbf{x}|. \quad (3.77)$$

This works for $\mathbf{x} \neq \mathbf{0}$ and gives

$$G(\mathbf{x}, \mathbf{y}) = \frac{1}{4\pi} \left[\frac{1}{|\mathbf{x} - \mathbf{y}|} - \frac{R}{|\mathbf{x}| |\mathbf{x}^* - \mathbf{y}|} \right], \quad \mathbf{x}^* = \frac{R^2}{|\mathbf{x}|^2} \mathbf{x}, \mathbf{x} \neq \mathbf{0}. \quad (3.78)$$

Since

$$|\mathbf{x}^* - \mathbf{y}| = |\mathbf{x}|^{-1} \left(R^4 - 2R^2 \mathbf{x} \cdot \mathbf{y} + |\mathbf{y}|^2 |\mathbf{x}|^2 \right)^{1/2},$$

when $\mathbf{x} \rightarrow \mathbf{0}$ we have

$$\varphi(\mathbf{x}, \mathbf{y}) = \frac{1}{4\pi} \frac{R}{|\mathbf{x}| |\mathbf{x}^* - \mathbf{y}|} \rightarrow \frac{1}{4\pi R}$$

and therefore we can define

$$G(\mathbf{0}, \mathbf{y}) = \frac{1}{4\pi} \left[\frac{1}{|\mathbf{y}|} - \frac{1}{R} \right].$$

Remark 3.37. Formula (3.71) actually defines the Green function for the Laplace operator in a domain $\Omega \subset \mathbb{R}^n$, with $n \geq 2$.

3.7.3 Green's representation formula

From Theorem 3.36 we know that every smooth function u can be written as the sum of the volume (Newtonian) potential of $-\Delta u$, the single layer potential of $\partial_\nu u$ and the double layer potential of u . Suppose u solves the Dirichlet problem

$$\begin{cases} \Delta u = f & \text{in } \Omega \\ u = g & \text{on } \partial\Omega. \end{cases} \quad (3.79)$$

Then (3.65) gives, for $\mathbf{x} \in \Omega$,

$$u(\mathbf{x}) = - \int_{\Omega} \Phi(\mathbf{x} - \mathbf{y}) f(\mathbf{y}) d\mathbf{y} + \\ + \int_{\partial\Omega} \Phi(\mathbf{x} - \boldsymbol{\sigma}) \partial_\nu u(\boldsymbol{\sigma}) d\sigma - \int_{\partial\Omega} g(\boldsymbol{\sigma}) \partial_\nu \Phi(\mathbf{x} - \boldsymbol{\sigma}) d\sigma. \quad (3.80)$$

This representation formula for u is not satisfactory, since it involves the data f and g but also the normal derivative $\partial_\nu u$, which is unknown. To get rid of $\partial_\nu u$, let $G(\mathbf{x}, \mathbf{y}) = \Phi(\mathbf{x} - \mathbf{y}) - \varphi(\mathbf{x}, \mathbf{y})$ be the Green function in Ω . Since $\varphi(\mathbf{x}, \cdot)$ is harmonic in Ω , we can apply (3.66) to u and $\varphi(\mathbf{x}, \cdot)$; we find

$$0 = \int_{\Omega} \varphi(\mathbf{x}, \mathbf{y}) f(\mathbf{y}) d\mathbf{y} + \\ - \int_{\partial\Omega} \varphi(\mathbf{x}, \boldsymbol{\sigma}) \partial_\nu u(\boldsymbol{\sigma}) d\sigma + \int_{\partial\Omega} g(\boldsymbol{\sigma}) \partial_\nu \varphi(\mathbf{x}, \boldsymbol{\sigma}) d\sigma. \quad (3.81)$$

Adding (3.80), (3.81) and recalling that $\varphi(\mathbf{x}, \boldsymbol{\sigma}) = \Phi(\mathbf{x} - \boldsymbol{\sigma})$ on $\partial\Omega$, we obtain:

Theorem 3.38. *Let Ω be a smooth domain and u be a smooth solution of (3.79). Then:*

$$u(\mathbf{x}) = - \int_{\Omega} f(\mathbf{y}) G(\mathbf{x}, \mathbf{y}) d\mathbf{y} - \int_{\partial\Omega} g(\boldsymbol{\sigma}) \partial_\nu G(\mathbf{x}, \boldsymbol{\sigma}) d\sigma. \quad (3.82)$$

Thus, the solution of the Dirichlet problem (3.79) can be written as the sum of the two Green's potentials in the right hand side of (3.82) and it is known as soon as the Green function in Ω is known. In particular, if u is harmonic, then

$$u(\mathbf{x}) = - \int_{\partial\Omega} g(\boldsymbol{\sigma}) \partial_\nu G(\mathbf{x}, \boldsymbol{\sigma}) d\sigma. \quad (3.83)$$

Comparing with (3.43), we deduce that

$$-\partial_\nu G(\mathbf{x}, \boldsymbol{\sigma}) d\sigma$$

represents the *harmonic measure* in Ω . The function

$$P(\mathbf{x}, \boldsymbol{\sigma}) = -\partial_\nu G(\mathbf{x}, \boldsymbol{\sigma})$$

is called **Poisson's kernel**. Since $G(\mathbf{x}, \cdot) > 0$ inside Ω and vanishes on $\partial\Omega$, P is *nonnegative* (actually positive).

On the other hand, the formula

$$u(\mathbf{x}) = - \int_{\Omega} f(\mathbf{y}) G(\mathbf{x}, \mathbf{y}) d\mathbf{y} \quad (3.84)$$

gives the solution of the Poisson equation $\Delta u = f$ in Ω , vanishing on $\partial\Omega$. From the positivity of G we have that:

$$f \geq 0 \quad \text{in } \Omega \text{ implies } u \leq 0 \text{ in } \Omega,$$

which is another form of the maximum principle.

- *Poisson's kernel and Poisson's formula* ($n = 3$). From (3.78) we can compute Poisson's kernel for the sphere $B_R(\mathbf{0})$ in \mathbb{R}^3 . We have, recalling that $\mathbf{x}^* = R^2 |\mathbf{x}|^{-2} \mathbf{x}$, if $\mathbf{x} \neq \mathbf{0}$,

$$\nabla_{\mathbf{y}} \left[\frac{1}{|\mathbf{x} - \mathbf{y}|} - \frac{R}{|\mathbf{x}| |\mathbf{x}^* - \mathbf{y}|} \right] = \frac{\mathbf{x} - \mathbf{y}}{|\mathbf{x} - \mathbf{y}|^3} - \frac{R}{|\mathbf{x}|} \frac{\mathbf{x}^* - \mathbf{y}}{|\mathbf{x}^* - \mathbf{y}|^3}.$$

If $\boldsymbol{\sigma} \in \partial B_R(\mathbf{0})$, from (3.75) and (3.77), we have $|\mathbf{x}^* - \boldsymbol{\sigma}| = R |\mathbf{x}|^{-1} |\mathbf{x} - \boldsymbol{\sigma}|$, therefore

$$\nabla_{\mathbf{y}} G(\mathbf{x}, \boldsymbol{\sigma}) = \frac{1}{4\pi} \left[\frac{\mathbf{x} - \boldsymbol{\sigma}}{|\mathbf{x} - \boldsymbol{\sigma}|^3} - \frac{|\mathbf{x}|^2}{R^2} \frac{\mathbf{x}^* - \boldsymbol{\sigma}}{|\mathbf{x} - \boldsymbol{\sigma}|^3} \right] = - \frac{\boldsymbol{\sigma}}{4\pi |\mathbf{x} - \boldsymbol{\sigma}|^3} \left[1 - \frac{|\mathbf{x}|^2}{R^2} \right].$$

Since on $\partial B_R(\mathbf{0})$ the exterior unit normal is $\boldsymbol{\nu}(\boldsymbol{\sigma}) = \boldsymbol{\sigma}/R$, we have

$$P(\mathbf{x}, \boldsymbol{\sigma}) = -\partial_{\boldsymbol{\nu}} G(\mathbf{x}, \boldsymbol{\sigma}) = -\nabla_{\mathbf{y}} G(\mathbf{x}, \boldsymbol{\sigma}) \cdot \boldsymbol{\nu}(\boldsymbol{\sigma}) = \frac{R^2 - |\mathbf{x}|^2}{4\pi R} \frac{1}{|\mathbf{x} - \boldsymbol{\sigma}|^3}.$$

As a consequence, we obtain Poisson's formula

$$u(\mathbf{x}) = \frac{R^2 - |\mathbf{x}|^2}{4\pi R} \int_{\partial B_R(\mathbf{0})} \frac{g(\boldsymbol{\sigma})}{|\mathbf{x} - \boldsymbol{\sigma}|^3} d\sigma \quad (3.85)$$

for the unique solution of the Dirichlet problem $\Delta u = 0$ in $B_R(\mathbf{0})$ and $u = g$ on $\partial B_R(\mathbf{0})$.

3.7.4 The Neumann function

We can find a representation formula for the solution of a Neumann problem as well. Let u be a smooth solution of the problem

$$\begin{cases} \Delta u = f & \text{in } \Omega \\ \partial_{\boldsymbol{\nu}} u = h & \text{on } \partial\Omega, \end{cases} \quad (3.86)$$

where f and h have to satisfy the solvability condition

$$\int_{\partial\Omega} h(\sigma) d\sigma = \int_{\Omega} f(y) dy, \quad (3.87)$$

keeping in mind that u is uniquely determined up to an additive constant. From Theorem 3.36 we can write

$$\begin{aligned} u(x) &= - \int_{\Omega} \Phi(x-y) f(y) dy + \\ &+ \int_{\partial\Omega} h(\sigma) \Phi(x-\sigma) d\sigma - \int_{\partial\Omega} u(\sigma) \partial_{\nu} \Phi(x-\sigma) d\sigma \end{aligned} \quad (3.88)$$

and this time we should get rid of the second integral, containing the unknown datum u on $\partial\Omega$. Mimicking what we have done for the Dirichlet problem, we try to find an analog of the Green's function, that is a function $N = N(x, y)$ given by

$$N(x, y) = \Phi(x-y) - \psi(x, y),$$

where, for x fixed, ψ is a solution of

$$\begin{cases} \Delta_y \psi = 0 & \text{in } \Omega \\ \partial_{\nu} \psi(x, \sigma) = \partial_{\nu} \Phi(x-\sigma) & \text{on } \partial\Omega, \end{cases}$$

in order to have $\partial_{\nu} N(x, \sigma) = 0$ on $\partial\Omega$. But this Neumann problem has no solution, because the solvability condition

$$\int_{\partial\Omega} \partial_{\nu} \Phi(x-\sigma) d\sigma = 0$$

is not satisfied. In fact, letting $u \equiv 1$ in (3.65), we get

$$\int_{\partial\Omega} \partial_{\nu} \Phi(x-\sigma) d\sigma = -1. \quad (3.89)$$

Thus, taking into account (3.89), we require ψ to satisfy

$$\begin{cases} \Delta_y \psi = 0 & \text{in } \Omega \\ \partial_{\nu} \psi(x, \sigma) = \partial_{\nu} \Phi(x-\sigma) + \frac{1}{|\partial\Omega|} & \text{on } \partial\Omega. \end{cases} \quad (3.90)$$

In this way,

$$\int_{\partial\Omega} \left(\partial_{\nu} \Phi(x-\sigma) + \frac{1}{|\partial\Omega|} \right) d\sigma = 0$$

and (3.90) is solvable. Note that, with this choice of ψ , we have

$$\partial_{\nu} N(x, \sigma) = -\frac{1}{|\partial\Omega|} \quad \text{on } \partial\Omega. \quad (3.91)$$

Applying now (3.66) to u and $\psi(\mathbf{x}, \cdot)$, we find:

$$0 = \int_{\Omega} \psi(\mathbf{y}) f(\mathbf{y}) d\mathbf{y} - \int_{\partial\Omega} \psi(\mathbf{x}, \boldsymbol{\sigma}) h(\boldsymbol{\sigma}) d\sigma + \int_{\partial\Omega} u(\boldsymbol{\sigma}) \partial_{\boldsymbol{\nu}} \psi(\boldsymbol{\sigma}) d\sigma. \quad (3.92)$$

Adding (3.92) to (3.88) and using (3.91) we obtain:

Theorem 3.39. *Let Ω be a smooth domain and u be a smooth solution of (3.86). Then:*

$$u(\mathbf{x}) - \frac{1}{|\partial\Omega|} \int_{\partial\Omega} u(\boldsymbol{\sigma}) d\sigma = \int_{\partial\Omega} h(\boldsymbol{\sigma}) N(\mathbf{x}, \boldsymbol{\sigma}) d\sigma - \int_{\Omega} f(\mathbf{y}) N(\mathbf{x}, \mathbf{y}) d\mathbf{y}.$$

Thus, the solution of the Neumann problem (3.86) can also be written as the sum of two potentials, up to the additive constant $c = \frac{1}{|\partial\Omega|} \int_{\partial\Omega} u(\boldsymbol{\sigma}) d\sigma$, the mean value of u .

The function N is called *Neumann function* (also Green's function for the Neumann problem) and it is defined up to an additive constant.

3.8 Uniqueness in Unbounded Domains

3.8.1 Exterior problems

Boundary value problems in unbounded domains occur in important applications, for instance in the motion of fluids past an obstacle, capacity problems or scattering of acoustic or electromagnetic waves.

As in the case of Poisson's equation in all \mathbb{R}^n , a problem in an unbounded domain requires suitable conditions at infinity to be well posed. Consider for example the Dirichlet problem

$$\begin{cases} \Delta u = 0 & \text{in } |\mathbf{x}| > 1 \\ u = 0 & \text{on } |\mathbf{x}| = 1. \end{cases} \quad (3.93)$$

For every real number a ,

$$u(\mathbf{x}) = a \log |\mathbf{x}| \quad \text{and} \quad u(\mathbf{x}) = a(1 - 1/|\mathbf{x}|)$$

are solutions to (3.93) in dimension two and three, respectively. Thus there is no uniqueness.

To restore uniqueness, a typical requirement in two dimensions is that u be bounded, while in three dimensions is that $u(\mathbf{x})$ has a (finite) limit²⁸, say u_{∞} , as $|\mathbf{x}| \rightarrow \infty$. Under these conditions, in both cases we select a unique solution.

The problem (3.93) is an *exterior Dirichlet problem*. Given a bounded domain Ω , we call *exterior of Ω* the set $\Omega_e = \mathbb{R}^n \setminus \overline{\Omega}$.

²⁸ That is: $\forall \varepsilon > 0$, $|u(\mathbf{x}) - u_{\infty}| < \varepsilon$ if $|\mathbf{x}| > N_{\varepsilon}$.

Without loss of generality, we will assume that $\mathbf{0} \in \Omega$ and, for simplicity, we will consider only *connected* exterior sets, i.e. **exterior domains**. Note that $\partial\Omega_e = \partial\Omega$.

As we have seen in several occasions, maximum principles are very useful to prove uniqueness. In exterior three dimensional domains we have:

Theorem 3.40. *Let $\Omega_e \subset \mathbb{R}^3$ be an exterior domain and $u \in C^2(\Omega_e) \cap C(\overline{\Omega}_e)$, be harmonic in Ω_e and vanishing as $|\mathbf{x}| \rightarrow \infty$. If $u \geq 0$ (resp. $u \leq 0$) on $\partial\Omega_e$ then $u \geq 0$ (resp. $u \leq 0$) in Ω_e .*

Proof. Let $u \geq 0$ on $\partial\Omega_e$. Fix $\varepsilon > 0$ and choose r_0 so large that $\Omega \subset \{|\mathbf{x}| < r\}$ and $u \geq -\varepsilon$ on $\{|\mathbf{x}| = r\}$, for every $r \geq r_0$. This is possible since $u(\mathbf{x}) \rightarrow 0$ as $|\mathbf{x}| \rightarrow +\infty$. Applying Theorem 3.7, p. 124, in the bounded set

$$\Omega_{e,r} = \Omega_e \cap \{|\mathbf{x}| < r\},$$

we get $u \geq -\varepsilon$ in $\Omega_{e,r}$. Since ε is arbitrary and r may be taken as large as we like, we deduce that $u \geq 0$ in Ω_e .

The argument for the case $u \leq 0$ on $\partial\Omega_e$ is similar and we leave the details to the reader. \square

An immediate consequence is the following uniqueness result in dimension $n = 3$ (for $n = 2$, see Problem 3.16):

Theorem 3.41. *Let $\Omega_e \subset \mathbb{R}^3$ be an exterior domain. Then there exists at most one solution $u \in C^2(\Omega_e) \cap C(\overline{\Omega}_e)$ of the Dirichlet problem*

$$\begin{cases} \Delta u = f & \text{in } \Omega_e \\ u = g & \text{on } \partial\Omega_e \\ u(\mathbf{x}) \rightarrow u_\infty & \text{as } |\mathbf{x}| \rightarrow \infty. \end{cases}$$

Proof. Apply Theorem 3.40 to the difference of two solutions. \square

We point out another interesting consequence of Theorem 3.40 and Lemma 3.16, p. 133: a harmonic function vanishing at infinity, for $|\mathbf{x}|$ large, is controlled by the fundamental solution. Actually, more is true:

Theorem 3.42. *Let u be harmonic in $\Omega_e \subset \mathbb{R}^3$ and $u(\mathbf{x}) \rightarrow 0$ as $|\mathbf{x}| \rightarrow \infty$. There exists a , such that, if $|\mathbf{x}| \geq 2a$,*

$$|u(\mathbf{x})| \leq \frac{a}{|\mathbf{x}|}, \quad |u_{x_j}(\mathbf{x})| \leq \frac{9a}{|\mathbf{x}|^2}, \quad |u_{x_j x_k}(\mathbf{x})| \leq \frac{81a}{|\mathbf{x}|^3}. \quad (3.94)$$

The proof shows how the constant a depends on u .

Proof. Choose $a \gg 1$ such that $\Omega \subset \{|\mathbf{x}| \leq a\}$ and $|u(\mathbf{x})| \leq 1$ if $|\mathbf{x}| \geq a$. Let $w(\mathbf{x}) = u(\mathbf{x}) - a/|\mathbf{x}|$. Then w is harmonic for $|\mathbf{x}| \geq a$, $w(\mathbf{x}) \leq 0$ on $|\mathbf{x}| = a$ and vanishes at infinity. Then, by Theorem 3.7, we deduce that

$$w(\mathbf{x}) \leq 0 \quad \text{for } \{|\mathbf{x}| \geq a\}. \quad (3.95)$$

Setting $v(\mathbf{x}) = a/|\mathbf{x}| - u(\mathbf{x})$, a similar argument gives $v(\mathbf{x}) \geq 0$ for $\{|\mathbf{x}| \geq a\}$. This and (3.95) implies $|u(\mathbf{x})| \leq a/|\mathbf{x}|$ in $\{|\mathbf{x}| \geq a\}$.

The gradient bound follows from (3.36). In fact, let $|\mathbf{x}| \geq 2a$. We can always find an integer $m \geq 2$ such that $ma \leq |\mathbf{x}| \leq (m+1)a$. From (3.35), we have

$$|u_{x_j}(\mathbf{x})| \leq \frac{3}{(m-1)a} \max_{\partial B_{(m-1)a}(\mathbf{x})} |u|.$$

On the other hand, since $\partial B_{(m-1)a}(\mathbf{x}) \subset \{|\mathbf{x}| \geq a\}$, we know that

$$\max_{\partial B_{(m-1)a}(\mathbf{x})} |u| \leq \frac{a}{|\mathbf{x}|}$$

and since $m \geq 2$, we have $m-1 \geq (m+1)/3 \geq |\mathbf{x}|/3a$. Thus

$$|u_{x_j}(\mathbf{x})| \leq \frac{3}{(m-1)} \frac{1}{|\mathbf{x}|} \leq \frac{9a}{|\mathbf{x}|^2}.$$

Similarly we can prove

$$|u_{x_j x_k}(\mathbf{x})| \leq \frac{81a}{|\mathbf{x}|^3}. \quad \square$$

The estimates (3.94) ensures the validity of the Green identity

$$\int_{\Omega_e} \nabla u \cdot \nabla v \, d\mathbf{x} = \int_{\partial\Omega_e} v \partial_\nu u \, d\sigma \quad (3.96)$$

for any pair $u, v \in C^2(\Omega_e) \cap C^1(\overline{\Omega}_e)$, harmonic in Ω_e and vanishing at infinity. To see this, apply the identity (3.96) in the bounded domain $\Omega_{e,r} = \Omega_e \cap \{|\mathbf{x}| < r\}$. Then, let $r \rightarrow \infty$ to get (3.96) in Ω_e .

In turn, via the Green identity (3.96), we can prove an appropriate version of Theorem 3.41 for the exterior Robin/Neumann problem

$$\begin{cases} \Delta u = f & \text{in } \Omega_e \\ \partial_\nu u + ku = g & \text{on } \partial\Omega_e, (k \geq 0) \\ u \rightarrow u_\infty & \text{as } |\mathbf{x}| \rightarrow \infty. \end{cases} \quad (3.97)$$

Observe that $k \neq 0$ corresponds to the Robin problem while $k = 0$ corresponds to the Neumann problem.

Theorem 3.43. *Let $\Omega_e \subset \mathbb{R}^3$ be an exterior domain. Then there exists at most one solution $u \in C^2(\Omega_e) \cap C^1(\overline{\Omega}_e)$ of problem (3.97).*

Proof. Suppose u, v are solutions of (3.97) and let $w = u - v$. Then w is harmonic in Ω_e , $\partial_\nu w + kw = 0$ on $\partial\Omega_e$ and $w \rightarrow 0$ as $|\mathbf{x}| \rightarrow \infty$.

Apply the identity (3.96) with $u = v = w$. Since $\partial_\nu w = -kw$ on $\partial\Omega_e$ we have:

$$\int_{\Omega_e} |\nabla w|^2 \, d\mathbf{x} = \int_{\partial\Omega_e} w \partial_\nu w \, d\sigma = - \int_{\partial\Omega_e} kw^2 \, d\sigma \leq 0.$$

Thus $\nabla w = 0$ and w is constant, because Ω_e is connected. But w vanishes at infinity so that $w = 0$. \square

3.9 Surface Potentials

In this section we go back to examine the meaning and the main properties of the surface potentials appearing in the identity (3.65). A remarkable consequence is the possibility to convert a boundary value problem into a **boundary integral equation**. This kind of formulation can be obtained for more general operators and more general problems, as soon as a fundamental solution is known. Thus it constitutes a flexible method with important implications. In particular, it constitutes the theoretical basis for the so called *boundary element method*, which may offer several advantages from the point of view of the computational cost in numerical approximations, due to a dimension reduction. Here we present the integral formulations of the main boundary value problems and state some basic results²⁹.

3.9.1 The double and single layer potentials

The last integral in (3.65) is of the form

$$\mathcal{D}(\mathbf{x};\mu) = \int_{\partial\Omega} \mu(\boldsymbol{\sigma}) \partial_\nu \Phi(\mathbf{x} - \boldsymbol{\sigma}) d\sigma \quad (3.98)$$

and it is called the **double layer potential** of μ . In three dimensions it may represent the electrostatic potential generated by a *dipole* distribution³⁰ of moment $\mu/4\pi$ on $\partial\Omega$.

To get a clue of the main features of $\mathcal{D}(\mathbf{x};\mu)$, it is useful to look at the particular case $\mu(\boldsymbol{\sigma}) \equiv 1$, that is

$$\mathcal{D}(\mathbf{x};1) = \int_{\partial\Omega} \partial_\nu \Phi(\mathbf{x} - \boldsymbol{\sigma}) d\sigma. \quad (3.99)$$

Inserting $u \equiv 1$ into (3.65), we get

$$\mathcal{D}(\mathbf{x};1) = -1, \quad \text{for every } \mathbf{x} \in \Omega. \quad (3.100)$$

On the other hand, if $\mathbf{x} \in \mathbb{R}^n \setminus \overline{\Omega}$ is fixed, $\Phi(\mathbf{x} - \cdot)$ is harmonic in Ω and can be inserted into (3.66), with $u \equiv 1$; the result is

$$\mathcal{D}(\mathbf{x};1) = 0, \quad \text{for every } \mathbf{x} \in \mathbb{R}^n \setminus \overline{\Omega}. \quad (3.101)$$

²⁹ See e.g. [23], R. Courant and D. Hilbert, 1953, or [5], R.B. Guenter and J.W. Lee, 1998, for the proofs.

³⁰ For every $\boldsymbol{\sigma} \in \partial\Omega$, let $-q(\boldsymbol{\sigma})$, $q(\boldsymbol{\sigma})$ two charges placed at the points $\boldsymbol{\sigma}$, $\boldsymbol{\sigma} + h\nu(\boldsymbol{\sigma})$, respectively. If $h > 0$ is very small, the pair of charges constitutes a *dipole of axis* $\nu(\boldsymbol{\sigma})$. The induced potential $u_h(\mathbf{x}, \boldsymbol{\sigma})$ at \mathbf{x} is given by, setting $4\pi q(\boldsymbol{\sigma})h = \mu(\boldsymbol{\sigma})$,

$$u(\mathbf{x}, \boldsymbol{\sigma}) = 4\pi q(\boldsymbol{\sigma}) [\Phi(\mathbf{x} - (\boldsymbol{\sigma} + h\nu)) - \Phi(\mathbf{x} - \boldsymbol{\sigma})] = \mu(\boldsymbol{\sigma}) \left[\frac{\Phi(\mathbf{x} - (\boldsymbol{\sigma} + h\nu)) - \Phi(\mathbf{x} - \boldsymbol{\sigma})}{h} \right].$$

Since h is very small, we can write, at first order of approximation, $u_h(\mathbf{x}) \simeq \mu(\boldsymbol{\sigma}) \partial_\nu \Phi(\mathbf{x} - \boldsymbol{\sigma})$. Integrating on $\partial\Omega$ we obtain $\mathcal{D}(\mathbf{x};\mu)$.

What happens for $\mathbf{x} \in \partial\Omega$? First of all we have to check that $\mathcal{D}(\mathbf{x}; 1)$ is well defined (i.e. finite) on $\partial\Omega$. Indeed the singularity of $\partial_\nu \Phi(\mathbf{x} - \boldsymbol{\sigma})$ becomes critical when $\mathbf{x} \in \partial\Omega$, since as $\boldsymbol{\sigma} \rightarrow \mathbf{x}$, the order of infinity equals the topological dimension of $\partial\Omega$. For instance, in the two dimensional case, we have

$$\mathcal{D}(\mathbf{x}; 1) = -\frac{1}{2\pi} \int_{\partial\Omega} \partial_\nu \log |\mathbf{x} - \boldsymbol{\sigma}| d\sigma = -\frac{1}{2\pi} \int_{\partial\Omega} \frac{(\mathbf{x} - \boldsymbol{\sigma}) \cdot \nu(\boldsymbol{\sigma})}{|\mathbf{x} - \boldsymbol{\sigma}|^2} d\sigma.$$

The order of infinity of the integrand is one and the boundary $\partial\Omega$ is a curve, a one dimensional object. In the three dimensional case we have

$$\mathcal{D}(\mathbf{x}; 1) = \frac{1}{4\pi} \int_{\partial\Omega} \frac{\partial}{\partial \nu} \frac{1}{|\mathbf{x} - \boldsymbol{\sigma}|} d\sigma = \frac{1}{4\pi} \int_{\partial\Omega} \frac{(\mathbf{x} - \boldsymbol{\sigma}) \cdot \nu(\boldsymbol{\sigma})}{|\mathbf{x} - \boldsymbol{\sigma}|^3} d\sigma.$$

The order of infinity of the integrand is two and $\partial\Omega$ is a bidimensional surface.

However, if we assume that Ω is a C^2 domain, then it can be proved that $\mathcal{D}(\mathbf{x}; 1)$ is well defined and finite on $\partial\Omega$.

To compute the value of $\mathcal{D}(\mathbf{x}; 1)$ on $\partial\Omega$, we first observe that the formulas (3.100) and (3.101) follow immediately from the geometric interpretation of the integrand in $\mathcal{D}(\mathbf{x}; 1)$. Precisely, in dimension two, consider the quantity

$$d\sigma^* = -\frac{(\mathbf{x} - \boldsymbol{\sigma}) \cdot \nu(\boldsymbol{\sigma})}{r_{x\sigma}^2} d\sigma = \frac{(\boldsymbol{\sigma} - \mathbf{x}) \cdot \nu(\boldsymbol{\sigma})}{r_{x\sigma}^2} d\sigma$$

where we keep the notation $r_{x\sigma} = |\mathbf{x} - \boldsymbol{\sigma}|$. We have (see Fig. 3.10),

$$\frac{(\boldsymbol{\sigma} - \mathbf{x}) \cdot \nu(\boldsymbol{\sigma})}{r_{x\sigma}} = \cos \varphi$$

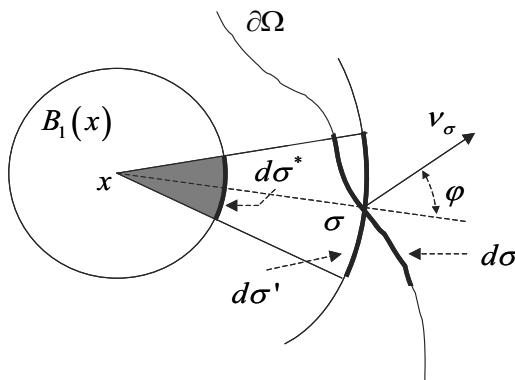


Fig. 3.10 Geometrical interpretation of the integrand in $\mathcal{D}(\mathbf{x}, 1)$, $n = 2$

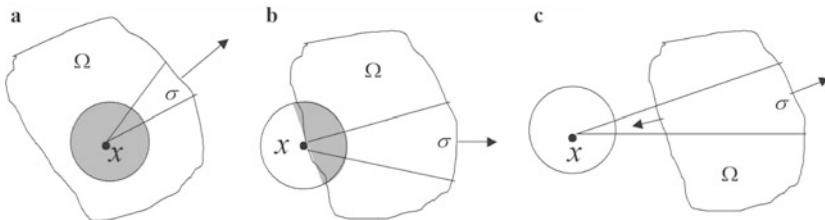


Fig. 3.11 Values of $\int_{\partial\Omega} d\sigma^*$ for $n = 2$. (a) 2π ; (b) π ; (c) 0

and therefore

$$d\sigma' = \frac{(\boldsymbol{\sigma} - \mathbf{x}) \cdot \boldsymbol{\nu}(\boldsymbol{\sigma})}{r_{x\sigma}} d\sigma = \cos \varphi \, d\sigma$$

is the projection of the length element $d\sigma$ on the circle $\partial B_{r_{x\sigma}}(\mathbf{x})$, up to an error of lower order. Then, $d\sigma^* = \frac{d\sigma'}{r_{x\sigma}}$ is the projection of $d\sigma$ on $\partial B_1(\mathbf{x})$.

Integrating on $\partial\Omega$, the contributions to $d\sigma^*$ sum up to 2π if $\mathbf{x} \in \Omega$ (case a) of Fig. 3.11 and to 0 if $\mathbf{x} \in \mathbb{R}^n \setminus \overline{\Omega}$, due to the sign compensations induced by the orientation of $\boldsymbol{\nu}(\boldsymbol{\sigma})$ (case c) of Fig. 3.11). Thus

$$\int_{\partial\Omega} d\sigma^* = \begin{cases} 2\pi & \text{if } \mathbf{x} \in \Omega \\ 0 & \text{if } \mathbf{x} \in \mathbb{R}^n \setminus \overline{\Omega}, \end{cases}$$

which are equivalent to (3.100) and (3.101), since

$$\mathcal{D}(\mathbf{x}; 1) = -\frac{1}{2\pi} \int_{\partial\Omega} d\sigma^*.$$

The case b) in Fig. 3.11 corresponds to $\mathbf{x} \in \partial\Omega$. It should be now intuitively clear that the point \mathbf{x} “sees” a total angle of only π radians and therefore $\mathcal{D}(\mathbf{x}; 1) = -1/2$.

The same kind of considerations hold in dimension three; this time, the quantity (see Fig. 3.12)

$$d\sigma^* = -\frac{(\mathbf{x} - \boldsymbol{\sigma}) \cdot \boldsymbol{\nu}(\boldsymbol{\sigma})}{r_{x\sigma}^3} d\sigma = \frac{(\boldsymbol{\sigma} - \mathbf{x}) \cdot \boldsymbol{\nu}(\boldsymbol{\sigma})}{r_{x\sigma}^3} d\sigma$$

is the projection on $\partial B_1(\mathbf{x})$ (*solid angle*) of the surface element $d\sigma$. Integrating over $\partial\Omega$, the contributions to $d\sigma^*$ sum up to 4π if $\mathbf{x} \in \Omega$ and to 0 if $\mathbf{x} \in \mathbb{R}^3 \setminus \overline{\Omega}$.

If $\mathbf{x} \in \partial\Omega$, \mathbf{x} “sees” a total solid angle of measure 2π . Since

$$\mathcal{D}(\mathbf{x}; 1) = -\frac{1}{4\pi} \int_{\partial\Omega} d\sigma^*,$$

we find again the values $-1, 0, -1/2$ in the three cases, respectively.

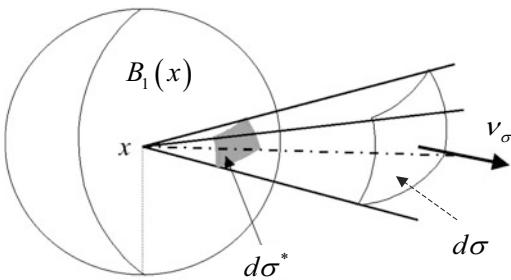


Fig. 3.12 The solid angle $d\sigma^*$, projected from $d\sigma$

We gather the above results in the following Gauss Lemma.

Lemma 3.44. *Let $\Omega \subset \mathbb{R}^n$ be a bounded, C^2 -domain. Then*

$$\mathcal{D}(\mathbf{x}; 1) = \int_{\partial\Omega} \partial_\nu \Phi(\mathbf{x} - \boldsymbol{\sigma}) d\sigma = \begin{cases} -1 & \mathbf{x} \in \Omega \\ -\frac{1}{2} & \mathbf{x} \in \partial\Omega \\ 0 & \mathbf{x} \in \mathbb{R}^n \setminus \overline{\Omega}. \end{cases}$$

Thus, when $\mu \equiv 1$, the double layer potential has a jump discontinuity across $\partial\Omega$. Indeed, if $\mathbf{x} \in \partial\Omega$,

$$\lim_{\mathbf{z} \rightarrow \mathbf{x}, \mathbf{z} \in \mathbb{R}^n \setminus \overline{\Omega}} \mathcal{D}(\mathbf{z}; 1) = \mathcal{D}(\mathbf{x}; 1) + \frac{1}{2}$$

and

$$\lim_{\mathbf{z} \rightarrow \mathbf{x}, \mathbf{z} \in \Omega} \mathcal{D}(\mathbf{z}; 1) = \mathcal{D}(\mathbf{x}; 1) - \frac{1}{2}.$$

These formulas are the key for understanding the general properties of $\mathcal{D}(\mathbf{x}; \mu)$, that we state in the following theorem.

Theorem 3.45. *Let $\Omega \subset \mathbb{R}^n$ be a bounded, C^2 domain and μ a continuous function on $\partial\Omega$. Then, $D(\mathbf{x}; \mu)$ is harmonic in $\mathbb{R}^n \setminus \partial\Omega$ and the following jump relations hold for every $\mathbf{z} \in \partial\Omega$:*

$$\lim_{\mathbf{x} \rightarrow \mathbf{z}, \mathbf{x} \in \mathbb{R}^n \setminus \overline{\Omega}} \mathcal{D}(\mathbf{x}; \mu) = \mathcal{D}(\mathbf{z}; \mu) + \frac{1}{2}\mu(\mathbf{z}) \quad (3.102)$$

and

$$\lim_{\mathbf{x} \rightarrow \mathbf{z}, \mathbf{x} \in \Omega} \mathcal{D}(\mathbf{x}; \mu) = \mathcal{D}(\mathbf{z}; \mu) - \frac{1}{2}\mu(\mathbf{z}). \quad (3.103)$$

Proof (Sketch). If $\mathbf{x} \notin \partial\Omega$ there is no problem in differentiating under the integral sign and, for $\boldsymbol{\sigma}$ fixed on $\partial\Omega$, the function

$$\mathbf{x} \mapsto \partial_\nu \Phi(\mathbf{x} - \boldsymbol{\sigma}) = \nabla_y \Phi(\mathbf{x} - \boldsymbol{\sigma}) \cdot \boldsymbol{\nu}(\boldsymbol{\sigma})$$

is harmonic. Thus $\mathcal{D}(\mathbf{x}; \mu)$ is harmonic in $\mathbb{R}^n \setminus \partial\Omega$.

Consider (3.102). This is not an elementary formula and we cannot take the limit under the integral sign, once more because of the critical singularity of $\partial_\nu \Phi(\mathbf{z} - \boldsymbol{\sigma})$ when $\mathbf{z} \in \partial\Omega$.

Let $\mathbf{x} \in \mathbb{R}^n \setminus \overline{\Omega}$. From Gauss Lemma 3.44, we have $\int_{\partial\Omega} \partial_\nu \Phi(\mathbf{x} - \boldsymbol{\sigma}) d\sigma = 0$ and therefore we can write

$$\mathcal{D}(\mathbf{x}; \mu) = \int_{\partial\Omega} \partial_\nu \Phi(\mathbf{x} - \boldsymbol{\sigma}) [\mu(\boldsymbol{\sigma}) - \mu(\mathbf{z})] d\sigma. \quad (3.104)$$

Now, when $\boldsymbol{\sigma}$ is near \mathbf{z} , the smoothness of $\partial\Omega$ and the continuity of μ mitigate the singularity of $\partial_\nu \Phi(\mathbf{x} - \boldsymbol{\sigma})$ and allows to take the limit under the integral sign. Thus

$$\lim_{\mathbf{x} \rightarrow \mathbf{z}} \int_{\partial\Omega} \partial_\nu \Phi(\mathbf{x} - \boldsymbol{\sigma}) [\mu(\boldsymbol{\sigma}) - \mu(\mathbf{z})] d\sigma = \int_{\partial\Omega} \partial_\nu \Phi(\mathbf{z} - \boldsymbol{\sigma}) [\mu(\boldsymbol{\sigma}) - \mu(\mathbf{z})] d\sigma.$$

Exploiting once more Gauss Lemma 3.44, we have

$$= \int_{\partial\Omega} \partial_\nu \Phi(\mathbf{z} - \boldsymbol{\sigma}) \mu(\boldsymbol{\sigma}) d\sigma - \mu(\mathbf{z}) \int_{\partial\Omega} \partial_\nu \Phi(\mathbf{z} - \boldsymbol{\sigma}) d\sigma = \mathcal{D}(\mathbf{z}; \mu) + \frac{1}{2} \mu(\mathbf{z}).$$

The proof of (3.103) is similar. \square

The second integral in (3.65) is of the form

$$\mathcal{S}(\mathbf{x}; \psi) = \int_{\partial\Omega} \Phi(\mathbf{x} - \boldsymbol{\sigma}) \psi(\boldsymbol{\sigma}) d\sigma$$

and it is called the **single layer potential of ψ** .

In three dimensions it represents the electrostatic potential generated by a charge distribution of density $\varphi/4\pi$ on $\partial\Omega$. If Ω is a C^2 -domain and ψ is continuous on $\partial\Omega$, then \mathcal{S} is *continuous across $\partial\Omega$* and

$$\Delta \mathcal{S} = 0 \quad \text{in } \mathbb{R}^n \setminus \partial\Omega,$$

because there is no problem in differentiating under the integral sign.

Since the flux of an electrostatic potential undergoes a jump discontinuity across a charged surface, we expect a jump discontinuity of the normal derivative of \mathcal{S} across $\partial\Omega$. Precisely

Theorem 3.46. *Let $\Omega \subset \mathbb{R}^n$ be a bounded, C^2 -domain and ψ a continuous function on $\partial\Omega$. Then, $\mathcal{S}(\mathbf{x}; \psi)$ is harmonic in $\mathbb{R}^n \setminus \partial\Omega$, continuous across $\partial\Omega$ and the following jump relations hold for every $\mathbf{z} \in \partial\Omega$, where $\mathbf{x}_h = \mathbf{z} + h\boldsymbol{\nu}(\mathbf{z})$:*

$$\lim_{h \rightarrow 0^+} \nabla \mathcal{S}(\mathbf{x}_h; \psi) \cdot \boldsymbol{\nu}(\mathbf{z}) = \int_{\partial\Omega} \nabla_{\mathbf{z}} \Phi(\mathbf{z} - \boldsymbol{\sigma}) \cdot \boldsymbol{\nu}(\mathbf{z}) \psi(\boldsymbol{\sigma}) d\sigma - \frac{1}{2} \psi(\mathbf{z}) \quad (3.105)$$

and

$$\lim_{h \rightarrow 0^-} \nabla \mathcal{S}(\mathbf{x}_h; \psi) \cdot \boldsymbol{\nu}(\mathbf{z}) = \int_{\partial\Omega} \nabla_{\mathbf{z}} \Phi(\mathbf{z} - \boldsymbol{\sigma}) \cdot \boldsymbol{\nu}(\mathbf{z}) \psi(\boldsymbol{\sigma}) d\sigma + \frac{1}{2} \psi(\mathbf{z}). \quad (3.106)$$

3.9.2 The integral equations of potential theory

By means of the jump relations (3.103)–(3.106) of the double and single layer potentials, we can reduce the main boundary value problems of potential theory into integral equations of a special form. Let $\Omega \subset \mathbb{R}^n$ be a smooth domain and $g \in C(\partial\Omega)$. We first show the reduction procedure for the *interior Dirichlet problem*

$$\begin{cases} \Delta u = 0 & \text{in } \Omega \\ u = g & \text{on } \partial\Omega. \end{cases} \quad (3.107)$$

The starting point is once more the identity (3.65), which gives for the solution u of (3.107) the representation

$$u(\mathbf{x}) = \int_{\partial\Omega} \Phi(\mathbf{x} - \boldsymbol{\sigma}) \partial_\nu u(\boldsymbol{\sigma}) d\sigma - \int_{\partial\Omega} g(\boldsymbol{\sigma}) \partial_\nu \Phi(\mathbf{x} - \boldsymbol{\sigma}) d\sigma.$$

In Sect. 3.5.3 we used the Green function to get rid of the single layer potential containing the unknown $\partial_\nu u$. Here we adopt a different strategy: we forget the single layer potential and try to represent u in the form of a double layer potential, by choosing an appropriate density. In other words, *we seek a continuous function μ on $\partial\Omega$, such that the solution u of (3.107) is given by*

$$u(\mathbf{x}) = \int_{\partial\Omega} \mu(\boldsymbol{\sigma}) \partial_\nu \Phi(\mathbf{x} - \boldsymbol{\sigma}) d\sigma = \mathcal{D}(\mathbf{x}; \mu). \quad (3.108)$$

The function u given by (3.108) is harmonic in Ω , therefore we have only to check the boundary condition

$$\lim_{\mathbf{x} \rightarrow \mathbf{z} \in \partial\Omega} u(\mathbf{x}) = g(\mathbf{z}).$$

Letting $\mathbf{x} \rightarrow \mathbf{z} \in \partial\Omega$, with $\mathbf{x} \in \Omega$, and taking into account the jump relation (3.102), we obtain for μ the integral equation

$$\int_{\partial\Omega} \mu(\boldsymbol{\sigma}) \partial_\nu \Phi(\mathbf{z} - \boldsymbol{\sigma}) d\sigma - \frac{1}{2}\mu(\mathbf{z}) = g(\mathbf{z}), \quad \mathbf{z} \in \partial\Omega. \quad (3.109)$$

If $\mu \in C(\partial\Omega)$ solves (3.109), then (3.108) is the solution of (3.107) in $C^2(\Omega) \cap C(\overline{\Omega})$. The following theorem holds.

Theorem 3.47. *Let $\Omega \subset \mathbb{R}^n$ be a bounded, C^2 domain and g a continuous function on $\partial\Omega$. Then, the integral equation (3.109) has a unique solution $\mu \in C(\partial\Omega)$ and the solution $u \in C^2(\Omega) \cap C(\overline{\Omega})$ of the Dirichlet problem (3.107) can be represented as the double layer potential of μ .*

We consider now the *interior Neumann problem*

$$\begin{cases} \Delta u = 0 & \text{in } \Omega \\ \partial_\nu u = g & \text{on } \partial\Omega, \end{cases} \quad (3.110)$$

where $g \in C(\partial\Omega)$ satisfies the solvability condition

$$\int_{\partial\Omega} g \, d\sigma = 0. \quad (3.111)$$

This time we seek a continuous function ψ on $\partial\Omega$, such that the solution u of (3.110) is given in the form

$$u(\mathbf{x}) = \int_{\partial\Omega} \psi(\boldsymbol{\sigma}) \Phi(\mathbf{x} - \boldsymbol{\sigma}) \, d\sigma = \mathcal{S}(\mathbf{x}; \psi). \quad (3.112)$$

The function u given by (3.112) is harmonic in Ω . We check the boundary condition in the form (see (3.107))

$$\lim_{h \rightarrow 0^-} \nabla u(\mathbf{x}_h) \cdot \boldsymbol{\nu}(\mathbf{z}) = g(\mathbf{z}).$$

Letting $h \rightarrow 0^-$ and taking into account the jump relation (3.106), we obtain for ψ the integral equation

$$\int_{\partial\Omega} \psi(\boldsymbol{\sigma}) \nabla_{\mathbf{z}} \Phi(\mathbf{z} - \boldsymbol{\sigma}) \cdot \boldsymbol{\nu}(\mathbf{z}) \, d\sigma + \frac{1}{2} \psi(\mathbf{z}) = g(\mathbf{z}), \quad \mathbf{z} \in \partial\Omega. \quad (3.113)$$

If $\psi \in C(\partial\Omega)$ solves (3.113), then (3.112) is a solution of (3.110) in the sense specified above.

It turns out that the general solution of (3.113) has the form

$$\psi = \bar{\psi} + C_0 \psi_0 \quad C_0 \in \mathbb{R},$$

where $\bar{\psi}$ is a particular solution of (3.113) and ψ_0 is a solution of the homogeneous equation

$$\int_{\partial\Omega} \psi_0(\boldsymbol{\sigma}) \nabla_{\mathbf{z}} \Phi(\mathbf{z} - \boldsymbol{\sigma}) \cdot \boldsymbol{\nu}(\mathbf{z}) \, d\sigma + \frac{1}{2} \psi_0(\mathbf{z}) = 0, \quad \mathbf{z} \in \partial\Omega. \quad (3.114)$$

As expected, we have infinitely many solutions to the Neumann problem. Observe that

$$\mathcal{S}(\mathbf{x}; \psi_0) = \int_{\partial\Omega} \psi_0(\boldsymbol{\sigma}) \Phi(\mathbf{x} - \boldsymbol{\sigma}) \, d\sigma$$

is harmonic in Ω , with vanishing normal derivative on $\partial\Omega$, because of (3.106) and (3.114). Consequently, $\mathcal{S}(\mathbf{x}; \psi_0)$ is constant and the following theorem holds.

Theorem 3.48. *Let $\Omega \subset \mathbb{R}^n$ be a bounded, C^2 -domain and g be a continuous function on $\partial\Omega$ satisfying (3.111). Then, the Neumann problem (3.110) has infinitely many solutions u of the form*

$$u(\mathbf{x}) = \mathcal{S}(\mathbf{x}; \bar{\psi}) + C,$$

where $\bar{\psi}$ is a particular solution of (3.113) and C is an arbitrary constant.

Another advantage of the method is that, in principle, exterior problems can be treated as the interior problems, with the same level of difficulty³¹. It is enough to use the exterior jump conditions (3.102), (3.105) and proceed in the same way (see Problem 3.18).

As an example of an elementary application of the method we solve the interior Neumann problem for the disc.

- *The Neumann problem for the disc.* Let $B_R = B_R(\mathbf{0}) \subset \mathbb{R}^2$ and consider the Neumann problem

$$\begin{cases} \Delta u = 0 & \text{in } B_R \\ \partial_{\nu} u = g & \text{on } \partial B_R, \end{cases}$$

where $g \in C(\partial B_R)$ satisfies the solvability condition (3.111). We know that u is unique up to an additive constant. We want to express the solution as a single layer potential:

$$u(\mathbf{x}) = -\frac{1}{2\pi} \int_{\partial B_R} \psi(\boldsymbol{\sigma}) \log |\mathbf{x} - \boldsymbol{\sigma}| d\sigma. \quad (3.115)$$

The Neumann condition $\partial_{\nu} u = g$ on ∂B_R translates into the following integral equation for the density ψ :

$$-\frac{1}{2\pi} \int_{\partial B_R} \frac{(\mathbf{z} - \boldsymbol{\sigma}) \cdot \boldsymbol{\nu}(\mathbf{z})}{|\mathbf{z} - \boldsymbol{\sigma}|^2} \psi(\boldsymbol{\sigma}) d\sigma + \frac{1}{2} \psi(\mathbf{z}) = g(\mathbf{z}), \quad \mathbf{z} \in \partial B_R. \quad (3.116)$$

On ∂B_R we have $\boldsymbol{\nu}(\mathbf{z}) = \mathbf{z}/R$, $(\mathbf{z} - \boldsymbol{\sigma}) \cdot \mathbf{z} = R^2 - \mathbf{z} \cdot \boldsymbol{\sigma}$ and moreover, $|\mathbf{z} - \boldsymbol{\sigma}|^2 = 2(R^2 - \mathbf{z} \cdot \boldsymbol{\sigma})$. Hence, (3.116) becomes

$$-\frac{1}{4\pi R} \int_{\partial B_R} \psi(\boldsymbol{\sigma}) d\sigma + \frac{1}{2} \psi(\mathbf{z}) = g(\mathbf{z}), \quad \mathbf{z} \in \partial B_R. \quad (3.117)$$

The solutions of the homogeneous equation ($g = 0$) are constant functions $\psi_0(x) = C$. A particular solution $\bar{\psi}$, with

$$\int_{\partial B_R} \bar{\psi}(\boldsymbol{\sigma}) d\sigma = 0,$$

is given by $\bar{\psi}(\mathbf{z}) = 2g(\mathbf{z})$.

Thus, the general solution of (3.117) is $\psi(\mathbf{z}) = 2g(\mathbf{z}) + C$, where C is an arbitrary constant, and the solution of the Neumann problem is given, up to an additive constant, by

$$u(\mathbf{x}) = -\frac{1}{\pi} \int_{\partial B_R} g(\boldsymbol{\sigma}) \log |\mathbf{x} - \boldsymbol{\sigma}| d\sigma.$$

³¹ Some care is required when solving the exterior Dirichlet problem via a double layer potential, due to the rapid vanishing of its kernel. See e.g. [5], Guenter and Lee, 1998.

Remark 3.49. The integral equations (3.109) and (3.113) are of the form

$$\int_{\partial\Omega} K(\mathbf{z}, \boldsymbol{\sigma}) \rho(\boldsymbol{\sigma}) d\sigma \pm \frac{1}{2} \rho(\mathbf{z}) = g(\mathbf{z}) \quad (3.118)$$

which are called *Fredholm integral equations of the first kind*. Their solution is based on the following so called **Fredholm alternative**: either equation (3.118) has exactly one solution for every $g \in C(\partial\Omega)$, or the homogeneous equation

$$\int_{\partial\Omega} K(\mathbf{z}, \boldsymbol{\sigma}) \phi(\boldsymbol{\sigma}) d\sigma \pm \frac{1}{2} \phi(\mathbf{z}) = 0$$

has a finite number ϕ_1, \dots, ϕ_N of non trivial, linearly independent solutions.

In this last case equation (3.118) is not always solvable and we have:

(a) The **adjoint** homogeneous equation

$$\int_{\partial\Omega} K(\boldsymbol{\sigma}, \mathbf{z}) \phi^*(\boldsymbol{\sigma}) d\sigma \pm \frac{1}{2} \phi^*(\mathbf{z}) = 0$$

has N nontrivial linearly independent solutions $\phi_1^*, \dots, \phi_N^*$.

(b) Equation (3.118) has a solution if and only if g satisfies the following N compatibility conditions:

$$\int_{\partial\Omega} \phi_j^*(\boldsymbol{\sigma}) g(\boldsymbol{\sigma}) d\sigma = 0, \quad j = 1, \dots, N. \quad (3.119)$$

(c) If g satisfies (3.119), the general solution of (3.118) is given by

$$\rho = \bar{\rho} + C_1 \phi_1 + \dots + C_N \phi_N$$

where $\bar{\rho}$ is a particular solution of equation (3.118) and C_1, \dots, C_N are arbitrary real constants.

The analogy with the solution of a system of linear algebraic equations should be evident. We will come back to Fredholm's alternative in Chap. 6.

Problems

3.1. Let B_R be the unit disc centered at $(0, 0)$. Use the method of separation of variables to solve the problem

$$\begin{cases} \Delta u = f & \text{in } B_R \\ u = 1 & \text{on } \partial B_R. \end{cases}$$

Find an explicit formula when $f(x, y) = y$.

[Hint: Use polar coordinates; expand $f = f(r, \cdot)$ in sine Fourier series in $[0, 2\pi]$ and derive a set of ordinary differential equations for the Fourier coefficients of $u(r, \cdot)$].

3.2. Let $C_{1,2} = \{(r, \theta) \in \mathbb{R}^2; 1 < r < 2\}$. Examine the solvability of the Neumann problem

$$\begin{cases} \Delta u = -1 & \text{in } C_{1,2} \\ \partial_\nu u = \cos \theta & \text{on } r = 1 \\ \partial_\nu u = \lambda(\cos \theta)^2 & \text{on } r = 2 \end{cases} \quad (\lambda \in \mathbb{R})$$

and write an explicit formula for the solution, when it exists.

3.3 Schwarz reflection principle. Let

$$B_1^+ = \{(x, y) \in \mathbb{R}^2 : x^2 + y^2 < 1, y > 0\}$$

and $u \in C^2(B_1^+) \cap C(\overline{B}_1^+)$, harmonic B_1^+ , $u(x, 0) = 0$. Show that the function

$$U(x, y) = \begin{cases} u(x, y) & y \geq 0 \\ -u(x, -y) & y < 0, \end{cases}$$

obtained from u by odd reflection with respect to y , is harmonic in all B_1 . State and prove the *Schwarz reflection principle* in dimension three.

[Hint: Let v be the solution of $\Delta v = 0$ in B_1 , $v = U$ on ∂B_1 . Define $w(x, y) = v(x, y) + v(x, -y)$ and show that $w \equiv 0$].

3.4. Let $\Omega \subseteq \mathbb{R}^n$ be a smooth bounded domain. Let $u \in C^2(\Omega) \cap C^1(\overline{\Omega})$ be a solution of the Robin problem

$$\begin{cases} \Delta u = 0, & \text{in } \Omega \\ \partial_\nu u + \alpha u = g & \text{on } \partial\Omega, \end{cases}$$

where $\alpha > 0$ and $g \in C(\partial\Omega)$. Show that if $g > 0$ on $\partial\Omega$, then $u > 0$ in $\overline{\Omega}$.

3.5. Let $\Omega \subseteq \mathbb{R}^n$ be a domain and u be harmonic in Ω . Use the analyticity Theorem 3.17, p. 134, to show that, if u has a **local** maximum or minimum in Ω , then it is constant.

3.6. Let u be harmonic in \mathbb{R}^3 such that

$$\int_{\mathbb{R}^3} |u(\mathbf{x})|^2 d\mathbf{x} < \infty.$$

Show that $u \equiv 0$.

[Hint: Write the mean value formula in a ball $B_R(\mathbf{0})$ for u . Use the Schwarz inequality and let $R \rightarrow +\infty$].

3.7. Let u be harmonic in \mathbb{R}^n and \mathbf{M} an orthogonal matrix of order n . Using the mean value property, show that $v(\mathbf{x}) = u(\mathbf{M}\mathbf{x})$ is harmonic in \mathbb{R}^n .

3.8. Let $u \in C(\Omega)$, $\Omega \subseteq \mathbb{R}^n$. Using the notations of Sect. 3.4.1, prove the following statements:

- a) For all $B_r(\mathbf{x}) \subset \subset \Omega$, $u(\mathbf{x}) \leq S(u; \mathbf{x}, r)$ if and only if $u(\mathbf{x}) \leq A(u; \mathbf{x}, r)$.
- b) If $u \in C(\Omega)$ is subharmonic in Ω then $r \mapsto S(u; \mathbf{x}, r)$ and $r \mapsto A(u; \mathbf{x}, r)$ are nondecreasing functions.
- c) If u is harmonic in Ω then u^2 is subharmonic in Ω .
- d) If $u \in C^2(\Omega)$, u is subharmonic if and only if $\Delta u \geq 0$ in Ω .

Moreover:

- e) Let u be subharmonic in Ω and $F : \mathbb{R} \rightarrow \mathbb{R}$, smooth. Show that if F is convex and increasing, then $F \circ u$ subharmonic.

3.9. Torsion problem. Let $\Omega \subset \mathbb{R}^2$ be a bounded domain and $v \in C^2(\Omega) \cap C^1(\bar{\Omega})$ be a solution of

$$\begin{cases} v_{xx} + v_{yy} = -2 & \text{in } \Omega \\ v = 0 & \text{on } \partial\Omega. \end{cases}$$

Show that $u = |\nabla v|^2$ attains its maximum on $\partial\Omega$.

3.10. Let Ω be a bounded domain and $u \in C(\bar{\Omega})$ be subharmonic and bounded.

- a) Assume that, for every $\mathbf{p} \in \partial\Omega \setminus \{\mathbf{p}_1, \dots, \mathbf{p}_N\}$, $\lim_{\mathbf{x} \rightarrow \mathbf{p}} u(\mathbf{x}) \leq 0$. Prove that $u \leq 0$ in Ω .
- b) Deduce that if u is harmonic and bounded in Ω and $\lim_{\mathbf{x} \rightarrow \mathbf{p}} u(\mathbf{x}) = 0$ for every $\mathbf{p} \in \partial\Omega \setminus \{\mathbf{p}_1, \dots, \mathbf{p}_N\}$, then $u \equiv 0$ in Ω .

[Hint: a) ($n > 2$) Let $w(\mathbf{x}) = u(\mathbf{x}) - \varepsilon \sum_{j=1}^N |\mathbf{x} - \mathbf{p}_j|^{2-n}$, $\varepsilon > 0$. Observe that w is subharmonic in Ω and there exist N balls $B_{\rho_j}(\mathbf{p}_j)$, with ρ_j small, such that $w < 0$ in each $B_{\rho_j}(\mathbf{p}_j)$. Moreover, on the boundary of $\Omega \setminus \cup_{j=1}^N B_{\rho_j}(\mathbf{p}_j)$, $w \leq 0$. Conclude by using the maximum principle].

3.11. Let $f \in C^2(\mathbb{R}^2)$ with compact support K and

$$u(\mathbf{x}) = -\frac{1}{2\pi} \int_{\mathbb{R}^2} \log |\mathbf{x} - \mathbf{y}| f(\mathbf{y}) d\mathbf{y}.$$

Show that

$$u(\mathbf{x}) = -\frac{M}{2\pi} \log |\mathbf{x}| + O(|\mathbf{x}|^{-1}), \quad \text{as } |\mathbf{x}| \rightarrow +\infty$$

where $M = \int_{\mathbb{R}^2} f(\mathbf{y}) d\mathbf{y}$.

[Hint: Write $\log |\mathbf{x} - \mathbf{y}| = \log(|\mathbf{x} - \mathbf{y}| / |\mathbf{x}|) + \log |\mathbf{x}|$ and show that, if $\mathbf{y} \in K$ then $|\log(|\mathbf{x} - \mathbf{y}| / |\mathbf{x}|)| \leq C / |\mathbf{x}|$].

3.12. Prove the representation formula (3.65) in dimension two.

3.13 Compute the Green function for the disc of radius R .

[Answer:

$$G(\mathbf{x}, \mathbf{y}) = -\frac{1}{2\pi} [\log |\mathbf{x} - \mathbf{y}| - \log(\frac{|\mathbf{x}|}{R} |\mathbf{x}^* - \mathbf{y}|)],$$

where $\mathbf{x}^* = R^2 \mathbf{x} |\mathbf{x}|^{-2}$, $\mathbf{x} \neq \mathbf{0}$. Moreover $G(0, \mathbf{y}) = -\frac{1}{2\pi} (\log |\mathbf{y}| - \log R)$.

3.14. Let $\Omega \subset \mathbb{R}^n$ be a bounded, smooth domain and G be the Green function in Ω . Prove that, for every $\mathbf{x}, \mathbf{y} \in \Omega$, $\mathbf{x} \neq \mathbf{y}$:

- (a) $G(\mathbf{x}, \mathbf{y}) > 0$.
- (b) $G(\mathbf{x}, \mathbf{y}) = G(\mathbf{y}, \mathbf{x})$.
- (c) $G(\mathbf{x}, \mathbf{y}) \leq \Phi(\mathbf{x} - \mathbf{y})$.

[Hint: (a) Let $B_r(\mathbf{x}) \subset \Omega$ and let w be the harmonic function in $\Omega \setminus \overline{B}_r(\mathbf{x})$ such that $w = 0$ on $\partial\Omega$ and $w = 1$ on $\partial B_r(\mathbf{x})$. Show that, for every r small enough, $G(\mathbf{x}, \cdot) > w(\cdot)$ in $\Omega \setminus \overline{B}_r(\mathbf{x})$.

(b) For fixed $\mathbf{x} \in \Omega$, define $w_1(\mathbf{y}) = G(\mathbf{x}, \mathbf{y})$ and $w_2(\mathbf{y}) = G(\mathbf{y}, \mathbf{x})$. Apply Green's identity (3.66) in $\Omega \setminus B_r(\mathbf{x})$ to w_1 and w_2 . Let $r \rightarrow 0$].

3.15. Determine the Green function for the half plane $\mathbb{R}_+^2 = \{(x, y); y > 0\}$ and (formally) derive the Poisson formula

$$u(x, y) = \frac{1}{\pi} \int_{\mathbb{R}} \frac{y}{(x - \xi)^2 + y^2} u(\xi, 0) d\xi$$

for a bounded harmonic function in \mathbb{R}_+^2 .

3.16. Prove that the exterior Dirichlet problem in the plane has a unique bounded solution $u \in C^2(\Omega_e) \cap C(\overline{\Omega}_e)$, through the following steps. Let w be the difference of two solutions. Then w is harmonic Ω_e , vanishes on $\partial\Omega_e$ and is bounded, say $|w| \leq M$.

1) Assume that the $\mathbf{0} \in \Omega$. Let $B_a(\mathbf{0})$ and $B_R(\mathbf{0})$ such that $B_a(\mathbf{0}) \subset \Omega \subset B_R(\mathbf{0})$ and define

$$u_R(\mathbf{x}) = M \frac{\ln |\mathbf{x}| - \ln |a|}{\ln R - \ln a}.$$

Use the maximum principle to show that $w \leq u_R$ in the ring $C_{a,R} = \{\mathbf{x} \in \mathbb{R}^2; a < |\mathbf{x}| < R\}$.

- 2) Let $R \rightarrow \infty$ and deduce that $w \leq 0$ in Ω_e .
- 3) Proceed similarly to show that $w \geq 0$ in Ω_e .

3.17. Find the Poisson formula for the circle B_R , by representing the solution of $\Delta u = 0$ in B_R , $u = g$ on ∂B_R , as a double layer potential.

3.18. Consider the following exterior Neumann-Robin problem in \mathbb{R}^3 :

$$\begin{cases} \Delta u = 0 & \text{in } \Omega_e \\ \partial_\nu u + ku = g & \text{on } \partial\Omega_e, (k \geq 0) \\ u \rightarrow 0 & \text{as } |\mathbf{x}| \rightarrow \infty. \end{cases} \quad (3.120)$$

(a) Show that the condition $\int_{\partial\Omega} gd\sigma = 0$ is necessary for the solvability of (3.120), if $k = 0$ and $|u(\mathbf{x})| \leq |\mathbf{x}|^{-1-\varepsilon}$ for $|\mathbf{x}|$ large, with $\varepsilon > 0$.

(b) Represent the solution as a single layer potential and derive the integral equations for the unknown density.

[Hint: (a) Show that, for R large,

$$\int_{\partial\Omega} g \, d\sigma = \int_{\{|\mathbf{x}|=R\}} \partial_\nu u \, d\sigma.$$

Then, use the proof of Theorem 3.42, p. 164, and let $R \rightarrow \infty$].

3.19. Solve (formally) the Neumann problem in the half space \mathbb{R}_+^3 , using a single layer potential.

3.20. Let $B = B_1(\mathbf{0}) \subset \mathbb{R}^2$. To complete the proof of Theorem 3.12 we must show that, if $g \in C(\partial B)$ and u is given by formula (3.22), with $R = 1$ and $\mathbf{p} = \mathbf{0}$, then

$$\lim_{\mathbf{x} \rightarrow \boldsymbol{\xi}} u(\mathbf{x}) = g(\boldsymbol{\xi}) \quad \text{for every } \boldsymbol{\xi} \in \partial B.$$

Fill in the details in the following steps and conclude the proof.

1. First show that

$$\frac{1 - |\mathbf{x}|^2}{2\pi} \int_{\partial B} \frac{1}{|\mathbf{x} - \boldsymbol{\sigma}|^2} d\sigma = 1$$

and, therefore, that

$$u(\mathbf{x}) - g(\boldsymbol{\xi}) = \frac{1 - |\mathbf{x}|^2}{2\pi} \int_{\partial B} \frac{g(\boldsymbol{\sigma}) - g(\boldsymbol{\xi})}{|\mathbf{x} - \boldsymbol{\sigma}|^2} d\sigma.$$

2. For $\delta > 0$, write

$$\begin{aligned} u(\mathbf{x}) - g(\boldsymbol{\xi}) &= \frac{1 - |\mathbf{x}|^2}{2\pi} \int_{\partial B \cap \{|\boldsymbol{\sigma} - \boldsymbol{\xi}| < \delta\}} \dots d\sigma + \frac{1 - |\mathbf{x}|^2}{2\pi} \int_{\partial B \cap \{|\boldsymbol{\sigma} - \boldsymbol{\xi}| > \delta\}} \dots d\sigma \\ &\equiv I + II. \end{aligned}$$

Fix $\varepsilon > 0$ and use the continuity of g to show that, if δ is small enough, then $|I| < \varepsilon$.

3. Show that, if $|\mathbf{x} - \boldsymbol{\xi}| < \delta/2$ and $|\boldsymbol{\sigma} - \boldsymbol{\xi}| > \delta$, then $|\mathbf{x} - \boldsymbol{\sigma}| > \delta/2$ and therefore $\lim_{\mathbf{x} \rightarrow \boldsymbol{\xi}} II = 0$.

- 3.21.** Consider the equation

$$Lu \equiv \Delta u + k^2 u = 0 \quad \text{in } \mathbb{R}^3,$$

known as *Helmoltz's or reduced wave equation*.

- (a) Show that the radial solutions $u = u(r)$, $r = |\mathbf{x}|$, satisfying the *outgoing Sommerfeld condition*

$$u_r + iku = O\left(\frac{1}{r^2}\right) \quad \text{as } r \rightarrow \infty,$$

are of the form

$$\varphi(r; k) = c \frac{e^{-ikr}}{r} \quad c \in \mathbb{C}.$$

- (b) For f smooth and compactly supported in \mathbb{R}^3 define the potential

$$U(\mathbf{x}) = c_0 \int_{\mathbb{R}^3} f(\mathbf{y}) \frac{e^{-ik|\mathbf{x}-\mathbf{y}|}}{|\mathbf{x}-\mathbf{y}|} d\mathbf{y}.$$

Select the constant c_0 such that $LU(\mathbf{x}) = -f(\mathbf{x})$.

[Answer (b): $c_0 = (4\pi)^{-1}$].

Chapter 4

Scalar Conservation Laws and First Order Equations

4.1 Introduction

In the main part of this chapter, from Sects. 4.1 to 4.6, we consider equations of the form

$$u_t + q(u)_x = 0, \quad x \in \mathbb{R}, t > 0. \quad (4.1)$$

In general, $u = u(x, t)$ represents the *density* or the *concentration* of a physical quantity Q (e.g. a mass) and $q(u)$ is its *flux function*¹. Equation (4.1) constitutes a *link* between density and flux and expresses a (**scalar**) **conservation law** for the following reason.

If we consider a control interval (x_1, x_2) , the integral

$$\int_{x_1}^{x_2} u(x, t) dx$$

gives the amount of Q contained inside (x_1, x_2) at time t . A *conservation law* states that, without sources or sinks, the rate of change of Q inside (x_1, x_2) is determined by the net flux through the end points of the interval. If the flux is modelled by a function $q = q(u)$, the law translates into the equation

$$\frac{d}{dt} \int_{x_1}^{x_2} u(x, t) dx = -q(u(x_2, t)) + q(u(x_1, t)), \quad (4.2)$$

where we assume that $q > 0$ ($q < 0$) for a flux along the positive (negative) direction of the x axis. If u and q are smooth functions, eq. (4.2) can be rewritten in the form

$$\int_{x_1}^{x_2} [u_t(x, t) + q(u(x, t))_x] dx = 0$$

which implies (4.1), due to the arbitrariness of the interval (x_1, x_2) .

¹ That is the rate of change of Q . The dimensions of q are $[mass] \times [time]^{-1}$.

At this point we have to decide which type of flux function we are dealing with, or, in other words, we have to establish a *constitutive relation for q* .

In the next section we go back to the model of pollution in a channel, considered in Subsect. 2.5.2, neglecting the diffusion and choosing for q a linear function of u , namely:

$$q(u) = vu,$$

where v is constant. The result is a *pure transport* model, in which the vector \mathbf{v} is the *advection² velocity*. In the sequel, we shall use a nonlinear model from traffic dynamics, with speed v depending on u , to introduce and motivate some important concepts and methods, such as the *method of the characteristics*.

The conservation law (4.1) occurs, for instance, in 1-dimensional fluid dynamics where it often describes the formation and propagation of special solution waves, called *shock* and *rarefaction waves*. A shock wave undergoes a *jump discontinuity* and an important question is how to reinterpret the differential equation (4.1) in order to admit discontinuous solutions. This leads to the concept of solution in a generalized or weak sense.

A typical problem associated with equation (4.1) is the *initial value problem*:

$$\begin{cases} u_t + q(u)_x = 0 \\ u(x, 0) = g(x) \end{cases} \quad (4.3)$$

where $x \in \mathbb{R}$. Sometimes x varies in a half-line or in a finite interval; as we shall see later, in these cases some other conditions have to be added to obtain a well posed problem.

The conservation law (4.1) is a particular case of first order quasilinear equation of the type

$$a(x, y, u)u_x + b(x, y, u)u_y = c(x, y, u),$$

that we consider in Sect. 4.7. In particular, for this kind of equations, we study the Cauchy problem, extending the method of the characteristics. Finally, in Sect. 4.8, we further generalize to fully nonlinear equations of the form

$$F(u_x, u_y, u, x, y) = 0,$$

concluding with a simple application to geometrical optics.

4.2 Linear Transport Equation

4.2.1 Pollution in a channel

Let us go back to the simple model for the evolution of a pollutant concentration in a narrow channel, considered in Subsect. 2.5.2.

² Advection is usually synonymous of *linear convection*.

When diffusion and transport are both relevant, we have derived the equation

$$c_t = Dc_{xx} - vc_x,$$

where c is the concentration of the pollutant and $v\mathbf{i}$ is the stream velocity (v constant). We want to discuss here the case of the *pure transport* equation

$$c_t + vc_x = 0 \quad (4.4)$$

that is, when $D = 0$. Introducing the vector

$$\mathbf{v} = v\mathbf{i} + \mathbf{j},$$

equation (4.4) can be written in the form

$$vc_x + c_t = \nabla c \cdot \mathbf{v} = 0,$$

pointing out the orthogonality of ∇c and \mathbf{v} . But ∇c is orthogonal to the level lines of c , along which c is constant. Therefore the level lines of c are the straight lines parallel to \mathbf{v} , of equation

$$x = vt + x_0.$$

These straight lines are called **characteristics**. Let us compute c along the characteristic $x = vt + x_0$, by letting

$$w(t) = c(x_0 + vt, t).$$

Since³

$$\dot{w}(t) = vc_x(x_0 + vt, t) + c_t(x_0 + vt, t),$$

from equation (4.4) we derive the *ordinary differential equation* $\dot{w}(t) = 0$, confirming that c is constant along the characteristic.

We want to determine the evolution of the concentration c , by knowing its initial profile

$$c(x, 0) = g(x). \quad (4.5)$$

The method to compute the solution at a point (\bar{x}, \bar{t}) , $\bar{t} > 0$, is very simple. Let $x = vt + x_0$ be the equation of the characteristic passing through (\bar{x}, \bar{t}) .

Then, we move back in time along this characteristic, from (\bar{x}, \bar{t}) until the point $(x_0, 0)$ of intersection with the x -axis (see Fig. 4.1).

Since c is constant along the characteristic and $c(x_0, 0) = g(x_0)$, it must be

$$c(\bar{x}, \bar{t}) = g(x_0) = g(\bar{x} - v\bar{t}).$$

Thus, if $g \in C^1(\mathbb{R})$, the unique solution of the initial value problem (4.4), (4.5) is given by

$$c(x, t) = g(x - vt). \quad (4.6)$$

³ The *dot* denotes derivative with respect to time.

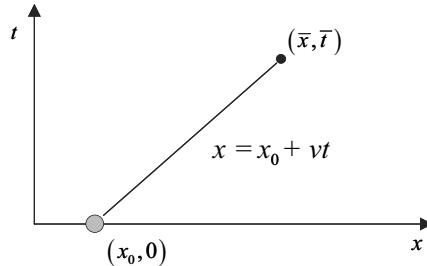


Fig. 4.1 Characteristic line for the linear transport problem

The solution (4.6) represents a *travelling wave*, moving with speed v in the positive (negative) x -direction if $v > 0$ ($v < 0$). In Fig. 4.2, an initial profile $g(x) = \sin(\pi x) \chi_{[0,1]}(x)$ is *transported* in the plane x, t along the straight-lines $x + t = \text{constant}$, i.e. with speed 1 in the negative x -direction.

4.2.2 Distributed source

In presence of an external distributed source along the channel of intensity $f = f(x, t)$ (measured in concentration per unit time), the conservation law (4.2) has to be changed into

$$\frac{d}{dt} \int_{x_1}^{x_2} u(x, t) dx = -q(u(x_2, t)) + q(u(x_1, t)) + \int_{x_1}^{x_2} f(x, t) dx. \quad (4.7)$$

In the pure advection case, (4.7) leads to the nonhomogeneous equation

$$c_t + vc_x = f, \quad (4.8)$$

with the initial condition

$$c(x, 0) = g(x). \quad (4.9)$$

Again, to compute the value of the solution u at a point (\bar{x}, \bar{t}) is an easy task. Let $x = x_0 + vt$ be the characteristic passing through (\bar{x}, \bar{t}) and compute u along this characteristic, setting $w(t) = c(x_0 + vt, t)$. From (4.8), w satisfies the ordinary differential equation

$$\dot{w}(t) = vc_x(x_0 + vt, t) + c_t(x_0 + vt, t) = f(x_0 + vt, t)$$

with the initial condition

$$w(0) = g(x_0).$$

Thus

$$w(t) = g(x_0) + \int_0^t f(x_0 + vs, s) ds.$$

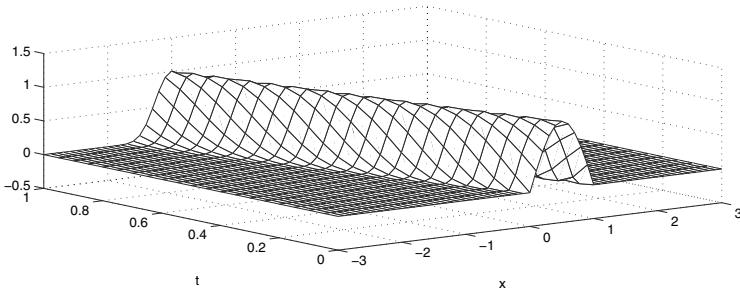


Fig. 4.2 Travelling wave solution of the linear transport equation

Letting $t = \bar{t}$ and recalling that $x_0 = \bar{x} - v\bar{t}$, we get

$$c(\bar{x}, \bar{t}) = w(\bar{t}) = g(\bar{x} - v\bar{t}) + \int_0^{\bar{t}} f(\bar{x} - v(\bar{t} - s), s) ds. \quad (4.10)$$

Since (\bar{x}, \bar{t}) is arbitrary, if g and f are reasonably smooth functions, (4.10) is the desired solution.

Proposition 4.1. *Let $g \in C^1(\mathbb{R})$ and $f, f_x \in C(\mathbb{R} \times (0, +\infty))$. The unique solution of the initial value problem*

$$\begin{cases} c_t + vc_x = f(x, t) & x \in \mathbb{R}, t > 0 \\ c(x, 0) = g(x) & x \in \mathbb{R} \end{cases}$$

is given by the formula

$$c(x, t) = g(x - vt) + \int_0^t f(x - v(t - s), s) ds. \quad (4.11)$$

Remark 4.2. Formula (4.11) can be derived using the *Duhamel method*, as in Subsect. 2.2.8 (see Problem 4.1).

4.2.3 Extinction and localized source

We now consider two other situations, of practical interest in several contexts. Suppose that, due e.g. to *biological decomposition*, the pollutant concentration c decays at the rate

$$r(x, t) = -\gamma c(x, t), \quad \gamma > 0.$$

Without external sources and diffusion, the mathematical model governing the evolution of c reads

$$c_t + vc_x = -\gamma c,$$

with the initial condition

$$c(x, 0) = g(x).$$

Setting

$$u(x, t) = c(x, t) e^{\frac{\gamma}{v}x}, \quad (4.12)$$

we have

$$u_x = \left(c_x + \frac{\gamma}{v} c \right) e^{\frac{\gamma}{v}x} \quad \text{and} \quad u_t = c_t e^{\frac{\gamma}{v}x}$$

and therefore the equation for u is

$$u_t + vu_x = 0$$

with the initial condition

$$u(x, 0) = g(x) e^{\frac{\gamma}{v}x}.$$

From Proposition 4.1, we get

$$u(x, t) = g(x - vt) e^{\frac{\gamma}{v}(x-vt)}$$

and, from (4.12),

$$c(x, t) = g(x - vt) e^{-\gamma t}.$$

Thus, the solution takes the form of a *damped travelling wave*.

We now examine the effect of a source of pollutant placed at a certain point of the channel, e.g. at $x = 0$. Typically, one may think of waste material from industrial machineries. Before the machines start working, for instance before time $t = 0$, we assume that the channel is clean. We want to determine the pollutant concentration, supposing that, at the point $x = 0$, c is kept at a constant level $\beta > 0$, for $t > 0$. Then, we have the *boundary condition*

$$c(0, t) = \beta, \quad \text{for } t > 0 \quad (4.13)$$

and the initial condition

$$c(x, 0) = 0, \quad \text{for } x > 0. \quad (4.14)$$

Thus the relevant domain for our problem is the first quadrant $x > 0, t > 0$.

As before, let $u(x, t) = c(x, t) e^{\frac{\gamma}{v}x}$, which is a solution of $u_t + vu_x = 0$. Then:

$$u(0, t) = c(0, t) = \beta, \quad \text{for } t > 0 \quad (4.15)$$

$$u(x, 0) = c(x, 0) e^{\frac{\gamma}{v}x} = 0, \quad \text{for } x > 0. \quad (4.16)$$

Since u is constant along the characteristics u must be of the form

$$u(x, t) = u_0(x - vt) \quad (4.17)$$

where u_0 is to be determined from the boundary condition (4.15) and the initial condition (4.16).

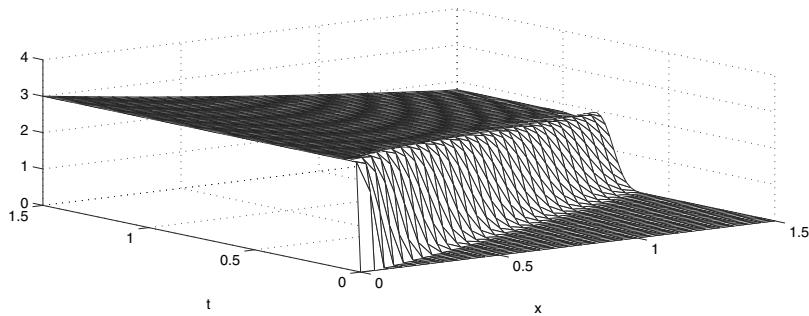


Fig. 4.3 Propagation of a discontinuity

Now, a characteristic $x = vt + x_0$, $x_0 > 0$, leaving from the x -axis, carries *zero data* and hence we deduce that $u = 0$ in all the lower sector $x > vt$. This means that the pollutant has not yet reached the point x , at time t , if $x > vt$.

To compute u for $x < vt$, observe that a characteristic leaving the t -axis, i.e. from a point $(0, t)$, carries the datum β . Therefore, $u = \beta$ in the upper sector $x < vt$.

Recalling (4.12), we can write our original solution in the compact form

$$c(x, t) = \beta \mathcal{H}(vt - x) e^{-\frac{\gamma}{v}x}$$

where \mathcal{H} is the Heaviside function.

Observe that in $(0, 0)$ there is a jump discontinuity which is *transported along the straight line* $x = vt$. Figure 4.3 shows the solution for $\beta = 3$, $\gamma = 0.7$, $v = 2$.

4.2.4 Inflow and outflow characteristics. A stability estimate

The domain in the localized source problem is the quadrant $x > 0, t > 0$. To uniquely determine the solution we have used the initial data on the x -axis, $x > 0$, and the boundary data on the t -axis, $t > 0$. The problem is therefore well posed. This is due to the fact that, since $v > 0$, when time increases, *all* the characteristics carry the information (the data) from the boundary of the quadrant *towards its interior* $x > 0, t > 0$. We say in this case that the characteristics are **inflow characteristics**.

More generally, consider the equation

$$u_t + au_x = f(x, t)$$

in the domain $x > 0, t > 0$, where a is a constant ($a \neq 0$). The characteristics are the lines

$$x - at = \text{constant}$$

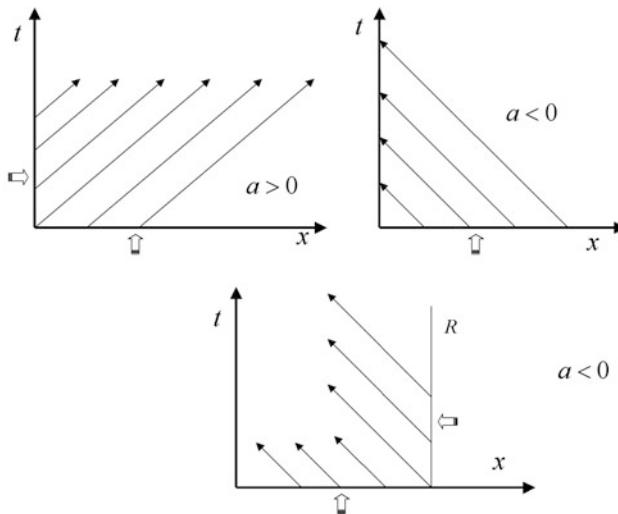


Fig. 4.4 The white arrows indicate where the data should be assigned

as shown in Fig. 4.4. If $a > 0$, we are in the case of the pollutant model: all the characteristics **are inflow** and **the data must be assigned on both semi-axes**.

If $a < 0$, the characteristics leaving the x -axis are **inflow**, while those leaving the t -axis are **outflow**. In this case the initial data alone are sufficient to uniquely determine the solution, while **no data has to be assigned on the semi-axis $x = 0, t > 0$** .

Coherently, a problem in the half-strip $0 < x < R, t > 0$, besides the initial data, requires a data assignment on the inflow boundary, namely (Fig. 4.4):

$$\begin{cases} u(0, t) = h_0(t) & \text{if } a > 0 \\ u(R, t) = h_R(t) & \text{if } a < 0. \end{cases}$$

The resulting initial-boundary value problem is well posed, since the solution is uniquely determined at every point in the strip by its values along the characteristics. Moreover, a stability estimate can be proved as follows. Consider, for instance, the case $a > 0$ and the problem⁴

$$\begin{cases} u_t + au_x = 0 & 0 < x < R, t > 0 \\ u(0, t) = h(t) & t > 0 \\ u(x, 0) = g(x) & 0 < x < R. \end{cases} \quad (4.18)$$

Multiply the differential equation by u and write

$$uu_t + auu_x = \frac{1}{2} \frac{d}{dt} u^2 + \frac{a}{2} \frac{d}{dx} u^2 = 0.$$

⁴ For the case $u_t + au_x = f \neq 0$, see Problem 4.2.

Integrating in x over $(0, R)$, we get:

$$\frac{d}{dt} \int_0^R u^2(x, t) dx + a [u^2(R, t) - u^2(0, t)] = 0.$$

Now use the data $u(0, t) = h(t)$ and the positivity of a , to obtain

$$\frac{d}{dt} \int_0^R u^2(x, t) dx \leq ah^2(t).$$

Integrating in t we have, using the initial condition $u(x, 0) = g(x)$,

$$\int_0^R u^2(x, t) dx \leq \int_0^R g^2(x) dx + a \int_0^t h^2(s) ds. \quad (4.19)$$

Now, let u_1 and u_2 be solutions of problem (4.18) with initial data g_1, g_2 and boundary data h_1, h_2 on $x = 0$. Then, by linearity, $w = u_1 - u_2$ is a solution of problem (4.18) with initial data $g_1 - g_2$ and boundary data $h_1 - h_2$ on $x = 0$. Applying the inequality (4.19) to w we have

$$\int_0^R [u_1(x, t) - u_2(x, t)]^2 dx \leq \int_0^R [g_1(x) - g_2(x)]^2 dx + a \int_0^t [h_1(s) - h_2(s)]^2 ds.$$

Thus, a least-squares distance of the solutions is controlled by a least-squares distance of the data, at each time. In this sense, the solution of problem (4.18) depends continuously on the initial data and on the boundary data on $x = 0$. We point out that the values of u on $x = R$ do not appear in (4.19).

4.3 Traffic Dynamics

4.3.1 A macroscopic model

From far away, an intense traffic on a highway can be considered as a fluid flow and described by means of macroscopic variables such as the *density* of cars⁵ ρ , their *average speed* v and their *flux function*⁶ q . The three (more or less regular) functions ρ , u and q are linked by the simple convection relation

$$q = vp.$$

To construct a model for the evolution of ρ we assume the following hypotheses.

1. *There is only one lane and overtaking is not allowed.* This is realistic, for instance, for the traffic in a tunnel (see Problem 4.6). Multi-lanes models with

⁵ Average number of cars per unit length.

⁶ Cars per unit time.

overtaking are beyond the scope of this introduction. However, the model we will present is often in agreement with observed dynamics also in this case.

2. *No car “sources” or “sinks”.* We consider a road section without exit/entrance gates.

3. *The average speed is not constant and depends only on the density,* that is

$$v = v(\rho).$$

This rather controversial assumption implies that, at a given density, the speed is uniquely determined and that a density change causes an immediate speed variation. Clearly

$$v'(\rho) = \frac{dv}{d\rho} \leq 0$$

since we expect the speed to decrease as the density increases.

As in Sect. 4.1, from hypotheses **2** and **3** we derive the conservation law:

$$\rho_t + q(\rho)_x = 0, \quad (4.20)$$

where

$$q(\rho) = v(\rho)\rho.$$

We need a constitutive relation for $v = v(\rho)$. When ρ is small, it is reasonable to assume that the average speed v is more or less equal to the maximal velocity v_m , given by the speed limit. When ρ increases, traffic slows down and stops at the maximum density ρ_m (bumper-to-bumper traffic). We adopt the simplest model consistent with the above considerations, namely

$$v(\rho) = v_m \left(1 - \frac{\rho}{\rho_m}\right), \quad (4.21)$$

so that

$$q(\rho) = v_m \rho \left(1 - \frac{\rho}{\rho_m}\right). \quad (4.22)$$

Since

$$q(\rho)_x = q'(\rho) \rho_x = v_m \left(1 - \frac{2\rho}{\rho_m}\right) \rho_x,$$

equation (4.20) becomes

$$\underbrace{\rho_t + v_m \left(1 - \frac{2\rho}{\rho_m}\right) \rho_x}_{{q}'(\rho)} = 0. \quad (4.23)$$

According to the terminology in Sect. 1.1, this is a quasilinear equation. We also point out that

$$q''(\rho) = -\frac{2v_m}{\rho_m} < 0,$$

so that q is strictly *concave*. We couple the equation (4.23) with the initial condition

$$\rho(x, 0) = g(x). \quad (4.24)$$

4.3.2 The method of characteristics

We want to solve the initial value problem (4.23), (4.24). To compute the density ρ at a point (x, t) , we follow the idea we already exploited in the linear transport case without external sources: *to connect the point (x, t) with a point $(x_0, 0)$ on the x -axis, through a curve along which ρ is constant* (Fig. 4.5).

Clearly, if we manage to find such a curve, that we call **characteristic based at $(x_0, 0)$** , the value of ρ at (x, t) is given by $\rho(x_0, 0) = g(x_0)$. Moreover, if this procedure can be repeated for every point (x, t) , $x \in \mathbb{R}$, $t > 0$, then we can compute ρ everywhere in the upper halfplane and the problem is completely solved. This is the *method of characteristics*.

Adopting a slightly different point of view, we can implement the above idea as follows: assume that $x = x(t)$ is the equation of the characteristic based at the point $(x_0, 0)$ and that along $x = x(t)$ we *always observe the same initial density $g(x_0)$* . In other words, we have

$$\rho(x(t), t) = g(x_0) \quad (4.25)$$

for every $t > 0$. If we differentiate the identity (4.25), we get

$$\frac{d}{dt}\rho(x(t), t) = \rho_x(x(t), t)\dot{x}(t) + \rho_t(x(t), t) = 0 \quad (t > 0). \quad (4.26)$$

On the other hand, (4.23) yields

$$\rho_t(x(t), t) + q'g(x_0)\rho_x(x(t), t) = 0,$$

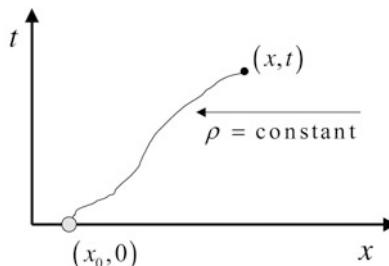


Fig. 4.5 Characteristic curve

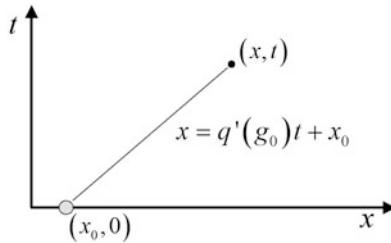


Fig. 4.6 Characteristic straight line ($g_0 = g(x_0)$)

so that, subtracting (4.27) from (4.26), we obtain

$$\rho_x(x(t), t)[\dot{x}(t) - q'(g(x_0))] = 0. \quad (4.27)$$

Assuming $\rho_x(x(t), t) \neq 0$, we deduce

$$\dot{x}(t) = q'(g(x_0)).$$

Since $x(0) = x_0$ we find

$$x(t) = q'(g(x_0))t + x_0. \quad (4.28)$$

Thus, the characteristics are **straight lines** with slope $q'(g(x_0))$. Different values of x_0 give, in general, different values of the slope (Fig. 4.6).

We can now derive a formula for ρ . To compute $\rho(x, t)$, $t > 0$, we move back in time along the characteristic through (x, t) until its base point $(x_0, 0)$. Then $\rho(x, t) = g(x_0)$. From (4.28) we have, since $x(t) = x$,

$$x_0 = x - q'(g(x_0))t$$

and finally

$$\rho(x, t) = g(x - q'(g(x_0))t). \quad (4.29)$$

Formula (4.29) represents a **travelling wave** (or a signal, a disturbance) **propagating with speed** $q'(g(x_0))$ along the positive x -direction.

We emphasize that $q'(g(x_0))$ is the *local wave speed* and it must not be confused with the traffic velocity. In fact, in general,

$$\frac{dq}{d\rho} = \frac{d(\rho v)}{d\rho} = v + \rho \frac{dv}{d\rho} \leq v$$

since $\rho \geq 0$ and $\frac{dv}{d\rho} \leq 0$.

The different nature of the two speeds becomes more evident if we observe that the wave speed *may be negative* as well. This means that, while the traffic advances along the positive x -direction, the disturbance given by the travelling wave may propagate in the opposite direction. Indeed, in our model (4.22), $\frac{dq}{d\rho} < 0$ when $\rho > \frac{\rho_m}{2}$.

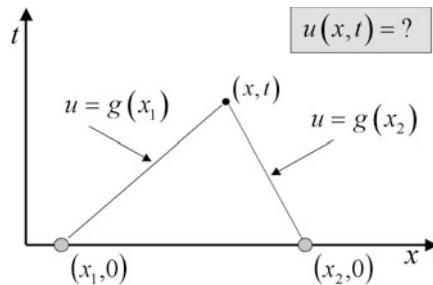


Fig. 4.7 Intersection of characteristics

Formula (4.29) seems to be rather satisfactory, since, apparently, it gives the solution of the initial value problem (4.23), (4.24) at every point. Actually, a more accurate analysis shows that, even if the initial datum g is smooth, the solution may develop a singularity in finite time (e.g. a jump discontinuity). When this occurs, the method of characteristics does not work anymore and formula (4.29) is not effective. A typical case is described in Fig. 4.7: two characteristics based at different points $(x_1, 0)$ and $(x_2, 0)$ intersect at the point (x, t) and the value $u(x, t)$ is not uniquely determined as soon as $g(x_1) \neq g(x_2)$.

In this case we have to weaken the concept of solution and the computation technique. We will come back on these questions later. For the moment, we analyze the method of characteristics in some remarkable cases.

4.3.3 The green light problem

Suppose that bumper-to-bumper traffic is standing at a red light, placed at $x = 0$, while the road ahead is empty. Accordingly, the initial density profile is

$$g(x) = \begin{cases} \rho_m & \text{for } x \leq 0 \\ 0 & \text{for } x > 0. \end{cases}$$

At time $t = 0$, the traffic light turns green and we want to describe the car flow evolution for $t > 0$. At the beginning, only the cars closer to the light start moving while most of them remain standing.

Since $q'(\rho) = v_m \left(1 - \frac{2\rho}{\rho_m}\right)$, the local wave speed is given by

$$q'(g(x_0)) = \begin{cases} -v_m & \text{for } x_0 \leq 0 \\ v_m & \text{for } x_0 > 0 \end{cases}$$

and the characteristics are the straight lines

$$\begin{aligned} x &= -v_m t + x_0 && \text{if } x_0 \leq 0 \\ x &= v_m t + x_0 && \text{if } x_0 > 0. \end{aligned}$$

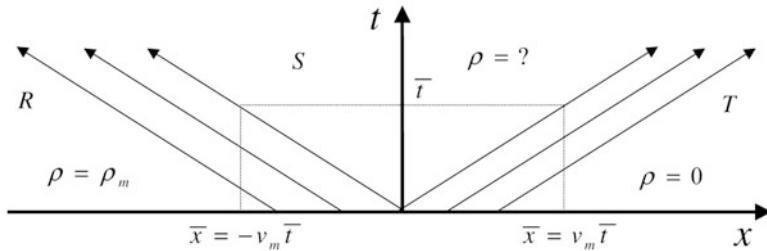


Fig. 4.8 Characteristic for the green light problem

The lines $x = v_m t$ and $x = -v_m t$ partition the upper half-plane in the three regions R , S and T , shown in Fig. 4.8.

Inside R we have $\rho(x, t) = \rho_m$, while inside T we have $\rho(x, t) = 0$. Consider the points on the horizontal line $t = \bar{t}$. At the points $(x, \bar{t}) \in T$ the density is zero: the traffic has not yet arrived in x at time $t = \bar{t}$. The front car is located at the point

$$\bar{x} = v_m \bar{t}$$

which moves at the maximum speed, since ahead the road is empty.

The cars placed at the points $(x, \bar{t}) \in R$ are still standing. The first car that starts moving at time $t = \bar{t}$ is at the point

$$\bar{x} = -v_m \bar{t}.$$

In particular, it follows that *the green light signal propagates back through the traffic at speed v_m .*

What is the value of the density inside the sector S ? No characteristic extends into S , due to the discontinuity of the initial data at the origin, and the method as it stands does not give any information on the value of ρ inside S .

A strategy that may give a reasonable answer is the following:

- a) Approximate the initial data by a continuous function g_ε , which converges to g as $\varepsilon \rightarrow 0$ at every point x , except 0.
- b) Construct the solution ρ_ε of the ε -problem by the method of characteristics.
- c) Let $\varepsilon \rightarrow 0$ and check that the limit of ρ_ε is a solution of the original problem.

Clearly we run the risk of constructing many solutions, each one depending on the way we regularize the initial data, but for the moment we are satisfied if we construct at least one solution.

- a) Let us choose as g_ε the function (Fig. 4.9)

$$g_\varepsilon(x) = \begin{cases} \rho_m & x \leq 0 \\ \rho_m(1 - \frac{x}{\varepsilon}) & 0 < x < \varepsilon \\ 0 & x \geq \varepsilon. \end{cases}$$

When $\varepsilon \rightarrow 0$, $g_\varepsilon(x) \rightarrow g(x)$ for every $x \neq 0$.

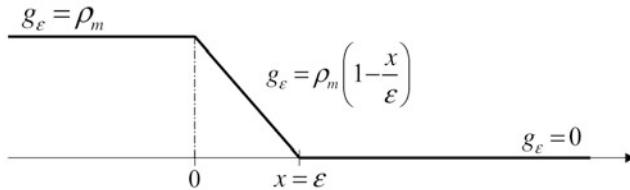


Fig. 4.9 Smoothing of the initial data in the green light problem

b) The characteristics for the ε -problem are:

$$\begin{aligned} x &= -v_m t + x_0 && \text{if } x_0 < 0 \\ x &= -v_m \left(1 - 2\frac{x_0}{\varepsilon}\right) t + x_0 && \text{if } 0 \leq x_0 < \varepsilon \\ x &= v_m t + x_0 && \text{if } x_0 \geq \varepsilon, \end{aligned}$$

since, for $0 \leq x_0 < \varepsilon$,

$$q'(g_\varepsilon(x_0)) = v_m \left(1 - \frac{2g_\varepsilon(x_0)}{\rho_m}\right) = -v_m \left(1 - 2\frac{x_0}{\varepsilon}\right).$$

We say that the characteristics in the region $-v_m t < x < v_m t + \varepsilon$ form a *rarefaction fan* (Fig. 4.10).

Clearly, $\rho_\varepsilon(x, t) = 0$ for $x \geq v_m t + \varepsilon$ and $\rho_\varepsilon(x, t) = \rho_m$ for $x \leq -v_m t$. Let now (x, t) belong to the region

$$-v_m t < x < v_m t + \varepsilon.$$

Solving for x_0 in the equation of the characteristic $x = -v_m \left(1 - 2\frac{x_0}{\varepsilon}\right) t + x_0$, we find

$$x_0 = \varepsilon \frac{x + v_m t}{2v_m t + \varepsilon}.$$

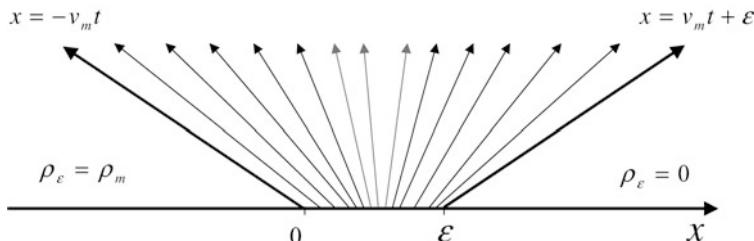


Fig. 4.10 Fanlike characteristics

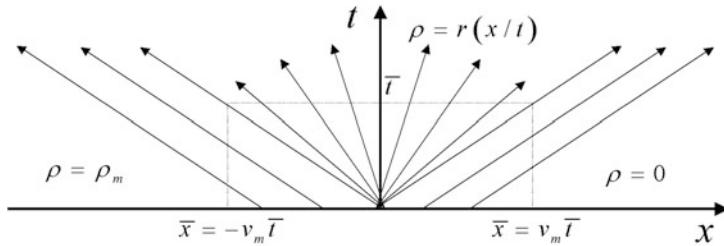


Fig. 4.11 Characteristics in a rarefaction wave

Then

$$\rho_\varepsilon(x, t) = g_\varepsilon(x_0) = \rho_m \left(1 - \frac{x_0}{\varepsilon}\right) = \rho_m \left(1 - \frac{x + v_m t}{2v_m t + \varepsilon}\right). \quad (4.30)$$

c) Letting $\varepsilon \rightarrow 0$ in (4.30) we obtain

$$\rho(x, t) = \begin{cases} \rho_m & \text{for } x \leq -v_m t \\ \frac{\rho_m}{2} \left(1 - \frac{x}{v_m t}\right) & \text{for } -v_m t < x < v_m t \\ 0 & \text{for } x \geq v_m t \end{cases}. \quad (4.31)$$

It is easy to check that ρ is a solution of the equation (4.23) in the regions R, S, T . For fixed t , the function ρ decreases linearly from ρ_m to 0, as x varies from $-v_m t$ to $v_m t$. Moreover, ρ is constant on the fan of straight lines

$$x = ht, \quad -v_m < h < v_m.$$

These type of solutions are called **rarefaction** or **simple waves** (centered at the origin). The characteristics and a typical profile are shown in Figs. 4.11 and 4.12.

The formula for $\rho(x, t)$ in the sector S can be obtained, a posteriori, by a formal procedure that emphasizes its structure. The equation of the characteristics

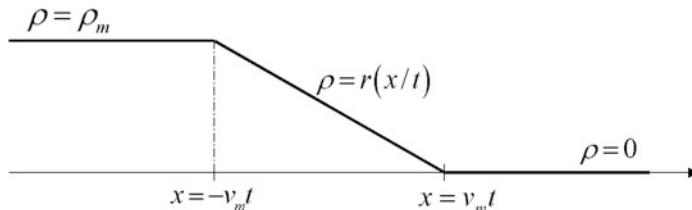


Fig. 4.12 Profile of a rarefaction wave at time t

based on $(x_0, 0)$ can be written in the form

$$x = v_m \left(1 - \frac{2g(x_0)}{\rho_m} \right) t + x_0 = v_m \left(1 - \frac{2\rho(x, t)}{\rho_m} \right) t + x_0$$

because $\rho(x, t) = g(x_0)$ along it. Inserting $x_0 = 0$ we obtain

$$x = v_m \left(1 - \frac{2\rho(x, t)}{\rho_m} \right) t.$$

Solving for ρ we find

$$\rho(x, t) = \frac{\rho_m}{2} \left(1 - \frac{x}{v_m t} \right) \quad (t > 0). \quad (4.32)$$

Since $v_m \left(1 - \frac{2\rho}{\rho_m} \right) = q'(\rho)$, we see that (4.32) is equivalent to

$$\rho(x, t) = r \left(\frac{x}{t} \right)$$

where $r = (q')^{-1}$ is the inverse function of q' .

Indeed the general form of a rarefaction wave, centered at a point (x_0, t_0) , is given by

$$\rho(x, t) = r \left(\frac{x - x_0}{t - t_0} \right).$$

We have constructed a continuous solution ρ of the green light problem, *connecting the two constant states* ρ_m and 0 by a rarefaction wave. However, it is not clear in which sense ρ is a solution across the lines $x = \pm v_m t$, since, there, its derivatives undergo a jump discontinuity. Also, it is not clear whether or not (4.31) is the only solution. We will return on these important points.

- *The vehicle paths.* We now examine the path of a vehicle, initially located at $x = -a < 0$. This vehicle, being inside the bumper-to-bumper region, will not move until time $\tau = a/v_m$, at which the green light signal reaches it. From this moment, the vehicle enters the rarefaction region and moves with speed $v(\rho) = v_m (1 - \rho/\rho_m)$, where ρ is given by (4.32). Thus, denoting by $x = x(t)$ the vehicle position, we must solve the initial value problem

$$\begin{cases} \dot{x} = \frac{v_m}{2} + \frac{1}{2} \frac{x}{t} \\ x(\tau) = -a. \end{cases}$$

Dividing both sides of the ODE by \sqrt{t} , we may write the differential equation in the form

$$\frac{d}{dt} \left(\frac{x}{\sqrt{t}} \right) = \frac{v_m}{2\sqrt{t}}.$$

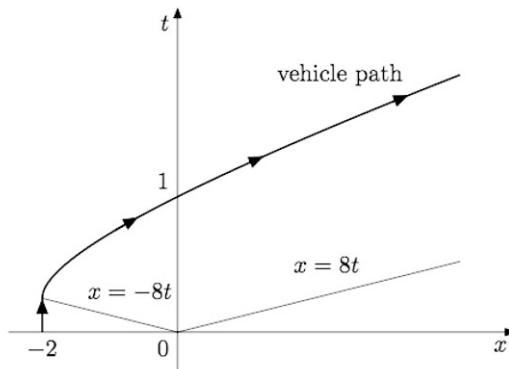


Fig. 4.13 The path of a vehicle starting at $x = -2$

Integrating over (τ, t) we find, since $v(\tau) = a$,

$$\frac{x(t)}{\sqrt{t}} + \frac{a}{\sqrt{\tau}} = v_m(\sqrt{t} - \sqrt{\tau}).$$

Inserting $\tau = a/v_m$, we finally get

$$x(t) = v_m t - 2\sqrt{av_m t}. \quad (4.33)$$

Since $v_m t - 2\sqrt{av_m t} < v_m t$, the path will never cross the characteristic $x = v_m t$ and hence will never leave the rarefaction region. Thus, after time $\tau = a/v_m$ the path is given by the portion of parabola (4.33), as shown in Fig. 4.13, where $a = 2$ and $v_m = 8$.

Suppose that the light turns on red at time t_{red} . Will our driver be able to go through the traffic light?

We only have to compute how much time it takes to him/her to reach the origin. Since $x(t) = 0$ at time $t_d = 4a/v_m$, our driver will be able to pass through the traffic light if $t_{red} < t_d$.

4.3.4 Traffic jam ahead

We now assume that the initial density profile is

$$g(x) = \begin{cases} \frac{1}{8}\rho_m & \text{for } x < 0 \\ \rho_m & \text{for } x > 0. \end{cases}$$

For $x > 0$, the density is maximal and therefore the traffic is bumper-to-bumper. The cars on the left move with speed $v = \frac{7}{8}v_m$, so that we expect a congestion

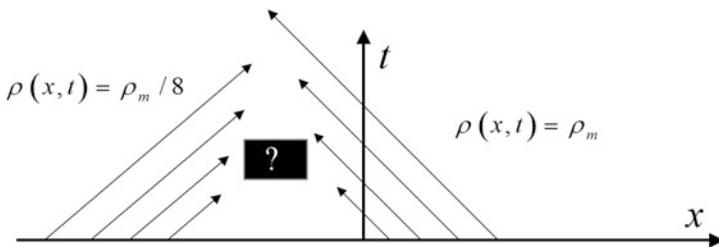


Fig. 4.14 Expecting a discontinuity

propagating back into the traffic. We have

$$q'(g(x_0)) = \begin{cases} \frac{3}{4}v_m & \text{if } x_0 < 0 \\ -v_m & \text{if } x_0 > 0 \end{cases}$$

and therefore the characteristics are

$$\begin{aligned} x &= \frac{3}{4}v_m t + x_0 && \text{if } x_0 < 0 \\ x &= -v_m t + x_0 && \text{if } x_0 > 0. \end{aligned}$$

The configuration in Fig. 4.14 shows that the characteristics intersect somewhere in finite time and the theory predicts that ρ becomes a "multivalued" function of position. In other words, ρ should assume two different values at the same point, which clearly makes no sense in our situation. Therefore we have to admit solutions with jump discontinuities, but then we have to reexamine the derivation of the conservation law, because the smoothness assumption for ρ does not hold anymore.

Thus, let us go back to the conservation of cars in integral form (see (4.2)):

$$\frac{d}{dt} \int_{x_1}^{x_2} \rho(x, t) dx = -q(\rho(x_2, t)) + q(\rho(x_1, t)), \quad (4.34)$$

valid in any control interval (x_1, x_2) . Suppose now that ρ is a smooth function except along a smooth curve

$$x = s(t), \quad t \in [t_1, t_2],$$

on which ρ undergoes a *jump discontinuity*.

For fixed t , let (x_1, x_2) be an interval containing the discontinuity point $x = s(t)$. From (4.34) we have

$$\frac{d}{dt} \left\{ \int_{x_1}^{s(t)} \rho(y, t) dy + \int_{s(t)}^{x_2} \rho(y, t) dy \right\} + q(\rho(x_2, t)) - q(\rho(x_1, t)) = 0. \quad (4.35)$$

The fundamental theorem of calculus gives

$$\frac{d}{dt} \int_{x_1}^{s(t)} \rho(y, t) dy = \int_{x_1}^{s(t)} \rho_t(y, t) dy + \rho_-(s(t), t) \dot{s}(t)$$

and

$$\frac{d}{dt} \int_{s(t)}^{x_2} \rho(y, t) dy = \int_{s(t)}^{x_2} \rho_t(y, t) dy - \rho_+(s(t), t) \dot{s}(t),$$

where

$$\rho_+(s(t), t) = \lim_{y \rightarrow s(t)^+} \rho(y, t), \quad \rho_-(s(t), t) = \lim_{y \rightarrow s(t)^-} \rho(y, t).$$

Hence, equation (4.35) becomes

$$\int_{x_1}^{x_2} \rho_t(y, t) dy + [\rho_-(s(t), t) - \rho_+(s(t), t)] \dot{s}(t) = q(\rho(x_1, t)) - q(\rho(x_2, t)).$$

Letting $x_2 \rightarrow s(t)^+$ and $x_1 \rightarrow s(t)^-$, we obtain

$$[\rho_-(s(t), t) - \rho_+(s(t), t)] \dot{s}(t) = q(\rho_-(s(t), t)) - q(\rho_+(s(t), t)),$$

that is:

$$\dot{s} = \frac{q(\rho_+(s, t)) - q(\rho_-(s, t))}{\rho_+(s, t) - \rho_-(s, t)}. \quad (4.36)$$

The relation (4.36) is an ordinary differential equation for s and it is known as **Rankine-Hugoniot jump condition**. If (4.36) holds, the discontinuity line is called **shock curve** or simply a shock, and the propagating discontinuity is called **shock wave**⁷. The jump

$$\rho_+(s, t) - \rho_-(s, t)$$

is called the *strength* of the shock.

The Rankine-Hugoniot condition gives the *shock speed* \dot{s} as the quotient of the flux jump over the density jump.

To determine the shock curve we need to know its initial point and the values of ρ from both sides of the curve.

Let us apply the above considerations to the traffic problem of Subsect. 4.3.4. We have

$$\rho_+ = \rho_m, \quad \rho_- = \frac{\rho_m}{8},$$

while

$$q(\rho_+) = 0, \quad q(\rho_-) = \frac{7}{64} v_m \rho_m$$

⁷ Actually, the solution ρ itself is often called *shock wave*.

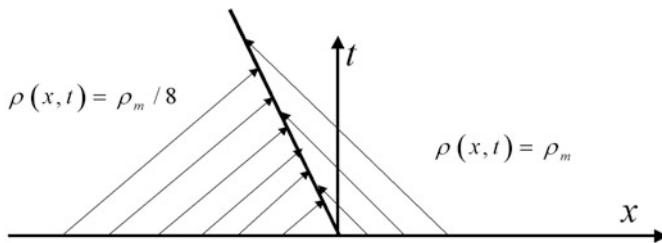


Fig. 4.15 Shock wave

and (4.36) gives⁸

$$\dot{s}(t) = \frac{q(\rho_+) - q(\rho_-)}{\rho_+ - \rho_-} = -\frac{1}{8}v_m.$$

Since clearly $s(0) = 0$, the shock curve is the straight line

$$x = -\frac{1}{8}v_mt.$$

Note that *the slope is negative: the shock propagates back with speed $-\frac{1}{8}v_m$* , as it is revealed by the brake lights of the cars, slowing down because of a traffic jam ahead.

As a consequence, the solution of our problem is given by the following formula (Fig. 4.15)

$$\rho(x, t) = \begin{cases} \frac{1}{8}\rho_m & x < -\frac{1}{8}v_mt \\ \rho_m & x > -\frac{1}{8}v_mt. \end{cases}$$

This time we say that the two constant states $\frac{1}{8}\rho_m$ and ρ_m are connected by a **shock wave**.

4.4 Weak (or Integral) Solutions

4.4.1 The method of characteristics revisited

The method of characteristics applied to the problem

$$\begin{cases} u_t + q(u)_x = 0 \\ u(x, 0) = g(x) \end{cases} \quad (4.37)$$

⁸ In the present case the following simple formula holds:

$$\frac{q(w) - q(z)}{w - z} = v_m \left(1 - \frac{w + z}{\rho_m} \right).$$

gives the travelling wave (see (4.29) with $x_0 = \xi$)

$$u(x, t) = g(x - q'(g(\xi))t) \quad \left(q' = \frac{dq}{du} \right) \quad (4.38)$$

with local speed $q'(g(\xi))$, in the positive x -direction. Since $u(x, t) \equiv g(\xi)$ along the characteristic

$$x = q'(g(\xi))t + \xi \quad (4.39)$$

based at $(\xi, 0)$, from (4.38) we obtain that u is implicitly defined by the equation

$$G(x, t, u) \equiv u - g(x - q'(u)t) = 0. \quad (4.40)$$

If g and q' are smooth functions, the Implicit Function Theorem implies that equation (4.40) defines u as a function of (x, t) , as long as the condition

$$G_u(x, t, u) = 1 + tq''(u)g'(x - q'(u)t) \neq 0 \quad (4.41)$$

holds. Since along the characteristic (4.39) we have $g'(x - q'(u)t) = g'(\xi)$ and $q''(u) = q''(g(\xi))$, an immediate consequence is that if $q'' \circ g$ and g' have the same sign, the solution given by the method of characteristics is defined and smooth for all times $t \geq 0$. Precisely, we have:

Proposition 4.3. *Assume that $q \in C^2(\mathbb{R})$, $g \in C^1(\mathbb{R})$ and $g'(\xi)q''(g(\xi)) \geq 0$ in \mathbb{R} . Then formula (4.40) defines the unique solution u of problem (4.37) in the half-plane $t \geq 0$. Moreover, $u \in C^1(\mathbb{R} \times [0, \infty))$.*

Thus, if $q'' \circ g$ and g' have the same sign, the characteristics cannot intersect. This is not surprising since

$$g'(\xi)q''(g(\xi)) = \frac{d}{d\xi}q'(g(\xi)),$$

so that the condition $g'(\xi)q''(g(\xi)) \geq 0$ means the the characteristics have increasing slope, preventing any intersection. Note that in the ε -approximation of the green light problem, q is concave and g_ε is decreasing. Although g_ε is not smooth, the characteristics do not intersect and ρ_ε is well defined for all times $t > 0$. In the limit as $\varepsilon \rightarrow 0$, the discontinuity of g reappears and the fan of characteristics produces a rarefaction wave.

What happens if $q'' \circ g$ and g' have a different sign over an interval $[a, b]$? Proposition 4.3 still holds for small times, since $G_u \sim 1$ if $t \sim 0$, but when time goes on, we expect the formation of a shock. Indeed, suppose, for instance, that q is concave and g is increasing. The family of characteristics based on a point in the interval $[a, b]$ is

$$x = q'(g(\xi))t + \xi, \quad \xi \in [a, b]. \quad (4.42)$$

Since $q'(g(\xi))$ decreases in $[a, b]$, we expect an intersection of characteristics along a shock curve. The main question is to find the positive time t_s (*breaking time*) and the location x_s of **first appearance of the shock**.

According to the above discussion, the breaking time must coincide with the first time t at which the expression

$$G_u(x, t, u) = 1 + tq''(u)g'(x - q'(u)t)$$

becomes zero. Computing G_u along the characteristic (4.42), we have $u = g(\xi)$ and

$$G_u(x, t, u) = 1 + tq''(g(\xi))g'(\xi).$$

Assume that the positive function

$$z(\xi) = -q''(g(\xi))g'(\xi)$$

attains its maximum only at the point $\xi_M \in [a, b]$. Then $z(\xi_M) > 0$ and

$$t_s = \min_{\xi \in [a, b]} \frac{1}{z(\xi)} = \frac{1}{z(\xi_M)}. \quad (4.43)$$

Since x_s belongs to the characteristics $x = q'(g(\xi_M))t + \xi_M$, we find

$$x_s = \frac{q'(g(\xi_M))}{z(\xi_M)} + \xi_M. \quad (4.44)$$

The point (x_s, t_s) has an interesting geometrical meaning. In fact, it turns out that if $g'(\xi)q''(g(\xi)) < 0$ in $[a, b]$, the family of characteristics (4.42) admits an envelope⁹ and (x_s, t_s) is the point on the envelope with minimum time coordinate (see Problem 4.7).

Example 4.4. Consider the initial value problem

$$\begin{cases} u_t + (1 - 2u)u_x = 0 \\ u(x, 0) = \frac{1}{2}\arctan(\pi x). \end{cases} \quad (4.45)$$

We have $q(u) = u - u^2$, $q'(u) = 1 - 2u$, $q''(u) = -2$, and $g(\xi) = \frac{1}{2}\arctan(\pi\xi)$, $g'(\xi) = \pi/2(1 + \pi^2\xi^2)$. Therefore, the function

$$z(\xi) = -q''(g(\xi))g'(\xi) = \frac{\pi}{(1 + \pi^2\xi^2)}$$

⁹ Recall that the envelope of a family of curves $\phi(x, t, \xi) = 0$, depending on the parameter ξ , is a curve $\psi(x, t) = 0$ tangent at each one of its points to a curve of the family. If the family of curves $\phi(x, t, \xi) = 0$ has an envelope, its parametric equations are obtained by solving the system

$$\begin{cases} \phi(x, t, \xi) = 0 \\ \phi_\xi(x, t, \xi) = 0 \end{cases}$$

with respect to x and t .

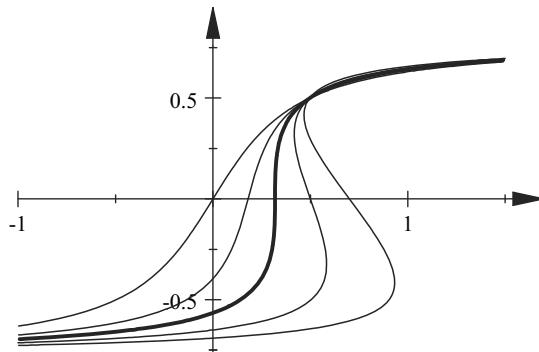


Fig. 4.16 The graphs implicitly defined by equation (4.46) for $t = 0, 0.18, 1/\pi, 0.5, 0.7$

has a maximum at $\xi_M = 0$ and $z(0) = \pi$. The breaking-time is $t_s = 1/\pi$ and

$$x_s = q'(g(\xi_M))t_s + \xi_M = 1/\pi.$$

Thus, the shock curve starts from $(1/\pi, 1/\pi)$. For $0 \leq t < 1/\pi$ the solution u is smooth and implicitly defined by the equation

$$G(x, t, u) = u - \frac{1}{2} \arctan [\pi x - \pi (1 - 2u) t] = 0. \quad (4.46)$$

After $t = 1/\pi$, equation (4.46) defines u as a multivalued function of (x, t) and does not define a solution anymore. Figure 4.16 shows what happens for $t = 0, 0.5, 1/\pi, 0.7, 0.18$. The thick line corresponds to graph of u at the breaking time $t = 1/\pi$.

How does the solution of Example 4.4 evolve after $t = 1/\pi$? We have to insert a shock into the multivalued graph in Fig. 4.16, in a way that the conservation law is preserved. We will see that the correct insertion point is prescribed by the Rankine-Hugoniot condition. It turns out that this corresponds to cutting off from the multivalued profile the two **equal area** lobes A and B as described in Fig. 4.17, at time $t = 0.7$ (*G. B. Whitham equal area rule*¹⁰).

4.4.2 Definition of weak solution

We have seen that the method of characteristics is not sufficient, in general, to determine the solution of an initial value problem for all times $t > 0$. In the green light problem a rarefaction wave was used to construct the solution in a region not covered by characteristics. In the traffic jam case the solution undergoes a shock, propagating according to the Rankine-Hugoniot condition.

¹⁰ See [30], Whitham, 1974.

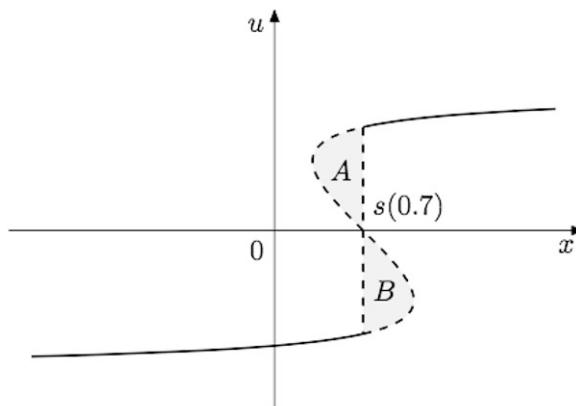


Fig. 4.17 Inserting a shock by the Whitham *equal-area rule* in Example 4.4

Some questions naturally arise.

Q1. *In which sense is the differential equation satisfied across a shock or, more generally, across a separation curve where the constructed solution is not differentiable?*

A way to “solve” the problem would be simply not to care about those points. However, in this case it would be possible to construct solutions that do not have any connection with the physical meaning of the conservation law.

Q2. *Is the solution unique?*

Q3. *If there is no uniqueness, is there a criterion to select the “physically correct” solution?*

To answer, we need first of all to introduce a more flexible notion of solution, in which the derivatives of the solution are not directly involved. Let us go back to the problem

$$\begin{cases} u_t + q(u)_x = 0 & x \in \mathbb{R}, t > 0 \\ u(x, 0) = g(x) & x \in \mathbb{R} \end{cases} \quad (4.47)$$

where q is a smooth function and g is bounded. Assume for the moment that u is a smooth solution, at least of class C^1 in $\mathbb{R} \times [0, \infty)$. We say that u is a **classical solution**.

Let v be a smooth function, with compact support in $\mathbb{R} \times [0, \infty)$. Thus v vanishes outside a compact set contained in $\mathbb{R} \times [0, \infty)$. We call v a *test function*. Multiplying the differential equation by v and integrating on $\mathbb{R} \times (0, \infty)$, we get

$$\int_0^\infty \int_{\mathbb{R}} [u_t + q(u)_x] v \, dx dt = 0. \quad (4.48)$$

The idea is to carry the derivatives onto the test function v via an integration by parts. If we integrate by parts the first term with respect to t we obtain¹¹:

$$\begin{aligned} \int_0^\infty \int_{\mathbb{R}} u_t v \, dx dt &= - \int_0^\infty \int_{\mathbb{R}} u v_t \, dx dt - \int_{\mathbb{R}} u(x, 0) v(x, 0) \, dx \\ &= - \int_0^\infty \int_{\mathbb{R}} u v_t \, dx dt - \int_{\mathbb{R}} g(x) v(x, 0) \, dx. \end{aligned}$$

Integrating by parts the second term in (4.48) with respect to x , we have:

$$\int_0^\infty \int_{\mathbb{R}} q(u)_x v \, dx dt = - \int_0^\infty \int_{\mathbb{R}} q(u) v_x \, dx dt.$$

Then, equation (4.48) becomes

$$\int_0^\infty \int_{\mathbb{R}} [uv_t + q(u)v_x] \, dx dt + \int_{\mathbb{R}} g(x) v(x, 0) \, dx = 0. \quad (4.49)$$

We have obtained an integral equation, valid **for every test function v** . Observe that no derivative of u appears in (4.49).

On the other hand, suppose that a smooth function u satisfies (4.49) *for every test function v* . Integrating by parts in the reverse order, we arrive to the equation

$$\int_0^\infty \int_{\mathbb{R}} [u_t + q(u)_x] v \, dx dt + \int_{\mathbb{R}} [g(x) - u(x, 0)] v(x, 0) \, dx = 0, \quad (4.50)$$

which is true for every test function v .

If we choose v vanishing for $t = 0$, then the second integral is zero and the arbitrariness of v implies

$$u_t + q(u)_x = 0 \text{ in } \mathbb{R} \times (0, +\infty). \quad (4.51)$$

Choosing now v not vanishing for $t = 0$, from (4.50) and (4.51), we get

$$\int_{\mathbb{R}} [g(x) - u(x, 0)] v(x, 0) \, dx = 0.$$

Once more, the arbitrariness of v implies

$$u(x, 0) = g(x) \text{ in } \mathbb{R}.$$

Therefore u is a classical solution of problem (4.47).

Conclusion: *a function $u \in C^1(\mathbb{R} \times [0 \times \infty))$ is a classical solution of problem (4.47) if and only if the equation (4.49) holds for every test function v .*

¹¹ Since v is compactly supported and u, v are smooth, there is no problem in exchanging the order of integration.

But (4.49) makes perfect sense for u merely bounded, and therefore it constitutes an alternative **weak or integral** formulation of problem (4.47). This motivates the following definition.

Definition 4.5. *A function u , bounded in $\mathbb{R} \times [0, \infty)$, is called a weak (or integral) solution of problem (4.47) if equation (4.49) holds for every test function v in $\mathbb{R} \times [0, \infty)$.*

We point out that a weak solution may be discontinuous, since the definition requires only that it is bounded.

4.4.3 Piecewise smooth functions and the Rankine-Hugoniot condition

Definition 4.5 looks rather satisfactory, because of its flexibility. However we have to understand which information is hidden in the integral formulation about the behavior of weak solutions across a discontinuity curve. First we introduce the class of piecewise- C^1 functions:

Definition 4.6. *We say that $u : \mathbb{R} \times [0, \infty) \rightarrow \mathbb{R}$ is a piecewise- C^1 function if there exist a finite number of C^1 curves Γ_j , $j = 1, \dots, N$, of equation $x = s_j(t)$, defined in some interval $I_j \subseteq [0, +\infty)$, such that:*

- a) *Outside of $U_{j=1}^N \Gamma_j$, u is C^1 .*
- b) *u is C^1 up to each Γ_j from both sides.*

Thus, across each Γ_j , either u is continuous or it undergoes a jump discontinuity. Moreover, the jumps

$$u_+(s_j(t), t) - u_-(s_j(t), t)$$

are continuous on I_j , $j = 1, \dots, N$.

We now show that, in the class of piecewise- C^1 functions, the only admissible discontinuities in the half plane $t > 0$ are those prescribed by the Rankine-Hugoniot condition, as it is stated in the following theorem. Here the initial datum plays no role, so we examine weak solutions of the differential equation only, by considering test functions supported in $\mathbb{R} \times (0, \infty)$.

Theorem 4.7. *Let $u : \mathbb{R} \times [0, \infty) \rightarrow \mathbb{R}$ be a piecewise- C^1 function. Then u is a weak solution of problem (4.46) if and only if:*

1. *u is a classical solution of (4.47) in the region where u is C^1 .*
2. *The Rankine-Hugoniot jump condition holds along each discontinuity line $\Gamma_j \cap (\mathbb{R} \times (0, \infty))$:*

$$\dot{s}_j(t) = \frac{q(u_+(s_j(t), t)) - q(u_-(s_j(t), t))}{u_+(s_j(t), t) - u_-(s_j(t), t)}.$$

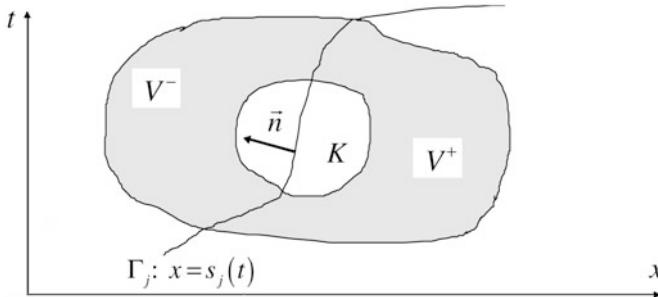


Fig. 4.18 Domain splitted by a discontinuity curve

Proof. Assume u is a weak solution of (4.47). We have already seen in the previous subsection that u is a classical solution of $u_t + q(u)_x = 0$ in any open set where is C^1 . To prove 2, let Γ_j be one of the discontinuity curves, of equation $x = s_j(t)$. Let V be an open set, contained in the half-plane $t > 0$, partitioned into two disjoint open domains V^+ and V^- by Γ_j as in Fig. 4.18.

In both V^+ and V^- u is a classical solution. Choose now a test function v , supported in a compact set $K \subset V$, such that $K \cap \Gamma_j$ is non empty. Since $v(x, 0) = 0$, we can write:

$$\begin{aligned} 0 &= \int_0^\infty \int_{\mathbb{R}} [uv_t + q(u)v_x] dx dt \\ &= \int_{V^+} [uv_t + q(u)v_x] dx dt + \int_{V^-} [uv_t + q(u)v_x] dx dt. \end{aligned}$$

Integrating by parts and observing that $v = 0$ on $\partial V^+ \setminus \Gamma_j$, we have:

$$\begin{aligned} &\int_{V^+} [uv_t + q(u)v_x] dx dt = \\ &= - \int_{V^+} [u_t + q(u)_x] v dx dt + \int_{\Gamma_j} [u_+ n_2 + q(u_+)n_1] v dl \\ &= \int_{\Gamma_j} [u_+ n_2 + q(u_+)n_1] v dl \end{aligned}$$

where u_+ denotes the value of u on Γ_j , from the V^+ side, $\mathbf{n} = (n_1, n_2)$ is the outward unit normal vector on ∂V^+ and dl denotes the arc length on Γ_j . Similarly, since \mathbf{n} is inward with respect to V^- :

$$\int_{V^-} [uv_t + q(u)v_x] dx dt = - \int_{\Gamma_j} [u_- n_2 + q(u_-)n_1] v dl$$

where u_- denotes the value of u on Γ_j , from the V^- side. Therefore we deduce that

$$\int_{\Gamma_j} \{[q(u_+) - q(u_-)] n_1 + [u_+ - u_-] n_2\} v dl = 0.$$

Since u is a piecewise- C^1 function, the jumps $[q(u_+) - q(u_-)]$ and $[u_+ - u_-]$ are continuous along Γ_j , and hence the arbitrariness of v yields

$$[q(u_+) - q(u_-)] n_1 + [u_+ - u_-] n_2 = 0 \quad (4.52)$$

on Γ_j . If u is continuous across Γ_j , (4.52) is automatically satisfied. If $u_+ \neq u_-$ we write the relation (4.52) more explicitly. Since $x = s_j(t)$ on Γ_j , we have

$$\mathbf{n} = (n_1, n_2) = \frac{1}{\sqrt{1 + (\dot{s}_j(t))^2}} (-1, \dot{s}_j(t)).$$

Hence (4.52) becomes, after simple calculations,

$$\dot{s}_j(t) = \frac{q[u_+(s_j(t), t)] - q[u_-(s_j(t), t)]}{u_+(s_j(t), t) - u_-(s_j(t), t)} \quad (4.53)$$

which is the Rankine-Hugoniot condition along Γ_j .

On the other hand, if u satisfies conditions 1 and 2, it is easy to check that u is a weak solution of problem (4.47). \square

By Theorem 4.7, any function constructed by connecting classical solutions and rarefaction waves in a continuous way is a weak solutions. The same is true for shock waves, since they satisfy the Rankine-Hugoniot condition. Thus, the solutions of the green light and of the traffic jam problems are precisely weak solutions.

Definition 4.5 gives a satisfactory answer to question Q1. The other two questions require a deeper analysis, as the following example shows.

Example 4.8. Non uniqueness. Imagine a flux of particles along the x -axis, each one moving with constant speed. Suppose that $u = u(x, t)$ represents the velocity field, which gives the speed of the particle located at x at time t . If $x = x(t)$ is the path of a particle, its velocity at time t is given by

$$\dot{x}(t) = u(x(t), t) \equiv \text{constant}.$$

Thus, we have

$$\begin{aligned} 0 &= \frac{d}{dt}u(x(t), t) = u_t(x(t), t) + u_x(x(t), t)\dot{x}(t) \\ &= u_t(x(t), t) + u_x(x(t), t)u(x(t), t). \end{aligned}$$

Therefore $u = u(x, t)$ satisfies Burgers equation

$$u_t + uu_x = u_t + \left(\frac{u^2}{2}\right)_x = 0 \quad (4.54)$$

which is a conservation law with $q(u) = u^2/2$. Note that q is strictly convex: $q'(u) = u$ and $q''(u) = 1$. We couple (4.54) with the initial condition $u(x, 0) = g(x)$, where

$$g(x) = \begin{cases} 0 & x < 0 \\ 1 & x > 0. \end{cases}$$

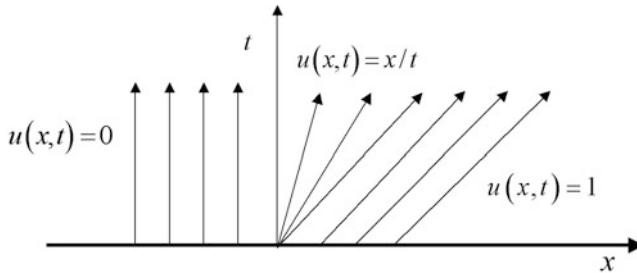


Fig. 4.19 The rarefaction wave in Example 4.8

The characteristics are the straight lines

$$x = g(x_0)t + x_0. \quad (4.55)$$

Therefore, $u = 0$ if $x < 0$ and $u = 1$ if $x > t$. The region $S = \{0 < x < t\}$ is not covered by characteristics. As in the green light problem, we connect the states 0 and 1 through a *rarefaction wave*. Since $q'(u) = u$, we have $r(s) = (q')^{-1}(s) = s$, so that we construct the weak solution (Fig. 4.19)

$$u(x, t) = \begin{cases} 0 & x \leq 0 \\ x/t & 0 < x < t \\ 1 & x \geq t. \end{cases} \quad (4.56)$$

However, u **is not the unique weak solution!** There exists also a *shock wave* solution. In fact, since

$$u_- = 0, u_+ = 1, q(u_-) = 0, q(u_+) = \frac{1}{2},$$

the Rankine-Hugoniot condition yields

$$\dot{s}(t) = \frac{q(u_+) - q(u_-)}{u_+ - u_-} = \frac{1}{2}.$$

Given the discontinuity at $x = 0$ of the initial data, the shock curve starts at $s(0) = 0$ and it is the straight line $x = \frac{t}{2}$. Hence, the function

$$w(x, t) = \begin{cases} 0 & x < \frac{t}{2} \\ 1 & x > \frac{t}{2} \end{cases}$$

is another weak solution (Fig. 4.20).

As we shall see, this shock wave has to be considered as a not physically acceptable solution.

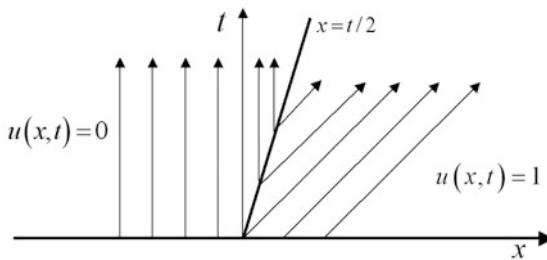


Fig. 4.20 A *nonphysical* shock

4.5 An Entropy Condition

The previous example shows that the answer to question **Q2** is *negative* and question **Q3** becomes relevant. We need a criterion to establish which is the physically correct solution.

Let us go back to classical solutions. From Proposition 4.3, p. 200, we have seen that the equation

$$G(x, t, u) \equiv u - g(x - q'(u)t) = 0$$

defines the unique classical solution u of problem (4.47), at least for small times. The Implicit Function Theorem gives

$$u_x(x, t) = -\frac{G_x(x, t, u)}{G_u(x, t, u)} = \frac{g'(x - q'(u)t)}{1 + tg'(x - q'(u)t)q''(u)}.$$

If we assume

$$g'(\xi) > 0, q''(u) \geq q''_{\min} > 0, \quad \text{for all } \xi \in \mathbb{R}, u \in \mathbb{R},$$

we get

$$u_x(x, t) \leq \frac{E}{t}$$

where $E = \frac{1}{q''_{\min}}$.

Using the mean value theorem we deduce the following condition¹²: there exists $E \geq 0$, such that, for every $x, z \in \mathbb{R}$, $z > 0$, and every $t > 0$,

$$u(x+z, t) - u(x, t) \leq \frac{E}{t}z. \quad (4.57)$$

Inequality (4.57) does not involve any derivative of u and makes perfect sense for discontinuous solutions as well. It turns out (see Theorem 4.10 below) that it

¹² For a suitable z^* between 0 and z , one has:

$$u(x+z, t) - u(x, t) = u_x(x+z^*)z.$$

constitutes an effective criterion to select physically admissible shocks. By analogy with gas dynamics, where a condition like (4.57) implies that the entropy increases across a shock, it is called **entropy condition**¹³. Actually, several types of entropy conditions have been formulated in the literature; another one similar to equation (4.61) below, is introduced in Subsect. 4.5.4.

A weak solution satisfying (4.57) is said to be an **entropic solution** and a number of consequences follows directly from it.

- The function

$$x \longmapsto u(x, t) - \frac{E}{t} x$$

is *decreasing*. In fact, let $x + z = x_2$, $x = x_1$ and $z > 0$. Then $x_2 > x_1$ and (4.57) is equivalent to

$$u(x_2, t) - \frac{E}{t} x_2 \leq u(x_1, t) - \frac{E}{t} x_1. \quad (4.58)$$

A remarkable consequence is that, for every $t > 0$, $u(\cdot, t)$ is continuous or it has at most a countable number of jump discontinuities.

- If x is a jump discontinuity point for $u(\cdot, t)$, then

$$u_+(x, t) < u_-(x, t), \quad (4.59)$$

where

$$u_{\pm}(x, t) = \lim_{y \rightarrow x^{\pm}} u(y, t).$$

To show it, choose $x_1 < x < x_2$ and let x_1 and x_2 both go to x in (4.58).

- Since q is **strictly convex**, (4.59) is equivalent to

$$q'(u_+) < \sigma(u_-, u_+) < q'(u_-), \quad (4.60)$$

where we have set

$$\sigma(u_-, u_+) = \frac{q(u_+) - q(u_-)}{u_+ - u_-}$$

and also to

$$\sigma(u_-, u_+) < \sigma(u_-, u_*), \quad (4.61)$$

for every state u_* between u_- and u_+ .

- Finally, if $x = s(t)$ is a shock curve, by the Rankine-Hugoniot jump condition, (4.60) reads

$$q'(u_+(s, t)) < \dot{s} < q'(u_-(s, t)) \quad (4.62)$$

along the curve.

¹³ The denomination is due to *Courant and Friedrichs, 1948*.

Remark 4.9. If

$$g' < 0, \quad q'' < q''_{\max} < 0,$$

the inequality (4.57) changes into

$$u(x+z, t) - u(x, t) \geq -\frac{E}{t}z$$

with $E = \frac{1}{|q''_{\max}|}$. Thus, at a jump discontinuity (x, t) we have

$$u_+(x, t) > u_-(x, t)$$

while (4.60), (4.61) and (4.62) remain unchanged.

The condition (4.62) is called **entropy inequality** and it has a remarkable meaning: *the states on the left (right) of the shock curve tends to propagate faster (slower) than the shock*, giving rise to a phenomenon called *self-sharpening*, since no state can cross the shock curve (see Example 4.4, p. 201). Thus the characteristics **hit forward** in time the shock line. In other words, it is not possible to go back in time along a characteristic and hit a shock line, expressing a sort of **irreversibility** after a shock.

The geometrical meaning of conditions (4.60) and (4.61) is illustrated in Fig. 4.21 in the concave case. In both cases (concave or convex), (4.61) implies that going along the chord AB from A to B one "sees" the graph of q on the left. Equivalently, in the concave case, we have $u_+ > u_-$, $u_+ < u_-$ in the convex case.

The above considerations lead us to select the entropy solutions as the only physically meaningful ones.

On the other hand, if the characteristics hit a shock curve backward in time, the discontinuity wave is to be considered *non-physical*.

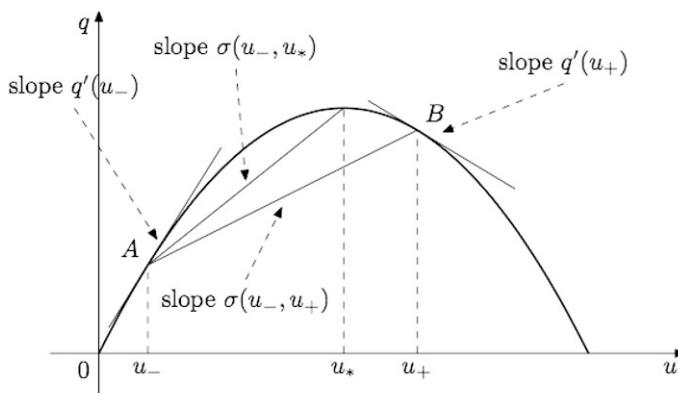


Fig. 4.21 Geometrical meaning of the entropy inequality

Thus, in Example 4.8, p. 207, the solution w represents a non-physical shock since it does not satisfy the entropy condition. The correct solution is therefore the simple wave (4.56).

The following result holds¹⁴.

Theorem 4.10. *If $q \in C^2(\mathbb{R})$ is strictly convex or concave and g is bounded, there exists a unique entropy solution of the problem*

$$\begin{cases} u_t + q(u)_x = 0 & x \in \mathbb{R}, t > 0 \\ u(x, 0) = g(x) & x \in \mathbb{R} \end{cases} \quad (4.63)$$

and $|u(x, t)| \leq \sup_{\mathbb{R}} |g| = M$. Moreover, let v be another entropic solution of (4.63) with initial data h such that $\sup_{\mathbb{R}} |h| \leq M$. Then the following stability estimates holds, for every $x_1, x_2 \in \mathbb{R}$, $t > 0$:

$$\int_{x_1}^{x_2} |u(x, t) - v(x, t)| dx \leq \int_{x_1 - At}^{x_2 + At} |g(x) - h(x)| dx \quad (4.64)$$

where $A = \max |q'|$ over the interval $[-M, M]$.

4.6 The Riemann problem

4.6.1 Convex/concave flux function

We now apply Theorem 4.10 to solve explicitly problem (4.47) with initial data

$$g(x) = \begin{cases} u_R & x > 0 \\ u_L & x < 0, \end{cases} \quad (4.65)$$

where u_R and u_L are constant states, $u_R \neq u_L$. This problem is known as **Riemann problem**, and it is particularly important also in the construction of numerical approximation methods.

Assume that q is *strictly* convex or concave. We have:

Theorem 4.11. *Let $q \in C^2(\mathbb{R})$ be strictly convex and $q'' \geq h > 0$ or strictly concave and $q'' \leq -h < 0$.*

- a) *If $q'' \geq h$ and $u_L > u_R$ or $q'' \leq -h$ and $u_- < u_+$, the unique entropy solution is given by the shock wave*

$$u(x, t) = \begin{cases} u_R & x > \sigma(u_L, u_R)t \\ u_L & x < \sigma(u_L, u_R)t \end{cases} \quad (4.66)$$

¹⁴ For the proof, see e.g. [18], Smoller, 1983.

where

$$\sigma(u_L, u_R) = \frac{q(u_R) - q(u_L)}{u_R - u_L}.$$

- b) If $q'' \geq h$ and $u_L \leq u_R$ or $q'' \leq -h$ and $u_L > u_R$, the unique entropy solution is given by the rarefaction wave

$$u(x, t) = \begin{cases} u_L & x < q'(u_L)t \\ r\left(\frac{x}{t}\right) & q'(u_L)t < x < q'(u_R)t \\ u_R & x > q'(u_R)t \end{cases}$$

where $r = (q')^{-1}$ is the inverse function of q' .

Proof. We consider only the convex case; the concave case is perfectly similar.

a) Since clearly $u_+ = u_R$ and $u_- = u_L$, the discontinuous solution (4.66) satisfies the Rankine Hugoniot condition and therefore it is clearly a weak solution. Moreover, since $u_R < u_L$ the entropy condition (4.57) holds as well, so that u is the unique entropic solution of problem (4.65) by Theorem 4.10.

b) Since

$$r(q'(u_R)) = u_R \quad \text{and} \quad r(q'(u_L)) = u_L,$$

u is continuous in the half-plane $t > 0$. We now check that u satisfies the equation

$$u_t + q(u)_x = 0$$

in the region

$$S = \left\{ (x, t) : q'(u_L) < \frac{x}{t} < q'(u_R) \right\}.$$

Let $u(x, t) = r\left(\frac{x}{t}\right)$. We have:

$$u_t + q(u)_x = -r'\left(\frac{x}{t}\right)\frac{x}{t^2} + q'(r)r'\left(\frac{x}{t}\right)\frac{1}{t} = r'\left(\frac{x}{t}\right)\frac{1}{t}\left[q'(r) - \frac{x}{t}\right] \equiv 0.$$

Thus, u is a weak solution in the upper half-plane.

Let us check the entropy condition. We consider only the case

$$q'(u_L)t \leq x < x + z \leq q'(u_R)t$$

leaving the other ones to the reader. Since $q'' \geq h > 0$, we have

$$r'(s) = \frac{1}{q''(r)} \leq \frac{1}{h} \quad (s = q'(r)),$$

so that, for a suitable z^* , $0 < z^* < z$,

$$\begin{aligned} u(x+z) - u(x, t) &= r\left(\frac{x+z}{t}\right) - r\left(\frac{x}{t}\right) \\ &= r'\left(\frac{x+z^*}{t}\right)\frac{z}{t} \leq \frac{1}{h}\frac{z}{t} \end{aligned}$$

which is the entropy condition with $E = 1/h$. \square

4.6.2 Vanishing viscosity method

There is another instructive and perhaps more natural way to construct entropic solutions of the conservation law

$$u_t + q(u)_x = 0, \quad (4.67)$$

the so called *vanishing viscosity method*. This method consists in viewing equation (4.67) as the limit for $\varepsilon \rightarrow 0^+$ of the equation

$$u_t + q(u)_x = \varepsilon u_{xx}, \quad (4.68)$$

which is a second order conservation law with flux function

$$\tilde{q}(u, u_x) = q(u) - \varepsilon u_x, \quad (4.69)$$

where ε is a *small positive* number. Although we recognize εu_{xx} as a diffusion term, this kind of model arises mostly in fluid dynamics where u is the fluid velocity and ε is its *viscosity*, from which comes the name of the method.

There are several good reasons in favor of this approach. First of all, a small amount of diffusion or viscosity makes the mathematical model more realistic in most applications. Note that εu_{xx} becomes relevant only when u_{xx} is large, that is in a region where u_x changes rapidly and a shock occurs. For instance in our model of traffic dynamics, it is natural to assume that drivers would slow down when they see increased (relative) density ahead. Thus, an appropriate model for their velocity is

$$\tilde{v}(\rho) = v(\rho) - \varepsilon \frac{\rho_x}{\rho}$$

which corresponds to the flux function $\tilde{q}(\rho) = \rho v(\rho) - \varepsilon \rho_x$ for the flow-rate of cars.

Another reason comes from the fact that shocks constructed by the vanishing viscosity method are *physical shocks*, since they satisfy the entropy inequality. This is due, in principle, to the time irreversibility present in a diffusion equation that, in the limit, selects the physical admissible solution.

As for the heat equation, we expect to obtain smooth solutions of (4.68) even with discontinuous initial data. On the other hand, the nonlinear term may force the evolution towards a shock wave.

Here we are interested in solutions of (4.68) connecting two constant states u_L and u_R , that is, satisfying the conditions

$$\lim_{x \rightarrow -\infty} u(x, t) = u_L, \quad \lim_{x \rightarrow +\infty} u(x, t) = u_R. \quad (4.70)$$

Since we are looking for shock waves, it is reasonable to seek a solution depending only on a coordinate $\xi = x - vt$ moving with the (unknown) shock speed v . Thus,

let us look for *bounded travelling waves* solution of (4.68) of the form

$$u(x, t) = U(x - vt) \equiv U(\xi),$$

with

$$U(-\infty) = u_L \quad \text{and} \quad U(+\infty) = u_R \quad (4.71)$$

and $u_L \neq u_R$. We have

$$u_t = -v \frac{dU}{d\xi}, \quad u_x = \frac{dU}{d\xi}, \quad u_{xx} = \frac{d^2U}{d\xi^2},$$

so that we obtain for U the ordinary differential equation

$$(q'(U) - v) \frac{dU}{d\xi} = \varepsilon \frac{d^2U}{d\xi^2}$$

which can be integrated to yield

$$q(U) - vU + A = \varepsilon \frac{dU}{d\xi}$$

where A is an arbitrary constant. Assuming that $\frac{dU}{d\xi} \rightarrow 0$ as $\xi \rightarrow \pm\infty$ and using (4.71) we get

$$q(u_L) - vu_L + A = 0 \quad \text{and} \quad q(u_R) - vu_R + A = 0. \quad (4.72)$$

Subtracting these two equations we find

$$v = \frac{q(u_R) - q(u_L)}{u_R - u_L} \equiv \bar{v} \quad (4.73)$$

and then

$$A = \frac{-q(u_R)u_L + q(u_L)u_R}{u_R - u_L} \equiv \bar{A}.$$

Thus, if there exists a travelling wave solution satisfying conditions (4.70), it moves with a speed \bar{v} predicted by the Rankine-Hugoniot formula. Still it is not clear whether such travelling wave solution exists. To verify this, examine the equation

$$\varepsilon \frac{dU}{d\xi} = q(U) - \bar{v}U + \bar{A}. \quad (4.74)$$

From (4.72), equation (4.74) has the two equilibria $U = u_R$ and $U = u_L$. Moreover, conditions (4.71) require u_R to be *asymptotically stable* and u_L *unstable*. At this point, we need to have information on the shape of q .

Assume $q'' < 0$. Then the phase diagram for eq. (4.74) is described in Fig. 4.22, for the two cases $u_L > u_R$ and $u_L < u_R$. Between u_L and u_R , $q(U) - \bar{v}U + \bar{A} > 0$ and, as the arrows indicate, U is *increasing*. We see that only the case $u_L < u_R$

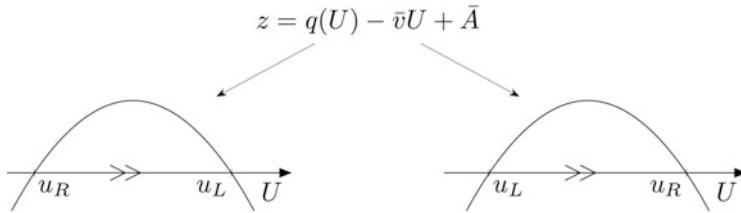


Fig. 4.22 Case (b) only is compatible with conditions (4.70)

is compatible with conditions (4.71) and this corresponds precisely to a shock formation for the non diffusive conservation law. Thus,

$$q'(u_L) - \bar{v} > 0 \quad \text{and} \quad q'(u_R) - \bar{v} < 0$$

or

$$q'(u_R) < \bar{v} < q'(u_L) \quad (4.75)$$

which is *the entropy inequality*.

Similarly, if $q'' > 0$, a travelling wave solution connecting the two states u_R and u_L exists only if $u_L > u_R$ and (4.75) holds.

Let us see what happens when $\varepsilon \rightarrow 0$. Assume $q'' < 0$. For ε small, we expect that our travelling wave increases abruptly from a value $U(\xi_1)$ close to u_L to a value $U(\xi_2)$ close to u_R , within a narrow region called the *transition layer*. For instance, we may choose ξ_1 and ξ_2 such that

$$U(\xi_2) - U(\xi_1) \geq (1 - \beta)(u_R - u_L)$$

with a positive β , very close to 0. We call the number $\varkappa = \xi_2 - \xi_1$ *thickness* of the transition layer. To compute it, we separate the variables U and ξ in (4.74) and integrate over (ξ_1, ξ_2) ; this yields

$$\xi_2 - \xi_1 = \varepsilon \int_{U(\xi_1)}^{U(\xi_2)} \frac{ds}{q(s) - \bar{v}s + \bar{A}}.$$

Thus, the thickness of the transition layer is proportional to ε . As $\varepsilon \rightarrow 0$, the transition region becomes narrower and narrower, and eventually a shock wave that satisfies the entropy inequality is obtained.

This phenomenon is clearly seen in the important case of *viscous Burgers equation* that we examine in more details in the next subsection.

Example 4.12. Burgers shock solution. Let us determine a travelling wave solution of the viscous Burgers equation

$$u_t + uu_x = \varepsilon u_{xx} \quad (4.76)$$

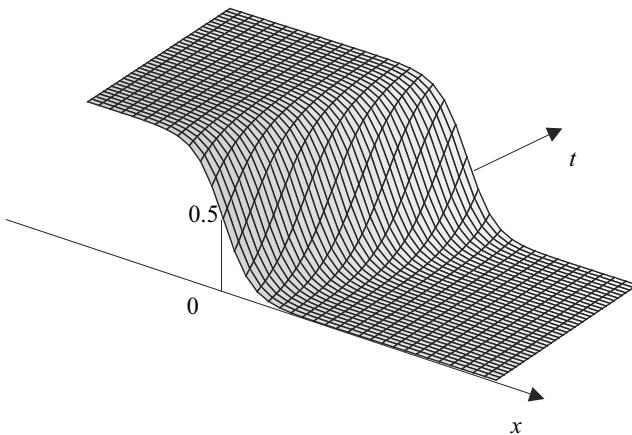


Fig. 4.23 The travelling wave in Example 4.12, with $c = 1$

connecting the states $u_L = 1$ and $u_R = 0$. Note that

$$q(u) = u^2/2$$

is convex. Then $\bar{v} = 1/2$ and $\bar{A} = 0$. Equation (4.74) becomes

$$2\varepsilon \frac{dU}{d\xi} = U^2 - U$$

that can be easily integrated to give (recall that $0 < U < 1$),

$$U(\xi) = \frac{1}{1 + c \exp\left(\frac{\xi}{2\varepsilon}\right)} \quad (c > 0).$$

Thus we find a family of travelling waves given by (see Fig. 4.23)

$$u(x, t) = U\left(x - \frac{t}{2}\right) = \frac{1}{1 + c \exp\left(\frac{2x-t}{4\varepsilon}\right)}. \quad (4.77)$$

When $\varepsilon \rightarrow 0^+$,

$$u(x, t) \rightarrow w(x, t) = \begin{cases} 0 & x > t/2 \\ 1 & x < t/2 \end{cases}$$

which is the entropic shock solution for the non viscous Burgers equation with initial data 1 if $x < 0$ and 0 if $x > 0$.

4.6.3 The viscous Burgers equation

The viscous Burgers equation is one of the most celebrated examples of nonlinear diffusion equation. It arose (*Burgers, 1948*) as a simplified form of the Navier-Stokes equation, in an attempt to study some aspects of turbulence. It appears also in gas dynamics, in the theory of sound waves and in traffic flow modelling, and it constitutes a basic example of competition between *dissipation* (due to linear diffusion) and *steepening* (shock formation due to the nonlinear transport term uu_x).

The success of Burgers equation is largely due to the rather surprising fact that the initial value problem can be solved analytically. In fact, via the so called *Hopf-Cole transformation*, Burgers equation is converted into the heat equation. Let us see how this can be done. First, let us write Burgers equation in the form

$$\frac{\partial u}{\partial t} + \frac{\partial}{\partial x} \left(\frac{1}{2} u^2 - \varepsilon u_x \right) = 0.$$

Then, the planar vector field $(-u, \frac{1}{2}u^2 - \varepsilon u_x)$ is curl-free and therefore there exists a potential $\psi = \psi(x, t)$ such that

$$\psi_x = -u \quad \text{and} \quad \psi_t = \frac{1}{2}u^2 - \varepsilon u_x.$$

Thus, ψ solves the equation

$$\psi_t = \frac{1}{2}\psi_x^2 + \varepsilon\psi_{xx}. \quad (4.78)$$

Now we try to get rid of the quadratic term by letting $\psi = g(\varphi)$ and suitably choosing g . We have:

$$\psi_t = g'(\varphi)\varphi_t, \quad \psi_x = g'(\varphi)\varphi_x, \quad \psi_{xx} = g''(\varphi)(\varphi_x)^2 + g'(\varphi)\varphi_{xx}.$$

Substituting into (4.78) we find

$$g'(\varphi)[\varphi_t - \varepsilon\varphi_{xx}] = \left[\frac{1}{2}(g'(\varphi))^2 + \varepsilon g''(\varphi) \right] (\varphi_x)^2.$$

Hence, if we choose $g(s) = 2\varepsilon \log s$, then the right hand side vanishes and we are left with

$$\varphi_t - \varepsilon\varphi_{xx} = 0. \quad (4.79)$$

Thus

$$\psi = 2\varepsilon \log \varphi$$

and from $u = -\psi_x$ we obtain

$$u = -2\varepsilon \frac{\varphi_x}{\varphi}, \quad (4.80)$$

which is the *Hopf-Cole transformation*.

An initial data

$$u(x, 0) = u_0(x) \quad (4.81)$$

transforms into an initial data of the form¹⁵

$$\varphi_0(x) = \exp \left\{ - \int_a^x \frac{u_0(z)}{2\varepsilon} dz \right\} \quad (a \in \mathbb{R}). \quad (4.82)$$

If (see Theorem 2.12, p. 78)

$$\frac{1}{x^2} \int_a^x u_0(z) dz \rightarrow 0 \quad \text{as } |x| \rightarrow \infty,$$

the initial value problem (4.79), (4.82) has a unique smooth solution in the half-plane $t > 0$, given by

$$\varphi(x, t) = \frac{1}{\sqrt{4\pi\varepsilon t}} \int_{-\infty}^{+\infty} \varphi_0(y) \exp \left(-\frac{(x-y)^2}{4\varepsilon t} \right) dy.$$

This solution is continuous with its x -derivative up to $t = 0$ at any continuity point of u_0 .

Consequently, from (4.80), problem

$$\begin{cases} u_t + uu_x = \varepsilon u_{xx} & x \in \mathbb{R}, t > 0, \\ u(x, 0) = u_0(x) & x \in \mathbb{R} \end{cases} \quad (4.83)$$

has a unique smooth solution in the half-plane $t > 0$, continuous up to $t = 0$ at any continuity point of u_0 , given by

$$u(x, t) = \frac{\int_{-\infty}^{+\infty} \varphi_0(y) \frac{x-y}{t} \exp \left(-\frac{(x-y)^2}{4\varepsilon t} \right) dy}{\int_{-\infty}^{+\infty} \varphi_0(y) \exp \left(-\frac{(x-y)^2}{4\varepsilon t} \right) dy}. \quad (4.84)$$

We use formula (4.84) to solve an initial pulse problem.

Example 4.13. Initial pulse (see Fig. 4.24). Consider problem (4.76), (4.76) with the initial condition

$$u_0(x) = M\delta(x)$$

¹⁵ The choice of a is arbitrary and does not affect the value of u .

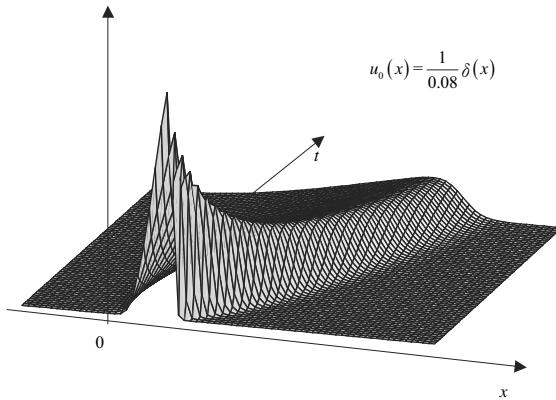


Fig. 4.24 Evolution of an initial pulse for the viscous Burgers equation ($M = 1, \varepsilon = 0.04$)

where δ denotes the Dirac measure at the origin. We have, choosing $a = 1$ in formula (4.82),

$$\varphi_0(x) = \exp \left\{ - \int_1^x \frac{u_0(y)}{2\varepsilon} dy \right\} = \begin{cases} 1 & x > 0 \\ \exp \left(\frac{M}{2\varepsilon} \right) & x < 0. \end{cases}$$

Formula (4.84), gives, after some routine calculations,

$$u(x, t) = \sqrt{\frac{4\varepsilon}{\pi t}} \frac{\exp \left(-\frac{x^2}{4\varepsilon t} \right)}{\frac{2}{\exp(M/2\varepsilon) - 1} + \operatorname{erfc} \left(\frac{x}{\sqrt{4\varepsilon t}} \right)}$$

where

$$\operatorname{erfc}(x) = \frac{2}{\sqrt{\pi}} \int_x^{+\infty} e^{-z^2} dz \quad (4.85)$$

is the *complementary error function*.

4.6.4 Flux function with inflection points

When the flux curve exhibits a finite number inflection points we have to find other criteria to select physically admissible discontinuities. In these cases, limiting ourselves to the class of piecewise- C^1 weak solutions, a slight generalization of (4.61), that we call *condition E*, acts as a proper entropy condition. It needs:

$$\sigma(u_-, u_+) \leq \sigma(u_-, u_*) \quad (4.86)$$

along a discontinuity curve, for every state u_* between u_- and u_+ . Observe that now we do not know *a priori* which is the greater value between u_- and u_+ . Also note that the equality sign in (4.86) includes also the linear case $q(u) = vu$, v constant. In this case, the jump discontinuity propagates with the same speed of the states on both sides: the corresponding wave is called *contact discontinuity* (see Fig. 4.3, p. 185). The following uniqueness theorem holds¹⁶:

Theorem 4.14. *There exists at most a piecewise- C^1 weak solution of problem (4.47), satisfying condition (4.86).*

Consider now a Riemann problem with states u_L for $x < 0$ and u_R for $x > 0$. According to (4.86), the states u_L, u_R can be connected by an admissible shock as long as the *chord* joining the two points $(u_L, q(u_L))$ and $(u_R, q(u_R))$ lies above (resp. below) the graph of q if $u_R < u_L$ (resp. $u_R > u_L$).

On the other hand, we know that the states u_L, u_R can be connected by a rarefaction wave as long as the *arc* joining the two points $(u_L, q(u_L))$ and $(u_R, q(u_R))$ is strictly *concave* (resp. *convex*) if $u_R < u_L$ (resp. $u_R > u_L$).

An immediate consequence is that, if q has inflection points, in general the Riemann problem cannot be solved in by a single wave.

We first examine the **case $u_R < u_L$** . Referring, for instance, to Figure 4.25, the states u_L and u_R cannot be directly connected by a single shock since the *chord* joining $(u_R, q(u_R))$, $(u_L, q(u_L))$ would cross the flux curve, violating the entropy condition (4.86).

Also we cannot use a single rarefaction wave since q' is not invertible between u_R and u_L . Thus, to construct the entropic solution, the right (and only!) way to proceed is to introduce suitable intermediate states that can be connected either by an admissible discontinuity (entropic shock or contact discontinuity) or by a rarefaction wave.

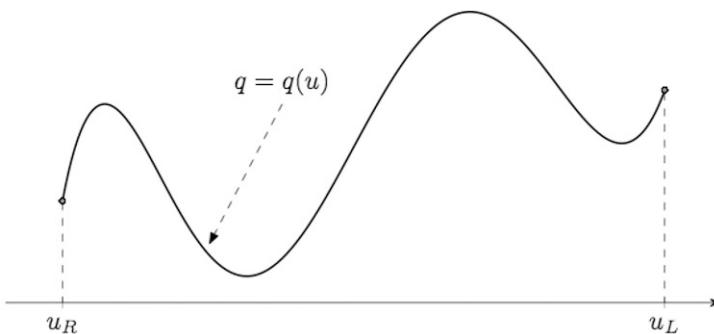


Fig. 4.25 Flux function with 3 inflection points

¹⁶ See O.A. Oleinik, Uniqueness and stability of the generalized solution of the Cauchy Problem for a quasilinear equation, Amer. Math. Soc. Transl., Ser 2, 33, 285–290, 1963.

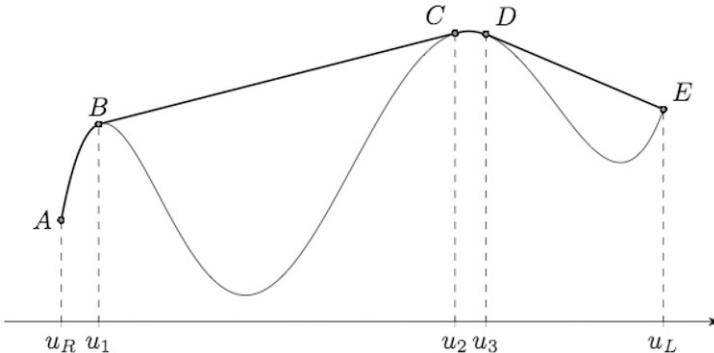


Fig. 4.26 Upper concave envelope of the function in Fig. 4.25

To this purpose we introduce the *upper concave envelope* q^c of q in $[u_R, u_L]$, that is the lowest concave function that lives over q in the interval $[u_R, u_L]$. Figure 4.26 shows q^c for the flux function in Fig. 4.25, which has 3 inflection points.

We use this example to describe the construction of the entropic solution. We see that the interval $[u_R, u_L]$ is decomposed into the subintervals $[u_R, u_1]$ and $[u_2, u_3]$, in which q_c is strictly concave (*rarefaction intervals*), alternating with the subintervals $[u_1, u_2]$ and $[u_3, u_L]$ where the graph of q^c is a straight line (*shock intervals*). Then our entropic solution has the following structure, where r denotes the inverse of q' on each rarefaction interval:

- in the sector $t > 0, x < q'(u_3)t = \sigma(u_L, u_3)t$, we have $u = u_L$;
 - across the shock line $x = \sigma(u_L, u_3)t$, u jumps down from $u = u_L$ to $u = u_3$.
- Note that only the characteristics of equation $x = q'(u_L)t + \xi$, $\xi < 0$, coming from

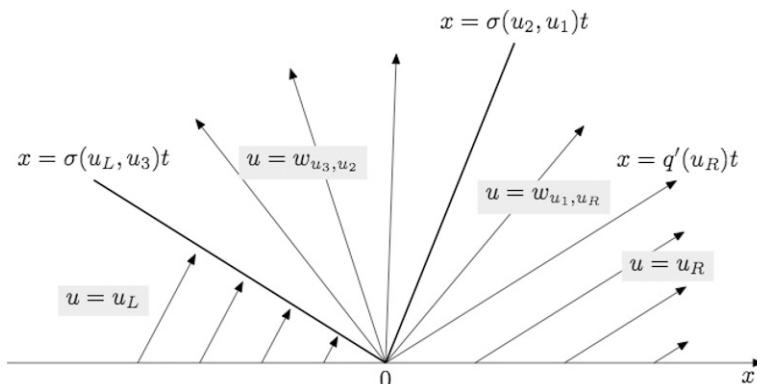


Fig. 4.27 Back semishock followed by a rarefaction wave, a contact discontinuity and then by a front rarefaction wave. Characteristics configuration

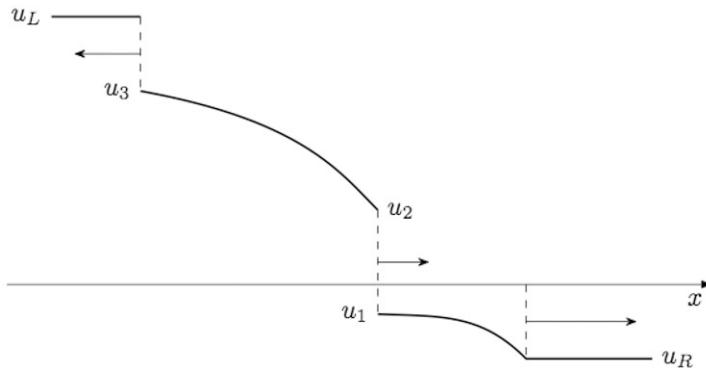


Fig. 4.28 Profile at a fixed time of the solution in Fig. 4.27

the left, impinge on the shock line. For this reason this kind of discontinuity is more properly called a *semishock*;

- the states u_3 and u_2 are connected by the rarefaction wave $w_{u_3,u_2}(x,t) = r(x/t)$, defined in the sector

$$\sigma(u_L, u_3)t < x < \sigma(u_2, u_1)t;$$

- across the shock line $x = \sigma(u_2, u_1)t$, u jumps down from $u = u_2$ to $u = u_1$. Since

$$\sigma(u_2, u_1) = q'(u_2) = q'(u_1),$$

the jump discontinuity moves at the same speed of the states at both sides and therefore it is a *contact discontinuity*;

- the states u_1 and u_R are connected by the rarefaction wave $w_{u_1,u_R}(x,t) = r(x/t)$, defined in the sector

$$\sigma(u_2, u_1)t < x < q'(u_R)t;$$

- finally, in the sector $t > 0$, $q'(u_R)t < x$, $u = u_R$.

The characteristic configuration and the profile of u at time t are described in Fig. 4.27 and 4.28, respectively.

The **case** $u_L < u_R$ can be treated similarly. This time one needs to introduce the lower convex envelope q^c of q . The interval $[u_L, u_R]$ is decomposed into subintervals in which q_c is strictly convex (*rarefaction intervals*), alternating with subintervals where the graph of q_c is a straight line (*shock intervals*). Then the construction of the entropic solution of the Riemann problem proceed as in the previous case (see Problem 4.13).

4.7 An Application to a Sedimentation Problem

The following example is taken from sedimentation theory¹⁷. A large number of particles, whose concentration is denoted by c , is suspended in a fluid inside a cylindrical container and falls with speed $v = v(c)$. Denote by $\Phi(c)$ the downward flux of particles, that is the number of particles that cross a horizontal plane, per unit area, per unit time. Then

$$\Phi(c) = cv(c).$$

We want to analyze the sedimentation process of the suspension under the following assumptions:

- i) The concentration is uniform over horizontal planes; the sediment is incompressible.
- ii) With reference to Fig. 4.29, initially, the concentration is $c = 0$ at the top $h + h_0$, $c = c_0$ from height h_0 to $h + h_0$, and $c = c_{\max}$, the maximum possible concentration, from the bottom to the height h_0 .

Then we may adopt a one dimensional model and write $c = c(z, t)$ where the z axis is placed along the axis of the cylinder, with origin at the height $h + h_0$ and oriented downward. Neglecting the part of the container from the bottom to the height h_0 , which actually plays no role, the usual argument based of conservation of mass leads to the equation

$$c_t + \Phi(c)_z = 0, \quad 0 < z < h, t > 0,$$

with the initial condition

$$c(z, 0) = c_0, \quad 0 < z < h$$

and the boundary conditions

$$c(0, t) = 0, \quad c(h, t) = c_{\max}, \quad t > 0.$$

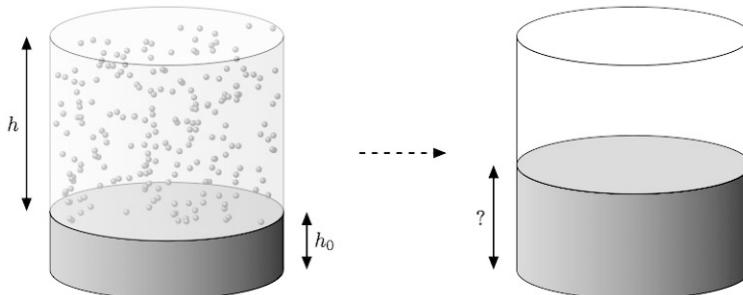


Fig. 4.29 Sedimentation of a suspension

¹⁷ See [27], Rhee, Aris, Amundson, 1986.

We expect that the system evolves towards a final stationary configuration. We want to describe the evolution process, to determine the final concentration profile and the time at which it is achieved.

Of course, to answer the above questions we need a constitutive equation for the speed $v(c)$. For simplicity, we shall adopt the following law:

$$v(c) = v_s \left(1 - \frac{c}{c_{\max}}\right) \left(1 - \beta \frac{c}{c_{\max}}\right) \quad (4.87)$$

where $1/2 < \beta < 1$. Here v_s is given by Stokes' law for the terminal velocity, that a particle attains when falling freely in a fluid. Formula (4.87) reflects the fact that at low concentration the particle speed decreases linearly from v_s while at higher concentrations v decreases faster than linearly. Note that, since $\beta < 1$, the flux curve

$$\Phi(c) = cv_s \left(1 - \frac{c}{c_{\max}}\right) \left(1 - \beta \frac{c}{c_{\max}}\right)$$

exhibits an inflexion point between 0 and c_{\max} .

It is convenient to switch to dimensionless variables by setting

$$x = \frac{z}{h}, \tau = \frac{v_0 t}{h}, u = \frac{c}{c_{\max}}. \quad (4.88)$$

The equation for u becomes

$$u_\tau + q(u)_x = 0$$

with flux function (see Fig. 4.30)

$$q(u) = \frac{\Phi(c_{\max}u)}{v_0 c_{\max}} = u(1-u)(1-\beta u), \quad (4.89)$$

initial condition

$$u(x, 0) = \frac{c_0}{c_{\max}} \equiv u_0, \quad 0 < x < 1,$$

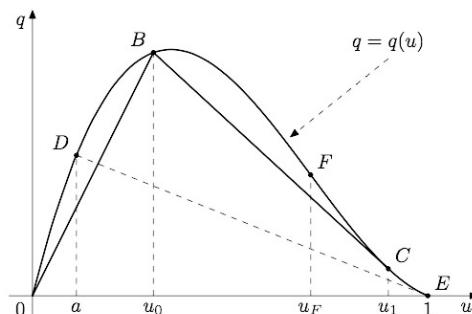


Fig. 4.30 The flux function for the sedimentation problem in dimensionless variables

and boundary conditions

$$u(0, \tau) = 0, \quad u(h, \tau) = 1, \quad \tau > 0. \quad (4.90a)$$

The curve $q = q(u)$ has an inflection point at

$$u_F = \frac{1 + \beta}{3\beta}$$

and intersects the tangent line issued from $E = (1, 0)$, at the point $D = (a, q(a))$, where $a = \frac{1-\beta}{\beta}$. Moreover,

$$q'(u) = 1 - 2(1 + \beta)u + 3\beta u^2$$

and

$$\sigma(u_-, u_+) = \frac{q(u_-) - q(u_+)}{u_- - u_+} = 1 - (1 + \beta)(u_- + u_+) + \beta(u_-^2 + u_- u_+ + u_+^2).$$

The evolution of u depends very much from the location of u_0 with respect to a and u_F . In fact, with reference to Fig. 4.30, we distinguish the following cases:

- (a) $0 < u_0 \leq a$
- (b) $a < u_0 < u_F$
- (c) $u_F \leq u_0 < 1$.

The most interesting one is case (b) and we analyze in details only this case, which is described in Fig. 4.30. For the other two cases we refer to Problem 4.13.

Thus, let $a < u_0 < u_F$. For the sake of coherence with the conventions adopted in the previous sections, in the plane of the variables x and τ , the x -axis will be the horizontal one, oriented as usual, keeping in mind that *moving in the positive direction corresponds to falling for the suspension*. We construct the solution in several steps.

- *Connecting 0 to u_0 through a shock.* Since u_0 stays on the concave arc DF on the flux curve, the two states 0 and u_0 are connected by a *front*¹⁸ *shock* Γ_0 of equation

$$x = \sigma(0, u_0)\tau = \frac{q(u_0)}{u_0}\tau = (1 - u_0)(1 - \beta u_0)\tau.$$

This shock represents the downward motion of the upper surface of the suspension at the constant speed $\sigma(0, u_0)$.

- *Connecting u_0 with an intermediate state u_1 through a semishock.* Now we would like to connect the two states u_0 and 1. This is not possible by using a single shock since the *chord BE* would cross the flux curve, violating the entropy condition (4.86). We have seen in Sect. 4.5.4 that the right way to construct an entropy solution is to introduce a state u_1 , intermediate between u_0 and 1, such that the corresponding cord BC is tangent at C to the flux curve (see Fig. 4.30). In this way we can connect u_0 and u_1 through an admissible back¹⁹ shock and then connect u_1 and 1 by means of a rarefaction wave, since the arc CE is strictly convex.

¹⁸ That is with positive speed.

¹⁹ That is with negative speed.

The intermediate state u_1 can be found by solving the equation $q'(u_1) = \sigma(u_0, u_1)$. This gives $u_1 = (1 + \beta - \beta u_0)/2\beta$. Note that $q'(u_1)$ is negative. The shock curve Γ_1 that connects the states u_0 and u_1 has equation

$$x = q'(u_1)\tau + 1.$$

Since Γ_1 coincides with the characteristic $x = q'(u_1)\tau + 1$, this back shock is a *semishock* and models the upward motion of a sediment layer of concentration u_1 , at the constant speed $q'(u_1)$.

- *Connecting u_1 to 1 through a rarefaction wave.* As we have already noted, due to the strict convexity of the arc CE , we can connect the states u_1 and 1 through a rarefaction wave $u(x, t) = (q')^{-1}(\frac{x-1}{\tau})$, centered at $(1, 0)$, which is delimited by the right most characteristic

$$x = q'(1)\tau + 1 = -(1 - \beta)\tau + 1$$

and it fans backward up to the characteristic

$$x = q'(u_1)\tau + 1,$$

which coincide with Γ_1 . This corresponds to the build-up from the bottom of sediment layers of concentrations from 1 to u_1 , at speed increasing from $q'(1)$ to $q'(u_1)$.

- *Shock merging and interaction with the rarefaction wave.* The front shock Γ_0 and the back shock Γ_1 merge at the point of coordinates

$$\tau_0 = \frac{u_0}{q(u_0) - u_0 q'(u_1)} = \frac{4\beta}{(1 + \beta - \beta u_0)^2}, \quad x_0 = (1 - u_0)(1 - \beta u_0)\tau_0. \quad (4.91)$$

After τ_0 , the state $u = 0$ interacts directly with the rarefaction wave. This gives rise to a new shock arc Γ_2 , starting from (x_0, τ_0) . See Fig. 4.31.

To determine the equation of Γ_2 , rather than computing directly the values of $u(x, \tau) = (q')^{-1}(\frac{x-1}{\tau})$, it is easier to look for a parametric representation $x = x(u)$, $\tau = \tau(u)$, where the values of u on Γ_2 , carried from the right by the characteristics of the rarefaction wave, play the role of the parameter. This can be done observing that Γ_2 is described by the system of the following two equations:

$$\frac{dx}{d\tau} = \sigma(0, u) = \frac{q(u)}{u}, \quad (\text{Rankine-Hugoniot}) \quad (4.92)$$

$$x = q'(u)\tau + 1, \quad (\text{characteristic of the rarefaction wave}). \quad (4.93)$$

Since $\sigma(0, u)$ is positive and decreasing, Γ_2 is a *front* shock modeling a deceleration in the fall of the upper surface of the suspension, from speed $\sigma(0, u_1)$ to 0.

Differentiating the second equation with respect to u , we get

$$\frac{dx}{du} = q''(u)\tau + q'(u)\frac{d\tau}{du}.$$

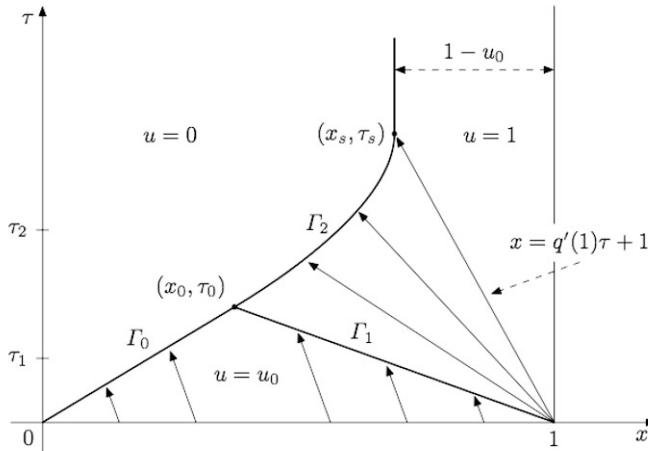


Fig. 4.31 Characteristics for the sedimentation problem (case (a))

Using (4.92), we have

$$\frac{dx}{du} = \frac{q(u)}{u} \frac{d\tau}{du},$$

which gives

$$\frac{q(u)}{u} \frac{d\tau}{du} = q''(u)\tau + q'(u) \frac{d\tau}{du}$$

and finally

$$\frac{1}{\tau} \frac{d\tau}{du} = \frac{uq''(u)}{q(u) - uq'(u)}$$

with the initial condition $\tau(u_1) = \tau_0$, since the characteristic $x = q'(u_1)\tau + 1$ carries the value u_1 up to the point (x_0, τ_0) . Integrating between τ_0 and τ , we find²⁰

$$\log\left(\frac{\tau}{\tau_0}\right) = \log\left(\frac{q(u_1) - u_1 q'(u_1)}{q(u) - u q'(u)}\right).$$

Recalling (4.91), we find, for $u_1 < u \leq 1$,

$$\tau = \tau_0 \frac{q(u_1) - u_1 q'(u_1)}{q(u) - u q'(u)} = \tau_0 \frac{(1 + \beta - 2\beta u_1)u_1^2}{(1 + \beta - 2\beta u)u^2} = \frac{u_0}{u^2(1 + \beta - 2\beta u)}$$

and

$$x = q'(u)\tau + 1 = \frac{u_0(1 - 2(1 + \beta)u + 3\beta u^2)}{u^2(1 + \beta - 2\beta u)} + 1.$$

These two equations give the parametric representation of Γ_2 .

²⁰ Note that $q'(u) < 0$ on Γ_2 .

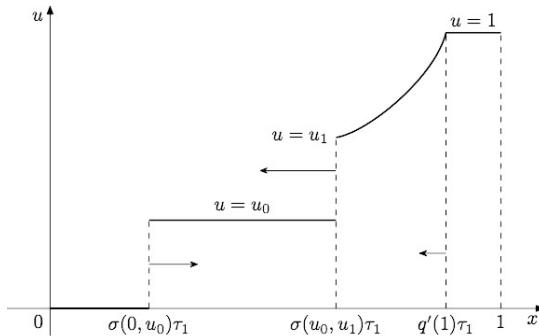


Fig. 4.32 Profile of the solution at time τ_1 , $0 < \tau_1 < \tau_s$

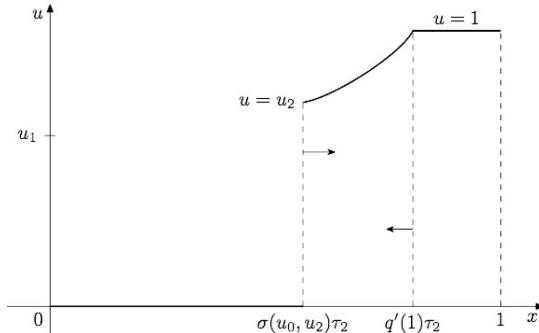


Fig. 4.33 Profile of the solution at time τ_2 , $\tau_0 < \tau_2 < \tau_s$

- *The final stationary profile.* The interaction between $u = 0$ and the rarefaction wave ceases when the speed $dx/d\tau$ of Γ_2 reaches zero. This occurs when the characteristic from the right carries the value $u = 1$, that is at the point S of coordinates

$$\tau_s = -\frac{u_0}{q'(1)} = \frac{u_0}{1-\beta}, \quad x_s = 1 - u_0.$$

The process of sedimentation is therefore completed at $\tau = \tau_s$. After τ_s , the shock become stationary since $\sigma(0, 1) = 0$. The final configuration is

$$u(x) = \begin{cases} 0 & 0 < x \leq 1 - u_0 \\ 1 & 1 - u_0 < x \leq 1. \end{cases}$$

Figures 4.32 and 4.33 show the profile of the solution at times $\tau = \tau_1$, before the shock merging, and $\tau = \tau_2$, $\tau_0 < \tau_2 < \tau_s$, respectively.

4.8 The Method of Characteristics for Quasilinear Equations

In this section we apply the method of characteristics to general quasilinear equations. We consider the case of two independent variables, where the intuition is supported by the geometric interpretation.

4.8.1 Characteristics

We consider equations of the form

$$a(x, y, u)u_x + b(x, y, u)u_y = c(x, y, u), \quad (4.94)$$

where

$$u = u(x, y)$$

and a, b, c are *continuously differentiable functions*.

The solutions of (4.94) can be constructed via geometric arguments. The tangent plane to the graph of a solution u at a point (x_0, y_0, z_0) has equation

$$u_x(x_0, y_0)(x - x_0) + u_y(x_0, y_0)(y - y_0) - (z - z_0) = 0$$

and the vector

$$\mathbf{n}_0 = (u_x(x_0, y_0), u_y(x_0, y_0), -1)$$

is *normal* to the plane. Introducing the vector

$$\mathbf{v}_0 = (a(x_0, y_0, z_0), b(x_0, y_0, z_0), c(x_0, y_0, z_0)),$$

equation (4.94) implies that

$$\mathbf{n}_0 \cdot \mathbf{v}_0 = 0.$$

Thus, \mathbf{v}_0 is tangent to the graph of u (Fig. 4.34). In other words, (4.94) says that, at every point (x, y, z) , the graph of any solution is tangent to the vector field

$$\mathbf{v}(x, y, z) = (a(x, y, z), b(x, y, z), c(x, y, z)).$$

In this case we say that the graph of a solution is an **integral surface** of the vector field \mathbf{v} .

Now, we can construct an integral surfaces of \mathbf{v} as union of **integral curves** of \mathbf{v} , that is curves tangent to \mathbf{v} at every point. These curves are solutions of the system

$$\frac{dx}{dt} = a(x, y, z), \quad \frac{dy}{dt} = b(x, y, z), \quad \frac{dz}{dt} = c(x, y, z) \quad (4.95)$$

and are called **characteristics**. Note that $z = z(t)$ gives the values u along the projected characteristic $(x(t), y(t))$, that is

$$z(t) = u(x(t), y(t)). \quad (4.96)$$

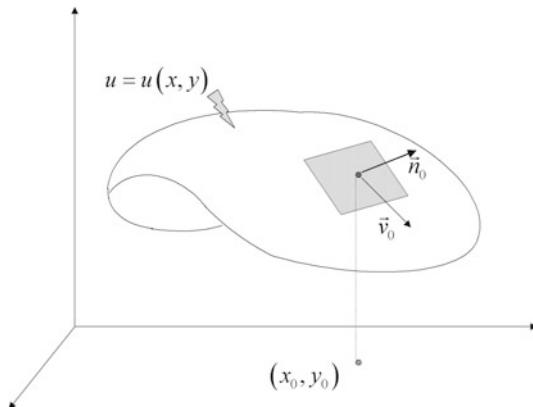


Fig. 4.34 Integral surface

In fact, differentiating (4.96) and using (4.95) and (4.94), we have

$$\begin{aligned} \frac{dz}{dt} &= u_x(x(t), y(t)) \frac{dx}{dt} + u_y(x(t), y(t)) \frac{dy}{dt} \\ &= a(x(t), y(t), z(t)) u_x(x(t), y(t)) + b((x(t), y(t), z(t)) u_y(x(t), y(t)) \\ &= c(x(t), y(t), z(t)). \end{aligned}$$

Thus, along a characteristic the partial differential equation (4.94) degenerates into an ordinary differential equation.

In the case of a conservation law (with $t = y$)

$$u_y + q'(u) u_x = 0 \quad \left(q'(u) = \frac{dq}{du} \right),$$

we have introduced the notion of characteristic in a slightly different way, but we shall see later that there is no contradiction.

The following proposition is a consequence of the above geometric reasoning and of the existence and uniqueness theorem for system of ordinary differential equations²¹.

Proposition 4.15. a) *Let the surface S be the graph of a C^1 function $u = u(x, y)$. If S is union of characteristics then u is a solution of the equation (4.94).*

b) *Every integral surface S of the vector field \mathbf{v} is union of characteristics. Namely: every point of S belongs exactly to one characteristic, entirely contained in S .*

c) *Two integral surfaces intersecting at one point intersect along the whole characteristic passing through that point.*

²¹ See [23], Courant–Hilbert, vol. 2, 1953.

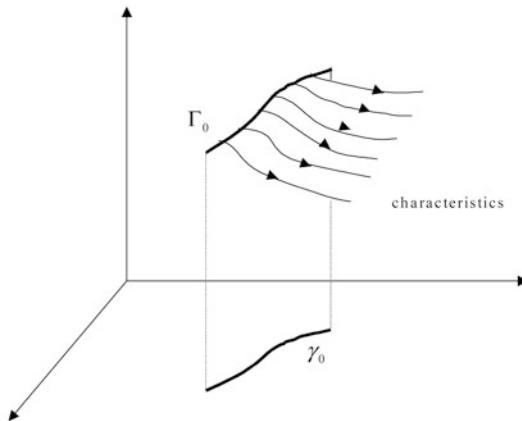


Fig. 4.35 Characteristics flowing out of Γ_0

4.8.2 The Cauchy problem

Proposition 4.15 gives a characterization of the integral surfaces as a union of characteristics. The problem is how to construct such unions to get a smooth surface. One way to proceed is to look for solutions u whose values are prescribed on a curve γ_0 , contained in the x, y plane.

In other words, suppose that

$$x(s) = f(s), \quad y(s) = g(s), \quad s \in I \subseteq \mathbb{R},$$

is a parametrization of γ_0 . We look for a solution u of (4.94) such that

$$u(f(s), g(s)) = h(s), \quad s \in I, \tag{4.97}$$

where $h = h(s)$ is a given function. We assume that I is a neighborhood of $s = 0$, and that f, g, h are continuously differentiable in I .

The system (4.94), (4.97) is called **Cauchy problem**. If we consider the three-dimensional curve Γ_0 given by the parametrization

$$x(s) = f(s), \quad y(s) = g(s), \quad z(s) = h(s),$$

then, solving the Cauchy problem (4.94), (4.97) amounts to determining an integral surface containing Γ_0 .

The data are often assigned in the form of *initial values*

$$u(x, 0) = h(x),$$

with y playing the role of “time”. In this case, γ_0 is the axis $y = 0$ and x plays the role of the parameter s . Then a parametrization of Γ_0 is given by

$$x = x, \quad y = 0, \quad z(x) = h(x).$$

By analogy, we often refer to Γ_0 as to the *initial curve*. The strategy to solve a Cauchy problem comes from its geometric meaning: since the graph of a solution $u = u(x, y)$ is a union of characteristics, we determine those flowing out from Γ_0 (Fig. 4.35) by solving the system

$$\frac{dx}{dt} = a(x, y, z), \quad \frac{dy}{dt} = b(x, y, z), \quad \frac{dz}{dt} = c(x, y, z), \quad (4.98)$$

with the family of initial conditions

$$x(0) = f(s), \quad y(0) = g(s), \quad z(0) = h(s), \quad (4.99)$$

parametrized by $s \in I$. The union of the characteristics of this family should give the graph of u . Why only *should*? We will come back later to this question.

Under our hypotheses, the Cauchy problem (4.98), (4.99) has exactly one solution

$$x = X(s, t), \quad y = Y(s, t), \quad z = Z(s, t) \quad (4.100)$$

in a neighborhood of $t = 0$, for every $s \in I$.

A couple of questions arise:

- a) Does the system of the three equations (4.100) define a function $z = u(x, y)$?
- b) Even if the answer to question a) is positive, is $z = u(x, y)$ the unique solution of the Cauchy problem (4.94), (4.97)?

Let us reason in a neighborhood of $s = t = 0$, setting

$$X(0, 0) = f(0) = x_0, \quad Y(0, 0) = g(0) = y_0, \quad Z(0, 0) = h(0) = z_0.$$

The answer to question a) is positive if we can solve for s and t the first two equations in (4.100), and find $s = S(x, y)$ and $t = T(x, y)$ of class C^1 in a neighborhood of (x_0, y_0) , such that

$$S(x_0, y_0) = 0, \quad T(x_0, y_0) = 0.$$

Then, from the third equation $z = Z(s, t)$, we get

$$z = Z(S(x, y), T(x, y)) = u(x, y). \quad (4.101)$$

From the *Inverse Function Theorem*, the system

$$\begin{cases} X(s, t) = x \\ Y(s, t) = y \end{cases}$$

defines

$$s = S(x, y) \quad \text{and} \quad t = T(x, y)$$

in a neighborhood of (x_0, y_0) if the Jacobian

$$J(0, 0) = \begin{vmatrix} X_s(0, 0) & Y_s(0, 0) \\ X_t(0, 0) & Y_t(0, 0) \end{vmatrix} \neq 0. \quad (4.102)$$

From (4.98) and (4.99), we have

$$X_s(0, 0) = f'(0), \quad Y_s(0, 0) = g'(0)$$

and

$$X_t(0, 0) = a(x_0, y_0, z_0), \quad Y_t(0, 0) = b(x_0, y_0, z_0),$$

so that (4.102) becomes

$$J(0, 0) = \begin{vmatrix} f'(0) & g'(0) \\ a(x_0, y_0, z_0) & b(x_0, y_0, z_0) \end{vmatrix} \neq 0 \quad (4.103)$$

or

$$b(x_0, y_0, z_0) f'(0) \neq a(x_0, y_0, z_0) g'(0). \quad (4.104)$$

Condition (4.104) means that *the vectors*

$$(a(x_0, y_0, z_0), b(x_0, y_0, z_0)) \quad \text{and} \quad (f'(0), g'(0))$$

are not parallel.

In conclusion: *if condition (4.103) holds, then (4.101) is a well defined C^1 -function.*

Now consider question **b).** The above construction of u implies that the surface $z = u(x, y)$ contains Γ_0 and all the characteristics flowing out from Γ_0 , so that u is a solution of the Cauchy problem (4.94), (4.97). Moreover, by Proposition 4.15 c), two integral surfaces containing Γ_0 must contain the same characteristics and therefore coincide.

We summarize everything in the following theorem, recalling that

$$(x_0, y_0, z_0) = (f(0), g(0), h(0)).$$

Theorem 4.16. *Let a, b, c be C^1 -functions in a neighborhood of (x_0, y_0, z_0) and f, g, h be C^1 -functions in I . If $J(0, 0) \neq 0$, then, in a neighborhood of (x_0, y_0) , there exists a unique C^1 -solution $u = u(x, y)$ of the Cauchy problem*

$$\begin{cases} a(x, y, u) u_x + b(x, y, u) u_y = c(x, y, u) \\ u(f(s), g(s)) = h(s). \end{cases} \quad (4.105)$$

Moreover, u is defined by the parametric equations (4.101).

Remark 4.17. If a, b, c and f, g, h are C^k -functions, $k \geq 2$, then u is a C^k -function as well.

It remains to examine what happens when $J(0, 0) = 0$, that is when the vectors $(a(x_0, y_0, z_0), b(x_0, y_0, z_0))$ and $(f'(0), g'(0))$ are parallel.

Suppose that there exists a C^1 -solution u of the Cauchy problem (4.105). Differentiating the second equation in (4.105) we get

$$h'(s) = u_x(f(s), g(s)) f'(s) + u_y(f(s), g(s)) g'(s). \quad (4.106)$$

Computing at $x = x_0$, $y = y_0$, $z = z_0$ and $s = 0$, we obtain

$$\begin{cases} a(x_0, y_0, z_0) u_x(x_0, y_0) + b(x_0, y_0, z_0) u_y(x_0, y_0) = c(x_0, y_0, z_0) \\ f'(0) u_x(x_0, y_0) + g'(0) u_y(x_0, y_0) = h'(0). \end{cases} \quad (4.107)$$

Since u is a solution of the Cauchy problem, the vector

$$(u_x(x_0, y_0), u_y(x_0, y_0))$$

is a solution of the algebraic system (4.107). But then, from Linear Algebra, we know that the condition

$$\text{rank} \begin{pmatrix} a(x_0, y_0, z_0) & b(x_0, y_0, z_0) & c(x_0, y_0, z_0) \\ f'(0) & g'(0) & h'(0) \end{pmatrix} = 1 \quad (4.108)$$

must hold and therefore the two vectors

$$(a(x_0, y_0, z_0), b(x_0, y_0, z_0), c(x_0, y_0, z_0)) \quad \text{and} \quad (f'(0), g'(0), h'(0)) \quad (4.109)$$

must be parallel. This is equivalent to saying that Γ_0 is parallel to the characteristic curve at (x_0, y_0, z_0) . When this occurs, we say that Γ_0 is **characteristic at the point (x_0, y_0, z_0)** .

Conclusion: *If $J(0, 0) = 0$, a necessary condition for the existence of a C^1 -solution $u = u(x, y)$ of the Cauchy problem in a neighborhood of (x_0, y_0) is that Γ_0 be characteristic at (x_0, y_0, z_0) .*

Now, assume Γ_0 itself is a characteristic and let $P_0 = (x_0, y_0, z_0) \in \Gamma_0$. If we choose a curve Γ^* transversal²² to Γ_0 at P_0 , by Theorem 4.16 there exists a unique integral surface containing Γ^* and, by Proposition 4.15c), p. 231, this surface contains Γ_0 . In this way we can construct infinitely many smooth solutions.

We point out that the condition (4.108) is compatible with the existence of a C^1 -solution only if Γ_0 is characteristic at P_0 . On the other hand, it may occur that $J(0, 0) = 0$, that Γ_0 is non characteristic at P_0 and that solutions of the Cauchy problem exist anyway; clearly, these solutions **cannot** be of class C^1 (see Example 4.18 below).

Let us summarize the steps to solve the Cauchy problem (4.105):

Step 1. Determine the solution

$$x = X(s, t), \quad y = Y(s, t), \quad z = Z(s, t) \quad (4.110)$$

of the characteristic system

$$\frac{dx}{dt} = a(x, y, z), \quad \frac{dy}{dt} = b(x, y, z), \quad \frac{dz}{dt} = c(x, y, z), \quad (4.111)$$

²² That is, with tangent vectors not colinear.

with initial conditions

$$x(s, 0) = f(s), \quad y(s, 0) = g(s), \quad z(s, 0) = h(s), \quad s \in I.$$

Step 2. Compute $J(s, t)$ on the initial curve Γ_0 , that is:

$$J(s, 0) = \begin{vmatrix} f'(s) & g'(s) \\ X_t(0, s) & Y_t(0, s) \end{vmatrix}.$$

The following cases may occur:

2a. $J(s, 0) \neq 0$, for every $s \in I$. This means that Γ_0 does not have characteristic points. Then, in a neighborhood of Γ_0 , there exists a unique solution $u = u(x, y)$ of the Cauchy problem, defined by the parametric equations (4.110).

2b. $J(s_0, 0) = 0$ for some $s_0 \in I$, and Γ_0 is characteristic at the point $P_0 = (f(s_0), g(s_0), h(s_0))$. A C^1 -solution may exist in a neighborhood of P_0 only if the rank condition (4.108) holds at P_0 .

2c. $J(s_0, 0) = 0$ for some $s_0 \in I$ and Γ_0 is **not** characteristic at P_0 . There are no C^1 -solutions in a neighborhood of P_0 . There may exist less regular solutions.

2d. Γ_0 is a characteristic. Then there exist infinitely many C^1 -solutions in a neighborhood of Γ_0 .

Example 4.18. Consider the nonhomogeneous Burgers equation

$$uu_x + u_y = 1. \quad (4.112)$$

As in Example 4.8, if y is the time variable y , $u = u(x, y)$ represents a *velocity field* of a flux of particles along the x -axis. Equation (4.112) states that the acceleration of each particle is equal to 1. Assume that

$$u(x, 0) = h(x), \quad x \in \mathbb{R}.$$

The characteristics are solutions of the system

$$\frac{dx}{dt} = z, \quad \frac{dy}{dt} = 1, \quad \frac{dz}{dt} = 1$$

and the initial curve Γ_0 has the parametrization

$$x = f(s) = s, \quad y = g(s) = 0, \quad z = h(s), \quad s \in \mathbb{R}.$$

The characteristics flowing out from Γ_0 are

$$X(s, t) = s + \frac{t^2}{2} + th(s), \quad Y(s, t) = t, \quad Z(s, t) = t + h(s).$$

Since

$$J(s, t) = \begin{vmatrix} 1 + th'(s) & 0 \\ t + h(s) & 1 \end{vmatrix} = 1 + th'(s),$$

we have $J(s, 0) = 1$ and we are in the case **2a**: in a neighborhood of Γ_0 there exists a unique C^1 -solution. If, for instance, $h(s) = s$, we find the solution

$$u = y + \frac{2x - y^2}{2 + 2y}, \quad (x \in \mathbb{R}, y \geq -1).$$

Now consider the same equation with initial condition

$$u\left(\frac{y^2}{4}, y\right) = \frac{y}{2},$$

equivalent to assigning the values of u on the parabola $x = \frac{y^2}{4}$. A parametrization of Γ_0 is given by

$$x = s^2, \quad y = 2s, \quad z = s, \quad s \in \mathbb{R}.$$

Solving the characteristic system with these initial conditions, we find

$$X(s, t) = s^2 + ts + \frac{t^2}{2}, \quad Y(s, t) = 2s + t, \quad Z(s, t) = s + t. \quad (4.113)$$

Observe that Γ_0 does not have any characteristic point, since its tangent vector $(2s, 2, 1)$ is never parallel to the characteristic direction $(s, 1, 1)$. However

$$J(s, t) = \begin{vmatrix} 2s + t & 2 \\ s + t & 1 \end{vmatrix} = -t$$

which vanishes for $t = 0$, i.e. exactly on Γ_0 . We are in the case **2c**. Solving for s and t , $t \neq 0$, in the first two equations (4.113), and substituting into the third one, we find

$$u(x, y) = \frac{y}{2} \pm \sqrt{x - \frac{y^2}{4}}.$$

We have found two solutions of the Cauchy problem, satisfying the differential equation in the region $x > \frac{y^2}{4}$. However, these solutions are not smooth in a neighborhood of Γ_0 , since on Γ_0 they are not differentiable.

- *Conservation laws.* According to the new definition, the characteristics of the equation

$$u_y + q'(u)u_x = 0, \quad \left(q'(u) = \frac{dq}{du} \right),$$

with initial conditions

$$u(x, 0) = g(x),$$

are the three-dimensional solution curves of the system

$$\frac{dx}{dt} = q'(z), \quad \frac{dy}{dt} = 1, \quad \frac{dz}{dt} = 0$$

with initial conditions

$$x(s, 0) = s, \quad y(s, 0) = 0, \quad z(s, 0) = g(s), \quad s \in \mathbb{R}.$$

Integrating, we find

$$z = g(s), \quad x = q'(g(s))t + s, \quad y = t.$$

The *projections* of these straight-lines on the x, y plane are

$$x = q'(g(s))y + s,$$

which coincide with the “old characteristics”, called *projected characteristics* in the general quasilinear context.

- *Linear equations.* Consider a linear equation

$$a(x, y)u_x + b(x, y)u_y = 0. \quad (4.114)$$

Introducing the vector $\mathbf{w}(x, y) = (a(x, y), b(x, y))$, we may write equation (4.114) in the form

$$D_{\mathbf{w}}u = \nabla u \cdot \mathbf{w} = 0.$$

Thus, every solution of the (4.114) is constant along the integral lines of the vector field \mathbf{w} , i.e. along the *projected characteristics*, solutions of the reduced characteristic system

$$\frac{dx}{dt} = a(x, y), \quad \frac{dy}{dt} = b(x, y), \quad (4.115)$$

locally equivalent to the ordinary differential equation

$$b(x, y)dx - a(x, y)dy = 0.$$

If a *first integral*²³ $\psi = \psi(x, y)$ of the system (4.115) is known, then the family of the projected characteristics is given in implicit form by

$$\psi(x, y) = k, \quad k \in \mathbb{R}$$

and the general solution of (4.114) is given by the formula

$$u(x, y) = G(\psi(x, y)),$$

²³ We recall that a *first integral* (also called *constant of motion*) for a system of ODE $\frac{d\mathbf{x}}{dt} = \mathbf{f}(\mathbf{x})$, is a C^1 -function $\varphi = \varphi(\mathbf{x})$, which is constant along the trajectories of the system, i.e. such that $\nabla \varphi \cdot \mathbf{f} \equiv 0$.

where $G = G(r)$ is an arbitrary C^1 -function, that can be determined by the Cauchy data.

Example 4.19. Let us solve the problem

$$\begin{cases} yu_x + xu_y = 0 \\ u(x, 0) = x^4. \end{cases}$$

Here $\mathbf{w} = (y, x)$ and the reduced characteristic system is

$$\frac{dx}{dt} = y, \quad \frac{dy}{dt} = x,$$

locally equivalent to

$$xdx - ydy = 0.$$

Integrating, we find that the projected characteristics are the hyperbolas

$$\psi(x, y) = x^2 - y^2 = k$$

and therefore $\psi(x, y) = x^2 - y^2$ is a first integral. Then, the general solution of the equation is

$$u(x, y) = G(x^2 - y^2).$$

Imposing the Cauchy condition, we have

$$G(x^2) = x^4$$

from which $G(r) = r^2$. The solution of the Cauchy problem is

$$u(x, y) = (x^2 - y^2)^2.$$

4.8.3 Lagrange method of first integrals

We have seen that, in the linear case, we can construct a *general solution*, depending on an arbitrary function, from the knowledge of a first integral for the reduced characteristic system. The method can be extended to equations of the form

$$a(x, y, u)u_x + b(x, y, u)u_y = c(x, y, u). \quad (4.116)$$

We say that two *first integrals* $\varphi = \varphi(x, y, u)$ are *independent* if $\nabla\varphi$ and $\nabla\psi$ are nowhere colinear. Then:

Theorem 4.20. *Let $\varphi = \varphi(x, y, u)$, $\psi = \psi(x, y, u)$ be two independent first integrals of the characteristic system and $F = F(h, k)$ be a C^1 -function. If*

$$F_h\varphi_u + F_k\psi_u \neq 0,$$

the equation

$$F(\varphi(x, y, u), \psi(x, y, u)) = 0$$

defines the general solution of (4.116) in implicit form.

Proof. It is based on the following two observations. First, the function

$$w = F(\varphi(x, y, u), \psi(x, y, u)) \quad (4.117)$$

is a first integral. In fact,

$$\nabla w = F_h \nabla \varphi + F_k \nabla \psi$$

so that

$$\nabla w \cdot (a, b, c) = F_h \nabla \varphi \cdot (a, b, c) + F_k \nabla \psi \cdot (a, b, c) \equiv 0$$

since φ and ψ are *first integrals*. Moreover, by hypothesis,

$$w_u = F_h \varphi_u + F_k \psi_u \neq 0.$$

Second, if w is a first integral and $w_u \neq 0$, then, the equation

$$w(x, y, u) = 0 \quad (4.118)$$

defines implicitly an integral surface $u = u(x, y)$ of (4.116). In fact, since w is a first integral, it satisfies the equation

$$a(x, y, u) w_x + b(x, y, u) w_y + c(x, y, u) w_u = 0. \quad (4.119)$$

Moreover, from the Implicit Function Theorem, we have

$$u_x = -\frac{w_x}{w_u}, \quad u_y = -\frac{w_y}{w_u},$$

and from (4.119) we get easily (4.116). □

Remark 4.21. As a by-product of the proof, we have that the general solution of the three-dimensional homogeneous equation (4.119) is given by (4.117).

Remark 4.22. The search for first integrals is sometimes simplified by writing the characteristic system in the form

$$\frac{dx}{a(x, y, u)} = \frac{dy}{b(x, y, u)} = \frac{du}{c(x, y, u)}.$$

Example 4.23. Consider again the nonhomogeneous Burgers equation

$$uu_x + u_y = 1$$

with initial condition

$$u\left(\frac{1}{2}y^2, y\right) = y.$$

A parametrization of the initial curve Γ_0 is

$$x = \frac{1}{2}s^2, \quad y = s, \quad z = s$$

and therefore Γ_0 is a characteristic. We are in the case **2d**. Let us use the Lagrange method. To find two independent first integrals, we write the characteristic system in the form

$$\frac{dx}{z} = dy = dz$$

or

$$dx = zdz, \quad dy = dz.$$

Integrating these two equations, we get

$$x - \frac{1}{2}z^2 = c_2, \quad y - z = c_1$$

and therefore

$$\varphi(x, y, z) = x - \frac{1}{2}z^2, \quad \psi(x, y, z) = y - z$$

are two first integral. Since

$$\nabla\varphi(x, y, z) = (1, 0, -z)$$

and

$$\nabla\psi(x, y, z) = (0, 1, -1)$$

we see that they are also independent. Thus, the general solution of Burgers equation is given by

$$F\left(x - \frac{1}{2}z^2, y - z\right) = 0$$

where F is an arbitrary C^1 -function.

Finally, to satisfy the initial condition, it is enough to choose F such that $F(0, 0) = 0$. As expected, there exist infinitely many solutions of the Cauchy problem.

4.8.4 Underground flow

We apply the methodology presented in the previous sections to a model for the underground flow of a fluid (like water). In the wet region, only a fraction of any control volume is filled with fluid. This fraction, denoted by ϕ , is called *porosity* and, in general, it depends on position, temperature and pressure. Here, we assume that ϕ depends on position only: $\phi = \phi(x, y, z)$.

If ρ is the fluid density and $\mathbf{q} = (q_1, q_2, q_3)$ is the flux vector (the volumetric flow rate of the fluid), the conservation of mass yields, in this case,

$$\phi\rho_t + \operatorname{div}(\rho\mathbf{q}) = 0.$$

For \mathbf{q} the following modified *Darcy's law* is commonly adopted:

$$\mathbf{q} = -\frac{k}{\mu} (\nabla p + \rho \mathbf{g})$$

where p is the pressure and \mathbf{g} is the gravity acceleration; $k > 0$ is the *permeability* of the medium and μ is the fluid *viscosity*. Thus, we have:

$$\phi \rho_t - \operatorname{div} \left[\frac{\rho k}{\mu} (\nabla p + \rho \mathbf{g}) \right] = 0. \quad (4.120)$$

Now, suppose that two *immiscible* fluids, of density ρ_1 and ρ_2 , flow underground. Immiscible means that the two fluids cannot dissolve one into the other or chemically interact. In particular, the conservation law holds for the mass of each fluid. Thus, if we denote by S_1 and S_2 the fractions (*saturations*) of the available space filled by the two fluids, respectively, we can write

$$\phi (S_1 \rho_1)_t + \operatorname{div}(\rho_1 \mathbf{q}_1) = 0 \quad (4.121)$$

$$\phi (S_2 \rho_2)_t + \operatorname{div}(\rho_2 \mathbf{q}_2) = 0. \quad (4.122)$$

We assume that $S_1 + S_2 = 1$, i.e. that the medium is completely saturated and that capillarity effects are negligible. We set $S_1 = S$ and $S_2 = 1 - S$. The Darcy law for the two fluids becomes

$$\mathbf{q}_1 = -k \frac{k_1}{\mu_1} (\nabla p + \rho_1 \mathbf{g})$$

$$\mathbf{q}_2 = -k \frac{k_2}{\mu_2} (\nabla p + \rho_2 \mathbf{g})$$

where k_1, k_2 are the *relative permeability coefficients*, in general depending on S .

We make now some other simplifying assumptions:

- a) Gravitational effects are negligible.
- b) $k, \phi, \rho_1, \rho_2, \mu_1, \mu_2$ are constant.
- c) $k_1 = k_1(S)$ and $k_2 = k_2(S)$ are known functions.

Equations (4.121) and (4.122) become:

$$\phi S_t + \operatorname{div} \mathbf{q}_1 = 0, \quad -\phi S_t + \operatorname{div} \mathbf{q}_2 = 0, \quad (4.123)$$

while the Darcy laws take the form

$$\mathbf{q}_1 = -k \frac{k_1}{\mu_1} \nabla p, \quad \mathbf{q}_2 = -k \frac{k_2}{\mu_2} \nabla p. \quad (4.124)$$

Letting $\mathbf{q} = \mathbf{q}_1 + \mathbf{q}_2$ and adding the two equations in (4.123) we have

$$\operatorname{div} \mathbf{q} = 0.$$

Adding the two equations in (4.124) yields, setting $K(S) = \left(\frac{k_1(S)}{\mu_1} + \frac{k_2(S)}{\mu_2} \right)^{-1}$,

$$\nabla p = -\frac{1}{k} K(S) \mathbf{q}$$

from which

$$\operatorname{div} \nabla p = \Delta p = -\frac{1}{k} \mathbf{q} \cdot \nabla K(S).$$

From the first equations in (4.123) and (4.124) we get

$$\begin{aligned} \phi S_t &= -\operatorname{div} \mathbf{q}_1 = \frac{k}{\mu_1} [\nabla k_1(S) \cdot \nabla p + k_1(S) \Delta p] \\ &= -\frac{k}{\mu_1} - [K(S) \mathbf{q} \cdot \nabla k_1(S) - k_1(S) \mathbf{q} \cdot \nabla K(S)] \\ &= \mathbf{q} \cdot \nabla H(S) = H'(S) \mathbf{q} \cdot \nabla S \end{aligned}$$

where

$$H(S) = -\frac{1}{\mu_1} k_1(S) K(S).$$

When \mathbf{q} is known, the resulting *quasilinear* equation for the saturation S is

$$\phi S_t = H'(S) \mathbf{q} \cdot \nabla S,$$

known as the *Bukley-Leverett* equation.

In particular, if \mathbf{q} can be considered one-dimensional and constant, i.e. $\mathbf{q} = q\mathbf{i}$, we have

$$qH'(S) S_x + \phi S_t = 0$$

which of the form (4.116), with $u = S$ and $y = t$. The characteristic system is (see Remark 4.22, p. 240)

$$\frac{dx}{qH'(S) S} = \frac{dt}{\phi} = \frac{dS}{0}.$$

Two first integrals are

$$w_1 = \phi x - qH'(S) t \quad \text{and} \quad w_2 = S.$$

Thus, the general solution is given by

$$F(\phi x - qH'(S) St, S) = 0.$$

The choice

$$F(w_1, w_2) = w_2 - f(w_1),$$

yields $S = f(\phi x - qH'(S) t)$ that satisfies the initial condition $S(x, 0) = f(\phi x)$.

4.9 General First Order Equations

4.9.1 Characteristic strips

We extend the characteristic method to *nonlinear* equations of the form

$$F(x, y, u, u_x, u_y) = 0. \quad (4.125)$$

We assume that $F = F(x, y, u, p, q)$ is a smooth function of its arguments and, to avoid trivial cases, that $F_p^2 + F_q^2 \neq 0$. In the quasilinear case,

$$F(x, y, u, p, q) = a(x, y, u)p + b(x, y, u)q - c(x, y, u)$$

and

$$F_p = a(x, y, u), \quad F_q = b(x, y, u), \quad (4.126)$$

so that $F_p^2 + F_q^2 \neq 0$ says that a and b do not vanish simultaneously.

Equation (4.125) has a geometrical interpretation as well. Let $u = u(x, y)$ be a smooth solution and consider a point (x_0, y_0, z_0) on its graph. Equation (4.125) constitutes a link between the components u_x and u_y of the normal vector

$$\mathbf{n}_0 = (-u_x(x_0, y_0), -u_y(x_0, y_0), 1)$$

but it is a little more complicated than in the quasilinear case²⁴ and is not a priori clear what a characteristic system for equation (4.125) should be. Reasoning by analogy with the quasilinear case, from (4.126) we are led to the equations

$$\begin{aligned} \frac{dx}{dt} &= F_p(x, y, z, p, q) \\ \frac{dy}{dt} &= F_q(x, y, z, p, q), \end{aligned} \quad (4.127)$$

where $z(t) = u(x(t), y(t))$ and

$$p = p(t) = u_x(x(t), y(t)), \quad q = q(t) = u_y(x(t), y(t)). \quad (4.128)$$

Thus, taking account of (4.127), the equation for z is:

$$\frac{dz}{dt} = u_x \frac{dx}{dt} + u_y \frac{dy}{dt} = pF_p + qF_q. \quad (4.129)$$

²⁴ If, for instance $F_q \neq 0$, by the Implicit Function Theorem, the equation $F(x_0, y_0, z_0, p, q) = 0$ defines $q = q(p)$ so that

$$F(x_0, y_0, z_0, p, q(p)) = 0.$$

Therefore, the possible tangent plane to u at (x_0, y_0, z_0) form a one parameter family of planes, given by

$$p(x - x_0) + q(p)(y - y_0) - (z - z_0) = 0.$$

This family, in general, envelopes a cone with vertex at (x_0, y_0, z_0) , called *Monge cone*. Each possible tangent plane touches the Monge cone along a generatrix.

Equations (4.129) and (4.127) correspond to the characteristic system (4.95), but with two more unknowns: $p(t)$ and $q(t)$. We need two more equations. Proceeding formally, from (4.127) we can write

$$\begin{aligned}\frac{dp}{dt} &= u_{xx}(x(t), y(t)) \frac{dx}{dt} + u_{xy}(x(t), y(t)) \frac{dy}{dt} \\ &= u_{xx}(x(t), y(t)) F_p + u_{xy}(x(t), y(t)) F_q.\end{aligned}$$

We have to get rid of the second order derivatives. Since u is a solution of (4.125), the identity

$$F(x, y, u(x, y), u_x(x, y), u_y(x, y)) \equiv 0$$

holds. Partial differentiation with respect to x yields, since $u_{xy} = u_{yx}$:

$$F_x + F_u u_x + F_p u_{xx} + F_q u_{xy} \equiv 0.$$

Computing along $x = x(t)$, $y = y(t)$, we get

$$u_{xx}(x(t), y(t)) F_p + u_{xy}(x(t), y(t)) F_q = -F_x - p(t) F_u. \quad (4.130)$$

Thus, we deduce for p the following differential equation:

$$\frac{dp}{dt} = -F_x(x, y, z, p, q) - p F_u(x, y, z, p, q).$$

Similarly, we find

$$\frac{dq}{dt} = -F_y(x, y, z, p, q) - q F_u(x, y, z, p, q).$$

In conclusion, we are led to the following **characteristic system** of five autonomous equations:

$$\frac{dx}{dt} = F_p, \quad \frac{dy}{dt} = F_q, \quad \frac{dz}{dt} = p F_p + q F_q \quad (4.131)$$

and

$$\frac{dp}{dt} = -F_x - p F_u, \quad \frac{dq}{dt} = -F_y - q F_u. \quad (4.132)$$

Observe that $F = F(x, y, u, p, q)$ is a **first integral of** (4.131), (4.132). In fact

$$\begin{aligned}&\frac{d}{dt} F(x(t), y(t), z(t), p(t), q(t)) \\ &= F_x \frac{dx}{dt} + F_y \frac{dy}{dt} + F_u \frac{dz}{dt} + F_p \frac{dp}{dt} + F_q \frac{dq}{dt} \\ &= F_x F_p + F_y F_q + F_u (p F_p + q F_q) + F_p (-F_x - p F_p) + F_q (-F_y - q F_q) \\ &\equiv 0\end{aligned}$$

and therefore, if $F(x(t_0), y(t_0), z(t_0), p(t_0), q(t_0)) = 0$ at some t_0 , then

$$F(x(t), y(t), z(t), p(t), q(t)) \equiv 0. \quad (4.133)$$

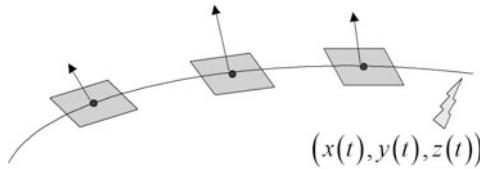


Fig. 4.36 Characteristic strip

Thus, the curve,

$$x = x(t), \quad y = y(t), \quad z = z(t),$$

still called a *characteristic curve*, lives on an integral surface, while

$$p = p(t), \quad q = q(t)$$

give the normal vector at each point, and it can be associated with a piece of the tangent plane, as shown in Fig. 4.36.

For this reason, a solution $(x(t), y(t), z(t), p(t), q(t))$ of (4.131), (4.132) is called *characteristic strip*.

4.9.2 The Cauchy Problem

As usual, the *Cauchy problem* consists in looking for a solution u of (4.125), assuming prescribed values on a given curve γ_0 in the x, y plane. If γ_0 has the parametrization

$$x = f(s), \quad y = g(s), \quad s \in I \subseteq \mathbb{R}$$

we want that

$$u(f(s), g(s)) = h(s), \quad s \in I,$$

where $h = h(s)$ is a given function. We assume that $0 \in I$ and that f, g, h are smooth functions in I .

Let Γ_0 be the *initial curve*, given by the parametrization

$$x = f(s), \quad y = g(s), \quad z = h(s). \quad (4.134)$$

Equations (4.134) only specify the “initial” points for x, y and z . To solve the characteristic system, we have first to complete Γ_0 into a *strip*

$$(f(s), g(s), h(s), \varphi(s), \psi(s))$$

where

$$\varphi(s) = u_x(f(s), g(s)) \quad \text{and} \quad \psi(s) = u_y(f(s), g(s)).$$

The two functions $\varphi(s)$ and $\psi(s)$ represent the initial values for p and q and cannot be chosen arbitrarily. In fact, a first condition that $\varphi(s)$ and $\psi(s)$ have to satisfy is (recall (4.133)):

$$F(f(s), g(s), h(s), \varphi(s), \psi(s)) \equiv 0. \quad (4.135)$$

A second condition comes from differentiating $h(s) = u(f(s), g(s))$. The result is the so called *strip condition*

$$h'(s) = \varphi(s)f'(s) + \psi(s)g'(s). \quad (4.136)$$

Now we are in position to give a (formal) procedure to construct a solution of our Cauchy problem: *Determine a solution $u = u(x, y)$ of*

$$F(x, y, u, u_x, u_y) = 0,$$

such that $u(f(s), g(s)) = h(s)$:

1. Solve for $\varphi(s)$ and $\psi(s)$ the (nonlinear) system

$$\begin{cases} F(f(s), g(s), h(s), \varphi(s), \psi(s)) = 0 \\ \varphi(s)f'(s) + \psi(s)g'(s) = h'(s). \end{cases} \quad (4.137)$$

2. Solve the characteristic system (4.131), (4.132) with initial conditions

$$x(0) = f(s), y(0) = g(s), z(0) = h(s), p(0) = \varphi(s), q(0) = \psi(s).$$

Suppose we find the solution

$$x = X(t, s), y = Y(t, s), z = Z(t, s), p = P(t, s), q = Q(t, s).$$

3. Solve $x = X(t, s), y = Y(t, s)$ for s, t in terms of x, y . Substitute $s = S(t, x)$ and $t = T(t, x)$ into $z = Z(t, s)$ to yield a solution $z = u(x, y)$.

Example 4.24. We want to solve the equation

$$u = u_x^2 - 3u_y^2$$

with initial condition $u(x, 0) = x^2$. We have $F(p, q) = p^2 - 3q^2 - u$ and the characteristic system is

$$\frac{dx}{dt} = 2p, \quad \frac{dy}{dt} = -6q, \quad \frac{dz}{dt} = 2p^2 - 6q^2 = 2z \quad (4.138)$$

$$\frac{dp}{dt} = p, \quad \frac{dq}{dt} = q. \quad (4.139)$$

A parametrization of the initial line Γ_0 is

$$f(s) = s, \quad g(s) = 0, \quad h(s) = s^2.$$

To complete the initial strip we solve the system

$$\begin{cases} \varphi^2 - 3\psi^2 = s^2 \\ \varphi = 2s. \end{cases}$$

There are two solutions:

$$\varphi(s) = 2s, \quad \psi(s) = \pm s.$$

The choice of $\psi(s) = s$ yields, integrating the equations (4.139),

$$P(s, t) = 2se^t, \quad Q(s, t) = se^t$$

whence, from (4.138),

$$X(t, s) = 4s(e^t - 1) + s, \quad Y(t, s) = -6s(e^t - 1), \quad Z(t, s) = s^2 e^{2t}.$$

Solving the first two equations for s, t and substituting into the third one, we get

$$u(x, y) = \left(x + \frac{y}{2} \right)^2.$$

The choice of $\psi(s) = -s$ yields

$$u(x, y) = \left(x - \frac{y}{2} \right)^2.$$

As the example shows, in general *there is no uniqueness*, unless system (4.137) has a unique solution. On the other hand, if this system has no (real) solution, then equation (4.142) has no solution as well.

Furthermore, observe that if $(x_0, y_0, z_0) = (f(0), g(0), h(0))$ and (p_0, q_0) is a solution of the system

$$\begin{cases} F(x_0, y_0, z_0, p_0, q_0) = 0 \\ p_0 f'(0) + q_0 g'(0) = h'(0), \end{cases} \quad (4.140)$$

by the Implicit Function Theorem, the condition

$$\begin{vmatrix} f'(0) & F_p(x_0, y_0, z_0, p_0, q_0) \\ g'(0) & F_q(x_0, y_0, z_0, p_0, q_0) \end{vmatrix} \neq 0 \quad (4.141)$$

ensures the existence of a solution $\varphi(s)$ and $\psi(s)$ of (4.137), in a neighborhood of $s = 0$. Condition (4.141) corresponds to (4.103) in the quasilinear case.

The following theorem summarizes the above discussion on the Cauchy problem

$$F(x, y, u, u_x, u_y) = 0 \quad (4.142)$$

with initial curve I_0 given by

$$x = f(s), \quad y = g(s), \quad z = h(s). \quad (4.143)$$

Theorem 4.25. *Assume that:*

- i) *F is twice continuously differentiable in a domain $D \subseteq \mathbb{R}^5$ and $F_p^2 + F_q^2 \neq 0$.*
- ii) *f, g, h are twice continuously differentiable in a neighborhood of $s = 0$.*
- iii) *(p_0, q_0) is a solution of the system (4.140) where $(x_0, y_0, z_0) = (f(0), g(0), h(0))$, and condition (4.141) holds.*

Then, in a neighborhood of (x_0, y_0) , there exists a C^2 solution $z = u(x, y)$ of the Cauchy problem (4.142), (4.143).

- *Geometrical optics.* The equation

$$c^2(u_x^2 + u_y^2) = 1 \quad (c > 0) \quad (4.144)$$

is called *eikonal equation* and arises in (two dimensional) geometrical optics. Indeed, assume that the level lines γ_t of equation

$$u(x, y) = t \quad (4.145)$$

represent the “wave fronts” of a wave perturbation (i.e. light) moving with time t and propagation speed c , that we assume to be constant. An orthogonal trajectory to the wave fronts coincides with a *light ray*. Therefore, a point $(x(t), y(t))$ on a ray satisfies the identity

$$u(x(t), y(t)) = t \quad (4.146)$$

and its velocity vector $\mathbf{v} = (\dot{x}, \dot{y})$ is parallel to ∇u . Therefore

$$\nabla u \cdot \mathbf{v} = |\nabla u| |\mathbf{v}| = c |\nabla u|.$$

On the other hand, differentiating (4.146) we get

$$\nabla u \cdot \mathbf{v} = u_x \dot{x} + u_y \dot{y} = 1,$$

from which the eikonal equation $c^2 |\nabla u|^2 = 1$.

Geometrically, if we fix a point (x_0, y_0, z_0) , equation $c^2(p^2 + q^2) = 1$ states that the family of planes

$$z - z_0 = p(x - x_0) + q(y - y_0),$$

tangent to an integral surface at (x_0, y_0, z_0) , all make a fixed angle $\theta = \arctan |\nabla u|^{-1} = \arctan c$ with the z -axis. This family envelopes the circular cone

$$(x - x_0)^2 + (y - y_0)^2 = c^2(z - z_0)^2$$

called the *light cone*, with opening angle 2θ .

The *eikonal equation* is of the form (4.125), with²⁵

$$F(x, y, u, p, q) = \frac{1}{2} [c^2(p^2 + q^2) - 1].$$

The characteristic system is²⁶:

$$\frac{dx}{d\tau} = c^2 p, \quad \frac{dy}{d\tau} = c^2 q, \quad \frac{dz}{d\tau} = c^2 p^2 + c^2 q^2 = 1 \quad (4.147)$$

²⁵ The factor $\frac{1}{2}$ is there for esthetic reasons.

²⁶ Using τ as a parameter along the characteristics.

and

$$\frac{dp}{d\tau} = 0, \quad \frac{dq}{d\tau} = 0. \quad (\text{c3})$$

An initial curve Γ_0

$$x = f(s), \quad y = g(s), \quad z = h(s),$$

can be completed into an initial strip, by solving for ϕ and ψ the system

$$\begin{cases} \phi(s)^2 + \psi(s)^2 = c^{-2} \\ \phi(s)f'(s) + \psi(s)g'(s) = h'(s). \end{cases} \quad (4.148)$$

This system has *two real and distinct solutions* if

$$f'(s)^2 + g'(s)^2 > c^2 h'(s)^2 \quad (4.149)$$

while it has *no real solutions* if²⁷

$$f'(s)^2 + g'(s)^2 < c^2 h'(s)^2. \quad (4.150)$$

If (4.149) holds, Γ_0 forms an angle greater than θ with the z -axis and therefore it is exterior to the light cone (Fig. 4.37). In this case we say that Γ_0 is *space-like* and we can find two different solutions of the Cauchy problem. If (4.150) holds, Γ_0 is contained in the light cone and we say it is *time-like*. The Cauchy problem does not have any solution.

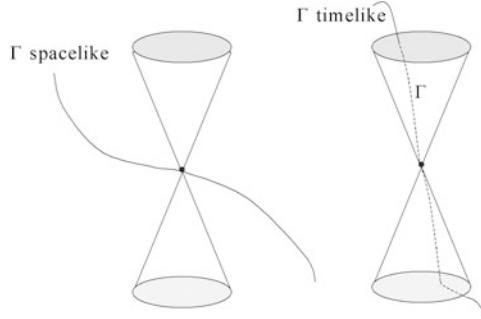


Fig. 4.37 Space-like and time-like initial curves

²⁷ System (4.148) is equivalent to finding the intersection between the circle $\xi^2 + \eta^2 = c^{-2}$ and the straight line $f'\xi + g'\eta = h'$. The distance of the center $(0, 0)$ from the line is given by

$$d = \frac{|h'|}{\sqrt{(f')^2 + (g')^2}}$$

so that there are 2 real intersections if $d < c^{-1}$, while there is no real intersection if $d > c^{-1}$.

Given a space-like curve Γ_0 and ϕ, ψ , solutions of the system (4.148), the corresponding characteristic strip is, for s fixed,

$$\begin{aligned} x(t) &= f(s) + c^2\phi(s)t, & y(t) &= g(s) + c^2\psi(s)t, & z(t) &= h(s) + t \\ p(t) &= \phi(s), & q(t) &= \psi(s). \end{aligned}$$

Observe that the point $(x(t), y(t))$ moves along the characteristic with speed

$$\sqrt{\dot{x}^2(t) + \dot{y}^2(t)} = \sqrt{\phi^2(s) + \psi^2(s)} = c$$

with direction $(\phi(s), \psi(s)) = (p(t), q(t))$. Therefore, the characteristic lines are coincident with the light rays. Moreover, we see that the fronts γ_t can be constructed from γ_0 by shifting any point on γ_0 along a ray at a distance ct . Thus, the wave fronts constitute a family of “parallel” curves.

Problems

4.1. Using Duhamel’s method (see Subsect. 2.8.3), solve the problem

$$\begin{cases} c_t + vc_x = f(x, t) & x \in \mathbb{R}, t > 0 \\ c(x, 0) = 0 & x \in \mathbb{R}. \end{cases}$$

Find an explicit formula when $f(x, t) = e^{-t} \sin x$.

[Hint: For fixed $s \geq 0$ and $t > s$, solve

$$\begin{cases} w_t + vw_x = 0 \\ w(x, s; s) = f(x, s). \end{cases}$$

Then, integrate w with respect to s over $(0, t)$].

4.2. Consider the following problem ($a > 0$):

$$\begin{cases} u_t + au_x = f(x, t) & 0 < x < R, t > 0 \\ u(0, t) = 0 & t > 0 \\ u(x, 0) = 0 & 0 < x < R. \end{cases}$$

Prove the stability estimate

$$\int_0^R u^2(x, t) dx \leq e^t \int_0^t \int_0^R f^2(x, s) dx ds, \quad t > 0.$$

[Hint: Multiply by u the equation. Use $a > 0$ and the inequality $2fu \leq f^2 + u^2$ to obtain

$$\frac{d}{dt} \int_0^R u^2(x, t) dx \leq \int_0^R f^2(x, t) dx + \int_0^R u^2(x, t) dx.$$

Prove that if $E(t)$ satisfies $E'(t) \leq G(t) + E(t)$, $E(0) = 0$, then $E(t) \leq e^t \int_0^t G(s) ds$].

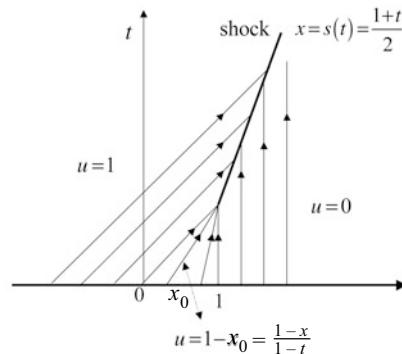


Fig. 4.38 The solution of Problem 4.3

4.3. Solve the Burgers equation $u_t + uu_x = 0$ with initial data

$$g(x) = \begin{cases} 1 & x \leq 0 \\ 1-x & 0 < x < 1 \\ 0 & x \geq 1. \end{cases}$$

[Answer: See Fig. 4.38].

4.4. Solve the problem

$$\begin{cases} ut + uu_x = 0 & x \in \mathbb{R}, t > 0 \\ u(x, 0) = g(x) & x \in \mathbb{R} \end{cases}$$

where

$$g(x) = \begin{cases} 0 & x < 0 \\ 1 & 0 < x < 1 \\ 0 & x > 1. \end{cases}$$

4.5. *A rather frequent traffic situation.* Consider the model of Sect. 4.3.1 and suppose that two traffic lights are placed at distance L , at the points $x = -\beta L$, $x = \alpha L$, with $\beta > 0$, $0 < \alpha < 1/2$, $\alpha + \beta = 1$. At time $t = 0$, the car distribution is given by

$$\begin{cases} \rho_m & x < -\beta L \\ 0 & -\beta L < x < 0 \\ \rho_m & 0 < x < \alpha L \\ 0 & x > \alpha L. \end{cases}$$

Thus, there is bumper-to-bumper congestion before the first light, a piece of clear road between $-\beta L$ and 0 , a congested queue of length αL before the second light, beyond which the road is clear.

a) Introduce dimensionless variables

$$u = \frac{\rho}{\rho_m}, z = \frac{x}{L}, \tau = \frac{v_m t}{L},$$

to get the equation $u_\tau + \tilde{q}(u)_z = 0$, where $\tilde{q}(u) = u(1-u)$, with initial condition

$$\begin{cases} 1 & z < -\beta \\ 0 & -\beta < z < 0 \\ 1 & 0 < z < \alpha \\ 0 & z > \alpha. \end{cases}$$

- b) Assuming that both lights turn on green at $\tau = 0$, determine the characteristic configuration and the shock curve.
- c) Compute the time τ_α at which the shock curve intersects the vertical line $z = \alpha$. Determine the path of the vehicles started from the queue in the road section between 0 and α . Deduce how long the second light should stay on green, in order to let all the cars in the queue to pass the light at $z = \alpha$.

[Partial answer: a) There is a shock starting $(0, 0)$ and two rarefaction waves centered at $(-\beta, 0)$ and $(\alpha, 0)$. The shock curve has equation

$$z = \begin{cases} 0 & 0 < \tau < \alpha \\ \alpha + \tau - 2\sqrt{\alpha\tau} & \alpha \leq \tau < 1/4\alpha \\ 2(1-2\alpha)\tau - (1-2\alpha) & 1/4\alpha \leq \tau. \end{cases}$$

The characteristic fans of the two rarefaction waves are delimited by the lines (from left to right) $z = -\beta - \tau$, $z = -\beta + \tau$ and $z = \alpha - \tau$, $z = \alpha + \tau$ respectively.

b) $\tau_\alpha = 1/(2-4\alpha)$. The green time must be greater than τ_α].

4.6. Traffic in a tunnel. A rather realistic model for the car speed in a very long tunnel is the following:

$$v(\rho) = \begin{cases} v_m & 0 \leq \rho \leq \rho_c \\ \lambda \log\left(\frac{\rho_m}{\rho}\right) & \rho_c \leq \rho \leq \rho_m \end{cases}$$

where $\lambda = \frac{v_m}{\log(\rho_m/\rho_c)}$.

Observe that v is continuous also at

$$\rho_c = \rho_m e^{-v_m/\lambda},$$

which represents a *critical density*: if $\rho \leq \rho_c$ the drivers are free to reach the speed limit. Typical values are: $\rho_c = 7$ car/Km, $v_m = 90$ Km/h, $\rho_m = 110$ car/Km, $v_m/\lambda = 2.75$.

Assume the entrance is placed at $x = 0$ and that the cars are waiting (with maximum density) the tunnel to open to the traffic at time $t = 0$. Thus, the initial density is

$$\rho = \begin{cases} \rho_m & x < 0 \\ 0 & x > 0. \end{cases}$$

- a) Determine density and car speed; draw their graphs.
- b) Determine and draw in the x, t plane the trajectory of a car initially at $x = x_0 < 0$, and compute the time it takes to enter the tunnel.

4.7. Consider the equation $u_t + q'(u)u_x = 0$, with initial condition $u(x, 0) = g(x)$. Assume that $g, q' \in C^1([a, b])$ and

$$g'(\xi)q''(g(\xi)) < 0$$

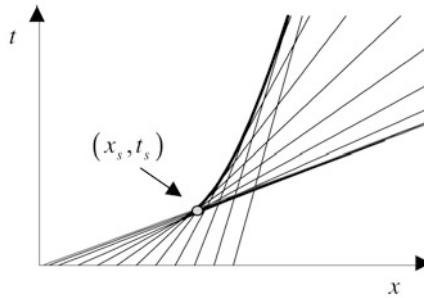


Fig. 4.39 Envelope of characteristics and point of shock formation

in $[a, b]$. Show that the family of characteristics

$$x = q'(u)t + \xi, \quad \xi \in [a, b] \quad (4.151)$$

admits an *envelope* and that the point (x_s, t_s) of shock formation, given by formulas (4.43) and (4.44), is the point on this envelope with minimum time coordinate (Fig. 4.39).

4.8. Find the solutions of the problems

$$\begin{cases} u_t \pm uu_x = 0 & t > 0, x \in \mathbb{R} \\ u(x, 0) = x & x \in \mathbb{R}. \end{cases}$$

4.9. Draw the characteristics and describe the evolution for $t \rightarrow +\infty$ of the solution of the problem

$$\begin{cases} u_t + uu_x = 0 & t > 0, x \in \mathbb{R} \\ u(x, 0) = \begin{cases} \sin x & 0 < x < \pi \\ 0 & x \leq 0 \text{ or } x \geq \pi. \end{cases} \end{cases}$$

4.10. Show that, for every $\alpha > 1$, the function

$$u_\alpha(x, t) = \begin{cases} -1 & 2x \leq -(1+\alpha)t \\ -\alpha & -(1+\alpha)t < 2x < 0 \\ \alpha & 0 < 2x < (\alpha+1)t \\ 1 & (\alpha+1)t \leq 2x \end{cases}$$

is a weak solution of the problem

$$\begin{cases} u_t + uu_x = 0 & t > 0, x \in \mathbb{R} \\ u(x, 0) = \begin{cases} -1 & x < 0 \\ 1 & x > 0. \end{cases} \end{cases}$$

Is it also an entropic solution, at least for some α ?

4.11. *Time irreversibility.* Let u_1 and u_2 be the weak solutions of the Burgers equation $u_t + (u^2/2)_x = 0$ with initial data

$$g_1(x) = \begin{cases} 1 & x < 1/2 \\ 0 & x > 1/2 \end{cases} \quad \text{and} \quad g_2(x) = \begin{cases} 1 & x < 0 \\ 1-x & 0 \leq x \leq 1 \\ 0 & x > 1. \end{cases}$$

Show that at time $t = 1$,

$$u_1(x, 1) = u_2(x, 2) = \begin{cases} 1 & x < 1 \\ 0 & x > 1. \end{cases}$$

Thus we cannot reconstruct the initial data of a weak solution u from the knowledge of u at a later time.

4.12. Using the Hopf-Cole transformation, solve the following problem for the viscous Burgers equation

$$\begin{cases} u_t + uu_x = \varepsilon u_{xx} & t > 0, x \in \mathbb{R} \\ u(0, x) = \mathcal{H}(x) & x \in \mathbb{R}, \end{cases}$$

where \mathcal{H} is the Heaviside function.

[Answer: The solution is

$$u(x, t) = \frac{1}{1 + \frac{\operatorname{erfc}(-x/\sqrt{4\varepsilon t})}{\operatorname{erfc}((x-t)/\sqrt{4\varepsilon t})} \exp\left(\frac{x-t/2}{2\varepsilon}\right)}$$

where $\operatorname{erfc}(s)$ is the *complementary error function* (4.85)].

4.13. Consider the conservation law $u_t + q(u)_x = 0$, where the graph of q is shown in Fig. 4.25). Exchange the roles of u_R and u_L , so that $u_l < u_R$, and construct the entropic solution of the Riemann problem.

[Hint: Since $u_L < u_R$, replace the graph of q by its lower convex envelope, shown in Fig. 4.40.]

4.14. Examine the sedimentation problem of Sect. 4.6 in the two cases

$$1) 0 < u_0 \leq a, \quad 2) u_F \leq u_0 < 1.$$

Determine the characteristics configuration as in Fig. 4.31, p. 228 and the solution profile at some significant time, as in Figs. 4.32, 4.33, p. 229.

4.15. Find the solution of the linear equation

$$u_x + xu_y = y$$

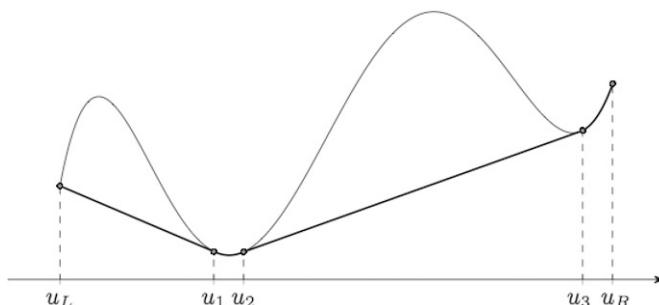


Fig. 4.40 Lower convex envelope of the flux function in Fig. 4.25

256 4 Scalar Conservation Laws and First Order Equations

satisfying the initial condition $u(0, y) = g(y)$, $y \in \mathbb{R}$, with

$$(a) \quad g(y) = \cos y \quad \text{and} \quad (b) \quad g(y) = y^2.$$

[Answer of (a): $u = xy - \frac{x^3}{3} + \cos\left(y - \frac{x^2}{2}\right)$].

4.16. Consider the linear equation

$$au_x + bu_y = c(x, y),$$

with the initial condition $u(x, 0) = h(x)$. Assume that a, b are constants ($b \neq 0$), and c, h are smooth and bounded.

- 1) Show that

$$u(x, y) = h(x - \gamma y) + \int_0^{y/b} c(a\tau + x - \gamma y, b\tau) d\tau, \quad \gamma = a/b.$$

- 2) Deduce that a jump discontinuity of h at x_0 propagates into a jump of the same size for u , along the characteristic of equation $x - \gamma y = x_0$.

4.17. Let

$$D = \{(x, y) : y > x^2\}$$

and $a = a(x, y)$ be a continuous function in \overline{D} .

- 1) Given $g \in C(\mathbb{R})$, check the solvability of the linear problem

$$\begin{cases} a(x, y)u_x - u_y = -u & (x, y) \in D \\ u(x, x^2) = g(x) & x \in \mathbb{R}. \end{cases}$$

- 2) Examine the cases

$$a(x, y) = y/2 \quad \text{and} \quad g(x) = \exp(-\gamma x^2),$$

where γ is a real parameter.

4.18. Solve the Cauchy problem

$$\begin{cases} xu_x - yu_y = u - y & x > 0, y > 0 \\ u(y^2, y) = y & y > 0. \end{cases}$$

May a solution exist in a neighborhood of the origin?

[Answer: $u(x, y) = (y + x^{2/3}y^{-1/3})/2$. No solution can exist in neighborhood of $(0, 0)$].

4.19. Consider a cylindrical pipe with axis along the x -axis, filled with a fluid moving along the positive direction. Let $\rho = \rho(x, t)$ and $q = \frac{1}{2}\rho^2$ be the fluid density and the flux function. Assume the walls of the pipe are composed by porous material, from which the fluid leaks at the rate $H = k\rho^2$, $k > 0$.

- a) Following the derivation of the conservation law given in the introduction, show that ρ satisfies the equation

$$\rho_t + \rho\rho_x = -k\rho^2.$$

- b) Solve the Cauchy problem with $\rho(x, 0) = 1$.

[Answer: b) $\rho(x, t) = 1/(1 + kt)$].

4.20. Solve the Cauchy problem

$$\begin{cases} u_x = -(u_y)^2 & x > 0, y \in \mathbb{R} \\ u(0, y) = 3y & y \in \mathbb{R}. \end{cases}$$

4.21. Solve the Cauchy problem

$$\begin{cases} u_x^2 + u_y^2 = 4u & (x, y) \in \mathbb{R}^2 \\ u(x, -1) = x^2 & x \in \mathbb{R}. \end{cases}$$

[Answer: $u(x, y) = x^2 + (y + 1)^2$].

4.22. Solve the Cauchy problem

$$\begin{cases} c^2(u_x^2 + u_y^2) = 1 & (x, y) \in \mathbb{R}^2 \\ u(\cos s, \sin s) = 0 & s \in \mathbb{R}. \end{cases}$$

[Answer: There are two solutions

$$u^\pm(x, y) = \frac{\pm 1}{c} \left\{ 1 - \sqrt{x^2 + y^2} \right\}$$

whose wave fronts are shown in Fig. 4.41].

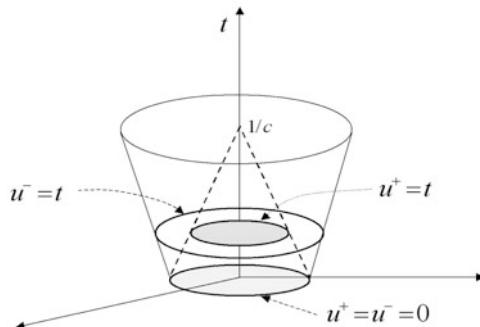


Fig. 4.41 Solutions of Problem 4.21

Chapter 5

Waves and Vibrations

5.1 General Concepts

5.1.1 Types of waves

Our daily experience deals with sound waves, electromagnetic waves (as radio or light waves), deep or surface water waves, elastic waves in solid materials. Oscillatory phenomena manifest themselves also in less macroscopic and known contexts and ways. This is the case, for instance, of rarefaction and shock waves in traffic dynamics or of electrochemical waves in human nervous system and in the regulation of the heart beat. In quantum physics, everything can be described in terms of wave functions, at a sufficiently small scale.

Although the above phenomena share many similarities, they show several differences as well. For example, progressive water waves propagate a disturbance, while standing waves do not. Sound waves need a supporting medium, while electromagnetic waves do not. Electrochemical waves interact with the supporting medium, in general modifying it, while water waves do not.

Thus, it seems too hard to give a general definition of *wave*, capable of covering all the above cases, so that we limit ourselves to introduce some terminology and general concepts, related to specific types of waves. We start with one-dimensional waves.

a. Progressive or travelling waves are disturbances described by a function of the following form:

$$u(x, t) = g(x - ct).$$

For $t = 0$, we have $u(x, 0) = g(x)$, which is the “initial” profile of the perturbation. This profile propagates without change of shape with speed $|c|$, in the positive (negative) x -direction if $c > 0$ ($c < 0$). We have already met this kind of waves in Chaps. 2 and 4.

b. Harmonic waves are particular progressive waves of the form

$$u(x, t) = A \exp\{i(kx - \omega t)\}, \quad A, k, \omega \in \mathbb{R}. \quad (5.1)$$

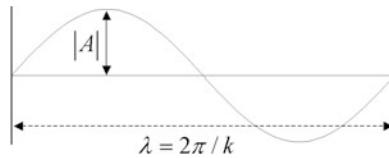


Fig. 5.1 Sinusoidal wave

It is understood that only the *real part* (or the imaginary part)

$$A \cos(kx - \omega t)$$

is of interest, but the complex notation may often simplify the computations. In (5.1) we distinguish, considering for simplicity ω and k positive:

- The wave *amplitude* $|A|$.
- The *wave number* k , which is the number of complete oscillations in the space interval $[0, 2\pi]$, and the *wavelength*

$$\lambda = \frac{2\pi}{k},$$

which is the distance between successive maxima (*crest*) or minima (*troughs*) of the waveform.

- The *angular frequency* ω , and the *frequency*

$$f = \frac{\omega}{2\pi}$$

which is the number of complete oscillations in one second (Hertz) at a fixed space position.

- The *wave or phase speed*

$$c_p = \frac{\omega}{k},$$

which is the crests (or troughs) speed.

c. Standing waves

$$u(x, t) = B \cos kx \cos \omega t.$$

In these disturbances, the basic sinusoidal wave, $\cos kx$, is modulated by the time dependent oscillation $B \cos \omega t$. A standing wave may be generated, for instance, by superposing two harmonic waves with the same amplitude, propagating in opposite directions:

$$A \cos(kx - \omega t) + A \cos(kx + \omega t) = 2A \cos kx \cos \omega t. \quad (5.2)$$

Consider now waves in dimension $n > 1$.

d. Plane waves. *Scalar* plane waves are of the form

$$u(\mathbf{x}, t) = f(\mathbf{k} \cdot \mathbf{x} - \omega t).$$

The disturbance propagates in the direction of \mathbf{k} with speed $c_p = \omega / |\mathbf{k}|$. The planes of equation

$$\theta(\mathbf{x}, t) = \mathbf{k} \cdot \mathbf{x} - \omega t = \text{constant}$$

constitute the *wave-fronts*. *Harmonic or monochromatic plane waves* have the form

$$u(\mathbf{x}, t) = A \exp\{i(\mathbf{k} \cdot \mathbf{x} - \omega t)\}.$$

Here \mathbf{k} is the *wave number* vector and ω is the *angular frequency*. The vector \mathbf{k} is orthogonal to the wave front and $|\mathbf{k}|/2\pi$ gives the number of waves per unit length. The scalar $\omega/2\pi$ still gives the number of complete oscillations in one second (Hertz) at a fixed space position.

e. Spherical waves are of the form

$$u(\mathbf{x}, t) = v(r, t)$$

where $r = |\mathbf{x} - \mathbf{x}_0|$ and $\mathbf{x}_0 \in \mathbb{R}^n$ is a fixed point. In particular $u(\mathbf{x}, t) = e^{i\omega t}v(r)$ represents a stationary spherical wave, while $u(\mathbf{x}, t) = v(r - ct)$ is a progressive wave whose wavefronts are the spheres $r - ct = \text{constant}$, moving with speed $|c|$ (outgoing if $c > 0$, incoming if $c < 0$).

5.1.2 Group velocity and dispersion relation

Many oscillatory phenomena can be modelled by linear equations whose solutions are superpositions of harmonic waves with angular frequency depending on the wave number:

$$\omega = \omega(k). \quad (5.3)$$

A typical example is the wave system produced by dropping a stone in a pond.

If ω is linear, e.g. $\omega(k) = ck$, $c > 0$, the crests move with speed c , independent of the wave number. However, if $\omega(k)$ is not proportional to k , the crests move with speed $c_p = \omega(k)/k$, that depends on the wave number. In other words, the crests move at different speeds for different wavelengths. As a consequence, the various components in a wave packet, given by the superposition of harmonic waves of different wavelengths, will eventually separate or *disperse*. For this reason, (5.3) is called **dispersion relation**.

In the theory of dispersive waves, the **group velocity**, given by

$$c_g = \omega'(k),$$

is a central notion, mainly for the following three reasons.

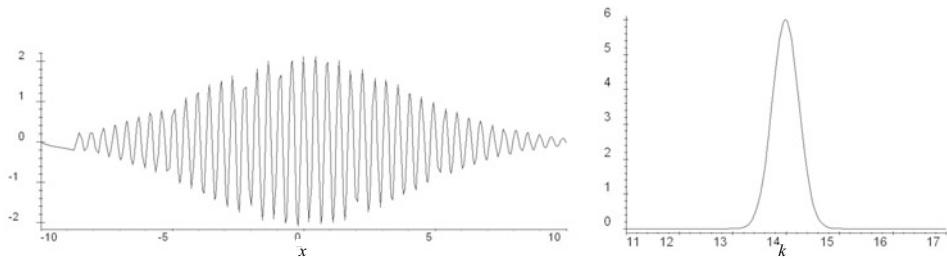


Fig. 5.2 Wave packet and its Fourier transform

1. *The group velocity is the speed at which an isolated wave packet moves as a whole.* A wave packet may be obtained by the superposition of dispersive harmonic waves, for instance through a Fourier integral of the form

$$u(x, t) = \int_{-\infty}^{+\infty} a(k) e^{i[kx - \omega(k)t]} dk \quad (5.4)$$

where only the real part has a physical meaning. Consider a localized wave packet, with wave number $k \approx k_0$, almost constant, and with amplitude slowly varying with x . Then, the packet contains a large number of crests and the amplitudes $|a(k)|$ of the various Fourier components are negligible except that in a small neighborhood of k_0 , say, $(k_0 - \delta, k_0 + \delta)$.

Figure 5.2 shows the initial profile of a Gaussian packet,

$$\operatorname{Re} u(x, 0) = \frac{3}{\sqrt{2}} \exp \left\{ -\frac{x^2}{32} \right\} \cos 14x,$$

slowly varying with x , with $k_0 = 14$, and its Fourier transform:

$$a(k) = 6 \exp \{-8(k - 14)^2\}.$$

As we can see, the amplitudes $|a(k)|$ of the various Fourier components are negligible except when k is near k_0 .

Then we may write

$$\omega(k) \approx \omega(k_0) + \omega'(k_0)(k - k_0) = \omega(k_0) + c_g(k - k_0)$$

and

$$u(x, t) \approx e^{i\{k_0 x - \omega(k_0)t\}} \int_{k_0 - \delta}^{k_0 + \delta} a(k) e^{i(k - k_0)(x - c_g t)} dk. \quad (5.5)$$

Thus, u turns out to be well approximated by the product of two waves. The first one is a pure harmonic wave with relatively short wavelength $2\pi/k_0$ and phase speed $\omega(k_0)/k_0$. The second one depends on x, t through the combination $x - c_g t$, and is a superposition of waves of very small wavenumbers $k - k_0$, which corre-

spond to very large wavelengths. We may interpret the second factor as a sort of envelope of the short waves of the packet, that is the packet as a whole, which therefore moves with the group speed.

2. *An observer traveling at the group velocity sees constantly waves of the same wavelength $2\pi/k$, after the transitory effects due to a localized initial perturbation (e.g. a stone thrown into a pond). In other words, c_g is the propagation speed of the wave numbers.*

Imagine dropping a stone into a pond. The water perturbation looks complicated, at the beginning. After a sufficiently long time, the various Fourier components will be quite dispersed and the perturbation will appear as a slowly modulated wave train, almost sinusoidal near every point, with a *local wave number* $k(x, t)$ and a *local frequency* $\omega(x, t)$. If the water is deep enough, we expect that, at each fixed time t , the wavelength increases with the distance from the stone (longer waves move faster, see Subsect. 5.11.4) and that, at each fixed point x , the wavelength tends to decrease with time.

Thus, the essential features of the wave system can be observed at a relatively long distance from the location of the initial disturbance and after some time has elapsed.

We may assume that the free surface displacement u is given by a Fourier integral of the form (5.4) and we are interested on the behavior of u for $t \gg 1$. An important tool comes from the method of stationary phase¹ which gives an asymptotic formula for integrals of the form

$$I(t) = \int_{-\infty}^{+\infty} f(k) e^{it\varphi(k)} dk \quad (5.6)$$

as $t \rightarrow +\infty$. We can put u into the form (5.6) by writing

$$u(x, t) = \int_{-\infty}^{+\infty} a(k) e^{it[k\frac{x}{t} - \omega(k)]} dk,$$

by moving from the origin at a fixed speed V (thus $x = Vt$) and then defining

$$\varphi(k) = kV - \omega(k).$$

Assume for simplicity that φ has only one stationary point k_0 , that is

$$\omega'(k_0) = V,$$

and that $\omega''(k_0) \neq 0$. Then, according to the *method of stationary phase*, we can write

$$u(Vt, t) = \sqrt{\frac{\pi}{|\omega''(k_0)|}} \frac{a(k_0)}{\sqrt{t}} \exp\{it[k_0V - \omega(k_0)]\} + O(t^{-1}). \quad (5.7)$$

¹ See Sect. 5.11.6.

Thus, if we allow errors of order t^{-1} , by moving with speed $V = \omega'(k_0) = c_g$, the same wave number k_0 always appears at the position $x = c_g t$. Note that the amplitude decreases like $t^{-1/2}$ as $t \rightarrow +\infty$. This is a typical attenuation effect of dispersion.

3. Energy is transported at the group velocity by waves of wavelength $2\pi/k$. In a wave packet like (5.5), the energy is proportional to²

$$\int_{k_0-\delta}^{k_0+\delta} |a(k)|^2 dk \simeq 2\delta |a(k_0)|^2$$

so that it moves at the same speed of k_0 , that is c_g .

Since the energy travels at the group velocity, there are significant differences in the wave system according to the sign of $c_g - c_p$, as we will see in Sect. 5.11.

5.2 Transversal Waves in a String

5.2.1 The model

We now derive a classical model for the small transversal vibration of a tightly stretched horizontal string (e.g. the string of a guitar). We assume the following hypotheses:

1. *Vibrations of the string have small amplitude.* This entails that the changes in the slope of the string from the horizontal equilibrium position are very small.
2. *Each point of the string undergoes vertical displacements only.* Horizontal displacements can be neglected.
3. *The vertical displacement of a point depends on time and on its position on the string.* If we denote by u the vertical displacement of a point located at x when the string is at rest, then we have $u = u(x, t)$ and, according to 1, $|u_x(x, t)| \ll 1$.
4. *The string is perfectly flexible.* This means that it offers no resistance to bending. In particular, the stress at any point on the string can be modelled by a tangential³ force \mathbf{T} of magnitude τ , called *tension*. Fig. 5.3 shows how the forces due to the tension acts at the end points of a small segment of the string.
5. *Friction is negligible.*

Under the above assumptions, the equation of motion of the string can be derived from *conservation of mass* and *Newton's law*.

Let $\rho_0 = \rho_0(x)$ be the linear density of the string at rest and $\rho = \rho(x, t)$ be its density at time t . Consider an arbitrary part of the string between x and $x + \Delta x$ and denote by Δs the corresponding length element at time t . Then, the

² See e.g. [29], A. Segel, 1987.

³ Consequence of absence of distributed moments along the string.

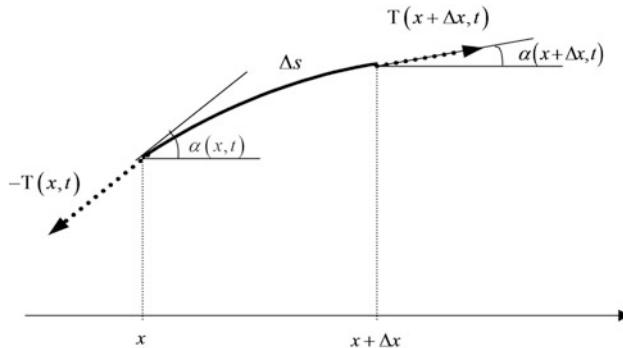


Fig. 5.3 Tension at the end points of a small segment of a string

conservation of mass yields

$$\rho_0(x) \Delta x = \rho(x, t) \Delta s. \quad (5.8)$$

To write Newton's law of motion we have to determine the forces acting on our small piece of string. Since the motion is vertical, the horizontal forces have to balance. On the other hand these forces come from the tension only, so that if $\tau(x, t)$ denotes the magnitude of the tension at x at time t , we can write (Fig. 5.3):

$$\tau(x + \Delta x, t) \cos \alpha(x + \Delta x, t) - \tau(x, t) \cos \alpha(x, t) = 0.$$

Dividing by Δx and letting $\Delta x \rightarrow 0$, we obtain

$$\frac{\partial}{\partial x} [\tau(x, t) \cos \alpha(x, t)] = 0,$$

from which

$$\tau(x, t) \cos \alpha(x, t) = \tau_0(t), \quad (5.9)$$

where $\tau_0(t)$ is *positive*.

The vertical forces are given by the vertical component of the tension and by body forces such as gravity and external loads. Using (5.9), the scalar vertical component of the tension at x , at time t , is given by:

$$\tau_{vert}(x, t) = \tau(x, t) \sin \alpha(x, t) = \tau_0(t) \tan \alpha(x, t) = \tau_0(t) u_x(x, t).$$

Therefore, the (scalar) vertical component of the force acting on our small piece of string, due to the tension, is

$$\tau_{vert}(x + \Delta x, t) - \tau_{vert}(x, t) = \tau_0(t) [u_x(x + \Delta x, t) - u_x(x, t)].$$

Denote by $f(x, t)$ the magnitude of the (vertical) body forces per unit mass. Then, using (5.8), the magnitude of the body forces acting on the string segment is given

by:

$$\int_x^{x+\Delta x} \rho(y, t) f(y, t) dy = \int_x^{x+\Delta x} \rho_0(y) f(y, t) dy.$$

Thus, using (5.8) again and observing that u_{tt} is the (scalar) vertical acceleration, Newton's law gives:

$$\int_x^{x+\Delta x} \rho_0(y) u_{tt}(y, t) dy = \tau_0(t) [u_x(x + \Delta x, t) - u_x(x, t)] + \int_x^{x+\Delta x} \rho_0(y) f(y, t) dy.$$

Dividing by Δx and letting $\Delta x \rightarrow 0$, we obtain the equation

$$u_{tt} - c^2(x, t) u_{xx} = f(x, t) \quad (5.10)$$

where

$$c^2(x, t) = \frac{\tau_0(t)}{\rho_0(x)}.$$

If the string is homogeneous, then ρ_0 is constant. If moreover the string is **perfectly elastic**⁴, then τ_0 is constant as well, since the horizontal tension is nearly the same as for the string at rest, in the horizontal position. We shall come back to eq. (5.10) shortly.

5.2.2 Energy

Suppose that a *perfectly flexible and elastic* string has length L at rest, in the horizontal position. We may identify its initial position with the segment $[0, L]$ on the x axis. Since $u_t(x, t)$ is the vertical velocity of the point located at x , the expression

$$E_{cin}(t) = \frac{1}{2} \int_0^L \rho_0 u_t^2 dx \quad (5.11)$$

represents the total **kinetic energy during the vibrations**. The string stores **potential energy** too, due to the work of elastic forces. These forces stretch an element of string of length Δx at rest, by

$$\Delta s - \Delta x = \int_x^{x+\Delta x} \sqrt{1 + u_x^2} dx - \Delta x = \int_x^{x+\Delta x} \left(\sqrt{1 + u_x^2} - 1 \right) dx \approx \frac{1}{2} u_x^2 \Delta x,$$

since⁵ $|u_x| \ll 1$. Thus, the work done by the elastic forces on that string element is

$$dW = \frac{1}{2} \tau_0 u_x^2 \Delta x.$$

⁴ For instance, guitar and violin strings are nearly homogeneous, perfectly flexible and elastic.

⁵ Recall that, at first order, if $\varepsilon \ll 1$, $\sqrt{1 + \varepsilon} - 1 \simeq \varepsilon/2$.

Summing all the contributions, the total **potential energy** is given by

$$E_{pot}(t) = \frac{1}{2} \int_0^L \tau_0 u_x^2 dx. \quad (5.12)$$

From (5.11) and (5.12) we find that the total mechanical energy is

$$E(t) = \frac{1}{2} \int_0^L [\rho_0 u_t^2 + \tau_0 u_x^2] dx. \quad (5.13)$$

Let us compute the variation of E . Taking the time derivative under the integral, we find (remember that $\rho_0 = \rho_0(x)$ and τ_0 is constant),

$$\dot{E}(t) = \int_0^L [\rho_0 u_t u_{tt} + \tau_0 u_x u_{xt}] dx.$$

Integrating by parts we get

$$\int_0^L \tau_0 u_x u_{xt} dx = \tau_0 [u_x(L, t) u_t(L, t) - u_x(0, t) u_t(0, t)] - \tau_0 \int_0^L u_t u_{xx} dx,$$

whence

$$\dot{E}(t) = \int_0^L [\rho_0 u_{tt} - \tau_0 u_{xx}] u_t dx + \tau_0 [u_x(L, t) u_t(L, t) - u_x(0, t) u_t(0, t)].$$

Using (5.10), we find:

$$\dot{E}(t) = \int_0^L \rho_0 f u_t dx + \tau_0 [u_x(L, t) u_t(L, t) - u_x(0, t) u_t(0, t)]. \quad (5.14)$$

In particular, if $f = 0$ and u is constant at the end points 0 and L (therefore $u_t(L, t) = u_t(0, t) = 0$), we deduce $\dot{E}(t) = 0$. This implies

$$E(t) = E(0)$$

which expresses the *conservation of energy*.

5.3 The One-dimensional Wave Equation

5.3.1 Initial and boundary conditions

Equation (5.10) is called the *one-dimensional wave equation*. The coefficient c has the dimensions of a velocity and in fact, we will shortly see that it represents the wave propagation speed along the string. When $f \equiv 0$, the equation is *homogeneous*

and the *superposition principle holds*: if u_1 and u_2 are solutions of

$$u_{tt} - c^2 u_{xx} = 0 \quad (5.15)$$

and a, b are (real or complex) scalars, then $au_1 + bu_2$ is a solution as well. More generally, if $u_k(x, t)$ is a family of solutions depending on the parameter k (integer or real) and $g = g(k)$ is a function rapidly vanishing at infinity, then

$$\sum_{k=1}^{\infty} u_k(x, t) g(k) \quad \text{and} \quad \int_{-\infty}^{+\infty} u_k(x, t) g(k) dk$$

are still solutions of (5.15).

Suppose we are considering the space-time region $0 < x < L$, $0 < t < T$. In a well posed problem for the (one-dimensional) heat equation it is appropriate to assign the initial profile of the temperature, because of the presence of a first order time derivative, and a boundary condition at both ends $x = 0$ and $x = L$, because of the second order spatial derivative.

By analogy with the Cauchy problem for second order ordinary differential equations, the presence of the second order time derivative in (5.10) suggests that not only the initial profile of the string has to be assigned, but the initial velocity has to be prescribed as well.

Thus, our initial (or Cauchy) data are

$$u(x, 0) = g(x), \quad u_t(x, 0) = h(x), \quad x \in [0, L].$$

The boundary data are formally similar to those for the heat equation:

- *Dirichlet data* describe the displacement at the end points of the string:

$$u(0, t) = a(t), \quad u(L, t) = b(t), \quad t > 0.$$

If $a(t) = b(t) \equiv 0$ (homogeneous data), both ends are fixed, with zero displacement.

• *Neumann data* describe an applied (scalar) vertical tension at the end points. As in the derivation of the wave equation, we may model this tension by $\tau_0 u_x$, so that the Neumann conditions take the form

$$\tau_0 u_x(0, t) = a(t), \quad \tau_0 u_x(L, t) = b(t), \quad t > 0.$$

In the special case of homogeneous data, $a(t) = b(t) \equiv 0$, both ends of the string are attached to a frictionless sleeve and are free to move vertically.

- *Robin data* describe a linear elastic attachment at the end points. One way to realize this type of boundary condition is to attach an end point to a linear spring⁶

⁶ That is a spring which obeys Hooke's law: the strain is a linear function of the stress.

whose other end is fixed. This translates into assigning

$$\tau_0 u_x(0, t) = ku(0, t), \quad \tau_0 u_x(L, t) = -ku(L, t), \quad t > 0,$$

where $k > 0$ is the elastic constant of the spring.

In several concrete situations, *mixed conditions* have to be assigned. For instance, Robin data at $x = 0$ and Dirichlet data at $x = L$.

- *Global Cauchy problem.* We may think of a string of infinite length and assign only the initial data

$$u(x, 0) = g(x), \quad u_t(x, 0) = h(x), \quad x \in \mathbb{R}.$$

Although physically unrealistic, it turns out that the solution of the global Cauchy problem is of fundamental importance. We shall solve it in Sect. 5.4.

Under reasonable assumptions on the data, the above problems are well posed. In the next section we use the method of separation of variables to show the well posedness for a Cauchy-Dirichlet problem.

Remark 5.1. Other kinds of problems for the wave equation are the so called *Goursat problem* and the *characteristic Cauchy problem*. Two examples are given in Problems 5.10, 5.11.

5.3.2 Separation of variables

Suppose that the vibration of a violin chord is modelled by the following Cauchy-Dirichlet problem

$$\begin{cases} u_{tt} - c^2 u_{xx} = 0 & 0 < x < L, t > 0 \\ u(0, t) = u(L, t) = 0 & t \geq 0 \\ u(x, 0) = g(x), u_t(x, 0) = h(x) & 0 \leq x \leq L \end{cases} \quad (5.16)$$

where $c^2 = \tau_0/\rho_0$ is constant.

We want to check whether this problem is *well posed*, that is, whether a solution exists, is unique and it is stable (i.e. it depends “continuously” on the data g and h). For the time being we proceed formally, without worrying too much about the correct hypotheses on g and h and the regularity of u .

- *Existence.* Since the boundary conditions are homogeneous⁷, we try to construct a solution by separation of variables.

Step 1. We start looking for nontrivial solutions of the form

$$U(x, t) = w(t)v(x)$$

⁷ Remember that this assumption is essential for using the separation of variables method.

with $v(0) = v(L) = 0$. Inserting U into the wave equation we find

$$0 = U_{tt} - c^2 U_{xx} = w''(t)v(x) - c^2 w(t)v''(x)$$

that is, separating the variables,

$$\frac{1}{c^2} \frac{w''(t)}{w(t)} = \frac{v''(x)}{v(x)}. \quad (5.17)$$

We have reached a familiar situation: (5.17) is an identity between two functions, one depending on t only and the other one depending on x only. Therefore the two sides of (5.17) must be both equal to the same constant, say λ . Thus, we are led to the equation

$$w''(t) - \lambda c^2 w(t) = 0 \quad (5.18)$$

and to the *eigenvalue problem*

$$v''(x) - \lambda v(x) = 0 \quad (5.19)$$

$$v(0) = v(L) = 0. \quad (5.20)$$

Step 2. Solution of the eigenvalue problem. There are three possibilities for the general integral of (5.19).

- a) If $\lambda = 0$, then $v(x) = A + Bx$ and the (5.20) imply $A = B = 0$.
- b) If $\lambda = \mu^2 > 0$, then $v(x) = Ae^{-\mu x} + Be^{\mu x}$ and again the (5.20) imply $A = B = 0$.
- c) If $\lambda = -\mu^2 < 0$, then $v(x) = A \sin \mu x + B \cos \mu x$. From (5.20) we get

$$\begin{aligned} v(0) &= B = 0 \\ v(1) &= A \sin \mu L + B \cos \mu L = 0, \end{aligned}$$

whence

$$A \text{ arbitrary}, B = 0, \mu L = m\pi, m = 1, 2, \dots .$$

Thus, only in case c) we find nontrivial solutions, given by

$$v_m(x) = A \sin \mu_m x, \quad \mu_m = \frac{m\pi}{L}. \quad (5.21)$$

Step 3. Insert $\lambda = -\mu_m^2 = -m^2\pi^2/L^2$ into (5.18). Then, the general solution is

$$w_m(t) = C \cos(\mu_m ct) + D \sin(\mu_m ct) \quad (C, D \in \mathbb{R}). \quad (5.22)$$

From (5.21) and (5.22) we construct the family of solutions

$$U_m(x, t) = [a_m \cos(\mu_m ct) + b_m \sin(\mu_m ct)] \sin \mu_m x, \quad m = 1, 2, \dots$$

where a_m and b_m are arbitrary constants.

U_m is called the m^{th} -**normal mode** of vibration or m^{th} -*harmonic*, and it represents a *standing wave* with frequency $m/2L$. The first harmonic and its frequency $1/2L$, the lowest possible, are said to be *fundamental*. All the other frequencies are *integral multiples* of the fundamental one. It seems that a violin chord produces good quality tones, pleasant to the ear, because of this reason. This is not the case, for instance, for a vibrating membrane like a drum, as we will see shortly.

Step 4. If the initial conditions are

$$u(x, 0) = A \sin \mu_m x \quad u_t(x, 0) = B \sin \mu_m x$$

then the solution of our problem is U_m , with $a_m = A$ and $b_m = B/(\mu_m c)$, and the string vibrates at its m^{th} -mode. In general, the solution is constructed by superposing the harmonics U_m through the formula

$$u(x, t) = \sum_{m=1}^{\infty} [a_m \cos(\mu_m ct) + b_m \sin(\mu_m ct)] \sin \mu_m x, \quad (5.23)$$

where the coefficients a_m and b_m have to be chosen in order to satisfy the initial conditions

$$u(x, 0) = \sum_{m=1}^{\infty} a_m \sin \mu_m x = g(x) \quad (5.24)$$

and

$$u_t(x, 0) = \sum_{m=1}^{\infty} c \mu_m b_m \sin \mu_m x = h(x), \quad (5.25)$$

for $0 \leq x \leq L$.

Looking at (5.24) and (5.25), it is natural to assume that both g and h have an expansion in Fourier sine series in the interval $[0, L]$. Let

$$\hat{g}_m = \frac{2}{L} \int_0^L g(x) \sin \mu_m x \, dx \quad \text{and} \quad \hat{h}_m = \frac{2}{L} \int_0^L h(x) \sin \mu_m x \, dx$$

be the Fourier sine coefficients of g and h . If we choose

$$a_m = \hat{g}_m, \quad b_m = \frac{\hat{h}_m}{\mu_m c}, \quad (5.26)$$

then (5.23) becomes

$$u(x, t) = \sum_{m=1}^{\infty} \left[\hat{g}_m \cos(\mu_m ct) + \frac{\hat{h}_m}{\mu_m c} \sin(\mu_m ct) \right] \sin \mu_m x \quad (5.27)$$

and satisfies (5.24) and (5.25).

Although every U_m is a smooth solution of the wave equation, in principle (5.27) is only a formal solution, unless we may differentiate term by term twice with respect to both x and t , obtaining

$$(\partial_{tt} - c^2 \partial_{xx}^2)u(x, t) = \sum_{m=1}^{\infty} (\partial_{tt} - c^2 \partial_{xx}^2)U_m(x, t) = 0. \quad (5.28)$$

This is possible if \hat{g}_m and \hat{h}_m vanish sufficiently fast as $m \rightarrow +\infty$. In fact, differentiating term by term twice, we have

$$u_{xx}(x, t) = - \sum_{m=1}^{\infty} \left[\mu_m^2 \hat{g}_m \cos(\mu_m ct) + \frac{\mu_m \hat{h}_m}{c} \sin(\mu_m ct) \right] \sin \mu_m x \quad (5.29)$$

and

$$u_{tt}(x, t) = - \sum_{m=1}^{\infty} \left[\mu_m^2 \hat{g}_m c^2 \cos(\mu_m ct) + \mu_m \hat{h}_m c \sin(\mu_m ct) \right] \sin \mu_m x. \quad (5.30)$$

Thus, if, for instance,

$$|\hat{g}_m| \leq \frac{C}{m^4} \quad \text{and} \quad |\hat{h}_m| \leq \frac{C}{m^3}, \quad (5.31)$$

then

$$|\mu_m^2 \hat{g}_m| \leq \frac{C\pi^2}{L^2 m^2}, \quad \text{and} \quad |\mu_m \hat{h}_m| \leq \frac{C\pi}{L m^2}$$

so that, by the Weierstrass test, the series in (5.29), (5.30) converge uniformly in $[0, L] \times [0, +\infty)$. Since also the series (5.27) is uniformly convergent in $[0, L] \times [0, +\infty)$, differentiation term by term is allowed and u is a C^2 solution of the wave equation.

Under which assumptions on g and h does formula (5.31) hold?

Let $g \in C^4([0, L])$, $h \in C^3([0, L])$ and assume the following compatibility conditions:

$$\begin{aligned} g(0) &= g(L) = g''(0) = g''(L) = 0 \\ h(0) &= h(L) = 0. \end{aligned} \quad (5.32)$$

Then (5.31) hold⁸.

⁸ It is an exercise on integration by parts. For instance, if $f \in C^4([0, L])$ and $f(0) = f(L) = f''(0) = f''(L) = 0$, then, integrating by parts four times, we have

$$\hat{f}_m = \int_0^L f(x) \sin\left(\frac{m\pi}{L}\right) dx = \frac{1}{m^4} \int_0^L f^{(4)}(x) \sin\left(\frac{m\pi}{L}\right) dx$$

and

$$|\hat{f}_m| \leq \max |f^{(4)}| \frac{L}{m^4}.$$

Moreover, under the same assumptions, u attains the initial and boundary date in the pointwise sense. Indeed it is not difficult to check that $u(y, s) \rightarrow 0$ as $u(y, s) \rightarrow (0, t)$ or $u(y, s) \rightarrow (L, t)$, for every $t \geq 0$, and

$$u(y, t) \rightarrow g(x), u_t(y, t) \rightarrow h(x), \quad \text{as } (y, t) \rightarrow (x, 0), \quad (5.33)$$

for every $x \in [0, L]$. We conclude that u is a solution of problem (5.16).

- *Uniqueness.* To show that (5.27) is the unique solution of problem (5.16), we use the conservation of energy. Let u and v be solutions of (5.16). Then $w = u - v$ is a solution of the same problem with zero initial and boundary data. We want to show that $w \equiv 0$.

Formula (5.13) gives, for the total mechanical energy,

$$E(t) = E_{cin}(t) + E_{pot}(t) = \frac{1}{2} \int_0^L [\rho_0 w_t^2 + \tau_0 w_x^2] dx.$$

In our case we have

$$\dot{E}(t) = 0,$$

since $f = 0$ and $w_t(L, t) = w_t(0, t) = 0$, whence

$$E(t) = E(0)$$

for every $t \geq 0$. Since, in particular, $w_t(x, 0) = w_x(x, 0) = 0$, we have

$$E(t) = E(0) = 0$$

for every $t \geq 0$. On the other hand, $E_{cin}(t) \geq 0$, $E_{pot}(t) \geq 0$, so that we deduce

$$E_{cin}(t) = 0, E_{pot}(t) = 0,$$

which force $w_t = w_x = 0$. Therefore w is constant and since $w(x, 0) = 0$, we conclude that $w(x, t) = 0$ for every $t \geq 0$.

- *Stability.* We want to show that small perturbations on the data yield small perturbations on the solution. Clearly, we need to establish how we intend to measure the distance for the data and for the corresponding solutions. For the initial data, we use the *least square distance*, given by⁹

$$\|g_1 - g_2\|_0 = \left(\int_0^L |g_1(x) - g_2(x)|^2 dx \right)^{1/2}.$$

For functions depending also on time, we define

$$\|u - v\|_{0,\infty} = \sup_{t>0} \left(\int_0^L |u(x, t) - v(x, t)|^2 dx \right)^{1/2}$$

which measures the maximum in time of the *least squares distance* in space.

⁹ The symbol $\|g\|$ denotes a *norm* of g . See Chap. 6.

Now, let u_1 and u_2 be solutions of problem (5.16), corresponding to the data g_1, h_1 and g_2, h_2 , respectively. Their difference $w = u_1 - u_2$ is a solution of the same problem with Cauchy data $g = g_1 - g_2$ and $h = h_1 - h_2$. From (5.27) we know that

$$w(x, t) = \sum_{m=1}^{\infty} \left[\hat{g}_m \cos(\mu_m c t) + \frac{\hat{h}_m}{\mu_m c} \sin(\mu_m c t) \right] \sin \mu_m x.$$

From Parseval's identity (A.8) and the elementary inequality $(a + b)^2 \leq 2(a^2 + b^2)$, $a, b \in \mathbb{R}$, we can write

$$\begin{aligned} \int_0^L |w(x, t)|^2 dx &= \frac{L}{2} \sum_{m=1}^{\infty} \left[\hat{g}_m \cos(\mu_m c t) + \frac{\hat{h}_m}{\mu_m c} \sin(\mu_m c t) \right]^2 \\ &\leq L \sum_{m=1}^{\infty} \left[\hat{g}_m^2 + \left(\frac{\hat{h}_m}{\mu_m c} \right)^2 \right]. \end{aligned}$$

Since $\mu_m \geq \pi/L$, using Parseval's identity for g and h , we obtain

$$\begin{aligned} \int_0^L |w(x, t)|^2 dx &\leq L \max \left\{ 1, \left(\frac{L}{\pi c} \right)^2 \right\} \sum_{m=1}^{\infty} [\hat{g}_m^2 + \hat{h}_m^2] \\ &= 2 \max \left\{ 1, \left(\frac{L}{\pi c} \right)^2 \right\} [\|g\|_0^2 + \|h\|_0^2] \end{aligned}$$

whence the stability estimate

$$\|u_1 - u_2\|_{0,\infty}^2 \leq 2 \max \left\{ 1, \left(\frac{L}{\pi c} \right)^2 \right\} [\|g_1 - g_2\|_0^2 + \|h_1 - h_2\|_0^2]. \quad (5.34)$$

Thus, “close” data produce “close” solutions.

Summarizing, we have proved the following result:

Theorem 5.2. *Let $g \in C^4([0, L])$ and $h \in C^3([0, L])$ satisfy conditions (5.32). Then, problem (5.34) has a unique solution $u \in C^2([0, L] \times (0, +\infty))$ given by (5.27). Moreover u satisfies the stability estimate*

$$\|u\|_{0,\infty}^2 \leq 2 \max \left\{ 1, \frac{L}{\pi c} \right\} [\|g\|_0^2 + \|h\|_0^2].$$

Remark 5.3. From (5.27), the chord vibration is given by the superposition of harmonics corresponding to the non-zero Fourier coefficients of the initial data. The complex of such harmonics determines a particular feature of the emitted sound, known as the *timbre*, a sort of signature of the musical instrument!

Remark 5.4. The hypotheses we have made on g and h are unnaturally restrictive. For example, if we pluck a violin chord at a point, the initial profile is continuous but has a corner at that point and cannot be even C^1 . A physically realistic assumption for the initial profile g is *continuity*.

Similarly, if we are willing to model the vibration of a chord set into motion by a strike of a little hammer, we should allow discontinuity in the initial velocity. Thus it is realistic to assume *h bounded*.

Under these weaker hypotheses, the separation of variables method does not work. On the other hand, we have already faced a similar situation in Chap. 4, where the necessity to admit discontinuous solutions of a conservation law has led to a more general and flexible formulation of the initial value problem. Also for the wave equation it is possible to introduce suitable *weak* formulations of the various initial-boundary value problems, in order to include realistic initial data and solutions with a low degree of regularity. A first attempt is shown in Subsect. 5.4.2. A weak formulation, more suitable also for numerical methods, is treated in Chap. 10.

5.4 The d'Alembert Formula

5.4.1 The homogeneous equation

In this section we establish the celebrated formula of d'Alembert for the solution of the following global Cauchy problem:

$$\begin{cases} u_{tt} - c^2 u_{xx} = 0 & x \in \mathbb{R}, t > 0 \\ u(x, 0) = g(x), \quad u_t(x, 0) = h(x) & x \in \mathbb{R}. \end{cases} \quad (5.35)$$

To find the solution of (5.35), we first factorize the wave equation in the following way:

$$(\partial_t - c\partial_x)(\partial_t + c\partial_x)u = 0. \quad (5.36)$$

Now, let

$$v = u_t + cu_x. \quad (5.37)$$

Then v solves the linear transport equation

$$v_t - cv_x = 0,$$

whence

$$v(x, t) = \psi(x + ct),$$

where ψ is a differentiable arbitrary function. From (5.37) we have

$$u_t + cu_x = \psi(x + ct)$$

and formula (4.11), p. 183, yields

$$u(x, t) = \int_0^t \psi(x - c(t-s) + cs) \, ds + \varphi(x - ct),$$

where φ is another arbitrary differentiable function.

Letting $x - ct + 2cs = y$, we find

$$u(x, t) = \frac{1}{2c} \int_{x-ct}^{x+ct} \psi(y) \, dy + \varphi(x - ct). \quad (5.38)$$

To determine ψ and φ , we impose the initial conditions

$$u(x, 0) = \varphi(x) = g(x) \quad (5.39)$$

and

$$u_t(x, 0) = \psi(x) - c\varphi'(x) = h(x),$$

whence

$$\psi(x) = h(x) + cg'(x). \quad (5.40)$$

Inserting (5.40) and (5.39) into (5.38) we get:

$$\begin{aligned} u(x, t) &= \frac{1}{2c} \int_{x-ct}^{x+ct} [h(y) + cg'(y)] \, dy + g(x - ct) \\ &= \frac{1}{2c} \int_{x-ct}^{x+ct} h(y) \, dy + \frac{1}{2} [g(x + ct) - g(x - ct)] + g(x - ct) \end{aligned}$$

and, finally, the **d'Alembert** formula

$$u(x, t) = \frac{1}{2} [g(x + ct) + g(x - ct)] + \frac{1}{2c} \int_{x-ct}^{x+ct} h(y) \, dy. \quad (5.41)$$

- *Uniqueness and stability.* If $g \in C^2(\mathbb{R})$ and $h \in C^1(\mathbb{R})$, formula (5.41) defines a C^2 -solution in the half-plane $\mathbb{R} \times [0, +\infty)$. On the other hand, a C^2 -solution u in $\mathbb{R} \times [0, +\infty)$ has to be given by (5.41), just because we can apply to u the procedure we have used to solve the Cauchy problem. Thus the solution is *unique*. However, observe that *no regularizing effect* takes place here: the solution u remains no more than C^2 for any $t > 0$. Thus, there is a striking difference with diffusion phenomena, governed by the heat equation.

Furthermore, let u_1 and u_2 be the solutions corresponding to the data g_1, h_1 and g_2, h_2 , respectively. Then, if all the data are bounded, the d'Alembert formula for $u_1 - u_2$ yields, for every $x \in \mathbb{R}$ and $t \in [0, T]$,

$$|u_1(x, t) - u_2(x, t)| \leq \|g_1 - g_2\|_\infty + T \|h_1 - h_2\|_\infty$$

where

$$\|g_1 - g_2\|_\infty = \sup_{x \in \mathbb{R}} |g_1(x) - g_2(x)|, \quad \|h_1 - h_2\|_\infty = \sup_{x \in \mathbb{R}} |h_1(x) - h_2(x)|.$$

Therefore, we have stability in *pointwise uniform sense*, at least for finite time.

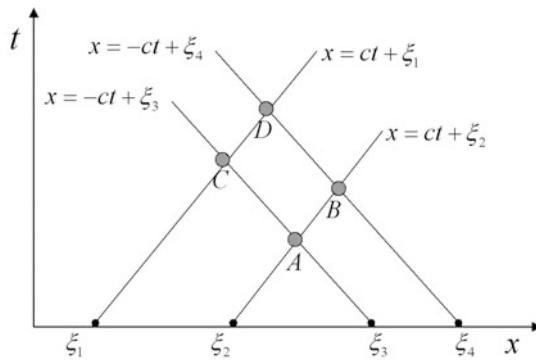


Fig. 5.4 Characteristic parallelogram

- *Progressive waves.* Rearranging the terms in (5.41), we may write u in the form¹⁰

$$u(x, t) = F(x + ct) + G(x - ct) \quad (5.42)$$

which gives u as a *superposition of two progressive waves moving at constant speed c in the negative and positive x -direction, respectively*. Thus, these waves are not dispersive.

The two terms in (5.42) are respectively constant along the two families of straight lines γ^+ and γ^- given by

$$x + ct = \text{constant}, \quad x - ct = \text{constant}.$$

These lines are called *characteristics*¹¹ and carry important information, as we will see in the next subsection.

- *Characteristic parallelogram.* An interesting consequence of (5.42) comes from looking at Fig. 5.4. Consider the *characteristic parallelogram* with vertices at the point A, B, C, D . From (5.42) we have

$$\begin{aligned} F(A) &= F(C), \quad G(A) = G(B) \\ F(D) &= F(B), \quad G(D) = G(C). \end{aligned}$$

¹⁰ For instance:

$$F(x + ct) = \frac{1}{2}g(x + ct) + \frac{1}{2c} \int_0^{x+ct} h(y) dy$$

and

$$G(x - ct) = \frac{1}{2}g(x - ct) + \frac{1}{2c} \int_{x-ct}^0 h(y) dy.$$

¹¹ In fact they are the *characteristics* for the two first order factors in the factorization (5.36). See Sect. 4.2.

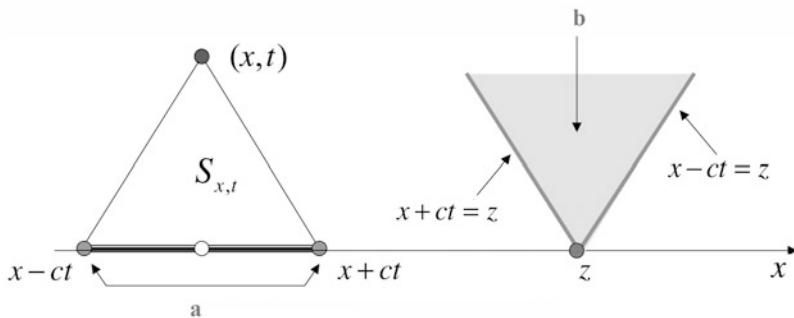


Fig. 5.5 (a) Domain of dependence of (x, t) ; (b) range of influence of z

Summing these relations we get

$$[F(A) + G(A)] + [F(D) + G(D)] = [F(C) + G(C)] + [F(B) + G(B)]$$

which is equivalent to

$$u(A) + u(D) = u(C) + u(B). \quad (5.43)$$

Thus, knowing u at three points of a characteristic parallelogram, we can compute u at the fourth one.

Actually, eq. (5.43) leads to a generalized formulation of the wave equation, as it is shown in Problem 5.6.

• *Domains of dependence and of influence.* From d'Alembert formula it follows that the value of u at the point (x, t) depends on the values of g at the points $x - ct$ and $x + ct$ and on the values of h over the whole interval

$$[x - ct, x + ct].$$

This interval is called **domain of dependence of (x, t)** (Fig. 5.5).

From a different perspective, the values of g and h at a point z affect the value of u at the points (x, t) in the sector

$$z - ct \leq x \leq z + ct,$$

which is called **range of influence of z** (Fig. 5.5). This entails that a disturbance initially localized at z is not felt at a point x until time

$$t = \frac{|x - z|}{c}.$$

Remark 5.5. Differentiating the last term in (5.41) with respect to time we get:

$$\begin{aligned} \frac{\partial}{\partial t} \frac{1}{2c} \int_{x-ct}^{x+ct} h(y) dy &= \frac{1}{2c} [ch(x+ct) - (-c)h(x-ct)] \\ &= \frac{1}{2} [h(x+ct) + h(x-ct)] \end{aligned}$$

which has the form of the first term with g replaced by h . It follows that if w_h denotes the solution of the problem

$$\begin{cases} w_{tt} - c^2 w_{xx} = 0 & x \in \mathbb{R}, t > 0 \\ w(x, 0) = 0, w_t(x, 0) = h(x) & x \in \mathbb{R} \end{cases} \quad (5.44)$$

then, d'Alembert formula can be written in the form

$$u(x, t) = \frac{\partial}{\partial t} w_g(x, t) + w_h(x, t). \quad (5.45)$$

Actually, (5.45) can be established without relying on d'Alembert formula, as we will see later.

5.4.2 Generalized solutions and propagation of singularities

In Remark 5.4, p. 275, we have emphasized the necessity of a weak formulation of problem (5.35) to include more physically realistic data. On the other hand, observe that d'Alembert formula makes perfect sense even for g continuous and h bounded. The question is in which sense the resulting function satisfies the wave equation, since, in principle, it is not even differentiable, only continuous. There are several ways to weaken the notion of solution to include this case; here, for instance, we mimic what we did for conservation laws.

Assuming for the moment that u is a smooth solution of the global Cauchy problem, we multiply the wave equation by a C^2 -test function v , defined in $\mathbb{R} \times [0, +\infty)$ and compactly supported. Integrating over $\mathbb{R} \times [0, +\infty)$ we obtain

$$\int_0^\infty \int_{\mathbb{R}} [u_{tt} - c^2 u_{xx}] v \, dx dt = 0.$$

Let us now integrate by parts both terms twice, in order to transfer all the derivatives from u to v . This yields

$$\int_0^\infty \int_{\mathbb{R}} c^2 u_{xx} v \, dx dt = \int_0^\infty \int_{\mathbb{R}} c^2 u v_{xx} \, dx dt,$$

since v is zero outside a compact subset of $\mathbb{R} \times [0, +\infty)$, and

$$\begin{aligned}\int_0^\infty \int_{\mathbb{R}} u_{tt} v \, dx dt &= - \int_{\mathbb{R}} u_t(x, 0) v(x, 0) \, dx - \int_0^\infty \int_{\mathbb{R}} u_t v_t \, dx dt \\ &= - \int_{\mathbb{R}} [u_t(x, 0) v(x, 0) - u(x, 0) v_t(x, 0)] \, dx + \int_0^\infty \int_{\mathbb{R}} u v_{tt} \, dx dt.\end{aligned}$$

Using the Cauchy data $u(x, 0) = g(x)$ and $u_t(x, 0) = h(x)$, we obtain the integral equation

$$\int_0^\infty \int_{\mathbb{R}} u[v_{tt} - c^2 v_{xx}] \, dx dt + \int_{\mathbb{R}} [g(x) v_t(x, 0) - h(x) v(x, 0)] \, dx = 0. \quad (5.46)$$

Note that (5.46) makes perfect sense for u continuous, g continuous and h bounded. Conversely, if u is a C^2 function that satisfies (5.46) **for every** test function v , then it turns out¹² that u is a solution of problem (5.35).

Thus we may adopt the following definition.

Definition 5.6. Let $g \in C(\mathbb{R})$ and h be bounded in \mathbb{R} . We say that $u \in C(\mathbb{R} \times [0, +\infty))$ is a **generalized solution** of problem (5.35) if (5.46) holds **for every** test function v .

If g is continuous and h is bounded, it can be shown that formula (5.42) constitutes precisely a generalized solution.

Figure 5.6 shows the wave propagation along a chord of infinite length, plucked at the origin and originally at rest, modelled by the solution of the problem

$$\begin{cases} u_{tt} - u_{xx} = 0 & x \in \mathbb{R}, t > 0 \\ u(x, 0) = g(x), u_t(x, 0) = 0 & x \in \mathbb{R} \end{cases}$$

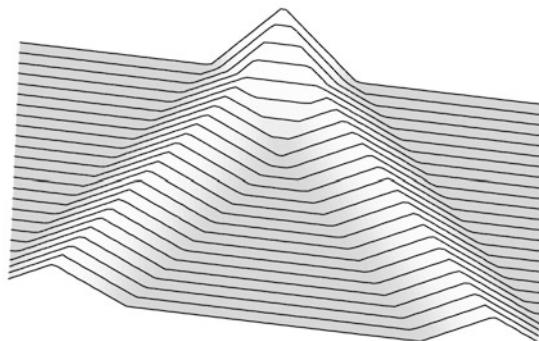


Fig. 5.6 Chord plucked at the origin ($c = 1$)

¹² We invite the reader to check it.

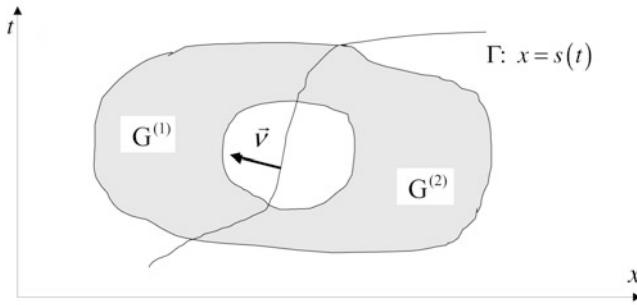


Fig. 5.7 Line of discontinuity of first derivatives

where g has a triangular profile. As we see, this generalized solution displays lines of discontinuities of the first derivatives, while outside these lines it is smooth.

We want to show that these lines are *characteristics*. More generally, consider a region $G \subset \mathbb{R} \times (0, +\infty)$, divided into two domains $G^{(1)}$ and $G^{(2)}$ by a smooth curve Γ of equation $x = s(t)$, as in Fig. 5.7. Let

$$\nu = \nu_1 \mathbf{i} + \nu_2 \mathbf{j} = \frac{1}{\sqrt{1 + (\dot{s}(t))^2}} (-\mathbf{i} + \dot{s}(t) \mathbf{j}) \quad (5.47)$$

be the unit normal to Γ , pointing inward to $G^{(1)}$.

Given any function f defined in G , we denote by $f^{(1)}$ and $f^{(2)}$ its restriction to the closure of $G^{(1)}$ and $G^{(2)}$, respectively, and we use the symbol

$$[f(s(t), t)] = f^{(1)}(s(t), t) - f^{(2)}(s(t), t)$$

for the jump of f across Γ , or simply $[f]$ when there is no risk of confusion.

Now, let u be a generalized solution of Cauchy problem (5.35), of class C^2 in the closure of both¹³ $G^{(1)}$ and $G^{(2)}$, whose first derivatives undergo a jump discontinuity on Γ . We want to prove that:

Proposition 5.7. Γ is a characteristic.

Proof. First of all observe that, from our hypotheses, we have $[u] = 0$ and $[u_x], [u_t] \neq 0$. Moreover, the jumps $[u_x]$ and $[u_t]$ are continuous along Γ .

By analogy with conservation laws, we expect that the integral formulation (5.46) should imply a sort of Rankine-Hugoniot condition, relating the jumps of the derivatives with the slope of Γ , and expressing the balance of linear momentum across Γ .

In fact, let v be a test function with compact support in G . Inserting v into (5.46), we can write

$$0 = \int_G (c^2 uv_{xx} - uv_{tt}) dx dt = \int_{G^{(2)}} (\dots) dx dt + \int_{G^{(1)}} (\dots) dx dt. \quad (5.48)$$

¹³ That is, the first and second derivatives of u extend continuously up to Γ , from both sides, separately.

From Gauss formula, since $v = 0$ on ∂G (dl denotes the arc length on Γ),

$$\begin{aligned} & \int_{G^{(2)}} \left(c^2 u^{(2)} v_{xx} - u^{(2)} v_{tt} \right) dx dt \\ &= \int_{\Gamma} (\nu_1 c^2 u^{(2)} v_x - \nu_2 u^{(2)} v_t) dl - \int_{G^{(2)}} (c^2 u_x^{(2)} v_x - u_t^{(2)} v_t) dx dt \\ &= \int_{\Gamma} (\nu_1 c^2 v_x - \nu_2 v_t) u^{(2)} dl - \int_{\Gamma} (\nu_1 c^2 u_x^{(2)} - \nu_2 u_t^{(2)}) v dl, \end{aligned}$$

since $\int_{G^{(2)}} (c^2 u_{xx}^{(2)} - u_{tt}^{(2)}) v dx dt = 0$. Similarly,

$$\begin{aligned} & \int_{G^{(1)}} (c^2 u^{(1)} v_{xx} - u^{(1)} v_{tt}) dx dt \\ &= - \int_{\Gamma} (\nu_1 c^2 v_x - \nu_2 v_t) u^{(1)} dl + \int_{\Gamma} (\nu_1 c^2 u_x^{(1)} - \nu_2 u_t^{(1)}) v dl, \end{aligned}$$

since $\int_{G^{(2)}} (c^2 u_{xx}^{(1)} - u_{tt}^{(1)}) v dx dt = 0$ as well.

Thus, since $[u] = 0$ on Γ , or more explicitly $[u(s(t), t)] \equiv 0$, (5.48) yields

$$\int_{\Gamma} (c^2 [u_x] \nu_1 - [u_t] \nu_2) v dl = 0.$$

Due to the arbitrariness of v and the continuity of $[u_x]$ and $[u_t]$ on Γ , we deduce

$$c^2 [u_x] \nu_1 - [u_t] \nu_2 = 0, \quad \text{on } \Gamma,$$

or, recalling (5.47),

$$\dot{s} = -c^2 \frac{[u_x]}{[u_t]} \quad \text{on } \Gamma, \tag{5.49}$$

which is the analogue of the Rankine-Hugoniot condition for conservation laws.

On the other hand, differentiating $[u(s(t), t)] \equiv 0$ we obtain

$$\frac{d}{dt} [u(s(t), t)] = [u_x(s(t), t)] \dot{s}(t) + [u_t(s(t), t)] \equiv 0$$

or

$$\dot{s} = -\frac{[u_t]}{[u_x]} \quad \text{on } \Gamma. \tag{5.50}$$

Equations (5.49) and (5.50) entail

$$\dot{s}(t) = \pm c$$

which yields

$$s(t) = \pm ct + \text{constant}$$

showing that Γ is a characteristic. □

5.4.3 The fundamental solution

It is rather instructive to solve the global Cauchy problem with $g \equiv 0$ and a special h : the Dirac delta at a point ξ , that is $h(x) = \delta(x - \xi)$. This problem models, for instance, the vibrations of a violin string generated by a unit impulse localized at ξ (a strike of a sharp hammer). The corresponding solution is called the **fundamen-**

tal solution and plays the same role of the fundamental solution for the diffusion equation.

Certainly, the *Dirac delta* is a quite unusual datum, out of reach of the theory we have developed so far. Therefore, we proceed formally. Thus, let $K = K(x, \xi, t)$ denote our fundamental solution and apply d'Alembert formula; we find

$$K(x, \xi, t) = \frac{1}{2c} \int_{x-ct}^{x+ct} \delta(y - \xi) dy, \quad (5.51)$$

which at first glance looks like a math-*UFO*. To get a more explicit formula, we first compute $\int_{-\infty}^x \delta(y) dy$. To do it, recall that (see Subsect. 2.3.3), if \mathcal{H} is the Heaviside function and

$$I_\varepsilon(y) = \frac{\mathcal{H}(y + \varepsilon) - \mathcal{H}(y - \varepsilon)}{2\varepsilon} = \begin{cases} \frac{1}{2\varepsilon} & -\varepsilon \leq y < \varepsilon \\ 0 & \text{everywhere else} \end{cases} \quad (5.52)$$

is the unit impulse of extent ε , then $\lim_{\varepsilon \downarrow 0} I_\varepsilon(y) = \delta(y)$. Then it seems appropriate to compute $\int_{-\infty}^x \delta(y) dy$ by means of the formula

$$\int_{-\infty}^x \delta(y) dy = \lim_{\varepsilon \downarrow 0} \int_{-\infty}^x I_\varepsilon(y) dy.$$

Now, we have:

$$\int_{-\infty}^x I_\varepsilon(y) dy = \begin{cases} 0 & x \leq -\varepsilon \\ (x + \varepsilon)/2\varepsilon & -\varepsilon < x < \varepsilon \\ 1 & x \geq \varepsilon. \end{cases}$$

Letting $\varepsilon \rightarrow 0$ we deduce that (the value at zero is irrelevant)

$$\int_{-\infty}^x \delta(y) dy = \mathcal{H}(x), \quad (5.53)$$

which actually is not surprising, if we remember that $\mathcal{H}' = \delta$. Everything works nicely. Let us go back to our math-*UFO*, by now . . . identified; we write

$$\int_{x-ct}^{x+ct} \delta(y - \xi) dy = \lim_{\varepsilon \downarrow 0} \int_{-\infty}^{x+ct} I_\varepsilon(y - \xi) dy - \lim_{\varepsilon \downarrow 0} \int_{-\infty}^{x-ct} I_\varepsilon(y - \xi) dy.$$

Then, using (5.51), (5.52) and (5.53), we conclude:

$$K(x, \xi, t) = \frac{1}{2c} \{ \mathcal{H}(x - \xi + ct) - \mathcal{H}(x - \xi - ct) \}. \quad (5.54)$$

Figure 5.8 shows the graph of $K(x, \xi, t)$, with $c = 1$. Note how the initial discontinuity at $x = \xi$ propagates along the characteristics $x = \xi \pm t$.

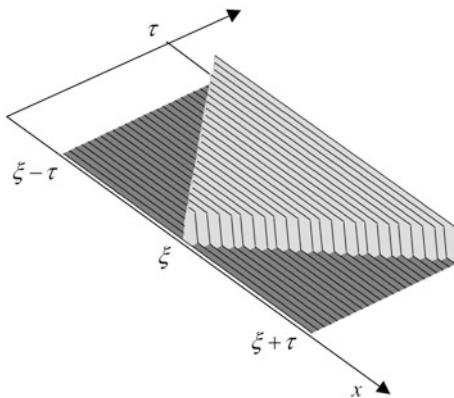


Fig. 5.8 The fundamental solution $K(x, \xi, t)$, $0 < t \leq \tau$

We have found the fundamental solution (5.54) through d'Alembert formula. Conversely, using the fundamental solution we may derive d'Alembert formula. Namely, consider the solution w_h of the Cauchy problem (5.44), with data (see Remark 5.5, p. 278)

$$w(x, 0) = 0, \quad w_t(x, 0) = h(x), \quad x \in \mathbb{R}.$$

We may write

$$h(x) = \int_{-\infty}^{+\infty} \delta(x - \xi) h(\xi) d\xi$$

looking at $h(x)$ as a superposition of impulses $\delta(x - \xi) h(\xi)$, concentrated at ξ . Then, we may construct w_h by superposing the solutions of the same problem with data $\delta(x - \xi) h(\xi)$ instead of h . But these solutions are given by

$$K(x, \xi, t) h(\xi)$$

and therefore we obtain

$$w_h(x, t) = \int_{-\infty}^{+\infty} K(x, \xi, t) h(\xi) d\xi.$$

More explicitly, from (5.54):

$$\begin{aligned} w_h(x, t) &= \frac{1}{2c} \int_{-\infty}^{+\infty} \{H(x - \xi + ct) - H(x - \xi - ct)\} h(\xi) d\xi \\ &= \frac{1}{2c} \int_{-\infty}^{x+ct} h(\xi) d\xi - \frac{1}{2c} \int_{-\infty}^{x-ct} h(\xi) d\xi \\ &= \frac{1}{2c} \int_{x-ct}^{x+ct} h(y) dy. \end{aligned}$$

At this point, (5.45) yields d'Alembert formula.

We shall use this method to construct the solution of the global Cauchy problem in dimension 3.

5.4.4 Nonhomogeneous equation. Duhamel's method

To solve the nonhomogeneous problem

$$\begin{cases} u_{tt} - c^2 u_{xx} = f(x, t) & x \in \mathbb{R}, t > 0 \\ u(x, 0) = 0, u_t(x, 0) = 0 & x \in \mathbb{R}, \end{cases} \quad (5.55)$$

we use the Duhamel's method (see Subsect. 2.2.8). For $s \geq 0$ fixed, let $w = w(x, t; s)$ be the solution of problem

$$\begin{cases} w_{tt} - c^2 w_{xx} = 0 & x \in \mathbb{R}, t \geq s \\ w(x, s; s) = 0, w_t(x, s; s) = f(x, s) & x \in \mathbb{R}. \end{cases} \quad (5.56)$$

Since the wave equation is invariant under (time) translations, from (5.41) we get

$$w(x, t; s) = \frac{1}{2c} \int_{x-c(t-s)}^{x+c(t-s)} f(y, s) dy.$$

Then, the solution of (5.55) is given by

$$u(x, t) = \int_0^t w(x, t; s) ds = \frac{1}{2c} \int_0^t ds \int_{x-c(t-s)}^{x+c(t-s)} f(y, s) dy = \frac{1}{2c} \int_{S_{x,t}} f(y, s) dy ds. \quad (5.57)$$

where $S_{x,t}$ is the *triangular sector* in Fig. 5.5. In fact, $u(x, 0) = 0$ and

$$u_t(x, t) = w(x, t; t) + \int_0^t w_t(x, t; s) ds = \int_0^t w_t(x, t; s) ds$$

since $w(x, t; t) = 0$. Thus $u_t(x, 0) = 0$. Moreover,

$$u_{tt}(x, t) = w_t(x, t; t) + \int_0^t w_{tt}(x, t; s) ds = f(x, t) + \int_0^t w_{tt}(x, t; s) ds$$

and

$$u_{xx}(x, t) = \int_0^t w_{xx}(x, t; s) ds.$$

Therefore, since $w_{tt} - c^2 w_{xx} = 0$,

$$u_{tt}(x, t) - c^2 u_{xx}(x, t) = f(x, t) + \int_0^t \{w_{tt}(x, t; s) ds - c^2 w_{xx}(x, t; s)\} ds = f(x, t).$$

Everything works and gives the unique solution in $C^2(\mathbb{R} \times [0, +\infty))$, under rather natural hypotheses on f : we require f and f_x to be continuous in $\mathbb{R} \times [0, +\infty)$.

Finally note from (5.57) that the value of u at the point (x, t) depends on the values of the forcing term f in all the triangular sector $S_{x,t}$.

5.4.5 Dissipation and dispersion

Dissipation and dispersion effects are quite important in wave propagation phenomena. Let us go back to our model for the vibrating string, with fixed end points, assuming that its weight is negligible and that there are no external loads.

- *External damping.* External factors of dissipation like friction due to the medium may be included into the model through some empirical constitutive law. We may assume, for instance, a *linear law* of friction expressing a force per unit mass proportional to the speed of vibration. Then, a force given by $-k\rho_0 u_t \Delta x \mathbf{j}$, where $k > 0$ is a damping constant, acts on the segment of string between x and $x + \Delta x$. The final equation takes the form

$$\rho_0 u_{tt} - \tau_0 u_{xx} + k\rho_0 u_t = 0. \quad (5.58)$$

The same calculations in Subsect. 5.2.2 yield

$$\dot{E}(t) = - \int_0^L k\rho_0 u_t^2 dx = -kE_{cin}(t) \leq 0 \quad (5.59)$$

which shows a rate of energy dissipation proportional to the kinetic energy.

For equation (5.58), the usual initial-boundary value problems are still well posed under reasonable assumptions on the data. In particular, the uniqueness of the solution follows from (5.59), since $E(0) = 0$ implies $E(t) = 0$ for all $t > 0$.

- *Internal damping.* The derivation of the wave equation in Subsect. 5.2.1 leads to

$$\rho_0 u_{tt} = (\tau_{vert})_x$$

where τ_{vert} is the (scalar) vertical component of the tension. The hypothesis of vibrations of small amplitude corresponds to taking

$$\tau_{vert} \simeq \tau_0 u_x, \quad (5.60)$$

where τ_0 is the (scalar) horizontal component of the tension. In other words, we assume that the vertical forces due to the tension at two end points of a string element are proportional to the relative displacement of these points. On the other hand, the string vibrations convert kinetic energy into heat, because of the friction among the string particles. The amount of heat increases with the speed of vibration while, at the same time, the vertical tension decreases. Thus, the vertical tension depends not only on the relative displacements u_x , but also on how fast these displacements change with time¹⁴. Hence, we modify (5.60) by inserting a term proportional to u_{xt} :

$$\tau_{vert} = \tau_0 u_x + \gamma u_{xt} \quad (5.61)$$

where γ is a *positive* constant. The positivity of γ follows from the fact that energy dissipation lowers the vertical tension, so that the slope u_x decreases if $u_x > 0$ and

¹⁴ In the movie *The Legend of 1900* there is a spectacular demo of this phenomenon.

increases if $u_x < 0$. Using the law (5.61), we derive the third order equation

$$\rho_0 u_{tt} - \tau_0 u_{xx} - \gamma u_{xxt} = 0. \quad (5.62)$$

In spite of the presence of the term u_{xxt} , the usual initial-boundary value problems are again well posed under reasonable assumptions on the data. In particular, uniqueness of the solution follows once again from dissipation of energy, since, in this case,

$$\dot{E}(t) = - \int_0^L \gamma u_{xt}^2 dx \leq 0.$$

- *Dispersion.* When the string is under the action of a vertical elastic restoring force proportional to u , the equation of motion becomes

$$u_{tt} - c^2 u_{xx} + \lambda u = 0 \quad (\lambda > 0) \quad (5.63)$$

known as the *linearized Klein-Gordon equation*. To emphasize the effect of the zero order term λu , let us seek for *harmonic waves solutions* of the form

$$u(x, t) = A e^{i(kx - \omega t)}.$$

Inserting u into (5.63) we find the *dispersion relation*

$$\omega^2 - c^2 k^2 = \lambda \implies \omega(k) = \pm \sqrt{c^2 k^2 + \lambda}.$$

Thus, these waves are dispersive with phase and group velocities given respectively by

$$c_p(k) = \frac{\sqrt{c^2 k^2 + \lambda}}{|k|}, \quad c_g = \frac{d\omega}{dk} = \frac{c^2 |k|}{\sqrt{c^2 k^2 + \lambda}}.$$

Observe that $c_g < c_p$.

A wave packet solution can be obtained by an integration over all possible wave numbers k :

$$u(x, t) = \int_{-\infty}^{+\infty} A(k) e^{i[kx - \omega(k)t]} dk \quad (5.64)$$

where $A(k)$ is the Fourier transform of the initial condition:

$$A(k) = \int_{-\infty}^{+\infty} u(x, 0) e^{-ikx} dx.$$

This entails that, even if the initial condition is *localized* inside a small interval, *all* the wavelengths contribute to the value of u . Although we observe a decaying in amplitude of order $t^{-1/2}$ (see formula (5.7), p. 263), these dispersive waves do not dissipate energy. For example, since the ends of the string are fixed, the total

mechanical energy is given by

$$E(t) = \frac{\rho_0}{2} \int_0^L (u_t^2 + c^2 u_x^2 + \lambda u^2) dx$$

and one may check that $\dot{E}(t) = 0$, $t > 0$.

5.5 Second Order Linear Equations

5.5.1 Classification

To derive formula (5.42), p. 277, we may use the characteristics in the following way. We change variables setting

$$\xi = x + ct, \quad \eta = x - ct \quad (5.65)$$

or

$$x = \frac{\xi + \eta}{2}, \quad t = \frac{\xi - \eta}{2c} \quad (5.66)$$

and define

$$U(\xi, \eta) = u\left(\frac{\xi + \eta}{2}, \frac{\xi - \eta}{2c}\right).$$

Then

$$U_\xi = \frac{1}{2}u_x + \frac{1}{2c}u_t$$

and, since $u_{tt} = c^2u_{xx}$, we obtain

$$U_{\xi\eta} = \frac{1}{4}u_{xx} - \frac{1}{4c}u_{xt} + \frac{1}{4c}u_{xt} - \frac{1}{4c^2}u_{tt} = 0.$$

The equation

$$U_{\xi\eta} = 0 \quad (5.67)$$

is called the *canonical* form of the wave equation; its solution is immediate:

$$U(\xi, \eta) = F(\xi) + G(\eta)$$

and going back to the original variables, (5.42) follows.

Consider now a general equation of the form:

$$au_{tt} + 2bu_{xt} + cu_{xx} + du_t + eu_x + hu = f \quad (5.68)$$

with (x, t) varying, in general, in a domain Ω . We assume that the coefficients a, b, c, d, e, h, f are smooth functions¹⁵ in Ω . The sum of second order terms

$$a(x, t)u_{tt} + 2b(x, t)u_{xt} + c(x, t)u_{xx} \quad (5.69)$$

is called **principal part** of eq. (5.68) and determines the *type* of equation according to the following classification. Consider the algebraic equation

$$H(p, q) = ap^2 + 2bpq + cq^2 = 1 \quad (a > 0) \quad (5.70)$$

in the plane p, q . If $b^2 - ac > 0$, (5.70) defines a hyperbola, if $b^2 - ac = 0$ a parabola and if $b^2 - ac < 0$ an ellipse. Accordingly, equation (5.68) is called:

- a) **Hyperbolic** when $b^2 - ac > 0$.
- b) **Parabolic** when $b^2 - ac = 0$.
- c) **Elliptic** when $b^2 - ac < 0$.

Note that the quadratic form $H(p, q)$ is, in the three cases, *indefinite, nonnegative, positive*, respectively. The above classification extends to equations in any number of variables, as we shall see later on.

It may happen that a single equation is of different type in different subdomains. For instance, the *Tricomi* equation $xu_{tt} - u_{xx} = 0$ is hyperbolic in the half plane $x > 0$, parabolic on $x = 0$ and elliptic in the half plane $x < 0$.

Basically, all the second order equations in two variables we have met so far are particular cases of (5.68). Specifically,

- The *wave* equation

$$u_{tt} - c^2u_{xx} = 0$$

is *hyperbolic*: $a(x, t) = 1$, $c(x, t) = -c^2$, and the other coefficients are zero.

- The *diffusion* equation

$$u_t - Du_{xx} = 0$$

is *parabolic*: $c(x, t) = -D$, $d(x, t) = 1$, and the other coefficients are zero.

- The *Laplace* equation (using y instead of t)

$$u_{xx} + u_{yy} = 0$$

is *elliptic*: $a(x, y) = 1$, $c(x, y) = 1$, and the other coefficients are zero.

May we reduce to a canonical form, similar to (5.67), the diffusion and the Laplace equations? Let us briefly examine why the change of variables (5.65) works for the wave equation.

Decompose the wave operator as follows

$$\partial_{tt} - c^2\partial_{xx} = (\partial_t + c\partial_x)(\partial_t - c\partial_x). \quad (5.71)$$

¹⁵ E.g. C^2 functions.

If we introduce the vectors $\mathbf{v} = (c, 1)$ and $\mathbf{w} = (-c, 1)$, then (5.71) can be written in the form

$$\partial_{tt} - c^2 \partial_{xx} = \partial_{\mathbf{v}} \partial_{\mathbf{w}}.$$

On the other hand, the characteristics

$$x + ct = \text{constant}, \quad x - ct = \text{constant}$$

of the two first order equations

$$\phi_t - c\phi_x = 0 \quad \text{and} \quad \psi_t + c\psi_x = 0,$$

corresponding to the two factors in (5.71), are straight lines in the direction of \mathbf{w} and \mathbf{v} , respectively. The change of variables

$$\xi = \phi(x, t) = x + ct \quad \eta = \psi(x, t) = x - ct$$

maps the straight lines $x + ct = 0$ and $x - ct = 0$ into $\xi = 0$ and $\eta = 0$, respectively. Moreover, recalling (5.66), we have

$$\partial_{\xi} = \frac{1}{2c} (\partial_t + c\partial_x) = \frac{1}{2c} \partial_{\mathbf{v}}, \quad \partial_{\eta} = \frac{1}{2c} (\partial_t - c\partial_x) = \frac{1}{2c} \partial_{\mathbf{w}}.$$

Thus, the wave operator is converted into a multiple of its canonical form:

$$\partial_{tt} - c^2 \partial_{xx} = \partial_{\mathbf{v}} \partial_{\mathbf{w}} = 4c^2 \partial_{\xi\eta}.$$

Once the characteristics are known, the change of variables (5.65) reduces the wave equation to the form (5.67).

Proceeding in the same way, for the diffusion operator we would have

$$\partial_{xx} = \partial_x \partial_x.$$

Therefore we find only one family of characteristics, given by

$$t = \text{constant}.$$

Thus, no change of variables is necessary and the diffusion equation is already in its canonical form.

For the Laplace operator we find

$$\partial_{xx} + \partial_{yy} = (\partial_y + i\partial_x)(\partial_y - i\partial_x)$$

and there are two families of *complex* characteristics given by

$$\phi(x, y) = x + iy = \text{constant}, \quad \psi(x, y) = x - iy = \text{constant}.$$

The change of variables

$$z = x + iy, \quad \bar{z} = x - iy$$

leads to the equation

$$\partial_{z\bar{z}}U = 0$$

whose general solution is

$$U(z, \bar{z}) = F(z) + G(\bar{z}).$$

This formula may be considered as a characterization of the harmonic functions in the complex plane.

It should be clear, however, that the characteristics for the diffusion and the Laplace equations do not play the same relevant role as they do for the wave equation.

5.5.2 Characteristics and canonical form

Let us go back to the equation in general form (5.68). Can we reduce its principal part to a canonical form? There are at least two substantial reasons to examine the question.

The first one is tied to the type of well posed problems associated with (5.68): which kind of data have to be assigned and where, in order to find a unique and stable solution? It turns out that hyperbolic, parabolic and elliptic equations share their well posed problems with their main prototypes: the wave, diffusion and Laplace equations, respectively. Also the choice of numerical methods depends very much on the type of problem to be solved.

The second reason comes from the different features the three types of equation exhibit. Hyperbolic equations model oscillatory phenomena with *finite speed of propagation of the disturbances*, while for parabolic equation, “information” travels with infinite speed. Finally, elliptic equations model stationary situations, with no evolution in time.

To obtain the canonical form of the principal part we try to apply the ideas at the end of the previous subsection. First of all, note that, if $a = c = 0$, the principal part is already in the form (5.67), so that we assume $a > 0$ (say). Now we decompose the differential operator in (5.69) into the product of two first order factors, as follows¹⁶:

$$a\partial_{tt} + 2b\partial_{xt} + c\partial_{xx} = a(\partial_t - \Lambda^+\partial_x)(\partial_t - \Lambda^-\partial_x) \quad (5.72)$$

¹⁶ Remember that

$$ax^2 + 2bxy + cy^2 = a(x - x_1)(x - x_2)$$

where

$$x_{1,2} = \left[-b \pm \sqrt{b^2 - ac} \right] / a.$$

where

$$\Lambda^{\pm} = \frac{-b \pm \sqrt{b^2 - ac}}{a}.$$

Case 1: $b^2 - ac > 0$, the equation is **hyperbolic**. The two factors in (5.72) represent derivatives along the direction fields

$$\mathbf{v}(x, t) = (-\Lambda^+(x, t), 1) \quad \text{and} \quad \mathbf{w}(x, t) = (-\Lambda^-(x, t), 1)$$

respectively, so that we may write

$$a\partial_{tt} + 2b\partial_{xt} + c\partial_{xx} = a\partial_{\mathbf{v}}\partial_{\mathbf{w}}.$$

The vector fields \mathbf{v} and \mathbf{w} are tangent at any point to the characteristics

$$\phi(x, t) = k_1 \quad \text{and} \quad \psi(x, t) = k_2 \quad (k_1, k_2 \in \mathbb{R}) \quad (5.73)$$

of the following *first-order* equations

$$\phi_t - \Lambda^+ \phi_x = 0 \quad \text{and} \quad \psi_t - \Lambda^- \psi_x = 0. \quad (5.74)$$

Note that we may write the two equations (5.74) in the compact form

$$av_t^2 + 2bv_x v_t + cv_x^2 = 0. \quad (5.75)$$

By analogy with the case of the wave equation, we expect that the change of variables

$$\xi = \phi(x, t), \quad \eta = \psi(x, t) \quad (5.76)$$

should straighten the characteristics, at least locally, converting $\partial_{\mathbf{v}}\partial_{\mathbf{w}}$ into a multiple of $\partial_{\xi\eta}$.

First of all, however, we have to make sure that the transformation (5.76) is *non-degenerate*, at least locally, or, in other words, that the Jacobian of the transformation does not vanish:

$$\phi_t \psi_x - \phi_x \psi_t \neq 0. \quad (5.77)$$

On the other hand, this follows from the fact that the vectors $\nabla\phi$ and $\nabla\psi$ are orthogonal to \mathbf{v} and \mathbf{w} , respectively, and that \mathbf{v} , \mathbf{w} are nowhere colinear (since $b^2 - ac > 0$).

Thus, at least locally, the inverse transformation

$$x = \Phi(\xi, \eta), \quad t = \Psi(\xi, \eta)$$

exists. Let

$$U(\xi, \eta) = u(\Phi(\xi, \eta), \Psi(\xi, \eta)) \quad \text{or} \quad u(x, t) = U(\phi(x, t), \psi(x, t)).$$

Then

$$u_x = U_\xi \phi_x + U_\eta \psi_x, \quad u_t = U_\xi \phi_t + U_\eta \psi_t$$

and moreover:

$$\begin{aligned} u_{tt} &= \phi_t^2 U_{\xi\xi} + 2\phi_t\psi_t U_{\xi\eta} + \psi_t^2 U_{\eta\eta} + \phi_{tt} U_\xi + \psi_{tt} U_\eta, \\ u_{xx} &= \phi_x^2 U_{\xi\xi} + 2\phi_x\psi_x U_{\xi\eta} + \psi_x^2 U_{\eta\eta} + \phi_{xx} U_\xi + \psi_{xx} U_\eta, \\ u_{xt} &= \phi_t\phi_x U_{\xi\xi} + (\phi_x\psi_t + \phi_t\psi_x) U_{\xi\eta} + \psi_t\psi_x U_{\eta\eta} + \phi_{xt} U_\xi + \psi_{xt} U_\eta. \end{aligned}$$

Then

$$au_{tt} + 2bu_{xy} + cu_{xx} = AU_{\xi\xi} + 2BU_{\xi\eta} + CU_{\eta\eta} + DU_\xi + EU_\eta,$$

where¹⁷

$$\begin{aligned} A &= a\phi_t^2 + 2b\phi_t\phi_x + c\phi_x^2, & C &= a\psi_t^2 + 2b\psi_t\psi_x + c\psi_x^2, \\ B &= a\phi_t\psi_t + b(\phi_x\psi_t + \phi_t\psi_x) + c\phi_x\psi_x, \\ D &= a\phi_{tt} + 2b\phi_{xt} + c\phi_{xx}, & E &= a\psi_{tt} + 2b\psi_{xt} + c\psi_{xx}. \end{aligned}$$

Now, we have

$$A = C = 0$$

since ϕ and ψ both satisfy (5.75), so that

$$au_{tt} + 2bu_{xy} + cu_{xx} = 2BU_{\xi\eta} + DU_\xi + EU_\eta.$$

We claim that $B \neq 0$; indeed, recalling that

$$\Lambda^+ \Lambda^- = c/a, \quad \Lambda^+ + \Lambda^- = -2b/a$$

and, from (5.74),

$$\phi_t = \Lambda^+ \phi_x, \quad \psi_t = \Lambda^- \psi_x,$$

after elementary computations we find

$$B = \frac{2}{a} (ac - b^2) \phi_x \psi_x.$$

From (5.77) we deduce that $B \neq 0$. Thus, (5.68) assumes the form

$$U_{\xi\eta} = \mathcal{F}(\xi, \eta, U, U_\xi, U_\eta)$$

which is its *canonical form*.

The curves (5.73) are called *characteristics* for (5.68) and are the solution curves of the ordinary differential equations

$$\frac{dx}{dt} = -\Lambda^+, \quad \frac{dx}{dt} = -\Lambda^-, \quad (5.78)$$

¹⁷ It is understood that all the functions are evaluated at $x = \Phi(\xi, \eta)$ and $t = \Psi(\xi, \eta)$.

respectively. Note that the two equations (5.78) can be put into the compact form

$$a \left(\frac{dx}{dt} \right)^2 - 2b \frac{dx}{dt} + c = 0. \quad (5.79)$$

Example 5.8. Consider the equation

$$xu_{tt} - (1 + x^2) u_{xt} = 0. \quad (5.80)$$

Since $b^2 - ac = (1 + x^2)/4 > 0$, (5.80) is hyperbolic in \mathbb{R}^2 . Equation (5.79) is

$$x \left(\frac{dx}{dt} \right)^2 + (1 + x^2) \frac{dx}{dt} = 0$$

which yields, for $x \neq 0$,

$$\frac{dx}{dt} = -\frac{1+x^2}{x} \quad \text{and} \quad \frac{dx}{dt} = 0.$$

Thus, the characteristics curves are:

$$\phi(x, t) = e^{2t}(1 + x^2) = k_1 \quad \text{and} \quad \psi(x, t) = x = k_2.$$

We set

$$\xi = e^{2t}(1 + x^2) \quad \text{and} \quad \eta = x.$$

After routine calculations, we find $D = E = 0$, so that the canonical form is

$$U_{\xi\eta} = 0.$$

The general solution of (5.80) is therefore

$$u(x, t) = F(e^{2t}(1 + x^2)) + G(x)$$

with F and G arbitrary C^2 functions.

Case 2: $b^2 - ac \equiv 0$, the equation is **parabolic**. There exists **only one** family of characteristics, given by $\phi(x, t) = k$, where ϕ is a solution of the first order equation

$$a\phi_t + b\phi_x = 0,$$

since $\Lambda^+ = \Lambda^- = -b/a$. If ϕ is known, choose any smooth function ψ such that $\nabla\phi$ and $\nabla\psi$ are linearly independent and

$$a\psi_t^2 + 2b\psi_t\psi_x + c\psi_x^2 = C \neq 0.$$

Set

$$\xi = \phi(x, t), \quad \eta = \psi(x, t)$$

and

$$U(\xi, \eta) = u(\Phi(\xi, \eta), \Psi(\xi, \eta)).$$

For the derivatives of U we can use the computations done in **case 1**. However, observe that, since $b^2 - ac = 0$ and $a\phi_t + b\phi_x = 0$, we have

$$\begin{aligned} B &= a\phi_t\psi_t + b(\phi_t\psi_x + \phi_x\psi_t) + c\phi_x\psi_x = \psi_t(a\phi_t + b\phi_x) + \psi_x(b\phi_t + c\phi_x) \\ &= b\psi_x \left(\phi_t + \frac{c}{b}\phi_x \right) = b\psi_x \left(\phi_t + \frac{b}{a}\phi_x \right) = \frac{b}{a}\psi_x(a\phi_t + b\phi_x) = 0. \end{aligned}$$

Thus, the equation for U becomes

$$CU_{\eta\eta} = \mathcal{F}(\xi, \eta, U, U_\xi, U_\eta)$$

which is the *canonical form*.

Example 5.9. The equation

$$u_{tt} - 6u_{xt} + 9u_{xx} = u$$

is parabolic. The family of characteristics is

$$\phi(x, t) = 3t + x = k.$$

Choose $\psi(x, t) = x$ and set

$$\xi = 3t + x, \quad \eta = x.$$

Since $\nabla\phi = (3, 1)$ and $\nabla\psi = (1, 0)$, the gradients are independent and we set

$$U(\xi, \eta) = u\left(\frac{\xi - \eta}{3}, \eta\right).$$

We have, $D = E = 0$, so that the equation for U is

$$U_{\eta\eta} - U = 0$$

whose general solution is

$$U(\xi, \eta) = F(\xi)e^{-\eta} + G(\xi)e^\eta$$

with F and G arbitrary C^2 functions. Finally, we find

$$u(x, t) = F(3t + x)e^{-x} + G(3t + x)e^x.$$

Case 3: $b^2 - ac < 0$, the equation is **elliptic**. In this case there are no real characteristics. If the coefficients a, b, c are analytic functions¹⁸ we can proceed as in

¹⁸ It means that they can be locally expanded in Taylor series.

Case 1, with two families of complex characteristics. This yields the canonical form

$$U_{zw} = \mathcal{G}(z, w, U, U_z, U_w) \quad z, w \in \mathbb{C}.$$

Letting

$$z = \xi + i\eta, \quad w = \xi - i\eta$$

and $\tilde{U}(\xi, \eta) = U(\xi + i\eta, \xi - i\eta)$ we can eliminate the complex variables arriving at the real canonical form

$$\tilde{U}_{\xi\xi} + \tilde{U}_{\eta\eta} = \tilde{\mathcal{G}}(\xi, \eta, \tilde{U}, \tilde{U}_\xi, \tilde{U}_\eta).$$

5.6 The Multi-dimensional Wave Equation ($n > 1$)

5.6.1 Special solutions

The wave equation

$$u_{tt} - c^2 \Delta u = f \quad (5.81)$$

constitutes a basic model for describing a remarkable number of oscillatory phenomena in dimension $n > 1$. Here $u = u(\mathbf{x}, t)$, $\mathbf{x} \in \mathbb{R}^n$ and, as in the one-dimensional case, c is the *speed of propagation*. If $f \equiv 0$, the equation is said *homogeneous* and the *superposition principle holds*. Let us examine some relevant solutions of (5.81).

- *Plane waves.* If $\mathbf{k} \in \mathbb{R}^n$ and $\omega^2 = c^2 |\mathbf{k}|^2$, the function

$$u(\mathbf{x}, t) = w(\mathbf{x} \cdot \mathbf{k} - \omega t)$$

is a solution of the homogeneous (5.81). Indeed,

$$u_{tt}(\mathbf{x}, t) - c^2 \Delta u(\mathbf{x}, t) = \omega^2 w''(\mathbf{x} \cdot \mathbf{k} - \omega t) - c^2 |\mathbf{k}|^2 w''(\mathbf{x} \cdot \mathbf{k} - \omega t) = 0.$$

We have already seen in Subsect. 5.1.1 that the planes

$$\mathbf{x} \cdot \mathbf{k} - \omega t = \text{constant}$$

constitute the wave fronts, moving at speed $c_p = \omega / |\mathbf{k}|$ in the \mathbf{k} direction. The scalar $\lambda = 2\pi / |\mathbf{k}|$ is the wavelength. If $w(z) = Ae^{iz}$, the wave is said *monochromatic* or *harmonic*.

- *Cylindrical waves* ($n = 3$) are of the form

$$u(\mathbf{x}, t) = w(r, t)$$

where $\mathbf{x} = (x_1, x_2, x_3)$, $r = \sqrt{x_1^2 + x_2^2}$. In particular, solutions of the form $u(\mathbf{x}, t) = e^{i\omega t} w(r)$ represent stationary cylindrical waves, that can be found by solving the

homogeneous equation (5.81), using the separation of variables method, in axially symmetric domains.

If the axis of symmetry is the x_3 axis, it is appropriate to use the cylindrical coordinates $x_1 = r \cos \theta$, $x_2 = r \sin \theta$, x_3 . Then, the wave equation becomes¹⁹

$$u_{tt} - c^2 \left(u_{rr} + \frac{1}{r} u_r + \frac{1}{r^2} u_{\theta\theta} + u_{x_3 x_3} \right) = 0.$$

Looking for standing waves of the form $u(r, t) = e^{i\lambda ct} w(r)$, $\lambda \geq 0$, we find, after dividing by $c^2 e^{i\lambda ct}$,

$$w''(r) + \frac{1}{r} w' + \lambda^2 w = 0.$$

This is a Bessel equation of zero order. We know from Sect. 2.7 that the only solutions bounded at $r = 0$ are

$$w(r) = a J_0(\lambda r), \quad a \in \mathbb{R},$$

where, we recall,

$$J_0(x) = \sum_{k=0}^{\infty} \frac{(-1)^k}{(k!)^2} \left(\frac{x}{2}\right)^{2k}$$

is the Bessel function of first kind of zero order. In this way we obtain waves of the form

$$u(r, t) = a J_0(\lambda r) e^{i\lambda ct}.$$

- *Spherical waves* ($n = 3$) are of the form

$$u(\mathbf{x}, t) = w(r, t)$$

where $\mathbf{x} = (x_1, x_2, x_3)$, $r = |\mathbf{x}| = \sqrt{x_1^2 + x_2^2 + x_3^2}$. In particular

$$u(\mathbf{x}, t) = e^{i\omega t} w(r)$$

represents a standing spherical wave and can be determined by solving the homogeneous equation (5.81) via the method of separation of variables in spherically symmetric domains. In this case, spherical coordinates

$$x_1 = r \cos \theta \sin \psi, \quad x_2 = r \sin \theta \sin \psi, \quad x_3 = r \cos \psi,$$

are appropriate and the wave equation becomes²⁰

$$\frac{1}{c^2} u_{tt} - u_{rr} - \frac{2}{r} u_r - \frac{1}{r^2} \left\{ \frac{1}{(\sin \psi)^2} u_{\theta\theta} + u_{\psi\psi} + \frac{\cos \psi}{\sin \psi} u_\psi \right\} = 0. \quad (5.82)$$

¹⁹ See Appendix C.

²⁰ See Appendix C.

Let us look for solution of the form

$$u(r, t) = e^{i\lambda ct} w(r), \quad \lambda \geq 0.$$

We find, after simplifying out $c^2 e^{i\lambda ct}$,

$$w''(r) + \frac{2}{r} w' + \lambda^2 w = 0$$

which can be written²¹

$$(rw)'' + \lambda^2 rw = 0.$$

Thus, $v = rw$ is solution of

$$v'' + \lambda^2 v = 0$$

which gives $v(r) = a \cos(\lambda r) + b \sin(\lambda r)$ and hence the attenuated spherical waves

$$w(r, t) = ae^{i\lambda ct} \frac{\cos(\lambda r)}{r}, \quad w(r, t) = be^{i\lambda ct} \frac{\sin(\lambda r)}{r}. \quad (5.83)$$

Let us now determine the general form of a spherical wave in \mathbb{R}^3 . Inserting $u(\mathbf{x}, t) = w(r, t)$ into (5.82) we obtain

$$w_{tt} - c^2 \left\{ w_{rr}(r) + \frac{2}{r} w_r \right\} = 0$$

which can be written in the form

$$(rw)_{tt} - c^2 (rw)_{rr} = 0. \quad (5.84)$$

Then, formula (5.42) gives

$$w(r, t) = \frac{F(r + ct)}{r} + \frac{G(r - ct)}{r} \equiv w_i(r, t) + w_o(r, t), \quad (5.85)$$

which represents the superposition of two attenuated progressive spherical waves. The wave fronts of w_o are the spheres $r - ct = k$, expanding as time goes on. Hence, w_o represents an *outgoing wave*. On the contrary, the wave w_i is *incoming*, since its wave fronts are the contracting spheres $r + ct = k$.

5.6.2 Well posed problems. Uniqueness

The well posed problems in dimension one, are still well posed in any number of dimensions. Let

$$Q_T = \Omega \times (0, T)$$

be a *space-time cylinder*, where Ω is a bounded C^1 or Lipschitz domain in \mathbb{R}^n . A solution $u(\mathbf{x}, t)$ is uniquely determined by assigning initial data and appropriate boundary conditions on the boundary $\partial\Omega$ of Ω .

²¹ Thanks to the providential presence of the factor 2 in the coefficient of w' .

More specifically, we may pose the following problems: *Determine $u = u(\mathbf{x}, t)$ such that:*

$$\begin{cases} u_{tt} - c^2 \Delta u = f & \text{in } Q_T \\ u(\mathbf{x}, 0) = g(\mathbf{x}), u_t(\mathbf{x}, 0) = h(\mathbf{x}) & \text{in } \Omega \\ + \text{boundary conditions} & \text{on } \partial\Omega \times [0, T] \end{cases} \quad (5.86)$$

where the boundary conditions can be:

- (a) $u = h$ (Dirichlet).
- (b) $\partial_\nu u = h$ (Neumann).
- (c) $\partial_\nu u + \alpha u = h$, $\alpha = \alpha(\sigma) > 0$ (Robin).
- (d) $u = h_1$ on $\Gamma_D \times [0, T]$ and $\partial_\nu u = h_2$ on $\Gamma_N \times [0, T]$ (mixed problem) with Γ_N a relatively open subset of $\partial\Omega$ and $\Gamma_D = \partial\Omega \setminus \Gamma_N$.

The *global Cauchy problem*

$$\begin{cases} u_{tt} - c^2 \Delta u = f & \mathbf{x} \in \mathbb{R}^n, t > 0 \\ u(\mathbf{x}, 0) = g(\mathbf{x}), u_t(\mathbf{x}, 0) = h(\mathbf{x}) & \mathbf{x} \in \mathbb{R}^n, \end{cases} \quad (5.87)$$

is quite important also in dimension $n > 1$. We will examine it with some details later on. Particularly relevant are the different features that the solutions exhibit for $n = 2$ and $n = 3$.

Under rather natural hypotheses on the data, problem (5.86) has at most one solution. To see it, we may use once again the conservation of energy, which is proportional to:

$$E(t) = \frac{1}{2} \int_{\Omega} \left\{ u_t^2 + c^2 |\nabla u|^2 \right\} d\mathbf{x}.$$

The growth rate is:

$$\dot{E}(t) = \int_{\Omega} \{ u_t u_{tt} + c^2 \nabla u_t \cdot \nabla u \} d\mathbf{x}.$$

Integrating by parts, we have

$$\int_{\Omega} c^2 \nabla u_t \cdot \nabla u \, d\mathbf{x} = c^2 \int_{\partial\Omega} u_\nu u_t \, d\sigma - \int_{\Omega} c^2 u_t \Delta u \, d\mathbf{x}$$

whence, since $u_{tt} - c^2 \Delta u = f$,

$$\dot{E}(t) = \int_{\Omega} \{ u_{tt} - c^2 \Delta u \} u_t \, d\mathbf{x} + c^2 \int_{\partial\Omega} u_\nu u_t \, d\sigma = \int_{\Omega} f u_t \, d\mathbf{x} + c^2 \int_{\partial\Omega} u_\nu u_t \, d\sigma. \quad (5.88)$$

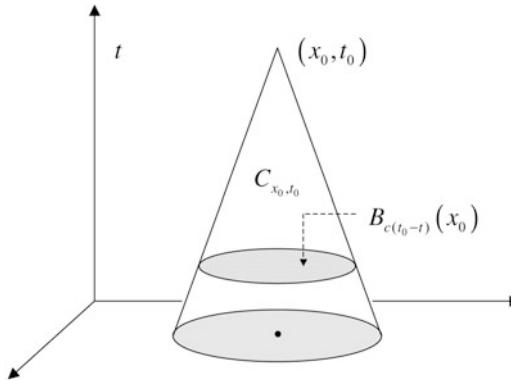


Fig. 5.9 Retrograde cone

It is now easy to prove the following result, where we use the symbol $C^{h,k}(D)$ to denote the set of functions h times continuously differentiable with respect to space and k times with respect to time in D .

Theorem 5.10. *Problem (5.86), coupled with one of the boundary conditions (a)–(d) above, has at most one solution in $C^{2,2}(Q_T) \cap C^{1,1}(\bar{Q}_T)$.*

Proof. Let u_1 and u_2 be solutions of the same problem, sharing the same data. Their difference $w = u_1 - u_2$ is a solution of the homogeneous equation, with zero initial/boundary data. We show that $w(\mathbf{x}, t) \equiv 0$.

In the case of Dirichlet, Neumann and mixed conditions, since either $w_\nu = 0$ or $w_t = 0$ on $\partial\Omega \times [0, T]$, we have $\dot{E}(t) = 0$. Thus, since $E(0) = 0$, we infer:

$$E(t) = \frac{1}{2} \int_{\Omega} \{w_t^2 + c^2 |\nabla w|^2\} d\mathbf{x} = 0, \quad \forall t > 0.$$

Therefore, for each $t > 0$, both w_t and $|\nabla w(\mathbf{x}, t)|$ vanish and hence $w(\mathbf{x}, t)$ is constant. Then $w(\mathbf{x}, t) \equiv 0$, since $w(\mathbf{x}, 0) = 0$. For the Robin problem, since α is time independent, we have, from (5.88),

$$\dot{E}(t) = -c^2 \int_{\partial\Omega} \alpha w w_t d\sigma = -\frac{c^2}{2} \frac{d}{dt} \int_{\partial\Omega} \alpha w^2 d\sigma$$

that is

$$\frac{d}{dt} \left\{ E(t) + \frac{c^2}{2} \int_{\partial\Omega} \alpha w^2 d\sigma \right\} = 0.$$

Hence,

$$E(t) + \frac{c^2}{2} \int_{\partial\Omega} \alpha w^2 d\sigma = k, \text{ constant.}$$

Since $E(0) = 0$ and $w(\mathbf{x}, 0) = 0$ it follows that $k = 0$. Since $\alpha \geq 0$, we easily conclude that $w \equiv 0$. \square

Uniqueness for the global Cauchy problem follows from another energy inequality, with more interesting consequences.

First a remark. For better clarity, let us consider the case $n = 2$. Suppose that a disturbance governed by the homogeneous wave equation ($f = 0$) is felt at \mathbf{x}_0 at time t_0 . Since the disturbances travel with speed c , $u(\mathbf{x}_0, t_0)$ is only affected by the values of the initial data in the circle $B_{ct_0}(\mathbf{x}_0)$. More generally, at time $t_0 - t$, $u(\mathbf{x}_0, t_0)$ is determined by the values of u in the circle $B_{c(t_0-t)}(\mathbf{x}_0)$. As t varies from 0 to t_0 , the union of the circles $B_{c(t_0-t)}(\mathbf{x}_0)$ in the \mathbf{x}, t space coincides with the so called *backward or retrograde cone with vertex at (\mathbf{x}_0, t_0)* and opening $\theta = \tan^{-1} c$, given by (see Fig. 5.9):

$$C_{\mathbf{x}_0, t_0} = \{(\mathbf{x}, t) : |\mathbf{x} - \mathbf{x}_0| \leq c(t_0 - t), 0 \leq t \leq t_0\}.$$

Thus, given a point \mathbf{x}_0 , it is natural to introduce an energy associated with its backward cone by the formula

$$e(t) = \frac{1}{2} \int_{B_{c(t_0-t)}(\mathbf{x}_0)} (u_t^2 + c^2 |\nabla u|^2) d\mathbf{x}.$$

It turns out that $e(t)$ is a decreasing function. Namely, going back to dimension n :

Lemma 5.11. *Let u be a C^2 -solution of the homogeneous wave equation in $\mathbb{R}^n \times [0, +\infty)$. Then*

$$\dot{e}(t) \leq 0.$$

Proof. We may write

$$e(t) = \frac{1}{2} \int_0^{c(t_0-t)} dr \int_{\partial B_r(\mathbf{x}_0)} (u_t^2 + c^2 |\nabla u|^2) d\sigma$$

so that

$$\dot{e}(t) = -\frac{c}{2} \int_{\partial B_{c(t_0-t)}(\mathbf{x}_0)} (u_t^2 + c^2 |\nabla u|^2) d\sigma + \int_{B_{c(t_0-t)}(\mathbf{x}_0)} (u_t u_{tt} + c^2 \nabla u \cdot \nabla u_t) d\mathbf{x}.$$

An integration by parts yields

$$\int_{B_{c(t_0-t)}(\mathbf{x}_0)} \nabla u \cdot \nabla u_t d\mathbf{x} = \int_{\partial B_{c(t_0-t)}(\mathbf{x}_0)} u_t u_\nu d\sigma - \int_{B_{c(t_0-t)}(\mathbf{x}_0)} u_t \Delta u d\mathbf{x}$$

whence

$$\begin{aligned} \dot{e}(t) &= \int_{B_{c(t_0-t)}(\mathbf{x}_0)} u_t (u_{tt} - c^2 \Delta u) d\mathbf{x} + \frac{c}{2} \int_{\partial B_{c(t_0-t)}(\mathbf{x}_0)} (2cu_t u_\nu - u_t^2 - c^2 |\nabla u|^2) d\sigma \\ &= \frac{c}{2} \int_{\partial B_{c(t_0-t)}(\mathbf{x}_0)} (2cu_t u_\nu - u_t^2 - c^2 |\nabla u|^2) d\sigma. \end{aligned}$$

Now, we have

$$|u_t u_\nu| \leq |u_t| |\nabla u|,$$

so that

$$2cu_t u_\nu - u_t^2 - c^2 |\nabla u|^2 \leq 2c |u_t| |\nabla u| - u_t^2 - c^2 |\nabla u|^2 = -(u_t - c |\nabla u|)^2 \leq 0$$

and therefore $\dot{e}(t) \leq 0$. \square

Two almost immediate consequences are stated in the following theorem:

Theorem 5.12. *Let $u \in C^2(\mathbb{R}^n \times [0, +\infty))$ be a solution of the Cauchy problem (5.87). If $g \equiv h \equiv 0$ in $B_{ct_0}(\mathbf{x}_0)$ and $f \equiv 0$ in $C_{\mathbf{x}_0, t_0}$ then $u \equiv 0$ in $C_{\mathbf{x}_0, t_0}$. Consequently, problem (5.87) has at most one solution in $C^2(\mathbb{R}^n \times [0, +\infty))$.*

5.7 Two Classical Models

5.7.1 Small vibrations of an elastic membrane

In Subsect. 5.2.3 we have derived a model for the small transversal vibrations of a string. Similarly, we may derive the equation governing the small transversal vibrations of a highly stretched membrane (think e.g. of a drum), at rest in the horizontal position. We briefly sketch the derivation leaving it to the reader to fill in the details. Assume the following hypotheses.

1. *The vibrations of the membrane are small and vertical.* This means that the changes from the plane horizontal shape are very small and horizontal displacements are negligible.
2. *The vertical displacement of a point of the membrane depends on time and on its position at rest.* Thus, if u denotes the vertical displacement of a point located at rest at (x, y) , we have

$$u = u(x, y, t).$$

3. *The membrane is perfectly flexible and elastic.* In other words, there is no resistance to bending. In particular, the stress in the membrane can be modelled by a tangential force \mathbf{T} of magnitude τ , called *tension*²². Perfect elasticity means that τ is a constant.
4. *Friction is negligible.*

Under the above assumptions, the equation of motion of the membrane can be derived from *conservation of mass* and *Newton's law*.

Let $\rho_0 = \rho_0(x, y)$ be the surface mass density of the membrane at rest and consider a small “rectangular” piece of membrane, with vertices at the points A, B, C, D of coordinates $(x, y), (x + \Delta x, y), (x, y + \Delta y)$ and $(x + \Delta x, y + \Delta y)$, respectively.

²² The tension \mathbf{T} has the following meaning. Consider a small region on the membrane, delimited by a closed curve γ . The material on one side of γ exerts on the material on the other side a *force per unit length* \mathbf{T} (*pulling*) along γ . A constitutive law for \mathbf{T} is

$$\mathbf{T}(x, y, t) = \tau(x, y, t) \mathbf{N}(x, y, t), \quad (x, y) \in \gamma,$$

where \mathbf{N} is the outward unit normal vector to γ , tangent to the membrane.

Again, the tangentiality of the tension force is due to the absence of distributed moments over the membrane.

Denote by ΔS the corresponding area at time t . Then, conservation of mass yields

$$\rho(x, y, t) \Delta S = \rho_0(x, y) \Delta x \Delta y. \quad (5.89)$$

To write Newton's law of motion we have to determine the forces acting on our small piece of membrane. Since the motion is vertical, the horizontal forces have to balance.

The vertical forces are given by body forces (e.g. gravity and external loads) and the vertical component of the tension.

Denote by $f(x, y, t) \mathbf{k}$ the resultant of the body forces per unit mass. Then, using (5.89), the body forces acting on the membrane element are given by:

$$\rho(x, y, t) f(x, y, t) \Delta S \mathbf{k} = \rho_0(x, y) f(x, y, t) \Delta x \Delta y \mathbf{k}.$$

Along the edges AB and CD , the tension is perpendicular to the x -axis and almost parallel to the y -axis. Its (scalar) vertical components are respectively given by

$$\tau_{vert}(x, y, t) \simeq \tau u_y(x, y, t) \Delta x, \quad \tau_{vert}(x, y + \Delta y, t) \simeq \tau u_y(x, y + \Delta y, t) \Delta x.$$

Similarly, along the edges AC and BD , the tension is perpendicular to the y -axis and almost parallel to the x -axis. Its (scalar) vertical components are respectively given by

$$\tau_{vert}(x, y, t) \simeq \tau u_x(x, y, t) \Delta y, \quad \tau_{vert}(x + \Delta x, y, t) \simeq \tau u_x(x + \Delta x, y, t) \Delta y.$$

Thus, using again (5.89) and observing that u_{tt} is the (scalar) vertical acceleration, Newton's law gives:

$$\begin{aligned} & \rho_0(x, y) \Delta x \Delta y u_{tt} = \\ & = \tau[u_y(x, y + \Delta y, t) - u_y(x, y, t)] \Delta x + \tau[u_x(x + \Delta x, y, t) - u_x(x, y, t)] \Delta y + \\ & + \rho_0(x, y) f(x, y, t) \Delta x \Delta y. \end{aligned}$$

Dividing for $\Delta x \Delta y$ and letting $\Delta x, \Delta y \rightarrow 0$, we obtain the equation

$$u_{tt} - c^2(u_{yy} + u_{xx}) = f(x, y, t) \quad (5.90)$$

where $c^2(x, y, t) = \tau/\rho_0(x, y)$.

- *Square membrane.* Consider a membrane occupying at rest a square of side a , pinned at the boundary. We want to study its vibrations when the membrane initially takes a horizontal position, with speed $h = h(x, y)$. If there is no external load and the weight of the membrane is negligible, the vibrations are governed by

the following initial-boundary value problem:

$$\begin{cases} u_{tt} - c^2 \Delta u = 0 & 0 < x < a, 0 < y < a, t > 0 \\ u(x, y, 0) = 0, u_t(x, y, 0) = h(x, y) & 0 < x < a, 0 < y < a \\ u(0, y, t) = u(a, y, t) = 0 & 0 \leq y \leq a, t \geq 0 \\ u(x, 0, t) = u(x, a, t) = 0 & 0 \leq x \leq a, t \geq 0. \end{cases}$$

The square shape of the membrane and the homogeneous boundary conditions suggest the use of the method of separation of variables. Let us seek for a solution under the form

$$u(x, y, t) = v(x, y) q(t)$$

with $v = 0$ at the boundary. Substituting into the wave equation, we find

$$q''(t) v(x, y) - c^2 q(t) \Delta v(x, y) = 0$$

and, separating the variables²³,

$$\frac{q''(t)}{c^2 q(t)} = \frac{\Delta v(x, y)}{v(x, y)} = -\lambda^2,$$

whence the equation

$$q''(t) + c^2 \lambda^2 q(t) = 0 \quad (5.91)$$

and the *eigenvalue problem*

$$\Delta v + \lambda^2 v = 0, \quad (5.92)$$

$$v(0, y) = v(a, y) = v(x, 0) = v(x, a) = 0, \quad 0 \leq x, y \leq a.$$

We first solve the eigenvalue problem, using once more time the method of separation of variables and setting $v(x, y) = X(x) Y(y)$, with the conditions

$$X(0) = X(a) = 0, \quad Y(0) = Y(a) = 0.$$

Substituting into (5.92), we obtain

$$\frac{Y''(y)}{Y(y)} + \lambda^2 = -\frac{X''(x)}{X(x)} = \mu^2$$

where μ is a new constant.

Letting $\nu^2 = \lambda^2 - \mu^2$, we have to solve the following two one-dimensional eigenvalue problems, in $0 < x < a$ and $0 < y < a$, respectively:

$$\begin{cases} X''(x) + \mu^2 X(x) = 0 \\ X(0) = X(a) = 0 \end{cases} \quad \begin{cases} Y''(y) + \nu^2 Y(y) = 0 \\ Y(0) = Y(a) = 0. \end{cases}$$

²³ The two ratios must be equal to the same constant. The choice of $-\lambda^2$ is guided by our former experience.

The solutions are:

$$\begin{aligned} X(x) &= A \sin(\mu_m x), & \mu_m &= \frac{m\pi}{a} \\ Y(y) &= B \sin(\nu_n y), & \nu_n &= \frac{n\pi}{a} \end{aligned}$$

where A, B are arbitrary constants and $m, n = 1, 2, \dots$. Since $\lambda^2 = \nu^2 + \mu^2$, we have

$$\lambda_{mn}^2 = \frac{\pi^2}{a^2} (m^2 + n^2), \quad m, n = 1, 2, \dots \quad (5.93)$$

corresponding to the eigenfunctions

$$v_{mn}(x, y) = \sin(\mu_m x) \sin(\nu_n y).$$

For $\lambda = \lambda_{mn}$, the general integral of (5.91) is

$$q_{mn}(t) = a \cos(c\lambda_{mn} t) + b \sin(c\lambda_{mn} t) \quad (a, b \in \mathbb{R}).$$

Thus we have found infinitely many special solutions to the wave equations, of the form

$$u_{mn} = [a \cos(c\lambda_{mn} t) + b \sin(c\lambda_{mn} t)] \sin(\mu_m x) \sin(\nu_n y),$$

which, moreover, vanish on the boundary.

Every u_{mn} is a standing wave and corresponds to a particular mode of vibration of the membrane. The *fundamental frequency* is $f_{11} = c\sqrt{2}/2a$, while the other frequencies are $f_{mn} = c\sqrt{m^2 + n^2}/2a$, which are **not** integer multiple of the fundamental one (as in the case of the vibrating string).

Going back to our problem, to find a solution which satisfies the initial conditions, we superpose the modes u_{mn} by defining

$$u(x, y, t) = \sum_{m,n=1}^{\infty} [a_{mn} \cos(c\lambda_{mn} t) + b_{mn} \sin(c\lambda_{mn} t)] \sin(\mu_m x) \sin(\nu_n y).$$

Since $u(x, y, 0) = 0$, we choose $a_{mn} = 0$ for every $m, n \geq 1$. From $u_t(x, y, 0) = h(x, y)$ we find the condition

$$\sum_{m,n=1}^{\infty} cb_{mn} \lambda_{mn} \sin(\mu_m x) \sin(\nu_n y) = h(x, y). \quad (5.94)$$

Therefore, we assume that h can be expanded in a double Fourier sine series as follows:

$$h(x, y) = \sum_{m,n=1}^{\infty} h_{mn} \sin(\mu_m x) \sin(\nu_n y),$$

where the coefficients h_{mn} are given by

$$h_{mn} = \frac{4}{a^2} \int_Q h(x, y) \sin\left(\frac{m\pi}{a}x\right) \sin\left(\frac{n\pi}{a}y\right) dx dy.$$

Then, if we choose $b_{mm} = h_{mm}/c\lambda_{mn}$, (5.94) is satisfied. Thus, we have constructed the *formal* solution

$$u(x, y, t) = \sum_{m,n=1}^{\infty} \frac{h_{mn}}{c\lambda_{mn}} \sin(c\lambda_{mn}t) \sin(\mu_m x) \sin(\nu_n y). \quad (5.95)$$

If the coefficients $h_{mm}/c\lambda_{mn}$ vanish fast enough as $m, n \rightarrow +\infty$, it can be shown that (5.95) gives the unique solution²⁴.

5.7.2 Small amplitude sound waves

Sound waves are small disturbances in the density and pressure of a compressible gas. In an isotropic gas, their propagation can be described in terms of a single scalar quantity. Moreover, due to the small amplitudes involved, it is possible to *linearize* the equations of motion, within a reasonable range of validity. Three are the relevant equations: two of them express *conservation of mass* and *balance of linear momentum*, the other one is a *constitutive* relation between density and pressure.

Conservation of mass provides the relation between the gas density $\rho = \rho(\mathbf{x}, t)$ and its velocity $\mathbf{v} = \mathbf{v}(\mathbf{x}, t)$:

$$\rho_t + \operatorname{div}(\rho\mathbf{v}) = 0. \quad (5.96)$$

The balance of linear momentum describes how the volume of gas occupying a region V reacts to the pressure exerted by the rest of the gas. Assuming that the viscosity of the gas is negligible, this force is given by the *normal pressure* $-p\boldsymbol{\nu}$ on the boundary of V ($\boldsymbol{\nu}$ is the exterior normal to ∂V).

Thus, if there are no significant external forces, the linear momentum equation is

$$\frac{D\mathbf{v}}{Dt} \equiv \mathbf{v}_t + (\mathbf{v} \cdot \nabla) \mathbf{v} = -\frac{1}{\rho} \nabla p. \quad (5.97)$$

The last equation is an empirical relation between p and ρ . Since the pressure fluctuations are very rapid, the compressions/expansions of the gas are *adiabatic*, *without any loss of heat*.

In these conditions, if $\gamma = c_p/c_v$ is the ratio of the specific heats of the gas ($\gamma \approx 1.4$ in air) then p/ρ^γ is constant, so that we can write

$$p = f(\rho) = C\rho^\gamma \quad (5.98)$$

with C constant.

²⁴ We leave it to the reader to find appropriate smoothness hypotheses on h , in order to ensure that (5.95) is the unique solution.

The system of equations (5.96), (5.97), (5.98) is quite complicated and extremely difficult to solve in its general form. Here we consider sound waves that are only small perturbations of static atmospheric conditions and this allows a major simplification. In fact, if we consider a static atmosphere, where ρ_0 and p_0 are constant density and pressure, with zero velocity field, we may write

$$\rho = (1 + s) \rho_0 \approx \rho_0,$$

where s is a small dimensionless quantity, called *condensation* and representing the relative variation of the density from equilibrium. Then, from (5.98), we have

$$p - p_0 \approx f'(\rho_0)(\rho - \rho_0) = s\rho_0 f'(\rho_0) \quad (5.99)$$

and

$$\nabla p \approx \rho_0 f'(\rho_0) \nabla s.$$

Now, if \mathbf{v} is also small, we may keep in (5.97) only first order terms in s and \mathbf{v} . Thus, we may neglect the convective acceleration $(\mathbf{v} \cdot \nabla) \mathbf{v}$ and approximate (5.97) and (5.96) with the linear equations

$$\mathbf{v}_t = -c_0^2 \nabla s \quad (5.100)$$

and

$$s_t + \operatorname{div} \mathbf{v} = 0, \quad (5.101)$$

respectively, where we have set $c_0^2 = f'(\rho_0) = C\gamma\rho_0^{\gamma-1}$.

Let us have a closer look on the implications of the above linearization. Suppose that V and S are average values of $|\mathbf{v}|$ and s , respectively. Moreover, let L and T typical order of magnitude for space and time in the wave propagation, such as wavelength and period. Rescale \mathbf{v} , s , \mathbf{x} and t as follows:

$$\boldsymbol{\xi} = \frac{\mathbf{x}}{L}, \quad \tau = \frac{t}{T}, \quad \mathbf{U}(\boldsymbol{\xi}, \tau) = \frac{\mathbf{v}(L\boldsymbol{\xi}, T\tau)}{V}, \quad \sigma(\boldsymbol{\xi}, \tau) = \frac{s(L\boldsymbol{\xi}, T\tau)}{S}. \quad (5.102)$$

Substituting (5.102) into (5.100) and (5.101), we obtain

$$\frac{V}{T} \mathbf{U}_\tau + \frac{c_0^2 S}{L} \nabla \sigma = \mathbf{0} \quad \text{and} \quad \frac{S}{T} \sigma_\tau + \frac{V}{L} \operatorname{div} \mathbf{U} = 0.$$

In this equations the coefficients must be of the same order of magnitude, therefore

$$\frac{V}{T} \approx \frac{c_0^2 S}{L} \quad \text{and} \quad \frac{S}{T} \approx \frac{V}{L}$$

which implies

$$\frac{L}{T} \approx c_0.$$

Hence, c_0 is a typical propagation speed, namely it is the **sound speed**. Now, the convective acceleration is negligible with respect to (say) \mathbf{v}_t , if

$$\frac{V^2}{L} \mathbf{U} \cdot \nabla \mathbf{U} \ll \frac{V}{T} \mathbf{U}_\tau \quad \text{or} \quad V \ll c_0.$$

Thus, if the gas speed is much smaller than the sound speed, our linearization makes sense. The ratio $\mathcal{M} = V/c_0$ is called **Mach number**.

The following theorem, in which we assume that both s and \mathbf{v} are smooth functions, follows from (5.100) and (5.101).

Theorem 5.13. a) *The condensation s is a solution of the wave equation*

$$s_{tt} - c_0^2 \Delta s = 0 \tag{5.103}$$

where $c_0 = \sqrt{f'(\rho_0)} = \sqrt{\gamma p_0 / \rho_0}$ is the speed of sound.

b) *If $\mathbf{v}(\mathbf{x}, 0) = \mathbf{0}$, there exists an acoustic potential ϕ such that $\mathbf{v} = \nabla \phi$. Moreover ϕ satisfies (5.103) as well.*

Proof. a) Taking the divergence on both sides of (5.100) and the t -derivative on both sides of (5.101) we get, respectively:

$$\operatorname{div} \mathbf{v}_t = -c_0^2 \Delta s$$

and

$$s_{tt} = -(\operatorname{div} \mathbf{v})_t.$$

Since $(\operatorname{div} \mathbf{v})_t = \operatorname{div} \mathbf{v}_t$, equation (5.103) follows.

b) We have (see (5.100))

$$\mathbf{v}_t = -c_0^2 \nabla s.$$

Let

$$\phi(\mathbf{x}, t) = -c_0^2 \int_0^t s(\mathbf{x}, z) dz.$$

Then

$$\phi_t = -c_0^2 s$$

and we may write (5.100) in the form

$$\frac{\partial}{\partial t} [\mathbf{v} - \nabla \phi] = \mathbf{0}.$$

Hence, since $\phi(\mathbf{x}, 0) = 0$, $\mathbf{v}(\mathbf{x}, 0) = \mathbf{0}$, we infer

$$\mathbf{v}(\mathbf{x}, t) - \nabla \phi(\mathbf{x}, t) = \mathbf{v}(\mathbf{x}, 0) - \nabla \phi(\mathbf{x}, 0) = \mathbf{0}.$$

Thus $\mathbf{v} = \nabla \phi$. Finally, from (5.101),

$$\phi_{tt} = -c_0^2 s_t = c_0^2 \operatorname{div} \mathbf{v} = c_0^2 \Delta \phi$$

which is (5.103). □

Once the potential ϕ is known, the velocity field \mathbf{v} , the condensation s and the pressure fluctuation $p - p_0$ can be recovered from the following formulas:

$$\mathbf{v} = \nabla\phi, \quad s = -\frac{1}{c_0^2}\phi_t, \quad p - p_0 = -\rho_0\phi_t.$$

Consider, for instance, a plane wave represented by the following potential:

$$\phi(\mathbf{x}, t) = w(\mathbf{x} \cdot \mathbf{k} - \omega t).$$

We know that if

$$c_0^2 |\mathbf{k}|^2 = \omega^2,$$

ϕ is a solution of (5.103). In this case, we have:

$$\mathbf{v} = w'\mathbf{k}, \quad s = \frac{\omega}{c_0^2}w', \quad p - p_0 = \rho_0\omega w'.$$

Example 5.14. Motion of a gas in a tube. Consider a straight cylindrical tube with axis along the x_1 -axis, filled with gas in the region $x_1 > 0$ of density ρ_0 , at pressure p_0 . A flat piston, whose face moves according to the relation $x_1 = h(t)$, sets the gas into motion. We assume that $|h(t)| \ll 1$ and $|h'(t)| \ll c_0$. Under these conditions, the motion of the piston generates sound waves of small amplitude and the acoustic potential ϕ is a solution of the homogeneous wave equation. To compute ϕ we need suitable boundary conditions. The continuity of the normal velocity of the gas at the contact surface with the piston gives

$$\phi_{x_1}(h(t), x_2, x_3, t) = h'(t).$$

Since $h(t) \sim 0$, we may approximate this condition by

$$\phi_{x_1}(0, x_2, x_3, t) = h'(t). \quad (5.104)$$

At the tube walls the normal velocity of the gas is zero, so that, if $\boldsymbol{\nu}$ denotes the outward unit normal vector on the tube wall, we have

$$\nabla\phi \cdot \boldsymbol{\nu} = 0. \quad (5.105)$$

Finally, since the waves are generated by the piston movement, we may look for an *outgoing plane wave*²⁵ solution of the form:

$$\phi(\mathbf{x}, t) = w(\mathbf{x} \cdot \mathbf{n} - c_0 t),$$

where \mathbf{n} is a unit vector. From (5.105) we have

$$\nabla\phi \cdot \boldsymbol{\nu} = w'(\mathbf{x} \cdot \mathbf{n} - c_0 t) \mathbf{n} \cdot \boldsymbol{\nu} = 0$$

²⁵ We do not expect *incoming* waves, which should be generated by sources placed far from the piston.

whence $\mathbf{n} \cdot \boldsymbol{\nu} = 0$ for every $\boldsymbol{\nu}$ orthogonal to the wall tube. Thus, we infer $\mathbf{n} = (1, 0, 0)$ and, as a consequence,

$$\phi(\mathbf{x}, t) = w(x_1 - c_0 t).$$

From (5.104) we get

$$w'(-ct) = h'(t)$$

so that (assuming $h(0) = 0$),

$$w(r) = -c_0 h\left(-\frac{r}{c_0}\right).$$

Hence, the acoustic potential is given by

$$\phi(\mathbf{x}, t) = -c_0 h\left(t - \frac{x_1}{c_0}\right),$$

which represents a *progressive wave* propagating along the tube. In this case:

$$\mathbf{v} = h'\left(t - \frac{x_1}{c_0}\right) \mathbf{i}, \quad s = \frac{1}{c_0} h'\left(t - \frac{x_1}{c_0}\right), \quad p = c_0 \rho_0 h'\left(t - \frac{x_1}{c_0}\right) + p_0.$$

5.8 The Global Cauchy Problem

5.8.1 Fundamental solution ($n = 3$) and strong Huygens' principle

In this section we consider the global Cauchy problem for the three-dimensional homogeneous wave equation:

$$\begin{cases} u_{tt} - c^2 \Delta u = 0 & \mathbf{x} \in \mathbb{R}^3, t > 0 \\ u(\mathbf{x}, 0) = g(\mathbf{x}), \quad u_t(\mathbf{x}, 0) = h(\mathbf{x}) & \mathbf{x} \in \mathbb{R}^3. \end{cases} \quad (5.106)$$

We know from Theorem 5.2 that problem (5.106) has at most one solution $u \in C^2(\mathbb{R}^3 \times [0, +\infty))$. Our purpose here is to show that the solution u exists and to express it in terms of the data g and h , through an explicit formula. Our derivation is rather heuristic so that, for the time being, we do not worry too much about the correct hypotheses on h and g , which we assume as smooth as we need to carry out the calculations.

First we need a lemma that reduces the problem to the case $g = 0$ (and which actually holds in any dimension). Denote by w_h the solution of the problem

$$\begin{cases} w_{tt} - c^2 \Delta w = 0 & \mathbf{x} \in \mathbb{R}^3, t > 0 \\ w(\mathbf{x}, 0) = 0, \quad w_t(\mathbf{x}, 0) = h(\mathbf{x}) & \mathbf{x} \in \mathbb{R}^3. \end{cases} \quad (5.107)$$

Lemma 5.15. *If $w_g \in C^3(\mathbb{R}^3 \times [0, +\infty))$, then $v = \partial_t w_g$ solves the problem*

$$\begin{cases} v_{tt} - c^2 \Delta v = 0 & \mathbf{x} \in \mathbb{R}^3, t > 0 \\ v(\mathbf{x}, 0) = g(\mathbf{x}), \quad v_t(\mathbf{x}, 0) = 0 & \mathbf{x} \in \mathbb{R}^3. \end{cases} \quad (5.108)$$

Therefore the solution of (5.106) is given by

$$u = \partial_t w_g + w_h. \quad (5.109)$$

Proof. Let $v = \partial_t w_g$. Differentiating the wave equation with respect to t , we have

$$0 = \partial_t(\partial_{tt} w_g - c^2 \Delta w_g) = (\partial_{tt} - c^2 \Delta) \partial_t w_g = v_{tt} - c^2 \Delta v.$$

Moreover,

$$v(\mathbf{x}, 0) = \partial_t w_g(\mathbf{x}, 0) = g(\mathbf{x}), \quad v_t(\mathbf{x}, 0) = \partial_{tt} w_g(\mathbf{x}, 0) = c^2 \Delta w_g(\mathbf{x}, 0) = 0.$$

Thus, v is a solution of (5.108) and $u = v + w_h$ is the solution of (5.106). \square

The lemma shows that, once the solution of (5.107) is determined, the solution of the complete problem (5.106) is given by (5.109).

Therefore, we focus on the solution of (5.107), first with a special choice of h , given by the three-dimensional Dirac measure at $\mathbf{0}$, $\delta_3(\mathbf{x})$. For example, in the case of sound waves, this initial datum models a sudden change of the air density, concentrated at the origin. If w represents the density variation with respect to a static atmosphere, then w solves the problem

$$\begin{cases} w_{tt} - c^2 \Delta w = 0 & \mathbf{x} \in \mathbb{R}^3, t > 0 \\ w(\mathbf{x}, 0) = 0, \quad w_t(\mathbf{x}, 0) = \delta_3(\mathbf{x}) & \mathbf{x} \in \mathbb{R}^3. \end{cases} \quad (5.110)$$

The solution of (5.110), which we denote by $K(\mathbf{x}, t)$, is called **fundamental solution** of the three-dimensional wave equation²⁶. To solve (5.110) we approximate the Dirac measure by the *fundamental* solution of the three-dimensional diffusion equation. Indeed, from Subsect. 2.3.4 (choosing $t = \varepsilon$, $D = 1$, $n = 3$), we know that

$$\Gamma(\mathbf{x}, \varepsilon) = \frac{1}{(4\pi\varepsilon)^{3/2}} \exp\left\{-\frac{|\mathbf{x}|^2}{4\varepsilon}\right\} \rightarrow \delta_3(\mathbf{x}) \quad \text{as } \varepsilon \rightarrow 0^+.$$

Denote by w_ε the solution of (5.110) with $\delta(\mathbf{x})$ replaced by $\Gamma(\mathbf{x}, \varepsilon)$. Since $\Gamma(\mathbf{x}, \varepsilon)$ is radially symmetric with pole at $\mathbf{0}$, we expect that w_ε shares the same type of symmetry and it is a spherical wave of the form

$$w_\varepsilon = w_\varepsilon(r, t), \quad r = |\mathbf{x}|.$$

²⁶ Or the *Green function* in \mathbb{R}^3 .

Thus, from (5.85) we may write

$$w_\varepsilon(r, t) = \frac{F(r + ct)}{r} + \frac{G(r - ct)}{r}. \quad (5.111)$$

The initial conditions require

$$F(r) + G(r) = 0 \quad \text{and} \quad c(F'(r) - G'(r)) = r\Gamma(r, \varepsilon)$$

or

$$F = -G \quad \text{and} \quad G'(r) = -r\Gamma(r, \varepsilon)/2c.$$

Integrating the second relation yields

$$G(r) = -\frac{1}{2c(4\pi\varepsilon)^{3/2}} \int_0^r s \exp\left\{-\frac{s^2}{4\varepsilon}\right\} ds = \frac{1}{4\pi c} \frac{1}{\sqrt{4\pi\varepsilon}} \left(\exp\left\{-\frac{r^2}{4\varepsilon}\right\} - 1\right)$$

and finally

$$w_\varepsilon(r, t) = \frac{1}{4\pi cr} \left\{ \frac{1}{\sqrt{4\pi\varepsilon}} \exp\left\{-\frac{(r-ct)^2}{4\varepsilon}\right\} - \frac{1}{\sqrt{4\pi\varepsilon}} \exp\left\{-\frac{(r+ct)^2}{4\varepsilon}\right\} \right\}.$$

Now observe that the function

$$\tilde{\Gamma}(r, \varepsilon) = \frac{1}{\sqrt{4\pi\varepsilon}} \exp\left\{-\frac{r^2}{4\varepsilon}\right\}$$

is the fundamental solution of the one-dimensional diffusion equation with $x = r$ and $t = \varepsilon$. We let now $\varepsilon \rightarrow 0^+$. Since $r + ct > 0$ for every $t > 0$,

$$\frac{1}{\sqrt{4\pi\varepsilon}} \exp\left\{-\frac{(r+ct)^2}{4\varepsilon}\right\} \rightarrow 0$$

while

$$\frac{1}{\sqrt{4\pi\varepsilon}} \exp\left\{-\frac{(r-ct)^2}{4\varepsilon}\right\} \rightarrow \delta(r - ct).$$

Therefore we conclude that, for $t > 0$,

$$K(\mathbf{x}, t) = \frac{\delta(|\mathbf{x}| - ct)}{4\pi cr}. \quad (5.112)$$

Moreover, we have $K(\mathbf{x}, 0) = 0$, since $w_\varepsilon(\mathbf{x}, 0) = 0$ for all ε , and, at least formally²⁷, $K_t(\mathbf{x}, 0) = \delta_3(\mathbf{x})$, since $\partial_t w_\varepsilon(\mathbf{x}, 0) \rightarrow \delta_3(\mathbf{x})$ as $\varepsilon \rightarrow 0$.

Let now the initial datum be concentrated at a point \mathbf{y} , that is $w_t(\mathbf{x}, 0) = \delta_3(\mathbf{x} - \mathbf{y})$. Due to the invariance under translation in space of the physical sys-

²⁷ A rigorous justification uses the theory of Schwartz distributions. See Example 7.27, p. 447 and Problem 7.17.

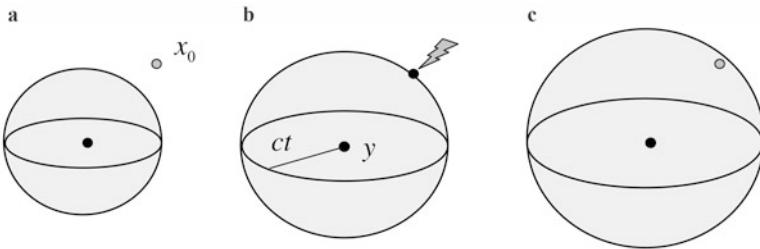


Fig. 5.10 Huygens principle. (a) $ct < |x_0 - y|$; (b) $ct = |x_0 - y|$; (c) $ct > |x_0 - y|$

tem, the corresponding fundamental solution is given by

$$K(\mathbf{x} - \mathbf{y}, t) = \frac{\delta(|\mathbf{x} - \mathbf{y}| - ct)}{4\pi c |\mathbf{x} - \mathbf{y}|} \quad (t > 0),$$

which represents an *outgoing travelling wave*, initially supported at \mathbf{y} and thereafter on the surface

$$\partial B_{ct}(\mathbf{y}) = \{\mathbf{x} : |\mathbf{x} - \mathbf{y}| = ct\}.$$

The union of the surfaces $\partial B_{ct}(\mathbf{y})$ is called the **support** of $K(\mathbf{x} - \mathbf{y}, t)$ and coincides with **the boundary of the forward space-time cone, with vertex at $(\mathbf{y}, 0)$** and opening $\theta = \tan^{-1} c$, given by

$$C_{\mathbf{y}, 0}^* = \{(\mathbf{x}, t) : |\mathbf{x} - \mathbf{y}| \leq ct, t > 0\}.$$

Following the terminology of Sect. 5.4, $\partial C_{\mathbf{y}, 0}^*$ constitutes the **range of influence of the point \mathbf{y}** .

The fact that the range of influence of the point \mathbf{y} is only the *boundary* of the forward cone and *not the full cone* has important consequences on the nature of the disturbances governed by the three-dimensional wave equation. The most striking phenomenon is that a perturbation generated at time $t = 0$ by a point source placed at \mathbf{y} is felt at the point \mathbf{x}_0 **only at time** $t_0 = |\mathbf{x}_0 - \mathbf{y}|/c$ (Fig. 5.10). This is known as *strong Huygens' principle* and explains why *sharp signals* are propagated from a point source.

We will shortly see that this is not the case in two dimensions.

5.8.2 The Kirchhoff formula

Using the fundamental solution as in Sect. 5.4.3, we may derive a formula for the solution of (5.107) with a general datum h . Since

$$h(\mathbf{x}) = \int_{\mathbb{R}^3} \delta_3(\mathbf{x} - \mathbf{y}) h(\mathbf{y}) d\mathbf{y},$$

we may see h as a superposition of the impulses $\delta_3(\mathbf{x} - \mathbf{y}) h(\mathbf{y})$, localized at \mathbf{y} , of strength $h(\mathbf{y})$. Accordingly, the solution of (5.107) is given by the superposition

of the corresponding solutions $K(\mathbf{x} - \mathbf{y}, t)h(\mathbf{y})$, that is, for $t > 0$,

$$w_h(\mathbf{x}, t) = \int_{\mathbb{R}^3} K(\mathbf{x} - \mathbf{y}, t)h(\mathbf{y}) d\mathbf{y} = \int_{\mathbb{R}^3} \frac{\delta(|\mathbf{x} - \mathbf{y}| - ct)}{4\pi c |\mathbf{x} - \mathbf{y}|} h(\mathbf{y}) d\mathbf{y}. \quad (5.113)$$

Thus w_h is the **\mathbf{x} - convolution of h with the fundamental solution**. Using the formula

$$\int_0^\infty \delta(r - ct) f(r) dr = f(ct)$$

we can write

$$w_h(\mathbf{x}, t) = \int_0^\infty \frac{\delta(r - ct)}{4\pi cr} dr \int_{\partial B_r(\mathbf{x})} h(\boldsymbol{\sigma}) d\sigma = \frac{1}{4\pi c^2 t} \int_{\partial B_{ct}(\mathbf{x})} h(\boldsymbol{\sigma}) d\sigma.$$

Lemma 5.15, p. 311, and the above intuitive argument lead to the following theorem:

Theorem 5.16 (Kirchhoff's formula). *Let $g \in C^3(\mathbb{R}^3)$ and $h \in C^2(\mathbb{R}^3)$. Then,*

$$u(\mathbf{x}, t) = \frac{\partial}{\partial t} \left[\frac{1}{4\pi c^2 t} \int_{\partial B_{ct}(\mathbf{x})} g(\boldsymbol{\sigma}) d\sigma \right] + \frac{1}{4\pi c^2 t} \int_{\partial B_{ct}(\mathbf{x})} h(\boldsymbol{\sigma}) d\sigma \quad (5.114)$$

is the unique solution $u \in C^2(\mathbb{R}^3 \times [0, +\infty))$ of problem (5.106).

Proof. Letting $\boldsymbol{\sigma} = \mathbf{x} + ct\boldsymbol{\omega}$, where $\boldsymbol{\omega} \in \partial B_1(\mathbf{0})$, we have $d\sigma = c^2 t^2 d\omega$ and we may write

$$w_g(\mathbf{x}, t) = \frac{1}{4\pi c^2 t} \int_{\partial B_{ct}(\mathbf{x})} g(\boldsymbol{\sigma}) d\sigma = \frac{t}{4\pi} \int_{\partial B_1(\mathbf{0})} g(\mathbf{x} + ct\boldsymbol{\omega}) d\omega.$$

Since $g \in C^3(\mathbb{R}^3)$, this formula shows that w_g satisfies the hypotheses of Lemma 5.15. Since $h \in C^2(\mathbb{R}^3)$, it follows that

$$w_h(\mathbf{x}, t) = \frac{1}{4\pi c^2 t} \int_{\partial B_{ct}(\mathbf{x})} h(\boldsymbol{\sigma}) d\sigma = \frac{t}{4\pi} \int_{\partial B_1(\mathbf{0})} h(\mathbf{x} + ct\boldsymbol{\omega}) d\omega$$

belongs to $C^2(\mathbb{R}^3 \times [0, +\infty))$ and therefore, to conclude the proof, it is enough to check that w_h solves problem (5.107). We have:

$$\partial_t w_h(\mathbf{x}, t) = \frac{1}{4\pi} \int_{\partial B_1(\mathbf{0})} h(\mathbf{x} + ct\boldsymbol{\omega}) d\omega + \frac{ct}{4\pi} \int_{\partial B_1(\mathbf{0})} \nabla h(\mathbf{x} + ct\boldsymbol{\omega}) \cdot \boldsymbol{\omega} d\omega. \quad (5.115)$$

Thus,

$$w_h(\mathbf{x}, 0) = 0 \quad \text{and} \quad \partial_t w_h(\mathbf{x}, 0) = h(\mathbf{x}).$$

Moreover, by Gauss' formula, we may write

$$\begin{aligned} \frac{ct}{4\pi} \int_{\partial B_1(\mathbf{0})} \nabla h(\mathbf{x} + ct\boldsymbol{\omega}) \cdot \boldsymbol{\omega} d\omega &= \frac{1}{4\pi ct} \int_{\partial B_{ct}(\mathbf{x})} \partial_\nu h(\boldsymbol{\sigma}) d\sigma \\ &= \frac{1}{4\pi ct} \int_{B_{ct}(\mathbf{x})} \Delta h(\mathbf{y}) d\mathbf{y} \\ &= \frac{1}{4\pi ct} \int_0^{ct} dr \int_{\partial B_r(\mathbf{x})} \Delta h(\boldsymbol{\sigma}) d\sigma, \end{aligned}$$

whence, from (5.115),

$$\begin{aligned}\partial_{tt} w_h(\mathbf{x}, t) &= \frac{c}{4\pi} \int_{\partial B_1(\mathbf{0})} \nabla h(\mathbf{x} + ct\boldsymbol{\omega}) \cdot \boldsymbol{\omega} d\omega - \frac{1}{4\pi c t^2} \int_{B_{ct}(\mathbf{x})} \Delta h(\mathbf{y}) d\mathbf{y} \\ &\quad + \frac{1}{4\pi t} \int_{\partial B_{ct}(\mathbf{x})} \Delta h(\boldsymbol{\sigma}) d\sigma \\ &= \frac{1}{4\pi t} \int_{\partial B_{ct}(\mathbf{x})} \Delta h(\boldsymbol{\sigma}) d\sigma.\end{aligned}$$

On the other hand,

$$\Delta w_h(\mathbf{x}, t) = \frac{t}{4\pi} \int_{\partial B_1(\mathbf{0})} \Delta h(\mathbf{x} + ct\boldsymbol{\omega}) d\omega = \frac{1}{4\pi c^2 t} \int_{\partial B_{ct}(\mathbf{x})} \Delta h(\boldsymbol{\sigma}) d\sigma$$

and therefore

$$\partial_{tt} w_h - c^2 \Delta w_h = 0.$$

□

Using formula (5.115) with g instead of h , we may write the Kirchhoff formula in the following form:

$$u(\mathbf{x}, t) = \frac{1}{4\pi c^2 t^2} \int_{\partial B_{ct}(\mathbf{x})} \{g(\boldsymbol{\sigma}) + \nabla g(\boldsymbol{\sigma}) \cdot (\boldsymbol{\sigma} - \mathbf{x}) + th(\boldsymbol{\sigma})\} d\sigma. \quad (5.116)$$

The presence of the gradient of g in (5.116) suggests that, unlike in the one-dimensional case, the solution u *may be more irregular than the data*. Indeed, if $g \in C^k(\mathbb{R}^3)$ and $h \in C^{k-1}(\mathbb{R}^3)$, $k \geq 2$, we can only guarantee that u is C^{k-1} and u_t is C^{k-2} at a later time.

Formula (5.116) makes perfect sense also for $g \in C^1(\mathbb{R}^3)$ and h bounded. Clearly, under these weaker hypotheses, (5.116) satisfies the wave equation in an appropriate generalized sense, as in Subsect. 5.4.2, for instance. In this case, scattered singularities in the initial data h may concentrate at later time on smaller sets, giving rise to stronger singularities (*focussing effect*, see Problem 5.16).

According to (5.116), $u(\mathbf{x}, t)$ depends upon the data g and h only on the surface $\partial B_{ct}(\mathbf{x})$, which therefore coincides with the **domain of dependence for** (\mathbf{x}, t) .

Assume that the support of g and h is a compact set D . Then $u(\mathbf{x}, t)$ is different from zero only for $t_{\min} < t < t_{\max}$ where t_{\min} and t_{\max} are the *first* and the *last* time t such that

$$D \cap \partial B_{ct}(\mathbf{x}) \neq \emptyset.$$

In other words, a disturbance, initially localized inside D , starts affecting the point \mathbf{x} at time t_{\min} and ceases to affect it after time t_{\max} . This is another way to express the *strong Huygens principle*.

Fix t and consider the union of all the spheres $\partial B_{ct}(\boldsymbol{\xi})$ as $\boldsymbol{\xi}$ varies on ∂D . The envelope of these surfaces constitutes the *wave front* and bounds the support of u , which spreads at speed c (see Problem 5.15).

5.8.3 The Cauchy problem in dimension 2

The solution of the Cauchy problem in two dimensions can be obtained from Kirchhoff formula, using the so called *Hadamard's method of descent*. Consider first the problem

$$\begin{cases} w_{tt} - c^2 \Delta w = 0 & \mathbf{x} \in \mathbb{R}^2, t > 0 \\ w(\mathbf{x}, 0) = 0, \quad w_t(\mathbf{x}, 0) = h(\mathbf{x}) & \mathbf{x} \in \mathbb{R}^2. \end{cases} \quad (5.117)$$

The key idea is to “embed” the two-dimensional problem (5.117) into a three-dimensional setting. More precisely, write the points in \mathbb{R}^3 as (\mathbf{x}, x_3) and set $h(\mathbf{x}, x_3) = h(\mathbf{x})$, where now $\mathbf{x} \in \mathbb{R}^2$. The solution U of the three-dimensional problem is given by Kirchhoff's formula:

$$U(\mathbf{x}, x_3, t) = \frac{1}{4\pi c^2 t} \int_{\partial B_{ct}(\mathbf{x}, x_3)} h \, d\sigma. \quad (5.118)$$

We claim that, since h does not depend on x_3 , U is independent of x_3 as well, and therefore the solution of (5.117) is given by (5.118) with, say, $x_3 = 0$.

To prove the claim, note that the spherical surface $\partial B_{ct}(\mathbf{x}, x_3)$ is the union of two hemispheres, whose equations are

$$y_3 = F_{\pm}(\mathbf{y}) = x_3 \pm \sqrt{c^2 t^2 - r^2},$$

where $r = |\mathbf{y} - \mathbf{x}|$. On both hemispheres we have:

$$\begin{aligned} d\sigma &= \sqrt{1 + |\nabla F_{\pm}|^2} \, d\mathbf{y} \\ &= \sqrt{1 + \frac{r^2}{c^2 t^2 - r^2}} \, d\mathbf{y} = \frac{ct}{\sqrt{c^2 t^2 - r^2}} \, d\mathbf{y} \end{aligned}$$

and therefore may write

$$U(\mathbf{x}, x_3, t) = \frac{1}{2\pi c} \int_{B_{ct}(\mathbf{x})} \frac{h(\mathbf{y})}{\sqrt{c^2 t^2 - |\mathbf{x} - \mathbf{y}|^2}} \, d\mathbf{y}.$$

Thus, U is independent of x_3 as claimed. From the above calculations, and recalling Lemma 5.15, p. 311, we deduce the following theorem.

Theorem 5.17 (Poisson's formula). *Let $g \in C^3(\mathbb{R}^2)$ and $h \in C^2(\mathbb{R}^2)$. Then,*

$$u(\mathbf{x}, t) = \frac{1}{2\pi c} \left\{ \frac{\partial}{\partial t} \int_{B_{ct}(\mathbf{x})} \frac{g(\mathbf{y}) \, d\mathbf{y}}{\sqrt{c^2 t^2 - |\mathbf{x} - \mathbf{y}|^2}} + \int_{B_{ct}(\mathbf{x})} \frac{h(\mathbf{y}) \, d\mathbf{y}}{\sqrt{c^2 t^2 - |\mathbf{x} - \mathbf{y}|^2}} \right\}$$

is the unique solution $u \in C^2(\mathbb{R}^2 \times [0, +\infty))$ of the problem

$$\begin{cases} u_{tt} - c^2 \Delta u = 0 & \mathbf{x} \in \mathbb{R}^2, t > 0 \\ u(\mathbf{x}, 0) = g(\mathbf{x}), \quad u_t(\mathbf{x}, 0) = h(\mathbf{x}) & \mathbf{x} \in \mathbb{R}^2. \end{cases}$$

Also Poisson's formula can be written in a somewhat more explicit form. Indeed, letting $\mathbf{y} - \mathbf{x} = ct\mathbf{z}$, we have

$$d\mathbf{y} = c^2 t^2 d\mathbf{z}, \quad |\mathbf{x} - \mathbf{y}|^2 = c^2 t^2 |\mathbf{z}|^2$$

whence

$$\int_{B_{ct}(\mathbf{x})} \frac{g(\mathbf{y})}{\sqrt{c^2 t^2 - |\mathbf{x} - \mathbf{y}|^2}} d\mathbf{y} = ct \int_{B_1(\mathbf{0})} \frac{g(\mathbf{x} + ct\mathbf{z})}{\sqrt{1 - |\mathbf{z}|^2}} d\mathbf{z}.$$

Then

$$\begin{aligned} & \frac{\partial}{\partial t} \int_{B_{ct}(\mathbf{x})} \frac{g(\mathbf{y})}{\sqrt{c^2 t^2 - |\mathbf{x} - \mathbf{y}|^2}} d\mathbf{y} \\ &= c \int_{B_1(\mathbf{0})} \frac{g(\mathbf{x} + ct\mathbf{z})}{\sqrt{1 - |\mathbf{z}|^2}} d\mathbf{z} + c^2 t \int_{B_1(\mathbf{0})} \frac{\nabla g(\mathbf{x} + ct\mathbf{z}) \cdot \mathbf{z}}{\sqrt{1 - |\mathbf{z}|^2}} d\mathbf{z} \end{aligned}$$

and, going back to the original variables, we obtain

$$u(\mathbf{x}, t) = \frac{1}{2\pi ct} \int_{B_{ct}(\mathbf{x})} \frac{g(\mathbf{y}) + \nabla g(\mathbf{y}) \cdot (\mathbf{y} - \mathbf{x}) + th(\mathbf{y})}{\sqrt{c^2 t^2 - |\mathbf{x} - \mathbf{y}|^2}} d\mathbf{y}. \quad (5.119)$$

Poisson's formula displays an important difference with respect to its three-dimensional analogue, Kirchhoff's formula. In fact *the domain of dependence* of the point (\mathbf{x}, t) is given by the **full circle**

$$B_{ct}(\mathbf{x}) = \{\mathbf{y}: |\mathbf{x} - \mathbf{y}| < ct\}.$$

This entails that a disturbance, initially localized at ξ , starts affecting the point \mathbf{x} at time $t_{\min} = |\mathbf{x} - \xi|/c$. However, this effect does not vanish for $t > t_{\min}$, since ξ still belongs to the circle $B_{ct}(\mathbf{x})$ after t_{\min} .

It is the phenomenon one may observe by placing a cork on still water and dropping a stone not too far away. The cork remains undisturbed until it is reached by the wave front, but its oscillations persist thereafter.

Thus, sharp signals do not exist in dimension two and *the strong Huygens principle does not hold*. It could be observed that if the initial data are compactly supported, then (5.119) shows a decay in time of order $1/t$ for any fixed \mathbf{x} , as $t \rightarrow \infty$. This decay becomes of order $1/t^2$ if $h \equiv 0$.

Remark 5.18. An examination of Poisson's formula reveals that the fundamental solution for the two dimensional wave equation is given by

$$K(\mathbf{x}, t) = \frac{1}{2\pi c} \frac{1}{\sqrt{c^2 t^2 - |\mathbf{x}|^2}} \chi_{B_{ct}(\mathbf{x})}, \quad (5.120)$$

where $\chi_{B_{ct}(\mathbf{x})}$ is the characteristic function of $B_{ct}(\mathbf{x})$. Its support is the **full forward space-time cone**, with vertex at $(\mathbf{0}, 0)$ and opening $\theta = \tan^{-1} c$, given by

$$C_{\mathbf{0},0}^* = \{(\mathbf{x}, t) : |\mathbf{x}| \leq ct, t \geq 0\}.$$

5.9 The Cauchy Problem with Distributed Sources

5.9.1 Retarded potentials ($n = 3$)

The solution of the nonhomogeneous Cauchy problem can be obtained via Duhamel's method. We give the details for $n = 3$ only (for $n = 2$, see Problem 5.18). By linearity it is enough to derive a formula for the solution of the problem with zero initial data:

$$\begin{cases} u_{tt} - c^2 \Delta u = f(\mathbf{x}, t) & \mathbf{x} \in \mathbb{R}^3, t > 0 \\ u(\mathbf{x}, 0) = 0, \quad u_t(\mathbf{x}, 0) = 0 & \mathbf{x} \in \mathbb{R}^3. \end{cases} \quad (5.121)$$

Assume that $f \in C^2(\mathbb{R}^3 \times [0, +\infty))$. For $s \geq 0$ fixed, let $w = w(\mathbf{x}, t; s)$ be the solution of the problem

$$\begin{cases} w_{tt} - c^2 \Delta w = 0 & \mathbf{x} \in \mathbb{R}^3, t \geq s \\ w(\mathbf{x}, s; s) = 0, \quad w_t(\mathbf{x}, s; s) = f(\mathbf{x}, s) & \mathbf{x} \in \mathbb{R}^3. \end{cases}$$

Since the wave equation is invariant under space-time translations, w is given by Kirchhoff's formula with t replaced by $t - s$:

$$w(\mathbf{x}, t; s) = \frac{1}{4\pi c^2(t-s)} \int_{\partial B_{c(t-s)}(\mathbf{x})} f(\boldsymbol{\sigma}, s) d\sigma.$$

Then,

$$u(\mathbf{x}, t) = \int_0^t w(\mathbf{x}, t; s) ds = \frac{1}{4\pi c^2} \int_0^t \frac{ds}{(t-s)} \int_{\partial B_{c(t-s)}(\mathbf{x})} f(\boldsymbol{\sigma}, s) d\sigma \quad (5.122)$$

is the unique solution $u \in C^2(\mathbb{R}^3 \times [0, +\infty))$ of (5.121)²⁸.

Formula (5.122) shows that $u(\mathbf{x}, t)$ depends on the values of f in the full **backward** cone

$$C_{\mathbf{x}, t} = \{(\mathbf{z}, s) : |\mathbf{z} - \mathbf{x}| \leq c(t-s), 0 \leq s \leq t\}.$$

Note that (5.122) may also be written in the form

$$u(\mathbf{x}, t) = \frac{1}{4\pi c^2} \int_{B_{ct}(\mathbf{x})} \frac{1}{|\mathbf{x} - \mathbf{y}|} f\left(\mathbf{y}, t - \frac{|\mathbf{x} - \mathbf{y}|}{c}\right) d\mathbf{y} \quad (5.123)$$

which is a so called *retarded potential*. In fact, the value of u at \mathbf{x} at time t does not depend on the source at the same time, but rather at the earlier time

$$t_{ret} = t - \frac{|\mathbf{x} - \mathbf{y}|}{c}.$$

²⁸ Check it, mimicking the proof in dimension one (Sect. 5.4.4).

The time lag $t - t_{ret}$ is precisely the time required by the source effect to propagate from \mathbf{y} to \mathbf{x} . This feature is clearly due to the finite speed c of propagation of the disturbances.

We can also write u in terms of the fundamental solution

$$K(\mathbf{x} - \mathbf{y}, t - s) = \frac{\delta(|\mathbf{x} - \mathbf{y}| - c(t - s))}{4\pi c |\mathbf{x} - \mathbf{y}|}.$$

Indeed, recalling formula (5.113), u can be expressed by the space-time convolution

$$u(\mathbf{x}, t) = \int_0^t \int_{\mathbb{R}^3} K(\mathbf{x} - \mathbf{y}, t - s) f(\mathbf{y}, s) d\mathbf{y} ds. \quad (5.124)$$

Important applications to Acoustics or Electromagnetism require more general f , modeling for instance various types of point sources. The method still works, as we shall see in the next example and, more significantly, in the next section.

- *Point source.* Let

$$f(\mathbf{x}, t) = \delta_3(\mathbf{x}) q(t).$$

f models a source concentrated at $\mathbf{x} = \mathbf{0}$, with strength $q = q(t)$, that we assume smoothly varying with time for $t \geq 0$, and *zero for* $t < 0$. Formally, substituting this source into (5.123) we obtain, since $\delta_3(\mathbf{y})$ requires to set $\mathbf{y} = \mathbf{0}$ inside the integral,

$$u(\mathbf{x}, t) = \frac{1}{4\pi c^2} \frac{1}{|\mathbf{x}|} q\left(t - \frac{|\mathbf{x}|}{c}\right). \quad (5.125)$$

Note that $u = 0$ for $|\mathbf{x}| > ct$. To justify the formula (5.125), we approximate δ_3 by a smooth function $\psi_\varepsilon = \psi_\varepsilon(\mathbf{x})$ supported in the ball $B_\varepsilon = B_\varepsilon(\mathbf{0})$, with unit strength, that is

$$\int_{B_\varepsilon} \psi_\varepsilon(\mathbf{x}) d\mathbf{x} = 1, \quad (5.126)$$

and such that

$$\psi_\varepsilon \rightarrow \delta_3 \text{ as } \varepsilon \rightarrow 0. \quad (5.127)$$

A function like ψ_ε is called an approximation of the identity and can be constructed in several ways (see Problem 7.1). Recall that the limit in (5.127) means

$$\int_{\mathbb{R}^n} \psi_\varepsilon(\mathbf{x}) \varphi(\mathbf{x}) d\mathbf{x} \rightarrow \varphi(\mathbf{0}) \quad \text{as } \varepsilon \rightarrow 0,$$

for every smooth function φ , compactly supported in \mathbb{R}^3 .

Let $f_\varepsilon(\mathbf{x}, t) = \psi_\varepsilon(\mathbf{x}) q(t) \mathcal{H}(t)$, where \mathcal{H} is the Heaviside function. Assume that $\varepsilon < |\mathbf{x}| < ct$ and $\varepsilon < ct - |\mathbf{x}|$. Then $B_\varepsilon \subset B_{ct}(\mathbf{x})$ and the integration in (5.123) can be restricted to B_ε . We find

$$u_\varepsilon(\mathbf{x}, t) = \frac{1}{4\pi c^2} \int_{B_\varepsilon} q\left(t - \frac{|\mathbf{x} - \mathbf{y}|}{c}\right) \frac{\psi_\varepsilon(\mathbf{y})}{|\mathbf{x} - \mathbf{y}|} d\mathbf{y}. \quad (5.128)$$

Letting $\varepsilon \rightarrow 0$ we get

$$u(\mathbf{x}, t) = \frac{1}{4\pi c^2} \frac{1}{|\mathbf{x}|} q \left(t - \frac{|\mathbf{x}|}{c} \right)$$

for all $\mathbf{x} \neq \mathbf{0}$, which is (5.125).

5.9.2 Radiation from a moving point source

A more interesting case occurs when a source (of acoustic waves, say) is moving. For simplicity we consider a point source of constant strength S , moving along the x_3 -axis with constant speed $v > 0$. We can model this kind of source by the formula

$$f(\mathbf{x}, t) = S\delta_3(\mathbf{x} - vt\mathbf{k}) = S\delta_3(x_1, x_2, x_3 - vt)$$

where \mathbf{k} is the unit vector along the axis x_3 . The idea is that the contributions of f come only from the points where δ_3 vanishes, that is from the points $\mathbf{x} = vt\mathbf{k}$. Here $\mathbf{x} \in \mathbb{R}^3$ and t goes (ideally) from $-\infty$ to $+\infty$.

Thus, we want to find a potential u such that

$$u_{tt} - c^2 \Delta u = S\delta_3(\mathbf{x} - vt\mathbf{k}) \quad \text{in } \mathbb{R}^3 \times \mathbb{R}. \quad (5.129)$$

This problem is rather anomalous, not only due to the presence of a travelling delta function as a source, but also because it is formulated in the entire space-time and there are not initial conditions. To find u we first proceed in a formal²⁹ way, expressing u as a retarded potential, obtained substituting

$$f(\mathbf{y}, t) = S\delta_3(\mathbf{y} - vt\mathbf{k})$$

into (5.123); we find

$$u(\mathbf{x}, t) = \frac{S}{4\pi c^2} \int \frac{1}{|\mathbf{x} - \mathbf{y}|} \delta_3 \left(\mathbf{y} - v(t - \frac{|\mathbf{x} - \mathbf{y}|}{c})\mathbf{k} \right) d\mathbf{y} \quad (5.130)$$

where the “integration” is extended over the set where the argument of δ_3 vanishes. It is actually rather delicate to interpret rigorously formula (5.130), since it involves the composition of the δ_3 measure with a nonlinear function of \mathbf{y} , smooth outside \mathbf{x} . Let us write (5.130) in a simpler form. To do it, we recall three properties of the δ_n measure, whose proof can be found in Chap. 7.

The first property states that

$$\delta_n(\mathbf{x}) = \delta(x_1)\delta(x_2)\cdots\delta(x_n)$$

where the right hand side denotes a *tensor product*³⁰ of delta’s. This formula simply means that the action of the n -dimensional delta measure can be decomposed into the actions of the 1-dimensional delta measures $\delta(x_1), \delta(x_2), \dots, \delta(x_n)$. In

²⁹ But, we believe, instructive.

³⁰ See Remark 5.21, p. 327.

formulas, this means, taking $n = 3$ for simplicity:

$$\int \delta_3(\mathbf{x}) \varphi(\mathbf{x}) d\mathbf{x} = \int \delta_1(x_1) \left(\int \delta(x_2) \left(\int \delta(x_3) \varphi(x_1, x_2, x_3) dx_3 \right) dx_2 \right) dx_1$$

for every smooth function φ , compactly supported.

The second one is that δ_n is an *even* distribution, i.e.

$$\delta_n(-\mathbf{x}) = \delta_n(\mathbf{x}). \quad (5.131)$$

The third one shows how δ_n behaves under a space dilation in \mathbb{R}^n . We need it in dimension $n = 1$:

$$\delta(ax) = \frac{1}{|a|} \delta(x), \quad \forall a \in \mathbb{R}. \quad (5.132)$$

Using the above properties, we can write our moving source in the following way:

$$S_3(\mathbf{x} - vt\mathbf{k}) = \delta_3(x_1, x_2, x_3 - vt) = \frac{S}{v} \delta(x_1) \delta(x_2) \delta\left(t - \frac{x_3}{v}\right).$$

Let us now go back to (5.130) and evaluate the action of $\delta(y_1) \delta(y_2)$, which demands to set $y_1 = y_2 = 0$ everywhere inside the integral. We get:

$$u(\mathbf{x}, t) = \frac{S}{4\pi c^2 v} \int \frac{1}{|\mathbf{x} - y_3 \mathbf{k}|} \delta\left(t - \frac{|\mathbf{x} - y_3 \mathbf{k}|}{c} - \frac{y_3}{v}\right) dy_3. \quad (5.133)$$

The “integration” in (5.133) is extended along the y_3 -axis, over the set where the argument of δ vanishes. As we shall see shortly, (5.133) can be handled by means of formula (7.24), in Proposition 7.29, p. 448, and allows us to derive a simple analytical expression for the moving source potential.

However, before computing explicitly u , we draw some interesting information from (5.133). In particular we may ask: *which points \mathbf{x} can be affected by the source radiation within time t ?* To answer, observe that the set where the argument of δ vanishes is given by

$$vt - y_3 = \frac{v}{c} |\mathbf{x} - y_3 \mathbf{k}|. \quad (5.134)$$

At a given point \mathbf{x} and at a given time t , $u(\mathbf{x}, t)$ is obtained by the superposition of the potentials generated by the moving source when this was located at the points $(0, 0, y_3)$, where y_3 is a solution of (5.134). Clearly we must have $y_3 \leq vt$. For instance, the perturbation originated at $y_3 = 0$ has reached, at time t , the sphere $\partial B_{ct}(\mathbf{0})$. More generally, formula (5.134) shows that the perturbation originated at any point $\mathbf{p} = (0, 0, y_3)$, with $y_3 < vt$, has reached, at time t , the points on the sphere $\partial B_{\rho_t}(\mathbf{p})$ with

$$\rho_t = \frac{c}{v} (vt - y_3).$$

We now distinguish two cases: the *subsonic case* $v < c$ and the *supersonic case* $v > c$. Set $m = v/c$.

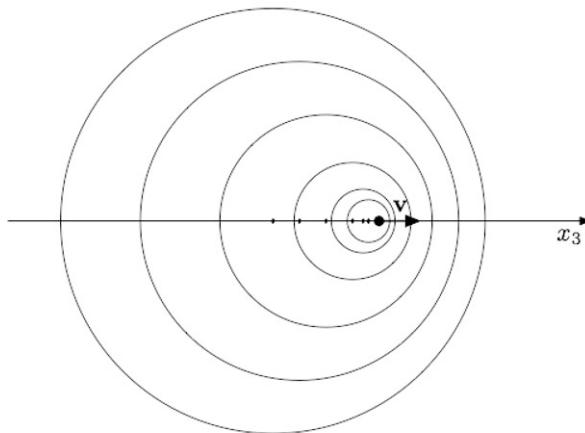


Fig. 5.11 The subsonic case

- The *subsonic case* $m < 1$. The radiation effect is described in Fig. 5.11, where the x_3 -axis is horizontal. Any point \mathbf{x} can be reached at time t by the perturbation originated by the source when it was located at a suitable point \mathbf{p} . A similar picture occurs when we hear the siren of an ambulance moving along a straight road at constant speed.

Let now compute explicitly the “integral” in (5.133). The first thing to do is to compute the roots of (5.134), for (\mathbf{x}, t) fixed. Squaring both sides, we get

$$v^2 t^2 - 2vty_3 + y_3^2 = \frac{v^2}{c^2} \left(|\mathbf{x}|^2 - 2x_3 y_3 + y_3^2 \right)$$

or

$$(1 - m^2) y_3^2 - 2(vt - m^2 x_3) y_3 + v^2 t^2 - m^2 |\mathbf{x}|^2 = 0.$$

Setting $\mathbf{x}' = (x_1, x_2)$, after some algebra we find the roots

$$y_3^\pm = \frac{1}{1 - m^2} \left[vt - m^2 x_3 \pm m \sqrt{|\mathbf{x}'|^2 (1 - m^2) + (vt - x_3)^2} \right]. \quad (5.135)$$

Since (5.134) must hold, we check if $vt - y_3^\pm \geq 0$. Thus we compute

$$vt - y_3^\pm = \frac{1}{1 - m^2} \left[m^2(x_3 - vt) \mp m \sqrt{|\mathbf{x}'|^2 (1 - m^2) + (vt - x_3)^2} \right].$$

Consider y_3^+ . Then the denominator is positive and the numerator is negative since it is less than

$$m^2(x_3 - vt) - m|vt - x_3| \leq (m^2 - m)|vt - x_3| \leq 0.$$

Thus $vt - y_3^+ \leq 0$. On the contrary, $vt - y_3^- \geq 0$ since $m^2(x_3 - vt) + m|vt - x_3| \geq 0$ and therefore the only admissible root is y_3^- .

Now we use formula (7.24), p. 448, with

$$g(y_3) = t - \frac{|\mathbf{x} - y_3 \mathbf{k}|}{c} - \frac{y_3}{v}.$$

Accordingly, we can write

$$\delta(g(y_3)) = \frac{\delta(y_3 - y_3^-)}{|g'(y_3^-)|}$$

and (5.133) yields

$$u(\mathbf{x}, t) = \frac{S}{4\pi c^2 v} \frac{1}{|\mathbf{x} - y_3^- \mathbf{k}| |g'(y_3^-)|}.$$

Now, $g'(y_3) = -\frac{y_3 - x_3}{c|\mathbf{x} - y_3 \mathbf{k}|} - \frac{1}{v}$ and, recalling (5.134),

$$\begin{aligned} |\mathbf{x} - y_3^- \mathbf{k}| |g'(y_3^-)| &= \frac{|c|\mathbf{x} - y_3^- \mathbf{k}| + v(y_3^- - x_3)|}{cv} = \frac{|(vt - y_3^-)c^2/v + v(y_3^- - x_3)|}{cv} \\ &= \frac{c}{v^2} |vt - m^2 x_3 - (1 - m^2) y_3^-| = \frac{1}{v} \sqrt{|\mathbf{x}'|^2 (1 - m^2) + (vt - x_3)^2}. \end{aligned}$$

In conclusion, we obtain

$$u(\mathbf{x}, t) = \frac{S}{4\pi c^2} \frac{1}{\sqrt{|\mathbf{x}'|^2 (1 - m^2) + (vt - x_3)^2}} \quad (5.136)$$

which is defined in $\mathbb{R}^3 \setminus (0, 0, vt)$.

- The *supersonic case* $m > 1$. In this case, the family of spheres $\partial B_{\rho_t}(\mathbf{p})$, centered at $\mathbf{p} = (0, 0, y_3)$, with radius $\rho_t = c/v(vt - y_3)$, $y_3 < vt$, has an envelope, given by the circular cone Γ around the x_3 axis, with vertex at $(0, 0, vt)$ and opening $\theta = \sin^{-1}(c/v)$. Thus, the equation of Γ , known as the **Mach cone**, is

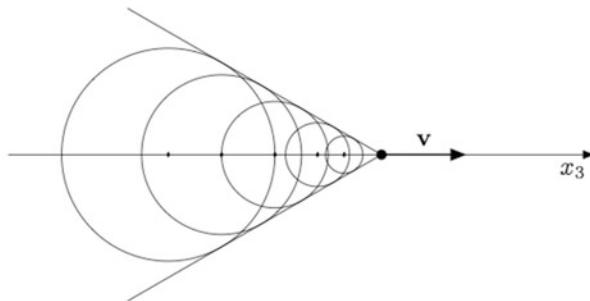
$$vt - x_3 = \sqrt{(m^2 - 1)} |\mathbf{x}'|, \quad (5.137)$$

since $\cot \theta = \sqrt{m^2 - 1}$.

The radiation effect is described in Fig. 5.12. Only the points inside Γ can be reached by the perturbation. Hence $u = 0$ outside Γ . A similar configuration obtains when a jet is flying at a supersonic speed.

To compute the “integral” in (5.133), we observe that both roots y_3^\pm given in (5.135) satisfy the condition $vt - y_3^\pm \geq 0$, since both the numerator and the denominator are negative (recall that $x_3 - vt \leq 0$). Thus both radiations from $(0, 0, y_3^+)$ and $(0, 0, y_3^-)$ contributes to the potential u and formula (7.24) gives

$$\delta(g(y_3)) = \frac{\delta(y_3 - y_3^-)}{|g'(y_3^-)|} + \frac{\delta(y_3 - y_3^+)}{|g'(y_3^+)|}.$$

**Fig. 5.12** The supersonic case

However, since

$$\begin{aligned} |\mathbf{x} - y_3^- \mathbf{k}| |g'(y_3^-)| &= |\mathbf{x} - y_3^+ \mathbf{k}| |g'(y_3^+)| \\ &= \frac{1}{v} \sqrt{(vt - x_3)^2 - |\mathbf{x}'|^2 (m^2 - 1)}, \end{aligned}$$

we finally obtain

$$u(\mathbf{x}, t) = \begin{cases} \frac{S}{2\pi c^2} \frac{1}{\sqrt{(vt - x_3)^2 - |\mathbf{x}'|^2 (m^2 - 1)}} & \text{inside } \Gamma \\ 0 & \text{outside } \Gamma \end{cases}. \quad (5.138)$$

Remark 5.19. On Γ the potential u becomes infinite. Thus, if we interpret u as a sound potential ϕ or a condensation s , as in Sect. 5.8.2, near Γ , the small amplitude assumption leading to the wave equation is not anymore consistent and formula is not the right one, for points near Γ .

Remark 5.20. If $S = ec^2/\epsilon$, where e is a particle charge and c is the speed of light, the potential (5.136) is a particular case of the so called (scalar) Lienard-Wiechert potential for a charge moving along a curve in a medium with dielectric constant ϵ .

What still remains to do is to **check rigorously that the potentials u , given in (5.136) and (5.138) satisfies (5.129) in a suitable sense**. This is done in Example 7.39, p. 452, in Chap. 7.

5.10 An Application to Thermoacoustic Tomography

Thermoacoustic (or photoacoustic) tomography (TAT) is a hybrid imaging technique used to visualize the capability of a medium to absorb non-ionizing radiations (of electromagnetic nature) and is largely exploited in several areas of applied sciences, notably to medical diagnostics.

A short electromagnetic impulse, such as a laser pulse of duration of about 20 picoseconds ($= 20 \cdot 10^{-12}$ seconds), is sent through a biological object. Part of the energy is absorbed throughout the body and the amount of absorption in a given part of it depends on the biological properties of that part. For instance, many tumoral cells absorb much more electromagnetic energy (in a proper energy band) than healthy cells. In this case, the knowledge of the *absorption map* of the body is of invaluable importance in cancer detection.

The optical pulse radiation causes tissues heating (in the range of millikelvin) with thermoelastic expansion/contraction, which in turn results in the generation of propagating ultrasound (pressure) waves (the photoacoustic effect). The pressure waves can be measured by means of transducers placed on a surface S surrounding the body. The goal is *to reconstruct the absorption map, or something related to, from these measurements*.

From a mathematical point of view, the reconstruction problem can be formulated as a so called *inverse problem*, typically *ill posed*. Several techniques have been employed to solve it, under suitable hypotheses³¹. Here we briefly sketch one of the most simple mathematical models, and describe the *time reversal method*, based, in dimension $n = 3$, on the strong Huygens principle and on the invariance of the wave equation under a time reversal. Indeed, throughout this section we shall work in dimension $n = 3$.

The mathematical model for the pressure

As usual we make use of general conservation laws and of suitable constitutive relations. The biological object is assimilated to an inviscid fluid, homogeneous and isotropic with respect to acoustic waves propagation³². We denote by $\rho(\mathbf{x}, t)$ and $p(\mathbf{x}, t)$ the density and the acoustic pressure of the medium, respectively.

We adopt the following assumptions:

- External forces are negligible.
- The velocity \mathbf{v} of the “fluid” is very small.
- The variations of pressure and density are very small:

$$\begin{aligned} p(\mathbf{x}, t) &= p_0 + \delta^* p(\mathbf{x}, t) \\ \rho(\mathbf{x}, t) &= \rho_0 + \delta^* \rho(\mathbf{x}, t) \end{aligned}$$

with $|\delta^* p| \ll 1$ and $|\delta^* \rho| \ll 1$. The general laws are: the *linearized mass conservation*

$$\rho_t + \operatorname{div}(\rho \mathbf{v}) \simeq \rho_t + \rho_0 \operatorname{div} \mathbf{v} = 0$$

³¹ We refer for instance to *P. Kuchment and L Kunyansky*, Mathematics of thermoacoustic tomography, European J. Appl. Math. 19, 191–224, 2008.

³² See [28], *O. Scherzer et al.*, Variational Methods in Imaging, Springer, 2008.

and the *linearized Euler equation*

$$\rho_0 \mathbf{v}_t \simeq \rho \mathbf{v}_t + (\mathbf{v} \cdot \nabla) \mathbf{v} = -\nabla p.$$

Combining the two laws, we get

$$\rho_{tt} - \Delta p = 0. \quad (5.139)$$

We now need to choose how to model:

1. The absorption of electromagnetic power.
 2. The relation between absorption and temperature variation.
 3. The effect of temperature variations on the thermal expansion of the absorbing tissue.
1. The absorption r is modelled by the heat source function (energy per unit time)

$$r(\mathbf{x}, t) = I_{em}(\mathbf{x}, t) \mu_{abs}(\mathbf{x})$$

where μ_{abs} is the *absorption map* and I_{em} is the *radiation intensity*. Given its impulsive nature, we may write

$$I_{em}(\mathbf{x}, t) \simeq J(\mathbf{x}) \delta(t) \quad (5.140a)$$

for a proper spatial intensity distribution $J(\mathbf{x})$. Thus we think that before time $t = 0$ nothing happens ($p \equiv 0$) and that the pulse is concentrated at time $t = 0$.

2. Heat conduction effects are negligible. Hence, we have

$$C(\mathbf{x}) T_t = r(\mathbf{x}, t),$$

where $C(\mathbf{x})$ is the specific heat at constant volume and T is the temperature.

3. We adopt the *linearized law of thermal expansion*, given by:

$$\beta(\mathbf{x}) T_t = \frac{1}{v_s^2} p_t - \rho_t$$

where β is the coefficient of *thermal expansion* and v_s is the sound speed (which we assume to be constant).

From 1, 2 and 3 we have:

$$\frac{1}{v_s^2} p_t - \rho_t = \frac{J(\mathbf{x}) \mu_{abs}(\mathbf{x}) \beta(\mathbf{x})}{C(\mathbf{x})} \delta(t) \equiv f(\mathbf{x}) \delta(t) \quad (5.141)$$

where f is called *energy deposition function*.

Taking the t -derivative and using (5.139), we formally obtain the equation

$$p_{tt} - v_s^2 \Delta p = v_s^2 f(\mathbf{x}) \delta'(t), \quad \mathbf{x} \in \mathbb{R}^n, t \in \mathbb{R} \quad (5.142)$$

where we think $p \equiv 0$ for $t < 0$. However, the meaning of $\delta'(t)$ has to be clarified.

As in Subsect. 2.3.3, we can define δ' through its action on test functions. Thus, take a smooth function φ vanishing outside a compact interval and write, after a formal integration by parts:

$$\int \delta'(t) \varphi(t) dt = - \int \delta(t) \varphi'(t) dt = -\varphi'(0). \quad (5.143)$$

Indeed, as we will see in Chap. 7, equation (5.143) gives the true meaning of δ' .

Still it is not clear how to solve (5.142). Fortunately, equation (5.142) has an equivalent formulation in more standard terms. Indeed it is possible to prove (see Problem 7.18) that p is a solution of (5.142) if and only if p solves the problem

$$\begin{cases} p_{tt} - \Delta p = 0 & \mathbf{x} \in \mathbb{R}^n, t > 0 \\ p(\mathbf{x}, 0) = v_s^2 f(\mathbf{x}), \quad p_t(\mathbf{x}, 0) = 0 & \mathbf{x} \in \mathbb{R}^n \end{cases}. \quad (5.144)$$

This is our model for the ultrasound waves.

Remark 5.21. As we shall see in Chap. 7, Subsect. 7.5.2, the “product” $J(\mathbf{x}) \delta(t)$ in formula (5.140a), takes the name of **direct** or **tensor product** between $J(\mathbf{x})$ and $\delta(t)$, and it should be more properly written as

$$J(\mathbf{x}) \otimes \delta(t).$$

Similarly, the two direct products on the right of (5.141) and (5.142) should be written as $f(\mathbf{x}) \otimes \delta(t)$ and $f(\mathbf{x}) \otimes \delta'(t)$, respectively. The symbol \otimes emphasizes that the two factors in the tensor product act on a test function $\psi(\mathbf{x}, t)$ separately, each one with respect to its own variables.

The inverse problem. The time reversal method

Assume now we are able to measure the pressure on a closed (nice) surface S surrounding the body, for $0 \leq t \leq T$. This means that we know p on the surface $S \times [0, T]$. Thus, supposing that the measured datum is g , we can add this information to our problem and write:

$$\begin{cases} p_{tt} - v_s^2 \Delta p = 0 & \mathbf{x} \in \mathbb{R}^n, t > 0 \\ p(\mathbf{x}, 0) = v_s^2 f(\mathbf{x}), \quad p_t(\mathbf{x}, 0) = 0 & \mathbf{x} \in \mathbb{R}^n \\ p(\boldsymbol{\sigma}, t) = g(\boldsymbol{\sigma}, t) & \boldsymbol{\sigma} \in S, \quad 0 \leq t \leq T. \end{cases} \quad (5.145)$$

Our problem is to reconstruct f from the knowledge of g . In fact, from the knowledge of f we can extract information on the absorption map through formula (5.141).³³

At first glance, it seems that our reconstruction problem is unsolvable, since many different initial data give rise to solutions with the same lateral data. Nevertheless, there are two elements that make the problem solvable. First, the wave

³³ See, for instance, L.V. Wang, H. Hu, *Biomedical Optics, Principles and Imaging*. Wiley-Interscience (2007).

equation in (5.145) is valid in the *whole space* \mathbb{R}^3 , not only in the region inside S , and second, f is *compactly supported* inside the body, hence *inside* S .

Now, the idea of the *time reversal method* is simple. Since the Huygens principle holds and f is compactly supported, after a long enough time, the pressure waves leave S . This means that, if T is large enough,

$$p(\mathbf{x}, T) = p_t(\mathbf{x}, T) = 0 \text{ inside } S.$$

But then p is a solution of the *final* problem

$$\begin{cases} p_{tt} - v_s^2 \Delta p = 0 & \text{inside } S \times (0, T) \\ p(\mathbf{x}, T) = p_t(\mathbf{x}, T) = 0 & \text{inside } S \\ p(\boldsymbol{\sigma}, t) = g(\boldsymbol{\sigma}, t) & \boldsymbol{\sigma} \in S, 0 \leq t \leq T. \end{cases} \quad (5.146)$$

Since the wave equation is invariant by time reversal, we know that, under reasonable assumptions on g , problem (5.146) has at most one solution. The existence and, above all, the stability estimates on the solution, follow by the methods we will develop in Chap. 10. Then, via standard numerical methods, one is able to compute $p(\mathbf{x}, 0) = f(\mathbf{x})$ and to solve the inverse problem.

5.11 Linear Water Waves

A great variety of interesting phenomena occurs in the analysis of water waves. Here we briefly analyze *surface water waves*, that is, disturbances of the free surface of an incompressible fluid, resulting from the balance between a restoring force, due to gravity and/or surface tension, and fluid inertia, due to an external action (such as wind, passage of a ship, sub-sea earthquakes). We will focus on the special case of *linear waves*, whose amplitude is small compared to wavelength, analyzing the dispersive relation in the approximation of deep water.

5.11.1 A model for surface waves

We start by deriving a basic model for surface water waves, assuming the following hypotheses.

1. The fluid has *constant density* ρ and *negligible viscosity*. In particular, the force exerted on a control fluid volume V by the rest of the fluid is given by the normal pressure³⁴ $-p\boldsymbol{\nu}$ on ∂V .
2. The atmospheric pressure above the fluid surface is constant (e.g. no wind) and we neglect the small air pressure variations, due to the fluid motion. In this case the fluid surface is called a **free surface**.

³⁴ $\boldsymbol{\nu}$ is the exterior normal unit vector to ∂V .

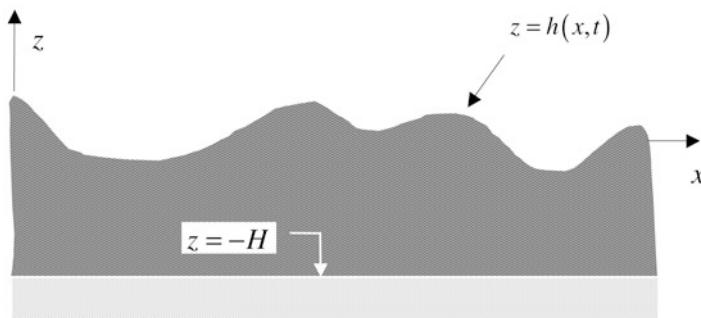


Fig. 5.13 Vertical section of the fluid region

3. The motion is *laminar* (no breaking waves or turbulence) and *two dimensional*. This means that in a suitable coordinate system x, z , where the coordinate x measures horizontal distance and z is a vertical coordinate (Fig. 5.13), we can describe the free surface by a function $z = h(x, t)$, while the velocity vector has the form $\mathbf{w} = u(x, z, t)\mathbf{i} + v(x, z, t)\mathbf{k}$.
4. The motion is **irrotational**, so that *there exists a (smooth) scalar potential* $\phi = \phi(x, z, t)$ such that:

$$\mathbf{w} = \nabla\phi = \phi_x\mathbf{i} + \phi_z\mathbf{k}.$$

The mathematical model governing the waves motion involves the unknowns h and ϕ , together with *initial conditions* and suitable *conditions at the boundary* of our relevant domain, composed by the free surface, the lower boundary and the lateral sides.

We assume that the side boundaries are so far away that their influence can be neglected. Therefore x varies all along the real axis.

Furthermore, we assume, for simplicity, that the lower boundary is flat, at the level $z = -H$.

Two equations for h and ϕ come from conservation of mass and balance of linear momentum, taking into account hypotheses 1 to 4 above.

The *mass conservation* gives:

$$\operatorname{div} \mathbf{w} = \Delta\phi = 0, \quad x \in \mathbb{R}, \quad -H < z < h(x, t). \quad (5.147)$$

Thus, ϕ is a harmonic function.

The *balance of linear momentum* yields:

$$\mathbf{w}_t + (\mathbf{w} \cdot \nabla)\mathbf{w} = \mathbf{g} - \frac{1}{\rho}\nabla p \quad (5.148)$$

where \mathbf{g} is the gravitational acceleration.

Let us rewrite (5.148) in terms of the potential ϕ . From the identity

$$\mathbf{w} \times \operatorname{curl} \mathbf{w} = \frac{1}{2} \nabla(|\mathbf{w}|^2) - (\mathbf{w} \cdot \nabla) \mathbf{w},$$

since $\operatorname{curl} \mathbf{w} = \mathbf{0}$, we obtain

$$(\mathbf{w} \cdot \nabla) \mathbf{w} = \frac{1}{2} \nabla(|\nabla \phi|^2).$$

Moreover, writing $\mathbf{g} = \nabla(-gz)$, (5.148) becomes

$$\frac{\partial}{\partial t}(\nabla \phi) + \frac{1}{2} \nabla(|\nabla \phi|^2) = -\frac{1}{\rho} \nabla p + \nabla(-gz)$$

or

$$\nabla \left\{ \phi_t + \frac{1}{2} |\nabla \phi|^2 + \frac{p}{\rho} + gz \right\} = 0.$$

As a consequence

$$\phi_t + \frac{1}{2} |\nabla \phi|^2 + \frac{p}{\rho} + gz = C(t)$$

with $C = C(t)$ is an arbitrary function. Since ϕ is uniquely defined up to an additive function of time, we can choose $C(t) = 0$ by adding to ϕ the function $\int_0^t C(s) ds$.

In this case, we obtain **Bernoulli's equation**

$$\phi_t + \frac{1}{2} |\nabla \phi|^2 + \frac{p}{\rho} + gz = 0. \quad (5.149)$$

We consider now the boundary conditions. We impose the so called **bed condition**, according to which the normal component of the velocity vanishes on the bottom; therefore

$$\phi_z(x, -H, t) = 0, \quad x \in \mathbb{R}. \quad (5.150)$$

More delicate is the condition to be prescribed on the free surface $z = h(x, t)$. In fact, since this surface is itself an unknown of the problem, we actually need *two conditions* on it.

The first one comes from Bernoulli's equation. Namely, the total pressure on the free surface is given by

$$p = p_{at} - \sigma h_{xx} \{1 + h_x^2\}^{-3/2}. \quad (5.151)$$

In (5.151) the term p_{at} is the atmospheric pressure, that we can take equal to zero, while the second term is due to the *surface tension*, as we will shortly see below.

Thus, inserting $z = h(x, t)$ and (5.151) into (5.149), we obtain the following **dynamic condition at the free surface**:

$$\phi_t + \frac{1}{2} |\nabla \phi|^2 - \frac{\sigma h_{xx}}{\rho \{1 + h_x^2\}^{3/2}} + gh = 0, \quad x \in \mathbb{R}, z = h(x, t). \quad (5.152)$$

A second condition follows imposing that the fluid particles on the free surface always remain there. If the particle path is described by the equations $x = x(t)$,

$z = z(t)$, this amounts to requiring that

$$z(t) - h(x(t), t) \equiv 0.$$

Differentiating the last equation yields

$$\dot{z}(t) - h_x(x(t), t) \dot{x}(t) - h_t(x(t), t) = 0$$

that is, since $\dot{x}(t) = \phi_x(x(t), z(t), t)$ and $\dot{z} = \phi_z(x(t), z(t), t)$,

$$\phi_z - h_t - \phi_x h_x = 0, \quad x \in \mathbb{R}, z = h(x, t), \quad (5.153)$$

which is known as the **kinematic condition at the free surface**.

Finally, we require a reasonable behavior of ϕ and h as $x \rightarrow \pm\infty$, for instance

$$\int_{\mathbb{R}} |\phi| < \infty, \quad \int_{\mathbb{R}} |h| < \infty \quad \text{and} \quad \phi, h \rightarrow 0 \quad \text{as } x \rightarrow \pm\infty. \quad (5.154)$$

Equation (5.147) and the boundary conditions (5.150), (5.152), (5.153) constitute our model for water waves. After a brief justification of formula (5.151), in the next subsection we go back to the above model, deriving a dimensionless formulation and a linearized version of it.

- *Effect of surface tension.* In a water molecule the two hydrogen atoms take an asymmetric position with respect to the oxygen atom. This asymmetric structure generates an electric dipole moment. Inside a bulk of water these moments balance, but on the surface they tend to be parallel and create a macroscopic inter-molecular force per unit length, confined to the surface, called *surface tension*.

The way this force manifests itself is similar to the action exerted on a small portion of an elastic material by the surrounding material and described by a *stress vector*, which is a force per unit area, on the boundary of the portion. Similarly, consider a small region on the water surface, delimited by a closed curve γ . The surface water on one side of γ exerts on the water on the other side a (*pulling*) *force per unit length* \mathbf{f} along γ .

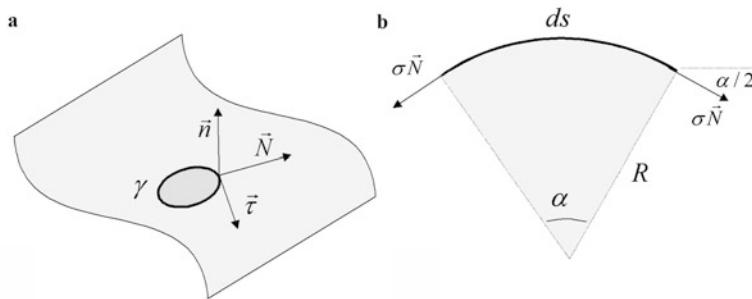
Let \mathbf{n} be a unit vector normal to the water surface and $\boldsymbol{\tau}$ be a unit tangent vector to γ (Fig. 5.14a) such that $\mathbf{N} = \boldsymbol{\tau} \times \mathbf{n}$ points outwards the region bounded by γ . A simple constitutive law for \mathbf{f} is

$$\mathbf{f}(\mathbf{x}, t) = \sigma(\mathbf{x}, t) \mathbf{N}(\mathbf{x}, t), \quad \mathbf{x} \in \gamma.$$

Thus, \mathbf{f} acts in the direction of \mathbf{N} ; its magnitude σ , independent of \mathbf{N} , is called **surface tension**.

Formula (5.151) is obtained by balancing the net vertical component of the force produced by surface tension with the difference of the pressure force across the surface.

Consider the section ds of a small surface element shown in Fig. 5.14b. A surface tension of magnitude σ acts tangentially at both ends. Up to higher order terms, the downward vertical component is given by $2\sigma \sin(\alpha/2)$. On the other hand, this force is equal to $(p_{at} - p)ds$ where p is the fluid pressure beneath the

Fig. 5.14 Surface tension $\sigma\mathbf{N}$

surface. Thus,

$$(p_{at} - p)ds = \pm 2\sigma \sin(\alpha/2),$$

where the “+” sign (“−” sign) corresponds to a *convex* (*concave*) surface element. Since we have $ds = R\alpha$ and, for small α , $2\sin(\alpha/2) \approx \alpha$, we may write

$$p_{at} - p \approx \pm \frac{\sigma}{R} = \sigma\kappa \quad (5.155)$$

where κ is the **curvature of the surface**. If the atmospheric pressure prevails, the curvature is positive and the surface is convex, otherwise the curvature is negative and the surface is concave, as in Fig. 5.14b. If the surface is described by $z = h(x, t)$, we have

$$\kappa = \frac{h_{xx}}{\{1 + h_x^2\}^{3/2}}$$

which, inserted into (5.155), gives (5.151).

5.11.2 Dimensionless formulation and linearization

The nonlinearities in (5.152), (5.153) and the fact that the free surface itself is an unknown of the problem make the above model quite difficult to analyze by elementary means. However, if we restrict our considerations to waves whose amplitude is much smaller than their wavelength, then both difficulties disappear. In spite of this simplification, the resulting theory has a rather wide range of applications, since it is not rare to observe waves with amplitude from 1 to 2 meters and a wavelength of up to a kilometer or more.

To perform a correct linearization procedure, we first introduce *dimensionless* variables. Denote by L , A and T , an average *wavelength*, *amplitude* and *period*³⁵, respectively. Set

$$\tau = \frac{t}{T}, \quad \xi = \frac{x}{L}, \quad \eta = \frac{z}{L}.$$

³⁵ That is, the time a crest takes to travel a distance of order L .

Since the dimensions of h and ϕ are, respectively [length] and $[length]^2 \times [time]^{-1}$, we may rescale ϕ and h by setting:

$$\begin{aligned}\Phi(\xi, \eta, \tau) &= \frac{T}{LA} \phi(L\xi, L\eta, T\tau), \\ \Gamma(\xi, \tau) &= \frac{1}{A} h(L\xi, L\eta, T\tau).\end{aligned}$$

In terms of these dimensionless variables, our model becomes, after elementary calculations:

$$\begin{aligned}\Delta\Phi = 0, & \quad -H_0 < \eta < \varepsilon\Gamma(\xi, \tau), \quad \xi \in \mathbb{R} \\ \Phi_\tau + \frac{\varepsilon}{2} |\nabla\Phi|^2 + \mathcal{F} \left\{ \Gamma - \mathcal{B}\Gamma_{\xi\xi} \left\{ 1 + \varepsilon^2\Gamma_\xi^2 \right\}^{3/2} \right\} = 0, & \quad \eta = \varepsilon\Gamma(\xi, \tau), \quad \xi \in \mathbb{R} \\ \Phi_\eta - \Gamma_\tau - \varepsilon\Phi_\xi\Gamma_\xi = 0 & \quad \eta = \varepsilon\Gamma(\xi, \tau), \quad \xi \in \mathbb{R} \\ \Phi_\eta(\xi, -H_0, \tau) = 0, & \quad \xi \in \mathbb{R},\end{aligned}$$

where we have emphasized the four dimensionless combinations³⁶

$$\varepsilon = \frac{A}{L}, \quad H_0 = \frac{H}{L}, \quad \mathcal{F} = \frac{gT^2}{L}, \quad \mathcal{B} = \frac{\sigma}{\rho g L^2}. \quad (5.156)$$

The parameter \mathcal{B} , called *Bond number*, measures the relevance of surface tension while \mathcal{F} , the *Froude number*, measures the relevance of gravity.

At this point, the assumption of *small amplitude compared to the wavelength*, translates simply into

$$\varepsilon = \frac{A}{L} \ll 1$$

and the linearization of the above system is achieved by letting $\varepsilon = 0$:

$$\begin{aligned}\Delta\Phi = 0, & \quad -H_0 < \eta < 0, \quad \xi \in \mathbb{R} \\ \Phi_\tau + \mathcal{F} \{ \Gamma - \mathcal{B}\Gamma_{\xi\xi} \} = 0, & \quad \eta = 0, \quad \xi \in \mathbb{R} \\ \Phi_\eta - \Gamma_\tau = 0, & \quad \eta = 0, \quad \xi \in \mathbb{R} \\ \Phi_\eta(\xi, -H_0, \tau) = 0, & \quad \xi \in \mathbb{R}.\end{aligned}$$

Going back to the original variables, we finally obtain the linearized system

$$\begin{cases} \Delta\phi = 0, & -H < z < 0, \quad x \in \mathbb{R} & (\text{Laplace}) \\ \phi_t + gh - \frac{\sigma}{\rho}h_{xx} = 0, & z = 0, \quad x \in \mathbb{R} & (\text{Bernoulli}) \\ \phi_z - h_t = 0, & z = 0, \quad x \in \mathbb{R} & (\text{kinematic}) \\ \phi_z(x, -H, t) = 0, & x \in \mathbb{R} & (\text{bed condition}). \end{cases} \quad (5.157)$$

³⁶ Note the reduction of the number of relevant parameters from seven (A , L , T , H , g , σ , ρ) to four.

It is possible to obtain an equation for ϕ only. Differentiate twice with respect to x the kinematic equation and use $\phi_{xx} = -\phi_{zz}$; this yields

$$h_{txx} = \phi_{zxx} = -\phi_{zzz}. \quad (5.158)$$

Differentiate Bernoulli's equation with respect to t , then use $h_t = \phi_z$ and (5.158). The result is:

$$\phi_{tt} + g\phi_z + \frac{\sigma}{\rho}\phi_{zzz} = 0, \quad z = 0, x \in \mathbb{R}. \quad (5.159)$$

5.11.3 Deep water waves

We solve now system (5.157) with the following initial conditions:

$$\phi(x, z, 0) = 0, \quad h(x, 0) = h_0(x), \quad h_t(x, 0) = 0. \quad (5.160)$$

Thus, initially ($t = 0$) the fluid velocity is zero and the free surface has been perturbed into a nonhorizontal profile h_0 , that we assume (for simplicity) *smooth, even* (i.e. $h_0(-x) = h_0(x)$) and *compactly supported*. In addition we consider the case of *deep water* ($H \gg 1$) so that the bed condition can be replaced by³⁷

$$\phi_z(x, z, t) \rightarrow 0 \quad \text{as } z \rightarrow -\infty. \quad (5.161)$$

The resulting initial-boundary value problem is not of the type we considered so far, but we are reasonably confident that it is well posed. Since x varies over all the real axis, we may use the Fourier transform with respect to x , setting

$$\hat{\phi}(k, z, t) = \int_{\mathbb{R}} e^{-ikx} \phi(x, z, t) dx, \quad \hat{h}(k, t) = \int_{\mathbb{R}} e^{-ikx} h(x, t) dx.$$

Note that, the assumptions on h_0 implies that $\hat{h}_0(k) = \hat{h}(k, 0)$ rapidly vanishes as $|k| \rightarrow \infty$ and $\hat{h}_0(-k) = \hat{h}_0(k)$. Moreover, since $\hat{\phi}_{xx} = -k^2 \hat{\phi}$, the Laplace equation transforms into the ordinary differential equation

$$\hat{\phi}_{zz} - k^2 \hat{\phi} = 0,$$

whose general solution is

$$\hat{\phi}(k, z, t) = A(k, t) e^{|k|z} + B(k, t) e^{-|k|z}. \quad (5.162)$$

From (5.161) we deduce $B(k, t) = 0$, so that

$$\hat{\phi}(k, z, t) = A(k, t) e^{|k|z}. \quad (5.163)$$

³⁷ For the case of finite depth see Problem 5.19.

Transforming (5.159) we get

$$\hat{\phi}_{tt} + g\hat{\phi}_z + \frac{\sigma}{\rho}\hat{\phi}_{zzz} = 0, \quad z = 0, k \in \mathbb{R}$$

and (5.163) yields for A the equation

$$A_{tt} + \left(g|k| + \frac{\sigma}{\rho}|k|^3 \right) A = 0.$$

Thus, we obtain

$$A(k, t) = a(k)e^{i\omega(k)t} + b(k)e^{-i\omega(k)t}$$

where (**dispersion relation**)

$$\omega(k) = \sqrt{g|k| + \frac{\sigma}{\rho}|k|^3},$$

and

$$\hat{\phi}(k, z, t) = \left\{ a(k)e^{i\omega(k)t} + b(k)e^{-i\omega(k)t} \right\} e^{|k|z}.$$

To determine $a(k)$ and $b(k)$, observe that the Bernoulli condition gives

$$\hat{\phi}_t(k, 0, t) + \left\{ g + \frac{\sigma}{\rho}k^2 \right\} \hat{h}(k, t) = 0, \quad k \in \mathbb{R} \quad (5.164)$$

from which

$$i\omega(k) \left\{ a(k)e^{i\omega(k)t} - b(k)e^{-i\omega(k)t} \right\} + \left(g + \frac{\sigma}{\rho}k^2 \right) \hat{h}(k, t) = 0, \quad k \in \mathbb{R}$$

and for $t = 0$

$$i\omega(k) \{a(k) - b(k)\} + \left(g + \frac{\sigma}{\rho}k^2 \right) \hat{h}_0(k) = 0. \quad (5.165)$$

Similarly, the kinematic condition gives

$$\hat{\phi}_z(k, 0, t) + \hat{h}_t(k, t) = 0, \quad k \in \mathbb{R}. \quad (5.166)$$

From (5.162) we have

$$\hat{\phi}_z(k, 0, t) = |k| \left\{ a(k)e^{i\omega(k)t} + b(k)e^{-i\omega(k)t} \right\}$$

and since $\hat{h}_t(k, 0) = 0$, we get, from (5.166) for $t = 0$ and $k \neq 0$,

$$a(k) + b(k) = 0. \quad (5.167)$$

From (5.165) and (5.167) we have ($k \neq 0$)

$$a(k) = -b(k) = \frac{i \left(g + \frac{\sigma}{\rho}k^2 \right)}{2\omega(k)} \hat{h}_0(k)$$

and therefore

$$\hat{\phi}(k, y, t) = \frac{i \left(g + \frac{\sigma}{\rho} k^2 \right)}{2\omega(k)} \left\{ e^{i\omega(k)t} - e^{-i\omega(k)t} \right\} e^{|k|z} \hat{h}_0(k).$$

From (5.164) we deduce:

$$\hat{h}(k, t) = \left(g + \frac{\sigma}{\rho} k^2 \right)^{-1} \hat{\phi}_t(k, 0, t) = \frac{1}{2} \left\{ e^{i\omega(k)t} + e^{-i\omega(k)t} \right\} \hat{h}_0(k)$$

and finally, transforming back³⁸

$$h(x, t) = \frac{1}{4\pi} \int_{\mathbb{R}} \left\{ e^{i(kx - \omega(k)t)} + e^{i(kx + \omega(k)t)} \right\} \hat{h}_0(k) dk. \quad (5.168)$$

5.11.4 Interpretation of the solution

We now examine some remarkable features of the solution (5.168). The surface displacement appears in **wave packet** form. The dispersion relation

$$\omega(k) = \sqrt{g|k| + \frac{\sigma}{\rho}|k|^3}$$

shows that each Fourier component of the initial free surface propagates both in the positive and negative x -directions. The phase and group velocities are (considering only $k > 0$, for simplicity)

$$c_p(k) = \frac{\omega(k)}{k} = \sqrt{\frac{g}{k} + \frac{\sigma k}{\rho}} \quad \text{and} \quad c_g(k) = \omega'(k) = \frac{g + 3\sigma k^2/\rho}{2\sqrt{gk + \sigma k^3/\rho}}.$$

Thus, we see that the speed of a wave of wavelength $\lambda = 2\pi/k$ **depends on its wavelength**. The fundamental parameter is

$$B^* = 4\pi^2 \mathcal{B} = \frac{\sigma k^2}{\rho g}$$

where \mathcal{B} is the Bond number. For water, under “normal” conditions,

$$\rho = 1 \text{ gr/cm}^3, \sigma = 72 \text{ gr/sec}^2, g = 980 \text{ cm/sec}^2 \quad (5.169)$$

so that $B^* = 1$ for wavelengths $\lambda \simeq 1.7$ cm. When $\lambda \gg 1.7$ cm, then $B^* < 1$, $k = \frac{2\pi}{\lambda} \ll 1$ and surface tension becomes negligible. This is the case of **gravity waves** (generated e.g. by dropping a stone into a pond) whose phase speed and

³⁸ Note that, since also $\omega(k)$ is even, we may write

$$h(x, t) = \frac{1}{2\pi} \int_{\mathbb{R}} \cos[kx - \omega(k)t] \hat{h}_0(k) dk.$$

group velocity are well approximated by

$$c_p \simeq \sqrt{\frac{g}{k}} = \sqrt{\frac{g\lambda}{2\pi}} \quad \text{and} \quad c_g \simeq \frac{1}{2}\sqrt{\frac{g}{k}} \simeq \frac{1}{2}c_p.$$

Thus, **longer waves move faster** and **energy is slower than the crests**.

On the other hand, if $\lambda \ll 1.7$ cm, then $B^* > 1$, $k = \frac{2\pi}{\lambda} \gg 1$ and this time surface tension prevails over gravity. In fact, short wavelengths are associated with relative high curvature of the free surface and high curvature is concomitant with large surface tension effects. This is the case of **capillarity waves** (generated e.g. by raindrops in a pond) and their speed and group velocity are well approximated by

$$c_p \simeq \sqrt{\frac{\sigma k}{\rho}} = \sqrt{\frac{2\pi\sigma}{\lambda\rho}} \quad \text{and} \quad c_g \simeq \frac{3}{2}\sqrt{\frac{\sigma k}{\rho}} \simeq \frac{3}{2}c_p.$$

Thus **shorter waves move faster** and **energy is faster than the crests**.

When both gravity and surface tension are relevant, Fig. 5.15 shows the graph of c_p^2 versus λ , for water, with the values (5.169):

$$c_p^2 = 156.97 \lambda + \frac{452.39}{\lambda}.$$

The main feature of this graph is the presence of the minimum

$$c_{\min} = 23 \text{ cm/sec}$$

corresponding just to the value $\lambda = 1.7$ cm. The consequence is curious: linear gravity and capillarity deep water waves can appear simultaneously only when the speed is greater than 23 cm/sec. A typical situation occurs when a small obstacle (e.g. a twig) moves at speed v in still water. The motion of the twig results in the

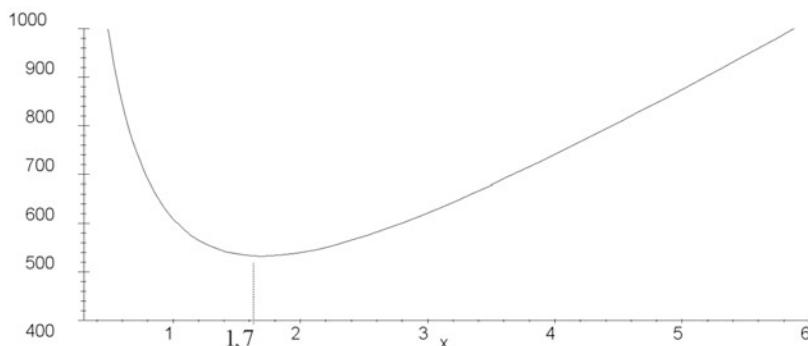


Fig. 5.15 c_p^2 versus λ

formation of a wave system that moves along with it, with gravity waves behind and capillarity waves ahead. In fact, the result above shows that this wave system can actually appear only if $v > 23$ cm/sec.

5.11.5 Asymptotic behavior

As we have already observed, the behavior of a wave packet is dominated for short times by the initial conditions and only after a relatively long time it is possible to observe the intrinsic features of the perturbation. For this reason, information about the asymptotic behavior of the packet as $t \rightarrow +\infty$ are important. Thus, we need a good asymptotic formula for the integral in (5.168) when $t \gg 1$.

For simplicity, we consider gravity waves only, for which

$$\omega(k) \simeq \sqrt{g|k|}.$$

Let us follow a particle $x = x(t)$ moving along the positive x -direction with constant speed $v > 0$, so that $x = vt$. Inserting $x = vt$ into (5.168) we find

$$\begin{aligned} h(vt, t) &= \frac{1}{4\pi} \int_{\mathbb{R}} e^{it(kv - \omega(k))} \hat{h}_0(k) dk + \frac{1}{4\pi} \int_{\mathbb{R}} e^{it(kv + \omega(k))} \hat{h}_0(k) dk \\ &\equiv h_1(vt, t) + h_2(vt, t). \end{aligned}$$

According to Theorem 5.22 in the next subsection (see also Remark 5.24, p. 341), with

$$\varphi(k) = kv - \omega(k),$$

if there exists exactly one stationary point for φ , i.e. only one point k_0 such that

$$\omega'(k_0) = v \quad \text{and} \quad \varphi''(k_0) = -\omega''(k_0) \neq 0,$$

we may estimate h_1 , for $t \gg 1$, by the following formula:

$$h_1(vt, t) = \frac{A(k_0)}{\sqrt{t}} \exp\{it[k_0v - \omega(k_0)]\} + O(t^{-1}) \quad (5.170)$$

where

$$A(k_0) = \hat{h}_0(k_0) \sqrt{\frac{1}{8\pi|\omega''(k_0)|}} \exp i\left\{-\frac{\pi}{4}\operatorname{sign} \omega''(k_0)\right\}.$$

We have ($k \neq 0$),

$$\omega'(k) = \frac{1}{2}\sqrt{g}|k|^{-1/2} \operatorname{sign}(k)$$

and

$$\omega''(k) = -\frac{\sqrt{g}}{4}|k|^{-3/2}.$$

Since $v > 0$, the equation $\omega'(k_0) = v$ gives the unique *point of stationary phase*

$$k_0 = \frac{g}{4v^2} = \frac{gt^2}{4x^2}.$$

Moreover,

$$k_0 v - \omega(k_0) = -\frac{g}{4v} = -\frac{gt}{4x}$$

and

$$\omega''(k_0) = -\frac{2v^3}{g} = -\frac{2x^3}{gt^3} < 0.$$

Hence, from (5.170), we find

$$h_1(vt, t) = \frac{1}{4} \hat{h}_0\left(\frac{g}{4v^2}\right) \sqrt{\frac{g}{\pi tv^3}} \exp i\left\{-\frac{gt}{4v} + \frac{\pi}{4}\right\} + O(t^{-1}).$$

Similarly, since

$$\hat{h}_0(k_0) = \hat{h}_0(-k_0),$$

we find

$$h_2(vt, t) = \frac{1}{4} \hat{h}_0\left(\frac{g}{4v^2}\right) \sqrt{\frac{g}{\pi tv^3}} \exp i\left\{\frac{gt}{4v} - \frac{\pi}{4}\right\} + O(t^{-1}).$$

Finally,

$$\begin{aligned} h(vt, t) &= h_1(vt, t) + h_2(vt, t) \\ &= \hat{h}_0\left(\frac{g}{4v^2}\right) \sqrt{\frac{g}{4\pi v^3 t}} \cos\left\{\frac{gt}{4v} - \frac{\pi}{4}\right\} + O(t^{-1}). \end{aligned}$$

This formula shows that, for large x and t , with $x/t = v$, constant, the wave packet is locally sinusoidal with wave number

$$k(x, t) = \frac{gt}{4vx} = \frac{gt^2}{4x^2}.$$

In other words, an observer moving at the constant speed $v = x/t$ sees a dominant wavelength $2\pi/k_0$, where k_0 is the solution of $\omega'(k_0) = x/t$. The amplitude decreases as $t^{-1/2}$. This is due to the dispersion of the various Fourier components of the initial configuration, after a sufficiently long time.

5.11.6 The method of stationary phase

The *method of stationary phase*, essentially due to Laplace, gives an asymptotic formula for integrals of the form

$$I(t) = \int_a^b f(k) e^{it\varphi(k)} dk \quad (-\infty \leq a < b \leq \infty)$$

as $t \rightarrow +\infty$. Actually, only the real part of $I(t)$, in which the factor $\cos[t\varphi(k)]$ appears, is of interest. Now, as t is very large and $\varphi(k)$ varies, we expect that $\cos[t\varphi(k)]$ oscillates eventually much more than f . For this reason, the contributions of the intervals where $\cos[t\varphi(k)] > 0$ will balance those in which $\cos[t\varphi(k)] < 0$, so that we expect that $I(t) \rightarrow 0$ as $t \rightarrow +\infty$, just as the Fourier coefficients of an integrable function tend to zero as the frequency goes to infinity.

To obtain information on the vanishing speed, assume φ is constant on a certain interval J . On this interval $\cos[t\varphi(k)]$ is constant as well and hence there are neither oscillations nor cancellations. Thus, it is reasonable that, for $t \gg 1$, the relevant contributions to $I(t)$ come from intervals where φ is constant or at least almost constant. The same argument suggests that eventually, a however small interval, containing a stationary point k_0 for φ , will contribute to the integral much more than any other interval without stationary points.

The method of stationary phase makes the above argument precise through the following theorem.

Theorem 5.22. Let f and φ belong to $C^2([a, b])$. Assume that

$$\varphi'(k_0) = 0, \varphi''(k_0) \neq 0 \quad \text{and} \quad \varphi'(k) \neq 0 \text{ for } k \neq k_0.$$

Then, as $t \rightarrow +\infty$

$$\int_a^b f(k) e^{it\varphi(k)} dk = \sqrt{\frac{2\pi}{|\varphi''(k_0)|}} \frac{f(k_0)}{\sqrt{t}} \exp \left\{ i \left[t\varphi(k_0) + \frac{\pi}{4} \operatorname{sign} \varphi''(k_0) \right] \right\} + O(t^{-1}).$$

First a lemma.

Lemma 5.23. Let f, φ as in Theorem 5.22. Let $[c, d] \subseteq [a, b]$ and assume that $|\varphi'(k)| \geq C > 0$ in $[c, d]$. Then

$$\int_c^d f(k) e^{it\varphi(k)} dk = O(t^{-1}) \quad t \rightarrow +\infty. \quad (5.171)$$

Proof. Integrating by parts we get (multiplying and dividing by φ'):

$$\int_c^d \frac{f}{\varphi'} \varphi' e^{it\varphi} dk = \frac{1}{it} \left\{ \frac{f(d) e^{it\varphi(d)}}{\varphi'(d)} - \frac{f(c) e^{it\varphi(c)}}{\varphi'(c)} - \int_c^d \frac{f' \varphi' - f \varphi''}{(\varphi')^2} e^{it\varphi} dk \right\}.$$

Thus, from $|e^{it\varphi(k)}| \leq 1$ and our hypotheses, we have

$$\left| \int_c^d f e^{it\varphi} dk \right| \leq \frac{1}{Ct} \left\{ |f(d)| + |f(c)| + \frac{1}{C} \int_c^d |f'\varphi' - f\varphi''| dk \right\} \leq \frac{K}{t}$$

which gives (5.171). \square

Proof of Theorem 5.22. Without loss of generality, we may assume $k_0 = 0$, so that $\varphi'(0) = 0$, $\varphi''(0) \neq 0$. From Lemma 5.23, it is enough to consider the integral

$$\int_{-\varepsilon}^{\varepsilon} f(k) e^{it\varphi(k)} dk$$

where $\varepsilon > 0$ is as small as we wish. We give the proof in the case φ is a quadratic polynomial, that is

$$\varphi(k) = \varphi(0) + Ak^2, \quad A = \frac{1}{2}\varphi''(0).$$

The other case can be reduced to this one by a suitable change of variables. Write

$$f(k) = f(0) + \frac{f(k) - f(0)}{k} k \equiv f(0) + q(k) k,$$

and observe that, since $f \in C^2([-\varepsilon, \varepsilon])$, $q'(k)$ is bounded in $[-\varepsilon, \varepsilon]$. Then, we have:

$$\int_{-\varepsilon}^{\varepsilon} f(k) e^{it\varphi(k)} dk = 2f(0)e^{it\varphi(0)} \int_0^{\varepsilon} e^{itAk^2} dk + e^{it\varphi(0)} \int_{-\varepsilon}^{\varepsilon} q(k) ke^{itAk^2} dk.$$

Now, an integration by parts shows that the second integral is $O(1/t)$ as $t \rightarrow \infty$ (the reader should check the details). In the first integral, if $A > 0$, let $tAk^2 = y^2$. Then

$$\int_0^{\varepsilon} e^{itAk^2} dk = \frac{1}{\sqrt{tA}} \int_0^{\varepsilon\sqrt{tA}} e^{iy^2} dy.$$

Since³⁹

$$\int_0^{\varepsilon\sqrt{tA}} e^{iy^2} dy = \frac{\sqrt{\pi}}{2} e^{i\frac{\pi}{4}} + O\left(\frac{1}{\varepsilon\sqrt{tA}}\right),$$

we get

$$\int_0^{\varepsilon} f(k) e^{it\varphi(k)} dk = \sqrt{\frac{2\pi}{|\varphi''(0)|}} \frac{f(0)}{\sqrt{t}} \exp\left\{i\left[\varphi(0)t + \frac{\pi}{4}\right]\right\} + O\left(\frac{1}{t}\right),$$

which proves the theorem when $A > 0$. The proof is similar if $A < 0$. \square

Remark 5.24. Theorem 5.22 holds for integrals extended over the whole real axis as well (actually this is the most interesting case) as long as, in addition, f is bounded, $|\varphi'(\pm\infty)| \geq C > 0$, and $\int_{\mathbb{R}} |f'\varphi' - f\varphi''| (\varphi')^{-2} dk < \infty$. Indeed, it is easy to check that Lemma 5.23 is true under these hypotheses and then the proof of Theorem 5.22 is exactly the same.

³⁹ Recall that $e^{i\pi/4} = (\sqrt{2} + i\sqrt{2})/2$. Moreover, the following formulas hold:

$$\left| \frac{\sqrt{\pi}}{2\sqrt{2}} - \int_0^{\lambda} \cos(y^2) dy \right| \leq \frac{\sqrt{\pi}}{\lambda}, \quad \left| \frac{\sqrt{\pi}}{2\sqrt{2}} - \int_0^{\lambda} \sin(y^2) dy \right| \leq \frac{\sqrt{\pi}}{\lambda}.$$

Problems

5.1. The chord of a guitar of length L is plucked at its middle point and then released. Write the mathematical model which governs the vibrations and solve it. Compute the energy $E(t)$.

5.2. Solve the problem

$$\begin{cases} u_{tt} - u_{xx} = 0 & 0 < x < 1, t > 0 \\ u(x, 0) = u_t(x, 0) = 0 & 0 \leq x \leq 1 \\ u_x(0, t) = 1, u(1, t) = 0 & t \geq 0. \end{cases}$$

5.3. *Forced vibrations.* Solve the problem.

$$\begin{cases} u_{tt} - u_{xx} = g(t) \sin x & 0 < x < \pi, t > 0 \\ u(x, 0) = u_t(x, 0) = 0 & 0 \leq x \leq \pi \\ u(0, t) = u(\pi, t) = 0 & t \geq 0. \end{cases}$$

[Answer: $u(x, t) = \sin x \int_0^t g(t - \tau) \sin \tau d\tau$].

5.4. *Equipartition of energy.* Let $u = u(x, t)$ be the solution of the global Cauchy problem for the equation $u_{tt} - c^2 u_{xx} = 0$, with initial data $u(x, 0) = g(x)$, $u_t(x, 0) = h(x)$. Assume that g and h are smooth functions with compact support contained in the interval (a, b) . Show that there exists T such that, for $t \geq T$,

$$E_{cin}(t) = E_{pot}(t).$$

5.5. Solve the global Cauchy problem for the equation $u_{tt} - c^2 u_{xx} = 0$, with the following initial data:

- a) $u(x, 0) = 1$ if $|x| < a$, $u(x, 0) = 0$ if $|x| > a$; $u_t(x, 0) = 0$.
- b) $u(x, 0) = 0$; $u_t(x, 0) = 1$ if $|x| < a$, $u_t(x, 0) = 0$ if $|x| > a$.

5.6. Check that formula (5.43) may be written in the following form:

$$u(x + c\tau - c\eta, t + \tau + \eta) - u(x + c\tau, t + \tau) - u(x - c\eta, t + \eta) + u(x, t) = 0 \quad (5.172)$$

for nonnegative τ, η . Show that if u is a C^2 function and satisfies (5.172) for all $(x, t) \in \mathbb{R} \times (0, +\infty)$ and all $\tau \geq 0, \eta \geq 0$, then $u_{tt} - c^2 u_{xx} = 0$. Thus, (5.172) can be considered as a weak formulation of the wave equation.

5.7. The small longitudinal free vibrations of an elastic bar are governed by the following equation

$$\rho(x) \sigma(x) \frac{\partial^2 u}{\partial t^2} = \frac{\partial}{\partial x} \left[E(x) \sigma(x) \frac{\partial u}{\partial x} \right] \quad (5.173)$$

where u is the longitudinal displacement, ρ is the linear density of the material, σ is the cross section of the bar and E is its *Young's modulus*⁴⁰.

⁴⁰ E is the proportionality factor in the *strain-stress* relation given by Hooke's law: T (strain) = E ε (stress). Here $\varepsilon \simeq u_x$. For steel, $E = 2 \times 10^{11}$ dine/cm², for aluminium, $E = 7 \times 10^{12}$ dine/cm².

Assume the bar has constant cross section but it is constructed by welding together two bars, of different (constant) Young's modulus E_1, E_2 and density ρ_1, ρ_2 , respectively.

Since the two bars are welded together, the displacement u is continuous across the junction, which we locate at $x = 0$. In this case:

- Give a weak formulation of the global initial value problem for equation (5.173).
- Deduce that the following jump condition must hold at $x = 0$:

$$E_1 u_x(0-, t) = E_2 u_x(0+, t) \quad t > 0. \quad (5.174)$$

(c) Let $c_j^2 = E_j / \rho_j$, $j = 1, 2$. A left incoming wave $u_{inc}(x, t) = \exp[i(x - c_1 t)]$ produces at the junction a reflected wave $u_{ref}(x, t) = a \exp[i(x + c_1 t)]$ and a transmitted wave $u_{tr}(x, t) = b \exp[i(x - c_2 t)]$. Determine a, b and interpret the result.

[Hint: (c) Look for a solution of the form $u = u_{inc} + u_{ref}$ for $x < 0$ and $u = u_{tr}$ for $x > 0$. Use the continuity of u and the jump condition (5.174)].

5.8. Determine the characteristics of Tricomi's equation

$$u_{tt} - tu_{xx} = 0.$$

[Answer: $3x \pm 2t^{3/2} = k$, for $t > 0$].

5.9. Classify the equation $t^2 u_{tt} + 2tu_{xt} + u_{xx} - u_x = 0$ and find the characteristics. After a reduction to canonical form, find the general solution.

[Answer: $U(x, t) = F(te^{-x}) + G(te^{-x})e^x$, with F, G arbitrary].

5.10. Consider the following *characteristic Cauchy problem*⁴¹ for the wave equation in the half-plane $x > t$:

$$\begin{cases} u_{tt} - u_{xx} = 0 & x > 0 \\ u(x, x) = f(x) & x \in \mathbb{R} \\ u_{\nu}(x, x) = g(x) & x \in \mathbb{R} \end{cases}$$

where $\nu = (1, -1) / \sqrt{2}$. Establish whether or not this problem is well posed.

5.11. Consider the following so called *Goursat problem*⁴² for the wave equation in the sector $-t < x < t$:

$$\begin{cases} u_{tt} - u_{xx} = 0 & -t < x < t \\ u(x, x) = f(x), u(x, -x) = g(x) & x > 0 \\ f(0) = g(0). \end{cases}$$

Establish whether or not this problem is well posed.

5.12. Ill posed non-characteristic Cauchy problem for the heat equation. Check that for every integer k , the function

$$u_k(x, t) = \frac{1}{k} (\cosh kx \cos kx \cos 2k^2 t - \sinh kx \sin kx \sin 2k^2 t)$$

⁴¹ Note that the data are the values of u and of the normal derivative on the characteristic $x = t$.

⁴² Note that the data are the values of u on the characteristics $x = t$ and $x = -t$, for $x > 0$.

solves $u_t = u_{xx}$ and the (noncharacteristic) initial conditions:

$$u(0, t) = \frac{1}{k} \cos 2k^2 t, \quad u_x(0, t) = 0.$$

Deduce that the corresponding Cauchy problem in the half-plane $x > 0$ is **ill posed**.

5.13. Circular membrane. A perfectly flexible and elastic membrane has at rest the shape of the circle $B_1 = \{(x, y) : x^2 + y^2 \leq 1\}$. If the boundary is fixed and there are no external loads, the vibrations of the membrane are governed by the following system:

$$\begin{cases} u_{tt} - c^2 (u_{rr} + \frac{1}{r} u_r + \frac{1}{r^2} u_{\theta\theta}) = 0 & 0 < r < 1, 0 \leq \theta \leq 2\pi, t > 0 \\ u(r, \theta, 0) = g(r, \theta), \quad u_t(r, 0) = h(r, \theta) & 0 < r < 1, 0 \leq \theta \leq 2\pi \\ u(1, \theta, t) = 0 & 0 \leq \theta \leq 2\pi, t \geq 0. \end{cases}$$

In the case $h = 0$ and $g = g(r)$, use the method of separation of variables to find the solution

$$u(r, t) = \sum_{n=1}^{\infty} a_n J_0(\lambda_n r) \cos \lambda_n t$$

where J_0 is the Bessel function of order zero, $\lambda_1, \lambda_2, \dots$ are the zeros of J_0 and the coefficients a_n are given by

$$a_n = \frac{2}{c_n^2} \int_0^1 g(s) J_0(\lambda_n s) s ds$$

where

$$c_n = \sum_{k=1}^{\infty} \frac{(-1)^k}{k! (k+1)!} \left(\frac{\lambda_n}{2}\right)^{2k+1}$$

(see Remark 2.10, p. 76).

5.14. Circular waveguide. Consider the equation $u_{tt} - c^2 \Delta u = 0$ in the cylinder

$$C_R = \{(r, \theta, z) : 0 \leq r \leq R, 0 \leq \theta \leq 2\pi, -\infty < z < +\infty\}.$$

Determine the axially symmetric solutions of the form $u(r, z, t) = v(r) w(z) h(t)$ satisfying the Neumann condition $u_r = 0$ on $r = R$.

[Answer: $u_n(r, z, t) = \exp\{-i(\omega t - kz)\} J_0(\mu_n r/R)$, $n \in \mathbb{N}$, where J_0 is the Bessel function, μ_n are its stationary points ($J'_0(\mu_n) = 0$) and

$$\frac{\omega^2}{c^2} = k^2 + \frac{\mu_n^2}{R^2}.$$

5.15. Let u be the solution of $u_{tt} - c^2 \Delta u = 0$ in $\mathbb{R}^3 \times (0, +\infty)$ with data

$$u(\mathbf{x}, 0) = g(\mathbf{x}) \quad \text{and} \quad u_t(\mathbf{x}, 0) = h(\mathbf{x}),$$

both supported in the sphere $\bar{B}_\rho(\mathbf{0})$. Describe the support of $u(\cdot, t)$ for $t > 0$.

[Answer: For $\rho \geq ct$, the support of $u(\cdot, t)$ is $\bar{B}_{\rho+ct}(\mathbf{0})$. For $\rho < ct$, the support of $u(\cdot, t)$ is the spherical shell $\bar{B}_{\rho+ct}(\mathbf{0}) \setminus B_{ct-\rho}(\mathbf{0})$, of width 2ρ , which expands at speed c .]

5.16. *Focussing effect.* Solve the problem

$$\begin{cases} w_{tt} - c^2 \Delta w = 0 & \mathbf{x} \in \mathbb{R}^3, t > 0 \\ w(\mathbf{x}, 0) = 0, \quad w_t(\mathbf{x}, 0) = h(|\mathbf{x}|) & \mathbf{x} \in \mathbb{R}^3 \end{cases}$$

where ($r = |\mathbf{x}|$)

$$h(r) = \begin{cases} 1 & 0 \leq r \leq 1 \\ 0 & r > 1. \end{cases}$$

Check that $w(r, t)$ displays a discontinuity at the origin at time $t = 1/c$.

5.17. Let u be a regular and bounded solution of $u_{tt} - c^2 \Delta u = 0$ in $\mathbb{R}^n \times (0, +\infty)$ with $u(\mathbf{x}, 0) = g(\mathbf{x})$, $u_t(\mathbf{x}, 0) = 0$ in \mathbb{R}^n ($n = 1, 2, 3$). Define

$$w(\mathbf{x}, t) = \frac{c}{\sqrt{4D\pi t}} \int_{\mathbb{R}} u(\mathbf{x}, s) \exp\left(-\frac{c^2 s^2}{4Dt}\right) ds.$$

Show (formally) that w solves the heat equation $w_t - D\Delta w = 0$ in $\mathbb{R}^n \times (0, +\infty)$ with $w(\mathbf{x}, 0) = g(\mathbf{x})$ in \mathbb{R}^n .

5.18. *Dissipative term.* Consider the problem

$$\begin{cases} u_{tt}(\mathbf{x}, t) + k u_t(\mathbf{x}, t) = c^2 \Delta u(\mathbf{x}, t) & \mathbf{x} \in \mathbb{R}^2, t > 0 \\ u(\mathbf{x}, 0) = 0, u_t(\mathbf{x}, 0) = g(\mathbf{x}) & \mathbf{x} \in \mathbb{R}^2. \end{cases}$$

- a) Determine $\alpha \in \mathbb{R}$ so that $v(\mathbf{x}, t) = e^{\alpha t} u(\mathbf{x}, t)$ solves an equation without first-order term (but with a zero-order one) on $\mathbb{R}^2 \times (0, +\infty)$.
- b) Find $\beta \in \mathbb{R}$ so that

$$w(x_1, x_2, x_3, t) = w(\mathbf{x}, x_3, t) = e^{\beta x_3} v(\mathbf{x}, t)$$

solves an equation with second-order terms only, on $\mathbb{R}^3 \times (0, +\infty)$.

- c) Determine the solution u of the original problem.

5.19. *Retarded potential* ($n = 2$).

- a) Show that the solution of the two-dimensional nonhomogeneous Cauchy problem with zero initial data is given by

$$u(\mathbf{x}, t) = \frac{1}{2\pi c} \int_0^t \int_{B_{c(t-s)}(\mathbf{x})} \frac{1}{\sqrt{c^2(t-s)^2 - |\mathbf{x} - \mathbf{y}|^2}} f(\mathbf{y}, s) d\mathbf{y} ds.$$

- b) *Point source.* Show that if

$$f(\mathbf{x}, t) = \delta_2(\mathbf{x}) q(t),$$

q smooth for $t \leq 0$ and zero for $t < 0$ then

$$u(\mathbf{x}, t) = \begin{cases} \frac{1}{2\pi c} \int_0^{t - \frac{|\mathbf{x}|}{c}} \frac{q(s)}{\sqrt{c^2(t-s)^2 - |\mathbf{x}|^2}} ds & \text{for } |\mathbf{x}| < ct, \\ 0 & \text{for } |\mathbf{x}| > ct. \end{cases}$$

5.20. For *linear gravity waves* ($\sigma = 0$), examine the case of uniform finite depth, replacing condition (5.161) by

$$\phi_z(x, -H, t) = 0$$

under the initial conditions (5.160).

(a) Write the dispersion relation.

Deduce that:

(b) The phase and group velocity have a finite upper bound.

(c) The square of the phase velocity in deep water ($H \gg \lambda$) is proportional to the wavelength.

(d) Linear shallow water waves ($H \ll \lambda$) are not dispersive.

[Answer: (a) $\omega^2 = gk \tanh(kH)$, (b) $c_p \max = \sqrt{kH}$, (c) $c_p^2 \sim g\lambda/2\pi$, (d) $c_p^2 \sim gH$].

5.21. Determine the travelling wave solutions of the linearized system (5.157) of the form

$$\phi(x, z, t) = F(x - ct) G(z).$$

Rediscover the dispersion relation found in Problem 5.19(a).

[Answer:

$$\phi(x, z, t) = \cosh k(z + H) \{A \cos k(x - ct) + B \sin k(x - ct)\},$$

A, B arbitrary constants and $c^2 = g \tanh(kH)/k$].

Chapter 6

Elements of Functional Analysis

6.1 Motivations

The main purpose in the previous chapters has been to introduce part of the basic and classical theory of some important equations of mathematical physics. The emphasis on phenomenological aspects and the connection with a probabilistic point of view should have conveyed to the reader some intuition and feeling about the interpretation and the limits of those models.

The few rigorous theorems and proofs we have presented had the role of bringing to light the main results on the qualitative properties of the solutions and justifying, partially at least, the well-posedness of the relevant boundary and initial/boundary value problems we have considered.

However, these purposes are somehow in competition with one of the most important role of modern mathematics, which is to reach a unifying vision of large classes of problems under a common structure, capable not only of increasing theoretical understanding, but also of providing the necessary flexibility to guide the numerical methods which will be used to compute approximate solutions.

This conceptual jump requires a change of perspective, based on the introduction of abstract methods, historically originating from the vain attempts to solve basic problems (e.g. in electrostatics) at the end of the 19th century. It turns out that the new level of knowledge opens the door to the solution of complex problems in modern technology.

These abstract methods, in which analytical and geometrical aspects fuse, are the core of the branch of Mathematics, called Functional Analysis.

It could be useful for understanding the subsequent development of the theory, to examine in an informal way how the main ideas come out, working on a couple of specific examples.

Let us go back to the derivation of the diffusion equation, in Subsect. 2.1.2. If the body is heterogeneous or anisotropic, may be with discontinuities in its thermal parameters (e.g. due to the mixture of two different materials), the Fourier

law of heat conduction for the flux function \mathbf{q} takes the form

$$\mathbf{q} = -\mathbf{A}\nabla u,$$

where the matrix $\mathbf{A} = \mathbf{A}(\mathbf{x})$ satisfies the condition

$$\mathbf{q} \cdot \nabla u = -\mathbf{A}\nabla u \cdot \nabla u \leq 0 \quad (\text{ellipticity condition}),$$

reflecting the tendency of heat to flow from hotter to cooler regions. If $\rho = \rho(\mathbf{x})$ and $c_v = c_v(\mathbf{x})$ are the density and the specific heat of the material, and $f = f(\mathbf{x})$ is the rate of external heat supply per unit volume, we are led to the diffusion equation

$$\rho c_v u_t - \operatorname{div}(\mathbf{A}\nabla u) = f.$$

In stationary conditions, $u(\mathbf{x}, t) = u(\mathbf{x})$, and we are reduced to

$$-\operatorname{div}(\mathbf{A}\nabla u) = f. \quad (6.1)$$

Since the matrix \mathbf{A} encodes the conductivity properties of the medium, we expect a low degree of regularity of \mathbf{A} , but then a natural question arises, since we cannot compute the divergence of $\mathbf{A}\nabla u$:

$$\text{what is the meaning of equation (6.1)?} \quad (6.2)$$

We have already faced similar situations in Sect. 4.4.2, where we have introduced discontinuous solutions of a conservation law, and in Sect. 5.4.2, where we have considered solutions of the wave equation with irregular initial data. Let us follow the same ideas.

Suppose we want to solve equation (6.1) in a bounded domain Ω , with zero boundary data (Dirichlet problem). Formally, we multiply the differential equation by a smooth test function v vanishing on $\partial\Omega$, and we integrate over Ω :

$$\int_{\Omega} -\operatorname{div}(\mathbf{A}\nabla u) v \, d\mathbf{x} = \int_{\Omega} f v \, d\mathbf{x}.$$

Using Gauss' formula we obtain the equation

$$\int_{\Omega} \mathbf{A}\nabla u \cdot \nabla v \, d\mathbf{x} = \int_{\Omega} f v \, d\mathbf{x} \quad \forall v \in \mathring{C}^1(\overline{\Omega}), \quad (6.3)$$

where $\mathring{C}^1(\overline{\Omega})$ is the set of functions in $C^1(\overline{\Omega})$, vanishing on $\partial\Omega$.

We call equation (6.3) the *weak* or *variational* formulation of our Dirichlet problem and we may say that $u \in \mathring{C}^1(\overline{\Omega})$ is a *weak* or *variational* solution of our Dirichlet problem if (6.3) holds for every $v \in \mathring{C}^1(\overline{\Omega})$.

Note that (6.3) makes perfect sense for \mathbf{A} and f bounded (possibly discontinuous). Fine, but:

Is the variational problem (6.3) well posed?

Things are not so straightforward, as we have experienced in Subsect. 4.4.3 and, actually, it turns out that the space $\mathring{C}^1(\overline{\Omega})$ is not the proper choice, although it seems to be the natural one. To see why, let us consider another example, somewhat more revealing.

Consider the equilibrium position of a stretched membrane having the shape of a square Ω , subject to an external load f (force per unit mass) and kept at level zero on $\partial\Omega$.

Since there is no time evolution, the position of the membrane may be described by a function $u = u(\mathbf{x})$, solution of the Dirichlet problem

$$\begin{cases} -\Delta u = f & \text{in } \Omega \\ u = 0 & \text{on } \partial\Omega. \end{cases} \quad (6.4)$$

For problem (6.4), eq. (6.3) becomes

$$\int_{\Omega} \nabla u \cdot \nabla v \, d\mathbf{x} = \int_{\Omega} fv \, d\mathbf{x} \quad \forall v \in \mathring{C}^1(\overline{\Omega}). \quad (6.5)$$

Now, this equation has an interesting physical interpretation. The integral in the left hand side represents the work done by the internal elastic forces, due to a *virtual displacement* v . On the other hand $\int_{\Omega} fv$ expresses the work done by the external forces to produce that virtual displacement.

Thus, the variational formulation (6.5) states that these two works balance, which constitutes a version of the *principle of virtual work*.

There is more, if we bring into play the energy. In fact, the *total potential energy* is proportional to

$$E(v) = \underbrace{\frac{1}{2} \int_{\Omega} |\nabla v|^2 \, d\mathbf{x}}_{\text{internal elastic energy}} - \underbrace{\int_{\Omega} fv \, d\mathbf{x}}_{\text{external potential energy}}. \quad (6.6)$$

Since nature likes to save energy, the equilibrium position u corresponds to the minimizer of (6.6) among all the *admissible* configurations $v \in \mathring{C}^1(\overline{\Omega})$. This fact is closely connected with the principle of virtual work and, actually, it is equivalent to it (see Subsect. 8.4.1).

Thus, changing point of view, instead of looking for a variational solution of (6.5) we may, equivalently, look for a minimizer of $E(v)$ over $\mathring{C}^1(\overline{\Omega})$.

However there is a drawback. It turns out that the minimum problem *does not have a solution*, except for some trivial cases. The reason is that we are looking in the wrong set of admissible functions.

Why $\mathring{C}^1(\overline{\Omega})$ is a wrong choice? To be minimalist, it is like looking for the minimizer of the function

$$f(x) = (x - \pi)^2$$

over the rational numbers \mathbb{Q} . Clearly $\inf_{x \in \mathbb{Q}} f(x) = 0$, but it is not the *minimum!* If we enlarge the numerical set from \mathbb{Q} to \mathbb{R} we clearly have $\min_{x \in \mathbb{R}} f(x) = f(\pi) = 0$.

Similarly, the space $\dot{C}^1(\overline{\Omega})$ is too narrow to have any hope of finding the minimizer there and we are forced to enlarge the set of admissible functions. To this purpose observe that $\dot{C}^1(\overline{\Omega})$ is not naturally tied to the physical meaning of $E(v)$, which is an energy and only requires *the gradient of u to be square integrable*, that is $|\nabla u| \in L^2(\Omega)$. There is no need of *a priori* continuity of the derivatives, actually neither of u . Thus, the correct functional setting turns out to be the so called *Sobolev space* $H_0^1(\Omega)$, whose elements are exactly the functions belonging to $L^2(\Omega)$, together with their first derivatives, vanishing on $\partial\Omega$. We could call them functions of finite energy!

Although we feel we are on the right track, there is a price to pay, to put everything in a rigorous perspective and avoid risks of contradiction or non-senses. In fact many questions arise immediately.

The first one is analogous to question (6.2): what do we mean by the *gradient* of a function which is only in $L^2(\Omega)$, maybe with a lot of discontinuities? Second: a function in $L^2(\Omega)$ is, in principle, well defined except on sets of measure zero. But, then, since $\partial\Omega$ is precisely a set of measure zero,

what does it mean “vanishing on $\partial\Omega$ ”?

We shall answer these questions in Chap. 7. We may anticipate that, for the first one, the idea is the same we used to define the *Dirac delta* as a derivative of the Heaviside function, resorting to a weaker notion of derivative (we shall say *in the sense of distributions*), based on the formula of integration by parts and on the introduction of a suitable set of test function.

For the second question, there is a way to introduce in a suitable coherent way a so called *trace operator* which associates to a function $u \in L^2(\Omega)$, with gradient in $L^2(\Omega)$, a function $u|_{\partial\Omega}$ representing its values on $\partial\Omega$ (see Subsect. 6.6.1). The elements of $H_0^1(\Omega)$ vanish on $\partial\Omega$ in the sense that they have zero trace.

Another question is:

what makes the space $H_0^1(\Omega)$ so special?

Here the conjunction between geometrical and analytical aspects comes into play. First of all, although it is an infinite-dimensional vector space, we may endow $H_0^1(\Omega)$ with a structure which reflects as much as possible the structure of a finite dimensional vector space like \mathbb{R}^n , where the life is obviously easier.

Indeed, in this vector space (thinking of \mathbb{R} as the scalar field) we may introduce an *inner product* given by

$$(u, v)_{H_0^1(\Omega)} = \int_{\Omega} \nabla u \cdot \nabla v$$

with the same properties of an inner product in \mathbb{R}^n . Then, it makes sense to talk about *orthogonality* between two functions u and v in $H^1(\Omega)$, expressed by the

vanishing of their inner product:

$$(u, v)_{H_0^1(\Omega)} = 0.$$

Having defined the inner product $(\cdot, \cdot)_{H_0^1(\Omega)}$, we may define the *size (norm)* of u by

$$\|u\|_{H_0^1(\Omega)} = \sqrt{(u, u)_{H_0^1(\Omega)}}$$

and the distance between u and v by

$$\text{dist}(u, v) = \|u - v\|_{H_0^1(\Omega)}.$$

Thus, we may say that a sequence $\{u_n\} \subset H^1(\Omega)$ converges to u in $H^1(\Omega)$ if $\text{dist}(u_n, u) \rightarrow 0$ as $n \rightarrow \infty$. It may be observed that all of this can be done, even more comfortably, in the space $\mathring{C}^1(\overline{\Omega})$. This is true, but with a key difference.

Let us use once more an analogy with an elementary fact. The minimizer of the function $f(x) = (x - \pi)^2$ does not exist among the rational numbers \mathbb{Q} , although it can be approximated as much as one likes by these numbers. If from a very practical point of view, rational numbers could be considered satisfactory enough, certainly it is not so from the point of view of the development of science and technology, since, for instance, no one could even conceive the achievements of *integral and differential calculus* without the real number system.

As \mathbb{R} is the *completion of \mathbb{Q}* , in the sense that \mathbb{R} contains all the limits of sequences in \mathbb{Q} that converge somewhere, the same is true for $H_0^1(\Omega)$ with respect to $\mathring{C}^1(\overline{\Omega})$. This makes $H_0^1(\Omega)$ a so called *Hilbert space* and gives it a big advantage with respect to $\mathring{C}^1(\overline{\Omega})$, which we illustrate going back to our membrane problem and precisely to equation (6.5). This time we use a geometrical interpretation.

In fact, (6.5) means that we are searching for an element u , whose inner product with any element v of $H_0^1(\Omega)$ reproduces “the action of f on v ”, given by the linear map

$$v \mapsto \int_{\Omega} fv.$$

This is a familiar situation in Linear Algebra. Any function $F : \mathbb{R}^n \rightarrow \mathbb{R}$, which is *linear*, that is such that

$$F(a\mathbf{x} + b\mathbf{y}) = aF(\mathbf{x}) + bF(\mathbf{y}) \quad \forall a, b \in \mathbb{R}, \forall \mathbf{x}, \mathbf{y} \in \mathbb{R}^n,$$

can be *expressed as the inner product with a unique representative vector $\mathbf{z}_F \in \mathbb{R}^n$* (Representation Theorem). This amounts to saying that there is exactly one solution \mathbf{z}_F of the equation

$$\mathbf{z} \cdot \mathbf{y} = F(\mathbf{y}), \quad \text{for every } \mathbf{y} \in \mathbb{R}^n. \tag{6.7}$$

The structure of the two equations (6.5), (6.7) is the same: on the left hand side there is an *inner product* and on the other one a *linear map*.

Thus another natural question arises:

Is there any analogue of the Representation Theorem in $H_0^1(\Omega)$?

The answer is yes (see Riesz's Theorem 6.44, p. 388) and holds in any Hilbert space, not only in $H_0^1(\Omega)$. In infinite dimensional spaces it is required the study of *linear, continuous functionals* and the related key concept of *dual space*. Then, an abstract result of geometrical nature, implies the well-posedness of a concrete boundary value problem.

What about equation (6.3)? If the matrix \mathbf{A} is symmetric and strictly positive, the left hand side of (6.3) *still defines an inner product* in $H_0^1(\Omega)$ and again Riesz's Theorem yields the well-posedness of the Dirichlet problem.

If \mathbf{A} is not symmetric, things change only a little. Various generalizations of Riesz's Theorem (e.g. the Lax-Milgram Theorem 6.39, p. 383) allow the unified treatment of more general problems, through their *weak* or *variational formulation*. Actually, as we have experienced with equation (6.3), the variational formulation is often the only way of formulating and solving a problem, without losing its original features.

The above arguments should have convinced the reader of the existence of a general Hilbert space structure, underlying a large class of problems arising in the applications. In this chapter we develop the tools of Functional Analysis, essential for a correct variational formulation of a wide variety of boundary value problems. The results we present constitute the theoretical basis for numerical methods such as *finite elements* or more generally, *Galerkin's methods*, and this makes the theory even more attractive and important.

More advanced results, related to general solvability questions and the so called spectral properties of elliptic operators (formalizing the notions of eigenvalues and eigenfunctions) are included in the final part of this chapter.

A final comment is in order. Look again at the minimization problem above. We have enlarged the class of admissible configurations from a class of quite smooth functions to a rather wide class of functions. What kind of solutions are we finding with these abstract methods? If the data (e.g. Ω and f , for the membrane) are regular, could the corresponding solutions be irregular? If yes, this does not sound too good! In fact, although we are working in a setting of possibly irregular configurations, it turns out that the solution actually possesses its natural degree of regularity, once more confirming the intrinsic coherence of the method.

It also turns out that the knowledge of the optimal regularity of the solution plays an important role in the error control for numerical methods. However, this part of the theory is rather technical and we do not have much space to treat it in detail. We shall only state some of the most common results.

The power of abstract methods is not restricted to stationary problems. As we shall see, Sobolev spaces depending on time can be introduced for the treatment of evolution problems, both of diffusive or wave propagation type (see Chap. 7).

In this introductory book, the emphasis is mainly to *linear* problems. However, in the last section, we present some of the most important tools of nonlinear analysis, the *Contractions principle*, the *Schauder* and the *Leray-Schauder* theorems. In Chaps. 9, 10 and 11, we will have occasion to apply these theorems to a variety of situations, in particular, to the stationary Navier-Stokes equation (Sect. 9.5).

6.2 Norms and Banach Spaces

It may be useful for future developments, to introduce *norm* and *distance* independently of an *inner product*, to emphasize better their axiomatic properties.

Let X be a linear space over the scalar field \mathbb{R} or \mathbb{C} . A *norm* in X , is a real function

$$\|\cdot\| : X \rightarrow \mathbb{R} \quad (6.8)$$

such that, for each scalar λ and every $x, y \in X$, the following properties hold:

- $N_1 : \|x\| \geq 0; \|x\| = 0$ if and only if $x = 0$ (positivity).
- $N_2 : \|\lambda x\| = |\lambda| \|x\|$ (homogeneity).
- $N_3 : \|x + y\| \leq \|x\| + \|y\|$ (triangular inequality).

A norm is introduced to measure the size (or the “length”) of each vector $x \in X$, so that properties N_1, N_2, N_3 should appear as natural requirements.

A *normed space* is a linear space X endowed with a norm $\|\cdot\|$. A norm induces a *distance* between two vectors given by

$$d(x, y) = \|x - y\|, \quad (6.9)$$

which makes X into a metric space.

More generally, a *metric space* is a set M , endowed with a distance $d : M \times M \rightarrow \mathbb{R}$, satisfying the following three properties; for every $x, y, z \in M$:

- $D_1 : d(x, y) \geq 0, d(x, y) = 0$, if and only if $x = y$ (positivity).
- $D_2 : d(x, y) = d(y, x)$ (symmetry).
- $D_3 : d(x, z) \leq d(x, y) + d(y, z)$ (triangular inequality).

Notice that a metric space *does not need to be a linear space*. We call

$$B_r(x) = \{y \in M : d(x, y) < r\}$$

the ball of radius r , centered at x . We say that a sequence $\{x_n\} \subset M$ converges to x in M , and we write $x_m \rightarrow x$ in M , if

$$d(x_m, x) \rightarrow 0 \quad \text{as } m \rightarrow \infty.$$

The limit is unique: if $d(x_m, x) \rightarrow 0$ and $d(x_m, y) \rightarrow 0$ as $m \rightarrow \infty$, then, from the triangular inequality, we have

$$d(x, y) \leq d(x_m, x) + d(x_m, y) \rightarrow 0 \text{ as } m \rightarrow \infty$$

which implies $x = y$.

As in Sect. 1.4, we can define a *topology in M*, induced by the metric, repeating verbatim all the definitions in that section and maintaining also the same properties. In particular, a set is closed (compact) if and only if it is sequentially closed (compact).

An important distinction is between convergent and Cauchy sequences. A sequence $\{x_m\} \subset M$ is a *Cauchy* (or *fundamental*) sequence if

$$d(x_m, x_k) \rightarrow 0 \quad \text{as } m, k \rightarrow \infty.$$

If $x_m \rightarrow x$ in M , from the triangular inequality, we may write

$$d(x_m, x_k) \leq d(x_m, x) + d(x_k, x) \rightarrow 0 \quad \text{as } m, k \rightarrow \infty,$$

and therefore

$$\{x_m\} \text{ convergent implies that } \{x_m\} \text{ is a Cauchy sequence.} \quad (6.10)$$

The converse is not true, in general. Take $M = \mathbb{Q}$, with the usual norm given by $|x|$. The sequence of rational numbers

$$x_m = \left(1 + \frac{1}{m}\right)^m$$

is a Cauchy sequence but it is *not* convergent in \mathbb{Q} , since its limit is the irrational number e .

A metric space in which every Cauchy sequence converges is called **complete**. The normed spaces which are complete with respect to the induced norm (6.9) deserves a special name.

Definition 6.1. *A complete, normed linear space is called **Banach space**.*

The notion of convergence (or of limit) can be extended to functions from a metric space into another, always reducing it to the convergence of distances, that are real functions.

Let $(M_1, d_1), (M_2, d_2)$ be two metric spaces and let $F : M_1 \rightarrow M_2$. We say that F is continuous at $x \in M_1$ if

$$d_2(F(y), F(x)) \rightarrow 0 \quad \text{when } d_1(x, y) \rightarrow 0$$

or, equivalently, if, for every sequence $\{x_m\} \subset M_1$,

$$d_1(x_m, x) \rightarrow 0 \quad \text{implies} \quad d_2(F(x_m), F(x)) \rightarrow 0.$$

F is *continuous in M* if it is *continuous at every* $x \in M$. In particular:

Proposition 6.2. *Every norm in a linear space X is continuous in X .*

Proof. Let $\|\cdot\|$ be a norm in X . From the triangular inequality, we may write,

$$\|y\| \leq \|y - x\| + \|x\| \quad \text{and} \quad \|x\| \leq \|y - x\| + \|y\|$$

whence

$$|\|y\| - \|x\|| \leq \|y - x\|.$$

Thus, if $\|y - x\| \rightarrow 0$ then

$$|\|y\| - \|x\|| \rightarrow 0,$$

which is the continuity of the norm. \square

Definition 6.3. *Two norms $\|\cdot\|_1$ and $\|\cdot\|_2$, defined in the same space X , are equivalent if there exist two positive numbers c_1, c_2 such that*

$$c_1 \|x\|_2 \leq \|x\|_1 \leq c_2 \|x\|_2 \quad \text{for every } x \in X.$$

Two equivalent norms induce the same topology. Some examples are in order.

Spaces of continuous functions. Let $\Omega \subset \mathbb{R}^n$ be a bounded open set.

The space $C(\overline{\Omega})$. The symbol $C(\overline{\Omega})$ denotes the set of (real or complex) continuous functions on $\overline{\Omega}$, endowed with the norm (called the maximum norm)

$$\|f\|_{C(\overline{\Omega})} = \max_{\overline{\Omega}} |f|.$$

A sequence $\{f_m\}$ converges to f in $C(\overline{\Omega})$ if

$$\max_{\overline{\Omega}} |f_m - f| \rightarrow 0,$$

that is, if f_m converges uniformly to f in $\overline{\Omega}$. Since a uniform limit of continuous functions is continuous, $C(\overline{\Omega})$ is a Banach space.

Note that other norms may be introduced in $C(\overline{\Omega})$, for instance the $L^2(\overline{\Omega})$ norm

$$\|f\|_{L^2(\overline{\Omega})} = \left(\int_{\overline{\Omega}} |f|^2 \right)^{1/2}.$$

Equipped with this norm $C(\overline{\Omega})$ is not complete. Let, for example $\Omega = (-1, 1) \subset \mathbb{R}$. The sequence

$$f_m(t) = \begin{cases} 0 & t \leq 0 \\ mt & 0 < t \leq \frac{1}{m} \\ 1 & t > \frac{1}{m} \end{cases} \quad (m \geq 1),$$

contained in $C([-1, 1])$, is a Cauchy sequence with respect to the L^2 norm. In fact

(letting $m > k$),

$$\begin{aligned}\|f_m - f_k\|_{L^2(A)}^2 &= \int_{-1}^1 |f_m(t) - f_k(t)|^2 dt = (m-k)^2 \int_0^{1/m} t^2 dt + \int_0^{1/k} (1-kt)^2 dt \\ &= \frac{(m-k)^2}{3m^3} + \frac{1}{3k} < \frac{1}{3} \left(\frac{1}{m} + \frac{1}{k} \right) \rightarrow 0 \quad \text{as } m, k \rightarrow \infty.\end{aligned}$$

However, f_n converges in $L^2(-1, 1)$ -norm (and pointwise) to the Heaviside function

$$\mathcal{H}(t) = \begin{cases} 1 & t \geq 0 \\ 0 & t < 0, \end{cases}$$

which is discontinuous at $t = 0$ and therefore does not belong to $C([-1, 1])$.

The spaces $C^k(\overline{\Omega})$, $k \geq 0$ integer. The symbol $C^k(\overline{\Omega})$ denotes the set of functions whose derivatives, up to the order k included, are continuous in Ω and can be extended continuously up to $\partial\Omega$.

To denote a derivative of order m , it is convenient to use the symbol (see Sect. 3.3.7)

$$D^\alpha = \frac{\partial^{\alpha_1}}{\partial x_1^{\alpha_1}} \cdots \frac{\partial^{\alpha_n}}{\partial x_n^{\alpha_n}},$$

where $\alpha = (\alpha_1, \dots, \alpha_n)$ is an n -uple of nonnegative integers (*multi-index*), of *length*

$$|\alpha| = \alpha_1 + \dots + \alpha_n = m.$$

We endow $C^k(\overline{\Omega})$ with the norm (*maximum norm of order k*)

$$\|f\|_{C^k(\overline{\Omega})} = \|f\|_{C(\overline{\Omega})} + \sum_{|\alpha|=1}^k \|D^\alpha f\|_{C(\overline{\Omega})}.$$

If $\{f_n\}$ is a Cauchy sequence in $C^k(\overline{\Omega})$, all the sequences $\{D^\alpha f_n\}$ with $0 \leq |\alpha| \leq k$ are Cauchy sequences in $C(\overline{\Omega})$. From the theorems on term by term differentiation of sequences (see Sect. 1.4), it follows that the resulting space is a Banach space.

The spaces $C^{0,\alpha}(\overline{\Omega})$, $0 \leq \alpha \leq 1$. We say that a function f is *Hölder continuous* in Ω , with exponent α , if

$$\sup_{\substack{\mathbf{x}, \mathbf{y} \in \Omega \\ \mathbf{x} \neq \mathbf{y}}} \frac{|f(\mathbf{x}) - f(\mathbf{y})|}{|\mathbf{x} - \mathbf{y}|^\alpha} \equiv C_H(f; \Omega) < \infty. \quad (6.11)$$

The quotient in the left hand side of (6.11) represents an “incremental quotient of order α ”. The number $C_H(f; \Omega)$ is called the *Hölder constant* of f in Ω . If $\alpha = 1$, f is Lipschitz continuous. Typical examples of Hölder continuous functions in \mathbb{R}^n with exponent α are the powers $f(\mathbf{x}) = |\mathbf{x}|^\alpha$.

We use the symbol $C^{0,\alpha}(\overline{\Omega})$ to denote the set of functions of $C(\overline{\Omega})$, Hölder continuous in Ω with exponent α . Endowed with the norm

$$\|f\|_{C^{0,\alpha}(\overline{\Omega})} = \|f\|_{C(\overline{\Omega})} + C_H(f; \Omega),$$

$C^{0,\alpha}(\overline{\Omega})$ becomes a Banach space.

The spaces $C^{k,\alpha}(\overline{\Omega})$, $k \geq 1$ integer, $0 \leq \alpha \leq 1$. $C^{k,\alpha}(\overline{\Omega})$ denotes the set of functions of $C^k(\overline{\Omega})$, whose derivatives up to order k included, are Hölder continuous in Ω with exponent α . Endowed with the norm

$$\|f\|_{C^{k,\alpha}(\overline{\Omega})} = \|f\|_{C^k(\overline{\Omega})} + \sum_{|\beta|=k} C_H(D^\beta f; \Omega)$$

$C^{k,\alpha}(\overline{\Omega})$ becomes a Banach space.

We shall see in Chaps. 7 and 8 the usefulness of these spaces.

Remark 6.4. With the introduction of *function spaces* we are actually making a step towards abstraction, regarding a function from a different perspective. In calculus we see it as a point map while here we have to consider it as a *single element* (or a point or a vector) of a *vector space*.

Summable and bounded functions. Let Ω be an *open set* in \mathbb{R}^n and $p \geq 1$ a real number. We denote by $L^p(\Omega)$ the set of functions $f : \Omega \rightarrow \mathbb{R}$ (or \mathbb{C}) such that $|f|^p$ is Lebesgue integrable in Ω . Identifying two functions f and g when they are *equal a.e.*¹ in Ω , $L^p(\Omega)$ becomes a Banach space² when equipped with the norm (*integral norm of order p*)

$$\|f\|_{L^p(\Omega)} = \left(\int_{\Omega} |f|^p \right)^{1/p}.$$

The identification of two functions equal a.e. amounts to saying that an element of $L^p(\Omega)$ is not a single function but, actually, an equivalence class of functions, different from one another only on subsets of measure zero. At first glance, this fact could be annoying, but after all, the situation is perfectly analogous to some familiar ones. Take for instance the *rational numbers*. A rational number is rigorously defined as an equivalent class of fractions: the fractions $2/3, 4/6, 8/12 \dots$ represent the *same* number, but in practice one uses the more convenient representative, namely $2/3$. Similarly, when dealing with a “function” in $L^p(\Omega)$, for most practical purposes, one may refer to the more convenient representative of the class. However, as we shall see later, some care is due, especially when pointwise properties are involved.

¹ A property is valid *almost everywhere* in a set Ω , *a.e.* in short, if it is true at all points in Ω , but for a subset of measure zero (see Appendix B).

² See e.g. [39], *Yoshida*, 1965.

The symbol $L^\infty(\Omega)$ denotes the set of *essentially bounded* functions in Ω . Recall³ that $f : \Omega \rightarrow \mathbb{R}$ (or \mathbb{C}) is essentially bounded if there exists M such that

$$|f(x)| \leq M \quad \text{a.e. in } \Omega. \quad (6.12)$$

The infimum of all numbers M with the property (6.12) is called *essential supremum of $|f|$* , and denoted by

$$\|f\|_{L^\infty(\Omega)} = \operatorname{ess\,sup}_{\Omega} |f|.$$

Note that the essential supremum may differ from the supremum. As an example take the characteristic function of the rational numbers $\chi_{\mathbb{Q}}$, equal to 1 on \mathbb{Q} and 0 on $\mathbb{R} \setminus \mathbb{Q}$. Then $\sup_{\mathbb{R}} \chi_{\mathbb{Q}} = 1$, but $\|\chi_{\mathbb{Q}}\|_{L^\infty(\mathbb{R})} = 0$, since \mathbb{Q} has zero Lebesgue measure.

If we identify two functions when they are equal a.e., $\|f\|_{L^\infty(\Omega)}$ is a norm in $L^\infty(\Omega)$, and $L^\infty(\Omega)$ becomes a Banach space. Hölder inequality (1.9), p. 11, for real functions, may be now rewritten in terms of norms as follows:

$$\left| \int_{\Omega} fg \right| \leq \|f\|_{L^p(\Omega)} \|g\|_{L^q(\Omega)}, \quad (6.13)$$

where $q = p/(p-1)$ is the *conjugate exponent* of p , allowing also the case $p = 1$, $q = \infty$.

Note that, if Ω has *finite measure* and $1 \leq p_1 < p_2 \leq \infty$, from (6.13) we have, choosing $g \equiv 1$, $p = p_2/p_1$ and $q = p_2/(p_2 - p_1)$:

$$\int_{\Omega} |f|^{p_1} \leq |\Omega|^{1/q} \|f\|_{L^{p_2}(\Omega)}^{p_1}$$

and therefore $L^{p_2}(\Omega) \subset L^{p_1}(\Omega)$. If the measure of Ω is *infinite*, this inclusion is not true, in general; for instance, $f \equiv 1$ belongs to $L^\infty(\mathbb{R})$ but is not in $L^p(\mathbb{R})$ for $1 \leq p < \infty$.

Remark 6.5. If Ω is bounded, $C(\overline{\Omega}) \subset L^p(\Omega)$ for every $1 \leq p \leq \infty$. If $p < \infty$, from Theorem B.11, p. 669, we deduce that $C(\overline{\Omega})$ is dense in $L^p(\Omega)$. This means that if $f \in L^p(\Omega)$, $1 \leq p < \infty$, there exists a sequence $\{f_k\} \subset C(\overline{\Omega})$ such that $f_k \rightarrow f$ in $L^p(\Omega)$. Thus $C(\overline{\Omega})$ is a **nonclosed** subspace of $L^p(\Omega)$, $1 \leq p < \infty$.

6.3 Hilbert Spaces

Let X be a linear space over \mathbb{R} . An *inner or scalar product* in X is a function

$$(\cdot, \cdot) : X \times X \rightarrow \mathbb{R}$$

³ See Appendix B.

with the following three properties. For every $x, y, z \in X$ and every scalar $\lambda, \mu \in \mathbb{R}$:

- $$\begin{aligned} H_1 : (x, x) &\geq 0 \text{ and } (x, x) = 0 \text{ if and only if } x = 0 && (\text{positivity}). \\ H_2 : (x, y) &= (y, x) && (\text{symmetry}). \\ H_3 : (\mu x + \lambda y, z) &= \mu(x, z) + \lambda(y, z) && (\text{bilinearity}). \end{aligned}$$

A linear space endowed with an inner product is called an *inner product space*. Property H_3 shows that the inner product is linear with respect to its first argument. From H_2 , the same is true for the second argument as well. Then, we say that (\cdot, \cdot) constitutes a *symmetric bilinear form* in X . When different inner product spaces are involved it may be necessary the use of notations like $(\cdot, \cdot)_X$, to avoid confusion.

If the scalar field is \mathbb{C} , then $(\cdot, \cdot) : X \times X \rightarrow \mathbb{C}$, and property H_2 has to be replaced by

$$H_2^* : (x, y) = \overline{(y, x)}$$

where the bar denotes complex conjugation. As a consequence, we have

$$(z, \mu x + \lambda y) = \overline{\mu}(z, x) + \overline{\lambda}(z, y)$$

and we say that (\cdot, \cdot) is *antilinear* with respect to its second argument or that it is a *sesquilinear form* in X .

An inner product induces a norm, given by

$$\|x\| = \sqrt{(x, x)}. \quad (6.14)$$

In fact, properties 1 and 2 in the definition of norm are immediate, while the triangular inequality is a consequence of the following important theorem.

Theorem 6.6. Let $x, y \in X$. Then:

(1) **Schwarz's inequality:**

$$|(x, y)| \leq \|x\| \|y\|. \quad (6.15)$$

Moreover equality holds in (6.15) if and only if x and y are linearly dependent.

(2) **Parallelogram law:**

$$\|x + y\|^2 + \|x - y\|^2 = 2\|x\|^2 + 2\|y\|^2.$$

Proof. (1) We mimic the finite dimensional proof. Let $t \in \mathbb{R}$ and $x, y \in X$. Using the properties of the inner product and (6.14), we may write:

$$0 \leq (tx + y, tx + y) = t^2\|x\|^2 + 2t(x, y) + \|y\|^2 \equiv P(t).$$

Thus, the second degree polynomial $P(t)$ is always nonnegative, whence

$$(x, y)^2 - \|x\|^2\|y\|^2 \leq 0$$

which is the Schwarz inequality. Equality is possible only if $tx + y = 0$, i.e. if x and y are linearly dependent.

(2) Just observe that

$$\|x \pm y\|^2 = (x \pm y, y \pm y) = \|x\|^2 \pm 2(x, y) + \|y\|^2. \quad (6.16)$$

□

The parallelogram law generalizes an elementary result in euclidean plane geometry: *in a parallelogram, the sum of the squares of the sides length equals the sum of the squares of the diagonals length.*

The Schwarz inequality implies that the inner product is continuous; in fact, writing

$$(w, z) - (x, y) = (w - x, z) + (x, z - y)$$

we have

$$|(w, z) - (x, y)| \leq \|w - x\| \|z\| + \|x\| \|z - y\|$$

so that, if $w \rightarrow x$ and $z \rightarrow y$, then $(w, z) \rightarrow (x, y)$.

Definition 6.7. Let H be an inner product space. We say that H is a **Hilbert space** if it is complete with respect to the norm (6.14), induced by the inner product.

Two Hilbert spaces H_1 and H_2 are *isometric* if there exists a one-to-one and onto linear map $L : H_1 \rightarrow H_2$, called *isometry*, which preserves the norm, that is:

$$\|x\|_{H_1} = \|Lx\|_{H_2}, \quad \forall x \in H_1. \quad (6.17)$$

An isometry preserves also the inner product since from

$$\|x - y\|_{H_1}^2 = \|L(x - y)\|_{H_2}^2 = \|Lx - Ly\|_{H_2}^2,$$

we get

$$\|x\|_{H_1}^2 - 2(x, y)_{H_1} + \|y\|_{H_1}^2 = \|Lx\|_{H_2}^2 - 2(Lx, Ly)_{H_2} + \|Ly\|_{H_2}^2$$

and from (6.17) we infer

$$(x, y)_{H_1} = (Lx, Ly)_{H_2}, \quad \forall x, y \in H_1.$$

Example 6.8. \mathbb{R}^n is a Hilbert space with respect to the usual inner product

$$(\mathbf{x}, \mathbf{y})_{\mathbb{R}^n} = \mathbf{x} \cdot \mathbf{y} = \sum_{j=1}^n x_j y_j, \quad \mathbf{x} = (x_1, \dots, x_n), \mathbf{y} = (y_1, \dots, y_n).$$

The induced norm is

$$|\mathbf{x}| = \sqrt{\mathbf{x} \cdot \mathbf{x}} = \sqrt{\sum_{j=1}^n x_j^2}.$$

More generally, if $\mathbf{A} = (a_{ij})_{i,j=1,\dots,n}$ is a square matrix of order n , *symmetric* and *positive*,

$$(\mathbf{x}, \mathbf{y})_{\mathbf{A}} = \mathbf{x} \cdot \mathbf{A} \mathbf{y} = \mathbf{A} \mathbf{x} \cdot \mathbf{y} = \sum_{i=1}^n a_{ij} x_i y_j \quad (6.18)$$

defines another scalar product in \mathbb{R}^n . Actually, *every* inner product in \mathbb{R}^n may be written in the form (11.57), with a suitable matrix \mathbf{A} .

\mathbb{C}^n is a Hilbert space with respect to the inner product

$$(\mathbf{x}, \mathbf{y})_{\mathbb{C}^n} = \sum_{j=1}^n x_j \bar{y}_j \quad \mathbf{x} = (x_1, \dots, x_n), \mathbf{y} = (y_1, \dots, y_n).$$

It is easy to show that every real (resp. complex) linear space of dimension n is isometric to \mathbb{R}^n (resp. \mathbb{C}^n).

Example 6.9. $L^2(\Omega)$ is a Hilbert space (perhaps the most important one) with respect to the inner product

$$(u, v)_{L^2(\Omega)} = \int_{\Omega} uv.$$

Example 6.10. Let $l_{\mathbb{C}}^2$ be the set of complex sequences $\mathbf{x} = \{x_m\}$ such that

$$\sum_{m \in \mathbb{Z}} |x_m|^2 < \infty.$$

For $\mathbf{x} = \{x_m\}$ and $\mathbf{y} = \{y_m\}$, define

$$(\mathbf{x}, \mathbf{y})_{l_{\mathbb{C}}^2} = \sum_{m \in \mathbb{Z}} x_i \bar{y}_j, \quad \mathbf{x} = \{x_n\}, \mathbf{y} = \{y_n\}.$$

Then $(\mathbf{x}, \mathbf{y})_{l_{\mathbb{C}}^2}$ is an inner product which makes $l_{\mathbb{C}}^2$ into a Hilbert space over \mathbb{C} . This space constitutes the discrete analogue of $L^2(0, 2\pi)$. Indeed, each $u \in L^2(0, 2\pi)$ has an expansion in Fourier series (Appendix A)

$$u(x) = \sum_{m \in \mathbb{Z}} \hat{u}_m e^{imx},$$

where

$$\hat{u}_m = \frac{1}{2\pi} \int_0^{2\pi} u(x) e^{-imx} dx.$$

Note that $\bar{\hat{u}}_m = \hat{u}_{-m}$, since u is a real function. From Parseval's identity, we have

$$(u, v)_{L^2(0, 2\pi)} = \int_0^{2\pi} uv = 2\pi \sum_{m \in \mathbb{Z}} \hat{u}_m \bar{\hat{v}}_{-m}$$

and (Bessel's equation)

$$\|u\|_{L^2(0, 2\pi)}^2 = \int_0^{2\pi} u^2 = 2\pi \sum_{m \in \mathbb{Z}} |\hat{u}_m|^2.$$

Example 6.11. A Sobolev space. It is possible to use the frequency space introduced in the previous example to define the derivatives of a function in $L^2(0, 2\pi)$ in a weak or generalized sense. Let $u \in C^1(\mathbb{R})$, 2π -periodic. The Fourier coefficients of u' are given by

$$\widehat{u'}_m = im\widehat{u}_m$$

and we may write

$$\|u'\|_{L^2(0,2\pi)}^2 = \int_0^{2\pi} (u')^2 = 2\pi \sum_{m \in \mathbb{Z}} m^2 |\widehat{u}_m|^2. \quad (6.19)$$

Thus, both sequences $\{\widehat{u}_m\}$ and $\{m\widehat{u}_m\}$ belong to $l_{\mathbb{C}}^2$. But the right hand side in (6.19) does not involve u' directly, so that it makes perfect sense to define

$$H_{per}^1(0, 2\pi) = \{u \in L^2(0, 2\pi) : \{\widehat{u}_m\}, \{m\widehat{u}_m\} \in l_{\mathbb{C}}^2\}$$

and introduce the inner product

$$(u, v)_{H_{per}^1(0,2\pi)} = (2\pi) \sum_{m \in \mathbb{Z}} (1 + m^2) \widehat{u}_m \widehat{v}_{-m}$$

which makes $H_{per}^1(0, 2\pi)$ into a Hilbert space. Since

$$\{m\widehat{u}_m\} \in l_{\mathbb{C}}^2,$$

with each $u \in H_{per}^1(0, 2\pi)$ is associated the function $v \in L^2(0, 2\pi)$ given by

$$v(x) = \sum_{m \in \mathbb{Z}} im\widehat{u}_m e^{imx}.$$

We see that v may be considered as a *generalized derivative* of u and $H_{per}^1(0, 2\pi)$ as the space of functions in $L^2(0, 2\pi)$, together with their first derivatives. Let $u \in H_{per}^1(0, 2\pi)$ and

$$u(x) = \sum_{m \in \mathbb{Z}} \widehat{u}_m e^{imx}.$$

Since

$$|\widehat{u}_m e^{imx}| = \frac{1}{m} m |\widehat{u}_m| \leq \frac{1}{2} \left(\frac{1}{m^2} + m^2 |\widehat{u}_m|^2 \right)$$

the Weierstrass test entails that the Fourier series of u converges uniformly in \mathbb{R} . Thus u has a continuous, 2π -periodic extension to all \mathbb{R} . Finally observe that, if we use the symbol u' also for the generalized derivative of u , the inner product in $H_{per}^1(0, 2\pi)$ can be written in the form

$$(u, v)_{H_{per}^1(0,2\pi)} = \int_0^{2\pi} (u'v' + uv).$$

6.4 Projections and Bases

6.4.1 Projections

Hilbert spaces are the ideal setting to solve problems in infinitely many dimensions. They unify through the inner product and the induced norm, both an analytical and a geometric structure. As we shall shortly see, we may coherently introduce the concepts of orthogonality, projection and basis, prove a infinite-dimensional Pythagoras' Theorem (an example is just Bessel's equality) and introduce other operations, extremely useful from both a theoretical and practical point of view.

As in the finite-dimensional case, two elements x, y belonging to an inner product space are called **orthogonal or normal** if $(x, y) = 0$, and we write $x \perp y$.

Now, if we consider a subspace V of \mathbb{R}^n , e.g. a hyperplane through the origin, every $\mathbf{x} \in \mathbb{R}^n$ has a unique orthogonal projection onto V . In fact, if $\dim V = k$ and the unit vectors $\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_k$ constitute an *orthonormal basis* in V , we may always find an orthonormal basis in \mathbb{R}^n , given by

$$\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_k, \mathbf{w}_{k+1}, \dots, \mathbf{w}_n,$$

where $\mathbf{w}_{k+1}, \dots, \mathbf{w}_n$ are suitable unit vectors. Thus, if

$$\mathbf{x} = \sum_{j=1}^k x_j \mathbf{v}_j + \sum_{j=k+1}^n x_j \mathbf{w}_j,$$

the projection of \mathbf{x} onto V is given by

$$P_V \mathbf{x} = \sum_{j=1}^k x_j \mathbf{v}_j.$$

On the other hand, the projection $P_V \mathbf{x}$ can be characterized through the following property, which does not involve a basis in \mathbb{R}^n : $P_V \mathbf{x}$ is *the point in V that minimizes the distance from \mathbf{x}* , that is

$$|P_V \mathbf{x} - \mathbf{x}| = \inf_{\mathbf{y} \in V} |\mathbf{y} - \mathbf{x}|. \quad (6.20)$$

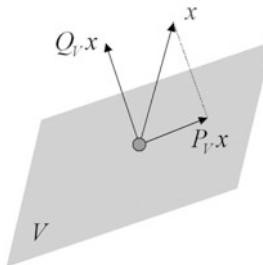
In fact, if $\mathbf{y} = \sum_{j=1}^k y_j \mathbf{v}_j$, we have

$$|\mathbf{y} - \mathbf{x}|^2 = \sum_{j=1}^k (y_j - x_j)^2 + \sum_{j=k+1}^n x_j^2 \geq \sum_{j=k+1}^n x_j^2 = |P_V \mathbf{x} - \mathbf{x}|^2. \quad (6.21)$$

In this case, the “infimum” in (6.20) is actually a “minimum”.

The uniqueness of $P_V \mathbf{x}$ follows from the fact that, if $\mathbf{y}^* \in V$ and

$$|\mathbf{y}^* - \mathbf{x}| = |P_V \mathbf{x} - \mathbf{x}|,$$

**Fig. 6.1** Projection Theorem

then, from (6.21) with $\mathbf{y} = \mathbf{y}^*$, we must have

$$\sum_{j=1}^k (y_j^* - x_j)^2 = 0,$$

whence $y_j^* = x_j$ for $j = 1, \dots, k$, and therefore $\mathbf{y}^* = P_V \mathbf{x}$. Since

$$(\mathbf{x} - P_V \mathbf{x}) \perp \mathbf{v}, \quad \forall \mathbf{v} \in V,$$

every $\mathbf{x} \in \mathbb{R}^n$ may be written in a unique way in the form

$$\mathbf{x} = \mathbf{y} + \mathbf{z}$$

with $\mathbf{y} \in V$ and $\mathbf{z} \in V^\perp$, where V^\perp denotes the subspace of the vectors orthogonal to V (see Fig. 6.1).

Then, we say that \mathbb{R}^n is *direct sum* of the subspaces V and V^\perp and we write

$$\mathbb{R}^n = V \oplus V^\perp.$$

Finally,

$$|\mathbf{x}|^2 = |\mathbf{y}|^2 + |\mathbf{z}|^2$$

which is Pythagoras' Theorem in \mathbb{R}^n .

We may extend all the above consideration to infinite-dimensional Hilbert spaces H , if we consider **closed subspaces** V of H . Here *closed* means with respect to the convergence induced by the norm in H . Recall that a subset $U \subset H$ is closed in H if it contains all the limit points of sequences in U . Observe that if V has *finite dimension* k , it is automatically closed, since it is isometric to \mathbb{R}^k (or \mathbb{C}^k). Also, a closed subspace of H is a Hilbert space as well, with respect to the inner product in H .

Unless stated explicitly, **from now on we consider Hilbert spaces over \mathbb{R}** (real Hilbert spaces), endowed with inner product (\cdot, \cdot) and induced norm $\|\cdot\|$.

Theorem 6.12. (Projection Theorem). *Let V be a closed subspace of a Hilbert space H . Then, for every $x \in H$, there exists a unique element $P_V x \in V$ such that*

$$\|P_V x - x\| = \inf_{v \in V} \|v - x\|. \quad (6.22)$$

Moreover, the following properties hold:

1. $P_V x = x$ if and only if $x \in V$.
2. Let $Q_V x = x - P_V x$. Then $Q_V x \in V^\perp$ and

$$\|x\|^2 = \|P_V x\|^2 + \|Q_V x\|^2.$$

Proof. Let

$$d = \inf_{v \in V} \|v - x\|.$$

By the definition of greatest lower bound, we may select a sequence $\{v_m\} \subset V$, such that $\|v_m - x\| \rightarrow d$ as $m \rightarrow \infty$. In fact, for every integer $m \geq 1$ there exists $v_m \in V$ such that

$$d \leq \|v_m - x\| < d + \frac{1}{m}. \quad (6.23)$$

Letting $m \rightarrow \infty$ in (6.23), we get $\|v_m - x\| \rightarrow d$.

We now show that $\{v_m\}$ is a Cauchy sequence. In fact, using the parallelogram law for the vectors $v_k - x$ and $v_m - x$, we obtain

$$\|v_k + v_m - 2x\|^2 + \|v_k - v_m\|^2 = 2\|v_k - x\|^2 + 2\|v_m - x\|^2. \quad (6.24)$$

Since $\frac{v_k + v_m}{2} \in V$, we may write

$$\|v_k + v_m - 2x\|^2 = 4 \left\| \frac{v_k + v_m}{2} - x \right\|^2 \geq 4d^2$$

whence, from (6.24):

$$\begin{aligned} \|v_k - v_m\|^2 &= 2\|v_k - x\|^2 + 2\|v_m - x\|^2 - \|v_k + v_m - 2x\|^2 \\ &\leq 2\|v_k - x\|^2 + 2\|v_m - x\|^2 - 4d^2. \end{aligned}$$

Letting $k, m \rightarrow \infty$, the right hand side goes to zero and therefore

$$\|v_k - v_m\| \rightarrow 0$$

as well. This proves that $\{v_m\}$ is a Cauchy sequence.

Since H is complete, v_m converges to an element $w \in H$ which belongs to V , because V is closed. Using the norm continuity (Proposition 6.2) we deduce

$$\|v_m - x\| \rightarrow \|w - x\| = d$$

so that w realizes the minimum distance from x among the elements in V .

We have to prove the uniqueness of w . Suppose $\bar{w} \in V$ is another element such that $\|\bar{w} - x\| = d$. The parallelogram law, applied to the vectors $w - x$ and $\bar{w} - x$, yields

$$\begin{aligned} \|w - \bar{w}\|^2 &= 2\|w - x\|^2 + 2\|\bar{w} - x\|^2 - 4 \left\| \frac{w + \bar{w}}{2} - x \right\|^2 \\ &\leq 2d^2 + 2d^2 - 4d^2 = 0 \end{aligned}$$

whence $w = \bar{w}$. We have proved that there exists a unique element $w = P_V x \in V$ such that

$$\|x - P_V x\| = d.$$

To prove 1, observe that, since V is closed, $x \in V$ if and only if $d = 0$, which means $x = P_V x$.

To show 2, let $Q_V x = x - P_V x$, $v \in V$ and $t \in \mathbb{R}$. Since $P_V x + tv \in V$ for every t , we have:

$$\begin{aligned} d^2 &\leq \|x - (P_V x + tv)\|^2 = \|Q_V x - tv\|^2 \\ &= \|Q_V x\|^2 - 2t(Q_V x, v) + t^2 \|v\|^2 \\ &= d^2 - 2t(Q_V x, v) + t^2 \|v\|^2. \end{aligned}$$

Erasing d^2 and dividing by $t > 0$, we get

$$(Q_V x, v) \leq \frac{t}{2} \|v\|^2$$

which forces $(Q_V x, v) \leq 0$; dividing by $t < 0$ we get

$$(Q_V x, v) \geq \frac{t}{2} \|v\|^2$$

which forces $(Q_V x, v) \geq 0$. Thus $(Q_V x, v) = 0$, which means $Q_V x \in V^\perp$ and implies that

$$\|x\|^2 = \|P_V x + Q_V x\|^2 = \|P_V x\|^2 + \|Q_V x\|^2,$$

concluding the proof. □

The elements $P_V x$, $Q_V x$ are called **orthogonal projections** of x onto V and V^\perp , respectively. The greatest lower bound (6.22) is actually a minimum. Moreover, thanks to properties 1, 2, we say that H is direct sum of V and V^\perp , that is:

$$H = V \oplus V^\perp.$$

Note that

$$V^\perp = \{0\} \quad \text{if and only if} \quad V = H.$$

Remark 6.13. Another characterization of $P_V x$ is the following (see Problem 6.4): $u = P_V x$ if and only if

$$\begin{cases} \mathbf{1.} \ u \in V \\ \mathbf{2.} \ (x - u, v) = 0, \forall v \in V. \end{cases}$$

Remark 6.14. It is useful to point out that, even if V is not a closed subspace of H , the subspace V^\perp is always closed. In fact, if $y_n \rightarrow y$ and $\{y_n\} \subset V^\perp$, we have, for every $x \in V$,

$$(y, x) = \lim (y_n, x) = 0$$

whence $y \in V^\perp$.

Example 6.15. Let $\Omega \subset \mathbb{R}^n$ be a set of finite measure. Consider in $L^2(\Omega)$ the 1-dimensional subspace V of the constant functions (a basis is given by $f \equiv 1$, for instance). Since it is finite-dimensional, V is closed in $L^2(\Omega)$. Given $f \in L^2(\Omega)$, to find the projection $P_V f$, we solve the minimization problem

$$\min_{\lambda \in \mathbb{R}} \int_{\Omega} (f - \lambda)^2.$$

Since

$$\int_{\Omega} (f - \lambda)^2 = \int_{\Omega} f^2 - 2\lambda \int_{\Omega} f + \lambda^2 |\Omega|,$$

we see that the minimizer is

$$\lambda = \frac{1}{|\Omega|} \int_{\Omega} f.$$

Therefore

$$P_V f = \frac{1}{|\Omega|} \int_{\Omega} f \quad \text{and} \quad Q_V f = f - \frac{1}{|\Omega|} \int_{\Omega} f.$$

Thus, the subspace V^\perp is given by the functions $g \in L^2(\Omega)$ with *zero mean value*. In fact these functions are orthogonal to $f \equiv 1$:

$$(g, 1)_{L^2(\Omega)} = \int_{\Omega} g = 0.$$

6.4.2 Bases

A Hilbert space H is said to be **separable** when there exists a *countable dense* subset of H .

Definition 6.16. An orthonormal basis in a separable Hilbert space H is sequence $\{w_k\}_{k \geq 1} \subset H$ such that⁴

$$\begin{cases} (w_k, w_j) = \delta_{kj} & k, j \geq 1, \\ \|w_k\| = 1 & k \geq 1 \end{cases}$$

and every $x \in H$ may be expanded in the form

$$x = \sum_{k=1}^{\infty} (x, w_k) w_k. \tag{6.25}$$

The series (6.25) is called **generalized Fourier series** and the numbers

$$c_k = (x, w_k)$$

are the *Fourier coefficients* of x with respect to the basis $\{w_k\}_{k \geq 1}$. Moreover

⁴ δ_{kj} is the Kronecker symbol.

(Pythagoras again!):

$$\|x\|^2 = \sum_{k=1}^{\infty} (x, w_k)^2.$$

Given an orthonormal sequence $\{w_k\}_{k \geq 1}$, the projection of $x \in H$ onto the subspace V spanned by, say, w_1, \dots, w_N , is given by

$$P_V x = \sum_{k=1}^N (x, w_k) w_k.$$

Thus, for any other linear combination

$$S_N = \sum_{k=1}^N a_j w_j$$

we have

$$\|x - P_V x\| \leq \|x - S_N\|. \quad (6.26)$$

An example of separable Hilbert space is $L^2(\Omega)$, $\Omega \subseteq \mathbb{R}^n$. In particular, the set of functions

$$\frac{1}{\sqrt{2\pi}}, \frac{\cos x}{\sqrt{\pi}}, \frac{\sin x}{\sqrt{\pi}}, \frac{\cos 2x}{\sqrt{\pi}}, \frac{\sin 2x}{\sqrt{\pi}}, \dots, \frac{\cos mx}{\sqrt{\pi}}, \frac{\sin mx}{\sqrt{\pi}}, \dots$$

constitutes an orthonormal basis in $L^2(0, 2\pi)$ (see Appendix A).

The following proposition is useful to check that an orthonormal sequence $\{w_k\}_{k \geq 1}$ is a basis for H .

Proposition 6.17. *An orthonormal sequence $\{w_k\}_{k \geq 1} \subset H$ is a basis for H if and only if one of the following conditions is satisfied.*

- i) *If $x \in H$ is orthogonal to w_k for every $k \geq 1$, then $x = 0$.*
- ii) *The set of all finite linear combination of the w'_k 's is dense in H .*

Proof. If $\{w_k\}_{k \geq 1} \subset H$ is a basis for H and $x \in H$ is orthogonal to w_k for every $k \geq 1$, from (6.25) we get $x = 0$. Viceversa, if $x \in H$ is orthogonal to w_k for every $k \geq 1$, and $x \neq 0$, clearly x cannot be generated by the series (6.25). Thus $\{w_k\}_{k \geq 1}$ is a basis if and only if i) holds.

We prove the second part. If $\{w_k\}_{k \geq 1} \subset H$ is a basis for H , every x can be approximated by the partial sums of the Fourier series (6.25) and therefore ii) holds. Viceversa, if ii) holds, given $x \in H$ and $\varepsilon > 0$, we can find $S_N = \sum_{k=1}^N a_j w_j$ such that $\|x - S_N\| \leq \varepsilon$. From (6.26) we infer that

$$\left\| x - \sum_{k=1}^N (x, w_k) w_k \right\| \leq \|x - S_N\| \leq \varepsilon.$$

Since ε was arbitrary, (6.25) holds for x and we conclude that $\{w_k\}_{k \geq 1}$ is a basis for H . \square

It turns out that:

Proposition 6.18. *Every separable Hilbert space H admits a countable orthonormal basis.*

Proof. Let $\{z_k\}_{k \geq 1}$ be dense in H . Disregarding, if necessary, those elements which are spanned by other elements in the sequence, we may assume that $\{z_k\}_{k \geq 1}$ constitutes an independent set, i.e. every finite subset of $\{z_k\}_{k \geq 1}$ is composed by independent elements.

Then, an orthonormal basis $\{w_k\}_{k \geq 1}$ is obtained by applying to $\{z_k\}_{k \geq 1}$ the following so called *Gram-Schmidt process*. First, we construct by induction a sequence $\{\tilde{w}\}_{k \geq 1}$ as follows. Let $\tilde{w}_1 = z_1$. Once \tilde{w}_{k-1} is known, we construct \tilde{w}_k , $k \geq 2$, by subtracting from z_k its components with respect to $\tilde{w}_1, \dots, \tilde{w}_{k-1}$:

$$\tilde{w}_k = z_k - \frac{(z_k, \tilde{w}_{k-1})}{\|\tilde{w}_{k-1}\|^2} \tilde{w}_{k-1} - \dots - \frac{(z_k, \tilde{w}_1)}{\|\tilde{w}_1\|^2} \tilde{w}_1.$$

In this way, \tilde{w}_k is orthogonal to $\tilde{w}_1, \dots, \tilde{w}_{k-1}$. Finally, set

$$w_k = \frac{\tilde{w}_k}{\|\tilde{w}_k\|}.$$

Note that z_k is a linear combination of w_1, w_2, \dots, w_k . Since $\{z_k\}_{k \geq 1}$ is dense in H , then the set of all finite linear combination of the w'_k s is dense in H as well. By Proposition 6.17, $\{w_k\}_{k \geq 1}$ is a countable orthonormal basis in H . \square

In the applications, orthonormal bases arise from solving particular boundary value problems, often in relation to the separation of variables method. Typical examples come from the vibrations of a nonhomogeneous string or from heat conduction in a rod with nonconstant thermal properties c_v, ρ, κ . The first example leads to the wave equation

$$\rho(x) u_{tt} - \tau u_{xx} = 0.$$

Setting $u(x, t) = v(x) z(t)$, we find for the spatial factor the equation

$$\tau v'' + \lambda \rho v = 0.$$

The second example leads to the heat equation

$$c_v \rho u_t - (\kappa u_x)_x = 0$$

and, separating the variables, we find for $v = v(x)$ the equation

$$(\kappa v')' + \lambda c_v \rho v = 0.$$

These equations are particular cases of a general class of ordinary differential equations of the form

$$(pu')' + qu + \lambda wu = 0, \quad (6.27)$$

called *Sturm-Liouville* equations. Usually one looks for solutions of (6.27) in an interval (a, b) , $-\infty \leq a < b \leq +\infty$, satisfying suitable conditions at the end points. The natural assumptions on p and q are $p \neq 0$ in (a, b) and p, q, p^{-1} locally integrable in (a, b) . The function w plays the role of a *weight function*, continuous in $[a, b]$ and positive in (a, b) .

In general, the resulting boundary value problem has nontrivial solutions only for particular values of λ , called *eigenvalues*. The corresponding solutions are called *eigenfunctions* and it turns out that, when suitably normalized, they constitute an orthonormal basis in the Hilbert space $L_w^2(a, b)$, the set of Lebesgue measurable functions in (a, b) , such that

$$\|u\|_{L_w^2}^2 = \int_a^b u^2(x) w(x) dx < \infty,$$

endowed with the inner product

$$(u, v)_{L_w^2} = \int_a^b u(x) v(x) w(x) dx.$$

We list below some examples⁵.

- Consider the problem

$$\begin{cases} (1-x^2) u'' - xu' + \lambda u = 0 & \text{in } (-1, 1) \\ u(-1) < \infty, \quad u(1) < \infty. \end{cases}$$

The differential equation is known as *Chebyshev's* equation and may be written in the form (6.27):

$$((1-x^2)^{1/2} u')' + \lambda (1-x^2)^{-1/2} u = 0,$$

which shows the proper weight function $w(x) = (1-x^2)^{-1/2}$. The eigenvalues are $\lambda_n = n^2$, $n = 0, 1, 2, \dots$. The corresponding eigenfunctions are the *Chebyshev polynomials* T_n , recursively defined by $T_0(x) = 1$, $T_1(x) = x$ and

$$T_{n+1} = 2xT_n - T_{n-1} \quad (n > 1).$$

For instance:

$$T_2(x) = 2x^2 - 1, \quad T_3(x) = 4x^3 - 3x, \quad T_4(x) = 8x^4 - 8x^2 - 1.$$

The normalized polynomials

$$\sqrt{1/\pi} T_0, \sqrt{2/\pi} T_1, \dots, \sqrt{2/\pi} T_n, \dots$$

constitute an orthonormal basis in $L_w^2(-1, 1)$.

⁵ For the proofs, see [23], *Courant-Hilbert*, vol. I, 1953.

- Consider the problem⁶

$$\left((1 - x^2) u' \right)' + \lambda u = 0 \quad \text{in } (-1, 1)$$

with weighted Neumann conditions

$$(1 - x^2) u'(x) \rightarrow 0 \quad \text{as } x \rightarrow \pm 1.$$

The differential equation is known as *Legendre's* equation. The eigenvalues are $\lambda_n = n(n+1)$, $n = 0, 1, 2, \dots$. The corresponding eigenfunctions are the *Legendre polynomials*, defined by $L_0(x) = 1$, $L_1(x) = x$,

$$(n+1)L_{n+1} = (2n+1)xL_n - nL_{n-1} \quad (n > 1)$$

or by *Rodrigues' formula*

$$L_n(x) = \frac{1}{2^n n!} \frac{d^n}{dx^n} (x^2 - 1)^n \quad (n \geq 0).$$

For instance, $L_2(x) = (3x^2 - 1)/2$, $L_3(x) = (5x^3 - 3x)/2$. The normalized polynomials

$$\sqrt{\frac{2n+1}{2}} L_n$$

constitute an orthonormal basis in $L^2(-1, 1)$ (here $w(x) \equiv 1$). Every function $f \in L^2(-1, 1)$ has an expansion

$$f(x) = \sum_{n=0}^{\infty} f_n L_n(x)$$

where $f_n = \frac{2n+1}{2} \int_{-1}^1 f(x) L_n(x) dx$, with convergence in $L^2(-1, 1)$.

- Consider the problem

$$\begin{cases} u'' - 2xu' + 2\lambda u = 0 & \text{in } (-\infty, +\infty) \\ e^{-x^2/2}u(x) \rightarrow 0 & \text{as } x \rightarrow \pm\infty. \end{cases}$$

The differential equation is known as *Hermite's* equation (see Problem 6.6) and may be written in the form (6.27):

$$(e^{-x^2} u')' + 2\lambda e^{-x^2} u = 0$$

which shows the proper weight function $w(x) = e^{-x^2}$. The eigenvalues are $\lambda_n = n$, $n = 0, 1, 2, \dots$. The corresponding eigenfunctions are the *Hermite polynomials*, defined by *Rodrigues' formula*

$$H_n(x) = (-1)^n e^{x^2} \frac{d^n}{dx^n} e^{-x^2} \quad (n \geq 0).$$

⁶ See also Problem 8.5.

For instance

$$H_0(x) = 1, \quad H_1(x) = 2x, \quad H_2(x) = 4x^2 - 2, \quad H_3(x) = 8x^3 - 12x.$$

The normalized polynomials $(\sqrt{\pi}2^n n!)^{-1/2} H_n$ constitute an orthonormal basis in $L_w^2(\mathbb{R})$, with $w(x) = e^{-x^2}$. Every $f \in L_w^2(\mathbb{R})$ has an expansion

$$f(x) = \sum_{n=0}^{\infty} f_n H_n(x)$$

where $f_n = (\sqrt{\pi}2^n n!)^{-1} \int_{\mathbb{R}} f(x) H_n(x) e^{-x^2} dx$, with convergence in $L_w^2(\mathbb{R})$.

- After separating variables in the model for the vibration of a circular membrane, the following *parametric Bessel's equation of order p* arises (see Problem 6.8):

$$x^2 u'' + xu' + (\lambda x^2 - p^2) u = 0 \quad x \in (0, a) \quad (6.28)$$

where $p \geq 0$, $\lambda \geq 0$, with the boundary conditions

$$u(0) \text{ finite}, \quad u(a) = 0. \quad (6.29)$$

Equation (6.28) may be written in Sturm-Liouville form as

$$(xu')' + \left(\lambda x - \frac{p^2}{x} \right) u = 0$$

which shows the proper weight function $w(x) = x$. The simple rescaling $z = \sqrt{\lambda}x$ reduces (6.28) to the *Bessel's equation of order p*

$$z^2 \frac{d^2 u}{dz^2} + z \frac{du}{dz} + (z^2 - p^2) u = 0, \quad (6.30)$$

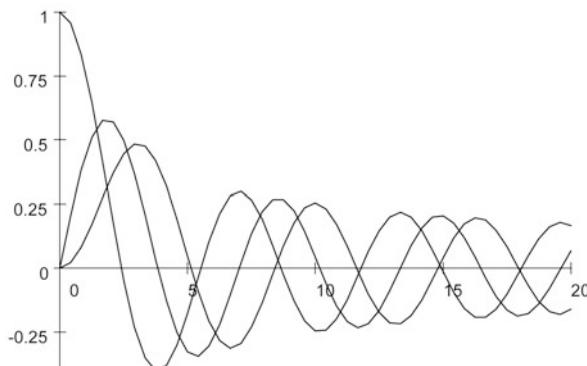


Fig. 6.2 Graphs of J_0, J_1 and J_2

where the dependence on the parameter λ is removed. The only bounded solutions of (6.30) are the *Bessel's functions of first kind and order p*, given by (see Fig. 6.2)

$$J_p(z) = \sum_{k=0}^{\infty} \frac{(-1)^k}{\Gamma(k+1)\Gamma(k+p+1)} \left(\frac{z}{2}\right)^{p+2k} \quad (6.31)$$

where

$$\Gamma(s) = \int_0^\infty e^{-t} t^{s-1} dt \quad (6.32)$$

is the Euler Γ -function. In particular, if $p = n \geq 0$, integer:

$$J_n(z) = \sum_{k=0}^{\infty} \frac{(-1)^k}{k!(k+n)!} \left(\frac{z}{2}\right)^{n+2k}.$$

For every p , there exists an infinite, increasing sequence $\{\alpha_{pj}\}_{j \geq 1}$ of positive zeroes of J_p :

$$J_p(\alpha_{pj}) = 0 \quad (j = 1, 2, \dots).$$

Then, the eigenvalues of problem (6.28), (6.29) are given by $\lambda_{pj} = \left(\frac{\alpha_{pj}}{a}\right)^2$, with corresponding eigenfunctions $u_{pj}(x) = J_p\left(\frac{\alpha_{pj}}{a}x\right)$. The normalized eigenfunctions

$$\frac{\sqrt{2}}{aJ_{p+1}(\alpha_{pj})} J_p\left(\frac{\alpha_{pj}}{a}x\right)$$

constitute an orthonormal basis in $L_w^2(0, a)$, with $w(x) = x$. Every function $f \in L_w^2(0, a)$ has an expansion in *Fourier-Bessel series*

$$f(x) = \sum_{j=1}^{\infty} f_j J_p\left(\frac{\alpha_{pj}}{a}x\right),$$

where

$$f_j = \frac{2}{a^2 J_{p+1}^2(\alpha_{pj})} \int_0^a x f(x) J_p\left(\frac{\alpha_{pj}}{a}x\right) dx,$$

convergent in $L_w^2(0, a)$.

6.5 Linear Operators and Duality

6.5.1 Linear operators

Let H_1 and H_2 be Hilbert spaces. A **linear operator** from H_1 into H_2 is a function

$$L : H_1 \rightarrow H_2$$

such that⁷, $\forall \alpha, \beta \in \mathbb{R}$ and $\forall x, y \in H_1$

$$L(\alpha x + \beta y) = \alpha Lx + \beta Ly.$$

For every linear operator we define its *Kernel*, $\mathcal{N}(L)$ and *Range*, $\mathcal{R}(L)$, as follows:

Definition 6.19. *The kernel of L is the pre-image of the null vector in H_2 :*

$$\mathcal{N}(L) = \{x \in H_1 : Lx = 0\}.$$

The range of L is the set of all outputs from points in H_1 :

$$\mathcal{R}(L) = \{y \in H_2 : \exists x \in H_1, Lx = y\}.$$

$\mathcal{N}(L)$ and $\mathcal{R}(L)$ are linear subspaces of H_1 and H_2 , respectively. If $\mathcal{N}(L) = \{0\}$ then we say that L is a *one-to-one* operator. In this case there exists an inverse operator $L^{-1} : \mathcal{R}(L) \rightarrow H_1$ such that $L \circ L^{-1} = I_{H_1}$, the identity operator in H_1 , and $L^{-1} \circ L = I_{\mathcal{R}(L)}$, the identity operator in $\mathcal{R}(L)$.

If $\mathcal{R}(L) = H_2$ we say that L is *onto*. If $L : H_1 \rightarrow H_2$ is one-to-one and onto, the equation

$$Lx = y \tag{6.33}$$

has one and only one solution $x \in H_1$ for every $y \in H_2$.

Our main objects will be linear bounded operators.

Definition 6.20. *A linear operator $L : H_1 \rightarrow H_2$ is bounded if there exists a number C such that*

$$\|Lx\|_{H_2} \leq C \|x\|_{H_1}, \quad \forall x \in H_1. \tag{6.34}$$

The number C controls the norm expansion operated by L on the elements of H_1 . In particular, if $C < 1$, L contracts the size of the vectors in H_1 .

If $x \neq 0$, using the linearity of L , we may write (6.34) in the form

$$\left\| L \left(\frac{x}{\|x\|_{H_1}} \right) \right\|_{H_2} \leq C,$$

which is equivalent to

$$\sup_{\|x\|_{H_1}=1} \|Lx\|_{H_2} = K(L) < \infty, \tag{6.35}$$

since $x/\|x\|_{H_1}$ is a unit vector in H_1 . Clearly $K(L) \leq C$.

⁷ Notation: if L is linear, when no confusion arises, we may write Lx instead of $L(x)$.

Proposition 6.21. *A linear operator $L : H_1 \rightarrow H_2$ is bounded if and only if it is continuous.*

Proof. Let L be bounded. From (6.34) we have, $\forall x, x_0 \in H_1$,

$$\|L(x - x_0)\|_{H_2} \leq C \|x - x_0\|_{H_1}$$

so that, if $\|x - x_0\|_{H_1} \rightarrow 0$, also $\|Lx - Lx_0\|_{H_2} = \|L(x - x_0)\|_{H_2} \rightarrow 0$. This shows the continuity of L .

Let L be continuous. In particular, L is continuous at $x = 0$ so that there exists δ such that

$$\|Lx\|_{H_2} \leq 1 \quad \text{if } \|x\|_{H_1} \leq \delta.$$

Choose now $y \in H_1$, with $\|y\|_{H_1} = 1$, and let $z = \delta y$. We have $\|z\|_{H_1} = \delta$ which implies

$$\delta \|Ly\|_{H_2} = \|Lz\|_{H_2} \leq 1$$

or

$$\|Ly\|_{H_2} \leq \frac{1}{\delta}$$

and (6.35) holds with $K \leq C = \frac{1}{\delta}$. □

Given two Hilbert spaces H_1 and H_2 , we denote by

$$\mathcal{L}(H_1, H_2)$$

the family of all linear bounded operators from H_1 into H_2 . If $H_1 = H_2$, we simply write $\mathcal{L}(H)$. $\mathcal{L}(H_1, H_2)$ becomes a linear space if we define, for $x \in H_1$ and $\lambda \in \mathbb{R}$,

$$\begin{aligned} (G + L)(x) &= Gx + Lx \\ (\lambda L)x &= \lambda Lx. \end{aligned}$$

Also, it is easy to check that the number $K(L)$ in (6.35) satisfies the properties of a norm and therefore we may choose it as a norm in $\mathcal{L}(H_1, H_2)$:

$$\|L\|_{\mathcal{L}(H_1, H_2)} = \sup_{\|x\|_{H_1}=1} \|Lx\|_{H_2}. \quad (6.36)$$

Thus, for every $L \in \mathcal{L}(H_1, H_2)$, we have

$$\|Lx\|_{H_2} \leq \|L\|_{\mathcal{L}(H_1, H_2)} \|x\|_{H_1}.$$

The resulting space is complete, so that:

Proposition 6.22. *Endowed with the norm (6.36), $\mathcal{L}(H_1, H_2)$ is a Banach space.*

Example 6.23. Let \mathbf{A} be an $m \times n$ real matrix. The map

$$L : \mathbf{x} \mapsto \mathbf{Ax}$$

is a linear operator from \mathbb{R}^n into \mathbb{R}^m . To compute $\|L\|_{\mathcal{L}(\mathbb{R}^n, \mathbb{R}^m)}$, note that

$$\|\mathbf{A}\mathbf{x}\|^2 = \mathbf{A}\mathbf{x} \cdot \mathbf{A}\mathbf{x} = \mathbf{A}^\top \mathbf{A}\mathbf{x} \cdot \mathbf{x}.$$

The matrix $\mathbf{A}^\top \mathbf{A}$ is symmetric and nonnegative and therefore, from Linear Algebra,

$$\sup_{\|\mathbf{x}\|=1} \mathbf{A}^\top \mathbf{A}\mathbf{x} \cdot \mathbf{x} = \Lambda_M$$

where Λ_M is the maximum eigenvalue of $\mathbf{A}^\top \mathbf{A}$. Thus, $\|L\|_{\mathcal{L}(\mathbb{R}^n, \mathbb{R}^m)} = \sqrt{\Lambda_M}$.

Example 6.24. Let V be a *closed subspace* of a Hilbert space H . The projections

$$x \mapsto P_V x, \quad x \mapsto Q_V x,$$

defined in Theorem 6.12, page 365, are bounded linear operators from H into H . In fact, from

$$\|x\|^2 = \|P_V x\|^2 + \|Q_V x\|^2,$$

it follows immediately that

$$\|P_V x\| \leq \|x\|, \quad \|Q_V x\| \leq \|x\|$$

so that (6.34) holds with $C = 1$. Since $P_V x = x$ when $x \in V$ and $Q_V x = x$ when $x \in V^\perp$, it follows that

$$\|P_V\|_{\mathcal{L}(H)} = \|Q_V\|_{\mathcal{L}(H)} = 1.$$

Finally, observe that

$$\mathcal{N}(P_V) = \mathcal{R}(Q_V) = V^\perp \quad \text{and} \quad \mathcal{N}(Q_V) = \mathcal{R}(P_V) = V.$$

Example 6.25. Let V and H be Hilbert spaces with⁸ $V \subset H$. Considering an element in V as an element of H , we define the operator $I_{V \rightarrow H} : V \rightarrow H$,

$$I_{V \rightarrow H}(u) = u,$$

which is called *embedding of V into H* . $I_{V \rightarrow H}$ is clearly a linear operator and it is also bounded if there exists a constant C such that

$$\|u\|_H \leq C \|u\|_V, \quad \text{for every } u \in V.$$

In this case, we say that V is *continuously embedded in H* and we write

$$V \hookrightarrow H.$$

For instance, $H_{per}^1(0, 2\pi) \hookrightarrow L^2(0, 2\pi)$.

⁸ The inner products in V and H may be different.

An important theorem in Functional Analysis states that⁹:

Theorem 6.26. *Let $L \in \mathcal{L}(H_1, H_2)$ be a one to one and onto. Then the inverse operator $L^{-1} : H_2 \rightarrow H_1$ is bounded.*

In other terms, Theorem 6.26 states that, if $L \in \mathcal{L}(H_1, H_2)$ and $x = L^{-1}y$ is the solution of equation $Lx = y$, the following estimate holds

$$\|x\|_{H_1} \leq \|L^{-1}\|_{\mathcal{L}(H_1, H_2)} \|y\|_{H_2}$$

for every $y \in H_2$.

6.5.2 Functionals and dual space

When $H_2 = \mathbb{R}$ (or \mathbb{C} , for complex Hilbert spaces), a linear operator $L : H \rightarrow \mathbb{R}$ takes the name of **functional**.

Definition 6.27. *The collection of all bounded linear functionals on a Hilbert space H is called the **dual space** of H and it is denoted by H^* (instead of $\mathcal{L}(H, \mathbb{R})$).*

Example 6.28. Let $H = L^2(\Omega)$, $\Omega \subseteq \mathbb{R}^n$ and fix $g \in L^2(\Omega)$. The functional defined by

$$L_g : f \longmapsto \int_{\Omega} fg$$

is linear and bounded. In fact, Schwarz's inequality yields

$$|L_g f| = \left| \int_{\Omega} fg \right| \leq \left(\int_{\Omega} |f|^2 \right)^{1/2} \left(\int_{\Omega} |g|^2 \right)^{1/2} = \|g\|_{L^2(\Omega)} \|f\|_{L^2(\Omega)}$$

so that $L_g \in L^2(\Omega)^*$ and $\|L_g\|_{L^2(\Omega)^*} \leq \|g\|_{L^2(\Omega)}$. Actually $\|L_g\|_{L^2(\Omega)^*} = \|g\|_{L^2(\Omega)}$ since, choosing $f = g$, we have

$$\|g\|_{L^2(\Omega)}^2 = L_g(g) \leq \|L_g\|_{L^2(\Omega)^*} \|g\|_{L^2(\Omega)}$$

whence also $\|L_g\|_{L^2(\Omega)^*} \geq \|g\|_{L^2(\Omega)}$.

Example 6.29. The functional in Example 6.28 is induced by the inner product with a fixed element in $L^2(\Omega)$. More generally, let H be a Hilbert space. For fixed $y \in H$, the functional

$$L_1 : x \longmapsto (x, y)$$

is continuous. In fact Schwarz's inequality yields

$$|(x, y)| \leq \|x\| \|y\|,$$

⁹ For the proof, see e.g. [39], Yoshida, 1971.

whence $L_1 \in H^*$ and $\|L_1\|_{H^*} \leq \|y\|$. Actually $\|L_1\|_{H^*} = \|y\|$ since, choosing $x = y$, we have

$$\|y\|^2 = |L_1 y| \leq \|L_1\|_{H^*} \|y\|,$$

or $\|L_1\|_{H^*} \geq \|y\|$. Observe that this argument provides the following alternative definition of the norm of an element $y \in H$:

$$\|y\| = \sup_{\|x\|=1} |(x, y)|. \quad (6.37)$$

To identify the dual space of a Hilbert space H is crucial in many instances. Example 6.29 shows that the inner product with a fixed element y in H defines an element of H^* , whose norm is exactly $\|y\|$. From Linear Algebra it is well known that *all* linear functionals in a finite-dimensional space can be represented in that way. Precisely, if L is linear in \mathbb{R}^n , there exists a vector $\mathbf{a} \in \mathbb{R}^n$ such that, for every $\mathbf{h} \in \mathbb{R}^n$,

$$L\mathbf{h} = \mathbf{a} \cdot \mathbf{h}$$

and $\|L\|_{(\mathbb{R}^n)^*} = |\mathbf{a}|$. The following theorem, known as **Riesz's representation Theorem**, says that an analogous result holds in Hilbert spaces.

Theorem 6.30. *Let H be a Hilbert space. For every $L \in H^*$ there exists a unique $u_L \in H$ such that:*

$$Lx = (u_L, x), \quad \forall x \in H.$$

Moreover, $\|L\|_{H^*} = \|u_L\|$.

Proof. Let \mathcal{N} be the kernel of L . If $\mathcal{N} = H$, then L is the *null operator* and $u_L = 0$. If $\mathcal{N} \subset H$, then \mathcal{N} is a *closed* subspace of H . In fact, if $\{x_n\} \subset \mathcal{N}$ and $x_n \rightarrow x$, then $0 = Lx_n \rightarrow Lx$ so that $x \in \mathcal{N}$; thus \mathcal{N} contains all its limit points and therefore is closed.

Then, by the Projection Theorem, there exists $z \in \mathcal{N}^\perp$, $z \neq 0$. Thus $Lz \neq 0$ and, given any $x \in H$, the element

$$w = x - \frac{Lx}{Lz} z$$

belongs to \mathcal{N} . In fact

$$Lw = L\left(x - \frac{Lx}{Lz} z\right) = Lx - \frac{Lx}{Lz} Lz = 0.$$

Since $z \in \mathcal{N}^\perp$, we have

$$0 = (z, w) = (z, x) - \frac{Lx}{Lz} \|z\|^2$$

which entails

$$Lx = \frac{L(z)}{\|z\|^2} (z, x).$$

Therefore if $u_L = L(z) \|z\|^{-2} z$, then $Lx = (u_L, x)$.

For the uniqueness, observe that, if $v \in H$ and

$$Lx = (v, x), \quad \text{for every } x \in H,$$

subtracting this equation from $Lx = (u_L, x)$, we infer

$$(u_L - v, x) = 0, \quad \text{for every } x \in H$$

which forces $v = u_L$.

To show $\|L\|_{H^*} = \|u_L\|$, use Schwarz's inequality

$$|(u_L, x)| \leq \|x\| \|u_L\|$$

to get

$$\|L\|_{H^*} = \sup_{\|x\|=1} |Lx| = \sup_{\|x\|=1} |(u_L, x)| \leq \|u_L\|.$$

On the other hand,

$$\|u_L\|^2 = (u_L, u_L) = Lu_L \leq \|L\|_{H^*} \|u_L\|$$

whence

$$\|u_L\| \leq \|L\|_{H^*}.$$

Thus $\|L\|_{H^*} = \|u_L\|$. □

The Riesz map $\mathcal{R}_H : H^* \rightarrow H$ given by

$$L \longmapsto u_L = \mathcal{R}_H L$$

is a *canonical isometry* and we say that u_L is the *Riesz element associated with* L , with respect to the scalar product (\cdot, \cdot) . Moreover, H^* endowed with the inner product

$$(L_1, L_2)_{H^*} = (u_{L_1}, u_{L_2}) \tag{6.38}$$

is clearly a Hilbert space. Thus, in the end, the Representation Theorem 6.30, allows the **identification of a Hilbert space with its dual**.

Example 6.31. $L^2(\Omega)$ is identified with its dual.

Remark 6.32. A few words about **notations**. The symbol (\cdot, \cdot) or $(\cdot, \cdot)_H$ denotes the inner product in a Hilbert space H . Let now $L \in H^*$. For the *action* of the functional L on an element $x \in H$ we used the symbol Lx . Sometimes, when it is useful or necessary to emphasize the *duality (or pairing)* between H and H^* , we shall use the notation $\langle L, x \rangle_*$ or even $\langle L, x \rangle_{H^*, H}$.

6.5.3 The adjoint of a bounded operator

The concept of *adjoint operator* extends the notion of transpose of an $m \times n$ matrix \mathbf{A} and plays a crucial role in several situations, for instance, in determining compatibility conditions for the solvability of several problems. The transpose \mathbf{A}^\top is characterized by the identity

$$(\mathbf{A}\mathbf{x}, \mathbf{y})_{\mathbb{R}^m} = (\mathbf{x}, \mathbf{A}^\top \mathbf{y})_{\mathbb{R}^n}, \quad \forall \mathbf{x} \in \mathbb{R}^n, \forall \mathbf{y} \in \mathbb{R}^m.$$

We extend precisely this relation to define the adjoint of a bounded linear operator. Let $L \in \mathcal{L}(H_1, H_2)$. If $y \in H_2$ is fixed, the real map

$$T_y : x \longmapsto (Lx, y)_{H_2}$$

defines an element of H_1^* . In fact

$$|T_y x| = |(Lx, y)_{H_2}| \leq \|Lx\|_{H_2} \|y\|_{H_2} \leq \|L\|_{\mathcal{L}(H_1, H_2)} \|y\|_{H_2} \|x\|_{H_1}$$

so that $\|T_y\|_{H_1^*} \leq \|L\|_{\mathcal{L}(H_1, H_2)} \|y\|_{H_2}$.

From Riesz's Theorem, there exists a unique $w \in H_1$ depending on y , which we denote by $w = L^*y$, such that

$$T_y x = (x, L^*y)_{H_1}, \quad \forall x \in H_1, \forall y \in H_2.$$

This defines L^* as an operator from H_2 into H_1 , which is called the (Hilbert) *adjoint of L* . Precisely:

Definition 6.33. *The operator $L^* : H_2 \rightarrow H_1$ defined by the identity*

$$(Lx, y)_{H_2} = (x, L^*y)_{H_1}, \quad \forall x \in H_1, \forall y \in H_2 \tag{6.39}$$

is called the adjoint of L .

Example 6.34. Let $L : H_1 \rightarrow H_2$ be an isometry. Then $L^* = L^{-1}$. In fact, since L preserves the inner product, we can write

$$(x, L^*y)_{H_1} = (Lx, y)_{H_2} = (x, L^{-1}y)_{H_1}, \quad \forall x \in H_1, \forall y \in H_2.$$

In particular, if $\mathcal{R}_H : H^* \rightarrow H$ is the Riesz map, then $\mathcal{R}_H^* = \mathcal{R}_H^{-1} : H \rightarrow H^*$.

Example 6.35. Let $T : L^2(0, 1) \rightarrow L^2(0, 1)$ be the linear map

$$Tu(x) = \int_0^x u(t) dt.$$

Schwarz's inequality gives

$$\left| \int_0^x u \right|^2 \leq x \int_0^x u^2,$$

whence

$$\|Tu\|_{L^2(0,1)}^2 = \int_0^1 |Tu|^2 = \int_0^1 \left| \int_0^x u \right|^2 dx \leq \int_0^1 (x \int_0^x u^2) dx \leq \frac{1}{2} \int_0^1 u^2 \leq \frac{1}{2} \|u\|_{L^2(0,1)}^2$$

and therefore T is bounded. To compute T^* , observe that

$$\begin{aligned} (Tu, v)_{L^2(0,1)} &= \int_0^1 [v(x) \int_0^x u(y) dy] dx = \text{exchanging the order of integration} \\ &= \int_0^1 [u(y) \int_x^1 v(x) dx] dy = (u, T^*v)_{L^2(0,1)}. \end{aligned}$$

Thus,

$$T^*v(x) = \int_x^1 v(t) dt.$$

Symmetric matrices correspond to selfadjoint operators. We say that L is **self-adjoint** if $H_1 = H_2$ and $L^* = L$. Then, (6.39) reduces to

$$(Lx, y) = (x, Ly).$$

An example of selfadjoint operator in a Hilbert space H is the projection P_V on a closed subspace of H ; in fact, recalling the Projection Theorem:

$$(P_Vx, y) = (P_Vx, P_Vy + Q_Vy) = (P_Vx, P_Vy) = (P_Vx + Q_Vx, P_Vy) = (x, P_Vy).$$

Important self-adjoint operators are associated with *inverses of differential operators*, as we will see in Chap. 8.

The following properties are immediate consequences of the definition of adjoint (for the proof, see Problem 6.11).

Proposition 6.36. *Let $L, L_1 \in \mathcal{L}(H_1, H_2)$ and $L_2 \in \mathcal{L}(H_2, H_3)$. Then:*

(a) $L^* \in \mathcal{L}(H_2, H_1)$. Moreover $(L^*)^* = L$ and

$$\|L^*\|_{\mathcal{L}(H_2, H_1)} = \|L\|_{\mathcal{L}(H_1, H_2)}.$$

(b) $(L_2 L_1)^* = L_1^* L_2^*$. In particular, if L is one-to-one and onto, then

$$(L^{-1})^* = (L^*)^{-1}.$$

The next theorem extends some relations, well known in the finite-dimensional case.

Theorem 6.37. *Let $L \in \mathcal{L}(H_1, H_2)$. Then*

- a) $\overline{\mathcal{R}(L)} = \mathcal{N}(L^*)^\perp$.
- b) $\mathcal{N}(L) = \mathcal{R}(L^*)^\perp$.

Proof. a) Let $z \in \mathcal{R}(L)$. Then, there exists $x \in H_1$ such that $z = Lx$ and, if $y \in \mathcal{N}(L^*)$, we have

$$(z, y)_{H_2} = (Lx, y)_{H_2} = (x, L^*y)_{H_1} = 0.$$

Thus, $\mathcal{R}(L) \subseteq \mathcal{N}(L^*)^\perp$. Since $\mathcal{N}(L^*)^\perp$ is closed¹⁰, it follows that

$$\overline{\mathcal{R}(L)} \subseteq \mathcal{N}(L^*)^\perp$$

as well. On the other hand, if $z \in \mathcal{R}(L)^\perp$, for every $x \in H_1$ we have

$$0 = (Lx, z)_{H_2} = (x, L^*z)_{H_1}$$

whence $L^*z = 0$. Therefore $\mathcal{R}(L)^\perp \subseteq \mathcal{N}(L^*)$, equivalent to $\mathcal{N}(L^*)^\perp \subseteq \overline{\mathcal{R}(L)}$.

b) Letting $L = L^*$ in a) we deduce

$$\overline{\mathcal{R}(L^*)} = \mathcal{N}(L)^\perp,$$

equivalent to $\mathcal{R}(L^*)^\perp = \mathcal{N}(L)$. □

6.6 Abstract Variational Problems

6.6.1 Bilinear forms and the Lax-Milgram Theorem

In the variational formulation of boundary value problems a key role is played by *bilinear forms*. Given two linear spaces V_1, V_2 , a **bilinear form** in $V_1 \times V_2$ is a function

$$a : V_1 \times V_2 \rightarrow \mathbb{R}$$

satisfying the following properties:

- i) For every $v \in V_2$, the function $u \mapsto a(u, v)$ is linear in V_1 .
- ii) For every $u \in V_1$, the function $v \mapsto a(u, v)$ is linear in V_2 .

When $V_1 = V_2$, we simply say that a is a *bilinear form in V* .

Remark 6.38. In complex inner product spaces we define *sesquilinear forms*, instead of bilinear forms, replacing ii) by:

ii)_{bis}) for every $x \in V_1$, the function $y \mapsto a(x, y)$ is *anti-linear*¹¹ in V_2 .

Here are some examples.

- A typical example of bilinear form in a Hilbert space is its inner product.
- The formula

$$a(u, v) = \int_a^b (p(x)u'v' + q(x)u'v + r(x)uv) dx$$

where p, q, r are bounded functions, defines a bilinear form in $C^1([a, b])$.

¹⁰ See Remark 6.14, p. 366.

¹¹ That is

$$a(x, \alpha y + \beta z) = \bar{\alpha}a(x, y) + \bar{\beta}a(x, z).$$

More generally, if Ω is a bounded domain in \mathbb{R}^n ,

$$a(u,v) = \int_{\Omega} (\alpha \nabla u \cdot \nabla v + u \mathbf{b}(\mathbf{x}) \cdot \nabla v + a_0(\mathbf{x}) uv) d\mathbf{x} \quad (\alpha > 0),$$

or

$$a(u,v) = \int_{\Omega} \alpha \nabla u \cdot \nabla v d\mathbf{x} + \int_{\partial\Omega} huv d\sigma \quad (\alpha > 0),$$

(\mathbf{b} , a_0 , h bounded) are bilinear forms in $C^1(\overline{\Omega})$.

- A bilinear form in $C^2(\overline{\Omega})$ involving higher order derivatives is

$$a(u,v) = \int_{\Omega} \Delta u \Delta v d\mathbf{x}.$$

Let V be a Hilbert space, a be a bilinear form in V and $F \in V^*$. Consider the following problem, called *abstract variational problem*:

$$\begin{cases} \text{Find } u \in V \text{ such that} \\ a(u,v) = Fv, \quad \forall v \in V. \end{cases} \quad (6.40)$$

As we shall see, many boundary values problems can be recast in this form. The fundamental result is the following one, known as the **Lax-Milgram Theorem**:

Theorem 6.39. *Let V be a (real) Hilbert space endowed with inner product (\cdot, \cdot) and norm $\|\cdot\|$. Let $a = a(u, v)$ be a bilinear form in V . If:*

- i) *a is continuous, i.e. there exists a constant M such that*

$$|a(u,v)| \leq M \|u\| \|v\|, \quad \forall u, v \in V.$$

- ii) *a is V -coercive, i.e. there exists a constant $\alpha > 0$ such that*

$$a(v,v) \geq \alpha \|v\|^2, \quad \forall v \in V, \quad (6.41)$$

then there exists a unique solution $\bar{u} \in V$ of problem (6.40). Moreover, the following stability estimate holds:

$$\|\bar{u}\| \leq \frac{1}{\alpha} \|F\|_{V^*}. \quad (6.42)$$

Remark 6.40. The coercivity inequality (6.41) may be considered as an abstract version of the *energy* or *integral estimates* we met in the previous chapters. Usually, it is the key estimate to prove in order to apply Theorem 6.39. We shall come back to the general solvability of a variational problem in Sect. 6.8, when a is not V -coercive.

Proof of Theorem 6.39. We split it into several steps.

1. Reformulation of problem (6.40). For every fixed $u \in V$, by the continuity of a , the linear map

$$v \mapsto a(u, v)$$

is bounded in V and therefore it defines an element of V^* . From Riesz's Representation Theorem, there exists a unique $A[u] \in V$ such that

$$a(u, v) = (A[u], v), \quad \forall v \in V. \quad (6.43)$$

Since $F \in V^*$ as well, there exists a unique $z_F \in V$ such that

$$Fv = (z_F, v), \quad \forall v \in V$$

and moreover $\|F\|_{V^*} = \|z_F\|$. Then, problem (6.40) can be recast in the following way:

$$\begin{cases} \text{Find } u \in V \text{ such that} \\ (A[u], v) = (z_F, v), \quad \forall v \in V \end{cases}$$

which, in turn, is equivalent to **finding u such that**

$$A[u] = z_F. \quad (6.44)$$

We want to show that (6.44) has exactly one solution which means that $A : V \rightarrow V$ is a *linear, continuous, one-to-one, surjective map*.

2. Linearity and continuity of A . We repeatedly use the definition of A and the bilinearity of a . To show linearity, we write, for every $u_1, u_2, v \in V$ and $\lambda_1, \lambda_2 \in \mathbb{R}$,

$$\begin{aligned} (A[\lambda_1 u_1 + \lambda_2 u_2], v) &= a(\lambda_1 u_1 + \lambda_2 u_2, v) = \lambda_1 a(u_1, v) + \lambda_2 a(u_2, v) \\ &= \lambda_1 (A[u_1], v) + \lambda_2 (A[u_2], v) = (\lambda_1 A[u_1] + \lambda_2 A[u_2], v) \end{aligned}$$

whence

$$A[\lambda_1 u_1 + \lambda_2 u_2] = \lambda_1 A[u_1] + \lambda_2 A[u_2].$$

Thus A is linear and we may write Au instead of $A[u]$. For the continuity, observe that

$$\|Au\|^2 = (Au, Au) = a(u, Au) \leq M \|u\| \|Au\|$$

whence

$$\|Au\| \leq M \|u\|.$$

3. A is one-to-one and has closed range, i.e.

$$\mathcal{N}(A) = \{0\} \quad \text{and} \quad \mathcal{R}(A) \text{ is a closed subspace of } V.$$

In fact, the coercivity of a yields

$$\alpha \|u\|^2 \leq a(u, u) = (Au, u) \leq \|Au\| \|u\|$$

whence

$$\|u\| \leq \frac{1}{\alpha} \|Au\|. \quad (6.45)$$

Thus, $Au = 0$ implies $u = 0$ and hence $\mathcal{N}(A) = \{0\}$.

To prove that $\mathcal{R}(A)$ is closed we have to consider a sequence $\{y_m\} \subset \mathcal{R}(A)$ such that

$$y_m \rightarrow y \in V$$

as $m \rightarrow \infty$, and show that $y \in \mathcal{R}(A)$. Since $y_m \in \mathcal{R}(A)$, there exists u_m such that $Au_m = y_m$. From (6.45) we infer

$$\|u_k - u_m\| \leq \frac{1}{\alpha} \|y_k - y_m\|$$

and therefore, since $\{y_m\}$ is convergent, $\{u_m\}$ is a Cauchy sequence. Since V is complete, there exists $u \in V$ such that

$$u_m \rightarrow u$$

and the continuity of A yields $y_m = Au_m \rightarrow Au$. Thus $Au = y$, so that $y \in \mathcal{R}(A)$ and $\mathcal{R}(A)$ is closed.

4. *A is surjective, that is $\mathcal{R}(A) = V$.* Suppose $\mathcal{R}(A) \subset V$. Since $\mathcal{R}(A)$ is a closed subspace, by the Projection Theorem there exists $z \neq 0$, $z \in \mathcal{R}(A)^\perp$. In particular, this implies

$$0 = (Az, z) = a(z, z) \geq \alpha \|z\|^2$$

whence $z = 0$. Contradiction. Therefore $\mathcal{R}(A) = V$.

5. *Solution of problem (6.40).* Since A is one-to-one and $\mathcal{R}(A) = V$, there exists exactly one solution $\bar{u} \in V$ of equation

$$Au = z_F.$$

From point **1**, \bar{u} is the unique solution of problem (6.40) as well.

6. *Stability estimate.* From (6.45) with $u = \bar{u}$, we obtain

$$\|\bar{u}\| \leq \frac{1}{\alpha} \|A\bar{u}\| = \frac{1}{\alpha} \|z_F\| = \frac{1}{\alpha} \|F\|_{V^*}$$

and the proof is complete. \square

Remark 6.41. Since a is a continuous bilinear form, we can define an operator $L \in \mathcal{L}(V, V^*)$ that associates to every $u \in V$ the functional $Lu \in V^*$ given by

$$v \mapsto a(u, v).$$

In other terms, L is uniquely defined by the relation

$$\langle Lu, v \rangle_* = a(u, v), \quad \forall u, v \in V.$$

Then the abstract variational problem is equivalent to the equation

$$Lu = F \quad \text{in } V^*$$

and by the Lax-Milgram Theorem L is a *continuous isomorphism between V and V^** . The inverse operator L^{-1} acts as the solution map $F \longmapsto u(F)$ and inequality (6.42) yields

$$\|L^{-1}\|_{\mathcal{L}(V^*, V)} \leq \frac{1}{\alpha}.$$

This inequality (or (6.42)) is called *stability estimate* for the following reason. The functional F , an element of V^* , encodes the “data” of the problem (6.40). If $F_1, F_2 \in V^*$ and u_1, u_2 are the corresponding solutions, inequality (6.42) gives

$$\|u_1 - u_2\| \leq \frac{1}{\alpha} \|F_1 - F_2\|_{V^*}.$$

Thus, close data imply close solutions. The stability constant $1/\alpha$ plays an important role, since it controls the norm-variation of the solutions in terms of the variations on the data, measured by $\|F_1 - F_2\|_{V^*}$. This entails, in particular, that the more the coerciveness constant α is large, the more “stable” is the solution.

Some applications require the solution to be in some Hilbert space W , while asking the variational equation

$$a(u, v) = Fv$$

to hold for every $v \in V$, with $V \neq W$. A variant of Theorem 6.39 deals with this asymmetric situation. Let $F \in V^*$ and $a = a(u, v)$ be a bilinear form in $W \times V$ satisfying the following three hypotheses:

i) *There exists M such that*

$$|a(u, v)| \leq M \|u\|_W \|v\|_V, \quad \forall u \in W, \forall v \in V.$$

ii) *There exists $\alpha > 0$ such that*

$$\sup_{\|v\|_V=1} a(u, v) \geq \alpha \|u\|_W, \quad \forall u \in W.$$

iii)

$$\sup_{w \in W} a(w, v) > 0, \quad \forall v \in V.$$

Condition ii) is an asymmetric coercivity, while iii) assures that, for every fixed $v \in V$, $a(v, \cdot)$ is positive at some point in W . We have the following theorem, due to Nečas (for the proof, see Problem 6.12):

Theorem 6.42. *If i), ii), iii) hold, there exists a unique $u \in W$ such that*

$$a(u, v) = Fv \quad \forall v \in V.$$

Moreover

$$\|u\|_W \leq \frac{1}{\alpha} \|F\|_{V^*}. \tag{6.46}$$

6.6.2 Minimization of quadratic functionals

When a is *symmetric*, i.e. if

$$a(u, v) = a(v, u), \quad \forall u, v \in V,$$

the abstract variational problem (6.40) is equivalent to a *minimization* problem. In fact, consider the quadratic functional

$$E(v) = \frac{1}{2}a(v, v) - Fv.$$

We have:

Theorem 6.43. *Let a be symmetric. Then \bar{u} is solution of problem (6.40) if and only if \bar{u} is a minimizer of E , that is*

$$E(\bar{u}) = \min_{v \in V} E(v).$$

Proof. For every $\varepsilon \in \mathbb{R}$ and every “variation” $v \in V$ we have

$$\begin{aligned} & E(\bar{u} + \varepsilon v) - E(\bar{u}) \\ &= \left\{ \frac{1}{2}a(\bar{u} + \varepsilon v, \bar{u} + \varepsilon v) - F(\bar{u} + \varepsilon v) \right\} - \left\{ \frac{1}{2}a(\bar{u}, \bar{u}) - F\bar{u} \right\} \\ &= \varepsilon \{a(\bar{u}, v) - Fv\} + \frac{1}{2}\varepsilon^2 a(v, v). \end{aligned}$$

Now, if \bar{u} is the solution of problem (6.40), then $a(\bar{u}, v) - Fv = 0$. Therefore

$$E(\bar{u} + \varepsilon v) - E(\bar{u}) = \frac{1}{2}\varepsilon^2 a(v, v) \geq 0$$

so that \bar{u} minimizes E . On the other hand, if \bar{u} is a minimizer of E , then

$$E(\bar{u} + \varepsilon v) - E(\bar{u}) \geq 0,$$

which entails

$$\varepsilon \{a(\bar{u}, v) - Fv\} + \frac{1}{2}\varepsilon^2 a(v, v) \geq 0$$

for every $\varepsilon \in \mathbb{R}$. This inequality forces

$$a(\bar{u}, v) - Fv = 0, \quad \forall v \in V, \tag{6.47}$$

and \bar{u} is a solution of problem (6.40)). \square

Letting $\varphi(\varepsilon) = E(\bar{u} + \varepsilon v)$, from the above calculations we have

$$\varphi'(0) = a(\bar{u}, v) - Fv.$$

Thus, the linear functional

$$v \mapsto a(\bar{u}, v) - Fv$$

appears as the **derivative of E** at \bar{u} along the direction v and we write

$$E'(\bar{u})v = a(\bar{u}, v) - Fv. \quad (6.48)$$

In Calculus of Variations E' is called **first variation** and denoted by δE . The *variational equation*

$$E'(u)v = a(u, v) - Fv = 0, \quad \forall v \in V \quad (6.49)$$

is called the **Euler equation** for the functional E .

Remark 6.44. A bilinear form a , symmetric and coercive, induces in V the inner product

$$(u, v)_a = a(u, v).$$

In this case, existence, uniqueness and stability for problem (6.40) follow directly from Riesz's Representation Theorem. In particular, *there exists a unique minimizer \bar{u} of E* .

6.6.3 Approximation and Galerkin method

The solution u of the abstract variational problem (6.40), satisfies the equation

$$a(u, v) = Fv \quad (6.50)$$

for *every* v in the Hilbert space V . In concrete applications, it is important to compute approximate solutions with a given degree of accuracy and the infinite dimension of V is the main obstacle. Often, however, V is separable and may be written as a *union of finite-dimensional subspaces*, so that, in principle, it could be reasonable to obtain approximate solutions by “projecting” equation (6.50) on those subspaces. This is the idea of **Galerkin's method**. In principle, the higher the dimension of the subspace the better should be the degree of approximation. More precisely, the idea is to construct a sequence $\{V_k\}$ of subspaces of V with the following properties:

- a) Every V_k is *finite-dimensional*: $\dim V_k = k$.
- b) $V_k \subset V_{k+1}$ (actually, not strictly necessary).
- c) $\overline{\cup V_k} = V$.

To realize the projection, assume that the vectors $\psi_1, \psi_2, \dots, \psi_k$ span V_k . Then, we look for an approximation of the solution u in the form

$$u_k = \sum_{j=1}^k c_j \psi_j, \quad (6.51)$$

by solving the problem

$$a(u_k, v) = Fv \quad \forall v \in V_k. \quad (6.52)$$

Since $\{\psi_1, \psi_2, \dots, \psi_k\}$ constitutes a basis in V_k , (6.52) amounts to requiring

$$a(u_k, \psi_r) = F\psi_r \quad r = 1, \dots, k. \quad (6.53)$$

Substituting (6.51) into (6.53), we obtain the k linear algebraic equations

$$\sum_{j=1}^k a(\psi_j, \psi_r) c_j = F\psi_r \quad r = 1, 2, \dots, k, \quad (6.54)$$

for the unknown coefficients c_1, c_2, \dots, c_k . Introducing the vectors

$$\mathbf{c} = \begin{pmatrix} c_1 \\ c_2 \\ \vdots \\ c_k \end{pmatrix}, \quad \mathbf{F} = \begin{pmatrix} F\psi_1 \\ F\psi_2 \\ \vdots \\ F\psi_k \end{pmatrix}$$

and the matrix $\mathbf{A} = (a_{rj})$, with entries

$$a_{rj} = a(\psi_j, \psi_r), \quad j, r = 1, \dots, k,$$

we may write (6.54) in the compact form

$$\mathbf{Ac} = \mathbf{F}. \quad (6.55)$$

The matrix \mathbf{A} is called *stiffness matrix* and clearly plays a key role in the numerical analysis of the problem.

If the bilinear form a is coercive, \mathbf{A} is *strictly positive*. In fact, let $\xi \in \mathbb{R}^k$. Then, by linearity and coercivity:

$$\begin{aligned} \mathbf{A}\xi \cdot \xi &= \sum_{r,j=1}^k a_{rj} \xi_r \xi_j = \sum_{r,j=1}^k a(\psi_j, \psi_r) \xi_r \xi_j \\ &= \sum_{r,j=1}^k a(\xi_j \psi_j, \xi_r \psi_r) = a\left(\sum_{i=1}^k \xi_j \psi_j, \sum_{j=1}^k \xi_r \psi_r\right) \\ &\geq \alpha \|\mathbf{v}\|^2 \end{aligned}$$

where

$$\mathbf{v} = \sum_{j=1}^k \xi_j \psi_j \in V_k.$$

Since $\{\psi_1, \psi_2, \dots, \psi_k\}$ is a basis in V_k , we have $\mathbf{v} = \mathbf{0}$ if and only if $\xi = \mathbf{0}$. Therefore \mathbf{A} is strictly positive and, in particular, nonsingular.

Thus, for each $k \geq 1$, there exists a unique solution $u_k \in V_k$ of (6.55). We want to show that $u_k \rightarrow u$, as $k \rightarrow \infty$, i.e. the convergence of the method, and give a control of the approximation error.

For this purpose, we prove the following lemma, known as *Céa's Lemma*, which also emphasizes the role of the continuity and the coercivity constants (M and α , respectively) of the bilinear form a .

Lemma 6.45. *Assume that the hypotheses of the Lax-Milgram Theorem hold and let u be the solution of problem (6.40). If u_k is the solution of problem (6.53), then*

$$\|u - u_k\| \leq \frac{M}{\alpha} \inf_{v \in V_k} \|u - v\|. \quad (6.56)$$

Proof. We have

$$a(u_k, v) = Fv, \quad \forall v \in V_k$$

and

$$a(u, v) = Fv, \quad \forall v \in V_k.$$

Subtracting the two equations we obtain

$$a(u - u_k, v) = 0, \quad \forall v \in V_k.$$

In particular, since $v - u_k \in V_k$, we have

$$a(u - u_k, v - u_k) = 0, \quad \forall v \in V_k,$$

which implies

$$\begin{aligned} a(u - u_k, u - u_k) &= a(u - u_k, u - v) + a(u - u_k, v - u_k) \\ &= a(u - u_k, u - v). \end{aligned}$$

Then, by the continuity and the coercivity of a ,

$$\alpha \|u - u_k\|^2 \leq a(u - u_k, u - u_k) \leq M \|u - u_k\| \|u - v\|,$$

whence,

$$\|u - u_k\| \leq \frac{M}{\alpha} \|u - v\|. \quad (6.57)$$

This inequality holds for every $v \in V_k$, with $\frac{M}{\alpha}$ independent of k . Therefore (6.57) still holds if we take in the right hand side the infimum over all $v \in V_k$. \square

Convergence of Galerkin's method. Since we have assumed that

$$\overline{\cup V_k} = V,$$

there exists a sequence $\{w_k\} \subset V_k$ such that $w_k \rightarrow u$ as $k \rightarrow \infty$. Lemma 6.45 gives, for every k :

$$\begin{aligned} \|u - u_k\| &\leq \frac{M}{\alpha} \inf_{v \in V_k} \|u - v\| \\ &\leq \frac{M}{\alpha} \|u - w_k\|, \end{aligned}$$

whence $\|u - u_k\| \rightarrow 0$.

6.7 Compactness and Weak Convergence

6.7.1 Compactness

The solvability of boundary value problems and the analysis of numerical methods involve several questions of convergence. In typical situations one is able to construct a sequence of approximations and the main task is to prove that this sequence converges to a solution of the problem in a suitable sense. It is often the case that, through *energy type estimates*, one is able to show that these sequences of approximations are *bounded* in some Hilbert space. How can we use this information? Although we cannot expect these sequences to converge, we may reasonably look for *convergent subsequences*, which is already quite satisfactory. In technical words, we are asking to our sequences to have a *compactness property*. Let us spend a few words on this important topological concept¹². Once more, the difference between finite and infinite dimension plays a big role.

Let X be a normed space. The general definition of compact set involves open coverings: an *open covering* of $E \subseteq X$ is a family of *open* sets whose union contains E .

Theorem 6.46. *We say that $E \subseteq X$ is compact if from every open covering of E it is possible to extract a finite subcovering of E .*

It is somewhat more convenient to work with **precompact** sets as well, that is sets whose **closure** is compact. In finite dimensional spaces the characterization of pre-compact sets is well known: $E \subset \mathbb{R}^n$ is pre-compact if and only if E is bounded. What about infinitely many dimensions? Let us introduce a characterization of precompact sets in normed spaces in terms of convergent sequences, much more comfortable to use. First, let us agree that a subset E of a normed space X is *sequentially precompact* (resp. *compact*), if for every sequence $\{x_k\} \subset E$ there exists a subsequence $\{x_{k_s}\}$, convergent in X (resp. in E). We have:

Theorem 6.47. *Let X be a normed space and $E \subset X$. Then E is precompact (compact) if and only if it is sequentially precompact (compact).*

While a compact set is always *closed and bounded* (see Problem 6.13), the following example exhibits a closed and bounded set which is *not* compact.

Example 6.48. Consider the real Hilbert space (see Problem 6.3)

$$l^2 = \left\{ \mathbf{x} = \{x_k\}_{k \geq 1} : \sum_{k=1}^{\infty} x_k^2 < \infty, x_k \in \mathbb{R} \right\}$$

¹² For the proofs, see e.g. [39], Yosida, 1971.

endowed with

$$(\mathbf{x}, \mathbf{y}) = \sum_{k=1}^{\infty} x_k y_k \quad \text{and} \quad \|\mathbf{x}\|^2 = \sum_{k=1}^{\infty} x_k^2.$$

Let $E = \{\mathbf{e}^k\}_{k \geq 1}$, where $\mathbf{e}^1 = \{1, 0, 0, \dots\}$, $\mathbf{e}^2 = \{0, 1, 0, \dots\}$, etc.. Observe that E constitutes an orthonormal basis in l^2 . Then, E is closed and bounded in l^2 . However, E is not sequentially compact. Indeed, if $j \neq k$,

$$\|\mathbf{e}^j - \mathbf{e}^k\| = \sqrt{2}$$

and therefore no subsequence of $\{\mathbf{e}^k\}_{k \geq 1}$ can be convergent.

Thus, in infinite-dimensions, closed and bounded does not imply compact. Actually, this can only happen in finite-dimensional spaces. In fact:

Theorem 6.49. *Let B be a Banach space. The unit ball $\{\mathbf{x} : \|\mathbf{x}\| \leq 1\}$ is compact if and only if B is finite-dimensional.*

6.7.2 Compactness in $C(\overline{\Omega})$ and in $L^p(\Omega)$

To recognize that a subset of a Banach or Hilbert space is compact is usually a hard task. A characterization of the compact subset in $C(\overline{\Omega})$, where Ω is a bounded domain is given by the following important theorem, known as the *Ascoli-Arzelà Theorem*¹³:

Theorem 6.50. *Let $\Omega \subset \mathbb{R}^n$ be a bounded domain and $E \subset C(\overline{\Omega})$. Then E is precompact in $C(\overline{\Omega})$ if and only if the following conditions hold.*

(i) *E is bounded, that is there exists $M > 0$ such that*

$$\|u\|_{C(\overline{\Omega})} \leq M, \quad \forall u \in E. \quad (6.58)$$

(ii) *E is equicontinuous, that is, for every $\varepsilon > 0$ there exists $\delta = \delta(\varepsilon) > 0$ such that if $\mathbf{x}, \mathbf{x} + \mathbf{h} \in \overline{\Omega}$ and $|\mathbf{h}| < \delta$ then*

$$|u(\mathbf{x} + \mathbf{h}) - u(\mathbf{x})| < \varepsilon, \quad \forall u \in E. \quad (6.59)$$

Observe that (ii) means

$$\lim_{\mathbf{h} \rightarrow \mathbf{0}} \|u(\cdot + \mathbf{h}) - u(\cdot)\|_{C(\overline{\Omega})} = 0, \quad \text{uniformly with respect to } u \in E.$$

¹³ Recall that the norm in $C(\overline{\Omega})$ is $\|f\|_{C(\overline{\Omega})} = \max_{\overline{\Omega}} |f|$.

In general, the condition (i) is rather easy to prove, while the equicontinuity condition (ii) could be difficult to check. A particularly significant case occurs when E is bounded in $C^{0,\alpha}(\overline{\Omega})$, $0 < \alpha \leq 1$ or in $C^1(\overline{\Omega})$. Then:

Corollary 6.51. *Let $\Omega \subset \mathbb{R}^n$ be a bounded domain and $E \subset C^{0,\alpha}(\overline{\Omega})$, $0 < \alpha \leq 1$. If there exists $M > 0$ such that*

$$\|u\|_{C^{0,\alpha}(\overline{\Omega})} \leq M, \quad \forall u \in E \quad (6.60)$$

then E è precompact in $C(\overline{\Omega})$.

Proof. Condition (6.60) means that

$$\|u\|_{C(\overline{\Omega})} + \sup_{\substack{\mathbf{x}, \mathbf{y} \in \overline{\Omega} \\ \mathbf{x} \neq \mathbf{y}}} \frac{|u(\mathbf{x}) - u(\mathbf{y})|}{|\mathbf{x} - \mathbf{y}|^\alpha} \leq M, \quad \forall u \in E.$$

Thus, in particular, $\|u\|_{C(\overline{\Omega})} \leq M$, which is (6.58). Moreover,

$$|u(\mathbf{x} + \mathbf{h}) - u(\mathbf{x})| \leq M |\mathbf{h}|^\alpha, \quad \forall \mathbf{x}, \mathbf{x} + \mathbf{h} \in \overline{\Omega}, \quad \forall u \in E \quad (6.61)$$

which gives the equicontinuity of E . \square

In $L^p(\Omega)$, $1 \leq p < \infty$, there exists an analogue of Corollary 6.51, due to M. Riesz, Fréchet and Kolmogoroff.

Theorem 6.52. *Let $\Omega \subset \mathbb{R}^n$ be a bounded domain and $S \subset L^p(\Omega)$, $1 \leq p < \infty$. If:*

i) *S is bounded: i.e. there exists K such that*

$$\|u\|_{L^p(\Omega)} \leq K, \quad \forall u \in S.$$

ii) *There exist α and L , positive, such that, if u is extended by zero outside Ω ,*

$$\|u(\cdot + \mathbf{h}) - u(\cdot)\|_{L^p(\Omega)} \leq L |\mathbf{h}|^\alpha, \quad \text{for every } \mathbf{h} \in \mathbb{R}^n \text{ and } u \in S,$$

then S is pre-compact.

The second condition expresses an *equicontinuity in norm* of all the elements in S . We shall meet this condition in Subsect. 7.10.1.

6.7.3 Weak convergence and compactness

We have seen that compactness in a normed space is equivalent to sequential compactness. In the applications, this translates into a very strong requirement for approximating sequences.

Fortunately, in normed spaces, and in particular in Hilbert spaces, there is another notion of convergence, much more flexible, which turns out to be perfectly adapted to the variational formulation of boundary value problems.

Let H be a Hilbert space with inner product (\cdot, \cdot) and norm $\|\cdot\|$. If $F \in H^*$ and $\|x_k - x\| \rightarrow 0$, then $Fx_k \rightarrow Fx$. In fact

$$|Fx_k - Fx| = |F(x_k - x)| \leq \|F\|_{H^*} \|x_k - x\|.$$

However, it could be that

$$Fx_k \rightarrow Fx$$

for every $F \in H^*$, even if $\|x_k - x\| \not\rightarrow 0$. Then, we say that x_k converges *weakly* to x . Precisely:

Definition 6.53. A sequence $\{x_k\} \subset H$ converges weakly to $x \in H$, and we write

$$x_k \rightharpoonup x$$

(with an “half arrow”), if

$$Fx_k \rightarrow Fx, \quad \forall F \in H^*.$$

To emphasize the difference, the convergence in norm is then called *strong convergence*. From Riesz’s Representation Theorem, it follows that $\{x_k\} \subset H$ converges weakly to $x \in H$ if and only if

$$(x_k, y) \rightarrow (x, y), \quad \forall y \in H.$$

The weak limit is unique, since $x_k \rightharpoonup x$ and $x_k \rightharpoonup z$ imply

$$(x - z, y) = 0 \quad \forall y \in H,$$

whence $x = z$. Moreover, Schwarz’s inequality gives

$$|(x_k - x, y)| \leq \|x_k - x\| \|y\|$$

so that *strong convergence* implies *weak convergence*, which should not be surprising. The two notions of convergence are equivalent in finite-dimensional spaces. It is not so in infinite dimensions, as the following example shows.

Example 6.54. Let $H = L^2(0, 2\pi)$. The sequence $v_k(x) = \cos kx$, $k \geq 1$, is weakly convergent to zero. In fact, for every $f \in L^2(0, 2\pi)$, Corollary A.2, p. 661, on the Fourier coefficients of f , implies that

$$(f, v_k)_{L^2(0, 2\pi)} = \int_0^{2\pi} f(x) \cos kx \, dx \rightarrow 0$$

as $k \rightarrow \infty$. However

$$\|v_k\|_{L^2(0,2\pi)} = \sqrt{\pi}$$

and therefore $\{v_k\}_{k \geq 1}$ does not converge strongly to zero.

Remark 6.55. If $L \in \mathcal{L}(H_1, H_2)$ and $x_k \rightharpoonup x$ in H_1 we cannot say that $Lx_k \rightarrow Lx$ in H_2 . However, if $F \in H_2^*$, then $F \circ L \in H_1^*$ and therefore

$$F(Lx_k) = (F \circ L)x_k \rightarrow (F \circ L)x = F(Lx).$$

Thus, if L is (strongly) continuous then it is *weakly sequentially continuous* as well.

Warning: Not always *strong implies weak!* Take a strongly closed set $E \subset H$. Thus E contains all the limits of strongly convergent sequences $\{x_k\} \subset E$. Can we deduce that E is *weakly sequentially closed* as well? The answer is *no*: if $x_k \rightharpoonup x$ (only weakly), since the convergence is not strong, we can not affirm that $x \in E$. Indeed, E is *not weakly sequentially closed*, in general¹⁴.

Take for instance

$$E = \{v \in L^2(0, 2\pi) : \|v\|_{L^2(0,2\pi)} = \sqrt{\pi}\}.$$

Then E is a strongly closed (and bounded) subset of $L^2(0, 2\pi)$. However, E contains the sequence $\{v_k\}$, where $v_k(x) = \cos kx$, and we have seen in Example 6.54 that $v_k \rightharpoonup 0 \notin E$. Hence E is *not weakly sequentially closed*.

We have observed that the norm in a Hilbert space is strongly continuous. With respect to weak convergence, the norm is only (sequentially) lower semicontinuous, as property 2 in the following theorem shows.

Theorem 6.56. Let $\{x_k\} \subset H$ such that $x_k \rightharpoonup x$. Then

- 1) $\{x_k\}$ is bounded.
- 2) $\|x\| \leq \liminf_{k \rightarrow \infty} \|x_k\|$.

We omit the proof¹⁵ of 1. For the second point, it is enough to observe that

$$\|x\|^2 = \lim_{k \rightarrow \infty} (x_k, x) \leq \|x\| \liminf_{k \rightarrow \infty} \|x_k\|$$

and simplify by $\|x\|$.

The usefulness of weak convergence is revealed by the following compactness result. Basically, it says that if we substitute *strong* with *weak* convergence, any bounded sequence in a Hilbert space is weakly pre-compact. Precisely:

¹⁴ See Problem 6.16.

¹⁵ It is a consequence of the Banach-Steinhaus Theorem. See e.g. [39], Yoshida, 1971.

Theorem 6.57. *Every bounded sequence in a Hilbert space H contains a subsequence which is weakly convergent to an element $x \in H$.*

Proof. We give it under the additional hypothesis that H is separable¹⁶. Thus, there exists a sequence $\{z_k\}$ dense in H . Let now $\{x_j\} \subset H$ be a bounded sequence: $\|x_j\| \leq M$, $\forall j \geq 1$. We split the proof into three steps.

1. We use a “diagonal” process, to construct a subsequence $\{x_s^{(s)}\}$, such that the real sequence $(x_s^{(s)}, z_k)$ is convergent for every fixed z_k . To do this, observe that the sequence $\{(x_j, z_1)\}$ is bounded in \mathbb{R} and therefore there exists $\{x_j^{(1)}\} \subset \{x_j\}$ such that $\{(x_j^{(1)}, z_1)\}$ is convergent. For the same reason, from $\{x_j^{(1)}\}$ we may extract a subsequence $\{x_j^{(2)}\}$ such that $\{(x_j^{(2)}, z_2)\}$ is convergent. By induction, we construct $\{x_j^{(k)}\}$ such that

$$\left\{ (x_j^{(k)}, z_k) \right\}$$

converges. Consider the diagonal sequence $\{x_s^{(s)}\}$, obtained by selecting $x_1^{(1)}$ from $\{x_j^{(1)}\}$, $x_2^{(2)}$ from $\{x_j^{(2)}\}$ and so on. Then,

$$\left\{ (x_s^{(s)}, z_k) \right\}$$

is convergent for every fixed $k \geq 1$.

2. We now use the density of $\{z_k\}$ in H , to show that $(x_s^{(s)}, z)$ converges for every $z \in H$. In fact, for fixed $\varepsilon > 0$ and $z \in H$, we may find z_k such that $\|z - z_k\| < \varepsilon$. Write

$$(x_s^{(s)} - x_m^{(m)}, z) = (x_s^{(s)} - x_m^{(m)}, z - z_k) + (x_s^{(s)} - x_m^{(m)}, z_k).$$

If j and m are large enough, we have

$$\left| (x_s^{(s)} - x_m^{(m)}, z_k) \right| < \varepsilon$$

since $(x_s^{(s)}, z_k)$ is convergent. Moreover, from Schwarz's inequality,

$$\left| (x_s^{(s)} - x_m^{(m)}, z - z_k) \right| \leq \|x_s^{(s)} - x_m^{(m)}\| \|z - z_k\| \leq 2M\varepsilon.$$

Thus, if j and m are large enough, we have

$$\left| (x_s^{(s)} - x_m^{(m)}, z) \right| \leq (2M + 1)\varepsilon,$$

hence the sequence $(x_s^{(s)} - x_m^{(m)}, z)$ is a Cauchy sequence in \mathbb{R} and therefore convergent.

3. From **2**, we may define a linear functional T in H by setting

$$Tz = \lim_{s \rightarrow \infty} (x_s^{(s)}, z).$$

¹⁶ There is actually no loss of generality since we can always replace H by the closure H_0 of the subspace generated by the finite linear combinations of elements of the sequence. Clearly H_0 is separable.

Since $\|x_s^{(s)}\| \leq M$, we have $|Tz| \leq M \|z\|$, whence $T \in H^*$. From the Riesz Representation theorem, there exists a unique $x_\infty \in H$ such that

$$Tz = (x_\infty, z), \quad \forall z \in H.$$

Thus

$$(x_s^{(s)}, z) \rightarrow (x_\infty, z), \quad \forall z \in H,$$

which means $x_s^{(s)} \rightharpoonup x_\infty$. \square

Example 6.58. Let $H = L^2(\Omega)$, $\Omega \subseteq \mathbb{R}^n$ and consider a sequence $\{u_k\}_{k \geq 1} \subset L^2(\Omega)$. To say that $\{u_k\}$ is bounded means that

$$\|u_k\|_{L^2(\Omega)} \leq M, \quad \text{for every } k \geq 1.$$

Theorem 6.57 implies the existence of a subsequence $\{u_{k_m}\}_{m \geq 1}$ and of $u \in L^2(\Omega)$ such that, as $m \rightarrow +\infty$,

$$\int_{\Omega} u_{k_m} v \rightarrow \int_{\Omega} uv, \quad \text{for every } v \in L^2(\Omega).$$

6.7.4 Compact operators

By definition, every operator in $\mathcal{L}(H_1, H_2)$ transforms bounded sets in H_1 into bounded sets in H_2 . The subclass of operators that transform *bounded sets* into *precompact sets* is particularly important.

Definition 6.59. Let H_1 and H_2 be Hilbert spaces and $L \in \mathcal{L}(H_1, H_2)$. We say that L is **compact** if, for every bounded $E \subset H_1$, the image $L(E)$ is **precompact** in H_2 .

An equivalent characterization of compact operators may be given in terms of weak convergence. Indeed, a linear operator is compact if and only if “it converts weak convergence into strong convergence”. Precisely:

Proposition 6.60. Let $L \in \mathcal{L}(H_1, H_2)$. L is compact if and only if, for every sequence $\{x_k\} \subset H_1$,

$$x_k \rightharpoonup 0 \quad \text{in } H_1 \quad \text{implies} \quad Lx_k \rightarrow 0 \quad \text{in } H_2. \quad (6.62)$$

Proof. Assume that (6.62) holds. Let $E \subset H_1$ be bounded, and $\{z_k\} \subset L(E)$. Then $z_k = Lx_k$ for some $x_k \in E$.

From Theorem 6.57, there exists a subsequence $\{x_{k_s}\}$ weakly convergent to $x \in H_1$. Then

$$y_s = x_{k_s} - x \rightharpoonup 0$$

in H_1 and, from (6.62), $Ly_s \rightarrow 0$ in H_2 , that is

$$z_{k_s} = Lx_{k_s} \rightarrow Lx \equiv z$$

in H_2 . Thus, $L(E)$ is sequentially precompact, and therefore precompact in H_2 .

Viceversa, let L be compact and $x_k \rightharpoonup 0$ in H_1 . Suppose $Lx_k \not\rightarrow 0$. Then, for some $\bar{\varepsilon} > 0$ and infinitely many indexes k_j , we have $\|Lx_{k_j}\| > \bar{\varepsilon}$. Since

$$x_{k_j} \rightharpoonup 0,$$

by Theorem 6.56, $\{x_{k_j}\}$ is bounded in H_1 , so that $\{Lx_{k_j}\}$ contains a subsequence (that we still call) $\{Lx_{k_j}\}$ strongly (and therefore weakly) convergent to some $y \in H_2$. On the other hand, we have $Lx_{k_j} \rightarrow 0$ as well, which entails $y = 0$. Thus

$$\|Lx_{k_j}\| \rightarrow 0.$$

Contradiction. □

Example 6.61. Let $H_{per}^1(0, 2\pi)$ be the Hilbert space introduced in Example 6.11, p. 362. The embedding of $H_{per}^1(0, 2\pi)$ into $L^2(0, 2\pi)$ is compact (see Problem 6.18).

Example 6.62. From Theorem 6.49, the identity operator $I : H \rightarrow H$ is compact if and only if $\dim H < \infty$. Also, any bounded operator with finite dimensional range is compact.

Example 6.63. Let $Q = (0, 1) \times (0, 1)$ and $g \in C(\overline{Q})$. Consider the integral operator

$$Tv(x) = \int_0^1 g(x, y) v(y) dy. \quad (6.63)$$

We want to show that T is compact from $L^2(0, 1)$ into $L^2(0, 1)$. In fact, for every $x \in (0, 1)$, Schwarz's inequality gives

$$|Tv(x)| \leq \int_0^1 |g(x, y) v(y)| dy \leq \|g(x, \cdot)\|_{L^2(0, 1)} \|v\|_{L^2(0, 1)}, \quad (6.64)$$

whence

$$\int_0^1 |Tv(x)|^2 dx \leq \|g\|_{L^2(Q)}^2 \|v\|_{L^2(0, 1)}^2$$

which implies that $Tv \in L^2(0, 1)$ and that T is bounded.

To check compactness, we use Proposition 6.60. Let $\{v_k\} \subset L^2(0, 1)$ such that $v_k \rightharpoonup 0$, that is

$$\int_0^1 v_k w \rightarrow 0, \quad \text{for every } w \in L^2(0, 1). \quad (6.65)$$

We have to show that $Tv_k \rightarrow 0$ in $L^2(0, 1)$. Being weakly convergent, $\{v_k\}$ is bounded so that

$$\|v_k\|_{L^2(0, 1)} \leq M, \quad (6.66)$$

for some M and every k . From (6.64) we have

$$|Tv_k(x)| \leq M \|g(x, \cdot)\|_{L^2(0, 1)}.$$

Moreover, inserting $w(\cdot) = g(x, \cdot)$ into (6.65), we infer that

$$Tv_k(x) = \int_0^1 g(x, y) v_k(y) dy \rightarrow 0, \quad \text{for every } x \in (0, 1).$$

From the Dominated Convergence Theorem¹⁷ we infer that $Tv_k \rightarrow 0$ in $L^2(0, 1)$. Therefore T is compact.

The following proposition is useful.

Proposition 6.64. *Let $L : H_1 \rightarrow H_2$ be compact. Then:*

- a) $L^* : H_2 \rightarrow H_1$ is compact.
- b) If $G \in \mathcal{L}(H_2, H_3)$ or $G \in \mathcal{L}(H_0, H_1)$, the operator $G \circ L$ or $L \circ G$ is compact.

Proof. a) We use Proposition 6.60. Let $\{x_k\} \subset H_2$ and $x_k \rightharpoonup 0$. Let us show that $\|L^*x_k\|_{H_1} \rightarrow 0$. We have:

$$\|L^*x_k\|_{H_1}^2 = (L^*x_k, L^*x_k)_{H_1} = (x_k, LL^*x_k)_{H_2}.$$

Since $L^* \in \mathcal{L}(H_2, H_1)$, we have

$$L^*x_k \rightharpoonup 0$$

in H_1 and the compactness of L entails

$$LL^*x_k \rightarrow 0$$

in H_2 . Since $\|x_k\| \leq M$, we finally have

$$\|L^*x_k\|_{H_1}^2 = (x_k, LL^*x_k)_{H_2} \leq M \|LL^*x_k\|_{H_2}^2 \rightarrow 0.$$

b) We leave it as an exercise. □

6.8 The Fredholm Alternative

6.8.1 Hilbert triplets

Let us go back to the variational problem

$$a(u, v) = Fv \quad \forall v \in V, \tag{6.67}$$

and suppose that Lax-Milgram Theorem 6.39, p. 383, cannot be applied, since, for instance, a is not V -coercive. In this situation it may happen that the problem does not have a solution, unless certain compatibility conditions on F are satisfied. A typical example is given by the Neumann problem

$$\begin{cases} -\Delta u = f & \text{in } \Omega \\ \partial_\nu u = g & \text{on } \partial\Omega. \end{cases}$$

A necessary and sufficient solvability condition is given by

$$\int_\Omega f + \int_{\partial\Omega} g = 0. \tag{6.68}$$

¹⁷ See Appendix B.

Moreover, if (6.68) holds, there are infinitely many solutions, differing among each other by an additive constant. Condition (6.68) has both a precise physical interpretation in terms of a resultant of forces at equilibrium and a deep mathematical meaning, with roots in Linear Algebra!

Indeed, the results we are going to present are extensions of well known facts concerning the solvability of linear algebraic systems of the form

$$\mathbf{A}\mathbf{x} = \mathbf{b} \quad (6.69)$$

where \mathbf{A} is an $n \times n$ matrix and $\mathbf{b} \in \mathbb{R}^n$. The following dichotomy holds: *either (6.69) has a unique solution for every \mathbf{b} or the homogeneous equation $\mathbf{A}\mathbf{x} = \mathbf{0}$ has nontrivial solutions.*

More precisely, system (6.69) is solvable if and only if \mathbf{b} belongs to the *column space of \mathbf{A}* , which is the orthogonal complement of $\mathcal{N}(\mathbf{A}^\top)$. If $\mathbf{w}_1, \dots, \mathbf{w}_s$ span $\mathcal{N}(\mathbf{A}^\top)$, this is equivalent to the s compatibility conditions, $0 \leq s \leq n$,

$$\mathbf{b} \cdot \mathbf{w}_j = \mathbf{0} \quad j = 1, \dots, s.$$

Finally, the kernels $\mathcal{N}(\mathbf{A})$ and $\mathcal{N}(\mathbf{A}^\top)$ have the same dimension and if $\mathbf{v}_1, \dots, \mathbf{v}_s$ span $\mathcal{N}(\mathbf{A})$, the general solution of (6.69) is given by

$$\mathbf{x} = \bar{\mathbf{x}} + \sum_{j=1}^s c_j \mathbf{v}_j$$

where $\bar{\mathbf{x}}$ is a particular solution of (6.69) and c_1, \dots, c_s are arbitrary constants.

The extension to infinite-dimensional spaces requires some care. In particular, in order to state an analogous dichotomy theorem for the variational problem (6.67), we need to clarify the general setting, to avoid confusion.

The problem involves two Hilbert spaces: V , the space where we seek the solution, and V^* , which the data F belongs to. Let us introduce a third space H , “intermediate” between V and V^* .

For better clarity, it is convenient to use the symbol $\langle \cdot, \cdot \rangle_*$ to denote the duality between V^* and V and, if strictly necessary, the symbol $\langle \cdot, \cdot \rangle_{V^*, V}$. Thus we write $\langle F, v \rangle_*$ for Fv .

In the applications to boundary value problems, usually $H = L^2(\Omega)$, with Ω bounded domain in \mathbb{R}^n , while V is a Sobolev space. In practice, we often meet a pair of Hilbert spaces V, H with the following properties:

1. $V \hookrightarrow H$, i.e. V is *continuously embedded in H* . Recall that this simply means that the embedding operator $I_{V \hookrightarrow H}$, from V into H , introduced in Example 6.25, p. 376, is continuous or, equivalently that there exists C such that

$$\|u\|_H \leq C \|u\|_V, \quad \forall u \in V. \quad (6.70)$$

2. V is dense in H .

Using Riesz's Representation Theorem we may identify H with H^* . Also, we may *continuously embed H into V^** , so that any element in H can be thought as an element of V^* . To see it, observe that, for any fixed $u \in H$, the functional $T_u : V \rightarrow \mathbb{R}$,

defined by

$$\langle T_u, v \rangle_* = (u, v)_H, \quad v \in V, \quad (6.71)$$

is continuous in V . In fact, the Schwarz inequality and (6.70) give

$$|(u, v)_H| \leq \|u\|_H \|v\|_H \leq C \|u\|_H \|v\|_V. \quad (6.72)$$

Thus, the map $u \mapsto T_u$ is continuous from H into V^* , with $\|T_u\|_{V^*} \leq C \|u\|_H$. Moreover, if $T_u = 0$, then

$$0 = \langle T_u, v \rangle_* = (u, v)_H, \quad \forall v \in V$$

which forces $u = 0$, by the density of V in H .

Thus, the map $u \mapsto T_u$ is one to one and defines the continuous embedding $I_{H \hookrightarrow V^*}$ of H into V^* . This allows the *identification* of $u \in H$ with $T_u \in V$. In particular, instead of (6.71), we can write

$$\langle T_u, v \rangle_* = \langle u, v \rangle_* = (u, v)_H, \quad \forall v \in V, \quad (6.73)$$

regarding u on the left in the last equality as an element of V^* and on the right as an element of H .

Finally, V (and therefore also H) is *dense*¹⁸ in V^* . Summarizing, we have

$$V \hookrightarrow H \hookrightarrow V^*$$

with *dense embeddings*. We call $\{V, H, V^*\}$ a **Hilbert triplet**.

Warning: Given the identification of H with H^* , the scalar product in H achieves a privileged role. As a consequence, a further identification of V with V^* through the inverse Riesz map \mathcal{R}_V^{-1} is forbidden, since it would give rise to a nonsense, unless $H = V$. Indeed, in the Hilbert triplet setting, we have the embedding J of V into V^* , defined by

$$I_{H \hookrightarrow V^*} \circ I_{V \hookrightarrow H}. \quad (6.74)$$

Accordingly, for $u \in V$ we have from (6.73),

$$\langle Ju, v \rangle_* = (u, v)_H, \quad \forall v \in V. \quad (6.75)$$

¹⁸ It is enough to show that the only element of V^* , orthogonal (in V^*) to T_v for every $v \in V$, is $F = 0$. Thus, let $F \in V^*$ and $(T_v, F)_{V^*} = 0$ for every v . Let $\mathcal{R}_V : V^* \rightarrow V$ be the Riesz's operator. Then we can write

$$0 = (T_v, F)_{V^*} = (\mathcal{R}_V T_v, \mathcal{R}_V F)_V = \langle T_v, \mathcal{R}_V F \rangle_* = (v, \mathcal{R}_V F)_H, \quad \forall v \in V$$

where the second equality comes from the definition (6.38) of the scalar product in V^* , the third one from the definition of \mathcal{R}_V and the fourth one from (6.73). Choosing in particular $v = \mathcal{R}_V F$, we conclude that $\|\mathcal{R}_V F\|_H = 0$ and therefore that $F = 0$, since \mathcal{R}_V is an isometry.

If now we identify $u \in V$ with an element of V^* through \mathcal{R}_V^{-1} , we would have

$$\langle \mathcal{R}_V^{-1}u, v \rangle_* = (u, v)_V, \quad \forall v \in V, \quad (6.76)$$

which is compatible with (6.75) only if $(u, v)_H = (u, v)_V$, for every $v \in V$. Since V is complete and dense in H , this forces $H = V$.

6.8.2 Solvability for abstract variational problems

To state the main result of this section we need to introduce weakly coercive forms and their adjoints.

Definition 6.65. We say that the bilinear form $a = a(u, v)$ is weakly coercive with respect to the pair (V, H) if there exist $\lambda_0 \in \mathbb{R}$ and $\alpha > 0$ such that

$$a(v, v) + \lambda_0 \|v\|_H^2 \geq \alpha \|v\|_V^2 \quad \forall v \in V.$$

We say that the number α is the weak coercivity constant of a .

The adjoint form a^* of a is given by $a^*(u, v) = a(v, u)$, obtained by interchanging the arguments in the analytical expression of a . In the applications to boundary value problems, a^* is associated with the so called *formal adjoint* of a differential operator (see Sect. 8.5.1).

We shall denote by $\mathcal{N}(a)$ and $\mathcal{N}(a^*)$, the set of solutions u and w , respectively, of the variational problems

$$a(u, v) = 0, \quad \forall v \in V \quad \text{and} \quad a^*(w, v) = 0, \quad \forall v \in V.$$

Observe that $\mathcal{N}(a)$ and $\mathcal{N}(a^*)$ are both subspaces of V , playing the role of *kernels* for a and a^* .

Theorem 6.66. Let $\{V, H, V^*\}$ be a Hilbert triplet, with V compactly embedded in H . Let a be a bilinear form in V , continuous and weakly coercive with respect to (V, H) . Then:

a) Either equation

$$a(u, v) = \langle F, v \rangle_*, \quad \forall v \in V, \quad (6.77)$$

has a unique solution u_F for every $F \in V^*$ and there exist a constant C , independent of u_F and F , such that

$$\|u_F\|_V \leq C \|F\|_{V^*}, \quad (6.78)$$

b) or

$$0 < \dim \mathcal{N}(a) = \dim \mathcal{N}(a_*) = d < \infty$$

and (6.77) is solvable if and only if $\langle F, w \rangle_* = 0$ for every $w \in \mathcal{N}(a^*)$.

The proof of Theorem 6.66 relies on a more general result, known as Fredholm's Alternative, stated in the next section.

Some comments are in order. The following dichotomy holds: either (6.77) has a unique solution for every $F \in V^*$ or the homogeneous equation $a(u, v) = 0$ has nontrivial solutions. In other words, existence of a solution to equation (6.77) for every $F \in V^*$ implies uniqueness; viceversa, uniqueness for equation (6.77) implies existence for every $F \in V^*$.

The same conclusions hold for the adjoint equation

$$a^*(u, v) = \langle F, v \rangle_*, \quad \forall v \in V.$$

If w_1, w_2, \dots, w_d span $\mathcal{N}(a^*)$, (6.77) is solvable if and only if the d compatibility conditions

$$\langle F, w_j \rangle_* = 0, \quad j = 1, \dots, d,$$

hold. In this case, equation (6.77) has infinitely many solutions given by

$$u = u_F + \sum_{j=1}^d c_j z_j,$$

where u_F is a particular solution of (6.77), $\{z_1, \dots, z_d\}$ span $\mathcal{N}(a)$ and c_1, \dots, c_d are arbitrary constants.

We shall apply Theorem 6.66 to boundary value problems in Chap. 8. Here is however a preliminary example.

Example 6.67. Let $V = H_{per}^1(0, 2\pi)$, $H = L^2(0, 2\pi)$ and assume that $w = w(t)$ is a positive, continuous function in $[0, 2\pi]$. We know (Example 6.61, p. 398) that V is compactly embedded in H . Moreover, V is dense in H . In fact, as we shall see in Chap. 7, $C_0(0, 2\pi)$, the set of continuous function, compactly supported in $(0, 2\pi)$, which is clearly contained in V , is dense in H . Thus $\{V, H, V^*\}$ is a Hilbert triplet.

Given $f \in H$, consider the variational problem

$$\int_0^{2\pi} u'v'w dt = \int_0^{2\pi} fv dt, \quad \forall v \in V. \quad (6.79)$$

The bilinear form $a(u, v) = \int_0^{2\pi} u'v'w dt$ is continuous in V but it is not V -coercive. In fact

$$|a(u, v)| \leq w_{\max} \|u'\|_{L^2(0, 2\pi)} \|v'\|_{L^2(0, 2\pi)} \leq w_{\max} \|u\|_{H^1(0, 2\pi)} \|v\|_{H^1(0, 2\pi)},$$

but $a(u, u) = 0$ if u is constant. However it is weakly coercive with respect to (V, H) , since

$$a(u, u) + \|u\|_{L^2(0, 2\pi)}^2 = \int_0^{2\pi} (u')^2 w dt + \int_0^{2\pi} u^2 dt \geq \min \{w_{\min}, 1\} \|u\|_{H^1(0, 2\pi)}^2.$$

Moreover,

$$\left| \int_0^{2\pi} fv dt \right| \leq \|f\|_{L^2(0, 2\pi)} \|v\|_{L^2(0, 2\pi)} \leq \|f\|_{L^2(0, 2\pi)} \|v\|_{H^1(0, 2\pi)}$$

hence the functional $F : v \mapsto \int_0^{2\pi} fv dt$ defines an element of V^* .

We are under the hypotheses of Theorem 6.66. The bilinear form is symmetric, so that $\mathcal{N}(a) = \mathcal{N}(a_*)$. The solutions of the homogeneous equation

$$a(u, v) = \int_0^{2\pi} u'v' w dt = 0, \quad \forall v \in V, \quad (6.80)$$

are the constant functions. In fact, letting $v = u$ in (6.80) we obtain

$$\int_0^{2\pi} (u')^2 w dt = 0$$

which forces $u(t) \equiv c$, constant, since $w > 0$. Then, $\dim \mathcal{N}(a) = 1$. Thus, from Theorem 6.66 we can draw the following conclusions: equation (6.79) is solvable if and only if

$$\langle F, 1 \rangle_* = \int_0^{2\pi} f dt = 0.$$

Moreover, in this case, (6.79) has infinitely many solutions of the form $u = \bar{u} + c$.

The variational problem has a simple interpretation as a boundary value problem. By an integration by parts, recalling that $v(0) = v(2\pi)$, we may rewrite (6.79) as

$$\int_0^{2\pi} [(-wu')' - f] v dt + v(0) [w(2\pi)u'(2\pi) - w(0)u'(0)] = 0, \quad \forall v \in V.$$

Choosing v vanishing at 0, we are left with

$$\int_0^{2\pi} [-(wu')' - f] v dt = 0, \quad \forall v \in V, v(0) = v(2\pi) = 0,$$

which forces

$$(u'w)' = -f.$$

Then

$$v(0) [w(2\pi)u'(2\pi) - w(0)u'(0)] = 0, \quad \forall v \in V,$$

which, in turn, forces

$$w(2\pi)u'(2\pi) = w(0)u'(0).$$

Thus, problem (6.79) constitutes the variational formulation of the following boundary value problem:

$$\begin{cases} (wu')' = -f & \text{in } (0, 2\pi) \\ u(0) = u(2\pi) \\ w(2\pi)u'(2\pi) = w(0)u'(0). \end{cases}$$

It is important to point out that the periodicity condition $u(0) = u(2\pi)$ is forced by the choice of the space V while the Neuman type periodicity condition is encoded in the variational equation (6.79).

6.8.3 Fredholm's alternative

We introduce some terminology. Let V_1, V_2 Hilbert spaces and $\Phi : V_1 \rightarrow V_2$. We say that Φ is a *Fredholm operator* if $\mathcal{N}(\Phi)$ and $\mathcal{R}(\Phi)^\perp$ have finite dimension. The *index of Φ* is the integer

$$\text{ind}(\Phi) = \dim \mathcal{N}(\Phi) - \dim \mathcal{R}(\Phi)^\perp = \dim \mathcal{N}(\Phi) - \dim \mathcal{N}(\Phi^*).$$

We have the following result, known as *Fredholm's Alternative Theorem*¹⁹:

Theorem 6.68. *Let V be a Hilbert space and $K \in \mathcal{L}(V)$ be a compact operator. Let I be the identity operator in V . Then*

$$\Phi = I - K$$

is a Fredholm operator with zero index. Moreover, $\Phi^ = I - K^*$,*

$$\mathcal{R}(\Phi) = \mathcal{N}(\Phi^*)^\perp \tag{6.81}$$

and

$$\mathcal{N}(\Phi) = \{0\} \iff \mathcal{R}(\Phi) = V. \tag{6.82}$$

Formula (6.81) implies that $\mathcal{R}(\Phi)$ is closed. Formula (6.82) shows that Φ is one-to-one if and only if it is onto. In this case, Theorem 6.26, p. 377, implies that both Φ^{-1} and $(\Phi^*)^{-1}$ belong to $\mathcal{L}(V)$. In other words, uniqueness for the equation

$$u - Ku = f \tag{6.83}$$

with $f = 0$, is equivalent to existence for every $f \in V$ and viceversa. Moreover, the solution of (6.83) satisfies the estimate

$$\|u\|_V \leq \|\Phi^{-1}\|_{\mathcal{L}(V)} \|f\|_V.$$

The same considerations hold for the adjoint $\Phi^* = I - K^*$ and the associated equation

$$v - K^*v = g.$$

If Φ is not one-to-one, let $d = \dim \mathcal{R}(\Phi)^\perp = \dim \mathcal{N}(\Phi^*) > 0$. Then, (6.81) says that equation (6.83) is solvable if and only if $f \perp \mathcal{N}(\Phi^*)$, that is, if and only if $(f, w)_V = 0$ for every solution w of

$$w - K^*w = 0. \tag{6.84}$$

If $\{w_1, w_2, \dots, w_d\}$ span $\mathcal{N}(\Phi^*)$, this is equivalent to saying that the d compatibility relations

$$(f, w_j)_V = 0, \quad j = 1, \dots, d,$$

are the necessary and sufficient conditions for the solvability of (6.83).

¹⁹ For the proof, see e.g. [32], Brezis, 2010.

Clearly, Theorem 6.68 holds for operators $K - \lambda I$ with $\lambda \neq 0$. The case $\lambda = 0$ cannot be included. Trivially, for the operator $K = 0$ (which is compact), we have $\mathcal{N}(K) = V$, hence, if $\dim V = \infty$, Theorem 6.68 does not hold. A more significant example is the one-dimensional range operator

$$Kx = L(x)x_0,$$

where $L \in \mathcal{L}(V)$ and x_0 is fixed in V . Assume $\dim V = \infty$. From Riesz's Representation Theorem, there exists $z \in V$ such that $Lx = (z, x)_V$, for every $x \in V$. Thus, $\mathcal{N}(K)$ is given by the subspace of the elements in V orthogonal to z , which has infinitely many dimensions.

- *Proof of Theorem 6.66 (sketch).* The strategy is to write equation

$$a(u, v) = \langle F, v \rangle_* \quad (6.85)$$

in the form

$$(I_V - K)u = g$$

where I_V is the identity operator in V and $K : V \rightarrow V$ is compact.

Let $J : V \rightarrow V^*$ be the embedding of V into V^* defined by (6.74), p. 401. Recall that J is the composition of the embeddings $I_{V \hookrightarrow H}$ and $I_{H \hookrightarrow V^*}$. Since $I_{V \hookrightarrow H}$ is compact and $I_{H \hookrightarrow V^*}$ is continuous, by Proposition 6.64, p. 399, we infer that J is **compact**. We write (6.85) in the form²⁰

$$a_{\lambda_0}(u, v) \equiv a(u, v) + \lambda_0(u, v)_H = \langle \lambda_0 Ju + F, v \rangle_*$$

where $\lambda_0 > 0$ is such that $a_{\lambda_0}(u, v)$ is coercive.

Since, for fixed $u \in V$, the linear map $v \mapsto a_{\lambda_0}(u, v)$ is continuous in V , there exists $L \in \mathcal{L}(V, V^*)$ such that

$$\langle Lu, v \rangle_* = a_{\lambda_0}(u, v), \quad \forall u, v \in V.$$

Thus, equation $a(u, v) = \langle F, v \rangle_*$ is equivalent to $\langle Lu, v \rangle_* = \langle \lambda_0 Ju + F, v \rangle_*$, $\forall v \in V$ and therefore to

$$Lu = \lambda_0 Ju + F. \quad (6.86)$$

Since a_{λ_0} is V -coercive, by the Lax-Milgram Theorem, the operator L is a continuous isomorphism between V and V^* and (6.86) can be written in the form

$$u - \lambda_0 L^{-1} Ju = L^{-1} F.$$

Letting $g = L^{-1} F \in V$ and $K = \lambda_0 L^{-1} J$, (6.86) becomes

$$(I_V - K)u = g$$

where $K : V \rightarrow V$.

Since J is compact and L^{-1} is continuous, K is compact. Applying the Fredholm Alternative Theorem and rephrasing the conclusions in terms of bilinear forms, we conclude the proof²¹. \square

²⁰ Recall that $\langle Ju, v \rangle_* = (u, v)_H$, from (6.75).

²¹ We omit the rather long and technical details.

6.9 Spectral Theory for Symmetric Bilinear Forms

6.9.1 Spectrum of a matrix

Let \mathbf{A} be an $n \times n$ matrix and $\lambda \in \mathbb{C}$. Then, either the equation

$$\mathbf{Ax} - \lambda \mathbf{x} = \mathbf{b}$$

has a unique solution for every \mathbf{b} or there exists $\mathbf{u} \neq \mathbf{0}$ such that

$$\mathbf{Au} = \lambda \mathbf{u}.$$

In the last case we say that λ, \mathbf{u} constitutes an *eigenvalue-eigenvector pair*. The set of eigenvalues of \mathbf{A} is called the *spectrum* of \mathbf{A} , denoted by $\sigma_P(\mathbf{A})$. If $\lambda \notin \sigma_P(\mathbf{A})$ the *resolvent matrix* $(\mathbf{A} - \lambda \mathbf{I})^{-1}$ is well defined. The set

$$\rho(\mathbf{A}) = \mathbb{C} \setminus \sigma_P(\mathbf{A})$$

is called the *resolvent set* of \mathbf{A} . If $\lambda \in \sigma_P(\mathbf{A})$, the kernel $\mathcal{N}(\mathbf{A} - \lambda \mathbf{I})$ is the subspace spanned by the eigenvectors corresponding to λ and it is called the *eigenspace* of λ . Note that $\sigma_P(\mathbf{A}) = \sigma_P(\mathbf{A}^\top)$.

The symmetric matrices are particularly important: all the eigenvalues $\lambda_1, \dots, \lambda_n$ are real (possibly of multiplicity greater than 1) and there exists in \mathbb{R}^n an orthonormal basis of eigenvectors $\mathbf{u}_1, \dots, \mathbf{u}_n$. The action of \mathbf{A} can be written as the sum of the projections into its eigenspaces, according to the following formula²², called *spectral decomposition* of \mathbf{A} :

$$\mathbf{A} = \lambda_1 \mathbf{u}_1 \mathbf{u}_1^\top + \lambda_2 \mathbf{u}_2 \mathbf{u}_2^\top + \dots + \lambda_n \mathbf{u}_n \mathbf{u}_n^\top.$$

We are going to extend these concepts in the Hilbert space setting. A motivation is the method of separation of variables.

6.9.2 Separation of variables revisited

Using the method of separation of variables, in the first chapters we have constructed solutions of boundary value problems by superposition of special solutions. However, explicit computations can be performed only when the geometry of the relevant domain is quite particular. What may we say in general? Let us consider an example from diffusion.

Suppose we have to solve the problem

$$\begin{cases} u_t = \Delta u & (x, y) \in \Omega, t > 0 \\ u(x, y, 0) = g(x, y) & (x, y) \in \Omega \\ u(x, y, t) = 0 & (x, y) \in \partial\Omega, t > 0 \end{cases}$$

²² Here \mathbf{u}_j are column vectors; $\mathbf{u}_j \mathbf{u}_j^\top$ is a matrix $n \times n$, sometime denoted by $\mathbf{u}_j \otimes \mathbf{u}_j$ (\mathbf{u}_j tensor \mathbf{u}_j).

where Ω is a bounded bi-dimensional domain. Let us look for solutions of the form

$$u(x, y, t) = v(x, y) w(t).$$

Substituting into the differential equation, with some elementary manipulations, we obtain

$$\frac{w'(t)}{w(t)} = \frac{\Delta v(x, y)}{v(x, y)} = -\lambda,$$

where λ is a constant, which leads to the two problems

$$w' + \lambda w = 0 \quad t > 0 \tag{6.87}$$

and

$$\begin{cases} -\Delta v = \lambda v & \text{in } \Omega \\ v = 0 & \text{on } \partial\Omega. \end{cases} \tag{6.88}$$

A number λ such that there exists a nontrivial solution v of (6.88) is called a *Dirichlet eigenvalue of the operator $-\Delta$ in Ω* and v is a corresponding *eigenfunction*. Now, the original problem can be solved if the following two properties hold:

a) There exists a sequence of (real) eigenvalues λ_k with corresponding eigenvectors u_k . Solving (6.87) for $\lambda = \lambda_k$ yields

$$w_k(t) = ce^{-\lambda_k t} \quad c \in \mathbb{R}.$$

b) The initial data g can be expanded in series of eigenfunctions:

$$u(x, y) = \sum g_k u_k(x, y).$$

Then, the solution is given by

$$u(x, y, t) = \sum g_k u_k(x, y) e^{-\lambda_k t}$$

where the series converges in some suitable sense.

Condition **b)** requires that the set of Dirichlet eigenfunctions of $-\Delta$ constitutes a basis in the space of initial data. This leads to the problem of determining the *spectrum* of a linear operator in a Hilbert space and, in particular, of self-adjoint compact operators. Indeed, it turns out that the solution map of a symmetric variational boundary value problem is often a self-adjoint compact operator. We will go back to the above problem in Sect. 8.3.4.

6.9.3 Spectrum of a compact self-adjoint operator

We define *resolvent* and *spectrum* for a bounded linear operator. Although the natural setting is the complex field \mathbb{C} , we limit ourselves to \mathbb{R} , mainly for simplicity but also because this is the interesting case for us.

Definition 6.69. Let H be a Hilbert space, $L \in \mathcal{L}(H)$, and I be the identity in H .

a) The resolvent set $\rho(L)$ of L is the set of real numbers λ such that $L - \lambda I$ is one-to-one and onto:

$$\rho(L) = \{\lambda \in \mathbb{R} : L - \lambda I \text{ is one-to-one and onto}\}.$$

b) The (real) spectrum $\sigma(L)$ of L is

$$\sigma(L) = \mathbb{R} \setminus \rho(L).$$

Remark 6.70. If $\lambda \in \rho(L)$, the resolvent $(L - \lambda I)^{-1}$ is bounded (see Theorem 6.26, p. 377).

If H is finite dimensional, any linear operator in $\mathcal{L}(H)$ is represented by a matrix, so that its spectrum is given by the set of its eigenvalues. In infinitely many dimensions the spectrum may be divided in three subsets. In fact, if $\lambda \in \sigma(L)$, different things can go wrong with $(L - \lambda I)^{-1}$. First of all, it may happen that $L - \lambda I$ is not one-to-one so that $(L - \lambda I)^{-1}$ does not even exist.

This means that

$$\mathcal{N}(L - \lambda I) \neq \emptyset$$

or, in other words, that the equation

$$Lu = \lambda u \tag{6.89}$$

has nontrivial solutions. Then, we say that λ is an *eigenvalue* of L and that the nonzero solutions of (6.89) are the *eigenvectors* corresponding to λ . The linear space $\mathcal{N}(L - \lambda I)$ spanned by these eigenvectors is called the *eigenspace* of λ . The dimension of $\mathcal{N}(L - \lambda I)$ is called the *multiplicity* of λ .

Definition 6.71. The set $\sigma_P(L)$ of the eigenvalues of L is called the point spectrum of L .

Other things can occur. $L - \lambda I$ is one-to-one, $\mathcal{R}(L - \lambda I)$ is dense in H , but $(L - \lambda I)^{-1}$ is unbounded. Then, we say that λ belongs to the *continuous spectrum* of L , denoted by $\sigma_C(L)$.

Finally, $L - \lambda I$ is one-to-one but $\mathcal{R}(L - \lambda I)$ is not dense in H . This defines the *residual spectrum* of L .

Example 6.72. Let $H = l^2$ and $L : l^2 \rightarrow l^2$ be the *shift* operator which maps $\mathbf{x} = \{x_1, x_2, \dots\} \in l^2$ into $\mathbf{y} = \{0, x_1, x_2, \dots\}$. We have

$$(L - \lambda I) \mathbf{x} = \{-\lambda x_1, x_1 - \lambda x_2, x_2 - \lambda x_3, \dots\}.$$

If $\lambda \neq 0$, then $\lambda \in \rho(L)$. In fact for every $\mathbf{z} = \{z_1, z_2, \dots\} \in l^2$,

$$(L - \lambda I)^{-1} \mathbf{z} = \left\{ -\frac{z_1}{\lambda}, -\frac{z_2}{\lambda} + \frac{z_1}{\lambda^2}, -\frac{z_3}{\lambda} + \frac{z_2}{\lambda^2}, \dots \right\}.$$

Since $\mathcal{R}(L)$ contains only sequences whose first element is zero, $\mathcal{R}(L)$ is *not dense in* l^2 , therefore $0 \in \sigma_R(L) = \sigma(L)$.

We are mainly interested in the spectrum of a *compact self-adjoint* operator. The following theorem is fundamental²³

Theorem 6.73. *Let K be a compact, selfadjoint operator on a separable, infinite dimensional Hilbert space H . Then:*

- a) $0 \in \sigma(K)$ and $\sigma(K) \setminus \{0\} = \sigma_P(K) \setminus \{0\}$. Moreover, every eigenvalue has finite multiplicity.
- b) $\sigma_P(K) \setminus \{0\}$ is a finite set or it can be ordered in a sequence

$$|\lambda_1| \geq |\lambda_2| \geq \cdots \geq |\lambda_m| \geq \cdots$$

with $\lambda_m \rightarrow 0$ as $m \rightarrow \infty$, and where each eigenvalue appears a number of times according to its multiplicity.

- c) H has an orthonormal basis $\{u_m\}$ consisting of eigenvectors of K .

A few comments are in order.

- If $\dim H = \infty$, the spectrum of a compact selfadjoint operator contains always $\lambda = 0$, which is not necessarily an eigenvalue²⁴. This follows immediately from Theorem 6.49, p. 392. Indeed if $0 \in \rho(K)$, $I_H = K^{-1}K$, would be a compact operator, contradicting Theorem 6.49. The other elements in $\sigma(K)$ are all eigenvalues, arranged in an infinite sequence $\{\lambda_m\}_{m \geq 1}$ converging to zero.
- Let u_j and u_k be eigenvectors corresponding to two *different* eigenvalues λ_j, λ_k , including possibly 0. Then $(u_j, u_k) = 0$. Indeed, from

$$Ku_j = \lambda_j u_j \quad \text{and} \quad Ku_k = \lambda_k u_k$$

we get

$$\lambda_j (u_j, u_k) = (Ku_j, u_k) = (u_j, Ku_k) = \lambda_k (u_j, u_k)$$

and hence $(u_j, u_k) = 0$, since $\lambda_j \neq \lambda_k$.

- If $\lambda \neq 0$ is an eigenvalue, the Fredholm Alternative Theorem 6.68 implies that the eigenspace $\mathcal{N}(K - \lambda I)$ has finite dimension and therefore every nonzero eigenvalue appears in the sequence $\{\lambda_m\}_{m \geq 1}$ only a finite number of times. For instance, if $\dim \mathcal{N}(K - \lambda_1 I) = d_1$, in the sequence $\{\lambda_m\}_{m \geq 1}$ we have $\lambda_1 = \lambda_2 = \cdots = \lambda_{d_1}$; if $\dim \mathcal{N}(K - \lambda_{d_1+1} I) = d_2$, we have

$$\lambda_{d_1+1} = \lambda_{d_1+2} = \cdots = \lambda_{d_1+d_2}$$

and so on.

²³ For the proof, see [32], Brezis, 2010.

²⁴ Actually, properties a) and b) hold for any compact operator.

- The separability of H comes into play only when $0 \in \sigma_P(K)$, to make sure that $\mathcal{N}(K)$ has a countable orthonormal basis. The union of the bases of $\mathcal{N}(K)$ and all $\mathcal{N}(K - \lambda_m I)$, $m \geq 1$, forms a basis in H .
- A consequence of Theorem 6.73 is the *spectral decomposition* formula for K . If $u \in H$ and $\{u_m\}$ is an orthonormal set of eigenvectors corresponding to all non-zero eigenvalues $\{\lambda_m\}_{m \geq 1}$, we can describe the action of K as follows:

$$Ku = \sum_{m \geq 1} (Ku, u_m) u_m = \sum_{m \geq 1} \lambda_m (u, u_m) u_m, \quad \forall u \in H. \quad (6.90)$$

6.9.4 Application to abstract variational problems

We now apply Theorems 6.66, and 6.73 to our abstract variational problems. The setting is the same of Theorem 6.66, given by a Hilbert triplet $\{V, H, V^*\}$, with compact embedding of V into H . We also assume that H is separable. Let a be a bilinear form in V , continuous and (V, H) -weakly coercive; in particular:

$$a_{\lambda_0}(v, v) \equiv a(v, v) + \lambda_0 \|v\|_H^2 \geq \alpha \|v\|_V^2, \quad \forall v \in V \quad (\alpha > 0). \quad (6.91)$$

The notion of *resolvent* and *spectrum* can be easily defined. Consider the problem

$$a(u, v) = \lambda(u, v)_H + \langle F, v \rangle_*, \quad \forall v \in V. \quad (6.92)$$

The *resolvent set* $\rho(a)$ is the set of real numbers λ such that (6.92) has a unique solution $u(F) \in V$, for every $F \in V^*$ and the solution map

$$S(a, \lambda) : F \longmapsto u(F) \quad (6.93)$$

is a continuous isomorphism between V^* and V .

The (real) *spectrum* is $\sigma(a) = \mathbb{R} \setminus \rho(a)$, while the *point spectrum* $\sigma_P(a)$ is the subset of the spectrum given by the *eigenvalues*, that is the numbers λ such that the homogeneous problem

$$a(u, v) = \lambda(u, v)_H, \quad \forall v \in V \quad (6.94)$$

has non-trivial solutions $u \in V$, called *eigenvectors* corresponding to λ . We call *eigenspace* of λ the space spanned by the corresponding eigenfunctions and we denote it by $\mathcal{N}(a, \lambda)$. Note that by the Alternative Theorem 6.66, $\sigma(a) = \sigma_P(a)$.

Theorem 6.74. *Let $\{V, H, V^*\}$ be a Hilbert triplet, with H separable and V compactly embedded in H . Assume that H is infinite dimensional. Let a be a symmetric bilinear form in V , continuous and weakly coercive ((6.91) holds). We have:*

- $\sigma(a) = \sigma_P(a) \subset (-\lambda_0, +\infty)$. The set of the eigenvalues is infinite and can be ordered in a nondecreasing sequence $\{\lambda_m\}_{m \geq 1}$, where each λ_m appears a (finite) number of times according to its multiplicity. Moreover, $\lambda_m \rightarrow +\infty$.

(b) If u_k, u_m are eigenvectors corresponding to two different eigenvalues λ_k and λ_m , respectively, then

$$a(u_k, u_m) = (u_k, u_m)_H = 0.$$

Moreover, H has an orthonormal basis $\{w_m\}$, with w_m corresponding to λ_m , for all $m \geq 1$.

(c) $\{w_m/\sqrt{\lambda_m + \lambda_0}\}$ constitutes an orthonormal basis in V , with respect to the scalar product

$$((u, v)) = a(u, v) + \lambda_0 (u, v)_H. \quad (6.95)$$

Before proving Theorem 6.74 we make some preliminary considerations.

Since $H \hookrightarrow V^*$, we can consider the map $S_{\lambda_0} \in \mathcal{L}(H)$, given by the restriction to H of the solution map $S(a_{\lambda_0}, 0)$. Let us explore the relation between $\sigma_P(S_{\lambda_0})$ and $\sigma_P(a_{\lambda_0})$.

First of all, note that $0 \notin \sigma_P(S_{\lambda_0})$, otherwise we should have $S_{\lambda_0}f = 0$ for some $f \in H$, $f \neq 0$, which leads to the contradiction

$$0 = a_{\lambda_0}(0, v) = (f, v)_H, \quad \forall v \in V.$$

Also, it is clear that $0 \in \rho(a_{\lambda_0})$, whence

$$\sigma_P(a_{\lambda_0}) \subset (0, +\infty). \quad (6.96)$$

Claim. $\lambda \in \sigma_P(a_{\lambda_0})$ if and only if $\mu = 1/\lambda \in \sigma_P(S_{\lambda_0})$. Moreover:

$$\mathcal{N}(a_{\lambda_0}, \lambda) = \mathcal{N}(S_{\lambda_0} - \mu I)$$

and from Theorem 6.66, p. 402, $\mathcal{N}(a_{\lambda_0}, \lambda)$ has dimension $d < \infty$; d is called the multiplicity of λ .

Proof. Let λ be an eigenvalue of a_{λ_0} and f be a corresponding eigenvector, that is,

$$a_{\lambda_0}(f, v) = \lambda (f, v)_H, \quad \forall v \in V. \quad (6.97)$$

Then $f \in V \subset H$ and setting $\mu = 1/\lambda$, equation (6.97) is equivalent to

$$a_{\lambda_0}(\mu f, v) = (f, v)_H$$

that is

$$S_{\lambda_0}f = \mu f. \quad (6.98)$$

Since (6.97) and (6.98) are equivalent, the claim follows. □

We are in position to prove the Theorem 6.74.

Proof. From (6.96) it follows that $\sigma(a) = \sigma_P(a) \subset (-\lambda_0, +\infty)$. Observe now that, since $S_{\lambda_0}(H) \subset V$ and V is compactly embedded into H , S_{λ_0} is compact as an operator from H into H . Also, by the symmetry of a , S_{λ_0} is selfadjoint, that is

$$(S_{\lambda_0}f, g)_H = (f, S_{\lambda_0}g)_H, \quad \text{for all } f, g \in H.$$

In fact, let $u = S_{\lambda_0}f$ and $w = S_{\lambda_0}g$. Then, for every $v \in V$,

$$a_{\lambda_0}(u, v) = (f, v)_H \quad \text{and} \quad a_{\lambda_0}(w, v) = (g, v)_H.$$

In particular,

$$a_{\lambda_0}(u, w) = (f, w)_H \quad \text{and} \quad a_{\lambda_0}(w, u) = (g, u)_H$$

so that, since $a_{\lambda_0}(u, w) = a_{\lambda_0}(w, u)$ and $(g, u)_H = (u, g)_H$, we can write

$$(S_{\lambda_0}f, g)_H = (u, g)_H = (f, w)_H = (f, S_{\lambda_0}g)_H.$$

Since we observed that $0 \notin \sigma_P(S_{\lambda_0})$, from Theorem 6.73 it follows that $\sigma(S_{\lambda_0}) \setminus \{0\} = \sigma_P(S_{\lambda_0})$ and the eigenvalues of S_{λ_0} form an infinite nonincreasing sequence $\{\mu_m\}$, with $\mu_m \rightarrow 0$.

Recalling the claim, $\sigma_P(a_{\lambda_0})$ consists of a sequence $\{\lambda_m^0\}$ that we can order in non-decreasing order, according to the multiplicity of each eigenvalue, with $\lambda_m^0 \rightarrow +\infty$. This proves *a*), with $\lambda_m = \lambda_m^0 - \lambda_0$.

To prove *b*), let u_k and u_m be eigenvectors corresponding to λ_k and λ_m , respectively, $\lambda_k \neq \lambda_m$. Then

$$a(u_k, u_m) = \lambda_k (u_k, u_m)_H \quad \text{and} \quad a(u_k, u_m) = \lambda_m (u_k, u_m)_H.$$

Subtracting the two equations we get

$$(\lambda_k - \lambda_m)(u_k, u_m)_H = 0$$

from which $(u_k, u_m)_H = 0 = a(u_k, u_m)$. Moreover, from *c*) in Theorem 6.73, H has an orthonormal basis $\{w_m\}_{m \geq 1}$ consisting of eigenvectors corresponding to $\{\lambda_m\}_{m \geq 1}$.

Finally, we have²⁵

$$a(w_m, w_k) = \lambda_m (w_m, w_k)_H = \lambda_m \delta_{mk}$$

so that

$$((w_m, w_k)) \equiv a_{\lambda_0}(w_m, w_k) = (\lambda_m + \lambda_0) \delta_{mk}.$$

This shows that $\{w_m / \sqrt{\lambda_m + \lambda_0}\}$ constitutes an orthonormal system in V , endowed with the scalar product (6.95). To check that it is indeed a basis, we show that the only element in V , orthogonal to every w_m , is $v = 0$. Thus, let $v \in V$ and assume that

$$((v, w_m)) \equiv a(v, w_m) + \lambda_0 (v, w_m)_H = 0, \quad \forall m \geq 1.$$

Since $a(v, w_m) = \lambda_m (v, w_m)_H$, we deduce that

$$(\lambda_m + \lambda_0)(v, w_m)_H = 0, \quad \forall m \geq 1.$$

Since $\{w_m\}$ is a basis in H , it follows that $v = 0$. The proof is complete. \square

Under the hypotheses of Theorem 6.74, every element $u \in H$ has the Fourier expansion

$$\sum_{m=1}^{\infty} c_m w_m \tag{6.99}$$

²⁵ δ_{mk} is the Kronecker symbol.

where $c_m = (u, w_m)_H$ are the Fourier coefficients of u and

$$\|u\|_H^2 = \sum_{m=1}^{\infty} c_m^2 < \infty.$$

It turns out that also the elements of V and V^* can be characterized in terms of their Fourier coefficients with respect to the basis $\{w_m\}_{m \geq 1}$. In fact the following theorem holds, where we adopt in V the inner product $((u, v)) = a_{\lambda_0}(u, v)$ and the induced norm $\|v\|_V = \sqrt{((v, v))}$.

Theorem 6.75. *Let $u \in H$. Then u belongs to V if and only if*

$$\|u\|_V^2 = ((u, u)) = \sum_{m=1}^{\infty} (\lambda_m + \lambda_0) c_m^2 < \infty. \quad (6.100)$$

Moreover $F \in V^*$ if and only if F has the expansion $F = \sum_{m=1}^{\infty} d_m w_m$ with

$$\sum_{m=1}^{\infty} \frac{d_m^2}{\lambda_m + \lambda_0} < \infty, \quad (6.101)$$

where $d_m = \langle F, w_m \rangle_*$.

Proof. We only prove (6.100). Let $\tilde{w}_m = w_m / \sqrt{\lambda_0 + \lambda_m}$. Since $\{\tilde{w}_m\}_{m \geq 1}$ is an orthonormal basis in V with respect to the scalar product (6.95), u belongs to V if and only if

$$\sum_{m=1}^{\infty} ((u, \tilde{w}_m))^2 < \infty.$$

Now,

$$((u, \tilde{w}_m)) = \frac{1}{\sqrt{\lambda_m + \lambda_0}} a_{\lambda_0}(u, w_m) = \sqrt{\lambda_m + \lambda_0} c_m$$

and (6.100) follows. \square

Using the above considerations, we can easily prove an important *variational* characterizations of the eigenvalues of a , in terms of the so called *Rayleigh quotient associated to a* , defined by

$$R(u) = \frac{a(u, u)}{\|u\|_H^2}. \quad (6.102)$$

Observe that minimizing R over $V \setminus \{0\}$ is equivalent to minimize $a(u, u)$ over V with the constraint $\|u\|_H^2 = 1$. We start with the variational principle for the first eigenvalue.

Theorem 6.76. *Assume a , V and H are as in Theorem 6.74.*

(1) *Let λ_1 be the first eigenvalue of a and u_1 be a corresponding eigenvector. Then*

$$\lambda_1 = R(u_1) = \min_{v \in V, v \neq 0} R(v) = \min_{v \in V, \|v\|_H=1} a(v, v). \quad (6.103)$$

(2) If $R(w) = \lambda_1 = \min_{v \in V, v \neq 0} R(v)$, then w is an eigenvector of a , corresponding to λ_1 , that is w is a solution of the equation

$$a(w, v) = \lambda_1 (w, v)_H \quad \forall v \in V.$$

Proof. (1) Let $\{\lambda_m\}_{m \geq 1}$ be the nondecreasing sequence of the eigenvalues of a , each one counted according its multiplicity, and $\{w_m\}_{m \geq 1}$ be a corresponding orthonormal basis in H of eigenvectors. For every $v \in V$ we can write

$$v = \sum_{m=1}^{\infty} (v, w_m)_H w_m$$

and, since $\lambda_m \geq \lambda_1$ for every $m > 1$,

$$a(v, v) = \sum_{m=1}^{\infty} \lambda_m (v, w_m)_H^2 \geq \lambda_1 \sum_{m=1}^{\infty} (v, w_m)_H^2 = \lambda_1 \|v\|_H^2. \quad (6.104)$$

Moreover,

$$a(w_1, w_1) = \lambda_1.$$

We infer that $\lambda_1 \leq R(v)$ for every $v \in V$ and that $\lambda_1 = R(w_1)$. Since $R(w_1) = R(u_1)$, (6.103) follows.

(2) Let d_1 be the dimension of the eigenspace corresponding to λ_1 , spanned by $\{w_1, \dots, w_{d_1}\}$. Then we have

$$\lambda_1 \sum_{m=1}^{\infty} (w, w_m)_H^2 = a(w, w) = \lambda_1 \sum_{m=1}^{d_1} (w, w_m)_H^2 + \sum_{m=d_1+1}^{\infty} \lambda_m (w, w_m)_H^2.$$

This forces

$$(\lambda_m - \lambda_1) (w, w_m)_H^2 = 0$$

for $m \geq d_1 + 1$, since $\lambda_m > \lambda_1$ when $m \geq d_1 + 1$. Then $w = \sum_{m=1}^{d_1} (w, w_m)_H w_m$ and therefore w is an eigenvector of a corresponding to λ_1 . \square

Also the other eigenvalues have a similar variational characterization. Assume a, V and H are as in Theorem 6.74. Let $W_k = \text{span}\{w_1, \dots, w_k\}$. We have:

Theorem 6.77. *The following characterization holds for the k -th eigenvalue, $k \geq 2$:*

$$\lambda_k = \min \{R(v) : v \neq 0, v \in V \cap W_{k-1}^\perp\}. \quad (6.105)$$

Moreover, the minimum is attained at any eigenvector corresponding to λ_k .

Proof. For every $v \in V \cap W_{k-1}^\perp$, we can write:

$$v = \sum_{m=1}^{\infty} (v, w_m)_H w_m = \sum_{m=k}^{\infty} (v, w_m)_H w_m.$$

Since $\lambda_m \geq \lambda_k$ for every $m \geq k$, we get

$$a(v, v) = \sum_{m=k}^{\infty} \lambda_m (v, w_m)_H^2 \geq \lambda_k \sum_{m=k}^{\infty} (v, w_m)_H^2 = \lambda_k \|v\|_H^2. \quad (6.106)$$

Moreover, if w is an eigenvector corresponding to λ_k , we have

$$w = \sum_{w_m \in \mathcal{N}(a, \lambda_k)} (w, w_m) w_m$$

and therefore

$$a(w, w) = \lambda_k \sum_{w_m \in \mathcal{N}(a, \lambda_k)} (w, w_m)_H^2 = \lambda_k \|w\|_H^2$$

so that $\lambda_k = R(w)$. □

There is another variational principle characterizing the eigenvalues in our functional setting, known as *the minimax property*²⁶, that avoids the use of any orthonormal basis of H . We refer to Problem 6.22 for both statement and proof.

6.10 Fixed Points Theorems

Fixed point theorems constitute a fundamental tool for solving *nonlinear* problems which can be reformulated by an equation of the form $F(x) = x$. A solution of this equation is a point which is mapped into itself and for this reason it is called *fixed point of F* .

In general, F is an operator defined in some metric or normed space X into itself. Here we consider two types of fixed point theorems:

1. Fixed points for maps that contract distances (so called *contractions*).
2. Fixed points involving *compactness and/or convexity*.

The most important theorem in the first class is the *Banach Contraction Theorem* whose natural setting is a complete metric space. This theorem gives a sufficient condition for the existence, uniqueness and stability of a *fixed point*, constructed through an iterative procedure, known as the *method of successive approximation*.

Among those in the second class, we consider the theorems of *Schauder* and of *Leray-Schauder*, stated in Banach spaces. These theorems do not give any information on the uniqueness of the fixed point.

We shall use a fixed point technique based on the above theorems in Sects. 9.5 and 11.2.

²⁶ Due to *Lord Rayleigh*, Theory of Sound, London, 1894.

6.10.1 The Contraction Mapping Theorem

Let (M, d) denote a metric space with distance d and

$$F : M \rightarrow M.$$

We say that F is a *strict contraction* if there exists a number ρ , $0 < \rho < 1$, such that

$$d(F(x), F(y)) \leq \rho d(x, y), \quad \forall x, y \in M. \quad (6.107)$$

Clearly, every contraction is continuous. For instance, a function $F : \mathbb{R}^n \rightarrow \mathbb{R}^n$ is a contraction if it is differentiable and $\sup_{\mathbb{R}^n} |\nabla F_j(\mathbf{x})| \leq \rho < 1/\sqrt{n}$, $j = 1, \dots, n$.

The inequality (6.107) expresses the fact that F contracts the distance between any two points, by at least a percentage $\rho < 1$. This fact confers to a recursive sequence $x_{m+1} = F(x_m)$, $m \geq 0$, starting from an arbitrary point $x_0 \in M$, a remarkable asymptotic stability, particularly useful in numerical approximation methods. Precisely, the following result holds.

Theorem 6.78. *Let (M, d) be a complete metric space and $F : M \rightarrow M$ be a strict contraction:*

$$d(F(x), F(y)) \leq \rho d(x, y), \quad \forall x, y \in M.$$

Then there exists a unique fixed point $x^ \in M$ of F . Moreover, for any choice of $x_0 \in M$, the recursive sequence*

$$x_{m+1} = F(x_m), \quad m \geq 0 \quad (6.108)$$

converges to x^ .*

Proof. Fix $x_0 \in M$. We show that the sequence (6.108) is a Cauchy sequence. In fact we have, for $j \geq 1$,

$$x_{j+1} = F(x_j), \quad x_j = F(x_{j-1})$$

so that

$$d(x_{j+1}, x_j) = d(F(x_j), F(x_{j-1})) \leq \rho d(x_j, x_{j-1}).$$

Iterating from $j - 1$ to $j = 1$, we get

$$d(x_{j+1}, x_j) \leq \rho^j d(x_1, x_0).$$

From the triangular inequality, if $m > k$,

$$d(x_m, x_k) \leq \sum_{j=k}^{m-1} d(x_{j+1}, x_j) \leq d(x_1, x_0) \sum_{j=k}^{m-1} \rho^j \leq d(x_1, x_0) \frac{\rho^k}{1-\rho}.$$

Thus, if $k \rightarrow \infty$, $d(x_m, x_k) \rightarrow 0$ and $\{x_m\}$ is a Cauchy sequence. By the completeness of (M, d) , there exists $x^* \in M$ such that $x_m \rightarrow x^*$. Letting $m \rightarrow \infty$ in the recursive relation (6.108), we obtain $x^* = F(x^*)$, i.e. x^* is a fixed point for F . If y^* is another fixed point for F , we have

$$d(x^*, y^*) = d(F(x^*), F(y^*)) \leq \rho d(x^*, y^*)$$

or

$$(1 - \rho) d(x^*, y^*) \leq 0$$

which implies $d(x^*, y^*) = 0$. Therefore $x^* = y^*$, i.e. the fixed is unique. \square

We emphasize that the initial point x_0 is arbitrarily chosen in M . From a different perspective, Theorem 6.78 affirms that the dynamical system defined by the recursive sequence (6.108) has a unique equilibrium point x^* . Moreover, x^* is asymptotically stable, with attraction basin coincident with M .

6.10.2 The Schauder Theorem

Schauder's theorem extends to a Banach space setting the following Brouwer Fixed Point Theorem, valid in \mathbb{R}^n :

Theorem 6.79. *Let $S \subset \mathbb{R}^n$ be a closed sphere and $T : S \rightarrow S$ be a continuous map. Then T has a fixed point \mathbf{x}^* .*

In dimension $n = 1$, Theorem 6.79 amounts to say that the graph of a continuous function $f : [0, 1] \rightarrow [0, 1]$ intersects the straight line $y = x$ at least one time. In dimension $n > 1$, there are several proofs, all rather nontrivial²⁷.

The Brower Theorem still holds if the sphere S is substituted by a set S' *homeomorphic* (i.e. *topologically equivalent*) to S . This means that there exists a one-to-one and onto function $\varphi : S \leftrightarrow S'$ (a *homeomorphism*) such that φ and φ^{-1} are continuous. A homeomorphism is a mathematical realization of a continuous deformation. Thus, any continuous deformation of S is homeomorphic to S .

The difficulty in extending Brower's Theorem to infinitely many dimensions relies on the fact that a closed sphere in an infinite dimensional Banach or Hilbert is *never compact*. A way overcome the difficulty is to consider *compact and convex sets*. Recall that E is *convex* if, for every $x, y \in E$, the line segment of endpoints x, y is contained in E .

Let X be a Banach space and $E \subset X$. The *convex envelope* of E , denoted by $co\{E\}$, is the smallest convex set containing E . In other words:

$$co\{E\} = \{\cap F : F \text{ convex}, E \subset F\}.$$

The symbol $\overline{co}\{E\}$ denotes the closure of $co\{E\}$, and is called the *closed convex envelope* of E . An important property of the convex envelopes is expressed in the following proposition²⁸.

Proposition 6.80. *If E is compact, then $\overline{co}\{E\}$ is compact.*

An important example of *compact and convex* set is the *closed convex envelope* of a finite number of points:

$$\overline{co}\{x_1, x_2, \dots, x_N\} = \left\{ \sum_{i=1}^N \lambda_i x_i : 0 \leq \lambda_i \leq 1, \sum_{i=1}^N \lambda_i = 1 \right\}.$$

²⁷ A particularly elegant proof can be found in P. Lax, The American Mathematical Monthly, vol. 106, No. 6.

²⁸ For the proof, see e.g. [38], A.E. Taylor, 1958.

Theorem 6.81 (Schauder). *Let X be a Banach space. We are given:*

- i) $A \subseteq X$, compact and convex.
- ii) $T : A \rightarrow A$, continuous.

Then T has a fixed point $x^ \in A$.*

Proof. The idea is to approximate T by means of operators acting on finite dimensional spaces in order to apply Brouwer's Theorem.

Since A is compact, for every $\varepsilon > 0$, we can find a covering of A consisting of a finite number n_ε of open balls $B_1 = B_\varepsilon(x_1), \dots, B_{n_\varepsilon} = B_\varepsilon(x_{n_\varepsilon})$ of radius ε . Let now

$$A_\varepsilon = \overline{\text{co}}\{x_1, x_2, \dots, x_{n_\varepsilon}\}.$$

Being A closed and convex, $A_\varepsilon \subseteq A$. Moreover, A_ε is homeomorphic to the closed unit ball in $\mathbb{R}^{M_\varepsilon}$, for a suitable M_ε , $M_\varepsilon \leq n_\varepsilon$. Define $P_\varepsilon : A \rightarrow A_\varepsilon$ by

$$P_\varepsilon(x) = \frac{\sum_{i=1}^{n_\varepsilon} \text{dist}(x, A - B_i)x_i}{\sum_{i=1}^{n_\varepsilon} \text{dist}(x, A - B_i)}, \quad x \in A.$$

Note that $\text{dist}(x, A - B_i) \neq 0$ if $x \notin B_i$, so that the denominator does not vanish, for every $x \in A$. Moreover, $P_\varepsilon(x)$ is given by a convex combination of the points x_i and hence $P_\varepsilon(x) \in A_\varepsilon$. P_ε is continuous, being a finite linear combination of distances and, for every $x \in A$, we have

$$\|P_\varepsilon(x) - x\| \leq \frac{\sum_{i=1}^{n_\varepsilon} \text{dist}(x, A - B_i) \|x_i - x\|}{\sum_{i=1}^{n_\varepsilon} \text{dist}(x, A - B_i)} < \varepsilon \quad (6.109)$$

since $\text{dist}(x, A - B_i) = 0$, if $x \notin B_i$.

Now, the operator

$$P_\varepsilon \circ T : A_\varepsilon \rightarrow A_\varepsilon$$

is continuous and, by Brouwer's Theorem, there exists a fixed point x_ε :

$$(P_\varepsilon \circ T)(x_\varepsilon) = x_\varepsilon.$$

Using the compactness of A , there exist a sequence $\{x_{\varepsilon_j}\}$ and a point $x^* \in A$ such that $x_{\varepsilon_j} \rightarrow x^*$ as $\varepsilon_j \rightarrow 0$. From (6.109) with $x = T(x_{\varepsilon_j})$, we get

$$\|x_{\varepsilon_j} - T(x_{\varepsilon_j})\| = \|(P_{\varepsilon_j} \circ T)(x_{\varepsilon_j}) - T(x_{\varepsilon_j})\| < \varepsilon_j.$$

Letting $\varepsilon_j \rightarrow 0$, by the continuity of the norm, we finally obtain $\|x^* - T(x^*)\| = 0$, i.e. $x^* = T(x^*)$. \square

The application of Schauder's Theorem requires the compactness of A , which is a strong requirement in infinitely many dimensions. In the applications to boundary value problems, it is much more convenient to formulate variants in which the

compactness is required to the image $T(A)$ or to the operator T itself, rather than to A . A first variant is the following.

Theorem 6.82. *Let X be a Banach space. Assume that:*

- i) $A \subset X$ is closed and convex.
- ii) $T : A \rightarrow A$ is continuous.
- iii) $\overline{T(A)}$ is compact in X .

Then T has a fixed point $x^ \in A$.*

Proof. Let $K = \overline{\text{co}}\{\overline{T(A)}\}$, i.e. the closed convex envelope of $\overline{T(A)}$. Since $\overline{T(A)}$ is compact, K is compact by Proposition 6.80 and $K \subseteq A$. Moreover $T(K) \subseteq T(A) \subseteq K$. Thus, the restriction $T : K \rightarrow K$ falls under the hypotheses of Theorem 6.81 and hence T has a fixed point $x^* \in K$. \square

A second variant uses the *compactness of T* . We recall that T is compact if the image of a bounded set has compact closure. We emphasize that if T is linear and compact then it is clearly continuous, but this is not the case if T is nonlinear²⁹.

Theorem 6.83. *Let X be a Banach space. Assume that:*

- i) $A \subset X$ is closed, bounded and convex.
- ii) $T : A \rightarrow A$ is a continuous and compact operator.

Then T has a fixed point $x^ \in A$.*

Proof. Observe that $\overline{T(A)}$ is compact in A and use Theorem 6.82. \square

6.10.3 The Leray-Schauder Theorem

A further variant of Schauder's Theorem requires the compactness of the operator T and the existence of a family of operators T_s , $0 \leq s \leq 1$, where $T_1 = T$ and T_0 is a compact operator which has a fixed point.

The simplest example is provided by the family $T_s = sT$, with $T_0 = 0$, which has the obvious fixed point $x = 0$. The main hypothesis in the following theorem asserts that, if the family of operators T_s , $0 \leq s \leq 1$, has fixed points, these points form a bounded set. It is also called *a priori estimate* for the fixed points of the family T_s ; “*a priori*” because we do not know whether or not they exist. Under this assumption, T has a fixed point.

Theorem 6.84 (Leray-Schauder). *Let X be a Banach space and $T : X \rightarrow X$ such that:*

- i) T is continuous and compact.

²⁹ Convince yourselves using functions of real variables.

ii) There exists M such that

$$\|x\| < M \quad (6.110)$$

for every solution (x, s) of the equation $x = sT(x)$ with $0 \leq s \leq 1$.

Then T has a fixed point.

Proof. Let

$$B_M = \{x \in X : \|x\| \leq M\}$$

and define $P : B_M \rightarrow B_M$ by

$$P(x) = \begin{cases} T(x) & \text{if } \|T(x)\| \leq M \\ M \frac{T(y)}{\|T(y)\|} & \text{if } \|T(y)\| > M. \end{cases}$$

B_M is closed, convex and bounded and P is continuous. Moreover, since $T(B_M)$ is precompact, $P(B_M)$ also is precompact. By Theorem 6.81, there exists $x^* \in B_M$ such that $P(x^*) = x^*$.

We claim that x^* is a fixed point for T . If not, we should have $P(x^*) \neq T(x^*)$. Therefore, it must be $\|T(x^*)\| > M$ and

$$x^* = P(x^*) = \frac{M}{\|T(x^*)\|} T(x^*) \quad (6.111)$$

or

$$x^* = sT(x^*), \quad \text{with } s = \frac{M}{\|T(x^*)\|} < 1.$$

From i) we infer $\|x^*\| < M$ while (6.111) implies $\|x^*\| = M$. Contradiction. \square

Problems

6.1. Heisenberg Uncertainty Principle. Let $\psi \in C^1(\mathbb{R})$ such that $x[\psi(x)]^2 \rightarrow 0$ as $|x| \rightarrow \infty$ and $\int_{\mathbb{R}} |\psi(x)|^2 dx = 1$. Show that

$$1 \leq 4 \int_{\mathbb{R}} x^2 |\psi(x)|^2 dx \int_{\mathbb{R}} |\psi'(x)|^2 dx.$$

(If ψ is a Schrödinger wave function, the first factor in the right hand side measures the spread of the density of a particle, while the second one measures the spread of its momentum).

6.2. Let H be a Hilbert space and $a(u, v)$ be a symmetric and nonnegative bilinear form in H :

$$a(u, v) = a(v, u) \quad \text{and} \quad a(u, v) \geq 0, \quad \forall u, v \in H.$$

Show that

$$|a(u, v)| \leq \sqrt{a(u, u)} \sqrt{a(v, v)}.$$

[Hint: Mimic the proof of Schwarz's inequality, see Theorem 6.6, p. 359].

6.3. Show the completeness of l^2 (see Example 6.48, p. 391).

[*Hint:* Take a Cauchy sequence $\{\mathbf{x}^k\}$ where $\mathbf{x}^k = \{x_m^k\}_{m \geq 1}$. In particular, $|x_m^k - x_m^h| \rightarrow 0$ as $h, k \rightarrow \infty$ and therefore $x_m^h \rightarrow x_m$ for every m . Define $\mathbf{x} = \{x_m\}_{m \geq 1}$ and show that $\mathbf{x}^k \rightarrow \mathbf{x}$ in l^2].

6.4. Let H be a Hilbert space and V be a closed subspace of H . Show that $u = P_V x$ if and only if

$$\begin{cases} \mathbf{1.} \ u \in V \\ \mathbf{2.} \ (x - u, v) = 0, \forall v \in V. \end{cases}$$

6.5. Let $f \in L^2(-1, 1)$. Find the polynomial of degree not greater than n that gives the best approximation of f in the least squares sense, that is, the polynomial p that minimizes

$$\int_{-1}^1 (f - q)^2$$

among all polynomials q with degree not greater than n .

[*Answer:*

$$p(x) = a_0 L_0(x) + a_1 L_1(x) + \dots + a_n L_n(x),$$

where L_n is the n -th Legendre polynomials and $a_j = (n+1/2)(f, L_n)_{L^2(-1,1)}$].

6.6. Hermite's equation and the quantum mechanics harmonic oscillator. Consider the equation

$$w'' + (2\lambda + 1 - x^2) w = 0, \quad x \in \mathbb{R} \quad (6.112)$$

with $w(x) \rightarrow 0$ as $x \rightarrow \pm\infty$.

a) Show that the change of variables $z = we^{x^2/2}$ transforms (6.112) into Hermite's equation for z :

$$z'' - 2xz' + 2\lambda z = 0$$

with $e^{-x^2/2}z(x) \rightarrow 0$ as $x \rightarrow \pm\infty$.

b) Consider the Schrödinger wave equation for the harmonic oscillator

$$\psi'' + \frac{8\pi^2 m}{h^2} (E - 2\pi^2 m\nu^2 x^2) \psi = 0 \quad x \in \mathbb{R}$$

where m is the mass of the particle, E is the total energy, h is the Plank constant and ν is the vibrational frequency. The physically admissible solutions are those satisfying the following conditions:

$$\psi \rightarrow 0 \text{ as } x \rightarrow \pm\infty \quad \text{and} \quad \|\psi\|_{L^2(\mathbb{R})} = 1.$$

Show that there is a solution if and only if

$$E = h\nu \left(n + \frac{1}{2} \right) \quad n = 0, 1, 2, \dots$$

and, for each n , the corresponding solution is given by

$$\psi_n(x) = k_n H_n \left(2\pi \sqrt{\nu m / h} x \right) \exp \left(-\frac{2\pi^2 \nu m}{h} x^2 \right)$$

where $k_n = \left(\frac{4\pi\nu m}{2^{2n}(n!)^2 h} \right)^{1/2}$ and H_n is the n -th Hermite polynomial.

6.7. Using separation of variables, solve the following steady state diffusion problem in three dimensions (r, θ, φ) spherical coordinates, $0 \leq \theta \leq 2\pi$, $0 \leq \varphi \leq \pi$:

$$\begin{cases} \Delta u = 0 & r < 1, 0 < \varphi < \pi \\ u(1, \varphi) = g(\varphi) & 0 \leq \varphi \leq \pi. \end{cases}$$

[Answer:

$$u(r, \varphi) = \sum_{n=0}^{\infty} a_n r^n L_n(\cos \varphi),$$

where L_n is the $n - th$ Legendre polynomial and

$$a_n = \frac{2n+1}{2} \int_{-1}^1 g(\cos^{-1} x) L_n(x) dx.$$

At a certain point, the change of variable $x = \cos \varphi$ is required].

6.8. The vertical displacement u of a circular membrane of radius a satisfies the bidimensional wave equation $u_{tt} = \Delta u$, with boundary condition $u(a, \theta, t) = 0$. Supposing the membrane initially at rest, write a formal solution of the problem.

[Answer:

$$u(r, \theta, t) = \sum_{p,j=0}^{\infty} J_p(\alpha_{pj} r) \{A_{pj} \cos p\theta + B_{pj} \sin p\theta\} \cos(\sqrt{\alpha_{pj}} t)$$

where the coefficients A_{pj} and B_{pj} are determined by the expansion of the initial condition $u(r, \theta, 0) = g(r, \theta)$ and $\{\alpha_{pj}\}_{j \geq 1}$ is the sequence of positive zeroes of the Bessel's functions of first kind and order J_p defined in (6.31), p. 373].

6.9. In calculus, we say that $f : \mathbb{R}^n \rightarrow \mathbb{R}$ is differentiable at \mathbf{x}_0 if there exists a linear mapping $L : \mathbb{R}^n \rightarrow \mathbb{R}$ such that

$$f(\mathbf{x}_0 + \mathbf{h}) - f(\mathbf{x}_0) = L\mathbf{h} + o(\|\mathbf{h}\|) \quad \text{as } \mathbf{h} \rightarrow \mathbf{0}.$$

Determine the Riesz elements associated with L , with respect to the inner products:

a) $(\mathbf{x}, \mathbf{y}) = \mathbf{x} \cdot \mathbf{y} = \sum_{j=1}^n x_j y_j$, b) $(\mathbf{x}, \mathbf{y})_{\mathbf{A}} = \mathbf{A}\mathbf{x} \cdot \mathbf{y} = \sum_{i,j=1}^n a_{ij} x_i y_j$,

where $\mathbf{A} = (a_{ij})$ is a positive and symmetric matrix.

6.10. Prove Proposition 6.36, p. 381.

[Hint: First show that $\|L^*\|_{\mathcal{L}(H_2, H_1)} \leq \|L\|_{\mathcal{L}(H_1, H_2)}$ and then that $(L^*)^* = L$. Reverse the role of L and L^* to show that $\|L^*\|_{\mathcal{L}(H_2, H_1)} \geq \|L\|_{\mathcal{L}(H_1, H_2)}$].

6.11. Prove Neças Theorem 6.42, p. 386.

[Hint: Try to follow the same steps in the proof of the Lax-Milgram Theorem].

6.12. Let $E \subset X$, where X is a Banach space. Prove the following facts:

a) If E is compact, then it is closed and bounded.

b) $E \subset F$ and F compact, if E closed then E is compact.

6.13. *Projection on a closed convex set.* Let H be a Hilbert space and $E \subset H$, closed and convex.

- a) Show that, for every $x \in H$, there is a unique element $P_E x \in E$ (called the *projection* of x on E) such that

$$\|P_E x - x\| = \inf_{v \in E} \|v - x\|. \quad (6.113)$$

- b) Show that $x^* = P_E x$ if and only if $x^* \in E$ and the following **variational inequality** holds:

$$(x^* - x, v - x^*) \geq 0, \quad \text{for every } v \in E. \quad (6.114)$$

- c) Give a geometrical interpretation of (6.114).

[Hint: a) Follow the proof of the Projection Theorem 6.12, p. 365. b) Let $0 \leq t \leq 1$ and define

$$\varphi(t) = \|x^* + t(v - x^*) - x\|^2 \quad v \in E.$$

Show that $x^* = P_E x$ if and only if $\varphi'(0) \geq 0$. Check that $\varphi'(0) \geq 0$ is equivalent to (6.114)].

6.14. Let $\Omega \subseteq \mathbb{R}^n$ be an open set, $a, b \in \mathbb{R}$ and

$$E = \{u \in L^2(\Omega) : a \leq u(\mathbf{x}) \leq b \quad \text{a.e. in } \Omega\}.$$

1. Show that E is a closed and convex subset of $L^2(\Omega)$.

2. Let $z \in L^2(\Omega)$ and $z^* = P_E z$ (see Problem 6.14). Show that the variational inequality (6.114) is equivalent, in this case, to the pointwise inequality

$$\forall v \in E : (z^*(\mathbf{x}) - z(\mathbf{x}))(v(\mathbf{x}) - z^*(\mathbf{x})) \geq 0, \quad \text{a.e. in } \Omega. \quad (6.115)$$

3. Derive the formulas

$$z^*(\mathbf{x}) = \max \{a, \min \{z(\mathbf{x}), b\}\} = \min \{b, \max \{a, z(\mathbf{x})\}\}. \quad (6.116)$$

[Hint: 1. Recall that if $u_k \rightarrow u$ in $L^2(\Omega)$, there exists a subsequence $\{u_{k_j}\}$, such that $u_{k_j} \rightarrow u$ a.e. in Ω .

2. The implication (6.115) \Rightarrow (6.114) is obvious. To prove the opposite one, argue by contradiction, assuming that (6.114) is true and there exists $\bar{v} \in E$ such that

$$(z^*(\mathbf{x}) - z(\mathbf{x}))(\bar{v}(\mathbf{x}) - z^*(\mathbf{x})) < 0$$

in $\Omega' \subset \Omega$, with $|\Omega'| > 0$. Let $v = \bar{v}$ in Ω' and $v = z^*$ in $\Omega \setminus \Omega'$ and deduce a contradiction with (6.114).

3. Show that (6.115) implies, a.e. in Ω , $z^*(\mathbf{x}) = z(\mathbf{x})$ if $a \leq z(\mathbf{x}) \leq b$, $z^*(\mathbf{x}) = a$ if $z(\mathbf{x}) < a$ and $z^*(\mathbf{x}) = b$, if $z(\mathbf{x}) > b\}$.

6.15. Let H be a Hilbert space and $E \subset H$, be closed and convex. Show that E is *weakly sequentially closed*.

[Hint: Let $\{x_k\} \subset E$ such that $x_k \rightharpoonup x$. Use (6.114) to show that $P_E x = x$, so that $x \in E$].

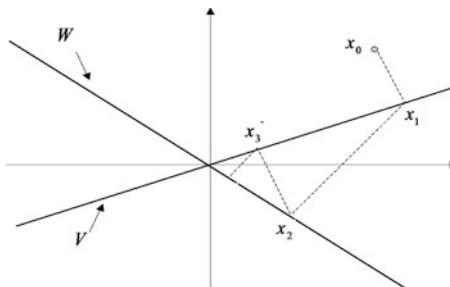


Fig. 6.3 The sequence of projections in problem 6.20 (a)

6.16. Let H be a Hilbert space and $\{x_k\} \subset E$ such that $x_k \rightharpoonup x$. Show that if:

i) $\|x_k\| \rightarrow \|x\|$ as $k \rightarrow \infty$ or ii) $\|x_k\| \leq \|x\|$ for every $k \geq 1$,

then $x_k \rightarrow x$.

6.17. Show that the embedding of $H_{per}^1(0, 2\pi)$ into $L^2(0, 2\pi)$ is compact.

[Hint: Let $\{u_k\} \subset H_{per}^1(0, 2\pi)$ with

$$\|u_k\|^2 = \sum_{m \in \mathbb{Z}} (1 + m^2) |\widehat{u_k}_m|^2 < M.$$

Show that, by a diagonal process, it is possible to select indexes k_j such that, for each m , $\widehat{u_{k_j}}_m$ converges to some number U_m . Let

$$u(x) = \sum_{m \in \mathbb{Z}} U_m e^{imx}$$

and show that $u_{k_j} \rightarrow u$ in $L^2(0, 2\pi)$].

6.18. Let $\{e_k\}_{k \geq 1}$ be an orthonormal basis in an infinite dimensional Hilbert space H . Show that $e_k \rightarrow 0$ as $k \rightarrow \infty$.

[Hint: Every $x \in H$ has the expansion $x = \sum_{k=1}^{\infty} (x, e_k) e_k$ with $\|x\|^2 = \dots$].

6.19. Let V and W be two closed subspaces of a Hilbert space H , with inner product (\cdot, \cdot) . Let $x_0 \in H$ and define the following sequence of projections (see Fig. 6.3):

$$x_{2n+1} = P_V(x_{2n}), \quad x_{2n+2} = P_W(x_{2n+1}), \quad n \geq 0.$$

Prove that:

- (a) If $V \cap W = \{0\}$ then $x_n \rightarrow 0$.
- (b) If $V \cap W \neq \{0\}$, then $x_n \rightarrow P_{V \cap W}(x_0)$

by filling in the details in the following steps.

1. Check that $\|x_{n+1}\|^2 = (x_{n+1}, x_n)$. By computing $\|x_{n+1} - x_n\|^2$, show that $\|x_n\|$ is decreasing (hence $\|x_n\| \rightarrow l \geq 0$) and $\|x_{n+1} - x_n\| \rightarrow 0$.

2. If $V \cap W = \{0\}$, show that if a subsequence $x_{2n_k} \rightharpoonup x$, then $x_{2n_k+1} \rightharpoonup x$ as well. Deduce that $x = 0$ (so that the entire sequence converges weakly to 0).

3. Show that

$$\|x_n\|^2 = (x_{n+1}, x_{n-2}) = (x_{n+2}, x_{n-3}) = \cdots = (x_{2n-1}, x_0)$$

and deduce that $x_n \rightarrow 0$.

4. If $V \cap W \neq \{0\}$, let $z_n = x_n - P_{V \cap W}(x_0)$ and reduce this case to the case (a).

6.20. Let $L : L^2(\mathbb{R}) \rightarrow L^2(\mathbb{R})$ be defined by $Lv(x) = v(-x)$ a.e.. Show that $\sigma(L) = \sigma_P(L) = \{1, -1\}$. Find the corresponding eigenfunctions.

6.21. *Minimax property of the eigenvalues.* Let V, H and a as in Theorem 6.74, p. 411. Denote by $R(v)$ the Rayleigh quotient (6.102) and by $\{\lambda_m\}_{m \geq 1}$ the nondecreasing sequence of eigenvalue of the bilinear form a .

1. Let v_1, \dots, v_{k-1} be $k-1$ linearly independent elements in V , $k \geq 2$. Set

$$\lambda_k^*(v_1, \dots, v_{k-1}) = \inf \{R(v) : (v, v_j)_H = 0, j = 1, \dots, k-1\}.$$

Show that $\lambda_k^* \leq \lambda_k$.

2. Show that

$$\max_{S_k} \lambda_k^*(v_1, \dots, v_{k-1}) = \lambda_k$$

where

$$S_k = \{(v_1, \dots, v_{k-1}), v_j \in V, j = 1, \dots, k-1, \text{ linearly independent}\}.$$

[Hint: 1. Show that it is possible to construct a linear combination of the first k eigenfunctions $\bar{v} = \sum_1^k c_j w_j$, which is orthogonal to each v_j , $j = 1, \dots, k-1$. Prove that $R(\bar{v}) \leq \lambda_k$. 2. Observe that $\lambda_k^*(w_1, \dots, w_{k-1}) = \lambda_k$].

6.22. Let (M, d) be a complete metrix space and $F(\cdot, \cdot) : M \times \mathbb{R} \rightarrow M$ with the following properties:

i) There exists ρ , $0 < \rho < 1$, such that

$$d(F(x, b), F(y, b)) \leq \rho d(x, y), \quad \forall x, y \in M, \forall b \in \mathbb{R}.$$

ii) There exists a number $L > 0$ such that

$$d(F(x, b_1), F(x, b_2)) \leq L |b_1 - b_2|, \quad \forall x \in M, \forall b_1, b_2 \in \mathbb{R}.$$

Show that the equation $F(x, b) = x$ has a unique fixed point $x^* = x^*(b)$ for every $b \in \mathbb{R}$ and

$$d(x(b_1), x(b_2)) \leq \frac{L}{1-\rho} |b_1 - b_2|, \quad \forall b_1, b_2 \in \mathbb{R}.$$

Chapter 7

Distributions and Sobolev Spaces

7.1 Distributions. Preliminary Ideas

We have seen the concept of *Dirac measure* arising in connection with the fundamental solutions of the diffusion and the wave equations. Another interesting situation is the following, where the Dirac measure models a mechanical impulse.

Consider a mass m moving along the x -axis with constant speed $v\vec{i}$ (see Fig. 7.1). At time $t = t_0$ an *elastic* collision with a vertical wall occurs. After the collision, the mass moves with opposite speed $-v\vec{i}$. If v_2, v_1 denote the scalar speeds at times t_1, t_2 , $t_1 < t_2$, by the laws of mechanics we should have

$$m(v_2 - v_1) = \int_{t_1}^{t_2} F(t) dt,$$

where F denotes the intensity of the force acting on m . When $t_1 < t_2 < t_0$ or $t_0 < t_1 < t_2$, then $v_2 = v_1 = v$ or $v_2 = v_1 = -v$, respectively, and therefore $F = 0$: no force is acting on m before and after the collision. However, if $t_1 < t_0 < t_2$, the left hand side is equal to $2mv \neq 0$. If we insist to model the intensity of the force by a function F , the integral in the right hand side is zero and we obtain a contradiction.

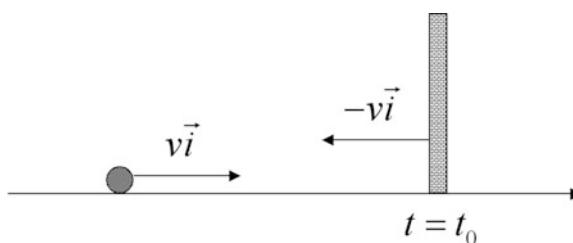


Fig. 7.1 Elastic collision at time $t = t_0$

Indeed, in this case, F is a force concentrated at time t_0 , of intensity $2mv$, that is

$$F(t) = 2mv \delta(t - t_0).$$

In this chapter we see how the Dirac delta is perfectly included in the theory of *distributions or Schwartz generalized functions*. We already mentioned in Subsect. 2.3.3 that the key idea in this theory is to describe a mathematical object through its action on smooth test functions φ , with compact support. In the case of the Dirac δ , such action is expressed by the formula (see Definition 2.8, p. 45)

$$\int \delta(x) \varphi(x) dx = \varphi(0)$$

where, we recall, the integral symbol is purely formal. As we shall shortly see, the appropriate notation is $\langle \delta, \varphi \rangle = \varphi(0)$.

Of course, by a principle of coherence, among the *generalized functions* we should be able to recover the *usual* functions of Analysis. This fact implies that the choice of the test functions cannot be arbitrary. In fact, let $\Omega \subseteq \mathbb{R}^n$ be a domain and take for instance a function $u \in L^2(\Omega)$. A natural way to define the *action* of u on a test φ is

$$\langle u, \varphi \rangle = (u, \varphi)_{L^2(\Omega)} = \int_{\Omega} u \varphi d\mathbf{x}.$$

If we let φ be varying over all $L^2(\Omega)$, we know from the last chapter, that $\langle u, \varphi \rangle$ identifies uniquely u . Indeed, if $v \in L^2(\Omega)$ is such that $\langle u, \varphi \rangle = \langle v, \varphi \rangle$ for every $\varphi \in L^2(\Omega)$, we have

$$0 = \langle u - v, \varphi \rangle = \int_{\Omega} (u - v) \varphi d\mathbf{x} \quad \forall \varphi \in L^2(\Omega) \quad (7.1)$$

which forces (why?) $u = v$ a.e. in Ω .

On the other hand, we cannot use L^2 -functions as test functions since, for instance, $\langle \delta, \varphi \rangle = \varphi(0)$ does not have any meaning.

We ask: is it possible to reconstruct u from the knowledge of $(u, \varphi)_{L^2(\Omega)}$, when φ varies on a set of *nice* functions?

Certainly this is impossible if we use only a restricted set of *test* functions. However, it is possible to recover u from the value of $(u, \varphi)_{L^2(\Omega)}$, when φ varies in a **dense** set in $L^2(\Omega)$. In fact, let $(u, \varphi)_{L^2(\Omega)} = (v, \varphi)_{L^2(\Omega)}$ for every test function. Given $\psi \in L^2(\Omega)$, there exists a sequence of test functions $\{\varphi_k\}$ such that $\|\varphi_k - \psi\|_{L^2(\Omega)} \rightarrow 0$. Then¹,

$$0 = \int_{\Omega} (u - v) \varphi_k d\mathbf{x} \rightarrow \int_{\Omega} (u - v) \psi d\mathbf{x}$$

¹ From

$$\left| \int_{\Omega} (u - v) (\varphi_k - \psi) d\mathbf{x} \right| \leq \|u - v\|_{L^2(\Omega)} \|\varphi_k - \psi\|_{L^2(\Omega)}.$$

so that (7.1) still holds for every $\psi \in L^2(\Omega)$ and $(u, \varphi)_{L^2(\Omega)}$ identifies a unique element in $L^2(\Omega)$.

Thus, the set of test functions must be *dense in* $L^2(\Omega)$ if we want L^2 -functions to be seen as *distributions*. In the next section we construct an appropriate set of test functions.

However, the main purpose of introducing the Schwartz distributions is not restricted to a mere extension of the notion of function but it relies on the possibility of broadening the domain of *calculus* in a significant way, opening the door to an enormous amount of new applications. Here the key idea is to use integration by parts to carry the derivatives onto the test functions. Actually, this is not a new procedure. For instance, we have used it in Subsect. 2.3.3, when have interpreted the Dirac delta at $x = 0$ as the derivative of the Heaviside function \mathcal{H} , (see formula (2.63) and footnote 19, page 46).

Also, the weakening of the notion of solution of conservation laws (Subsect. 4.4.2) or of the wave equation (Subsect. 5.4.2) follows more or less the same pattern.

In the first part of this chapter we give the basic concepts of the theory of Schwartz distributions, mainly finalized to the introduction of Sobolev spaces. The basic reference are the books [37] of *L. Schwartz, 1966*, or [34] of *Gelfand and Shilov, 1964*, to which we refer for the proofs we do not present here.

7.2 Test Functions and Mollifiers

Recall that, given a continuous function v , defined in a domain $\Omega \subseteq \mathbb{R}^n$, the *support* of v is given by the closure, in the relative topology of Ω , of the set of points where v is different from zero:

$$\text{supp}(v) = \Omega \cap \text{closure of } \{\mathbf{x} \in \Omega : v(\mathbf{x}) \neq 0\}.$$

Actually, the support or, better, the essential support, is defined also for measurable functions, not necessarily continuous in Ω . Namely, let Z be the union of the open sets on which $v = 0$ a.e. Then, $\Omega \setminus Z$ is called the *essential support* of v and we use the same symbol $\text{supp}(v)$ to denote it.

We say that v is *compactly supported* in Ω , if $\text{supp}(v)$ is a *compact* subset of Ω .

Definition 7.1. Denote by $C_0^\infty(\Omega)$ the set of functions belonging to $C^\infty(\Omega)$, compactly supported in Ω . We call test functions the elements of $C_0^\infty(\Omega)$.

The reader can easily check that the function given by

$$\eta(\mathbf{x}) = \begin{cases} c \exp\left(\frac{1}{|\mathbf{x}|^2 - 1}\right) & 0 \leq |\mathbf{x}| < 1 \\ 0 & |\mathbf{x}| \geq 1 \end{cases} \quad (c \in \mathbb{R}) \quad (7.2)$$

belongs to $C_0^\infty(\mathbb{R}^3)$.

The function (7.2) is a typical and important example of test function. Indeed, we will see below that many other test functions can be generated by convolution with (7.2).

Let us briefly recall the definition and the main properties of the convolution of two functions. Given two functions u and v defined in \mathbb{R}^n , the *convolution* $u * v$ of u and v is given by the formula:

$$(u * v)(\mathbf{x}) = \int_{\mathbb{R}^n} u(\mathbf{x} - \mathbf{y}) v(\mathbf{y}) d\mathbf{y} = \int_{\mathbb{R}^n} u(\mathbf{y}) v(\mathbf{x} - \mathbf{y}) d\mathbf{y}.$$

It can be proved that (*Young's Theorem*): if $u \in L^p(\mathbb{R}^n)$ and $v \in L^q(\mathbb{R}^n)$, $p, q \in [1, \infty]$, then $u * v \in L^r(\mathbb{R}^n)$ where $\frac{1}{r} = \frac{1}{p} + \frac{1}{q} - 1$ and

$$\|u * v\|_{L^r(\mathbb{R}^n)} \leq \|u\|_{L^p(\mathbb{R}^n)} \|v\|_{L^q(\mathbb{R}^n)}.$$

The convolution is a very useful device to regularize “wild functions”. Indeed, consider the function η defined in (7.2). We have:

$$\eta \geq 0 \quad \text{and} \quad \text{supp}(\eta) = \overline{B}_1(\mathbf{0}),$$

where, we recall, $B_R(\mathbf{0}) = \{\mathbf{x} \in \mathbb{R}^n : |\mathbf{x}| < R\}$. Choose

$$c = \left(\int_{B_1(\mathbf{0})} \exp\left(\frac{1}{|\mathbf{x}|^2 - 1}\right) d\mathbf{x} \right)^{-1}$$

so that $\int_{\mathbb{R}^n} \eta = 1$. Set, for $\varepsilon > 0$,

$$\eta_\varepsilon(\mathbf{x}) = \frac{1}{\varepsilon^n} \eta\left(\frac{\mathbf{x}}{\varepsilon}\right). \quad (7.3)$$

This function belongs to $C_0^\infty(\mathbb{R}^n)$ (and therefore to all $L^p(\mathbb{R}^n)$), with support equal to $\overline{B}_\varepsilon(\mathbf{0})$, and still $\int_{\mathbb{R}^n} \eta_\varepsilon = 1$.

Let now $f \in L^p(\Omega)$, $1 \leq p \leq \infty$. If we set $f \equiv 0$ outside Ω , we obtain a function in $L^p(\mathbb{R}^n)$, denoted by \tilde{f} , for which the convolution $\eta_\varepsilon * \tilde{f}$ is well defined in all \mathbb{R}^n :

$$\begin{aligned} f_\varepsilon(\mathbf{x}) &= (\eta_\varepsilon * \tilde{f})(\mathbf{x}) = \int_{\Omega} \eta_\varepsilon(\mathbf{x} - \mathbf{y}) f(\mathbf{y}) d\mathbf{y} \\ &= \int_{B_\varepsilon(\mathbf{0})} \eta(\mathbf{z}) \tilde{f}(\mathbf{x} - \mathbf{z}) d\mathbf{z}. \end{aligned}$$

Observe that, since $\int_{\mathbb{R}^n} \eta_\varepsilon = 1$, $\tilde{f} * \eta_\varepsilon$ may be considered as a *convex weighted average* of \tilde{f} and, as such, we expect a smoothing effect on \tilde{f} . Indeed, even if f is very irregular, f_ε is a C^∞ -function. For this reason η_ε is called a *mollifier*. Moreover, as $\varepsilon \rightarrow 0$, f_ε is an approximation of f in the sense explained in the following important lemma.

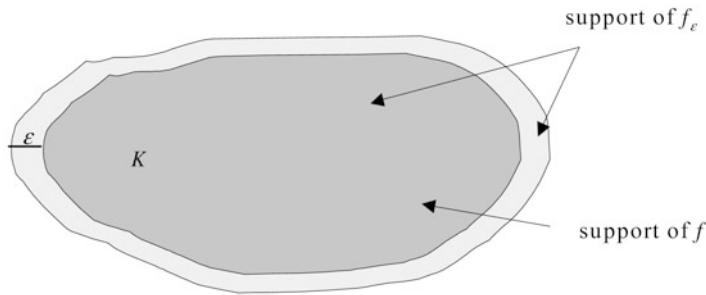


Fig. 7.2 Support of the convolution with a ε -mollifier

Lemma 7.2. Let $f \in L^p(\Omega)$, $1 \leq p \leq \infty$; then f_ε has the following properties:

- a. The support of f_ε is contained in a ε -neighborhood of the support of f (Fig. 7.2):

$$\text{supp}(f_\varepsilon) \subseteq \{\mathbf{x} \in \mathbb{R}^n : \text{dist}(\mathbf{x}, \text{supp}(f)) \leq \varepsilon\}.$$

- b. $f_\varepsilon \in C^\infty(\mathbb{R}^n)$ and, if $\text{supp}(f)$ is a compact set $K \subset \Omega$, then $f_\varepsilon \in C_0^\infty(\Omega)$, for $\varepsilon < \text{dist}(K, \partial\Omega)$.

- c. If $f \in C(\Omega)$, $f_\varepsilon \rightarrow f$ uniformly in every compact $K \subset \Omega$, as $\varepsilon \rightarrow 0$.

- d. If $1 \leq p < \infty$, then

$$\|f_\varepsilon\|_{L^p(\Omega)} \leq \|f\|_{L^p(\Omega)} \quad \text{and} \quad \|f_\varepsilon - f\|_{L^p(\Omega)} \rightarrow 0 \quad \text{as } \varepsilon \rightarrow 0.$$

Proof. a. Let $K = \text{supp}(f)$. If $|\mathbf{z}| \leq \varepsilon$ and $\text{dist}(\mathbf{x}, K) > \varepsilon$, then $\tilde{f}(\mathbf{x} - \mathbf{z}) = 0$ so that $f_\varepsilon(\mathbf{x}) = 0$.

b. Since $\mathbf{x} \mapsto \eta_\varepsilon(\mathbf{x} - \mathbf{y}) \in C_0^\infty(\mathbb{R}^n)$ for every \mathbf{y} , f_ε is continuous and there is no problem in differentiating under the integral sign, obtaining all the time a continuous function. Thus $f_\varepsilon \in C^\infty(\mathbb{R}^n)$. From a, if K is compact, the support of f_ε is compact as well and contained in Ω , if $\varepsilon < \text{dist}(K, \partial\Omega)$. Therefore $f_\varepsilon \in C_0^\infty(\Omega)$.

c. Since $\int_{\mathbb{R}^n} \eta_\varepsilon = 1$, We can write

$$f_\varepsilon(\mathbf{x}) - f(\mathbf{x}) = \int_{\{|\mathbf{z}| \leq \varepsilon\}} \eta(\mathbf{z}) [\tilde{f}(\mathbf{x} - \mathbf{z}) - f(\mathbf{x})] d\mathbf{z}.$$

If $\mathbf{x} \in K \subset \Omega$, K compact, and $\varepsilon < \frac{1}{2}\text{dist}(K, \partial\Omega)$, then $\tilde{f}(\mathbf{x} - \mathbf{z}) = f(\mathbf{x} - \mathbf{z})$ so that

$$|f_\varepsilon(\mathbf{x}) - f(\mathbf{x})| \leq \sup_{|\mathbf{z}| \leq \varepsilon} |f(\mathbf{x} - \mathbf{z}) - f(\mathbf{x})|.$$

Since f is uniformly continuous in every compact subset of Ω , we have that

$$\sup_{|\mathbf{z}| \leq \varepsilon} |f(\mathbf{x} - \mathbf{z}) - f(\mathbf{x})| \rightarrow 0,$$

uniformly in \mathbf{x} , as $\varepsilon \rightarrow 0$. Thus $f_\varepsilon \rightarrow f$ uniformly in K .

d. From Hölder's inequality, we have, for $q = p/(p-1)$,

$$\begin{aligned} f_\varepsilon(\mathbf{x}) &= \int_{\Omega} \eta_\varepsilon(\mathbf{x} - \mathbf{y}) f(\mathbf{y}) d\mathbf{y} = \int_{\Omega} (\eta_\varepsilon(\mathbf{x} - \mathbf{y}))^{1/q} (\eta_\varepsilon(\mathbf{x} - \mathbf{y}))^{1/p} f(\mathbf{y}) d\mathbf{y} \\ &\leq \left(\int_{\Omega} \eta_\varepsilon(\mathbf{x} - \mathbf{y}) |f(\mathbf{y})|^p d\mathbf{y} \right)^{1/p}. \end{aligned}$$

This inequality and Fubini's Theorem² yield

$$\|f_\varepsilon\|_{L^p(\Omega)} \leq \|f\|_{L^p(\Omega)}. \quad (7.4)$$

In fact:

$$\begin{aligned} \|f_\varepsilon\|_{L^p(\Omega)}^p &= \int_{\Omega} |f_\varepsilon(\mathbf{x})|^p d\mathbf{x} \leq \int_{\Omega} \left(\int_{\Omega} \eta_\varepsilon(\mathbf{x} - \mathbf{y}) |f(\mathbf{y})|^p d\mathbf{y} \right) d\mathbf{x} \\ &= \int_{\Omega} |f(\mathbf{y})|^p \left(\int_{\Omega} \eta_\varepsilon(\mathbf{x} - \mathbf{y}) d\mathbf{x} \right) d\mathbf{y} \leq \int_{\Omega} |f(\mathbf{y})|^p d\mathbf{y} = \|f\|_{L^p(\Omega)}^p. \end{aligned}$$

From Theorem B.11, p. 669, given any $\delta > 0$, there exists $g \in C_0(\Omega)$ such that $\|g - f\|_{L^p(\Omega)} < \delta$. Then, (7.4) implies

$$\|g_\varepsilon - f_\varepsilon\|_{L^p(\Omega)} \leq \|g - f\|_{L^p(\Omega)} < \delta.$$

Moreover, since the support of g is compact, $g_\varepsilon \rightarrow g$ uniformly in Ω , by c, so that, we have, in particular, $\|g_\varepsilon - g\|_{L^p(\Omega)} < \delta$, for ε small. Thus,

$$\|f - f_\varepsilon\|_{L^p(\Omega)} \leq \|f - g\|_{L^p(\Omega)} + \|g - g_\varepsilon\|_{L^p(\Omega)} + \|g_\varepsilon - f_\varepsilon\|_{L^p(\Omega)} \leq 3\delta.$$

This shows that $\|f - f_\varepsilon\|_{L^p(\Omega)} \rightarrow 0$ as $\varepsilon \rightarrow 0$. □

Remark 7.3. Let $f \in L_{loc}^1(\Omega)$, i.e. $f \in L^1(\Omega')$ for every³ $\Omega' \subset\subset \Omega$. The convolution $f_\varepsilon(\mathbf{x})$ is well defined if \mathbf{x} stays ε -away from $\partial\Omega$, that is if \mathbf{x} belongs to the set

$$\Omega_\varepsilon = \{\mathbf{x} \in \Omega : \text{dist}(\mathbf{x}, \partial\Omega) > \varepsilon\}.$$

Moreover, $f_\varepsilon \in C^\infty(\Omega_\varepsilon)$.

Remark 7.4. In general $\|f - f_\varepsilon\|_{L^\infty(\Omega)} \not\rightarrow 0$ as $\varepsilon \rightarrow 0$. However $\|f_\varepsilon\|_{L^\infty(\Omega)} \leq \|f\|_{L^\infty(\Omega)}$ is clearly true.

Example 7.5. Let $\Omega' \subset\subset \Omega$ and $f = \chi_{\Omega'}$ be the characteristic function of Ω' . Then, $f_\varepsilon = \chi_{\Omega'} * \eta_\varepsilon \in C_0^\infty(\Omega)$ as long as $\varepsilon < \text{dist}(\Omega', \partial\Omega)$. Note that $0 \leq f_\varepsilon \leq 1$. In fact

$$f_\varepsilon(\mathbf{x}) = \int_{\Omega} \eta_\varepsilon(\mathbf{x} - \mathbf{y}) \chi_{\Omega'}(\mathbf{y}) d\mathbf{y} = \int_{\Omega' \cap B_\varepsilon(\mathbf{x})} \eta_\varepsilon(\mathbf{x} - \mathbf{y}) d\mathbf{y} \leq \int_{B_\varepsilon(\mathbf{0})} \eta_\varepsilon(\mathbf{y}) d\mathbf{y} = 1.$$

² See Appendix B.

³ $\Omega' \subset\subset \Omega$ means that the closure of Ω' is a compact subset of Ω .

Moreover, $f \equiv 1$ in Ω'_ε . In fact, if $\mathbf{x} \in \Omega'_\varepsilon$, the ball $B_\varepsilon(\mathbf{x})$ is contained in Ω' and therefore

$$\int_{\Omega' \cap B_\varepsilon(\mathbf{x})} \eta_\varepsilon(\mathbf{x} - \mathbf{y}) d\mathbf{y} = \int_{B_\varepsilon(\mathbf{0})} \eta_\varepsilon(\mathbf{y}) d\mathbf{y} = 1.$$

A consequence of Lemma 7.2 is the following approximation theorem.

Theorem 7.6. $C_0^\infty(\Omega)$ is dense in $L^p(\Omega)$ for every $1 \leq p < \infty$.

Proof. Denote by $L_c^p(\Omega)$ the space of functions in $L^p(\Omega)$, with (essential) support compactly contained in Ω . Let $f \in L_c^p(\Omega)$ and $K = \text{supp}(f)$. From Lemma 7.2.a, we know that $\text{supp}(f_\varepsilon)$ is contained in an ε -neighborhood of K , which is still a compact subset of Ω , for ε small.

Since by Lemma 7.2.d, $f_\varepsilon \rightarrow f$ in $L^p(\Omega)$, we deduce that $C_0^\infty(\Omega)$ is dense in $L_c^p(\Omega)$, if $1 \leq p < \infty$. On the other hand, $L_c^p(\Omega)$ is dense in $L^p(\Omega)$; in fact, let $\{K_m\}$ be a sequence of compact subsets of Ω such that

$$K_m \subset K_{m+1} \quad \text{and} \quad \cup K_m = \Omega.$$

Denote by χ_{K_m} the characteristic function of K_m . Then, we have

$$\{\chi_{K_m} f\} \subset L_c^p(\Omega) \quad \text{and} \quad \|\chi_{K_m} f - f\|_{L^p} \rightarrow 0 \quad \text{as } m \rightarrow +\infty$$

by the Dominated Convergence Theorem⁴, since $|\chi_{K_m} f| \leq |f|$. □

7.3 Distributions

We now endow $C_0^\infty(\Omega)$ with a suitable notion of convergence. Recall that the symbol

$$D^\alpha = \frac{\partial^{\alpha_1}}{\partial x_1^{\alpha_1}} \cdots \frac{\partial^{\alpha_n}}{\partial x_n^{\alpha_n}}, \quad \alpha = (\alpha_1, \dots, \alpha_n),$$

denotes a derivative of order $|\alpha| = \alpha_1 + \dots + \alpha_n$. If $|\alpha| = 0$ we set $D^\alpha \varphi = \varphi$.

Definition 7.7. Let $\{\varphi_k\} \subset C_0^\infty(\Omega)$ and $\varphi \in C_0^\infty(\Omega)$. We say that

$$\varphi_k \rightarrow \varphi \quad \text{in } C_0^\infty(\Omega), \quad \text{as } k \rightarrow +\infty$$

if the following two properties hold:

1. There exists a compact set $K \subset \Omega$ containing the support of every φ_k .
2. $\forall \alpha = (\alpha_1, \dots, \alpha_n)$, $|\alpha| \geq 0$, $D^\alpha \varphi_k \rightarrow D^\alpha \varphi$ uniformly in Ω .

It is not difficult to show that the limit so defined is *unique*. The space $C_0^\infty(\Omega)$ is denoted by $\mathcal{D}(\Omega)$, when endowed with the above notion of convergence.

Following the discussion in the first section, we focus on the linear functionals in $\mathcal{D}(\Omega)$. If L is one of those, we shall use the bracket (or pairing) $\langle L, \varphi \rangle$ or

⁴ See Appendix B.

$\langle L, \varphi \rangle_{\mathcal{D}'(\Omega)}$, if there is any risk of confusion, to denote the action of L on a test function φ .

We say that a linear functional

$$L : \mathcal{D}(\Omega) \rightarrow \mathbb{R}$$

is *continuous* in $\mathcal{D}(\Omega)$ if

$$\langle L, \varphi_k \rangle \rightarrow \langle L, \varphi \rangle, \quad \text{whenever } \varphi_k \rightarrow \varphi \text{ in } \mathcal{D}(\Omega). \quad (7.5)$$

Note that, given the linearity of L , it is enough to check (7.5) in the case $\varphi = 0$.

Definition 7.8. A **distribution** in Ω is a linear continuous functional in $\mathcal{D}(\Omega)$. The set of distributions is denoted by $\mathcal{D}'(\Omega)$.

Two distributions L_1 and L_2 coincide when their action on every test function is the same, i.e. if

$$\langle L_1, \varphi \rangle = \langle L_2, \varphi \rangle, \quad \forall \varphi \in \mathcal{D}(\Omega).$$

To every $u \in L^2(\Omega)$ corresponds the functional I_u whose action on φ is

$$\langle I_u, \varphi \rangle = \int_{\Omega} u \varphi \, dx,$$

which is certainly continuous in $\mathcal{D}(\Omega)$. Therefore I_u is a distribution in $\mathcal{D}'(\Omega)$ and we have seen at the end of Sect. 7.1 that I_u may be identified with u .

Thus, the notion of distribution generalizes the notion of function (in $L^2(\Omega)$) and the pairing $\langle \cdot, \cdot \rangle$ between $\mathcal{D}(\Omega)$ and $\mathcal{D}'(\Omega)$ generalizes the inner product in $L^2(\Omega)$.

The functional I_u is actually well defined for any function $u \in L^1_{loc}(\Omega)$, that, therefore, can be identified with it. On the other hand, if u is a function and $u \notin L^1_{loc}$, u **cannot** represent a distribution. A typical example is $u(x) = 1/x$ which does not belong to $L^1_{loc}(\mathbb{R})$. However, there is a distribution closely related to $1/x$ as we show in Example 7.12, p. 436.

- **Measures.** Special classes of distributions are obtained by relaxing the condition 2 in Definition 7.7. In particular, let $L \in \mathcal{D}'(\Omega)$. If $\langle L, \varphi_k \rangle \rightarrow 0$ whenever the supports of the φ_k are contained in the same compact K and $\varphi_k \rightarrow 0$ uniformly, L is called a *measure*. Equivalently, L is a measure if and only if, for every φ there exists a constant C , depending on L and the support of φ , such that

$$|\langle L, \varphi \rangle| \leq C \max |\varphi|. \quad (7.6)$$

Every function $u \in L^1_{loc}(\Omega)$ is a measure since

$$\left| \int_{\Omega} u \varphi \right| \leq \|u\|_{L^1(\text{supp}(\varphi))} \max |\varphi|.$$

Another important measure is the Dirac delta, defined below.

Definition 7.9. The n -dimensional **Dirac delta** at the origin, denoted by δ_n , is the distribution in $\mathcal{D}'(\mathbb{R}^n)$, whose action is

$$\langle \delta_n, \varphi \rangle = \varphi(\mathbf{0}). \quad (7.7)$$

If $n = 1$ we simply write δ instead of δ_n .

Quite often, especially in the applied sciences, the Dirac delta at $\mathbf{0}$ is denoted by $\delta_n(\mathbf{x})$ and the formula (7.7) is written in the symbolic form

$$\int \delta_n(\mathbf{x}) \varphi(\mathbf{x}) = \varphi(\mathbf{0}).$$

Similarly, the Dirac delta at \mathbf{y} is denoted by $\delta_n(\mathbf{x} - \mathbf{y})$ and defined by the relation

$$\int \delta_n(\mathbf{x} - \mathbf{y}) \varphi(\mathbf{x}) = \varphi(\mathbf{y}).$$

These are the notations that we already used in the previous chapters.

$\mathcal{D}'(\Omega)$ is a linear space. Indeed if α, β are real (or complex) scalars, $\varphi \in \mathcal{D}(\Omega)$ and $L_1, L_2 \in \mathcal{D}'(\Omega)$, we define $\alpha L_1 + \beta L_2 \in \mathcal{D}'(\Omega)$ by means of the formula

$$\langle \alpha L_1 + \beta L_2, \varphi \rangle = \alpha \langle L_1, \varphi \rangle + \beta \langle L_2, \varphi \rangle.$$

In $\mathcal{D}'(\Omega)$ we may introduce a notion of (weak) convergence: $\{L_k\}$ converges to L in $\mathcal{D}'(\Omega)$ if

$$\langle L_k, \varphi \rangle \rightarrow \langle L, \varphi \rangle, \quad \forall \varphi \in \mathcal{D}(\Omega).$$

If $1 \leq p \leq \infty$, we have the **continuous embeddings**:

$$L^p(\Omega) \hookrightarrow L_{loc}^1(\Omega) \hookrightarrow \mathcal{D}'(\Omega).$$

This means that, if $u_k \rightarrow u$ in $L^p(\Omega)$ or in $L_{loc}^1(\Omega)$, then⁵ $u_k \rightarrow u$ in $\mathcal{D}'(\Omega)$ as well.

With respect to this type of convergence, $\mathcal{D}'(\Omega)$ possesses a *completeness* property that may be used to construct a distribution or to recognize that some linear functional in $\mathcal{D}(\Omega)$ is a distribution. Precisely, one can prove the following result.

Proposition 7.10. Let $\{F_k\} \subset \mathcal{D}'(\Omega)$ such that $\lim_{k \rightarrow \infty} \langle F_k, \varphi \rangle$ exists and is finite for all $\varphi \in \mathcal{D}(\Omega)$. Call $F(\varphi)$ this limit. Then, $F \in \mathcal{D}'(\Omega)$ and $F_k \rightarrow F$ in $\mathcal{D}'(\Omega)$.

⁵ For instance, let $\varphi \in \mathcal{D}(\Omega)$. We have, by Hölder's inequality:

$$\left| \int_{\Omega} (u_k - u) \varphi d\mathbf{x} \right| \leq \|u_k - u\|_{L^p(\Omega)} \|\varphi\|_{L^q(\Omega)}$$

where $q = p/(p-1)$. Then, if $\|u_k - u\|_{L^p(\Omega)} \rightarrow 0$, also $\int_{\Omega} (u_k - u) \varphi d\mathbf{x} \rightarrow 0$, showing the convergence of $\{u_k\}$ in $\mathcal{D}'(\Omega)$.

In particular, if the numerical series

$$\sum_{k=1}^{\infty} \langle F_k, \varphi \rangle$$

converges for all $\varphi \in \mathcal{D}(\Omega)$, then $\sum_{k=1}^{\infty} F_k = F \in \mathcal{D}'(\Omega)$.

Example 7.11. For every $\varphi \in \mathcal{D}(\mathbb{R})$, the numerical series

$$\sum_{k=-\infty}^{\infty} \langle \delta(x-k), \varphi \rangle = \sum_{k=-\infty}^{\infty} \varphi(k)$$

is convergent, since only a finite number of terms is different from zero⁶. From Proposition 7.10, we deduce that the series

$$\text{comb}(x) = \sum_{k=-\infty}^{\infty} \delta(x-k) \quad (7.8)$$

is convergent in $\mathcal{D}'(\mathbb{R})$ and its sum is a distribution called the **Dirac comb**. This name is due to the fact it models a train of impulses concentrated at the integers (see Fig. 7.3, using some ... fantasy).

Example 7.12. Let $h_r(x) = 1 - \chi_{[-r,r]}(x)$ be the characteristic function of the set $\mathbb{R} \setminus [-r, r]$. Define

$$p.v. \frac{1}{x} = \lim_{r \rightarrow 0} \frac{1}{x} h_r(x).$$

We want to show that $p.v. \frac{1}{x}$ defines a distribution in $\mathcal{D}'(\mathbb{R})$, called *principal value* of $\frac{1}{x}$. By Proposition 7.10 it is enough to check that, for all $\varphi \in \mathcal{D}(\mathbb{R})$, the limit

$$\lim_{r \rightarrow 0} \int_{\mathbb{R}} \frac{1}{x} h_r(x) \varphi(x) dx$$

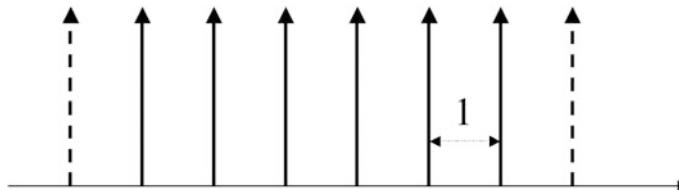


Fig. 7.3 A train of impulses

⁶ Only a finite number of integers k belongs to the support of φ .

is finite. Indeed, assume that $\text{supp}(\varphi) \subset [-a, a]$. Then,

$$\int_{\mathbb{R}} \frac{1}{x} h_r(x) \varphi(x) dx = \int_{\{r < |x| < a\}} \frac{\varphi(x)}{x} dx = \int_{\{r < |x| < a\}} \frac{\varphi(x) - \varphi(0)}{x} dx$$

since

$$\int_{\{r < |x| < a\}} \frac{\varphi(0)}{x} dx = 0,$$

due to the odd symmetry of $1/x$. Now, we have

$$\varphi(x) - \varphi(0) = \varphi'(0)x + o(x), \quad \text{as } x \rightarrow 0,$$

so that

$$\frac{\varphi(x) - \varphi(0)}{x} = \varphi'(0) + o(1), \quad \text{as } x \rightarrow 0.$$

This implies that $[\varphi(x) - \varphi(0)]/x$ is summable in $[-a, a]$ and therefore

$$\lim_{r \rightarrow 0} \int_{\{r < |x| < a\}} \frac{\varphi(x) - \varphi(0)}{x} dx = \int_{\{|x| < a\}} \frac{\varphi(x) - \varphi(0)}{x} dx$$

is a finite number. Thus, $p.v.\frac{1}{x} \in \mathcal{D}'(\mathbb{R})$ and the above computations yield

$$\langle p.v.\frac{1}{x}, \varphi \rangle = \lim_{r \rightarrow 0} \int_{\{r < |x|\}} \frac{\varphi(x)}{x} dx \equiv p.v. \int_{\mathbb{R}} \frac{\varphi(x)}{x} dx$$

where *p.v.* stays for *principal value*.

- *Support of a distribution.* The Dirac δ_n is *concentrated at a point*. More precisely, we say that its **support** coincides with a point. The support of a general distribution F may be defined in the following way. We want to characterize the smallest closed set outside of which F vanishes. However, we cannot proceed as in the case of a function, since a distribution is defined on the elements of $\mathcal{D}(\Omega)$, not on subsets of \mathbb{R}^n .

Thus, let us start saying that $F \in \mathcal{D}'(\Omega)$ vanishes in an open set $A \subset \Omega$ if

$$\langle F, \varphi \rangle = 0$$

for every $\varphi \in \mathcal{D}(\Omega)$ whose support is contained in A . Let \mathcal{A} be the *union of all open sets where F vanishes*. \mathcal{A} is open. Then, we define:

$$\text{supp}(F) = \Omega \setminus \mathcal{A}.$$

For example, $\text{supp}(\text{comb}) = \mathbb{Z}$.

Remark 7.13. Let $F \in \mathcal{D}'(\Omega)$ with compact support K . Then the bracket $\langle F, v \rangle$ is well defined for all $v \in C^\infty(\Omega)$, **not necessarily with compact support**. In fact, let $\varphi \in \mathcal{D}(\Omega)$, $0 \leq \varphi \leq 1$, such that $\varphi \equiv 1$ in an open neighborhood of K

(see Example 7.5, p. 432). Then $v\varphi \in \mathcal{D}(\Omega)$ and we can define

$$\langle F, v \rangle = \langle F, v\varphi \rangle.$$

Note that $\langle F, v\varphi \rangle$ is independent of the choice of φ . Indeed if ψ has the same property of φ , then

$$\langle F, v\varphi \rangle - \langle F, v\psi \rangle = \langle F, v(\varphi - \psi) \rangle = 0$$

since $\varphi - \psi = 0$ in an open neighborhood of K .

7.4 Calculus

7.4.1 The derivative in the sense of distributions

A central concept in the theory of the Schwartz distributions is the notion of *weak* or *distributional derivative*. Clearly we have to abandon the classical definition, since, for instance, we are going to define the derivative for a function $u \in L^1_{loc}$, which may be quite irregular.

The idea is to carry the derivative onto the test functions, as if we were using the integration by parts formula.

Let us start from a function $u \in C^1(\Omega)$. If $\varphi \in \mathcal{D}(\Omega)$, denoting by $\nu = (\nu_1, \dots, \nu_n)$ the outward normal unit vector to $\partial\Omega$, we have

$$\begin{aligned} \int_{\Omega} \varphi \partial_{x_i} u \, d\mathbf{x} &= \int_{\partial\Omega} \varphi u \, \nu_i \, d\mathbf{x} - \int_{\Omega} u \partial_{x_i} \varphi \, d\mathbf{x} \\ &= - \int_{\Omega} u \partial_{x_i} \varphi \, d\mathbf{x} \end{aligned}$$

since $\varphi = 0$ on $\partial\Omega$. The equation

$$\int_{\Omega} \varphi \partial_{x_i} u \, d\mathbf{x} = - \int_{\Omega} u \partial_{x_i} \varphi \, d\mathbf{x},$$

interpreted in $\mathcal{D}'(\Omega)$, becomes

$$\langle \partial_{x_i} u, \varphi \rangle = - \langle u, \partial_{x_i} \varphi \rangle. \quad (7.9)$$

Formula (7.9) shows that the action of $\partial_{x_i} u$ on the test function φ equals the action of u on the test function $-\partial_{x_i} \varphi$. On the other hand, formula (7.9) makes perfect sense if we replace u by any $F \in \mathcal{D}'(\Omega)$ and it is not difficult to check that it defines a continuous linear functional in $\mathcal{D}(\Omega)$. This leads to the following fundamental notion:

Definition 7.14. Let $F \in \mathcal{D}'(\Omega)$. The derivative $\partial_{x_i} F$ is the distribution defined by the formula

$$\langle \partial_{x_i} F, \varphi \rangle = -\langle F, \partial_{x_i} \varphi \rangle, \quad \forall \varphi \in \mathcal{D}(\Omega).$$

From (7.9), if $u \in C^1(\Omega)$, its derivatives in the sense of distributions coincide with the classical ones. This is the reason we keep the same notations in the two cases.

Note that the derivative of a distribution is **always defined!** Moreover, since any derivative of a distribution is a distribution, we deduce that **every distribution possesses derivatives of any order** (in $\mathcal{D}'(\Omega)$):

$$\langle D^\alpha F, \varphi \rangle = (-1)^{|\alpha|} \langle F, D^\alpha \varphi \rangle.$$

For example, the second order derivative

$$\partial_{x_i x_k} F = \partial_{x_i} (\partial_{x_k} F)$$

is defined by

$$\langle \partial_{x_i x_k} F, \varphi \rangle = \langle F, \partial_{x_i x_k} \varphi \rangle. \quad (7.10)$$

Not only. Since φ is smooth, then $\partial_{x_i x_k} \varphi = \partial_{x_k x_i} \varphi$ so that (7.10) yields

$$\partial_{x_i x_k} F = \partial_{x_k x_i} F.$$

Thus, for **all** $F \in \mathcal{D}'(\Omega)$ we may always reverse the order of differentiation *without any restriction*.

Example 7.15. Let $u(x) = \mathcal{H}(x)$, the Heaviside function. In $\mathcal{D}'(\mathbb{R})$ we have $\mathcal{H}' = \delta$. In fact, let $\varphi \in \mathcal{D}(\mathbb{R})$. By definition,

$$\langle \mathcal{H}', \varphi \rangle = -\langle \mathcal{H}, \varphi' \rangle.$$

On the other hand, $\mathcal{H} \in L^1_{loc}(\mathbb{R})$, hence

$$\langle \mathcal{H}, \varphi' \rangle = \int_{\mathbb{R}} \mathcal{H}(x) \varphi'(x) dx = \int_0^\infty \varphi'(x) dx = -\varphi(0)$$

whence

$$\langle \mathcal{H}', \varphi \rangle = \varphi(0) = \langle \delta, \varphi \rangle$$

or $\mathcal{H}' = \delta$. Note that $\mathcal{H}' = 0$ a.e. in the pointwise sense. Thus, the derivative in the a.e. sense *does not* coincide with the derivative in the distributional sense.

Example 7.16. Let $F = \delta_n$. Then $\partial_{x_j} \delta_n$ is defined in $\mathcal{D}'(\mathbb{R}^n)$ by

$$\langle \partial_{x_j} \delta_n, \varphi \rangle = -\langle \delta_n, \partial_{x_j} \varphi \rangle = -\partial_{x_j} \varphi(\mathbf{0}), \quad \forall \varphi \in \mathcal{D}'(\mathbb{R}^n).$$

Note that $\partial_{x_j} \delta_n$ is not a measure and has support at the origin.

Another aspect of the idyllic relationship between calculus and distributions is given by the following theorem, which expresses the continuity in $\mathcal{D}'(\Omega)$ of every derivative D^α .

Proposition 7.17. *If $F_k \rightarrow F$ in $\mathcal{D}'(\Omega)$ then $D^\alpha F_k \rightarrow D^\alpha F$ in $\mathcal{D}'(\Omega)$ for any multi-index α .*

Proof. $F_k \rightarrow F$ in $\mathcal{D}'(\Omega)$ means that $\langle F_k, \varphi \rangle \rightarrow \langle F, \varphi \rangle$, $\forall \varphi \in \mathcal{D}(\Omega)$. In particular, since $D^\alpha \varphi \in \mathcal{D}(\Omega)$,

$$\langle D^\alpha F_k, \varphi \rangle = (-1)^{|\alpha|} \langle F_k, D^\alpha \varphi \rangle \rightarrow (-1)^{|\alpha|} \langle F, D^\alpha \varphi \rangle = \langle D^\alpha F, \varphi \rangle. \quad \square$$

As a consequence, if $\sum_{k=1}^{\infty} F_k = F$ in $\mathcal{D}'(\Omega)$, then

$$\sum_{k=1}^{\infty} D^\alpha F_k = D^\alpha F \quad \text{in } \mathcal{D}'(\Omega).$$

Thus, term by term differentiation is **always** permitted in $\mathcal{D}'(\Omega)$.

More difficult is the proof of the following theorem, which expresses a well known fact for functions.

Proposition 7.18. *Let Ω be a domain in \mathbb{R}^n . If $F \in \mathcal{D}'(\Omega)$ and $\partial_{x_j} F = 0$ for every $j = 1, \dots, n$, then F is a constant function.*

7.4.2 Gradient, divergence, Laplacian

There is no problem to define *vector valued distributions*. The space of test functions is $\mathcal{D}(\Omega; \mathbb{R}^n)$, i.e. the set of vectors $\varphi = (\varphi_1, \dots, \varphi_n)$ whose components belong to $\mathcal{D}(\Omega)$.

A distribution $\mathbf{F} \in \mathcal{D}'(\Omega; \mathbb{R}^n)$ is given by $\mathbf{F} = (F_1, \dots, F_n)$ with $F_j \in \mathcal{D}'(\Omega)$, $j = 1, \dots, n$. The pairing between $\mathcal{D}(\Omega; \mathbb{R}^n)$ and $\mathcal{D}'(\Omega; \mathbb{R}^n)$ is defined by

$$\langle \mathbf{F}, \varphi \rangle = \sum_{i=1}^n \langle F_i, \varphi_i \rangle. \quad (7.11)$$

- The gradient of $F \in \mathcal{D}'(\Omega)$, $\Omega \subset \mathbb{R}^n$, is simply

$$\nabla F = (\partial_{x_1} F, \partial_{x_2} F, \dots, \partial_{x_n} F).$$

Clearly $\nabla F \in \mathcal{D}'(\Omega; \mathbb{R}^n)$. If $\varphi \in \mathcal{D}(\Omega; \mathbb{R}^n)$, we have

$$\langle \nabla F, \varphi \rangle = \sum_{i=1}^n \langle \partial_{x_i} F, \varphi_i \rangle = - \sum_{i=1}^n \langle F, \partial_{x_i} \varphi_i \rangle = -\langle F, \operatorname{div} \varphi \rangle$$

whence

$$\langle \nabla F, \varphi \rangle = -\langle F, \operatorname{div} \varphi \rangle \quad (7.12)$$

which shows the action of ∇F on φ .

- For $\mathbf{F} \in \mathcal{D}'(\Omega; \mathbb{R}^n)$, we set

$$\operatorname{div}\mathbf{F} = \sum_{i=1}^n \partial_{x_i} F_i.$$

Clearly $\operatorname{div}\mathbf{F} \in \mathcal{D}'(\Omega)$. If $\varphi \in \mathcal{D}(\Omega)$, then

$$\langle \operatorname{div}\mathbf{F}, \varphi \rangle = \left\langle \sum_{i=1}^n \partial_{x_i} F_i, \varphi \right\rangle = - \sum_{i=1}^n \langle F_i, \partial_{x_i} \varphi \rangle = -\langle \mathbf{F}, \nabla \varphi \rangle$$

whence

$$\langle \operatorname{div}\mathbf{F}, \varphi \rangle = -\langle \mathbf{F}, \nabla \varphi \rangle. \quad (7.13)$$

- The Laplace operator is defined in $\mathcal{D}'(\Omega)$ by

$$\Delta F = \sum_{i=1}^n \partial_{x_i x_i} F.$$

If $\varphi \in \mathcal{D}(\Omega)$, then

$$\langle \Delta F, \varphi \rangle = \langle F, \Delta \varphi \rangle.$$

Using (7.12), (7.13) we get

$$\langle \Delta F, \varphi \rangle = \langle F, \operatorname{div}\nabla \varphi \rangle = -\langle \nabla F, \nabla \varphi \rangle = \langle \operatorname{div}\nabla F, \varphi \rangle$$

whence $\Delta = \operatorname{div}\nabla$ also in $\mathcal{D}'(\Omega)$.

Example 7.19. Consider the **fundamental solution** for the Laplace operator in \mathbb{R}^3

$$u(\mathbf{x}) = \frac{1}{4\pi} \frac{1}{|\mathbf{x}|}.$$

Observe that $u \in L^1_{loc}(\mathbb{R}^3)$ so that $u \in \mathcal{D}'(\mathbb{R}^3)$. We want to show that, in $\mathcal{D}'(\mathbb{R}^3)$,

$$-\Delta u = \delta_3. \quad (7.14)$$

First of all, if $\Omega \subset \mathbb{R}^3$ and $\mathbf{0} \notin \Omega$, we know that u is *harmonic in Ω* , that is

$$\Delta u = 0 \quad \text{in } \Omega,$$

in the classical sense and therefore also in $\mathcal{D}'(\mathbb{R}^3)$. Thus, let $\varphi \in \mathcal{D}(\mathbb{R}^3)$ with $\mathbf{0} \in \operatorname{supp}(\varphi)$. We have, since $u \in L^1_{loc}(\mathbb{R}^3)$:

$$\langle \Delta u, \varphi \rangle = \langle u, \Delta \varphi \rangle = \frac{1}{4\pi} \int_{\mathbb{R}^3} \frac{1}{|\mathbf{x}|} \Delta \varphi(\mathbf{x}) d\mathbf{x}. \quad (7.15)$$

We would like to carry the Laplacian onto $1/|\mathbf{x}|$. However, this cannot be done directly, since the integrand is not continuous at $\mathbf{0}$. Therefore we exclude a small sphere $B_r = B_r(\mathbf{0})$ from our integration region and write

$$\int_{\mathbb{R}^3} \frac{1}{|\mathbf{x}|} \Delta \varphi(\mathbf{x}) d\mathbf{x} = \lim_{r \rightarrow 0} \int_{B_R \setminus B_r} \frac{1}{|\mathbf{x}|} \Delta \varphi(\mathbf{x}) d\mathbf{x} \quad (7.16)$$

where $B_R = B_R(\mathbf{0})$ is a sphere containing the support of φ . An integration by parts in the spherical shell $C_{R,r} = B_R \setminus B_r$ yields⁷

$$\int_{C_{R,r}} \frac{1}{|\mathbf{x}|} \Delta \varphi(\mathbf{x}) d\mathbf{x} = \int_{\partial B_r} \frac{1}{r} \partial_{\nu} \varphi(\mathbf{x}) d\sigma - \int_{C_{R,r}} \nabla \left(\frac{1}{|\mathbf{x}|} \right) \cdot \nabla \varphi(\mathbf{x}) d\mathbf{x}$$

where $\nu = -\frac{\mathbf{x}}{|\mathbf{x}|}$ is the *outward* normal unit vector on ∂B_r . Integrating once more by parts the last integral, we obtain:

$$\begin{aligned} \int_{C_{R,r}} \nabla \left(\frac{1}{|\mathbf{x}|} \right) \cdot \nabla \varphi(\mathbf{x}) d\mathbf{x} &= \int_{\partial B_r} \partial_{\nu} \left(\frac{1}{|\mathbf{x}|} \right) \varphi(\mathbf{x}) d\sigma - \int_{C_{R,r}} \Delta \left(\frac{1}{|\mathbf{x}|} \right) \varphi(\mathbf{x}) d\mathbf{x} \\ &= \int_{\partial B_r} \partial_{\nu} \left(\frac{1}{|\mathbf{x}|} \right) \varphi(\mathbf{x}) d\sigma, \end{aligned}$$

since $\Delta \left(\frac{1}{|\mathbf{x}|} \right) = 0$ inside $C_{R,r}$. From the above computations we infer

$$\int_{C_{R,r}} \frac{1}{|\mathbf{x}|} \Delta \varphi(\mathbf{x}) d\mathbf{x} = \int_{\partial B_r} \frac{1}{r} \partial_{\nu} \varphi(\mathbf{x}) d\sigma - \int_{\partial B_r} \partial_{\nu} \left(\frac{1}{|\mathbf{x}|} \right) \varphi(\mathbf{x}) d\sigma. \quad (7.17)$$

We have:

$$\frac{1}{r} \left| \int_{\partial B_r} \partial_{\nu} \varphi(\mathbf{x}) d\sigma \right| \leq \frac{1}{r} \int_{\partial B_r} |\partial_{\nu} \varphi(\mathbf{x})| d\sigma \leq 4\pi r \max_{\mathbb{R}^3} |\nabla \varphi|$$

and therefore

$$\lim_{r \rightarrow 0} \int_{\partial B_r} \frac{1}{r} \partial_{\nu} \varphi(\mathbf{x}) d\sigma = 0.$$

Moreover, since

$$\partial_{\nu} \left(\frac{1}{|\mathbf{x}|} \right) = \nabla \left(\frac{1}{|\mathbf{x}|} \right) \cdot \left(-\frac{\mathbf{x}}{|\mathbf{x}|^3} \right) = \left(-\frac{\mathbf{x}}{|\mathbf{x}|^3} \right) \cdot \left(-\frac{\mathbf{x}}{|\mathbf{x}|} \right) = \frac{1}{|\mathbf{x}|^2},$$

we may write

$$\int_{\partial B_r} \partial_{\nu} \left(\frac{1}{|\mathbf{x}|} \right) \varphi(\mathbf{x}) d\sigma = 4\pi \frac{1}{4\pi r^2} \int_{\partial B_r} \varphi(\mathbf{x}) d\sigma \rightarrow 4\pi \varphi(\mathbf{0}).$$

⁷ Recall that $\varphi = 0$ and $\nabla \varphi = \mathbf{0}$ on ∂B_R .

Thus, from (7.17) we get

$$\lim_{r \rightarrow 0} \int_{B_R \setminus B_r} \frac{1}{|\mathbf{x}|} \Delta \varphi(\mathbf{x}) d\mathbf{x} = -4\pi \varphi(\mathbf{0})$$

and finally (7.15) yields

$$\langle \Delta u, \varphi \rangle = -\varphi(\mathbf{0}) = -\langle \delta_3, \varphi \rangle$$

whence $-\Delta u = \delta_3$.

7.5 Operations with Distributions

7.5.1 Multiplication. Leibniz rule

Let us analyze the multiplication between two distributions. Does it make any sense to define, for instance, the product $\delta \cdot \delta = \delta^2$ as a distribution in $\mathcal{D}'(\mathbb{R})$?

Things are not so smooth. An idea for defining δ^2 may be the following: take a sequence $\{u_k\}$ of functions in $L^1_{loc}(\mathbb{R})$ such that $u_k \rightarrow \delta$ in $\mathcal{D}'(\mathbb{R})$, compute u_k^2 and set

$$\delta^2 = \lim_{k \rightarrow \infty} u_k^2 \quad \text{in } \mathcal{D}'(\mathbb{R}).$$

Since we may approximate δ in $\mathcal{D}'(\mathbb{R})$ in many ways (see Problem 7.2), it is necessary that the definition *does not depend* on the approximating sequence. In other words, to compute δ^2 we must be free to choose any approximating sequence of δ . However, this is illusory. Indeed choose

$$u_k = k \chi_{[0, 1/k]}.$$

We have $u_k \rightarrow \delta$ in $\mathcal{D}'(\mathbb{R})$ but, if $\varphi \in \mathcal{D}(\mathbb{R})$, by the Mean Value Theorem we have

$$\int_{\mathbb{R}} u_k^2 \varphi = k^2 \int_0^{1/k} \varphi = k \varphi(x_k)$$

for some $x_k \in [0, 1/k]$. Now, if $\varphi(0) > 0$, say, we deduce that

$$\int_{\mathbb{R}} u_k^2 \varphi \rightarrow +\infty, \quad k \rightarrow +\infty$$

so that $\{u_k^2\}$ *does not converge* in $\mathcal{D}'(\mathbb{R})$.

The method does not work and it seems that there is no other reasonable way to define δ^2 in $\mathcal{D}'(\Omega)$. Thus, we simply give up defining δ^2 as a distribution or, in general, the product of two distributions. However, if $F \in \mathcal{D}'(\Omega)$ and $u \in C^\infty(\Omega)$, we may define the product uF by the formula

$$\langle uF, \varphi \rangle = \langle F, u\varphi \rangle, \quad \forall \varphi \in \mathcal{D}(\Omega).$$

First of all, this makes sense since $u\varphi \in \mathcal{D}(\Omega)$. Also, if $\varphi_k \rightarrow \varphi$ in $\mathcal{D}(\Omega)$, then $u\varphi_k \rightarrow u\varphi$ in $\mathcal{D}(\Omega)$ and

$$\langle uF, \varphi_k \rangle = \langle F, u\varphi_k \rangle \rightarrow \langle F, u\varphi \rangle = \langle uF, \varphi \rangle,$$

so that uF is a well defined element of $\mathcal{D}'(\Omega)$.

Example 7.20. Let $u \in C^\infty(\mathbb{R})$. We have

$$u\delta = u(0)\delta.$$

Indeed, if $\varphi \in \mathcal{D}(\mathbb{R})$,

$$\langle u\delta, \varphi \rangle = \langle \delta, u\varphi \rangle = u(0)\varphi(0) = \langle u(0)\delta, \varphi \rangle.$$

Note that the product $u\delta$ makes sense even if u is only continuous. In particular

$$x\delta = 0.$$

The *Leibniz* rule holds: let $F \in \mathcal{D}'(\Omega)$ and $u \in C^\infty(\Omega)$; then

$$\partial_{x_i}(uF) = u \partial_{x_i}F + \partial_{x_i}u F. \quad (7.18)$$

In fact, let $\varphi \in \mathcal{D}(\Omega)$; we have:

$$\langle \partial_{x_i}(uF), \varphi \rangle = -\langle uF, \partial_{x_i}\varphi \rangle = -\langle F, u\partial_{x_i}\varphi \rangle$$

while

$$\begin{aligned} \langle u \partial_{x_i}F + \partial_{x_i}u F, \varphi \rangle &= \langle \partial_{x_i}F, u\varphi \rangle + \langle F, \varphi \partial_{x_i}u \rangle \\ &= -\langle F, \partial_{x_i}(u\varphi) \rangle + \langle F, \varphi \partial_{x_i}u \rangle = -\langle F, u\partial_{x_i}\varphi \rangle \end{aligned}$$

and (7.18) follows.

Example 7.21. From $x\delta = 0$ and Leibniz rule we obtain

$$\delta + x\delta' = 0.$$

More generally,

$$x^m \delta^{(k)} = 0 \quad \text{in } \mathcal{D}'(\mathbb{R}), \quad \text{if } 0 \leq k < m.$$

7.5.2 Composition

Composition in $\mathcal{D}'(\mathbb{R})$ requires caution as well. For instance, if $F = \delta$ and $v(x) = x^3$, is there a natural way to define $F \circ v$ as a distribution in $\mathcal{D}'(\mathbb{R})$?

As above, consider the sequence $u_k = k\chi_{[0,1/k]}$ and compute $w_k = u_k \circ v$. If $\varphi \in \mathcal{D}(\mathbb{R})$, we have

$$\int_{\mathbb{R}} w_k \varphi = k \int_{\mathbb{R}} \chi_{[0,1/k]}(x^3) \varphi(x) dx = k \int_0^{k^{-1/3}} \varphi(x) dx = k^{2/3} \varphi(x_k)$$

for some $x_k \in [0, 1/k]$. Then, if $\varphi(0) > 0$, $\int_{\mathbb{R}} w_k \varphi \rightarrow +\infty$ and $F \circ v$ does not make any sense. Thus, it seems hard to define the composition between two general distributions. To see what can be done, let us start analyzing the case of two functions.

Let $\Omega', \Omega \subseteq \mathbb{R}^n$ and $\psi : \Omega' \rightarrow \Omega$ be **one to one**, with ψ and ψ^{-1} of class C^∞ . If $F : \Omega \rightarrow \mathbb{R}$ is a C^1 -function we may consider the composition

$$w(\mathbf{x}) = F \circ \psi(\mathbf{x}) = F(\psi(\mathbf{x})).$$

For $\varphi \in \mathcal{D}(\Omega')$, we have, using the change of variables $\mathbf{y} = \psi(\mathbf{x})$:

$$\int_{\Omega'} F(\psi(\mathbf{x})) \varphi(\mathbf{x}) d\mathbf{x} = \int_{\Omega} F(\mathbf{y}) \varphi(\psi^{-1}(\mathbf{y})) |\det J_{\psi^{-1}}(\mathbf{y})| d\mathbf{y} \quad (7.19)$$

where $J_{\psi^{-1}}$ denotes the Jacobian of the transformation ψ^{-1} . Observe that

$$\varphi \circ \psi^{-1} \cdot |\det J_{\psi^{-1}}| \in \mathcal{D}(\Omega'),$$

with support given by $K = \psi(K')$, where K' is the support of φ . Thus, in terms of distributions, (7.19) writes

$$\langle F \circ \psi, \varphi \rangle_{\mathcal{D}'(\Omega')} = \langle F, \varphi \circ \psi^{-1} \cdot |\det J_{\psi^{-1}}| \rangle_{\mathcal{D}'(\Omega)}. \quad (7.20)$$

This formula makes sense also if $F \in \mathcal{D}'(\Omega)$ and it is not difficult to check that if $\varphi_k \rightarrow 0$ in $\mathcal{D}(\Omega')$ then $\varphi_k \circ \psi^{-1} \cdot |\det J_{\psi^{-1}}| \rightarrow 0$ in $\mathcal{D}(\Omega)$. Thus, (7.20) defines a distribution in $\mathcal{D}'(\Omega')$. Precisely:

Definition 7.22. If $F \in \mathcal{D}'(\Omega)$ and $\psi : \Omega' \rightarrow \Omega$ is one to one, with ψ and ψ^{-1} of class C^∞ , then formula (7.20) defines the composition $F \circ \psi$ as an element of $\mathcal{D}'(\Omega')$.

Let us check that (7.20) behaves well with respect to convergence in $\mathcal{D}'(\Omega')$.

Proposition 7.23. Let ψ be as in Definition 7.22. If $F_k \rightarrow F$ in $\mathcal{D}'(\Omega)$, then $F_k \circ \psi \rightarrow F \circ \psi$ in $\mathcal{D}'(\Omega')$.

Proof. Let $\varphi \in \mathcal{D}(\Omega')$. Then

$$\begin{aligned} \lim_{k \rightarrow \infty} \langle F_k \circ \psi, \varphi \rangle_{\mathcal{D}'(\Omega')} &= \lim_{k \rightarrow \infty} \langle F_k, \varphi \circ \psi^{-1} \cdot |\det J_{\psi^{-1}}| \rangle_{\mathcal{D}'(\Omega)} \\ &= \langle F, \varphi \circ \psi^{-1} \cdot |\det J_{\psi^{-1}}| \rangle_{\mathcal{D}'(\Omega)} \\ &= \langle F \circ \psi, \varphi \rangle_{\mathcal{D}'(\Omega')}. \end{aligned}$$

□

Abuses of notation are quite common, like $F(\psi(\mathbf{x}))$ to denote $F \circ \psi$. For instance, we have repeatedly used the (comfortable and incorrect) notation $\delta(\mathbf{x} - \mathbf{x}_0)$ instead of the (uncomfortable and correct) notation $\delta \circ \psi$, with $\psi(\mathbf{x}) = \mathbf{x} - \mathbf{x}_0$.

Example 7.24. In $\mathcal{D}'(\mathbb{R}^n)$, we have

$$\delta_n(a\mathbf{x}) = \frac{1}{|a|^n} \delta_n(\mathbf{x}).$$

Using formula (7.20) we may extend to distributions some properties, typical of functions. We list some of them.

We say that $F \in \mathcal{D}'(\mathbb{R}^n)$ is:

- *Radial*, if

$$F(\mathbf{Ax}) = F(\mathbf{x}), \quad \text{for every orthogonal matrix } \mathbf{A}.$$

- *Homogeneous of degree λ* , if

$$F(t\mathbf{x}) = t^\lambda F(\mathbf{x}), \quad \forall t > 0.$$

- *Even, odd* if, respectively,

$$F(-\mathbf{x}) = F(\mathbf{x}), \quad F(-\mathbf{x}) = -F(\mathbf{x}).$$

- *Periodic with period \mathbf{P}* , if

$$F(\mathbf{x} + \mathbf{P}) = F(\mathbf{x}).$$

Example 7.25.

a. $\delta_n \in \mathcal{D}'(\mathbb{R}^n)$ is radial, even and homogeneous of degree $\lambda = -n$.

b. $p.v.\frac{1}{x} \in \mathcal{D}'(\mathbb{R})$ is odd and homogeneous of degree $\lambda = -1$.

c. In $\mathcal{D}'(\mathbb{R})$, *comb* is periodic with period 1.

Example 7.26. Let $f : \mathbb{R} \rightarrow \mathbb{R}$ be one-to-one, with f and f^{-1} of class C^∞ and $f(x_0) = 0$. Then necessarily $f'(x_0) \neq 0$ and the following formula holds in $\mathcal{D}'(\mathbb{R})$:

$$\delta(f(x)) = \frac{\delta(x - x_0)}{|f'(x_0)|}. \tag{7.21}$$

In fact, let $\varphi \in \mathcal{D}(\mathbb{R})$. Setting $g = f^{-1}$, we have, since $\det J_g(y) = g'(y) = [f'(g(y))]^{-1}$:

$$\langle \delta \circ f, \varphi \rangle = \langle \delta, \varphi \circ g \cdot |\det J_g| \rangle = \langle \delta, \frac{\varphi \circ g}{|f' \circ g|} \rangle = \frac{\varphi(g(0))}{|f'(g(0))|} = \frac{\varphi(x_0)}{|f'(x_0)|}$$

and

$$\left\langle \frac{\delta(x - x_0)}{|f'(x_0)|}, \varphi \right\rangle = \frac{1}{|f'(x_0)|} \langle \delta, \varphi(x + x_0) \rangle = \frac{\varphi(x_0)}{|f'(x_0)|}.$$

Comparing the two formulas we get (7.21).

Example 7.27. Let $r > 0$. In $\mathcal{D}'(\mathbb{R}^3)$, we have:

$$\langle \delta(|\mathbf{x}| - r), \varphi(\mathbf{x}) \rangle = r^2 \int_{\partial B_1} \varphi(r\boldsymbol{\omega}) d\omega = \int_{\partial B_r} \varphi(\boldsymbol{\sigma}) d\sigma,$$

for every $\varphi \in \mathcal{D}'(\mathbb{R}^3)$. Indeed, the formula corresponds to the following formal change to spherical coordinates:

$$\int_{\mathbb{R}^3} \delta(|\mathbf{x}| - r) \varphi(\mathbf{x}) d\mathbf{x} = \int_0^{+\infty} \delta(\rho - r) \int_{\partial B_1} \varphi(\rho\boldsymbol{\sigma}) \rho^2 d\sigma d\rho = r^2 \int_{\partial B_1} \varphi(r\boldsymbol{\omega}) d\omega.$$

- *Restriction of a distribution.* It is possible to define the restriction of a distribution $F \in \mathcal{D}'(\Omega)$ to an open set $\Omega' \subset \Omega$ by simply defining

$$\langle F|_{\Omega'}, \varphi \rangle_{\mathcal{D}'(\Omega')} = \langle F, \tilde{\varphi} \rangle_{\mathcal{D}'(\Omega)} \quad (7.22)$$

where $\varphi \in \mathcal{D}(\Omega')$ and $\tilde{\varphi} \in \mathcal{D}(\Omega)$ is the extension to zero of φ outside Ω' . Clearly, (7.22) defines $F|_{\Omega'}$ as a distribution in $\mathcal{D}'(\Omega')$, since if $\varphi_k \rightarrow 0$ in $\mathcal{D}(\Omega')$ then $\tilde{\varphi}_k \rightarrow 0$ in $\mathcal{D}(\Omega)$. Also;

$$F_k \rightarrow F \text{ in } \mathcal{D}'(\Omega) \text{ implies } F_k|_{\Omega'} \rightarrow F|_{\Omega'} \text{ in } \mathcal{D}'(\Omega')$$

as it is easy to check.

Example 7.28. Let $\Omega \subset \mathbb{R}^n$ be an open set. If $\mathbf{0} \in \Omega$ then $\langle \delta|_{\Omega}, \varphi \rangle = \varphi(\mathbf{0})$ for all $\varphi \in \mathcal{D}(\Omega)$. If $\mathbf{0} \notin \Omega$ then $\langle \delta|_{\Omega}, \varphi \rangle = 0$ for all $\varphi \in \mathcal{D}'(\Omega)$ and therefore $\delta|_{\Omega}$ is the zero function.

We can generalize the formula (7.21) to smooth functions f having (finite or infinite) simple zeros without cluster points. First, let $f \in C^\infty(\mathbb{R})$ and define $\delta \circ f \in \mathcal{D}'(\mathbb{R})$ by the formula

$$\langle \delta \circ f, \varphi \rangle = \lim_{\varepsilon \rightarrow 0} \langle I_\varepsilon(f), \varphi \rangle \quad (7.23)$$

where $I_\varepsilon(t) = 1/\varepsilon$ for $|t| < \varepsilon/2$ and $I_\varepsilon(t) = 0$ otherwise, provided the limit exists for every test function $\varphi \in \mathcal{D}(\mathbb{R})$.

Proposition 7.29. Assume that f vanishes only at the points x_j , with $f'(x_j) \neq 0$, $j = 1, 2, \dots$. Then the limit in (7.23) exists and

$$\delta(f(x)) = \sum_{j \geq 1} \frac{\delta(x - x_j)}{|f'(x_j)|} \quad \text{in } \mathcal{D}'(\mathbb{R}). \quad (7.24)$$

Proof. Let $\varphi \in \mathcal{D}(\mathbb{R})$ with $\text{supp}(\varphi) = [a, b]$. Only a finite number of zeros of f belong to $[a, b]$, say x_1, \dots, x_N . Let ε so small that $|f(x)| < \varepsilon/2$ and $f'(x) \neq 0$ on the union of N disjoint intervals $I_j = (x_j - \eta, x_j + \eta)$, $j = 1, \dots, N$. Denote by f_j the restriction of f to I_j and let $g_j = f_j^{-1}$. We have :

$$\langle I_\varepsilon(f(x)), \varphi \rangle = \sum_{j=1}^N \int_{I_j} I_\varepsilon(f_j(x)) \varphi(x) dx \underset{y=f_j(x)}{=} \sum_{j=1}^N \frac{1}{\varepsilon} \int_{-\varepsilon/2}^{\varepsilon/2} \varphi(g_j(y)) |f'_j(g_j(y))|^{-1} dy.$$

Letting $\varepsilon \rightarrow 0$, we find, since $g_j(0) = x_j$:

$$\lim_{\varepsilon \rightarrow 0} \langle I_\varepsilon(f(x)), \varphi \rangle = \sum_{j=1}^N \varphi(g_j(0)) |f'_j(g_j(0))|^{-1} = \sum_{j=1}^N \frac{\varphi(x_j)}{|f'_j(x_j)|}$$

which in terms of distributions means (7.24). □

Example 7.30. Let $f(x) = x^2 - a^2$ ($a > 0$). Then

$$\delta(f(x)) = \frac{\delta(x+a) + \delta(x-a)}{2a}.$$

7.5.3 Division

The division in $\mathcal{D}'(\Omega)$ is rather delicate, even restricting to $F \in \mathcal{D}'(\Omega)$ and $u \in C^\infty(\Omega)$. To divide F by u means to find $G \in \mathcal{D}'(\Omega)$ such that $uG = F$. If u never vanishes there is no problem, since in this case $1/u \in C^\infty(\Omega)$ and the answer is simply

$$G = \frac{1}{u}F.$$

If u vanishes somewhere, things get complicated. We only consider a particular case in one dimension.

Let $I \subseteq \mathbb{R}$ be an **open interval** and $u \in C^\infty(I)$. If u vanishes at z , we say that z is a *zero of order $m(z)$* if the derivatives of u up to order $m(z) - 1$, included, vanish at z , while the derivative of order $m(z)$ does not vanish at z . For instance, $z = 0$ is a zero of order 3 for $u(x) = \sin x - x$.

One can prove the following proposition.

Proposition 7.31. Assume that u vanishes at isolated points z_1, z_2, \dots with order $m(z_1), m(z_2), \dots$. Then, the equation

$$uG = 0$$

has infinitely many solutions in $\mathcal{D}'(I)$, given by the following formula:

$$G = \sum_j \sum_{k=0}^{m(z_j)-1} c_{jk} \delta^{(k)}(x - z_j) \quad (7.25)$$

where c_{jk} are arbitrary constants and $\delta^{(k)}$ is the derivative of δ of order k .

Example 7.32. The solutions in $\mathcal{D}'(\mathbb{R})$ of the equation

$$xG = 0$$

are the distributions of the form $G = c\delta$, with $c \in \mathbb{R}$. To solve the equation

$$xG = 1, \quad (7.26)$$

we add to the solutions of the homogeneous equation $xG = 0$ a particular solution of (7.26). It turns out that one of these is

$$G_1 = p.v. \frac{1}{x}.$$

In fact, if $\varphi \in \mathcal{D}(\mathbb{R})$, from Example 7.12, p. 436, we get

$$\begin{aligned} \langle x \cdot (p.v. \frac{1}{x}), \varphi \rangle &= \langle p.v. \frac{1}{x}, x\varphi \rangle = \\ &= p.v. \int_{\mathbb{R}} \frac{x\varphi(x)}{x} dx = \int_{\mathbb{R}} \varphi(x) dx = \langle 1, \varphi \rangle \end{aligned}$$

whence

$$x \cdot p.v. \left(\frac{1}{x} \right) = 1.$$

Therefore, the general solution of (7.26) is

$$G = p.v. \frac{1}{x} + c\delta, \quad c \in \mathbb{R}. \quad (7.27)$$

7.5.4 Convolution

The convolution of two distributions may be defined with some restrictions as well. Let us see why. If $u, w \in L^1(\mathbb{R}^n)$ and $\varphi \in \mathcal{D}(\mathbb{R}^n)$ we may write:

$$\begin{aligned} \langle u * w, \varphi \rangle &= \int_{\mathbb{R}^n} \left[\int_{\mathbb{R}^n} u(\mathbf{x} - \mathbf{y}) w(\mathbf{y}) d\mathbf{y} \right] \varphi(\mathbf{x}) d\mathbf{x} \\ &= \int_{\mathbb{R}^n} \int_{\mathbb{R}^n} u(\mathbf{x}) w(\mathbf{y}) \varphi(\mathbf{x} + \mathbf{y}) d\mathbf{y} d\mathbf{x}. \end{aligned}$$

Now, the question is: may we give any meaning to this formula if u and v are generic distributions? The answer is negative, mainly because the function

$$\phi(\mathbf{x}, \mathbf{y}) = \varphi(\mathbf{x} + \mathbf{y})$$

does **not** have compact support⁸ in $\mathbb{R}^n \times \mathbb{R}^n$ (unless $\varphi \equiv 0$).

However, a modification of the above formula would give the possibility to define the convolution between two distributions, if *at least one of them has compact support*. Here we limit ourselves to define the convolution between a distribution T and a C^∞ -function u .

If $T \in L^1(\mathbb{R}^n)$, with compact support, then the usual definition of convolution is

$$(T * u)(\mathbf{x}) = \int_{\mathbb{R}^n} T(\mathbf{y}) u(\mathbf{x} - \mathbf{y}) d\mathbf{y} = \langle T, u(\mathbf{x} - \cdot) \rangle. \quad (7.28)$$

Since

$$\mathbf{y} \longmapsto u(\mathbf{x} - \mathbf{y})$$

is a C^∞ -function, recalling Remark 7.13, p. 437, the last bracket in (7.28) makes sense if T is a distribution with compact support as well. Precisely, we have:

Proposition 7.33. *Let $T \in \mathcal{D}'(\mathbb{R}^n)$, with compact support, and $u \in C^\infty(\mathbb{R}^n)$. Then, the following formula*

$$(T * u)(\mathbf{x}) = \langle T, u(\mathbf{x} - \cdot) \rangle \quad (7.29)$$

defines a C^∞ -function, called **convolution** of T and u .

Example 7.34. Let $u \in C^\infty(\mathbb{R}^n)$. Then

$$(\delta * u)(\mathbf{x}) = \langle \delta, u(\mathbf{x} - \cdot) \rangle = u(\mathbf{x})$$

i.e

$$\delta * u = u. \quad (7.30)$$

Thus, the Dirac distribution at zero, acts as the **identity** with respect to the convolution. Formula (7.30) actually holds for all $u \in \mathcal{D}'(\mathbb{R}^n)$. In particular:

$$\delta * \delta = \delta.$$

Proposition 7.35. *The convolution commutes with derivatives. Actually, we have:*

$$\partial_{x_j}(T * u) = \partial_{x_j} T * u = T * \partial_{x_j} u.$$

⁸ For instance: if $\varphi \in \mathcal{D}'(\mathbb{R})$ and $\text{supp}(\varphi) = [a, b]$, then the support of $\varphi(x + y)$ in \mathbb{R}^2 is the unbounded strip $a \leq x + y \leq b$.

In particular, if $T = \mathcal{H}$ and $u \in \mathcal{D}(\mathbb{R})$,

$$(\mathcal{H} * u)' = (\mathcal{H}' * u) = \delta * u = u.$$

Warning: The convolution of functions is associative. For distributions, the convolution is, in general, not *associative*. In fact, consider the three distributions $1, \delta', \mathcal{H}$; we have (formally):

$$\delta' * 1 = (\delta * 1)' = 1' = 0$$

whence

$$\mathcal{H} * (\delta' * 1) = \mathcal{H} * 0 = 0.$$

However,

$$(\mathcal{H} * \delta') * 1 = (\mathcal{H}' * \delta) * 1 = (\delta * \delta) * 1 = 1.$$

The problem is that two out of three factors (1 and \mathcal{H}) have noncompact support. If at least *two factors have compact support* one can show that the convolution is associative.

7.5.5 Tensor or direct product

The *tensor* or *direct* product of two distributions F, G , denoted by $F \otimes G$, is an (important) operation that enables us to construct a distribution in $\mathcal{D}'(\mathbb{R}^m \times \mathbb{R}^n)$ starting from two distributions in $\mathcal{D}'(\mathbb{R}^m)$ and $\mathcal{D}'(\mathbb{R}^n)$, respectively. To avoid confusions, we call \mathbf{x} the independent variable in \mathbb{R}^m and \mathbf{y} the one in \mathbb{R}^n . If we were dealing with two functions $u \in L^1_{loc}(\mathbb{R}^m)$ and $v \in L^1_{loc}(\mathbb{R}^n)$, the tensor product is defined simply by the formula

$$w(\mathbf{x}, \mathbf{y}) = (u \otimes v)(\mathbf{x}, \mathbf{y}) = u(\mathbf{x})v(\mathbf{y}).$$

As a distribution, the action of w on a test function $\varphi \in \mathcal{D}(\mathbb{R}^m \times \mathbb{R}^n)$ is given by (using Fubini's Theorem)

$$\begin{aligned} & \int_{\mathbb{R}^m \times \mathbb{R}^n} w(\mathbf{x}, \mathbf{y}) \varphi(\mathbf{x}, \mathbf{y}) d\mathbf{x} d\mathbf{y} = \\ & \int_{\mathbb{R}^m} u(\mathbf{x}) \left(\int_{\mathbb{R}^n} v(\mathbf{y}) \varphi(\mathbf{x}, \mathbf{y}) d\mathbf{y} \right) d\mathbf{x} = \int_{\mathbb{R}^n} v(\mathbf{y}) \left(\int_{\mathbb{R}^m} u(\mathbf{x}) \varphi(\mathbf{x}, \mathbf{y}) d\mathbf{x} \right) d\mathbf{y}. \end{aligned}$$

This formula can be extended to distributions. Indeed, let $F \in \mathcal{D}'(\mathbb{R}^m)$, $G \in \mathcal{D}'(\mathbb{R}^n)$ and $\varphi \in \mathcal{D}(\mathbb{R}^m \times \mathbb{R}^n)$. Then the two functions

$$\psi_{(1)}(\mathbf{x}) = \langle G, \varphi(\mathbf{x}, \cdot) \rangle \quad \text{and} \quad \psi_{(2)}(\mathbf{y}) = \langle F, \varphi(\cdot, \mathbf{y}) \rangle$$

are test functions in $\mathcal{D}(\mathbb{R}^m)$ and $\mathcal{D}(\mathbb{R}^n)$, respectively. The following Theorem holds⁹.

⁹ See [39], Yoshida, 1971.

Theorem 7.36. Given $F \in \mathcal{D}'(\mathbb{R}^m)$ and $G \in \mathcal{D}'(\mathbb{R}^n)$, there exists a unique distribution $W \in \mathcal{D}'(\mathbb{R}^m \times \mathbb{R}^n)$ such that, $\forall \varphi \in \mathcal{D}(\mathbb{R}^m \times \mathbb{R}^n)$:

$$\langle W, \varphi \rangle = \langle F, \psi_{(1)} \rangle = \langle G, \psi_{(2)} \rangle. \quad (7.31)$$

In particular, $\forall \varphi_1 \in \mathcal{D}(\mathbb{R}^m)$, $\forall \varphi_2 \in \mathcal{D}(\mathbb{R}^n)$,

$$\langle W, \varphi_1 \varphi_2 \rangle = \langle F, \varphi_1 \rangle \langle G, \varphi_2 \rangle.$$

The distribution defined in Theorem 7.36 is called the *tensor or direct product of the distributions* F and G , denoted by $F \otimes G$, as we have already mentioned.

There is no difficulty in defining the tensor product of any number k of distributions. It turns out that this product is *associative* and we can write without ambiguity

$$F_1 \otimes F_2 \otimes \cdots \otimes F_k.$$

Example 7.37. Let us check that $\delta_3(\mathbf{x}) = \delta(x_1) \otimes \delta(x_2) \otimes \delta(x_3)$. In fact, let $\varphi \in \mathcal{D}(\mathbb{R}^3)$. We have, on one side, $\langle \delta_3, \varphi \rangle = \varphi(\mathbf{0})$, on the other side

$$\begin{aligned} \langle \delta(x_1) \otimes \delta(x_2) \otimes \delta(x_3), \varphi(x_1, x_2, x_3) \rangle &= \langle \delta(x_1) \otimes \delta(x_2), \varphi(x_1, x_2, 0) \rangle \\ &= \langle \delta(x_1), \varphi(x_1, 0, 0) \rangle = \varphi(\mathbf{0}). \end{aligned}$$

Example 7.38. Let $g \in L^1_{loc}(\mathbb{R})$. Let us compute $g(x) \otimes \delta'(y)$. Let $\varphi \in \mathcal{D}(\mathbb{R} \times \mathbb{R})$. We have:

$$\begin{aligned} \langle g(x) \otimes \delta'(y), \varphi(x, y) \rangle &= \langle g(x), \langle \delta'(y), \varphi(x, y) \rangle \rangle = -\langle g(x), \langle \delta(y), \varphi_y(x, y) \rangle \rangle \\ &= -\langle g(x), \varphi_y(x, 0) \rangle = -\int_{\mathbb{R}} g(x) \varphi_y(x, 0) dx. \end{aligned}$$

Very often, in applied contexts, the symbol \otimes is omitted, as we actually did in Chap. 5, in Sect. 5.10 and Subsects. 5.11.1 and 5.11.2.

Example 7.39. In this example we go back to Subsect. 5.11.2 and prove that the potentials in formulas (5.136) and (5.138), pp. 323, 324, are solutions in the sense of distributions of the moving source equation, written with the proper symbology:

$$u_{tt} - c^2 \Delta u = S \delta(x_1) \otimes \delta(x_2) \otimes \delta(x_3 - vt) \quad \text{in } \mathcal{D}'(\mathbb{R}^3 \times \mathbb{R}), \quad (7.32)$$

in the subsonic and supersonic case, respectively. We put $\mathbf{x}' = (x_1, x_2)$.

- The **subsonic case** $m = v/c < 1$. The corresponding potential is given by

$$u(\mathbf{x}', x_3, t) = \frac{S}{4\pi c^2} \frac{1}{\sqrt{|x'|^2 (1 - m^2) + (vt - x_3)^2}}. \quad (7.33)$$

Since u is a travelling wave along the x_3 -axis, it is convenient to introduce the new coordinates $\xi' = \sqrt{1 - m^2} \mathbf{x}'$, $\xi_3 = vt - x_3$ and look at the function

$$U(\xi', \xi_3) = \frac{S}{4\pi c^2} \frac{1}{\sqrt{|\xi'|^2 + \xi_3^2}} = \frac{S}{4\pi c^2} \frac{1}{|\xi'|}.$$

Then $u(\mathbf{x}', x_3, t) = U(\sqrt{1 - m^2} \mathbf{x}', vt - x_3)$ and

$$u_{tt} = v^2 U_{\xi_3 \xi_3}, \quad u_{x_j x_j} = (1 - m^2) U_{\xi_j \xi_j}, \quad j = 1, 2, \quad u_{x_3 x_3} = U_{\xi_3 \xi_3}.$$

Inserting into (7.32), we get

$$(v^2 - c^2) U_{\xi_3 \xi_3} - c^2 (1 - m^2) \Delta_2 U = S \delta_2(\xi'/\sqrt{1 - m^2}) \otimes \delta(\xi_3), \quad \text{in } \mathcal{D}'(\mathbb{R}^3).$$

Dividing by c^2 , simplifying by $(1 - m^2)$ and recalling that

$$\delta_2(\xi'/\sqrt{1 - m^2}) = (1 - m^2) \delta_2(\xi'),$$

we conclude that u is a solution of (7.32) if and only if U satisfies the equation:

$$\Delta_3 U = -\frac{S}{c^2} \delta_2(\xi') \otimes \delta(\xi_3) = -\frac{S}{c^2} \delta_3(\xi), \quad \text{in } \mathcal{D}'(\mathbb{R}^3). \quad (7.34)$$

From Example 7.19, p. 441, we know that $-\frac{1}{4\pi|\xi|}$ is the fundamental solution of Δ_3 . Thus (7.34) follows and the subsonic case is completed.

- The **supersonic case** $m = v/c > 1$. The corresponding potential is given by

$$u(\mathbf{x}, t) = \begin{cases} \frac{S}{2\pi c^2} \frac{1}{\sqrt{(vt - x_3)^2 - |\mathbf{x}'|^2 (m^2 - 1)}} & vt - x_3 > |\mathbf{x}'| \sqrt{(m^2 - 1)} \\ 0 & vt - x_3 < |\mathbf{x}'| \sqrt{(m^2 - 1)}. \end{cases}$$

In this case, we introduce the coordinates $\xi' = \sqrt{m^2 - 1} \mathbf{x}'$, $\xi_3 = vt - x_3$ and look at the function

$$U(\xi', \xi_3) = \frac{S}{2\pi c^2} \frac{1}{\sqrt{\xi_3^2 - |\xi'|^2}}, \quad \text{for } \xi_3 > |\xi'|,$$

and $U = 0$ for $\xi_3 < |\xi'|$. Then $u(\mathbf{x}', x_3, t) = U(\sqrt{m^2 - 1} \mathbf{x}', vt - x_3)$ and

$$u_{tt} = v^2 U_{\xi_3 \xi_3}, \quad u_{x_j x_j} = (m^2 - 1) U_{\xi_j \xi_j}, \quad j = 1, 2, \quad u_{x_3 x_3} = U_{\xi_3 \xi_3}.$$

Inserting into (7.32), we get

$$(v^2 - c^2) U_{\xi_3 \xi_3} - c^2 (m^2 - 1) \Delta_2 U = S (m^2 - 1) \delta_2(\xi') \otimes \delta(\xi_3) \quad \text{in } \mathcal{D}'(\mathbb{R}^3).$$

Dividing by c^2 and simplifying by $(m^2 - 1)$, we conclude that u is a solution of (7.32) if and only if U satisfies the equation:

$$U_{\xi_3 \xi_3} - \Delta_2 U = \frac{S}{c^2} \delta_2(\xi') \otimes \delta(\xi_3) = \frac{S}{c^2} \delta_3(\xi) \quad \text{in } \mathcal{D}'(\mathbb{R}^3). \quad (7.35)$$

Recalling Remark 5.18, p. 317, we recognize that $(c^2/S)U$ is the fundamental solution of the two-dimensional wave equation (with $t = \xi_3, c = 1$) (see Problem 7.18) and therefore (7.35) is true. This completes the supersonic case.

7.6 Tempered Distributions and Fourier Transform

7.6.1 Tempered distributions

We now introduce the Fourier transform \hat{F} of a distribution. As usual, the idea is to define the action of \hat{F} by carrying the transform onto the test functions. However a problem immediately arises: if $\varphi \in \mathcal{D}(\mathbb{R}^n)$ is not identically zero, then

$$\widehat{\varphi}(\xi) = \int_{\mathbb{R}^n} e^{-i\mathbf{x} \cdot \xi} \varphi(\mathbf{x}) d\mathbf{x}$$

cannot belong¹⁰ to $\mathcal{D}(\mathbb{R}^n)$. Thus, it is necessary to choose a larger space of test functions. It turns out that the correct one consists in the set of functions *rapidly vanishing at ∞* , which obviously contains $\mathcal{D}(\mathbb{R}^n)$. It is convenient to consider *functions and distributions with complex values*.

Definition 7.40. Denote by $\mathcal{S}(\mathbb{R}^n)$ the space of functions $v \in C^\infty(\mathbb{R}^n)$ rapidly vanishing at infinity, i.e. such that

$$D^\alpha v(\mathbf{x}) = o(|\mathbf{x}|^{-m}), \quad |\mathbf{x}| \rightarrow \infty,$$

for all $m \in \mathbb{N}$ and every multi-index α .

¹⁰ Let $n = 1$ and $\varphi \in \mathcal{D}(\mathbb{R})$. Assume that

$$\text{supp}(\varphi) \subset (-a, a).$$

We may write

$$\widehat{\varphi}(\xi) = \int_{-a}^a e^{-ix\xi} \varphi(x) dx = \int_{-a}^a \sum_{n=0}^{\infty} \frac{(-ix\xi)^n}{n!} \varphi(x) dx = \sum_{n=0}^{\infty} \frac{(-i\xi)^n}{n!} \int_{-a}^a x^n \varphi(x) dx.$$

Since

$$\left| \int_{-a}^a x^n \varphi(x) dx \right| \leq \max |\varphi| a^n,$$

it follows that $\widehat{\varphi}$ is an *analytic* function in all \mathbb{C} . Therefore $\widehat{\varphi}$ cannot vanish outside a compact interval, unless $\widehat{\varphi} \equiv 0$. But then $\varphi \equiv 0$ as well.

Example 7.41. The function $v(\mathbf{x}) = e^{-|\mathbf{x}|^2}$ belongs to $\mathcal{S}(\mathbb{R}^n)$ while $v(\mathbf{x}) = e^{-|\mathbf{x}|^2} \sin(e^{|\mathbf{x}|^2})$ does not (why?).

We endow $\mathcal{S}(\mathbb{R}^n)$ with an “ad hoc” notion of convergence. If $\beta = (\beta_1, \dots, \beta_n)$ is a multi-index, we set

$$\mathbf{x}^\beta = x_1^{\beta_1} \cdots x_n^{\beta_n}.$$

Definition 7.42. Let $\{v_k\} \subset \mathcal{S}(\mathbb{R}^n)$ and $v \in \mathcal{S}(\mathbb{R}^n)$. We say that

$$v_k \rightarrow v \quad \text{in } \mathcal{S}(\mathbb{R}^n)$$

if for every pair of multi-indexes α, β ,

$$\mathbf{x}^\beta D^\alpha v_k \rightarrow \mathbf{x}^\beta D^\alpha v, \quad \text{uniformly in } \mathbb{R}^n.$$

Remark 7.43. If $\{v_k\} \subset \mathcal{D}(\mathbb{R}^n)$ and $v_k \rightarrow v$ in $\mathcal{D}(\mathbb{R}^n)$, then

$$v_k \rightarrow v \quad \text{in } \mathcal{S}(\mathbb{R}^n)$$

as well, since each v_k vanishes outside a common compact set so that the multiplication by \mathbf{x}^β does not have any influence.

The Fourier transform will be defined for distributions in $\mathcal{D}'(\mathbb{R}^n)$, continuous with respect to the convergence in Definition 7.42. These are the so called *tempered distributions*. Precisely:

Definition 7.44. We say that $T \in \mathcal{D}'(\mathbb{R}^n)$ is a tempered distribution if

$$\langle T, v_k \rangle \rightarrow 0$$

for all sequences $\{v_k\} \subset \mathcal{D}(\mathbb{R}^n)$ such that $v_k \rightarrow 0$ in $\mathcal{S}(\mathbb{R}^n)$. The set of tempered distributions is denoted by $\mathcal{S}'(\mathbb{R}^n)$.

So far, a tempered distribution T is only defined on $\mathcal{D}(\mathbb{R}^n)$. To define T on $\mathcal{S}(\mathbb{R}^n)$, first we observe that $\mathcal{D}(\mathbb{R}^n)$ is dense in $\mathcal{S}(\mathbb{R}^n)$.

In fact, given $v \in \mathcal{S}(\mathbb{R}^n)$, let

$$v_k(\mathbf{x}) = v(\mathbf{x}) \rho(|\mathbf{x}|/k)$$

where $\rho = \rho(s)$, $s \geq 0$, is a non-negative C^∞ -function, equal to 1 in $[0, 1]$ and zero for $s \geq 2$ (see Fig. 7.4). It can be checked that $\{v_k\} \subset \mathcal{D}(\mathbb{R}^n)$ and $v_k \rightarrow v$ in $\mathcal{S}(\mathbb{R}^n)$, since $\rho(|\mathbf{x}|/k)$ is equal to 1 for $\{|\mathbf{x}| < k\}$ and zero for $\{|\mathbf{x}| > 2k\}$.

Then, we set

$$\langle T, v \rangle = \lim_{k \rightarrow \infty} \langle T, v_k \rangle. \tag{7.36}$$

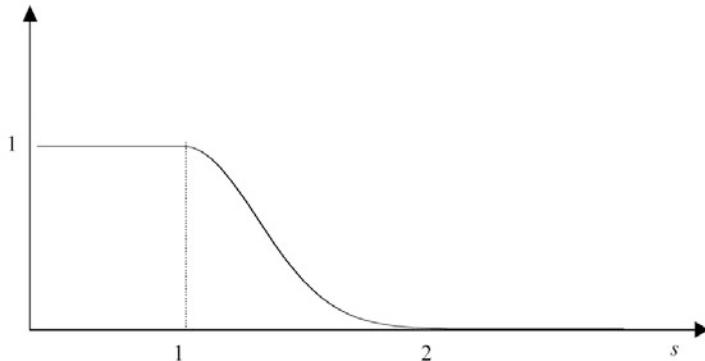


Fig. 7.4 A smooth decreasing and nonnegative function, equal to 1 in $[0, 1]$ and vanishing for $s \geq 2$

It can be shown that this limit exists and is finite, and it is independent of the approximating sequence $\{v_k\}$. Thus, a **tempered distribution is a continuous functional on $\mathcal{S}(\mathbb{R}^n)$** .

Example 7.45. We leave it as an exercise to show that the following distributions are tempered.

- a. Any polynomial.
- b. Any compactly supported distribution.
- c. Any periodic distribution (e.g. the Dirac comb).
- d. Any function $u \in L^p(\mathbb{R}^n)$, $1 \leq p \leq \infty$. Thus, we have

$$\mathcal{S}(\mathbb{R}^n) \subset L^p(\mathbb{R}^n) \subset \mathcal{S}'(\mathbb{R}^n).$$

On the contrary:

- e. $e^x \notin \mathcal{S}'(\mathbb{R})$ (why?).

Like $\mathcal{D}'(\Omega)$, $\mathcal{S}'(\mathbb{R}^n)$ possesses a *completeness* property that may be used to construct a tempered distribution or to recognize that some linear functional in $\mathcal{D}(\mathbb{R}^n)$ is a tempered distribution. First, we say that a sequence $\{T_k\} \subset \mathcal{S}'(\mathbb{R}^n)$ converges to T in $\mathcal{S}'(\mathbb{R}^n)$ if

$$\langle T_k, v \rangle \rightarrow \langle T, v \rangle, \quad \forall v \in \mathcal{S}(\mathbb{R}^n).$$

We have:

Proposition 7.46. Let $\{T_k\} \subset \mathcal{S}'(\mathbb{R}^n)$ such that

$$\lim_{k \rightarrow \infty} \langle T_k, v \rangle \text{ exists and is finite}, \quad \forall v \in \mathcal{S}(\mathbb{R}^n).$$

Then, this limit defines $T \in \mathcal{S}'(\mathbb{R}^n)$ and T_k converges to T in $\mathcal{S}'(\mathbb{R}^n)$.

Example 7.47. The Dirac comb is a tempered distribution. In fact, if $v \in \mathcal{S}(\mathbb{R})$, we have

$$\langle \text{comb}, v \rangle = \sum_{k=-\infty}^{\infty} v(k)$$

and the series is convergent since $v(k) \rightarrow 0$ more rapidly than $|k|^{-m}$ for every $m > 0$. From Proposition 7.46, $\text{comb} \in \mathcal{S}'(\mathbb{R})$.

Remark 7.48 (Convolution). If $T \in \mathcal{S}'(\mathbb{R}^n)$ and $v \in \mathcal{S}(\mathbb{R}^n)$, the convolution is well defined by formula (7.29). Then, $T * v \in \mathcal{S}'(\mathbb{R}^n)$ and coincides with a function in $C^\infty(\mathbb{R}^n)$.

7.6.2 Fourier transform in \mathcal{S}'

If $u \in L^1(\mathbb{R}^n)$, its Fourier transform is given by

$$\hat{u}(\xi) = \mathcal{F}[u](\xi) = \int_{\mathbb{R}^n} e^{-i\mathbf{x}\cdot\xi} u(\mathbf{x}) d\mathbf{x}.$$

It could be that, even if u is compactly supported, $\hat{u} \notin L^1(\mathbb{R}^n)$. For instance, if $p_a(x) = \chi_{[-a,a]}(x)$ then

$$\hat{p}_a(\xi) = 2 \frac{\sin(a\xi)}{\xi}$$

which is not¹¹ in $L^1(\mathbb{R})$. When also $\hat{u} \in L^1(\mathbb{R}^n)$, u can be reconstructed from \hat{u} through the following *inversion* formula:

Theorem 7.49. Let $u \in L^1(\mathbb{R}^n)$, $\hat{u} \in L^1(\mathbb{R}^n)$. Then

$$u(\mathbf{x}) = \frac{1}{(2\pi)^n} \int_{\mathbb{R}^n} e^{i\mathbf{x}\cdot\xi} \hat{u}(\xi) d\xi \equiv \mathcal{F}^{-1}[\hat{u}](\mathbf{x}). \quad (7.37)$$

In particular, the inversion formula (7.37) holds for $u \in \mathcal{S}(\mathbb{R}^n)$, since (exercise) $\hat{u} \in \mathcal{S}(\mathbb{R}^n)$ as well. Moreover, it can be proved that

$$u_k \rightarrow u \quad \text{in } \mathcal{S}(\mathbb{R}^n)$$

if and only if

$$\hat{u}_k \rightarrow \hat{u} \quad \text{in } \mathcal{S}(\mathbb{R}^n),$$

which means that

$$\mathcal{F}, \mathcal{F}^{-1} : \mathcal{S}(\mathbb{R}^n) \rightarrow \mathcal{S}(\mathbb{R}^n)$$

are *continuous, one-to-one operators in $\mathcal{S}(\mathbb{R}^n)$* .

¹¹ See Appendix B.

Now observe that, if $u, v \in \mathcal{S}(\mathbb{R}^n)$,

$$\begin{aligned}\langle \widehat{u}, v \rangle &= \int_{\mathbb{R}^n} \left(\int_{\mathbb{R}^n} e^{-i\mathbf{x} \cdot \boldsymbol{\xi}} u(\mathbf{x}) d\mathbf{x} \right) v(\boldsymbol{\xi}) d\boldsymbol{\xi} \\ &= \int_{\mathbb{R}^n} \left(\int_{\mathbb{R}^n} e^{-i\mathbf{x} \cdot \boldsymbol{\xi}} v(\boldsymbol{\xi}) d\boldsymbol{\xi} \right) u(\mathbf{x}) d\mathbf{x} = \langle u, \widehat{v} \rangle,\end{aligned}$$

so that the following *weak Parseval identity* holds:

$$\langle \widehat{u}, v \rangle = \langle u, \widehat{v} \rangle. \quad (7.38)$$

The key point is that the last bracket makes sense for $u = T \in \mathcal{S}'(\mathbb{R}^n)$ as well, and defines a tempered distribution. In fact:

Lemma 7.50. *Let $T \in \mathcal{S}'(\mathbb{R}^n)$. The linear functional*

$$v \mapsto \langle T, \widehat{v} \rangle, \quad \forall v \in \mathcal{S}(\mathbb{R}^n)$$

is a tempered distribution.

Proof. Let $v_k \rightarrow v$ in $\mathcal{D}(\mathbb{R}^n)$. Then $v_k \rightarrow v$ and $\widehat{v}_k \rightarrow \widehat{v}$ in $\mathcal{S}(\mathbb{R}^n)$ as well. Since $T \in \mathcal{S}'(\mathbb{R}^n)$, we have

$$\lim_{k \rightarrow \infty} \langle T, \widehat{v}_k \rangle = \langle T, \widehat{v} \rangle$$

so that $v \mapsto \langle T, \widehat{v} \rangle$ defines a distribution by Proposition 7.46. If $v_k \rightarrow 0$ in $\mathcal{S}(\mathbb{R}^n)$, then $\widehat{v}_k \rightarrow 0$ in $\mathcal{S}(\mathbb{R}^n)$ and $\langle T, \widehat{v}_k \rangle \rightarrow 0$. Thus, $v \mapsto \langle T, \widehat{v} \rangle$ is a tempered distribution. \square

We are now in position to define the Fourier transform of $T \in \mathcal{S}'(\mathbb{R}^n)$.

Definition 7.51. *Let $T \in \mathcal{S}'(\mathbb{R}^n)$. The Fourier transform $\widehat{T} = \mathcal{F}[T]$ is the tempered distribution defined by*

$$\langle \widehat{T}, v \rangle = \langle T, \widehat{v} \rangle, \quad \forall v \in \mathcal{S}(\mathbb{R}^n).$$

We see that the transform has been carried onto the test function $v \in \mathcal{S}(\mathbb{R}^n)$. As a consequence, all the properties valid for functions, continue to hold for tempered distributions. We list some of them. Let $T \in \mathcal{S}'(\mathbb{R}^n)$ and $v \in \mathcal{S}(\mathbb{R}^n)$.

1. Translation. If $\mathbf{a} \in \mathbb{R}^n$,

$$\mathcal{F}[T(\mathbf{x} - \mathbf{a})] = e^{-i\mathbf{a} \cdot \boldsymbol{\xi}} \widehat{T} \quad \text{and} \quad \mathcal{F}[e^{i\mathbf{a} \cdot \mathbf{x}} T] = \widehat{T}(\boldsymbol{\xi} - \mathbf{a}).$$

In fact ($v = v(\boldsymbol{\xi})$):

$$\begin{aligned}\langle \mathcal{F}[T(\mathbf{x} - \mathbf{a})], v \rangle &= \langle T(\mathbf{x} - \mathbf{a}), \widehat{v} \rangle = \langle T, \widehat{v}(\mathbf{x} + \mathbf{a}) \rangle \\ &= \langle T, \mathcal{F}[e^{-i\mathbf{a} \cdot \boldsymbol{\xi}} v] \rangle = \langle \widehat{T}, e^{-i\mathbf{a} \cdot \boldsymbol{\xi}} v \rangle = \langle e^{-i\mathbf{a} \cdot \boldsymbol{\xi}} \widehat{T}, v \rangle.\end{aligned}$$

2. Rescaling. If $h \in \mathbb{R}$, $h \neq 0$,

$$\mathcal{F}[T(h\mathbf{x})] = \frac{1}{|h|^n} \widehat{T}\left(\frac{\boldsymbol{\xi}}{h}\right).$$

In fact, using (7.20) first with $\psi(\mathbf{x}) = h\mathbf{x}$, then with $\psi^{-1}(\mathbf{x}) = h^{-1}\mathbf{x}$, we write:

$$\begin{aligned} \langle \mathcal{F}[T(h\mathbf{x})], v \rangle &= \langle T(h\mathbf{x}), \widehat{v} \rangle = \langle T, \frac{1}{|h|^n} \widehat{v}\left(\frac{\mathbf{x}}{h}\right) \rangle \\ &= \langle T, \mathcal{F}[v(h\boldsymbol{\xi})] \rangle = \langle \widehat{T}, v(h\boldsymbol{\xi}) \rangle = \langle \frac{1}{|h|^n} \widehat{T}\left(\frac{\boldsymbol{\xi}}{h}\right), v \rangle. \end{aligned}$$

In particular, choosing $h = -1$, it follows that if T is even (odd) then \widehat{T} is even (odd).

3. Derivatives:

$$a) \mathcal{F}[\partial_{x_j} T] = i\xi_j \widehat{T} \quad \text{and} \quad b) \mathcal{F}[x_j T] = i\partial_{\xi_j} \widehat{T}.$$

Namely:

$$\begin{aligned} \langle \mathcal{F}[\partial_{x_j} T], v \rangle &= \langle \partial_{x_j} T, \widehat{v} \rangle = -\langle T, \partial_{x_j} \widehat{v} \rangle \\ &= \langle T, \mathcal{F}[i\xi_j v] \rangle = \langle i\xi_j \widehat{T}, v \rangle. \end{aligned}$$

For the second formula, we have:

$$\begin{aligned} \langle \mathcal{F}[x_j T], v \rangle &= \langle x_j T, \widehat{v} \rangle = \langle T, x_j \widehat{v} \rangle \\ &= \langle T, -i\mathcal{F}[\partial_{\xi_j} v] \rangle = \langle -i\widehat{T}, \partial_{\xi_j} v \rangle = \langle i\partial_{\xi_j} \widehat{T}, v \rangle. \end{aligned}$$

4. Convolution¹². If $T \in \mathcal{S}'(\mathbb{R}^n)$ and $v \in \mathcal{S}(\mathbb{R}^n)$,

$$\mathcal{F}[T * v] = \widehat{T} \cdot \widehat{v}.$$

Example 7.52. We know that $\delta_n \in \mathcal{S}'(\mathbb{R}^n)$. We have:

$$\widehat{\delta}_n = 1, \quad \widehat{1} = (2\pi)^n \delta_n.$$

In fact:

$$\langle \widehat{\delta}_n, v \rangle = \langle \delta_n, \widehat{v} \rangle = \widehat{v}(\mathbf{0}) = \int_{\mathbb{R}^n} v(\boldsymbol{\xi}) d\boldsymbol{\xi} = \langle 1, v \rangle.$$

For the second formula, using (7.37) we have:

$$\begin{aligned} \langle \widehat{1}, v \rangle &= \langle 1, \widehat{v} \rangle = \int_{\mathbb{R}^n} \widehat{v}(\mathbf{x}) d\mathbf{x} = (2\pi)^n v(\mathbf{0}) \\ &= \langle (2\pi)^n \delta_n, v \rangle. \end{aligned}$$

Example 7.53. Transform of x_j :

$$\widehat{x}_j = i(2\pi)^n \partial_{\xi_j} \delta_n.$$

¹² We omit the proof.

Indeed, from 3, b) and Example 7.52, we may write

$$\widehat{x}_j = \mathcal{F}[x_j \cdot 1] = i\partial_{\xi_j} \widehat{1} = i(2\pi)^n \partial_{\xi_j} \delta_n.$$

7.6.3 Fourier transform in L^2

Since $L^2(\mathbb{R}^n) \subset \mathcal{S}'(\mathbb{R}^n)$, the Fourier transform is well defined for all functions in $L^2(\mathbb{R}^n)$. The following theorem holds, where \bar{v} denotes the complex conjugate of v .

Theorem 7.54. *$u \in L^2(\mathbb{R}^n)$ if and only if $\widehat{u} \in L^2(\mathbb{R}^n)$. Moreover, if $u, v \in L^2(\mathbb{R}^n)$, the following strong Parseval identity holds:*

$$\int_{\mathbb{R}^n} \widehat{u} \cdot \bar{v} = (2\pi)^n \int_{\mathbb{R}^n} u \cdot \bar{v}. \quad (7.39)$$

In particular

$$\|\widehat{u}\|_{L^2(\mathbb{R}^n)}^2 = (2\pi)^n \|u\|_{L^2(\mathbb{R}^n)}^2. \quad (7.40)$$

Formula (7.40) shows that *the Fourier transform is an isometry in $L^2(\mathbb{R}^n)$* (but for the factor $(2\pi)^n$).

Proof. Since $\mathcal{S}(\mathbb{R}^n)$ is dense in $L^2(\mathbb{R}^n)$, it is enough to prove (7.39) for $u, v \in \mathcal{S}(\mathbb{R}^n)$. Let $w = \bar{v}$. From (7.38) we have

$$\int_{\mathbb{R}^n} \widehat{u} \cdot w = \int_{\mathbb{R}^n} u \cdot \widehat{w}.$$

On the other hand,

$$\widehat{w}(\mathbf{x}) = \int_{\mathbb{R}^n} e^{-i\mathbf{x} \cdot \mathbf{y}} \bar{v}(\mathbf{y}) d\mathbf{y} = (2\pi)^n \overline{\mathcal{F}^{-1}[\bar{v}]}(\mathbf{x}) = (2\pi)^n \bar{v}(\mathbf{x})$$

and (7.39) follows. □

Example 7.55. Let us compute

$$\int_{\mathbb{R}} \left(\frac{\sin x}{x} \right)^2 dx.$$

We know that the Fourier transform of $p_1 = \chi_{[-1,1]}$ is

$$\hat{p}_1(\xi) = 2 \frac{\sin \xi}{\xi},$$

which belongs to $L^2(\mathbb{R})$. Thus, (7.40) yields

$$4 \int_{\mathbb{R}} \left(\frac{\sin \xi}{\xi} \right)^2 d\xi = 2\pi \int_{\mathbb{R}} (\chi_{[-1,1]}(x))^2 dx = 4\pi$$

whence

$$\int_{\mathbb{R}} \left(\frac{\sin x}{x} \right)^2 dx = \pi.$$

7.7 Sobolev Spaces

7.7.1 An abstract construction

Sobolev spaces constitute one of the most relevant functional settings for the treatment of boundary value problems. Here, we will be mainly concerned with Sobolev spaces based on $L^2(\Omega)$, developing only the theoretical elements we will need in the sequel¹³.

The following abstract theorem is a flexible tool for generating Sobolev Spaces. The ingredients of the construction are:

- The space $\mathcal{D}'(\Omega; \mathbb{R}^n)$, in particular, for $n = 1$, $\mathcal{D}'(\Omega)$.
- Two Hilbert spaces H and Z with $Z \hookrightarrow \mathcal{D}'(\Omega; \mathbb{R}^n)$ for some $n \geq 1$. In particular

$$v_k \rightarrow v \text{ in } Z \quad \text{implies} \quad v_k \rightarrow v \text{ in } \mathcal{D}'(\Omega; \mathbb{R}^n). \quad (7.41)$$

- A linear continuous operator $L : H \rightarrow \mathcal{D}'(\Omega; \mathbb{R}^n)$ (such as a gradient or a divergence).

We have:

Theorem 7.56. Define

$$W = \{v \in H : Lv \in Z\}$$

and

$$(u, v)_W = (u, v)_H + (Lu, Lv)_Z. \quad (7.42)$$

Then W is a Hilbert space with inner product given by (7.42). The embedding of W in H is continuous and the restriction of L to W is continuous from W into Z .

Proof. It is easy to check that (7.42) has all the properties of an inner product, with induced norm

$$\|u\|_W = \sqrt{\|u\|_H^2 + \|Lu\|_Z^2}.$$

Thus W is an inner-product space. It remains to check its completeness. Let $\{v_k\}$ be a Cauchy sequence in W . We must show that there exists $v \in H$ such that

$$v_k \rightarrow v \text{ in } H$$

and

$$Lv_k \rightarrow Lv \text{ in } Z.$$

¹³ We omit the most technical proofs, that can be found, for instance, in the classical books of Adams, 1975 [31], Maz'ya, 1985 [35] or Ziemer, 1989 [40].

Observe that $\{v_k\}$ and $\{Lv_k\}$ are Cauchy sequences in H and Z , respectively. Thus, there exist $v \in H$ and $z \in Z$ such that

$$v_k \rightarrow v \quad \text{in } H \quad \text{and} \quad Lv_k \rightarrow z \quad \text{in } Z.$$

The continuity of L and (7.41) yield

$$Lv_k \rightarrow Lv \quad \text{in } \mathcal{D}'(\Omega; \mathbb{R}^n) \quad \text{and} \quad Lv_k \rightarrow z \quad \text{in } \mathcal{D}'(\Omega; \mathbb{R}^n).$$

Since the limit of a sequence in $\mathcal{D}'(\Omega; \mathbb{R}^n)$ is unique, we infer that $Lv = z$. Therefore

$$Lv_k \rightarrow Lv \quad \text{in } Z$$

and W is a Hilbert space.

The continuity of the embedding $W \subset H$ follows from

$$\|u\|_H \leq \|u\|_W$$

while the continuity of $L|_W : W \rightarrow Z$ follows from

$$\|Lu\|_Z \leq \|u\|_W.$$

The proof is complete. □

7.7.2 The space $H^1(\Omega)$

Let $\Omega \subseteq \mathbb{R}^n$ be a domain. Choose in Theorem 7.56:

$$H = L^2(\Omega), \quad Z = L^2(\Omega; \mathbb{R}^n) \hookrightarrow \mathcal{D}'(\Omega; \mathbb{R}^n)$$

and $L : H \rightarrow \mathcal{D}'(\Omega; \mathbb{R}^n)$ given by

$$L = \nabla,$$

where the gradient is considered in the sense of distributions. Then, W is the **Sobolev space** of the functions in $L^2(\Omega)$, whose *first derivatives in the sense of distributions are functions in $L^2(\Omega)$* . For this space we use the symbol¹⁴ $H^1(\Omega)$. Thus:

$$H^1(\Omega) = \{v \in L^2(\Omega) : \nabla v \in L^2(\Omega; \mathbb{R}^n)\}.$$

In other words, if $v \in H^1(\Omega)$, every partial derivative $\partial_{x_i} v$ is a function $v_i \in L^2(\Omega)$. This means that

$$\langle \partial_{x_i} v, \varphi \rangle = -(v, \partial_{x_i} \varphi)_{L^2(\Omega)} = (v_i, \varphi)_{L^2(\Omega)}, \quad \forall \varphi \in \mathcal{D}(\Omega)$$

or, more explicitly,

$$\int_{\Omega} v(\mathbf{x}) \partial_{x_i} \varphi(\mathbf{x}) d\mathbf{x} = - \int_{\Omega} v_i(\mathbf{x}) \varphi(\mathbf{x}) d\mathbf{x}, \quad \forall \varphi \in \mathcal{D}(\Omega).$$

¹⁴ Also $H^{1,2}(\Omega)$ or $W^{1,2}(\Omega)$ are used.

In many applied situations, the Dirichlet integral

$$\int_{\Omega} |\nabla v|^2$$

represents an energy. The functions in $H^1(\Omega)$ are therefore associated with *configurations having finite energy*. From Theorem 7.56 and the separability of $L^2(\Omega)$, we have¹⁵:

Proposition 7.57. *$H^1(\Omega)$ is a separable Hilbert space, continuously embedded in $L^2(\Omega)$. The gradient operator is continuous from $H^1(\Omega)$ into $L^2(\Omega; \mathbb{R}^n)$.*

The inner product and the norm in $H^1(\Omega)$ are given, respectively, by

$$(u, v)_{H^1(\Omega)} = \int_{\Omega} uv \, d\mathbf{x} + \int_{\Omega} \nabla u \cdot \nabla v \, d\mathbf{x}$$

and

$$\|u\|_{H^1(\Omega)}^2 = \int_{\Omega} u^2 \, d\mathbf{x} + \int_{\Omega} |\nabla u|^2 \, d\mathbf{x}.$$

Example 7.58. Let $\Omega = B_{1/2}(\mathbf{0}) = \{\mathbf{x} \in \mathbb{R}^2 : |\mathbf{x}| < 1/2\}$ and

$$u(\mathbf{x}) = (-\log|\mathbf{x}|)^a, \quad \mathbf{x} \neq \mathbf{0}.$$

We have, using polar coordinates,

$$\int_{B_{1/2}(\mathbf{0})} u^2 \, d\mathbf{x} = 2\pi \int_0^{1/2} (-\log r)^{2a} r dr < \infty, \quad \text{for every } a \in \mathbb{R},$$

so that $u \in L^2(B_{1/2}(\mathbf{0}))$ for every $a \in \mathbb{R}$. Moreover:

$$u_{x_i} = -ax_i |\mathbf{x}|^{-2} (-\log|\mathbf{x}|)^{a-1}, \quad i = 1, 2,$$

and therefore

$$|\nabla u| = \left| a(-\log|\mathbf{x}|)^{a-1} \right| |\mathbf{x}|^{-1}.$$

Using polar coordinates, we get

$$\int_{B_{1/2}(\mathbf{0})} |\nabla u|^2 \, d\mathbf{x} = 2\pi a^2 \int_0^{1/2} |\log r|^{2a-2} r^{-1} dr.$$

¹⁵ If we associate with each element u of $H^1(\Omega)$ the vector $u, u_{x_1}, \dots, u_{x_n}$, we see that $H^1(\Omega)$ is isometric to a subspace of

$$L^2(\Omega) \times L^2(\Omega) \times \dots \times L^2(\Omega) = L^2(\Omega; \mathbb{R}^{n+1})$$

which is separable because $L^2(\Omega)$ is separable.

This integral is finite only if $2 - 2a > 1$ or $a < 1/2$. In particular, ∇u represents the gradient of u in the sense of distribution as well. We conclude that $u \in H^1(B_{1/2}(\mathbf{0}))$ only if $a < 1/2$. We point out that when $a > 0$, u is **unbounded** near $\mathbf{0}$.

We have affirmed that the Sobolev spaces constitute an adequate functional setting to solve boundary value problems. This point requires that we go more deeply into the arguments in Sect. 6.1 and that we make some necessary observations. When we write $f \in L^2(\Omega)$, we may think of a single function

$$f : \Omega \rightarrow \mathbb{R} \text{ (or } \mathbb{C}),$$

square summable in the Lebesgue sense. However, if we want to exploit the Hilbert space structure of $L^2(\Omega)$, we need to identify two functions when they are equal a.e. in Ω . Adopting this point of view, each element in $L^2(\Omega)$ is actually an *equivalence class* of which f is a *representative*. The drawback here is that it does not make sense anymore to compute the *value of f at a single point*, since a point is a set with measure zero!

The same considerations hold for “functions” in $H^1(\Omega)$, since

$$H^1(\Omega) \subset L^2(\Omega).$$

On the other hand, if we deal with a boundary value problem, it is clear that *we would like to compute the solution at any point in Ω* !

Even more important is the question of the *trace of a function on the boundary of a domain*. By *trace* of f on $\partial\Omega$ we mean the restriction of f to $\partial\Omega$. In a Dirichlet or Neumann problem we assign precisely the trace of the solution or of its normal derivative on $\partial\Omega$, which is a set with measure zero. Does this make any sense if $u \in H^1(\Omega)$?

It could be objected that, after all, one always works with a single representative and that the numerical approximation of the solution only involves a finite number of points, making meaningless the distinction between functions in $L^2(\Omega)$ or in $H^1(\Omega)$ or continuous. Then, why do we have to struggle to give a precise meaning to the trace of a function in $H^1(\Omega)$?

One reason comes from numerical analysis itself, in particular from the need to keep under control the approximation errors and to provide stability estimates.

Let us ask, for instance: if a Dirichlet data is known within an error of order ε in L^2 -norm on $\partial\Omega$, can we estimate in terms of ε the corresponding error in the solution?

If we are satisfied with an L^2 or an L^∞ norm *in the interior of the domain*, this kind of estimate may be available. But often, an energy estimate is required, involving the norm in $L^2(\Omega)$ of the gradient of the solution. In this case, the L^2 norm of the boundary data is not sufficient and it turns out that the exact information on the data, necessary to restore an energy estimate, is encoded in the trace characterization of Sect. 7.9.

We shall introduce the notion of *trace on $\partial\Omega$* for a function in $H^1(\Omega)$, using an approximation procedure with smooth functions. However, there are two cases, in which the trace problem may be solved quite simply: the one-dimensional case and the case of functions with zero trace. We start with the first case.

- *Characterization of $H^1(a, b)$.* As Example 7.58 shows, a function in $H^1(\Omega)$ may be unbounded. In dimension $n = 1$ this cannot occur. In fact, the elements in $H^1(a, b)$ are continuous functions¹⁶ in $[a, b]$.

Proposition 7.59. *Let $u \in L^2(a, b)$. Then $u \in H^1(a, b)$ if and only if u is continuous in $[a, b]$ and there exists $w \in L^2(a, b)$ such that*

$$u(y) = u(x) + \int_x^y w(s) ds, \quad \forall x, y \in [a, b]. \quad (7.43)$$

Moreover $u' = w$ (both a.e. and in the sense of distributions).

Proof. Assume that u is continuous in $[a, b]$ and that (7.43) holds with $w \in L^2(a, b)$. Choose $x = a$. Replacing, if necessary, u by $u - u(a)$, we may assume $u(a) = 0$, so that

$$u(y) = \int_a^y w(s) ds, \quad \forall x, y \in [a, b].$$

Let $\varphi \in \mathcal{D}(a, b)$. We have:

$$\langle u', \varphi \rangle = -\langle u, \varphi' \rangle = - \int_a^b u(s) \varphi'(s) ds = - \int_a^b \left[\int_a^s w(t) dt \right] \varphi'(s) ds =$$

(exchanging the order of integration)

$$= - \int_a^b \left[\int_t^b \varphi'(s) ds \right] w(t) dt = \int_a^b \varphi(t) w(t) dt = \langle w, \varphi \rangle.$$

Thus $u' = w$ in $\mathcal{D}'(a, b)$ and therefore $u \in H^1(a, b)$. From the Lebesgue Differentiation Theorem¹⁷ we deduce that $u' = w$ a.e. as well.

Viceversa, let $u \in H^1(a, b)$. Define

$$v(x) = \int_c^x u'(s) ds, \quad x \in [a, b]. \quad (7.44)$$

The function v is continuous in $[a, b]$ and the above proof shows that $v' = u'$ in $\mathcal{D}'(a, b)$. Then, by Proposition 7.18, p. 440,

$$u = v + C, \quad C \in \mathbb{R},$$

and therefore u is continuous in $[a, b]$ as well. Moreover, (7.44) yields

$$u(y) - u(x) = v(y) - v(x) = \int_x^y u'(s) ds$$

which is (7.43). □

¹⁶ Rigorously: every equivalence class in $H^1(a, b)$ has a representative which is continuous in $[a, b]$.

¹⁷ See Appendix B.

Since a function $u \in H^1(a, b)$ is continuous in $[a, b]$, the value $u(x_0)$ at every point $x_0 \in [a, b]$ makes perfect sense. In particular the *trace* of u at the end points of the interval is given by the values $u(a)$ and $u(b)$.

7.7.3 The space $H_0^1(\Omega)$

Let $\Omega \subseteq \mathbb{R}^n$ be a domain. We introduce an important subspace of $H^1(\Omega)$.

Definition 7.60. We denote by $H_0^1(\Omega)$ the closure of $\mathcal{D}(\Omega)$ in $H^1(\Omega)$.

Thus $u \in H_0^1(\Omega)$ if and only if there exists a sequence $\{\varphi_k\} \subset \mathcal{D}(\Omega)$ such that $\varphi_k \rightarrow u$ in $H^1(\Omega)$, that is, such that both $\|\varphi_k - u\|_{L^2(\Omega)} \rightarrow 0$ and $\|\nabla \varphi_k - \nabla u\|_{L^2(\Omega; \mathbb{R}^n)} \rightarrow 0$ as $k \rightarrow \infty$.

Since the test functions in $\mathcal{D}(\Omega)$ have zero trace on $\partial\Omega$, every $u \in H_0^1(\Omega)$ “inherits” this property and it is reasonable to consider the elements $H_0^1(\Omega)$ as the functions in $H^1(\Omega)$ with zero trace on $\partial\Omega$. Clearly, $H_0^1(\Omega)$ is a Hilbert subspace of $H^1(\Omega)$.

An important property that holds in $H_0^1(\Omega)$, particularly useful in the solution of boundary value problems, is expressed by the following inequality of Poincaré. Recall that the diameter of a set Ω is given by

$$\text{diam}(\Omega) = \sup_{\mathbf{x}, \mathbf{y} \in \Omega} |\mathbf{x} - \mathbf{y}|.$$

Theorem 7.61. Let $\Omega \subset \mathbb{R}^n$ be a bounded domain. There exists a positive constant C_P (a Poincaré’s constant) depending only on n and $\text{diam}(\Omega)$, such that, for every $u \in H_0^1(\Omega)$,

$$\|u\|_{L^2(\Omega)} \leq C_P \|\nabla u\|_{L^2(\Omega; \mathbb{R}^n)}. \quad (7.45)$$

Proof. We use a strategy which is rather common for proving formulas in $H_0^1(\Omega)$. First, we prove the formula for $v \in \mathcal{D}(\Omega)$; then, if $u \in H_0^1(\Omega)$, we select a sequence $\{v_k\} \subset \mathcal{D}(\Omega)$ converging to u in the $H^1(\Omega)$ norm as $k \rightarrow \infty$, that is

$$\|v_k - u\|_{L^2(\Omega)} \rightarrow 0, \quad \|\nabla v_k - \nabla u\|_{L^2(\Omega; \mathbb{R}^n)} \rightarrow 0.$$

In particular

$$\|v_k\|_{L^2(\Omega)} \rightarrow \|u\|_{L^2(\Omega)}, \quad \|\nabla v_k\|_{L^2(\Omega; \mathbb{R}^n)} \rightarrow \|\nabla u\|_{L^2(\Omega; \mathbb{R}^n)}.$$

Since (7.45) holds for every v_k , we have

$$\|v_k\|_{L^2(\Omega)} \leq C_P \|\nabla v_k\|_{L^2(\Omega; \mathbb{R}^n)}.$$

Letting $k \rightarrow \infty$ we obtain (7.45) for u . Thus, it is enough to prove (7.45) for $v \in \mathcal{D}(\Omega)$. Assume without loss of generality that $\mathbf{0} \in \Omega$, and set $\max_{\mathbf{x} \in \Omega} |\mathbf{x}| \leq M = \text{diam}(\Omega) < \infty$. Applying the Gauss Divergence Theorem, we can write

$$\int_{\Omega} \operatorname{div}(v^2 \mathbf{x}) d\mathbf{x} = 0, \quad (7.46)$$

since $v = 0$ on $\partial\Omega$. Now,

$$\operatorname{div}(v^2 \mathbf{x}) = 2v\nabla v \cdot \mathbf{x} + nv^2$$

so that (7.46) yields

$$\int_{\Omega} v^2 d\mathbf{x} = -\frac{2}{n} \int_{\Omega} v \nabla v \cdot \mathbf{x} d\mathbf{x}.$$

Since Ω is bounded, using Schwarz's inequality, we get

$$\int_{\Omega} v^2 d\mathbf{x} = \frac{2}{n} \left| \int_{\Omega} v \nabla v \cdot \mathbf{x} d\mathbf{x} \right| \leq \frac{2M}{n} \left(\int_{\Omega} v^2 d\mathbf{x} \right)^{1/2} \left(\int_{\Omega} |\nabla v|^2 d\mathbf{x} \right)^{1/2}.$$

Simplifying, it follows that

$$\|v\|_{L^2(\Omega)} \leq C_P \|\nabla v\|_{L^2(\Omega; \mathbb{R}^n)}$$

with $C_P = 2M/n$. \square

Inequality (7.45) implies that in $H_0^1(\Omega)$ the norm $\|u\|_{H^1(\Omega)}$ is equivalent to $\|\nabla u\|_{L^2(\Omega; \mathbb{R}^n)}$. Indeed

$$\|u\|_{H^1(\Omega)} = \sqrt{\|u\|_{L^2(\Omega)}^2 + \|\nabla u\|_{L^2(\Omega; \mathbb{R}^n)}^2}$$

and from (7.45),

$$\|\nabla u\|_{L^2(\Omega; \mathbb{R}^n)} \leq \|u\|_{H^1(\Omega)} \leq \sqrt{C_P^2 + 1} \|\nabla u\|_{L^2(\Omega; \mathbb{R}^n)}.$$

When Ω is bounded, unless explicitly stated, **we will choose in H_0^1**

$$(u, v)_{H_0^1(\Omega)} = (\nabla u, \nabla v)_{L^2(\Omega; \mathbb{R}^n)} \quad \text{and} \quad \|u\|_{H_0^1(\Omega)} = \|\nabla u\|_{L^2(\Omega; \mathbb{R}^n)}$$

as inner product and norm, respectively.

7.7.4 The dual of $H_0^1(\Omega)$

In the applications of the Lax-Milgram theorem to boundary value problems, the dual of $H_0^1(\Omega)$ plays an important role. In fact it deserves a special symbol.

Definition 7.62. We denote by $H^{-1}(\Omega)$ the dual of $H_0^1(\Omega)$ with the norm

$$\|F\|_{H^{-1}(\Omega)} = \sup \left\{ |Fv| : v \in H_0^1(\Omega), \|v\|_{H_0^1(\Omega)} \leq 1 \right\}.$$

The first thing to observe is that, since $\mathcal{D}(\Omega)$ is dense (by definition) and continuously embedded in $H_0^1(\Omega)$, $H^{-1}(\Omega)$ is a *space of distributions*. This means two things:

- a) If $F \in H^{-1}(\Omega)$, its restriction to $\mathcal{D}(\Omega)$ is a distribution.
- b) If $F, G \in H^{-1}(\Omega)$ and $F\varphi = G\varphi$ for every $\varphi \in \mathcal{D}(\Omega)$, then $F = G$.

To prove a) it is enough to note that if $\varphi_k \rightarrow \varphi$ in $\mathcal{D}(\Omega)$, then $\varphi_k \rightarrow \varphi$ in $H_0^1(\Omega)$ as well, and therefore $F\varphi_k \rightarrow F\varphi$. Thus $F \in \mathcal{D}'(\Omega)$.

To prove b), let $u \in H_0^1(\Omega)$ and $\varphi_k \rightarrow u$ in $H_0^1(\Omega)$, with $\varphi_k \in \mathcal{D}(\Omega)$. Then, since $F\varphi_k = G\varphi_k$, we may write

$$Fu = \lim_{k \rightarrow +\infty} F\varphi_k = \lim_{k \rightarrow +\infty} G\varphi_k = Gu, \quad \forall u \in H_0^1(\Omega),$$

whence $F = G$.

Thus, $H^{-1}(\Omega)$ is in *one-to-one* correspondence with a subspace of $\mathcal{D}'(\Omega)$ and in this sense we will write

$$H^{-1}(\Omega) \subset \mathcal{D}'(\Omega).$$

The following theorem gives a characterization of the elements of $H^{-1}(\Omega)$.

Theorem 7.63. $H^{-1}(\Omega)$ is the set of distributions of the form

$$F = f_0 + \operatorname{div} \mathbf{f} \tag{7.47}$$

where $f_0 \in L^2(\Omega)$ and $\mathbf{f} = (f_1, \dots, f_n) \in L^2(\Omega; \mathbb{R}^n)$. Moreover:

$$\|F\|_{H^{-1}(\Omega)} \leq \|f_0\|_{L^2(\Omega)} + \|\mathbf{f}\|_{L^2(\Omega; \mathbb{R}^n)}. \tag{7.48}$$

Proof. Let $F \in H^{-1}(\Omega)$. From Riesz's Representation Theorem, there exists a unique $u \in H_0^1(\Omega)$ such that

$$(u, v)_{H_0^1(\Omega)} = Fv, \quad \forall v \in H_0^1(\Omega).$$

Since Ω can be unbounded we use in $H_0^1(\Omega)$ the full innerproduct in $H^1(\Omega)$ and we get:

$$(u, v)_{H_0^1(\Omega)} = (\nabla u, \nabla v)_{L^2(\Omega; \mathbb{R}^n)} + (u, v)_{L^2(\Omega)} = -\langle \operatorname{div} \nabla u, v \rangle + (u, v)_{L^2(\Omega)}$$

in $\mathcal{D}'(\Omega)$, it follows that (7.47) holds with $f_0 = u$ and $\mathbf{f} = -\nabla u$. Moreover, $\|F\|_{H^{-1}(\Omega)} = \|u\|_{H_0^1(\Omega)} = (\|f_0\|_{L^2(\Omega)}^2 + \|\mathbf{f}\|_{L^2(\Omega; \mathbb{R}^n)}^2)^{1/2} \leq \|f_0\|_{L^2(\Omega)} + \|\mathbf{f}\|_{L^2(\Omega; \mathbb{R}^n)}$.

Viceversa, let $F = f_0 + \operatorname{div} \mathbf{f}$, with $f_0 \in L^2(\Omega)$ and $\mathbf{f} = L^2(\Omega; \mathbb{R}^n)$. Then $F \in \mathcal{D}'(\Omega)$ and, letting $\langle F, v \rangle = Fv$, we have;

$$Fv = \int_{\Omega} f_0 v \, d\mathbf{x} + \int_{\Omega} \mathbf{f} \cdot \nabla v \, d\mathbf{x}, \quad \forall v \in \mathcal{D}(\Omega).$$

From the Schwarz inequality we have

$$|Fv| \leq \left\{ \|f_0\|_{L^2(\Omega)} + \|\mathbf{f}\|_{L^2(\Omega; \mathbb{R}^n)} \right\} \|v\|_{H_0^1(\Omega)}. \tag{7.49}$$

Thus, F is continuous in the H_0^1 -norm. It remains to show that F has a unique continuous extension to all $H_0^1(\Omega)$. Take $u \in H_0^1(\Omega)$ and $\{v_k\} \subset \mathcal{D}(\Omega)$ such that $\|v_k - u\|_{H_0^1(\Omega)} \rightarrow 0$. Then, (7.49) yields

$$|Fv_k - Fv_h| \leq \left\{ \|f_0\|_{L^2(\Omega)} + \|\mathbf{f}\|_{L^2(\Omega; \mathbb{R}^n)} \right\} \|v_k - v_h\|_{H_0^1(\Omega)}.$$

Therefore $\{Fv_k\}$ is a Cauchy sequence in \mathbb{R} and converges to a limit we may denote by Fu , which is independent of the sequence approximating u , as it is not difficult to check.

Finally, since

$$|Fu| = \lim_{k \rightarrow \infty} |Fv_k| \quad \text{and} \quad \|u\|_{H_0^1(\Omega)} = \lim_{k \rightarrow \infty} \|v_k\|_{H_0^1(\Omega)},$$

from (7.49) we get:

$$|Fu| \leq \left\{ \|f_0\|_{L^2(\Omega)} + \|\mathbf{f}\|_{L^2(\Omega; \mathbb{R}^n)} \right\} \|u\|_{H_0^1(\Omega)}$$

showing that $F \in H^{-1}(\Omega)$ and that (7.48) holds. \square

Theorem 7.63 says that the elements of $H^{-1}(\Omega)$ are represented by a linear combination of functions in $L^2(\Omega)$ and their first derivatives (in the sense of distributions). In particular, $L^2(\Omega) \hookrightarrow H^{-1}(\Omega)$.

Example 7.64. If $n = 1$, the (restriction of the) Dirac δ belongs to $H^{-1}(-a, a)$. Indeed, we have $\delta = \mathcal{H}'$ where \mathcal{H} is the Heaviside function, and $\mathcal{H} \in L^2(-a, a)$. However, if $n \geq 2$ and $\mathbf{0} \in \Omega$, $\delta_n \notin H^{-1}(\Omega)$. For instance, let $n = 2$ and $\Omega = B_{1/2}(\mathbf{0})$. Suppose $\delta_2 \in H^{-1}(\Omega)$. Then we could write

$$|\varphi(\mathbf{0})| \leq K \|\varphi\|_{H_0^1(\Omega)}$$

and, using the density of $\mathcal{D}(\Omega)$ in $H_0^1(\Omega)$, this estimate should hold for any $u \in H_0^1(\Omega)$ as well. But this is impossible, since in $H_0^1(\Omega)$ there are functions which are unbounded near the origin, as we have seen in Example 7.58, p. 463.

Example 7.65. Let Ω be a smooth, bounded domain in \mathbb{R}^n . Let $u = \chi_\Omega$ be its characteristic function. Since $\chi_\Omega \in L^2(\mathbb{R}^n)$, the distribution $\mathbf{F} = \nabla \chi_\Omega$ belongs to $H^{-1}(\mathbb{R}^n; \mathbb{R}^n)$. The support of $\mathbf{F} = \nabla \chi_\Omega$ coincides with $\partial\Omega$ and its action on a test $\varphi \in \mathcal{D}(\mathbb{R}^n; \mathbb{R}^n)$ is described by the following formula:

$$\langle \nabla \chi_\Omega, \varphi \rangle = - \int_{\mathbb{R}^n} \chi_\Omega \operatorname{div} \varphi \, d\mathbf{x} = - \int_{\partial\Omega} \varphi \cdot \boldsymbol{\nu} \, d\sigma.$$

Thus $\langle \nabla \chi_\Omega, \varphi \rangle$ gives the inward flux of the vector field φ through $\partial\Omega$.

It is important to avoid confusion between $H^{-1}(\Omega)$ and $H^1(\Omega)^*$, the dual of $H^1(\Omega)$. Since, in general, $\mathcal{D}(\Omega)$ is **not dense** in $H^1(\Omega)$, the space $H^1(\Omega)^*$ is **not** a space of distributions. Indeed, although the restriction to $\mathcal{D}(\Omega)$ of every $T \in H^1(\Omega)^*$ is a distribution, this restriction does not identify T . As a simple example, take a constant $\mathbf{f} \in \mathbb{R}^n$ and define

$$T\varphi = \int_{\Omega} \mathbf{f} \cdot \nabla \varphi \, d\mathbf{x}.$$

Since

$$|T\varphi| \leq |\mathbf{f}| \|\nabla \varphi\|_{L^2(\Omega; \mathbb{R}^n)},$$

we infer that $T \in H^1(\Omega)^*$. However, the restriction of T to $\mathcal{D}(\Omega)$ is the zero distribution, since in $\mathcal{D}'(\Omega)$ we have

$$\langle T, \varphi \rangle = - \langle \operatorname{div} \mathbf{f}, \varphi \rangle = 0, \quad \forall \varphi \in \mathcal{D}(\Omega).$$

7.7.5 The spaces $H^m(\Omega)$, $m > 1$

By involving higher order derivatives, we may construct new Sobolev spaces. Let N be the number of multi-indexes $\alpha = (\alpha_1, \dots, \alpha_n)$ such that $|\alpha| = \sum_{i=1}^n \alpha_i \leq m$. Choose in Theorem 7.56

$$H = L^2(\Omega), \quad Z = L^2(\Omega; \mathbb{R}^N) \subset \mathcal{D}'(\Omega; \mathbb{R}^N),$$

and $L : L^2(\Omega) \rightarrow \mathcal{D}'(\Omega; \mathbb{R}^N)$ given by

$$Lv = \{D^\alpha v\}_{|\alpha| \leq m}.$$

Then W is the **Sobolev space** of the functions in $L^2(\Omega)$, whose *derivatives (in the sense of distributions) up to order m included, are functions in $L^2(\Omega)$* . For this space we use the symbol $H^m(\Omega)$. Thus:

$$H^m(\Omega) = \{v \in L^2(\Omega) : D^\alpha v \in L^2(\Omega), \quad \forall \alpha : |\alpha| \leq m\}.$$

From Theorem 7.56 and the separability of $L^2(\Omega)$, we deduce:

Proposition 7.66. *$H^m(\Omega)$ is a separable Hilbert space, continuously embedded in $L^2(\Omega)$. The operators D^α , $|\alpha| \leq m$, are continuous from $H^m(\Omega)$ into $L^2(\Omega)$.*

The inner product and the norm in $H^m(\Omega)$ are given, respectively, by

$$(u, v)_{H^m(\Omega)} = \sum_{|\alpha| \leq m} \int_{\Omega} D^\alpha u D^\alpha v \, d\mathbf{x}$$

and

$$\|u\|_{H^m(\Omega)}^2 = \sum_{|\alpha| \leq m} \int_{\Omega} |D^\alpha u|^2 \, d\mathbf{x}.$$

If $u \in H^m(\Omega)$, any derivative of u of order $k \leq m$ belongs to $H^{m-k}(\Omega)$ and $H^m(\Omega) \hookrightarrow H^{m-k}(\Omega)$, $k \geq 1$.

Example 7.67. Let $B_{1/2}(\mathbf{0}) \subset \mathbb{R}^3$ and consider $u(\mathbf{x}) = |\mathbf{x}|^{-a}$. It is easy to check (see Problem 7.23) that $u \in H^1(B_{1/2}(\mathbf{0}))$ if $a < 1/2$. The second order derivatives of u are given by:

$$u_{x_i x_j} = a(a+2)x_i x_j |\mathbf{x}|^{-a-4} - a\delta_{ij} |\mathbf{x}|^{-a-2}.$$

Then

$$|u_{x_i x_j}| \leq |a(a+2)| |\mathbf{x}|^{-a-2}$$

so that $u_{x_i x_j} \in L^2(B_{1/2}(0))$ if $2a + 4 < 3$, or $a < -\frac{1}{2}$. Thus $u \in H^2(B_{1/2}(\mathbf{0}))$ if $a < -1/2$.

7.7.6 Calculus rules

Most calculus rules in $H^m(\Omega)$ are formally similar to the classical ones, although some of their proofs are not so trivial. We list here a few of them.

- *Derivative of a product.* Let $u \in H^1(\Omega)$ and $v \in \mathcal{D}(\Omega)$. Then $uv \in H^1(\Omega)$ and

$$\nabla(uv) = u\nabla v + v\nabla u. \quad (7.50)$$

Formula (7.50) holds if both $u, v \in H^1(\Omega)$ as well. In this case, however,

$$uv \in L^1(\Omega) \quad \text{and} \quad \nabla(uv) \in L^1(\Omega; \mathbb{R}^n).$$

- *Change of variables.* Let $u \in H^1(\Omega)$ and $g : \Omega' \rightarrow \Omega$ be onto, one-to-one and Lipschitz, with g^{-1} Lipschitz. Then, the composition $u \circ g : \Omega' \rightarrow \mathbb{R}$ belongs to $H^1(\Omega')$ and

$$\partial_{x_i}[u \circ g](\mathbf{x}) = \sum_{k=1}^n \partial_{x_k} u(g(\mathbf{x})) \partial_{x_i} g_k(\mathbf{x}) \quad (7.51)$$

both a.e. in Ω and in $\mathcal{D}'(\Omega)$. In particular, the Lipschitz change of variables $\mathbf{y} = g(\mathbf{x})$ transforms $H^1(\Omega)$ into $H^1(\Omega')$.

- *Chain rule* (see Problem 7.30). Let $u \in H^1(\Omega)$ and $f : \mathbb{R} \rightarrow \mathbb{R}$ be Lipschitz. Assume that $f(0) = 0$ if Ω is unbounded. Then, the composition $f \circ u : \Omega \rightarrow \mathbb{R}$ belongs to $H^1(\Omega)$ and

$$\nabla[f \circ u] = f'(u)\nabla u \quad (7.52)$$

both a.e. in Ω and in $\mathcal{D}'(\Omega)$. If $f(0) = 0$ and $u \in H_0^1(\Omega)$ then $f \circ u \in H_0^1(\Omega)$.

In particular, choosing respectively

$$f(t) = |t|, \quad f(t) = \max\{t, 0\} \quad \text{and} \quad f(t) = -\min\{t, 0\},$$

we deduce the following result:

Proposition 7.68. *Let $u \in H^1(\Omega)$ (resp $H_0^1(\Omega)$). Then*

$$|u|, \quad u^+ = \max\{u, 0\} \quad \text{and} \quad u^- = -\min\{u, 0\}$$

all belong to $H^1(\Omega)$ (resp. to $H_0^1(\Omega)$). Moreover, the following formulas hold both a.e. in Ω and in $\mathcal{D}'(\Omega)$:

$$\nabla u^+ = \nabla u \chi_{\{u>0\}}, \quad \nabla u^- = \nabla u \chi_{\{u<0\}}$$

and

$$\nabla u = \nabla u^+ - \nabla u^-, \quad \nabla(|u|) = \text{sign}(u) \nabla u.$$

Finally, let Ω be a Lipschitz domain. If $u_k \rightharpoonup u$ (resp. $u_k \rightarrow u$) in $H^1(\Omega)$ then

$u_k^+ \rightharpoonup u^+$, $u_k^- \rightharpoonup u^-$, $|u_k| \rightharpoonup |u|$ (resp. $u_k^+ \rightarrow u^+$, $u_k^- \rightarrow u^-$, $|u_k| \rightarrow |u|$) in $H^1(\Omega)$.

Proof. The first part of the Proposition is a direct consequence of (7.52). We first prove the final assertions for $|u|$.

Let $u_k \rightarrow u$ in $H^1(\Omega)$. As we shall see later, by Rellich's Theorem 7.90, p. 487, the embedding of $H^1(\Omega)$ into $L^2(\Omega)$ is compact. Then, by Proposition 6.60, p. 397, $u_k \rightarrow u$ strongly in $L^2(\Omega)$. Since

$$\||u_k| - |u|\|_{L^2(\Omega)} \leq \|u_k - u\|_{L^2(\Omega)}$$

we see that $|u_k| \rightarrow |u|$ in $L^2(\Omega)$. On the other hand, $\{|u_k|\}$ is bounded in $H^1(\Omega)$ and, again by Rellich's Theorem, there exists a subsequence $\{|u_{k_j}|\}$ weakly convergent in $H^1(\Omega)$ and strongly in $L^2(\Omega)$. But then the limit must be $|u|$ and we conclude that the whole sequence $\{|u_k|\}$ converges weakly to $|u|$ in $H^1(\Omega)$.

Let now $u_k \rightarrow u$ in $H^1(\Omega)$. We have just seen that $\nabla(|u_k|) \rightharpoonup \nabla(|u|)$ in $L^2(\Omega)$. Moreover

$$|\nabla(|u_k|)| = |\nabla u_k| \rightarrow |\nabla u| = |\nabla(|u|)|$$

in $L^2(\Omega)$. We conclude that (see Problem 6.17) $\nabla(|u_k|) \rightarrow \nabla(|u|)$ in $L^2(\Omega)$ and therefore $|u_k| \rightarrow |u|$ in $H^1(\Omega)$.

The statements for u^+ and u^- follows from those for $|u|$ and the formulas

$$u^+ = \frac{u + |u|}{2} \text{ and } u^- = \frac{u - |u|}{2}.$$

□

Remark 7.69. As a consequence of Proposition 7.68, if $u \in H^1(\Omega)$ is constant in a measurable set $K \subseteq \Omega$, then $\nabla u = 0$ a.e. in K .

- *Derivative of a convolution.* Let $u \in H^1(\mathbb{R}^n)$ and $\varphi \in \mathcal{D}(\mathbb{R}^n)$. Then we know that $\varphi * u \in C^\infty(\mathbb{R}^n)$. We have:

$$\nabla(\varphi * u) = (\varphi * \nabla u). \quad (7.53)$$

Proof. Let $\mathbf{v} \in \mathcal{D}(\mathbb{R}^n, \mathbb{R}^n)$. We can write

$$\int_{\mathbb{R}^n} \nabla(\varphi * u) \cdot \mathbf{v} \, d\mathbf{x} = - \int_{\mathbb{R}^n} (\varphi * u) \operatorname{div} \mathbf{v} \, d\mathbf{x} = - \int_{\mathbb{R}^n} \left(\int_{\mathbb{R}^n} \varphi(\mathbf{x} - \mathbf{y}) u(\mathbf{y}) \, d\mathbf{y} \right) \operatorname{div} \mathbf{v}(\mathbf{x}) \, d\mathbf{x}.$$

On the other hand, integrating twice by parts, we get

$$\begin{aligned} \int_{\mathbb{R}^n} (\varphi * \nabla u) \cdot \mathbf{v} \, d\mathbf{x} &= \int_{\mathbb{R}^n} \left(\int_{\mathbb{R}^n} \varphi(\mathbf{x} - \mathbf{y}) \nabla u(\mathbf{y}) \, d\mathbf{y} \right) \cdot \mathbf{v}(\mathbf{x}) \, d\mathbf{x} \\ &= \int_{\mathbb{R}^n} \left(\int_{\mathbb{R}^n} \nabla_{\mathbf{y}} \varphi(\mathbf{x} - \mathbf{y}) u(\mathbf{y}) \, d\mathbf{y} \right) \cdot \mathbf{v}(\mathbf{x}) \, d\mathbf{x} \\ &= - \int_{\mathbb{R}^n} \nabla \left(\int_{\mathbb{R}^n} \varphi(\mathbf{x} - \mathbf{y}) u(\mathbf{y}) \, d\mathbf{y} \right) \cdot \mathbf{v}(\mathbf{x}) \, d\mathbf{x} \\ &= \int_{\mathbb{R}^n} \left(\int_{\mathbb{R}^n} \varphi(\mathbf{x} - \mathbf{y}) u(\mathbf{y}) \, d\mathbf{y} \right) \operatorname{div} \mathbf{v}(\mathbf{x}) \, d\mathbf{x}. \end{aligned}$$

Therefore

$$\int_{\mathbb{R}^n} \nabla(\varphi * u) \cdot \mathbf{v} \, d\mathbf{x} = \int_{\mathbb{R}^n} (\varphi * \nabla u) \cdot \mathbf{v} \, d\mathbf{x}$$

for every $\mathbf{v} \in \mathcal{D}(\mathbb{R}^n, \mathbb{R}^n)$ and (7.53) follows. \square

7.7.7 Fourier transform and Sobolev spaces

The spaces $H^m(\mathbb{R}^n)$, $m \geq 1$, may be defined in terms of the Fourier transform. In fact, by Theorem 7.54, p. 460,

$$u \in L^2(\mathbb{R}^n) \quad \text{if and only if} \quad \hat{u} \in L^2(\mathbb{R}^n)$$

and

$$\|u\|_{L^2(\mathbb{R}^n)}^2 = (2\pi)^{-n} \|\hat{u}\|_{L^2(\mathbb{R}^n)}^2.$$

It follows that, for every multi-index α with $|\alpha| \leq m$,

$$D^\alpha u \in L^2(\mathbb{R}^n) \quad \text{if and only if} \quad \xi^\alpha \hat{u} \in L^2(\mathbb{R}^n)$$

and

$$\|D^\alpha u\|_{L^2(\mathbb{R}^n)}^2 = (2\pi)^{-n} \|\xi^\alpha \hat{u}\|_{L^2(\mathbb{R}^n)}^2.$$

Observe now that for all α , $0 \leq |\alpha| \leq m$, there exists a positive constant C such that:

$$|\xi^\alpha|^2 \leq |\xi|^{2|\alpha|} \leq (1 + |\xi|^2)^m \in C \left(1 + \sum_{0 \leq |\alpha| \leq m} |\xi^\alpha|^2 \right), \quad \forall \xi \in \mathbb{R}^m.$$

The first two inequalities follow from $\alpha_j \leq |\alpha| \leq m$, while the third one from the binomial formula. Thus we obtain the following result.

Proposition 7.70. *Let $u \in L^2(\mathbb{R}^n)$. Then:*

- i) $u \in H^m(\mathbb{R}^n)$ if and only if $(1 + |\xi|^2)^{m/2} \hat{u} \in L^2(\mathbb{R}^n)$.
- ii) The norms

$$\|u\|_{H^m(\mathbb{R}^n)} \quad \text{and} \quad \|(1 + |\xi|^2)^{m/2} \hat{u}\|_{L^2(\mathbb{R}^n)}$$

are equivalent.

- *Sobolev spaces of real order.* The norm

$$\|u\|_{H^m(\mathbb{R}^n)} = \|(1 + |\xi|^2)^{m/2} \hat{u}\|_{L^2(\mathbb{R}^n)} \tag{7.54}$$

makes perfect sense even if m is not an integer and we are led to the following definition.

Definition 7.71. Let $s \in \mathbb{R}$, $0 < s < \infty$. We denote by $H^s(\mathbb{R}^n)$ the space of functions $u \in L^2(\mathbb{R}^n)$ such that $|\xi|^s \hat{u} \in L^2(\mathbb{R}^n)$.

Formally, saying that

$$|\xi_j|^s \hat{u} \in L^2(\mathbb{R}^n)$$

amounts to saying that a “derivative of order s ” of u belongs to $L^2(\mathbb{R}^n)$. Then

$$u \in H^s(\mathbb{R}^n)$$

if all the “derivatives of order s ” of u belong to $L^2(\mathbb{R}^n)$. We have:

Proposition 7.72. $H^s(\mathbb{R}^n)$ is a Hilbert space with inner product and norm given by

$$(u, v)_{H^s(\mathbb{R}^n)} = \int_{\mathbb{R}^n} \left(1 + |\xi|^2\right)^s \hat{u} \bar{\hat{v}} d\xi$$

and

$$\|u\|_{H^s(\mathbb{R}^n)} = \left\| \left(1 + |\xi|^2\right)^{s/2} \hat{u} \right\|_{L^2(\mathbb{R}^n)}.$$

The space $H^{1/2}(\mathbb{R}^n)$ of the L^2 -functions possessing “half derivatives” in $L^2(\mathbb{R}^n)$ plays an important role in Sect. 7.9.

7.8 Approximations by Smooth Functions and Extensions

7.8.1 Local approximations

The functions in $H^1(\Omega)$ may be quite irregular. However, using mollifiers, any $u \in H^1(\Omega)$ may be approximated *locally* by smooth functions, in the sense that the approximation holds in every compact subset of Ω .

Denote by $\eta_\varepsilon = \frac{1}{\varepsilon^n} \eta\left(\frac{x}{\varepsilon}\right)$ the mollifier introduced in Sect. 7.2 and by Ω_ε the set of points ε -away from $\partial\Omega$, i.e. (see Remark 7.3, p. 432):

$$\Omega_\varepsilon = \{x \in \Omega : \text{dist}(x, \partial\Omega) > \varepsilon\}.$$

For u defined in Ω , denote by \tilde{u} the zero extension of u outside Ω . We have:

Theorem 7.73. Let $u \in H^1(\Omega)$ and, for $\varepsilon > 0$, small, define

$$u_\varepsilon = \eta_\varepsilon * \tilde{u}.$$

Then, if $\varepsilon \rightarrow 0$,

1. $u_\varepsilon \in C^\infty(\mathbb{R}^n)$ and $u_\varepsilon \rightarrow u$ in $L^2(\Omega)$.
2. $\nabla u_\varepsilon \rightarrow \nabla u$ in $L^2(\Omega'; \mathbb{R}^n)$, for every $\Omega' \subset\subset \Omega$.

Proof. Property 1 follows from Lemma 7.2, **b** and **d**, p. 431. To prove 2, note first that, if $\varepsilon < \text{dist}(\Omega', \partial\Omega)$ and $\mathbf{x} \in \Omega'$, the function $\mathbf{y} \mapsto \eta_\varepsilon(\mathbf{x} - \mathbf{y})$ belongs to $\mathcal{D}(\Omega)$. Then, for every $j = 1, 2, \dots, n$, we have

$$\partial_{x_j} u_\varepsilon(\mathbf{x}) = \int_{\Omega} \partial_{x_j} \eta_\varepsilon(\mathbf{x} - \mathbf{y}) u(\mathbf{y}) d\mathbf{y} = - \int_{\Omega} \partial_{y_j} \eta_\varepsilon(\mathbf{x} - \mathbf{y}) u(\mathbf{y}) d\mathbf{y}$$

and finally, integrating by parts,

$$\partial_{x_j} u_\varepsilon(\mathbf{x}) = \int_{\Omega} \eta_\varepsilon(\mathbf{x} - \mathbf{y}) \partial_{y_j} u(\mathbf{y}) d\mathbf{y} = (\partial_{y_j} u)_\varepsilon(\mathbf{x}). \quad (7.55)$$

Then, 2 follows from property **d** of Lemma 7.2, applied to Ω' . \square

An almost immediate consequence of formula 7.55 is the following

Corollary 7.74. *If Ω is a domain and $\nabla u = \mathbf{0}$ a.e., then u is constant in Ω .*

7.8.2 Extensions and global approximations

By Theorem 7.73, we may approximate a function in $H^1(\Omega)$ by smooth functions, as long as we stay at positive distance from $\partial\Omega$. We wonder whether an approximation is possible in all $\overline{\Omega}$. First we give the following definition.

Definition 7.75. Denote by $\mathcal{D}(\overline{\Omega})$ the set of restrictions to $\overline{\Omega}$ of functions in $\mathcal{D}(\mathbb{R}^n)$.

Thus, $\varphi \in \mathcal{D}(\overline{\Omega})$ if there is $\psi \in \mathcal{D}(\mathbb{R}^n)$ such that $\varphi = \psi$ in $\overline{\Omega}$. Clearly, $\mathcal{D}(\overline{\Omega}) \subset C^\infty(\overline{\Omega})$. We want to establish whether

$$\mathcal{D}(\overline{\Omega}) \text{ is dense in } H^1(\Omega). \quad (7.56)$$

The case $\Omega = \mathbb{R}^n$ is special, since $\mathcal{D}(\Omega)$ coincides with $\mathcal{D}(\overline{\Omega})$. We have:

Theorem 7.76. $\mathcal{D}(\mathbb{R}^n)$ is dense in $H^1(\mathbb{R}^n)$. In particular $H^1(\mathbb{R}^n) = H_0^1(\mathbb{R}^n)$.

Proof. First observe that $H_c^1(\mathbb{R}^n)$, the subspace of functions with compact (essential) support in \mathbb{R}^n , is dense in $H^1(\mathbb{R}^n)$. In fact, let $u \in H^1(\mathbb{R}^n)$ and $v \in \mathcal{D}(\mathbb{R}^n)$, such that $0 \leq v \leq 1$ and $v \equiv 1$ if $|\mathbf{x}| \leq 1$. Define

$$u_s(\mathbf{x}) = v\left(\frac{\mathbf{x}}{s}\right) u(\mathbf{x}).$$

Then $u_s \in H_c^1(\mathbb{R}^n)$ and

$$\nabla u_s(\mathbf{x}) = v\left(\frac{\mathbf{x}}{s}\right) \nabla u(\mathbf{x}) + \frac{1}{s} u(\mathbf{x}) \nabla v\left(\frac{\mathbf{x}}{s}\right).$$

From the Dominated Convergence Theorem, it follows that¹⁸

$$u_s \rightarrow u \text{ in } H^1(\mathbb{R}^n), \quad \text{as } s \rightarrow \infty.$$

¹⁸ Observe that $|u_s| \leq |u|$ and $|\nabla u_s| \leq |\nabla u| + M|u|$, where $M = \max |\nabla v|$.

On the other hand $\mathcal{D}(\mathbb{R}^n)$ is dense in $H_c^1(\mathbb{R}^n)$. In fact, if $u \in H_c^1(\mathbb{R}^n)$, we have

$$u_\varepsilon = u * \eta_\varepsilon \in \mathcal{D}(\mathbb{R}^n)$$

and $u_\varepsilon \rightarrow u$ in $H^1(\mathbb{R}^n)$. \square

However, in general (7.56) is not true, as the following example shows.

Example 7.77. Consider, for instance,

$$\Omega = \{(\rho, \theta) : 0 < \rho < 1, 0 < \theta < 2\pi\}.$$

The domain Ω coincides with the open unit circle $B_1(0, 0)$, centered at the origin, without the radius

$$\{(\rho, \theta) : 0 < \rho < 1, \theta = 0\}.$$

The closure $\overline{\Omega}$ is given by the full closed circle. Let

$$u(\rho, \theta) = \rho^{1/2} \cos(\theta/2).$$

Then $u \in L^2(\Omega)$, since u is bounded. Moreover¹⁹,

$$|\nabla u|^2 = u_\rho^2 + \frac{1}{\rho^2} u_\theta^2 = \frac{1}{4\rho} \quad \text{in } \Omega,$$

so that $u \in H^1(\Omega)$. However, $u(\rho, 0+) = \rho^{1/2}$ while $u(\rho, 2\pi-) = -\rho^{1/2}$. Thus, u has a jump discontinuity across $\theta = 0$ and therefore no sequence of smooth functions in $\overline{\Omega}$ can converge to u in $H^1(\Omega)$.

The difficulty in Example 7.77 is that the domain Ω lies on both sides of part of its boundary (the radius $0 < \rho < 1, \theta = 0$). Thus, to have a hope that (7.56) is true, we have to avoid domains with this anomaly and consider domains with some degree of regularity.

Assume that Ω is a C^1 or even a Lipschitz domain. Theorem 7.76 suggests a strategy to prove (7.56): given $u \in H^1(\Omega)$, extend the definition of u to all \mathbb{R}^n in order to obtain a function in $H^1(\mathbb{R}^n)$ and then apply Theorem 7.76. The first thing to do is to introduce an *extension operator*:

Definition 7.78. We say that a linear operator $E : H^1(\Omega) \rightarrow H^1(\mathbb{R}^n)$ is an extension operator if, for all $u \in H^1(\Omega)$:

1. $E(u) = u$ in Ω .
2. If Ω is bounded, $E(u)$ is compactly supported.
3. E is continuous:

$$\|Eu\|_{H^1(\mathbb{R}^n)} \leq c(n, \Omega) \|u\|_{H^1(\Omega)}.$$

¹⁹ See Appendix C.

How do we construct E ? The first thing that comes into mind is to define $Eu = 0$ outside Ω (*trivial extension*). This certainly works if $u \in H_0^1(\Omega)$. In fact, it can be proved that $u \in H_0^1(\Omega)$ if and only if its trivial extension belongs to $H^1(\mathbb{R}^n)$.

However, the trivial extension works in this case only. For instance, let $u \in H^1(0, \infty)$ with $u(0) = a \neq 0$. Let Eu be the trivial extension of u . Then, in $\mathcal{D}'(\mathbb{R})$, $(Eu)' = u' + a\delta$ which is not even in $L^2(\mathbb{R})$.

Thus, we have to use another method. If Ω is a half space, that is

$$\Omega = \mathbb{R}_+^n = \{(x_1, \dots, x_n) : x_n > 0\}$$

an extension operator can be defined by a reflection method as follows:

- *Reflection method.* Let $u \in H^1(\mathbb{R}_+^n)$. Write $\mathbf{x} = (\mathbf{x}', x_n)$, $\mathbf{x}' \in \mathbb{R}^{n-1}$. We reflect in an even way with respect to the hyperplane $x_n = 0$, by setting $Eu = \tilde{u}$ where

$$\tilde{u}(\mathbf{x}) = u(\mathbf{x}', |x_n|).$$

Then, it is possible to prove that, in $\mathcal{D}'(\mathbb{R}^n)$:

$$\tilde{u}_{x_j}(\mathbf{x}) = \begin{cases} u_{x_j}(\mathbf{x}', |x_n|) & j < n \\ u_{x_n}(\mathbf{x}', |x_n|) \operatorname{sign} x_n & j = n. \end{cases} \quad (7.57)$$

It is now easy to check that E has the properties 1,2,3 listed above. In particular,

$$\|Eu\|_{H^1(\mathbb{R}^n)}^2 = 2\|u\|_{H^1(\mathbb{R}_+^n)}^2.$$

- *Extension operator for Lipschitz domains.* Suppose now that Ω is a bounded Lipschitz domain. To construct an extension operator we use two rather general ideas, which may be applied in several different contexts: *localization* and *reduction to the half space*.

Localization. It is based on the following lemma. Given a set K , by *open covering* of K we mean a collection \mathcal{O} of open sets, such that $K \subset \cup_{O \in \mathcal{O}} O$.

Lemma 7.79 (Partition of unity). *Let $K \subset \mathbb{R}^n$ be a compact set and O_1, \dots, O_N be an open covering of K . There exist functions ψ_1, \dots, ψ_N with the following properties:*

1. *For every $j = 1, \dots, N$, $\psi_j \in C_0^\infty(O_j)$ and $0 \leq \psi_j \leq 1$.*
2. *For every $\mathbf{x} \in K$, $\sum_{j=1}^N \psi_j(\mathbf{x}) = 1$.*

Proof (sketch). Since $K \subset \cup_{j=1}^N O_j$ and each O_j is open, we can find open sets $K_j \subset\subset O_j$ such that

$$K \subset \cup_{j=1}^N K_j.$$

Let χ_{K_j} be the characteristic function of K_j and η_ε the mollifier (7.3). Define $\varphi_{j,\varepsilon} = \eta_\varepsilon * \chi_{K_j}$. According to Example 7.3, p. 432, we may fix ε so small in order to have

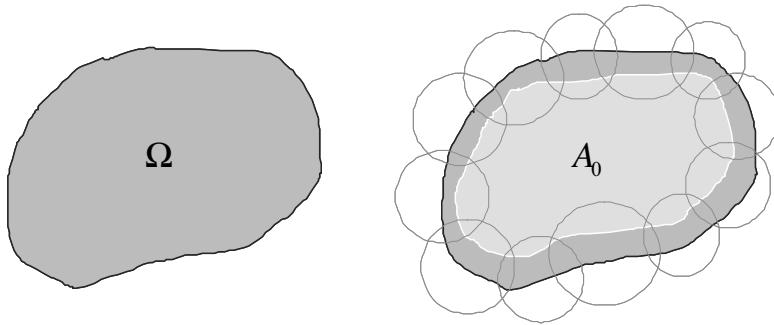


Fig. 7.5 A set Ω and an open covering of its closure

$\varphi_{j,\varepsilon} \in C_0^\infty(O_j)$ and $\varphi_{j,\varepsilon} > 0$ on K_j . Then the functions

$$\psi_j = \frac{\varphi_{j,\varepsilon}}{\sum_{s=1}^N \varphi_{s,\varepsilon}}$$

satisfy conditions **1** and **2**. □

The set of functions ψ_1, \dots, ψ_N is called *a partition of unity for K, associated with the covering O_1, \dots, O_N* . Now, if $u : K \rightarrow \mathbb{R}$, the localization procedure consists in writing

$$u = \sum_{j=1}^N \psi_j u \quad (7.58)$$

i.e. as a sum of functions $u_j = \psi_j u$ supported in O_j .

Reduction to a half space. Take an open covering of $\partial\Omega$ by N balls $B_j = B(\mathbf{x}_j)$, $j = 1, \dots, N$, centered at $\mathbf{x}_j \in \partial\Omega$ and such that $\partial\Omega \cap B_j$ is locally a graph of a Lipschitz function $y_n = \varphi_j(\mathbf{y}')$, $\mathbf{y}' \in \mathcal{N}_j$, where \mathcal{N}_j is a neighborhood of the origin in \mathbb{R}^{n-1} . This is possible, since $\partial\Omega$ is compact. Moreover, let $A_0 \subset \Omega$ be an open set containing $\Omega \setminus \cup_{j=1}^N B_j$ (Fig. 7.5).

Then, A_0, B_1, \dots, B_N is an open covering of $\overline{\Omega}$. Let $\psi_0, \psi_1, \dots, \psi_N$ be a partition of unity for $\overline{\Omega}$, associated with A_0, B_1, \dots, B_N .

For each B_j , $1 \leq j \leq N$, introduce the bi-Lipschitz transformation $\mathbf{z} = \Phi_j(\mathbf{x})$

$$\begin{cases} \mathbf{z}' = \mathbf{y}' \\ z_n = y_n - \varphi_j(\mathbf{y}') \end{cases} \quad (7.59)$$

Set (Fig. 7.6)

$$\Phi_j(B_j \cap \Omega) \equiv U_j \subset \mathbb{R}_+^n$$

and

$$\Phi_j(B_j \cap \partial\Omega) \equiv \Gamma_j \subset \partial\mathbb{R}_+^n = \{z_n = 0\}.$$

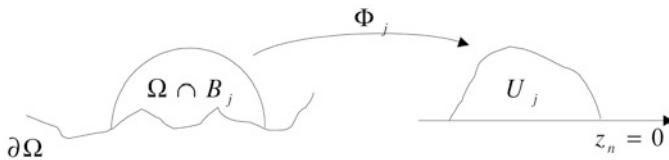


Fig. 7.6 The bi-Lipschitz transformation Φ_j flattens $B_j \cap \partial\Omega$

Let $u \in H^1(\Omega)$ and $u_j = \psi_j u$. Then, $w_j = u_j \circ \Phi_j^{-1}$ is supported in $U_j \cup \Gamma_j$, so that, by extending it to zero in $\mathbb{R}_+^n \setminus U_j$, we have $w_j \in H^1(\mathbb{R}_+^n)$.

The function $Ew_j = \tilde{w}_j$, obtained by the reflection method, belongs to $H^1(\mathbb{R}^n)$. Now we go back defining

$$Eu_j = \tilde{w}_j \circ \Phi_j, \quad 1 \leq j \leq N,$$

in B_j and $Eu_j = 0$ outside B_j . Finally, let $u_0 = \psi_0 u_0$ and let Eu_0 be the trivial extension of u_0 . Set

$$Eu = \sum_{j=0}^N Eu_j.$$

At this point, it is not difficult to show that E satisfies the requirements 1, 2, 3 of Definition 7.78. We have proved the following

Theorem 7.80. *Let Ω be either \mathbb{R}_+^n or a bounded, Lipschitz domain. Then, there exists an extension operator $E : H^1(\Omega) \rightarrow H^1(\mathbb{R}^n)$.*

An immediate consequence of Theorems 7.80 and 7.76, is the following global approximation result:

Theorem 7.81. *Let Ω be either \mathbb{R}_+^n or a bounded, Lipschitz domain. Then $D(\overline{\Omega})$ is dense in $H^1(\Omega)$. In other words, if $u \in H^1(\Omega)$, there exists a sequence $\{u_m\} \subset D(\overline{\Omega})$ such that*

$$\|u_m - u\|_{H^1(\Omega)} \rightarrow 0 \quad \text{as } m \rightarrow +\infty.$$

7.9 Traces

7.9.1 Traces of functions in $H^1(\Omega)$

The possibility of approximating any element $u \in H^1(\Omega)$ by smooth functions in $\overline{\Omega}$ represents a key tool for introducing the notion of *restriction of u on $\Gamma = \partial\Omega$* . Such restriction is called the **trace of u** on Γ and it will be an element of $L^2(\Gamma)$.

Observe that if $\Omega = \mathbb{R}_+^n$, then $\Gamma = \partial\mathbb{R}_+^n = \mathbb{R}^{n-1}$ and $L^2(\Gamma)$ is well defined. If Ω is a Lipschitz domain, we define $L^2(\Gamma)$ by localization. More precisely, let

B_1, \dots, B_N be an open covering of Γ by balls centered at points on Γ , as in Subsect. 7.8.2. If $g : \Gamma \rightarrow \mathbb{R}$, write

$$g = \sum_{j=1}^N \psi_j g$$

where ψ_1, \dots, ψ_N is a partition of unity for Γ , associated with B_1, \dots, B_N . Since $\Gamma \cap B_j$ is the graph of a Lipschitz function $y_n = \varphi_j(\mathbf{y}')$, $\mathbf{y}' \in \mathcal{N}_j$ (see p. 14), on $\Gamma \cap B_j$ there is a natural notion of “area element”, given by

$$d\sigma = \sqrt{1 + |\nabla \varphi_j|^2} d\mathbf{y}'.$$

Thus, for $1 \leq p < \infty$, we may define

$$\int_{\Gamma \cap B_j} \psi_j |g|^p d\sigma = \int_{\mathcal{N}_j} \psi_j (\varphi_j(\mathbf{y}')) |g(\varphi_j(\mathbf{y}'))|^p \sqrt{1 + |\nabla \varphi_j(\mathbf{y}')|^2} d\mathbf{y}'.$$

We say that $g \in L^p(\Gamma)$ if²⁰

$$\|g\|_{L^2(\Gamma)}^p = \int_{\Gamma} |g|^p d\sigma = \sum_{j=1}^N \int_{\Gamma \cap B_j} \psi_j |g|^p d\sigma < \infty. \quad (7.60)$$

With the norm (7.60) and the usual identification of functions if they are equal a.e. with respect to the surface measure $d\sigma$, $L^p(\Gamma)$ is a Banach space. $L^2(\Gamma)$ is a Hilbert space with respect to the inner product

$$(g, h)_{L^2(\Gamma)} = \sum_{j=1}^N \int_{\Gamma \cap B_j} \psi_j gh d\sigma.$$

Let us go back to our trace problem. We may consider $n > 1$, since there is no problem if $n = 1$. The strategy consists in the following two steps.

Let

$$\tau_0 : \mathcal{D}(\overline{\Omega}) \rightarrow L^2(\Gamma)$$

be the operator that associates to every function v its restriction $v|_{\Gamma}$ to Γ : $\tau_0 v = v|_{\Gamma}$. This makes perfect sense, since each $v \in \mathcal{D}(\overline{\Omega})$ is continuous on Γ .

First step. Show that $\|\tau_0 v\|_{L^2(\Gamma)} \leq c(\Omega, n) \|v\|_{H^1(\Omega)}$. Thus, τ_0 is continuous from $\mathcal{D}(\overline{\Omega}) \subset H^1(\Omega)$ into $L^2(\Gamma)$.

Second step. Extend τ_0 to all $H^1(\Omega)$ by using the density of $\mathcal{D}(\overline{\Omega})$ in $H^1(\Omega)$.

²⁰ Observe that the norm (7.60) depends on the particular covering and partition of unity. However, it can be shown that norms corresponding to different coverings and partitions of unity are all equivalent and induce the same topology on $L^2(\Gamma)$.

An elementary analogy may be useful. Suppose we have a function $f : \mathbb{Q} \rightarrow \mathbb{R}$ and we want to define the value of f at an irrational point x . What do we do? Since \mathbb{Q} is dense in \mathbb{R} , we select a sequence $\{r_k\} \subset \mathbb{Q}$ such that $r_k \rightarrow x$. Then we compute $f(r_k)$ and set $f(x) = \lim_{k \rightarrow \infty} f(r_k)$. Of course, we have to prove that the limit exists, by showing, for example, that $\{f(r_n)\}$ is a Cauchy sequence and that the limit does not depend on the approximating sequence $\{r_n\}$.

Theorem 7.82. *Let Ω be either \mathbb{R}_+^n or a bounded, Lipschitz domain. Then there exists a linear operator (trace operator) $\tau_0 : H^1(\Omega) \rightarrow L^2(\Gamma)$ such that:*

1. $\tau_0 u = u|_\Gamma$ if $u \in \mathcal{D}(\overline{\Omega})$.
2. $\|\tau_0 u\|_{L^2(\Gamma)} \leq c(\Omega, n) \|u\|_{H^1(\Omega)}$.

Proof. Let $\Omega = \mathbb{R}_+^n$. First, we prove inequality 2 for $u \in \mathcal{D}(\overline{\Omega})$. In this case $\tau_0 u = u(\mathbf{x}', 0)$ and we show that²¹

$$\int_{\mathbb{R}^{n-1}} |u(\mathbf{x}', 0)|^2 d\mathbf{x}' \leq 2 \|u\|_{H^1(\mathbb{R}_+^n)}^2, \quad \forall u \in \mathcal{D}(\overline{\Omega}). \quad (7.61)$$

For every $x_n \in (0, 1)$ we may write:

$$u^2(\mathbf{x}', 0) = u^2(\mathbf{x}', x_n) - \int_0^{x_n} \frac{d}{dt} u^2(\mathbf{x}', t) dt = u^2(\mathbf{x}', x_n) - 2 \int_0^{x_n} u(\mathbf{x}', t) u_{x_n}(\mathbf{x}', t) dt.$$

Since by Schwarz's inequality

$$\left(\int_0^1 |u(\mathbf{x}', t) u_{x_n}(\mathbf{x}', t)| dt \right)^2 \leq \int_0^1 u^2(\mathbf{x}', t) dt \int_0^1 u_{x_n}^2(\mathbf{x}', t) dt,$$

we deduce that (using the elementary inequality $2ab \leq a^2 + b^2$)

$$\begin{aligned} u^2(\mathbf{x}', 0) &\leq u^2(\mathbf{x}', x_n) + \int_0^1 u^2(\mathbf{x}', t) dt + \int_0^1 u_{x_n}^2(\mathbf{x}', t) dt \\ &\leq u^2(\mathbf{x}', x_n) + \int_0^1 u^2(\mathbf{x}', t) dt + \int_0^1 |\nabla u(\mathbf{x}', t)|^2 dt. \end{aligned}$$

Integrating both sides on \mathbb{R}^{n-1} with respect to \mathbf{x}' and on $(0, 1)$ with respect to x_n , we easily obtain (7.61).

Assume now $u \in H^1(\mathbb{R}_+^n)$. Since $\mathcal{D}(\overline{\Omega})$ is dense in $H^1(\mathbb{R}_+^n)$, we can select $\{u_k\} \subset \mathcal{D}(\overline{\Omega})$ such that $u_k \rightarrow u$ in $H^1(\mathbb{R}_+^n)$.

The linearity of τ_0 and estimate (7.61) yield

$$\|\tau_0 u_h - \tau_0 u_k\|_{L^2(\mathbb{R}^{n-1})} \leq \sqrt{2} \|u_h - u_k\|_{H^1(\mathbb{R}_+^n)}.$$

Since $\{u_k\}$ is a Cauchy sequence in $H^1(\mathbb{R}_+^n)$, we infer that $\{\tau_0 u_k\}$ is a Cauchy sequence in $L^2(\mathbb{R}^{n-1})$. Therefore, there exists $u_0 \in L^2(\mathbb{R}^{n-1})$ such that

$$\tau_0 u_k \rightarrow u_0 \quad \text{in } L^2(\mathbb{R}^{n-1}).$$

²¹ A more precise estimate is given in Problem 7.31.

The limiting element u_0 does not depend on the approximating sequence $\{u_k\}$. In fact, if $\{v_k\} \subset \mathcal{D}(\overline{\Omega})$ and $v_k \rightarrow u$ in $H^1(\mathbb{R}_+^n)$, then

$$\|v_k - u_k\|_{H^1(\mathbb{R}_+^n)} \rightarrow 0.$$

From

$$\|\tau_0 v_k - \tau_0 u_k\|_{L^2(\mathbb{R}^{n-1})} \leq \sqrt{2} \|v_k - u_k\|_{H^1(\mathbb{R}_+^n)}$$

it follows that $\tau_0 v_k \rightarrow u_0$ in $L^2(\mathbb{R}^{n-1})$ as well.

Thus, if $u \in H^1(\mathbb{R}_+^n)$, it makes sense to define $\tau_0 u = u_0$. It should be clear that τ_0 has the properties 1, 2.

If Ω is a bounded Lipschitz domain, the theorem can be proved once more by localization and reduction to a half space. We omit the details. \square

Definition 7.83. *The function $\tau_0 u$ is called the trace of u on Γ .*

The following integration by parts formula for functions in $H^1(\Omega)$ is a consequence of the trace Theorem 7.82.

Corollary 7.84. *Assume Ω is either \mathbb{R}_+^n or a bounded, Lipschitz domain. Let $u \in H^1(\Omega)$ and $\mathbf{v} \in H^1(\Omega; \mathbb{R}^n)$. Then*

$$\int_{\Omega} \nabla u \cdot \mathbf{v} \, d\mathbf{x} = - \int_{\Omega} u \operatorname{div} \mathbf{v} \, d\mathbf{x} + \int_{\Gamma} (\tau_0 u) (\tau_0 \mathbf{v}) \cdot \boldsymbol{\nu} \, d\sigma, \quad (7.62)$$

where $\boldsymbol{\nu}$ is the outward unit normal to Γ and $\tau_0 \mathbf{v} = (\tau_0 v_1, \dots, \tau_0 v_n)$.

Proof. Formula (7.62) holds if $u \in \mathcal{D}(\overline{\Omega})$ and $\mathbf{v} \in \mathcal{D}(\overline{\Omega}; \mathbb{R}^n)$. Let $u \in H^1(\Omega)$ and $\mathbf{v} \in H^1(\Omega; \mathbb{R}^n)$. Select $\{u_k\} \subset \mathcal{D}(\overline{\Omega})$, $\{\mathbf{v}_k\} \subset \mathcal{D}(\overline{\Omega}; \mathbb{R}^n)$ such that $u_k \rightarrow u$ in $H^1(\Omega)$ and $\mathbf{v}_k \rightarrow \mathbf{v}$ in $H^1(\Omega; \mathbb{R}^n)$. Then:

$$\int_{\Omega} \nabla u_k \cdot \mathbf{v}_k \, d\mathbf{x} = - \int_{\Omega} u_k \operatorname{div} \mathbf{v}_k \, d\mathbf{x} + \int_{\Gamma} (\tau_0 u_k) (\tau_0 \mathbf{v}_k) \cdot \boldsymbol{\nu} \, d\sigma.$$

Letting $k \rightarrow \infty$, by the continuity of τ_0 , we obtain (7.62). \square

It is not surprising that the kernel of τ_0 is precisely $H_0^1(\Omega)$:

$$\tau_0 u = 0 \iff u \in H_0^1(\Omega).$$

However, only the proof of the “ \Leftarrow ” part is trivial. The proof of the “ \Rightarrow ” part is rather technical and we omit it.

For the trace $\tau_0 u$ we also use the notation $u|_{\Gamma}$ and, if there is no risk of confusion, inside a boundary integral, we will write simply

$$\int_{\partial\Omega} u \, d\sigma.$$

In a similar way, we may define the trace of $u \in H^1(\Omega)$ on a relatively open subset $\Gamma_0 \subset \Gamma$, regular in the sense of Definition 1.5, p. 14.

Theorem 7.85. Assume Ω is either \mathbb{R}_+^n or a bounded, Lipschitz domain. Let Γ_0 be an open subset of Γ . Then there exists a trace operator $\tau_{\Gamma_0} : H^1(\Omega) \rightarrow L^2(\Gamma_0)$ such that:

1. $\tau_{\Gamma_0} u = u|_{\Gamma_0}$ if $u \in \mathcal{D}(\overline{\Omega})$.
2. $\|\tau_{\Gamma_0} u\|_{L^2(\Gamma_0)} \leq c(\Omega, n) \|u\|_{H^1(\Omega)}$.

The function $\tau_{\Gamma_0} u$ is called the *trace of u on Γ_0* , often denoted by $u|_{\Gamma_0}$. The kernel of τ_{Γ_0} is denoted by $H_{0,\Gamma_0}^1(\Omega)$:

$$\tau_{\Gamma_0} u = 0 \iff u \in H_{0,\Gamma_0}^1(\Omega).$$

This space can be characterized in another way. Let V_{0,Γ_0} be the set of functions in $\mathcal{D}(\overline{\Omega})$ vanishing in a neighborhood of Γ_0 . Then:

Proposition 7.86. $H_{0,\Gamma_0}^1(\Omega)$ is the closure of V_{0,Γ_0} in $H^1(\Omega)$.

7.9.2 Traces of functions in $H^m(\Omega)$

We have seen that a function $u \in H^m(\mathbb{R}_+^n)$, $m \geq 1$, has a trace on $\Gamma = \partial\mathbb{R}_+^n$. However, if $m = 2$, every derivative of u belongs to $H^1(\mathbb{R}_+^n)$, so that it has a trace on Γ . In particular, we may define the trace of $\partial_{x_n} u$ on Γ . Let

$$\tau_1 u = (\partial_{x_n} u)|_{\Gamma}.$$

In general, for $m \geq 2$, we may define the trace on Γ of the derivatives $\partial_{x_n}^j u = \frac{\partial^j u}{\partial x_n^j}$ for $j = 0, 1, \dots, m-1$ and set

$$\tau_j u = (\partial_{x_n}^j u)|_{\Gamma}.$$

In this way, we construct a linear operator $\tau : H^m(\mathbb{R}_+^n) \rightarrow L^2(\Gamma; \mathbb{R}^m)$, given by

$$\tau u = (\tau_0 u, \dots, \tau_{m-1} u).$$

From Theorem 7.82, τ satisfies the following conditions:

1. $\tau u = (u|_{\Gamma}, (\partial_{x_n} u)|_{\Gamma}, \dots, (\partial_{x_n}^{m-1} u)|_{\Gamma})$, if $u \in \mathcal{D}(\overline{\mathbb{R}_+^n})$.
2. $\|\tau u\|_{L^2(\Gamma; \mathbb{R}^m)} \leq c \|u\|_{H^m(\mathbb{R}_+^n)}$.

The operator τ associates to $u \in H^m(\mathbb{R}_+^n)$ the trace on Γ of u and its derivatives up to the order $m-1$, in the direction x_n . This direction corresponds to the *interior normal* to $\Gamma = \partial\mathbb{R}_+^n$.

Analogously, for a bounded domain Ω we may define the trace on Γ of the (interior or exterior) normal derivatives of u , up to order $m-1$. This requires Ω to be at least a C^m -domain. The following theorem holds, where ν denotes the exterior unit normal to $\partial\Omega$.

Theorem 7.87. Assume Ω is either \mathbb{R}_+^n or a bounded, C^m -domain, $m \geq 2$. Then there exists a trace operator $\tau : H^m(\Omega) \rightarrow L^2(\Gamma; \mathbb{R}^m)$ such that:

1. $\tau u = (u|_\Gamma, \frac{\partial u}{\partial \nu}|_\Gamma, \dots, \frac{\partial^{m-1} u}{\partial \nu^{m-1}}|_\Gamma)$ if $u \in \mathcal{D}(\overline{\Omega})$.
2. $\|\tau u\|_{L^2(\Gamma; \mathbb{R}^m)} \leq c(\Omega, n) \|u\|_{H^m(\Omega)}$.

Similarly, we may define a trace of the (interior or exterior) normal derivatives of u , up to order $m - 1$, on an open regular subset $\Gamma_0 \subset \Gamma$.

It turns out that the kernel of the operator τ is given by the *closure of* $D(\Omega)$ in $H^m(\Omega)$, denoted by $H_0^m(\Omega)$. Precisely:

$$\tau u = (0, \dots, 0) \iff u \in H_0^m(\Omega).$$

Clearly, $H_0^m(\Omega)$ is a Hilbert subspace of $H^m(\Omega)$. If $u \in H_0^m(\Omega)$, u and its normal derivatives up to order $m - 1$ have zero trace on Γ .

7.9.3 Trace spaces

The operator $\tau_0 : H^1(\Omega) \rightarrow L^2(\Gamma)$ is **not** surjective. In fact the image of τ_0 is *strictly contained in* $L^2(\Gamma)$. In other words, there are functions in $L^2(\Gamma)$ which are *not* traces of functions in $H^1(\Omega)$. So, the natural question is: which functions $g \in L^2(\Gamma)$ are traces of functions in $H^1(\Omega)$? The answer is not elementary: roughly speaking, we could characterize them as *functions possessing half derivatives in* $L^2(\Gamma)$. It is as if in the restriction to the boundary, a function of $H^1(\Omega)$ loses “one half of each derivative”. To give an idea of what this means, let us consider the case $\Omega = \mathbb{R}_+^n$. We have:

Theorem 7.88. $\text{Im } \tau_0 = H^{1/2}(\mathbb{R}^{n-1})$.

Proof First we show that $\text{Im } \tau_0 \subseteq H^{1/2}(\mathbb{R}^{n-1})$. Let $u \in H^1(\mathbb{R}_+^n)$ and extend it to all \mathbb{R}^n by even reflection with respect to the plane $x_n = 0$. We write the points in \mathbb{R}^n as $\mathbf{x} = (\mathbf{x}', x_n)$, with $\mathbf{x}' = (x_1, \dots, x_{n-1})$. Define $g(\mathbf{x}') = u(\mathbf{x}', 0)$. We show that $g \in H^{1/2}(\mathbb{R}^{n-1})$, that is

$$\|g\|_{H^{1/2}(\mathbb{R}^{n-1})}^2 = \int_{\mathbb{R}^{n-1}} (1 + |\boldsymbol{\xi}'|^2)^{1/2} |\widehat{g}(\boldsymbol{\xi}')|^2 d\boldsymbol{\xi}' < \infty.$$

First, we consider $u \in \mathcal{D}(\mathbb{R}^n)$ and express \widehat{g} in terms of \widehat{u} . By the Fourier inversion formula, we may write

$$u(\mathbf{x}', x_n) = \frac{1}{(2\pi)^n} \int_{\mathbb{R}^{n-1}} e^{i\mathbf{x}' \cdot \boldsymbol{\xi}'} \left(\int_{\mathbb{R}} \widehat{u}(\boldsymbol{\xi}', \xi_n) e^{ix_n \xi_n} d\xi_n \right) d\boldsymbol{\xi}'$$

so that

$$g(\mathbf{x}') = \frac{1}{(2\pi)^{n-1}} \int_{\mathbb{R}^{n-1}} e^{i\mathbf{x}' \cdot \boldsymbol{\xi}'} \left(\frac{1}{2\pi} \int_{\mathbb{R}} \widehat{u}(\boldsymbol{\xi}', \xi_n) d\xi_n \right) d\boldsymbol{\xi}'.$$

This shows that

$$\widehat{g}(\boldsymbol{\xi}') = \frac{1}{2\pi} \int_{\mathbb{R}} \widehat{u}(\boldsymbol{\xi}', \xi_n) d\xi_n.$$

Thus:

$$\|g\|_{H^{1/2}(\mathbb{R}^{n-1})}^2 = \frac{1}{(2\pi)^2} \int_{\mathbb{R}^{n-1}} (1 + |\xi'|^2)^{1/2} \left| \int_{\mathbb{R}} \hat{u}(\xi', \xi_n) d\xi_n \right|^2 d\xi'.$$

Note now the following two facts. First, from Schwarz's inequality, we may write

$$\begin{aligned} \left| \int_{\mathbb{R}} \hat{u}(\xi', \xi_n) d\xi_n \right| &\leq \int_{\mathbb{R}} (1 + |\xi'|^2)^{-1/2} (1 + |\xi'|^2)^{1/2} |\hat{u}(\xi', \xi_n)| d\xi_n \\ &\leq \left(\int_{\mathbb{R}} (1 + |\xi'|^2) |\hat{u}(\xi', \xi_n)|^2 d\xi_n \right)^{1/2} \left(\int_{\mathbb{R}} (1 + |\xi'|^2)^{-1} d\xi_n \right)^{1/2}. \end{aligned}$$

Second,²²

$$\int_{\mathbb{R}} (1 + |\xi'|^2)^{-1} d\xi_n = \int_{\mathbb{R}} (1 + |\xi'|^2 + \xi_n^2)^{-1} d\xi_n = \frac{\pi}{(1 + |\xi'|^2)^{1/2}}.$$

Thus,

$$\|g\|_{H^{1/2}(\mathbb{R}^{n-1})}^2 \leq \frac{1}{4\pi} \int_{\mathbb{R}^n} (1 + |\xi'|^2) |\hat{u}(\xi)|^2 d\xi = \frac{1}{4\pi} \|u\|_{H^1(\mathbb{R}^n)}^2 = \frac{1}{2\pi} \|u\|_{H^1(\mathbb{R}_+^n)}^2 < \infty.$$

Therefore, $g \in H^{1/2}(\mathbb{R}^{n-1})$. By the usual density argument, this is true for every $u \in H^1(\mathbb{R}_+^n)$ and shows that $\text{Im } \tau_0 \subseteq H^{1/2}(\mathbb{R}^{n-1})$.

To prove the opposite inclusion, take any $g \in H^{1/2}(\mathbb{R}^{n-1})$ and define

$$u(x', x_n) = \frac{1}{(2\pi)^{n-1}} \int_{\mathbb{R}^{n-1}} e^{-(1+|\xi'|^2)^{1/2} x_n} \hat{g}(\xi') e^{i\xi' \cdot \xi'} d\xi', \quad x_n \geq 0.$$

Then, clearly $u(x', 0) = g(x')$ and from (7.40),

$$\begin{aligned} \|u\|_{H^1(\mathbb{R}_+^n)}^2 &= \frac{1}{(2\pi)^{n-1}} \int_0^{+\infty} \int_{\mathbb{R}^{n-1}} (1 + |\xi'|^2) e^{-2(1+|\xi'|^2)^{1/2} x_n} |\hat{g}(\xi')|^2 d\xi' dx_n \\ &= \frac{1}{2^n \pi^{n-1}} \int_{\mathbb{R}^{n-1}} (1 + |\xi'|^2)^{1/2} |\hat{g}(\xi')|^2 d\xi' = \frac{1}{2^n \pi^{n-1}} \|g\|_{H^{1/2}(\mathbb{R}^{n-1})}^2. \end{aligned}$$

Therefore, $g \in \text{Im } \tau_0$ so that $H^{1/2}(\mathbb{R}^{n-1}) \subseteq \text{Im } \tau_0$. □

If Ω is a bounded, Lipschitz domain, it is possible to define $H^{1/2}(\Gamma)$ by localization and reduction to the half space, as we did for $L^2(\Gamma)$. In this way we can endow $H^{1/2}(\Gamma)$ with an inner product that makes it a Hilbert space, continuously embedded in $L^2(\Gamma)$. It turns out that $H^{1/2}(\Gamma)$ coincides with $\text{Im } \tau_0$:

$$H^{1/2}(\Gamma) = \{\tau_0 u : u \in H^1(\Omega)\}. \quad (7.63)$$

Actually, slightly changing our point of view, we take (7.63) as a definition of $H^{1/2}(\Gamma)$ and endow $H^{1/2}(\Gamma)$ with the norm

$$\|g\|_{H^{1/2}(\Gamma)} = \inf \left\{ \|w\|_{H^1(\Omega)} : w \in H^1(\Omega), \tau_0 w = g \right\}. \quad (7.64)$$

²² $\int_{\mathbb{R}} (a^2 + t^2)^{-1} dt = \left[\frac{1}{a} \arctan \left(\frac{t}{a} \right) \right]_{-\infty}^{+\infty} = \frac{\pi}{a}, \quad (a > 0).$

This norm is equal to the smallest among the norms of all elements in $H^1(\Omega)$ sharing the same trace g on Γ and takes into account that the trace operator τ_0 is not injective, since we know that $\text{Ker } \tau_0 = H_0^1(\Omega)$. In particular, the following *trace inequality holds*:

$$\|\tau_0 u\|_{H^{1/2}(\Gamma)} \leq \|u\|_{H^1(\Omega)}, \quad (7.65)$$

which means that the trace operator τ_0 is continuous from $H^1(\Omega)$ onto $H^{1/2}(\Gamma)$.

In a similar way, if Γ_0 is a relatively open and regular subset of Γ , we may define

$$H^{1/2}(\Gamma_0) = \{\tau_{\Gamma_0} u : u \in H^1(\Omega)\}, \quad (7.66)$$

endowed with the norm

$$\|g\|_{H^{1/2}(\Gamma_0)} = \inf \left\{ \|w\|_{H^1(\Omega)} : w \in H^1(\Omega), \tau_{\Gamma_0} w = g \right\}.$$

In particular, the following *trace inequality holds*:

$$\|\tau_{\Gamma_0} u\|_{H^{1/2}(\Gamma_0)} \leq \|u\|_{H^1(\Omega)}. \quad (7.67)$$

which means that the trace operator τ_{Γ_0} is continuous from $H^1(\Omega)$ onto $H^{1/2}(\Gamma_0)$.

Summarizing, we have:

Theorem 7.89. *The spaces $H^{1/2}(\Gamma)$ and $H^{1/2}(\Gamma_0)$ defined respectively by (7.63) and (7.66) are Hilbert spaces continuously embedded into $L^2(\Gamma)$ and $L^2(\Gamma_0)$.*

Proof. Consider $H^{1/2}(\Gamma)$. Let $H^*(\Omega)$ be the orthogonal complement to $H_0^1(\Omega)$ in $H^1(\Omega)$, that is

$$H^1(\Omega) = H_0^1(\Omega) \oplus H^*(\Omega).$$

Then $H^*(\Omega)$ is a Hilbert space. Take the restriction τ_0^* of τ_0 to $H^*(\Omega)$. Clearly τ_0^* is an isomorphism between $H^*(\Omega)$ and $H^{1/2}(\Gamma)$. Moreover τ_0^* is norm-preserving. To see it note that if $u \in H^*(\Omega)$ and $\tau_0^* u = g$, the counter-images of $\tau_0^{-1} g$ are of the form $w = u + v$ with $v \in H_0^1(\Omega)$. Since

$$\|u + v\|_{H^1(\Omega)}^2 = \|u\|_{H^1(\Omega)}^2 + \|v\|_{H^1(\Omega)}^2$$

we have

$$\|g\|_{H^{1/2}(\Gamma)} = \inf \left\{ \|w\|_{H^1(\Omega)} : w \in H^1(\Omega), \tau_0 w = g \right\} = \|u\|_{H^1(\Omega)}^2.$$

Thus τ_0^* is an isometry between $H^*(\Omega)$ and $H^{1/2}(\Gamma)$ and this implies that $H^{1/2}(\Gamma)$ is a Hilbert space with inner product

$$(g, h)_{H^{1/2}(\Gamma)} = ((\tau_0^*)^{-1} g, (\tau_0^*)^{-1} h)_{H^1(\Omega)}.$$

The same arguments work for $H^{1/2}(\Gamma_0)$. □

Finally, if Ω is \mathbb{R}_+^n or a bounded C^m domain, $m \geq 2$, the space of traces of functions in $H^m(\Omega)$ is the fractional order Sobolev space $H^{m-1/2}(\Gamma)$, still showing a loss of “half derivative”. Coherently, the trace of a normal derivative undergoes a loss of one more derivative and belongs to $H^{m-3/2}(\Gamma)$; the derivatives of order

$m - 1$ have traces in $H^{1/2}(\Gamma)$. Thus we obtain:

$$\tau : H^m(\Omega) \rightarrow \left(H^{m-1/2}(\Gamma), H^{m-3/2}(\Gamma), \dots, H^{1/2}(\Gamma) \right).$$

The kernel of τ is $H_0^m(\Omega)$.

7.10 Compactness and Embeddings

7.10.1 Rellich's theorem

Since

$$\|u\|_{L^2(\Omega)} \leq \|u\|_{H^1(\Omega)},$$

$H^1(\Omega)$ is *continuously embedded* in $L^2(\Omega)$ that is, if a sequence $\{u_k\}$ converges to u in $H^1(\Omega)$ it converges to u in $L^2(\Omega)$ as well.

If we assume that Ω is a **bounded**, **Lipschitz** domain, then the embedding of $H^1(\Omega)$ in $L^2(\Omega)$ is also **compact** (see also Problem 7.27). Thus, a bounded sequence $\{u_k\} \subset H^1(\Omega)$ has the following important property:

There exists a subsequence $\{u_{k_j}\}$ and $u \in H^1(\Omega)$, such that

- a. $u_{k_j} \rightarrow u$ in $L^2(\Omega)$.
- b. $u_{k_j} \rightharpoonup u$ in $H^1(\Omega)$ (i.e. u_{k_j} converges weakly²³ to u in $H^1(\Omega)$).

Actually, only property **a** follows from the compactness of the embedding. Property **b** expresses a general fact in every Hilbert space H : every bounded subset $E \subset H$ is *sequentially weakly compact* (Theorem 6.57, p. 395).

Theorem 7.90. *Let Ω be a bounded, Lipschitz domain. Then $H^1(\Omega)$ is compactly embedded in $L^2(\Omega)$.*

Proof. We use the compactness criterion expressed in Theorem 6.52, p. 393. First, observe that, for every $v \in \mathcal{D}(\mathbb{R}^n)$ we may write

$$v(\mathbf{x} + \mathbf{h}) - v(\mathbf{x}) = \int_0^1 \frac{d}{dt} v(\mathbf{x} + t\mathbf{h}) dt = \int_0^1 \nabla v(\mathbf{x} + t\mathbf{h}) \cdot \mathbf{h} dt$$

whence

$$|v(\mathbf{x} + \mathbf{h}) - v(\mathbf{x})|^2 = \left| \int_0^1 \nabla v(\mathbf{x} + t\mathbf{h}) \cdot \mathbf{h} dt \right|^2 \leq |\mathbf{h}|^2 \int_0^1 |\nabla v(\mathbf{x} + t\mathbf{h})|^2 dt.$$

Integrating on \mathbb{R}^n we find

$$\int_{\mathbb{R}^n} |v(\mathbf{x} + \mathbf{h}) - v(\mathbf{x})|^2 d\mathbf{x} \leq |\mathbf{h}|^2 \int_{\mathbb{R}^n} d\mathbf{x} \int_0^1 |\nabla v(\mathbf{x} + t\mathbf{h})|^2 dt \leq |\mathbf{h}|^2 \|\nabla v\|_{L^2(\mathbb{R}^n; \mathbb{R}^n)}^2$$

²³ See Sect. 6.7.

so that

$$\int_{\mathbb{R}^n} |v(\mathbf{x} + \mathbf{h}) - v(\mathbf{x})|^2 d\mathbf{x} \leq |\mathbf{h}|^2 \|\nabla v\|_{L^2(\mathbb{R}^n; \mathbb{R}^n)}^2. \quad (7.68)$$

Since $\mathcal{D}(\mathbb{R}^n)$ is dense in $H^1(\mathbb{R}^n)$, we infer that (7.68) holds for every $u \in H^1(\mathbb{R}^n)$ as well.

Let now $S \subset H^1(\Omega)$ be bounded, i.e. there exists a number M such that:

$$\|u\|_{H^1(\Omega)} \leq M, \quad \forall u \in S.$$

By Theorem 7.80, p. 479, every $u \in S$ has an extension $\tilde{u} \in H^1(\mathbb{R}^n)$, with support contained in an open set $\Omega' \supset \Omega$. Thus, $\tilde{u} \in H_0^1(\Omega')$ and moreover,

$$\|\nabla \tilde{u}\|_{L^2(\Omega'; \mathbb{R}^n)} \leq c \|\nabla u\|_{L^2(\Omega; \mathbb{R}^n)} \leq cM.$$

Denote by \tilde{S} the set of such extensions. Then (7.68) holds for every $\tilde{u} \in \tilde{S}$:

$$\int_{\Omega'} |\tilde{u}(\mathbf{x} + \mathbf{h}) - \tilde{u}(\mathbf{x})|^2 d\mathbf{x} \leq |\mathbf{h}|^2 \|\nabla \tilde{u}\|_{L^2(\mathbb{R}^n; \mathbb{R}^n)}^2 \leq c^2 M^2 |\mathbf{h}|^2.$$

Theorem 6.52, p. 393, implies that \tilde{S} is precompact in $L^2(\Omega')$, which implies that S is precompact in $L^2(\Omega)$. \square

7.10.2 Poincaré's inequalities

Under suitable hypotheses, the norm $\|u\|_{H^1(\Omega)}$ is equivalent to $\|\nabla u\|_{L^2(\Omega; \mathbb{R}^n)}$. This means that there exists a constant C_P , independent of u , such that

$$\|u\|_{L^2(\Omega)} \leq C_P \|\nabla u\|_{L^2(\Omega; \mathbb{R}^n)}. \quad (7.69)$$

Inequalities like (7.69) are called *Poincaré's inequalities* and play a big role in the variational treatment of boundary value problems, as we shall realize in the next chapter. We have already proved in Theorem 7.61, p. 466, that (7.69) holds if $u \in H_0^1(\Omega)$, that is for functions vanishing on $\Gamma = \partial\Omega$.

On the other hand, (7.69) cannot hold if $u = \text{constant} \neq 0$. Roughly speaking, the hypotheses that guarantee the validity of (7.69) require that u vanishes in some “nontrivial set”. For instance, if u belongs to one of the following two subspaces of $H^1(\Omega)$, then (7.69) holds:

1. $u \in H_{0,\Gamma_0}^1(\Omega)$ (u vanishes on a nonempty relatively open subset $\Gamma_0 \subset \Gamma$) or, more generally, $\int_{\Gamma_0} u = 0$.
2. $u \in H^1(\Omega)$ and $\int_E u = 0$ where $E \subseteq \Omega$ has positive measure, that is, $|E| = a > 0$. Note that, if $E = \Omega$, then u has mean value zero in Ω .

Theorem 7.91. *Let Ω be a bounded Lipschitz domain. Then, there exists a constant C_P such that*

$$\|u\|_{L^2(\Omega)} \leq C_P \|\nabla u\|_{L^2(\Omega; \mathbb{R}^n)} \quad (7.70)$$

for every u in $H^1(\Omega)$, belonging to one of the subspaces 1 or 2 above.

Proof. Assume $u \in H_{0,\Gamma_0}^1(\Omega)$. By contradiction suppose (7.70) is not true. This means that for every integer $j \geq 1$, there exists $u_j \in H_{0,\Gamma_0}^1(\Omega)$ such that

$$\|u_j\|_{L^2(\Omega)} > j \|\nabla u_j\|_{L^2(\Omega;\mathbb{R}^n)}. \quad (7.71)$$

Normalize u_j in $L^2(\Omega)$ by setting

$$w_j = \frac{u_j}{\|u_j\|_{L^2(\Omega)}}.$$

Then, from (7.71),

$$\|w_j\|_{L^2(\Omega)} = 1 \quad \text{and} \quad \|\nabla w_j\|_{L^2(\Omega;\mathbb{R}^n)} < \frac{1}{j} \leq 1.$$

Thus $\{w_j\}$ is bounded in $H^1(\Omega)$ and by Rellich's Theorem there exists a subsequence $\{w_{j_k}\}$ and $w \in H_{0,\Gamma_0}^1(\Omega)$ such that

- $w_{j_k} \rightarrow w$ in $L^2(\Omega)$.
- $\nabla w_{j_k} \rightarrow \nabla w$ in $L^2(\Omega;\mathbb{R}^n)$.

The continuity of the norm gives

$$\|w\|_{L^2(\Omega)} = \lim_{j_k \rightarrow +\infty} \|w_{j_k}\|_{L^2(\Omega)} = 1.$$

On the other hand, the weak sequential lower semicontinuity of the norm (Theorem 6.56, p. 395) yields,

$$\|\nabla w\|_{L^2(\Omega;\mathbb{R}^n)} \leq \liminf_{j_k \rightarrow \infty} \|\nabla w_{j_k}\|_{L^2(\Omega;\mathbb{R}^n)} = 0$$

so that $\nabla w = \mathbf{0}$ a.e. Since Ω is connected, w is constant by Corollary 7.74, p. 475, and since $w \in H_{0,\Gamma_0}^1(\Omega)$, we infer $w = 0$, in contradiction to $\|w\|_{L^2(\Omega)} = 1$.

The proof in the other cases is identical. \square

Remark 7.92. Let Ω be a bounded Lipschitz domain and $E \subset \Omega$, with $|E| = a > 0$. If $u \in H^1(\Omega)$, set

$$u_E = \frac{1}{|E|} \int_E u \quad (7.72)$$

and $w = u - u_E$. Then $\int_E w = 0$ and (7.70) holds for w . Thus, we get a Poincaré's inequality of the form:

$$\|u - u_E\|_{L^2(\Omega)} \leq C_P \|\nabla u\|_{L^2(\Omega;\mathbb{R}^n)}. \quad (7.73)$$

A similar inequality holds if E is replaced by nonempty relatively open subset $\Gamma_0 \subseteq \Gamma$. As a consequence, the following norms are all equivalent to the standard norm in $H^1(\Omega)$:

$$\sqrt{\int_{\Omega} |\nabla u|^2 + (u_E)^2}, \quad \sqrt{\int_{\Omega} |\nabla u|^2 + (u^2)_E} \quad (7.74)$$

or

$$\sqrt{\int_{\Omega} |\nabla u|^2 + (u_{\Gamma_0})^2}, \quad \sqrt{\int_{\Omega} |\nabla u|^2 + (u^2)_{\Gamma_0}}. \quad (7.75)$$

Let us prove, for instance, that

$$c \left\{ \int_{\Omega} |\nabla u|^2 + (u_E)^2 \right\} \leq \|u\|_{H^{1,2}(\Omega)}^2 \leq C \left\{ \int_{\Omega} |\nabla u|^2 + (u_E)^2 \right\} \quad (7.76)$$

where the constants c, C depends only on n, Ω and E . We have, from (7.73), using the inequality $2ab \leq \varepsilon^{-1}a^2 + \varepsilon b^2$,

$$\begin{aligned} \int_{\Omega} u^2 &\leq C_p^2 \int_{\Omega} |\nabla u|^2 + 2u_E \int_{\Omega} u \leq C_p^2 \int_{\Omega} |\nabla u|^2 + 2u_E |\Omega|^{1/2} \left(\int_{\Omega} u^2 \right)^{1/2} \\ &\leq C_p^2 \int_{\Omega} |\nabla u|^2 + \frac{|\Omega|}{\varepsilon} u_E^2 + \varepsilon \int_{\Omega} u^2. \end{aligned}$$

Choosing $\varepsilon = \frac{1}{2}$ we have

$$\int_{\Omega} u^2 \leq 2C_p^2 \int_{\Omega} |\nabla u|^2 + 4|\Omega| u_E^2$$

from which the inequality in the right hand side of (7.76) follows with $C(n, \Omega, E) = \max\{2C_p^2 + 1, 4|\Omega|\}$. The inequality on the left of (7.76) is easier and we leave it as an exercise. \square

7.10.3 Sobolev inequality in \mathbb{R}^n

From Proposition 7.59, p. 465, we know that the elements of $H^1(\mathbb{R})$ are continuous and (Problem 7.25) vanish as $x \rightarrow \pm\infty$. Moreover, if $u \in \mathcal{D}(\mathbb{R})$, we may write

$$u^2(x) = \int_{-\infty}^x \frac{d}{ds} u^2(s) ds = 2 \int_{-\infty}^x u(s) u'(s) ds.$$

Using Schwarz's inequality and $2ab \leq a^2 + b^2$, we get

$$u(x)^2 \leq 2 \|u\|_{L^2(\mathbb{R})} \|u'\|_{L^2(\mathbb{R})} \leq \|u\|_{L^2(\mathbb{R})}^2 + \|u'\|_{L^2(\mathbb{R})}^2 = \|u\|_{H^1(\mathbb{R})}^2.$$

Since $\mathcal{D}(\mathbb{R})$ is dense in $H^1(\mathbb{R})$, this inequality holds if $u \in H^1(\mathbb{R})$ as well. We have proved:

Proposition 7.93. *Let $u \in H^1(\mathbb{R})$. Then $u \in L^\infty(\mathbb{R})$ and*

$$\|u\|_{L^\infty(\mathbb{R})} \leq \|u\|_{H^1(\mathbb{R})}.$$

On the other hand, Example 7.58, p. 463, implies that, when $\Omega \subseteq \mathbb{R}^n$, $n \geq 2$,

$$u \in H^1(\Omega) \not\Rightarrow u \in L^\infty(\Omega).$$

However, it is possible to prove that u is actually p -summable with a suitable $p > 2$. Moreover, the L^p -norm of u can be estimated by the H^1 -norm of u .

To guess which exponent p is the correct one, assume that the inequality

$$\|u\|_{L^p(\mathbb{R}^n)} \leq c \|\nabla u\|_{L^2(\mathbb{R}^n)} \quad (7.77)$$

is valid **for every** $u \in \mathcal{D}(\mathbb{R}^n)$, where c may depend on p and n **but not on** u . We now use a typical “dimensional analysis” argument.

Inequality (7.77) must be *invariant under dilations* in the following sense. Let $u \in \mathcal{D}(\mathbb{R}^n)$ and for $\lambda > 0$ set

$$u_\lambda(\mathbf{x}) = u(\lambda\mathbf{x}).$$

Then $u_\lambda \in \mathcal{D}(\mathbb{R}^n)$ so that inequality (7.77) must be true for u_λ , with c **independent of** λ :

$$\|u_\lambda\|_{L^p(\mathbb{R}^n)} \leq c \|\nabla u_\lambda\|_{L^2(\mathbb{R}^n; \mathbb{R}^n)}. \quad (7.78)$$

Now,

$$\int_{\mathbb{R}^n} |u_\lambda|^p d\mathbf{x} = \int_{\mathbb{R}^n} |u(\lambda\mathbf{x})|^p d\mathbf{x} = \frac{1}{\lambda^n} \int_{\mathbb{R}^n} |u(\mathbf{y})|^p d\mathbf{y}$$

while

$$\int_{\mathbb{R}^n} |\nabla u_\lambda(\mathbf{x})|^2 d\mathbf{x} = \frac{1}{\lambda^{n-2}} \int_{\mathbb{R}^n} |\nabla u(\mathbf{y})|^2 d\mathbf{y}.$$

Therefore, (7.78) becomes

$$\frac{1}{\lambda^{n/p}} \left(\int_{\mathbb{R}^n} |u|^p d\mathbf{y} \right)^{1/p} \leq c(n, p) \frac{1}{\lambda^{(n-2)/2}} \left(\int_{\mathbb{R}^n} |\nabla u|^2 d\mathbf{y} \right)^{1/2}$$

or

$$\|u\|_{L^p(\mathbb{R}^n)} \leq c \lambda^{1 - \frac{n}{2} + \frac{n}{p}} \|\nabla u\|_{L^2(\mathbb{R}^n; \mathbb{R}^n)}.$$

The only way to get a constant independent of λ is to choose p such that

$$1 - \frac{n}{2} + \frac{n}{p} = 0.$$

Hence, if $n > 2$, the correct p is given by

$$p^* = \frac{2n}{n-2}$$

which is called the *Sobolev exponent* for $H^1(\mathbb{R}^n)$. The following theorem of *Sobolev, Gagliardo, Nirenberg* holds:

Theorem 7.94. *Let $u \in H^1(\mathbb{R}^n)$, $n \geq 3$. Then $u \in L^{p^*}(\mathbb{R}^n)$ with $p^* = \frac{2n}{n-2}$, and the following inequality holds.*

$$\|u\|_{L^{p^*}(\mathbb{R}^n)} \leq c \|\nabla u\|_{L^2(\mathbb{R}^n, \mathbb{R}^n)} \quad (7.79)$$

where $c = c(n)$.

In the case $n = 2$, the correct statement is:

Proposition 7.95. *Let $u \in H^1(\mathbb{R}^2)$. Then $u \in L^p(\mathbb{R})$ for $2 \leq p < \infty$, and*

$$\|u\|_{L^p(\mathbb{R}^2)} \leq c(p) \|u\|_{H^1(\mathbb{R}^2)}.$$

7.10.4 Bounded domains

We now consider bounded domains. In dimension $n = 1$, the elements of $H^1(a, b)$ are continuous in $[a, b]$ and therefore bounded as well. Furthermore, the following inequality holds:

$$\|v\|_{L^\infty(a,b)} \leq C^* \|v\|_{H^1(a,b)} \quad (7.80)$$

with

$$C^* = \sqrt{2} \max \left\{ (b-a)^{-1/2}, (b-a)^{1/2} \right\}.$$

Indeed, by Schwarz's inequality we have, for any $x, y \in [a, b]$, $y > x$:

$$u(y) = u(x) + \int_x^y u'(s) ds \leq u(x) + \sqrt{b-a} \|u'\|_{L^2(a,b)}$$

whence, using the elementary inequality $(A+B)^2 \leq 2A^2 + 2B^2$,

$$u(y)^2 \leq 2u(x)^2 + 2(b-a) \|u'\|_{L^2(a,b)}^2.$$

Integrating over (a, b) with respect to x we get

$$(b-a) u(y)^2 \leq 2 \|u\|_{L^2(a,b)}^2 + 2(b-a)^2 \|u'\|_{L^2(a,b)}^2$$

from which (7.80) follows easily.

In dimension $n \geq 2$, the improvement in summability is indicated in the following theorem.

Theorem 7.96. *Let Ω be a bounded, Lipschitz domain. Then:*

1. *If $n > 2$, $H^1(\Omega) \hookrightarrow L^p(\Omega)$ for $2 \leq p \leq \frac{2n}{n-2}$. Moreover, if $2 \leq p < \frac{2n}{n-2}$, the embedding of $H^1(\Omega)$ in $L^p(\Omega)$ is compact.*
2. *If $n = 2$, $H^1(\Omega) \hookrightarrow L^p(\Omega)$ for $2 \leq p < \infty$, with compact embedding.*

In the above cases

$$\|u\|_{L^p(\Omega)} \leq c(n, p, \Omega) \|u\|_{H^1(\Omega)}.$$

For instance, in the important case $n = 3$, we have

$$p^* = \frac{2n}{n-2} = 6.$$

Hence $H^1(\Omega) \hookrightarrow L^6(\Omega)$ and

$$\|u\|_{L^6(\Omega)} \leq c(\Omega) \|u\|_{H^1(\Omega)}.$$

When the embedding of $H^1(\Omega)$ in $L^p(\Omega)$ is compact, the Poincaré inequality in Theorem 7.91, p. 488 may be replaced by (see Problem 7.26)

$$\|u\|_{L^p(\Omega)} \leq c(n, p, \Omega) \|\nabla u\|_{L^2(\Omega; \mathbb{R}^n)}. \quad (7.81)$$

Theorem 7.96 shows what we can conclude about a H^1 -function with regards to further regularity. It is natural to expect something more for H^m -functions, with $m > 1$. In fact, the following result holds.

Theorem 7.97. *Let Ω be a bounded, Lipschitz domain, and $m > n/2$. Then*

$$H^m(\Omega) \hookrightarrow C^{k,\alpha}(\overline{\Omega}), \quad (7.82)$$

for k integer, $0 \leq k < m - \frac{n}{2}$ and $\alpha = 1/2$ if n is odd, $\alpha \in (0, 1)$ if n is even. The embedding in (7.82) is compact and

$$\|u\|_{C^{k,\alpha}(\overline{\Omega})} \leq c(n, m, \alpha, \Omega) \|u\|_{H^m(\Omega)}.$$

Theorem 7.97 implies that, in dimension $n = 2$, two derivatives occurs to get (Hölder) continuity:

$$H^2(\Omega) \hookrightarrow C^{0,\alpha}(\overline{\Omega}), \quad 0 < \alpha < 1.$$

In fact, n is even, $m = 2$, $m - \frac{n}{2} = 1$. Similarly

$$H^3(\Omega) \hookrightarrow C^{1,\alpha}(\overline{\Omega}), \quad 0 < \alpha < 1,$$

since $m - \frac{n}{2} = 2$. In dimension $n = 3$ we have

$$H^2(\Omega) \subset C^{0,1/2}(\overline{\Omega})$$

and

$$H^3(\Omega) \subset C^{1,1/2}(\overline{\Omega}),$$

since $m - \frac{n}{2} = \frac{1}{2}$ in the first case and $m - \frac{n}{2} = \frac{3}{2}$ in the second.

Remark 7.98. If $u \in H^m(\Omega)$ for any $m \geq 1$, then $u \in C^\infty(\overline{\Omega})$. This kind of results is very useful in the regularity theory for boundary value problems.

7.11 Spaces Involving Time

7.11.1 Functions with values into Hilbert spaces

The natural functional setting for evolution problems requires spaces which involve time. Given a function $u = u(\mathbf{x}, t)$, it is often convenient to separate the roles of space and time adopting the following point of view. Assume that $t \in [0, T]$ and that for every t , or at least for a.e. t , the function $u(\cdot, t)$ belongs to a *separable* Hilbert space V (e.g. $L^2(\Omega)$ or $H^1(\Omega)$).

Then, we may consider u as a function of the real variable t with values into V :

$$u : [0, T] \rightarrow V.$$

When we adopt this convention, we write $u(t)$ and $\dot{u}(t)$ instead of $u(\mathbf{x}, t)$ and $u_t(\mathbf{x}, t)$.

- *Spaces of continuous functions.* We start by introducing the set $C([0, T]; V)$ of the continuous functions $u : [0, T] \rightarrow V$, endowed with the norm

$$\|u\|_{C([0, T]; V)} = \max_{0 \leq t \leq T} \|u(t)\|_V.$$

Clearly $C([0, T]; V)$ is Banach space. We say that $v : [0, T] \rightarrow V$ is the (strong) derivative of u if

$$\left\| \frac{u(t+h) - u(t)}{h} - v(t) \right\|_V \rightarrow 0$$

for each $t \in [0, T]$. We write as usual u' or \dot{u} for the derivative of u .

The symbol $C^1([0, T]; V)$ denotes the Banach space of functions whose derivative exists and belongs to $C([0, T]; V)$, endowed with the norm

$$\|u\|_{C^1([0, T]; V)} = \|u\|_{C([0, T]; V)} + \|\dot{u}\|_{C([0, T]; V)}$$

Similarly we can define the spaces $C^k([0, T]; V)$ and $C^\infty([0, T]; V)$, while $\mathcal{D}(0, T; V)$ denotes the subspace of $C^\infty([0, T]; V)$ of the functions compactly supported in $(0, T)$.

- *Integrals and spaces of summable or integrable functions.* We can extend to these types of functions the notions of measurability and integral, without too much effort, following more or less the procedure outlined in Appendix B. First, we introduce the set of functions $s : [0, T] \rightarrow V$ which assume only a finite number of values. These functions are called *simple* and are of the form

$$s(t) = \sum_{j=1}^N \chi_{E_j}(t) u_j \quad (0 \leq t \leq T), \tag{7.83}$$

where, $u_1, \dots, u_N \in V$ and E_1, \dots, E_N are Lebesgue measurable, mutually disjoint subsets of $[0, T]$.

We say that $f : [0, T] \rightarrow V$ is *measurable* if there exists a sequence of simple functions $s_k : [0, T] \rightarrow V$ such that, as $k \rightarrow \infty$,

$$\|s_k(t) - f(t)\|_V \rightarrow 0, \quad \text{a.e. in } [0, T].$$

It is not difficult to prove that, if f is measurable and $v \in V$, the (real) function $t \mapsto (f(t), v)_V$ is Lebesgue measurable in $[0, T]$.

The notion of integral is defined first for simple functions. If s is given by (7.83), we define

$$\int_0^T s(t) dt = \sum_{j=1}^N |E_j| u_j.$$

Then:

Definition 7.99. We say that $f : [0, T] \rightarrow V$ is summable in $[0, T]$ if it is measurable and there exists a sequence $s_k : [0, T] \rightarrow V$ of simple functions such that

$$\int_0^T \|s_k(t) - f(t)\|_V dt \rightarrow 0, \quad \text{as } k \rightarrow +\infty. \quad (7.84)$$

If f is summable in $[0, T]$, we define the integral of f as follows:

$$\int_0^T f(t) dt = \lim_{k \rightarrow +\infty} \int_0^T s_k(t) dt, \quad \text{as } k \rightarrow +\infty. \quad (7.85)$$

Since for simple function it is easy to check that

$$\begin{aligned} \left\| \int_0^T [s_h(t) - s_k(t)] dt \right\|_V &\leq \int_0^T \|s_h(t) - s_k(t)\|_V dt \\ &\leq \int_0^T \|s_h(t) - f(t)\|_V dt + \int_0^T \|s_k(t) - f(t)\|_V dt, \end{aligned}$$

it follows from (7.84) that the sequence $\left\{ \int_0^T s_k(t) dt \right\}$ is a Cauchy sequence in V , so that the limit (7.85) is well defined and does not depend on the choice of the approximating sequence $\{s_k\}$. Moreover, the following important theorem holds, due to Bochner:

Theorem 7.100. A measurable function $f : [0, T] \rightarrow V$ is summable in $[0, T]$ if and only if the real function $t \mapsto \|f(t)\|_V$ is summable in $[0, T]$. Moreover

$$\left\| \int_0^T f(t) dt \right\|_V \leq \int_0^T \|f(t)\|_V dt \quad (7.86)$$

and

$$\left(u, \int_0^T f(t) dt \right)_V = \int_0^T (u, f(t))_V dt, \quad \forall u \in V. \quad (7.87)$$

The inequality (7.86) is well known in the case of real or complex functions. By Riesz's Representation Theorem, (7.87) shows that the action of any element of V^* commutes with the integral.

Once the definition of integral has been given, we can introduce the spaces $L^p(0, T; V)$, $1 \leq p \leq \infty$.

We define $L^p(0, T; V)$ as the set of measurable functions $u : [0, T] \rightarrow V$ such that:

if $1 \leq p < \infty$

$$\|u\|_{L^p(0, T; V)} = \left(\int_0^T \|u(t)\|_V^p dt \right)^{1/p} < \infty \quad (7.88)$$

while if $p = \infty$

$$\|u\|_{L^\infty(0, T; V)} = \operatorname{ess} \sup_{0 \leq t \leq T} \|u(t)\|_V < \infty.$$

Endowed with the above norms, $L^p(0, T; V)$ becomes a Banach space for $1 \leq p \leq \infty$. If $p = 2$, the norm (7.88) is induced by the inner product

$$(u, v)_{L^2(0, T; V)} = \int_0^T (u(t), v(t))_V dt$$

that makes $L^2(0, T; V)$ into a Hilbert space.

If $u \in L^1(0, T; V)$, the function $t \mapsto \int_0^t u(s) ds$ is continuous and

$$\frac{d}{dt} \int_0^t u(s) ds = u(t) \quad \text{a.e. in } (0, T).$$

Proposition 7.101. *Let $1 \leq p < \infty$. Then:*

- a) $\mathcal{D}(0, T; V)$ is dense in $L^p(0, T; V)$.
- b) $(L^p(0, T; V))^*$ is isometric to and can be identified with $L^q(0, T; V^*)$, $q = p/(p-1)$.

We conclude this subsection with a useful result (see Problem 7.32): the weak convergence in $L^2(0, T; V)$ preserves boundedness in $L^\infty(0, T; V)$.

Proposition 7.102. *Let $\{u_k\} \subset L^2(0, T; V)$, weakly convergent to u . Assume that*

$$\sup_{t \in [0, T]} \|u_k(t)\|_V \leq C$$

with C independent of k . Then, also,

$$\sup_{t \in [0, T]} \|u(t)\|_V \leq C.$$

7.11.2 Sobolev spaces involving time

To define *Sobolev spaces*, we need the notion of derivative in the sense of distributions for functions $u \in L^1_{loc}(0, T; V)$. In general, the weak derivative \dot{u} is defined by the linear operator

$$\langle \dot{u}, \varphi \rangle = - \int_0^T u(t) \dot{\varphi}(t) dt \quad (7.89)$$

from $\mathcal{D}(0, T)$ into V . As usual, the crochet $\langle \cdot, \cdot \rangle$ denotes the *action* of the distribution \dot{u} on the test function φ . Indeed, we have

$$\|\langle \dot{u}, \varphi \rangle\|_V = \left\| \int_0^T u(t) \dot{\varphi}(t) dt \right\|_V \leq \int_0^T |\dot{\varphi}(t)| \|u(t)\|_V dt \leq K \max |\dot{\varphi}(t)|$$

where K is the integral of $\|u(t)\|_V$ over the support of $\dot{\varphi}$. Therefore \dot{u} is continuous with respect to the convergence in $\mathcal{D}(0, T)$ and defines an element of $\mathcal{D}'(0, T; V)$.

Thus, we say that $\dot{u} \in L^1_{loc}(0, T; V)$ is the *derivative in the sense of distribution* of u if

$$\langle \dot{u}, \varphi \rangle = \int_0^T \varphi(t) \dot{u}(t) dt = - \int_0^T \dot{\varphi}(t) u(t) dt \quad (7.90)$$

for every $\varphi \in \mathcal{D}(0, T)$ or, equivalently, if

$$\int_0^T \varphi(t) (\dot{u}(t), v)_V dt = - \int_0^T \dot{\varphi}(t) (u(t), v)_V dt, \quad \forall v \in V. \quad (7.91)$$

We denote by $H^1(0, T; V)$ the *Sobolev space of the functions* $u \in L^2(0, T; V)$ such that $\dot{u} \in L^2(0, T; V)$. This is a Hilbert space with inner product

$$(u, v)_{H^1(0, T; V)} = \int_0^T \{(u(t), v(t))_V + (\dot{u}(t), \dot{v}(t))_V\} dt.$$

Since functions in $H^1(a, b)$ are continuous in $[a, b]$, it makes sense to consider the value of u at any point of $[a, b]$. In a certain way, the functions in $H^1(0, T; V)$ depends only on the real variable t , so that the following theorem is not surprising.

Theorem 7.103. *Let $u \in H^1(0, T; V)$. Then, $u \in C([0, T]; V)$ and*

$$\max_{0 \leq t \leq T} \|u(t)\|_V \leq C(T) \|u\|_{H^1(0, T; V)}.$$

Moreover, the fundamental theorem of calculus holds:

$$u(t) = u(s) + \int_s^t \dot{u}(r) dr, \quad 0 \leq s \leq t \leq T.$$

If V and W are separable Hilbert spaces with

$$V \hookrightarrow W$$

it makes sense to define the Sobolev space

$$H^1(0, T; V, W) = \{u \in L^2(0, T; V) : \dot{u} \in L^2(0, T; W)\},$$

where \dot{u} is intended in the sense of distribution in $\mathcal{D}'(0, T; W)$, endowed with the norm

$$\|u\|_{H^1(0, T; V, W)}^2 = \|u\|_{L^2(0, T; V)}^2 + \|\dot{u}\|_{L^2(0, T; W)}^2.$$

Indeed, the linear functional

$$\langle \dot{u}, \varphi \rangle = \int_0^T \varphi(t) \dot{u}(t) dt = - \int_0^T \dot{\varphi}(t) u(t) dt$$

defines a distribution $\dot{u} \in \mathcal{D}'(0, T; W)$. It is enough to observe that, since $V \hookrightarrow W$, we have, for some constant C ,

$$\|\langle \dot{u}, \varphi \rangle\|_W \leq \max |\dot{\varphi}(t)| \int_0^T \|u(t)\|_W dt \leq C \max |\dot{\varphi}(t)| \sqrt{T} \left(\int_0^T \|u(t)\|_V^2 dt \right)^{1/2}$$

and therefore \dot{u} is continuous with respect to the convergence in $\mathcal{D}(0, T)$.

This situation typically occurs in the applications to initial-boundary value problems. In fact, given a Hilbert triplet $\{V, H, V^*\}$, with V and H separable, we shall see in Chap. 10 that the natural functional setting for those problems is precisely the space $H^1(0, T; V, V^*)$. The following result is fundamental²⁴.

Theorem 7.104. *Let $\{V, H, V^*\}$ be a Hilbert triplet, with V and H separable. Then:*

- a) $C^\infty([0, T]; V)$ is dense in $H^1(0, T; V, V^*)$.
- b) $H^1(0, T; V, V^*) \hookrightarrow C([0, T]; H)$, that is

$$\|u\|_{C([0, T]; H)} \leq C(T) \|u\|_{H^1(0, T; V, V^*)}.$$

- c) If $u, v \in H^1(0, T; V, V^*)$, the following integration by parts formula holds:

$$\int_s^t \{\langle \dot{u}(r), v(r) \rangle_* + \langle \dot{v}(r), u(r) \rangle_*\} dr = (u(t), v(t))_H - (u(s), v(s))_H$$

for all $s, t \in [0, T]$.

²⁴ For the proof, see [24], *Dautray-Lions*, vol. 5, 1985.

Remark 7.105. From the integration by parts formula we infer,

$$\frac{d}{dt} (u(t), v(t))_H = \langle \dot{u}(t), v(t) \rangle_* + \langle \dot{v}(t), u(t) \rangle_*$$

a.e. $t \in [0, T]$ and (letting $u = v$)

$$\int_s^t \frac{d}{dt} \|u(r)\|_H^2 dt = \|u(t)\|_H^2 - \|u(s)\|_H^2. \quad (7.92)$$

Problems

7.1. A distribution $F \in \mathcal{D}'(\Omega)$ is said to be positive if $\langle F, \varphi \rangle \geq 0$ for every $\varphi \in \mathcal{D}(\Omega)$, $\varphi \geq 0$ in Ω . Show that if F is positive, then it is a measure.

[Hint: Let $\varphi \in \mathcal{D}(\Omega)$, $\varphi \geq 0$, with support(φ) = K . Let $\psi \in \mathcal{D}(\Omega)$ such that $\psi \geq 0$ and $\psi \equiv 1$ on K . Observe that $\psi \|\varphi\|_{L^\infty(\Omega)} \pm \varphi \geq 0$ and prove that (7.6), p. 434, holds].

7.2. Approximations of δ_n .

(a) Let $B_r = B_r(\mathbf{0}) \subset \mathbb{R}^n$. Show that, if χ_{B_r} is the characteristic function of B_r ,

$$\lim_{r \rightarrow 0} \frac{1}{|B_r|} \chi_{B_r} = \delta_n, \quad \text{in } \mathcal{D}'(\mathbb{R}^n).$$

(b) Let η_ε be the mollifier in (7.3). Show that $\lim_{\varepsilon \rightarrow 0} \eta_\varepsilon = \delta_n$, in $\mathcal{D}'(\mathbb{R}^n)$.

(c) Let $\Gamma_D(\mathbf{x}, t)$ be the fundamental solution of the heat equation $u_t = D\Delta u$. Show that, as $t \rightarrow 0^+$,

$$\Gamma_D(\cdot, t) \rightarrow \delta_n, \quad \text{in } \mathcal{D}'(\mathbb{R}^n).$$

(d) Let $f \in L^1(\mathbb{R}^n)$ with the following properties:

$$f \geq 0, \quad f(-\mathbf{x}) = f(\mathbf{x}), \quad \int_{\mathbb{R}^n} f = 1.$$

Define $f_k(\mathbf{x}) = k^n f(k\mathbf{x})$. Show that

$$\lim_{k \rightarrow +\infty} f_k = \delta_n, \quad \text{in } \mathcal{D}'(\mathbb{R}^n).$$

7.3. Show that the series

$$\sum_{k=1}^{\infty} c_k \sin kx$$

converges in $\mathcal{D}'(\mathbb{R})$ if the numerical sequence $\{c_k\}$ is *slowly increasing*, i.e. if there exists $p \in \mathbb{R}$ such that $|c_k| \leq k^p$ as $k \rightarrow \infty$.

7.4. Show that if $F \in \mathcal{D}'(\mathbb{R}^n)$, $v \in \mathcal{D}(\mathbb{R}^n)$ and v vanishes in an open set containing the support of F , then $\langle F, v \rangle = 0$. Is it true that $\langle F, v \rangle = 0$ if v vanishes *only* on the support of F ?

7.5. Let $u(x) = |x|$ and $S(x) = \operatorname{sign}(x)$. Prove that $u' = S$ in $\mathcal{D}'(\mathbb{R})$.

7.6. Prove that $x^m \delta^{(k)} = 0$ in $\mathcal{D}'(\mathbb{R})$, if $0 \leq k < m$.

7.7. Let $u(x) = \ln|x|$. Then, $u' = p.v.\frac{1}{x}$, in $\mathcal{D}'(\mathbb{R})$.

[Hint: Write

$$\langle u', \varphi \rangle = -\langle u, \varphi' \rangle = - \int_{\mathbb{R}} \ln|x| \varphi'(x) dx = - \lim_{\varepsilon \rightarrow 0} \int_{\{|x|>\varepsilon\}} \ln|x| \varphi'(x) dx$$

and integrate by parts].

7.8. Let $n = 3$ and $\mathbf{F} \in \mathcal{D}'(\Omega; \mathbb{R}^3)$. Define $\operatorname{curl} \mathbf{F} \in \mathcal{D}'(\Omega; \mathbb{R}^3)$ by the formula

$$\operatorname{curl} \mathbf{F} = (\partial_{x_2} F_3 - \partial_{x_3} F_2, \partial_{x_3} F_1 - \partial_{x_1} F_3, \partial_{x_1} F_2 - \partial_{x_2} F_1).$$

Check that, for all $\boldsymbol{\varphi} \in \mathcal{D}(\Omega; \mathbb{R}^3)$,

$$\langle \operatorname{curl} \mathbf{F}, \boldsymbol{\varphi} \rangle = \langle \mathbf{F}, \operatorname{curl} \boldsymbol{\varphi} \rangle.$$

7.9. Show that if $u(x_1, x_2) = \frac{1}{2\pi} \ln(x_1^2 + x_2^2)$, then

$$-\Delta u = \delta_2, \quad \text{in } \mathcal{D}'(\mathbb{R}^2).$$

7.10. Show that if $u_k \rightarrow u$ in $L^p(\mathbb{R}^n)$ then $u_k \rightarrow u$ in $\mathcal{S}'(\mathbb{R}^n)$.

7.11. Solve the equation $x^2 G = 0$ in $\mathcal{D}'(\mathbb{R})$.

7.12. Let $u \in C^\infty(\mathbb{R})$ with compact support in $[0, 1]$. Compute $\operatorname{comb} * u$.

[Answer: $\sum_{k=-\infty}^{+\infty} u(x - k)$, the periodic extension of u over \mathbb{R}].

7.13. Let $\mathcal{H} = \mathcal{H}(x)$ be the Heaviside function. Prove that

$$a) \mathcal{F}[\operatorname{sign}(x)] = \frac{2}{i} p.v.\frac{1}{\xi}, \quad b) \mathcal{F}[\mathcal{H}] = \pi\delta + \frac{1}{i} p.v.\frac{1}{\xi}.$$

[Hint: a) Let $u(x) = \operatorname{sign}(x)$. Note that $u' = 2\delta$. Transform this equation to obtain $\xi\hat{u}(\xi) = -2i$. Solve this equation using formula (7.27), and recall that \hat{u} is odd while δ is even. b) Write $\mathcal{H}(x) = \frac{1}{2} + \frac{1}{2}\operatorname{sign}(x)$ and use a)].

7.14. Show that $\operatorname{comb}_{\mathcal{P}}(x) = \sum_{k=-\infty}^{+\infty} \delta(x - k\mathcal{P})$ is periodic of period \mathcal{P} . Compute the Fourier transform of $\operatorname{comb}_{\mathcal{P}}$.

7.15. Compute $\delta \circ f$ where $f(x) = \sin x$.

[Answer: $\sum_{k=-\infty}^{+\infty} \delta(x - k\pi) \equiv \operatorname{comb}_\pi(x)$].

7.16. Let $v \in S(\mathbb{R})$ and define $\varphi(x) = \sum_{k=-\infty}^{+\infty} v(x + 2\pi k)$.

- a) Show that φ is well defined in \mathbb{R} , 2π -periodic and has the Fourier expansion

$$\varphi(x) = \frac{1}{2\pi} \sum_{n=-\infty}^{+\infty} \hat{v}(n) e^{inx}.$$

- b) Deduce the **Poisson's formula**

$$\sum_{k=-\infty}^{+\infty} \hat{v}(k) = 2\pi \sum_{k=-\infty}^{+\infty} v(2\pi k). \quad (7.93)$$

- c) Show that $\mathcal{F}[comb] = \frac{1}{2\pi} comb_{2\pi}$.

7.17. a) Use Proposition 7.10, p. 435, to show that

$$F_r(\mathbf{x}) = \frac{\delta(|\mathbf{x}| - r)}{|\mathbf{x}|} \quad \text{and} \quad G_r(\mathbf{x}) = \frac{\partial_r \delta(|\mathbf{x}| - r)}{|\mathbf{x}|}$$

are well defined as distributions in $\mathcal{D}'(\mathbb{R}^3)$.

- b) Use Example 7.27, p. 447, to show that $F_r \rightarrow 0$ and $G_r \rightarrow 4\pi\delta_3$ in $\mathcal{D}'(\mathbb{R}^3)$, as $r \rightarrow 0$.

7.18. Let $g, h \in C_0^\infty(\mathbb{R}^n)$.

- a) Show that $u \in C^2(\mathbb{R}^n \times [0, +\infty))$ is a solution of the global Cauchy problem

$$\begin{cases} u_{tt} - c^2 \Delta u = 0, & \mathbf{x} \in \mathbb{R}^n, t > 0 \\ u(\mathbf{x}, 0) = g(\mathbf{x}), \quad u_t(\mathbf{x}, 0) = h(\mathbf{x}), & \mathbf{x} \in \mathbb{R}^n \end{cases} \quad (n = 1, 2, 3) \quad (7.94)$$

if and only if $u^0(\mathbf{x}, t)$, the extension of u to zero for $t < 0$, is a solution in $\mathcal{D}'(\mathbb{R}^n \times \mathbb{R})$ of the equation

$$u_{tt}^0 - c^2 \Delta u^0 = g(\mathbf{x}) \otimes \delta'(t) + h(\mathbf{x}) \otimes \delta(t). \quad (7.95)$$

- b) Let $K = K(\mathbf{x}, t)$ be the fundamental solution of the wave equation in one of the dimensions $n = 1, 2, 3$. Deduce from a) that $K^0(\mathbf{x}, t)$ satisfies the equation

$$K_{tt}^0 - c^2 \Delta K^0 = \delta_n(\mathbf{x}) \otimes \delta(t), \quad \text{in } \mathcal{D}'(\mathbb{R}^n \times \mathbb{R}).$$

[Hint: a) Let u satisfy (7.95). This means that, for every $\varphi \in \mathcal{D}(\mathbb{R}^n \times \mathbb{R})$,

$$\int_{\mathbb{R}} \int_{\mathbb{R}^n} u^0 (\varphi_{tt} - c^2 \Delta \varphi) d\mathbf{x} dt = - \int_{\mathbb{R}^n} g(\mathbf{x}) \varphi_t(\mathbf{x}, 0) d\mathbf{x} + \int_{\mathbb{R}^n} h(\mathbf{x}) \varphi(\mathbf{x}, 0) d\mathbf{x}. \quad (7.96)$$

Integrate twice by parts the first integral, taking into account that $u^0 = u^0(\mathbf{x}, t) = \mathcal{H}(t) u(\mathbf{x}, t)$, to get,

$$\begin{aligned} & \int_0^{+\infty} \int_{\mathbb{R}^n} (u_{tt} - c^2 \Delta u) \varphi d\mathbf{x} dt \\ &= \int_{\mathbb{R}^n} [u(\mathbf{x}, 0) - g(\mathbf{x})] \varphi_t(\mathbf{x}, 0) d\mathbf{x} - \int_{\mathbb{R}^n} [u_t(\mathbf{x}, 0) - h(\mathbf{x})] \varphi(\mathbf{x}, 0) d\mathbf{x}. \end{aligned} \quad (7.97)$$

Choose arbitrarily $\varphi \in \mathcal{D}(\mathbb{R}^n \times (0, +\infty))$ to recover the wave equation. Then choose $\psi_0, \psi_1 \in \mathcal{D}(\mathbb{R}^n)$ and $b(t) \in C_0^\infty(\mathbb{R})$ such that $b(0) = 1$ and $b'(0) = 0$. Observe that the function

$$\psi(\mathbf{x}, t) = (\psi_0(\mathbf{x}) + t\psi_1(\mathbf{x})) b(t)$$

belongs to $\mathcal{D}(\mathbb{R}^n \times \mathbb{R})$. Insert ψ into (7.97) and use the arbitrariness of ψ_0, ψ_1 to deduce that u satisfy also the initial conditions. Viveversa, let u satisfy (7.94). A double integration by parts gives, using the initial conditions,

$$0 = \int_0^{+\infty} \int_{\mathbb{R}^n} u(\varphi_{tt} - c^2 \Delta \varphi) d\mathbf{x} dt + \int_{\mathbb{R}^n} g(\mathbf{x}) \varphi_t(\mathbf{x}, 0) d\mathbf{x} - \int_{\mathbb{R}^n} h(\mathbf{x}) \varphi(\mathbf{x}, 0) d\mathbf{x}$$

which is equivalent to (7.96).

- b) Recall that $K(\mathbf{x}, 0) = 0, K_t(\mathbf{x}, 0) = \delta_3(\mathbf{x})$.

7.19. Let $g \in C_0^\infty(\mathbb{R}^n), f \in C_0^\infty(\mathbb{R}^n \times (0, +\infty))$.

- a) Show that $u \in C^{2,1}(\mathbb{R}^n \times [0, +\infty))$ is a solution of the global Cauchy problem

$$\begin{cases} u_t - \Delta u = f(\mathbf{x}, t), & \mathbf{x} \in \mathbb{R}^n, t > 0 \\ u(\mathbf{x}, 0) = g(\mathbf{x}), & \mathbf{x} \in \mathbb{R}^n \end{cases} \quad (7.98)$$

if and only if $u^0(\mathbf{x}, t)$, the extension of u to zero for $t < 0$, is a solution of the equation

$$u_t^0 - \Delta u^0 = f^0(\mathbf{x}, t) + g(\mathbf{x}) \otimes \delta(t) \quad \text{in } \mathcal{D}'(\mathbb{R}^n \times \mathbb{R}). \quad (7.99)$$

- b) Let $\Gamma = \Gamma(\mathbf{x}, t)$ be the fundamental solution of the heat equation in dimension $n = 1, 2, 3$. Deduce from a) that $\Gamma^0(\mathbf{x}, t)$ satisfies

$$\Gamma_t^0 - \Delta \Gamma^0 = \delta_n(\mathbf{x}) \otimes \delta(t) \quad \text{in } \mathcal{D}'(\mathbb{R}^n \times \mathbb{R}).$$

7.20. Choose in Theorem 7.56, p. 461,

$$H = L^2(\Omega; \mathbb{R}^n), \quad Z = L^2(\Omega) \subset \mathcal{D}'(\Omega)$$

and $L : H \rightarrow \mathcal{D}'(\Omega)$ given by $L = \operatorname{div}$. Identify the resulting space W .

7.21. Let X and Z be Banach spaces with $Z \hookrightarrow \mathcal{D}'(\Omega; \mathbb{R}^n)$ (e.g. $L^p(\Omega)$ or $L^p(\Omega; \mathbb{R}^n)$). Let $L : X \rightarrow \mathcal{D}'(\Omega; \mathbb{R}^n)$ be a linear continuous operator (e.g. a gradient or a divergence). Define

$$W = \{v \in X : Lv \in Z\}$$

with norm

$$\|u\|_W^2 = \|u\|_X^2 + \|Lu\|_Z^2.$$

Prove that W is a Banach space, continuously embedded in X .

7.22. The Sobolev spaces $W^{1,p}$. Let $\Omega \subseteq \mathbb{R}^n$ be a domain. For $p \geq 1$, define

$$W^{1,p}(\Omega) = \{v \in L^p(\Omega) : \nabla v \in L^p(\Omega; \mathbb{R}^n)\}.$$

Using the result of Problem 7.21, show that $W^{1,p}(\Omega)$ is a Banach space.

7.23. Let $\Omega = B_{1/2}(\mathbf{0}) \subset \mathbb{R}^n, n > 2$, and $u(\mathbf{x}) = |\mathbf{x}|^{-a}, \mathbf{x} \neq \mathbf{0}$. Determine for which values of a , $u \in H^2(\Omega)$.

7.24. Prove that $C_0^\infty(\Omega)^+ = \{u \in C_0^\infty(\Omega); u \geq 0 \text{ in } \Omega\}$ is dense in

$$H_0^1(\Omega)^+ = \{u \in H_0^1(\Omega); u \geq 0 \text{ a.e. in } \Omega\}.$$

7.25. Let $u \in H^s(\mathbb{R})$. Prove that, if $s > 1/2$, $u \in C(\mathbb{R})$ and $u(x) \rightarrow 0$ as $x \rightarrow \pm\infty$.

[Hint: Show that $\hat{u} \in L^1(\mathbb{R})$].

7.26. Let u and Ω satisfy the hypotheses of Theorem 7.91, p. 488. Prove that, if the embedding $W^{1,p}(\Omega) \hookrightarrow L^p(\Omega)$ is compact, $1 < p < +\infty$ then

$$\|u\|_{L^p(\Omega)} \leq c(n, p, \Omega) \|\nabla u\|_{L^2(\Omega; \mathbb{R}^n)}.$$

7.27. Let Ω be a bounded domain (not necessarily Lipschitz). Show that $H_0^1(\Omega)$ is compactly embedded in $L^2(\Omega)$.

[Hint: Extend u by zero outside Ω].

7.28. Let

$$H_{0,a}^1(a, b) = \{u \in H^1(a, b) : u(a) = 0\}.$$

Show that Poincaré's inequality holds in $H_{0,a}^1(a, b)$.

7.29. Let $n > 1$ and

$$\Omega = \{(\mathbf{x}', x_n) : \mathbf{x}' \in \mathbb{R}^{n-1}, 0 < x_n < d\}.$$

Show that the Poincaré's inequality holds in $H_0^1(\Omega)$.

7.30. Prove the chain rule in Subsec. 7.7.6 under the further assumption that $f \in C^1(\mathbb{R})$ and that Ω is a bounded Lipschitz domain.

[Hint: There exists $\{u_m\} \subset C^1(\overline{\Omega})$, $u_m \rightarrow u$ in $H^1(\Omega)$ and a.e. in Ω , as $m \rightarrow \infty$. From $|f(u_m) - f(u)| \leq \max |f'| |u_m - u|$ deduce that $f(u_m) \rightarrow f(u)$ in $L^2(\Omega)$ and a.e. in Ω , as $m \rightarrow \infty$. Using the dominated convergence Theorem, show that

$$f'(u_m) \partial_{x_j} u_m - f'(u) \partial_{x_j} u \rightarrow 0$$

in $L^2(\Omega)$, as $m \rightarrow \infty$].

7.31. Modifying slightly the proof of Theorem 7.82, p. 481, to prove that for every $\varepsilon > 0$

$$\|\tau_0 u\|_{L^2(\mathbb{R}^{n-1})}^2 \leq \frac{1+\varepsilon}{\varepsilon} \|u\|_{L^2(\mathbb{R}_+^n)}^2 + \varepsilon \|\nabla u\|_{L^2(\mathbb{R}_+^n; \mathbb{R}^n)}^2. \quad (7.100)$$

7.32. Let Ω be a bounded, Lipschitz domain and let Γ_0 a relatively open subset of $\partial\Omega$.

a) Show that

$$(u, v)_{\tilde{H}^1(\Omega)} = \int_{\Gamma_0} u|_{\Gamma_0} v|_{\Gamma_0} d\sigma + \int_{\Omega} \nabla u \cdot \nabla v d\mathbf{x}$$

is an inner product in $H^1(\Omega)$.

b) Show that the induced norm is equivalent to $\|u\|_{H^1(\Omega)}$.

[Hint: See Remark 7.92, p. 489].

7.33. Let Ω be a bounded Lipschitz domain and $\tau_0 : H^1(\Omega) \rightarrow L^2(\Gamma)$ be the trace operator. Show that $\tau_0(\mathcal{D}(\overline{\Omega}))$ is dense in $H^{1/2}(\Gamma, \Gamma = \partial\Omega)$.

7.34. Let Ω be a bounded Lipschitz domain, $\Gamma = \partial\Omega$ and $g : \Gamma \rightarrow \mathbb{R}$ be Lipschitz continuous with Lipschitz constant L , that is

$$|g(\sigma_1) - g(\sigma_2)| \leq L|\sigma_1 - \sigma_2|$$

for every $\sigma_1, \sigma_2 \in \Gamma$.

- a) Show that the function $\tilde{g} : \mathbb{R}^n \rightarrow \mathbb{R}$, defined by

$$\tilde{g}(\mathbf{x}) = \min_{\sigma \in \Gamma} \{g(\sigma) + L|\mathbf{x} - \sigma|\}$$

is Lipschitz continuous in \mathbb{R}^n with Lipschitz constant L and $\tilde{g} = g$ on Γ . Thus \tilde{g} is a Lipschitz extension g in \mathbb{R}^n .

- b) Deduce that $g \in H^{1/2}(\Gamma)$.

[Hint: a) Let $\tilde{g}(\mathbf{x}) = g(\sigma_x) + L|\mathbf{x} - \sigma_x|$. Observe that

$$\begin{aligned} g(\mathbf{x}) - g(\mathbf{y}) &= g(\sigma_x) + L|\mathbf{x} - \sigma_x| - g(\sigma_y) - L|\mathbf{y} - \sigma_y| \\ &\leq L(|\mathbf{x} - \sigma_x| - |\mathbf{y} - \sigma_y|) \leq \dots \end{aligned}$$

7.35. Prove Proposition 7.102, p. 496.

[Hint: Recall that a sequence of real functions $\{g_k\}$ convergent to g in $L^1(0, T)$, has a subsequence converging a.e. to the same limit (Theorem B.4). Apply this result to the sequence $g_k(t) = (u_k(t), v)_V$ and observe that $|g_k(t)| \leq C \|v\|_V$].

Chapter 8

Variational Formulation of Elliptic Problems

8.1 Elliptic Equations

Poisson's equation $\Delta u = f$ is the simplest among the *elliptic equations*, according to the classification in Sect. 5.5, at least in dimension two. This type of equations plays an important role in the modelling of a large variety of phenomena, often of stationary nature. Typically, in drift, diffusion and reaction models like those considered in Chap. 2, a stationary solution corresponds to a steady state, with no more dependence on time.

Elliptic equations appear in the theory of electrostatic and electromagnetic potentials or in the search of vibration modes of elastic structures as well (e.g. through the method of separation of variables for the wave equation). Let us define precisely what we mean by *elliptic equation* in dimension n .

Let $\Omega \subseteq \mathbb{R}^n$ be a domain, $\mathbf{A}(\mathbf{x}) = (a_{ij}(\mathbf{x}))$ a square matrix of order n , $\mathbf{b}(\mathbf{x}) = (b_1(\mathbf{x}), \dots, b_n(\mathbf{x}))$, $\mathbf{c}(\mathbf{x}) = (c_1(\mathbf{x}), \dots, c_n(\mathbf{x}))$ vector fields in \mathbb{R}^n , $a = a(\mathbf{x})$ and $f = f(\mathbf{x})$ real functions. An equation of the form

$$-\sum_{i,j=1}^n \partial_{x_i} (a_{ij}(\mathbf{x}) u_{x_j}) + \sum_{i=1}^n \partial_{x_i} (b_i(\mathbf{x}) u) + \sum_{i=1}^n c_i(\mathbf{x}) u_{x_i} + a(\mathbf{x}) u = f(\mathbf{x}) \quad (8.1)$$

or

$$-\sum_{i,j=1}^n a_{ij}(\mathbf{x}) u_{x_i x_j} + \sum_{i=1}^n b_i(\mathbf{x}) u_{x_i} + a(\mathbf{x}) u = f(\mathbf{x}) \quad (8.2)$$

is said to be **elliptic in Ω** if \mathbf{A} is **positive** in Ω , i.e. if the following *ellipticity condition* holds:

$$\sum_{i,j=1}^n a_{ij}(\mathbf{x}) \xi_i \xi_j > 0, \quad \forall \mathbf{x} \in \Omega, \forall \boldsymbol{\xi} \in \mathbb{R}^n, \boldsymbol{\xi} \neq \mathbf{0}.$$

We say that (8.1) is in **divergence form** since it may be written as

$$\underbrace{-\operatorname{div}(\mathbf{A}(\mathbf{x}) \nabla u)}_{\text{diffusion}} + \underbrace{\operatorname{div}(\mathbf{b}(\mathbf{x}) u) + \mathbf{c}(\mathbf{x}) \cdot \nabla u}_{\text{transport}} + \underbrace{a(\mathbf{x}) u}_{\text{reaction}} = \underbrace{f(\mathbf{x})}_{\text{external source}} \quad (8.3)$$

which emphasizes the particular structure of the highest order terms. Usually, the first term models the diffusion in heterogeneous or anisotropic media, when the constitutive law for the flux function \mathbf{q} is given by the Fourier or the Fick law:

$$\mathbf{q} = -\mathbf{A}\nabla u.$$

Here u may represent, for instance, a temperature or the concentration of a substance. Thus, the term $-\operatorname{div}(\mathbf{A}\nabla u)$ is associated with thermal or molecular diffusion. The matrix \mathbf{A} is called *diffusion matrix*; the dependence of \mathbf{A} on \mathbf{x} denotes anisotropic diffusion.

The examples in Chap. 2 explain the meaning of the other terms in equation (8.3). In particular, $\operatorname{div}(\mathbf{b}u)$ models *convection or transport* and corresponds to a flux function given by

$$\mathbf{q} = \mathbf{b}u.$$

The vector \mathbf{b} has the dimensions of a **velocity**. Think, for instance, of the fumes emitted by a factory installations, which diffuse and are transported by the wind. In this case \mathbf{b} is the wind velocity. Note that, if $\operatorname{div}\mathbf{b} = 0$, then $\operatorname{div}(\mathbf{b}u)$ reduces to $\mathbf{b} \cdot \nabla u$ which is of the same form of the third term $\mathbf{c} \cdot \nabla u$.

The term $a u$ models *reaction*. If u is the concentration of a substance, a represents the rate of decomposition ($a > 0$) or growth ($a < 0$).

Finally, f represents an external action, distributed in Ω , e.g. the rate of heat per unit mass supplied by an external source.

If the entries a_{ij} of the matrix \mathbf{A} and the component b_j of \mathbf{b} are all differentiable, we may compute the divergence of both $\mathbf{A}\nabla u$ and $\mathbf{b}u$, and reduce (8.1) to the *non-divergence form*

$$-\sum_{i,j=1}^n a_{ij}(\mathbf{x}) u_{x_i x_j} + \sum_{k=1}^n \tilde{b}_k(\mathbf{x}) u_{x_k} + \tilde{c}(\mathbf{x}) u = f(\mathbf{x})$$

where

$$\tilde{b}_k(\mathbf{x}) = -\sum_{i=1}^n \partial_{x_i} a_{ik}(\mathbf{x}) + b_k(\mathbf{x}) + c_k(\mathbf{x}) \quad \text{and} \quad \tilde{c}(\mathbf{x}) = \operatorname{div}\mathbf{b}(\mathbf{x}) + a(\mathbf{x}).$$

However, when a_{ij} or b_j are *not differentiable*, we must keep the divergence form and interpret the differential equation (8.3) in a suitable weak sense.

A *non-divergence form equation* is also associated with diffusion phenomena through stochastic processes which generalize the Brownian motion, called *diffusion processes* (see e.g. [47], Øksendal, 1985).

In the next section we give a brief account of the various notions of solution available for these kinds of equations. Since the Poisson equation is both in divergence and non-divergence form, we use the Dirichlet problem for this equation as a model problem.

8.2 Notions of Solutions

Assume we are given a domain $\Omega \subset \mathbb{R}^n$ and two real functions $f : \Omega \rightarrow \mathbb{R}, g : \partial\Omega \rightarrow \mathbb{R}$. We want to determine a function u satisfying the equation

$$\begin{cases} -\Delta u = f & \text{in } \Omega \\ u = g & \text{on } \partial\Omega. \end{cases} \quad (8.4)$$

Let us examine what we mean by *solving* the above problem. The obvious part is the final goal: we would like to show *existence, uniqueness and stability* of the solution; then, based on these results, we want to *compute* the solution by Numerical Analysis methods.

Less obvious is the *meaning of solution*. In fact, in principle, every problem may be formulated in several ways and a different notion of solution is associated with each way. What is important in the applications is to select the “most efficient notion” for the problem under examination, where “efficiency” may stand for the best compromise between *simplicity of both formulation and theoretical treatment, sufficient flexibility and generality, adaptability to numerical methods*.

Here is a (nonexhaustive!) list of various notions of solution for problem (8.4).

- **Classical** solutions are twice continuously differentiable functions; the differential equation and the boundary condition is satisfied in the usual pointwise sense.
- **Strong** solutions belong to the Sobolev space $H^2(\Omega)$. Thus, they possess derivatives in $L^2(\Omega)$ up to the second order, in the sense of distributions.

The differential equation is satisfied in the pointwise sense, a.e. with respect to the Lebesgue measure in Ω , while the boundary condition is satisfied in the sense of traces.

- **Distributional** solutions belong to $L_{loc}^1(\Omega)$ and the equation holds in the sense of distributions, i.e.:

$$\int_{\Omega} -u\Delta\varphi d\mathbf{x} = \int_{\Omega} f\varphi d\mathbf{x}, \quad \forall \varphi \in \mathcal{D}(\Omega).$$

The boundary condition is satisfied in a very weak sense.

- **Variational** (or **weak**) solutions belong to the Sobolev space $H^1(\Omega)$. The boundary value problem is recast within the framework of the abstract variational theory developed in Sect. 6.6. Often the new formulation represents a version of the *principle of virtual work*.
- **Viscosity** solutions. In the simplest case, a viscosity solution is a *continuous function* in $\overline{\Omega}$, satisfying the following conditions: for every $\mathbf{x}_0 \in \Omega$, and every smooth function φ such that $\varphi(\mathbf{x}_0) = u(\mathbf{x}_0)$ and

$$\varphi(\mathbf{x}) \geq u(\mathbf{x}) \quad (\text{resp. } \varphi(\mathbf{x}) \leq u(\mathbf{x}))$$

in a neighborhood of \mathbf{x}_0 , it holds

$$-\Delta\varphi(\mathbf{x}_0) \geq f(\mathbf{x}_0) \quad (\text{resp. } \leq).$$

As we see, the action of the differential operator is carried onto the test function, in a proper pointwise fashion.¹

Clearly, all these notions of solution must be connected by a *coherence principle*, which may be stated as follows: if all the data (domain, coefficients, boundary data, forcing terms) and the solution are smooth, *all the above notions must be equivalent*. Thus, the *non-classical* notions constitute a generalization of the classical one.

An important task, with consequences in the error control in numerical methods, is to establish the optimal degree of regularity of a non-classical solution.

More precisely, let u be a non-classical solution of the Poisson problem (8.4). The question is:

how much does the regularity of f, g and of the domain Ω affect the regularity of the solution?

An exhaustive answer requires rather complicated tools. In the sequel we shall indicate only the most relevant results.

The theory for classical and strong solutions is well established and can be found, e.g. in the book [3] of *Gilbarg-Trudinger*, 1998. From the numerical point of view, the *method of finite differences* best fits the differential structure of the problem and aims at approximating classical solutions.

The distributional theory is well developed, is quite general, but is not the most appropriate framework for solving boundary value problems. Indeed, the sense in which the boundary values are attained is one of the most delicate points, when one is willing to widen the notion of solution.

The notion of viscosity solution is designed to deal with fully nonlinear operators, when no integration by parts is available. Important examples come from Stochastic Optimal Control and Dynamic Programming. For the basic theory, we refer to the by now classical paper of *M. Crandall, H. Ishii, P.L. Lions*, User's Guide to Viscosity Solutions of Second Order Partial Differential equations, Bull. A.M.S., 1983.

For our purposes, the most convenient notion of solution is the variational (or weak) one: it leads to a quite flexible formulation, with a sufficiently high degree of generality, and a basic theory solely relying on the Lax-Milgram Theorem (Sect. 6.6). Moreover, the analogy (and often the coincidence) with the principle of virtual work indicates a direct connection with the physical interpretation.

¹ To get a clue of the meaning of viscosity solution, look at the problem $u'' = 0$, in the interval $(0, 1)$, $u(0) = 0$, $u(1) = 1$, whose solution is clearly $u(x) = x$. The conditions in the notion of viscosity solution prescribe that, if a parabola $\varphi = \varphi(x)$ touches from above (below) the graph of $u(x) = x$ at a point x_0 , then $\varphi''(x_0) \geq 0$ (≤ 0).

Finally, the variational formulation is the most natural to implement the *Galerkin method* (*finite elements*, *spectral elements*, etc...), widely used in the numerical approximation of the solutions of boundary value problems.

Thus, we shall develop the basic theory of elliptic equations in divergence form, recasting the most common boundary value problems within the functional framework of the abstract variational problems of Sect. 6.6. To present the main ideas behind the variational formulation, we start from the Poisson equation (with zero order term). All the results we present hold in any dimension $n \geq 1$. In dimension $n = 1$, Ω is an interval $(a, b) \subset \mathbb{R}$ and $\partial_{\mathbf{v}} u$ at the boundary means $-u_x(a)$ and $u_x(b)$.

8.3 Problems for the Poisson Equation

In this section we examine the weak formulation of the most common boundary value problems for the equation $-\Delta u + a(\mathbf{x}) u = f$. In general one can proceed along the following steps.

1. Select a space of *smooth test functions, compatible with the boundary conditions*. Multiply the differential equation by a *test function* and integrate over the domain Ω .
2. Assume all the data of the problem are smooth. Integrate by parts the diffusion term, using the boundary conditions and obtaining an *integral equation (the variational formulation)*.
3. Check that the choice of the test functions and the consequent variational formulation are the correct ones; that is, recover the original formulation by integrating by parts in reverse order. This means that for classical solutions the two formulations are equivalent (coherence principle).
4. Interpret the integral equation as an *abstract variational problem* (Sect. 6.6) in a suitable Hilbert space. In general, this is a Sobolev space, given by the closure of the space of test functions in a suitable norm.

At this point, the main tools for solving the abstract variational problem are the Lax-Milgram Theorem 6.39, p. 383 (or the Riesz Representation Theorem 6.30, p. 378, if the bilinear form is symmetric) and the Fredholm's Alternative Theorem 6.66, p. 402.

8.3.1 Dirichlet problem

Let $\Omega \subset \mathbb{R}^n$ be a *bounded domain*. We examine the following problem:

$$\begin{cases} -\Delta u + a(\mathbf{x}) u = f & \text{in } \Omega \\ u = 0 & \text{on } \partial\Omega. \end{cases} \quad (8.5)$$

To achieve a weak formulation, let us follows the steps 1–4 above.

1. We first assume that a and f are smooth and that $u \in C^2(\Omega) \cap C(\overline{\Omega})$ is a classical solution of (8.5). We select $C_0^1(\Omega)$ as the space of test functions, having continuous first derivatives and compact support in Ω . In particular, *they vanish in a neighborhood of $\partial\Omega$* . Let $v \in C_0^1(\Omega)$ and multiply the differential equation by v . We get

$$\int_{\Omega} \{-\Delta u + a(\mathbf{x})u - f\} v \, d\mathbf{x} = 0. \quad (8.6)$$

2. We Integrate by parts and use the boundary condition. We obtain

$$\int_{\Omega} \{\nabla u \cdot \nabla v + a(\mathbf{x})uv\} \, d\mathbf{x} = \int_{\Omega} fv \, d\mathbf{x}, \quad \forall v \in C_0^1(\Omega). \quad (8.7)$$

Thus (8.5) implies (8.7).

3. On the other hand, assume (8.7) is true. Integrating by parts in the reverse order we return to (8.6), which entails $-\Delta u + a(\mathbf{x})u - f = 0$ in Ω , due to the arbitrariness of v . Thus, *for classical solutions, the two formulations (8.5) and (8.7) are equivalent*.

4. Observe that (8.7) only involves first order derivatives of the solution and of the test function. Then, enlarging the space of test functions to $H_0^1(\Omega)$, the closure of $C_0^1(\Omega)$ in the norm $\|u\|_{H_0^1(\Omega)} = \|\nabla u\|_{L^2(\Omega; \mathbb{R}^n)}$, we may state the **weak** formulation of problem (8.5) as follows:

Determine $u \in H_0^1(\Omega)$ such that

$$\int_{\Omega} \{\nabla u \cdot \nabla v + a(\mathbf{x})uv\} \, d\mathbf{x} = \int_{\Omega} fv \, d\mathbf{x}, \quad \forall v \in H_0^1(\Omega). \quad (8.8)$$

Introducing the bilinear form

$$B(u, v) = \int_{\Omega} \{\nabla u \cdot \nabla v + a(\mathbf{x})uv\} \, d\mathbf{x}$$

and the linear functional

$$Lv = \int_{\Omega} fv \, d\mathbf{x},$$

equation (8.8) corresponds to the abstract variational problem

$$B(u, v) = Lv, \quad \forall v \in H_0^1(\Omega).$$

The well-posedness of this problem follows from the Lax-Milgram Theorem under the hypothesis $a(\mathbf{x}) \geq 0$ a.e. in Ω . Recall that a Poincaré inequality holds in $H_0^1(\Omega)$, (see Theorem 7.61, p. 466):

$$\|u\|_{L^2(\Omega)} \leq C_P \|\nabla u\|_{L^2(\Omega; \mathbb{R}^n)}.$$

We have:

Theorem 8.1. Assume that $f \in L^2(\Omega)$ and that $0 \leq a(\mathbf{x}) \leq \alpha_0$ a.e. in Ω . Then, problem (8.8) has a unique solution $u \in H_0^1(\Omega)$. Moreover

$$\|\nabla u\|_{L^2(\Omega; \mathbb{R}^n)} \leq C_P \|f\|_{L^2(\Omega)}.$$

Proof. We check that the hypotheses of the Lax-Milgram Theorem hold, with $V = H_0^1(\Omega)$.

Continuity of the bilinear form B . The Schwarz and Poincaré inequalities yield:

$$\begin{aligned} |B(u, v)| &\leq \|\nabla u\|_{L^2(\Omega; \mathbb{R}^n)} \|\nabla v\|_{L^2(\Omega; \mathbb{R}^n)} + \alpha_0 \|u\|_{L^2(\Omega)} \|v\|_{L^2(\Omega)} \\ &\leq (1 + C_P^2 \alpha_0) \|\nabla u\|_{L^2(\Omega; \mathbb{R}^n)} \|\nabla v\|_{L^2(\Omega; \mathbb{R}^n)} \end{aligned}$$

so that B is continuous in $H_0^1(\Omega)$.

Coercivity of B . It follows from

$$B(u, u) = \int_{\Omega} |\nabla u|^2 d\mathbf{x} + \int_{\Omega} a(\mathbf{x}) u^2 d\mathbf{x} \geq \|\nabla u\|_{L^2(\Omega; \mathbb{R}^n)}^2$$

since $a \geq 0$.

Continuity of L . The Schwarz and Poincaré inequalities give

$$|Lv| = \left| \int_{\Omega} fv d\mathbf{x} \right| \leq \|f\|_{L^2(\Omega)} \|v\|_{L^2(\Omega)} \leq C_P \|f\|_{L^2(\Omega)} \|\nabla v\|_{L^2(\Omega; \mathbb{R}^n)}.$$

Hence $L \in H^{-1}(\Omega)$ and $\|L\|_{H^{-1}(\Omega)} \leq C_P \|f\|_{L^2(\Omega)}$. The conclusions follow from the Lax-Milgram Theorem. \square

Remark 8.2. Suppose that $a = 0$ and that u represents the equilibrium position of an elastic membrane. Then $B(u, v)$ represents the work done by the elastic internal forces, due to a *virtual displacement* v . On the other hand Lv expresses the work done by the external forces. The weak formulation (8.8) states that these two works balance, which constitutes a version of the *principle of virtual work*.

Furthermore, due to the symmetry of B , the solution u of the problem **minimizes in $H_0^1(\Omega)$ the Dirichlet functional**

$$E(u) = \underbrace{\frac{1}{2} \int_{\Omega} |\nabla u|^2 d\mathbf{x}}_{\text{internal elastic energy}} - \underbrace{\int_{\Omega} fu d\mathbf{x}}_{\text{external potential energy}}$$

which represents the **total potential energy**. Equation (8.8) constitutes the Euler equation for E .

Thus, in agreement with the principle of virtual work, u minimizes the potential energy among all the admissible configurations.

Similar observations can be made for the other types of boundary conditions.

8.3.2 Neumann, Robin and mixed problems

Let $\Omega \subset \mathbb{R}^n$ be a bounded, Lipschitz domain. We examine the following Neumann problem:

$$\begin{cases} -\Delta u + a(\mathbf{x})u = f & \text{in } \Omega \\ \partial_{\nu} u = g & \text{on } \partial\Omega \end{cases} \quad (8.9)$$

where ν denotes the outward normal unit vector to $\partial\Omega$. As for the Dirichlet problem, to derive a variational formulation, we follows the steps **1–4**.

1. Assume that a , f and g are smooth and that $u \in C^2(\Omega) \cap C^1(\overline{\Omega})$ is a classical solution of (8.9). Since there is no particularly convenient way to choose the space of test functions in order to incorporate the Neumann boundary condition, we choose $C^1(\overline{\Omega})$ as the space of test functions, having continuous first derivatives up to $\partial\Omega$. Let $v \in C^1(\overline{\Omega})$, arbitrary, and multiply the Poisson equation by v . Integrating over Ω , we get

$$\int_{\Omega} \{-\Delta u + au\} v \, d\mathbf{x} = \int_{\Omega} fv \, d\mathbf{x}. \quad (8.10)$$

2. An integration by parts gives

$$-\int_{\partial\Omega} \partial_{\nu} u \, vd\sigma + \int_{\Omega} \{\nabla u \cdot \nabla v + auv\} \, d\mathbf{x} = \int_{\Omega} fv \, d\mathbf{x}, \quad \forall v \in C^1(\overline{\Omega}). \quad (8.11)$$

Using the Neumann condition we may write

$$\int_{\Omega} \{\nabla u \cdot \nabla v + auv\} \, d\mathbf{x} = \int_{\Omega} fv \, d\mathbf{x} + \int_{\partial\Omega} gv \, d\sigma, \quad \forall v \in C^1(\overline{\Omega}). \quad (8.12)$$

Thus (8.9) implies (8.12).

3. On the other hand, suppose that (8.12) is true. Integrating by parts in the reverse order, we find

$$\int_{\Omega} \{-\Delta u + au - f\} v \, d\mathbf{x} + \int_{\partial\Omega} \partial_{\nu} u \, vd\sigma = \int_{\partial\Omega} gv \, d\sigma, \quad (8.13)$$

for every $\forall v \in C^1(\overline{\Omega})$. Since $C_0^1(\Omega) \subset C^1(\overline{\Omega})$ we may insert any $v \in C_0^1(\Omega)$ into (8.13), to get

$$\int_{\Omega} \{-\Delta u + au - f\} v \, d\mathbf{x} = 0.$$

The arbitrariness of $v \in C_0^1(\Omega)$ entails $-\Delta u + au - f = 0$ in Ω . Therefore (8.13) becomes

$$\int_{\partial\Omega} \partial_{\nu} u \, vd\sigma = \int_{\partial\Omega} gv \, d\sigma, \quad \forall v \in C^1(\overline{\Omega})$$

and the arbitrariness of $v \in C^1(\overline{\Omega})$ forces $\partial_\nu u = g$, recovering the Neumann condition as well². Thus, *for classical solutions, the two formulations (8.9) and (8.12) are equivalent.*

4. Recall now that, by Theorem 7.81, p. 479, $C^1(\overline{\Omega})$ is dense in $H^1(\Omega)$, which therefore constitutes the natural Sobolev space for the Neumann problem. Then, enlarging the space of test functions to $H^1(\Omega)$, we may give the **weak** formulation of problem (8.9) as follows:

Determine $u \in H^1(\Omega)$ such that

$$\int_{\Omega} \{\nabla u \cdot \nabla v + auv\} d\mathbf{x} = \int_{\Omega} fv d\mathbf{x} + \int_{\partial\Omega} gv d\sigma, \quad \forall v \in H^1(\Omega). \quad (8.14)$$

We point out that the Neumann condition is encoded in (8.14) and not forced as for the Dirichlet boundary conditions. Since we used the density of $C^1(\overline{\Omega})$ in $H^1(\Omega)$ and the trace of v on $\partial\Omega$, some regularity of the domain is needed, even in the variational formulation (Lipschitz is enough). Introducing the bilinear form

$$B(u, v) = \int_{\Omega} \{\nabla u \cdot \nabla v + auv\} d\mathbf{x} \quad (8.15)$$

and the linear functional

$$Lv = \int_{\Omega} fv d\mathbf{x} + \int_{\partial\Omega} gv d\sigma, \quad (8.16)$$

(8.14) may be formulated as the abstract variational problem

$$B(u, v) = Lv, \quad \forall v \in H^1(\Omega).$$

The following theorem states the well-posedness of this problem under reasonable hypotheses on the data. Recall from Theorem 7.82, p. 481, the *trace inequality*

$$\|v\|_{L^2(\partial\Omega)} \leq \overline{C}(n, \Omega) \|v\|_{H^1(\Omega)}. \quad (8.17)$$

Theorem 8.3. *Let $\Omega \subset \mathbb{R}^n$ be a bounded, Lipschitz domain, $f \in L^2(\Omega)$, $g \in L^2(\partial\Omega)$ and*

$$0 < c_0 \leq a(\mathbf{x}) \leq \alpha_0 \quad a.e. \text{ in } \Omega.$$

Then, problem (8.14) has a unique solution $u \in H^1(\Omega)$. Moreover,

$$\|u\|_{H^1(\Omega)} \leq \frac{1}{\min\{1, c_0\}} \left\{ \|f\|_0 + \overline{C} \|g\|_{L^2(\partial\Omega)} \right\}.$$

Proof. We check that the hypotheses of the Lax-Milgram Theorem hold, with $V = H^1(\Omega)$.

² If Ω is smooth, the set of the restrictions to $\partial\Omega$ of functions in $C^1(\overline{\Omega})$ is dense in $L^2(\partial\Omega)$.

Continuity of the bilinear form B . The Schwarz inequality yields:

$$\begin{aligned} |B(u, v)| &\leq \|\nabla u\|_{L^2(\Omega; \mathbb{R}^n)} \|\nabla v\|_{L^2(\Omega; \mathbb{R}^n)} + \alpha_0 \|u\|_{L^2(\Omega)} \|v\|_{L^2(\Omega)} \\ &\leq (1 + \alpha_0) \|u\|_{H^1(\Omega)} \|v\|_{H^1(\Omega)} \end{aligned}$$

so that B is continuous in $H^1(\Omega)$.

Coercivity of B . It follows from

$$B(u, u) = \int_{\Omega} |\nabla u|^2 d\mathbf{x} + \int_{\Omega} au^2 d\mathbf{x} \geq \min \{1, c_0\} \|u\|_{H^1(\Omega)}^2$$

since

$$a(\mathbf{x}) \geq c_0 > 0 \quad \text{a.e. in } \Omega.$$

Continuity of L . From Schwarz's inequality and (8.17) we get:

$$\begin{aligned} |Lv| &\leq \left| \int_{\Omega} fv d\mathbf{x} \right| + \left| \int_{\partial\Omega} gv d\sigma \right| \leq \|f\|_{L^2(\Omega)} \|v\|_{L^2(\Omega)} + \|g\|_{L^2(\partial\Omega)} \|v\|_{L^2(\partial\Omega)} \\ &\leq \left\{ \|f\|_{L^2(\Omega)} + \overline{C} \|g\|_{L^2(\partial\Omega)} \right\} \|v\|_{H^1(\Omega)}. \end{aligned}$$

Therefore L is continuous in $H^1(\Omega)$ with

$$\|L\|_{H^1(\Omega)^*} \leq \|f\|_{L^2(\Omega)} + \overline{C} \|g\|_{L^2(\partial\Omega)}.$$

The conclusion follows from the Lax-Milgram Theorem. □

Remark 8.4. The condition $a(\mathbf{x}) \geq c_0 > 0$ for a.e. $\mathbf{x} \in \Omega$, in Theorem 8.3, may be relaxed by asking that³

$$a(\mathbf{x}) \geq 0 \quad \text{and} \quad \int_{\Omega} a(\mathbf{x}) d\mathbf{x} = I_0 > 0.$$

Indeed, in this case,

$$a(\mathbf{x}) \geq c_1 > 0$$

on some set E of positive measure. By Remark 7.92, p. 489, we can write

$$B(u, u) \geq \min \{1, c_1\} \left(\int_{\Omega} |\nabla u|^2 d\mathbf{x} + \int_E u^2 d\mathbf{x} \right) \geq C \|u\|_{H^1(\Omega)}^2$$

where C depends on $c_1, |\Omega|$ and $|E|$, recovering the coercivity of B .

If $a \equiv 0$, neither the existence nor the uniqueness of a solution is guaranteed. Two solutions with the same Neumann data differ by a constant. A way to restore uniqueness is to select a solution with, e.g., zero mean value, that is

$$\int_{\Omega} u(\mathbf{x}) d\mathbf{x} = 0.$$

³ Useful in some control problems.

The existence of a solution requires the following **compatibility condition** on the data f and g :

$$\int_{\Omega} f \, d\mathbf{x} + \int_{\partial\Omega} g \, d\sigma = 0, \quad (8.18)$$

obtained by substituting $v = 1$ into the equation

$$\int_{\Omega} \nabla u \cdot \nabla v \, dx = \int_{\Omega} fv \, d\mathbf{x} + \int_{\partial\Omega} gv \, d\sigma.$$

Note that, since Ω is bounded, the function $v = 1$ belongs to $H^1(\Omega)$.

If $a \equiv 0$ and (8.18) does not hold, problem (8.9) has no solution. Viceversa, we shall see later that, if this condition is fulfilled, a solution exists.

If $g = 0$, (8.18) has a simple interpretation. Indeed problem (8.9) is a model for the equilibrium configuration of a membrane whose boundary is free to slide along a vertical guide. The compatibility condition $\int_{\Omega} fd\mathbf{x} = 0$ expresses the obvious fact that, at equilibrium, the resultant of the external loads must vanish.

Robin problem. The same arguments leading to the weak formulation of the Neumann problem (8.9) may be used for the problem

$$\begin{cases} -\Delta u + a(\mathbf{x})u = f & \text{in } \Omega \\ \partial_{\nu} u + hu = g & \text{on } \partial\Omega, \end{cases} \quad (8.19)$$

inserting into (8.11) the Robin condition

$$\partial_{\nu} u = -hu + g, \quad \text{on } \partial\Omega.$$

We obtain the following **variational formulation**:

Determine $u \in H^1(\Omega)$ such that

$$\int_{\Omega} \{\nabla u \cdot \nabla v + auv\} \, d\mathbf{x} + \int_{\partial\Omega} huv \, d\sigma = \int_{\Omega} fv \, d\mathbf{x} + \int_{\partial\Omega} g \, d\sigma, \quad \forall v \in H^1(\Omega).$$

We have:

Theorem 8.5. Let Ω , f , g and a be as in Theorem 8.3, and $0 \leq h(\mathbf{x}) \leq h_0$ a.e. on $\partial\Omega$. Then, problem (8.19) has a unique weak solution $u \in H^1(\Omega)$. Moreover

$$\|u\|_{H^1(\Omega)} \leq \frac{1}{\min\{1, c_0\}} \left\{ \|f\|_{L^2(\Omega)} + \overline{C} \|g\|_{L^2(\partial\Omega)} \right\}.$$

Proof. Introducing the bilinear form

$$\tilde{B}(u, v) = B(u, v) + \int_{\partial\Omega} huv \, d\sigma,$$

the variational formulation becomes

$$\tilde{B}(u, v) = Lv, \quad \forall v \in H^1(\Omega)$$

where B and L are defined in (8.15) and (8.16), respectively.

From the Schwarz inequality and (8.17), we infer

$$\left| \int_{\partial\Omega} huv \, d\sigma \right| \leq h_0 \|u\|_{L^2(\partial\Omega)} \|v\|_{L^2(\partial\Omega)} \leq \bar{C}^2 h_0 \|u\|_{H^1(\Omega)} \|v\|_{H^1(\Omega)}.$$

On the other hand, the positivity of α , a_0 and h entails that

$$\tilde{B}(u, u) \geq B(u, u) \geq \min \{1, c_0\} \|u\|_{H^1(\Omega)}^2.$$

The conclusions follow easily. □

Remark 8.6. As in Theorem 8.3, we may relax the condition $a_0(\mathbf{x}) \geq c_0 > 0$ for a.e. $\mathbf{x} \in \Omega$, by asking that $a(\mathbf{x}) \geq 0$, a.e. in Ω , $h \geq 0$ a.e. on $\partial\Omega$, and

$$\int_{\Omega} a(\mathbf{x}) \, dx + \int_{\partial\Omega} h \, d\sigma = q > 0.$$

Mixed Dirichlet-Neumann problem. Let Γ_D and Γ_N be nonempty relatively open subsets of $\partial\Omega$, with $\Gamma_N = \partial\Omega \setminus \overline{\Gamma_D}$, regular in the sense of Definition 1.5, p. 14. Consider the problem

$$\begin{cases} -\Delta u + a(\mathbf{x}) u = f & \text{in } \Omega \\ u = 0 & \text{on } \Gamma_D \\ \partial_{\nu} u = g & \text{on } \Gamma_N. \end{cases}$$

The correct functional setting is $H_{0,\Gamma_D}^1(\Omega)$, i.e. the set of functions in $H^1(\Omega)$ with zero trace on Γ_D . From Theorem 7.91, p. 488, the Poincaré inequality holds in $H_{0,\Gamma_D}^1(\Omega)$ and therefore we may choose the norm

$$\|u\|_{H_{0,\Gamma_D}^1(\Omega)} = \|\nabla u\|_{L^2(\Omega; \mathbb{R}^n)}.$$

From (8.10) and the Gauss formula, we obtain, since $u = 0$ on Γ_D ,

$$-\int_{\Gamma_N} \partial_{\nu} u v \, d\sigma + \int_{\Omega} \{\nabla u \cdot \nabla v + a u v\} \, d\mathbf{x} = \int_{\Omega} f v \, d\mathbf{x}, \quad \forall v \in C^1(\overline{\Omega}).$$

The Neumann condition on Γ_N , yields the following **variational formulation**:

Determine $u \in H_{0,\Gamma_D}^1(\Omega)$ such that, $\forall v \in H_{0,\Gamma_D}^1(\Omega)$,

$$\int_{\Omega} \nabla u \cdot \nabla v \, d\mathbf{x} + \int_{\Omega} a u v \, d\mathbf{x} = \int_{\Omega} f v \, d\mathbf{x} + \int_{\Gamma_N} g v \, d\sigma.$$

Using the *trace* inequality (Theorem 7.85, p. 483)

$$\|v\|_{L^2(\Gamma_N)} \leq \tilde{C} \|v\|_{H^1(\Omega)}, \tag{8.20}$$

the proof of the next theorem follows the usual pattern.

Theorem 8.7. Let $\Omega \subset \mathbb{R}^n$ be a bounded Lipschitz domain. Assume $f \in L^2(\Omega)$, $g \in L^2(\Gamma_N)$ and $0 \leq a(\mathbf{x}) \leq \alpha_0$ a.e. in Ω . Then the mixed problem has a unique solution $u \in H_{0,\Gamma_D}^1(\Omega)$. Moreover:

$$\|\nabla u\|_{L^2(\Omega; \mathbb{R}^n)} \leq C_P \|f\|_{L^2(\Omega)} + \tilde{C} \|g\|_{L^2(\Gamma_N)}.$$

8.3.3 Eigenvalues and eigenfunctions of the Laplace operator

In Sect. 6.9.2 we have seen how the efficacy of the separation of variables method for a given problem relies on the existence of a basis of eigenfunctions associated with that problem. The abstract results in Sect. 6.9.4, concerning the spectrum of a weakly coercive bilinear form, constitute the appropriate tools for analyzing the spectral properties of uniformly elliptic operators and in particular of the Laplace operator. It is important to point out that the spectrum of a differential operator must be associated with specific homogeneous boundary conditions.

Thus, for instance, we may consider the *Dirichlet eigenfunctions* for the Laplace operator in a bounded domain Ω , i.e. the nontrivial solutions of the problem

$$\begin{cases} -\Delta u = \lambda u & \text{in } \Omega \\ u = 0 & \text{on } \partial\Omega. \end{cases} \quad (8.21)$$

A weak solution of problem (8.21) is a function $u \in H_0^1(\Omega)$ such that

$$(\nabla u, \nabla v)_{L^2(\Omega; \mathbb{R}^n)} = \lambda (u, v)_{L^2(\Omega)}, \quad \forall v \in H_0^1(\Omega).$$

Since Ω is bounded, the bilinear form is $H_0^1(\Omega)$ -coercive so that Theorem 6.74, p. 411, gives:

Theorem 8.8. Let Ω be a bounded domain. Then:

- a) There exists in $L^2(\Omega)$ an orthonormal basis $\{u_k\}_{k \geq 1}$ consisting of Dirichlet eigenfunctions for the Laplace operator.
- b) The corresponding eigenvalues $\{\lambda_k\}_{k \geq 1}$ are all positive and may be arranged in a nondecreasing sequence

$$0 < \lambda_1 < \lambda_2 \leq \cdots \leq \lambda_k \leq \cdots \quad (8.22)$$

with $\lambda_k \rightarrow +\infty$. Every eigenvalue appears in the sequence (8.22) a finite number of times, according to its multiplicity.

- c) The sequence $\{u_k/\sqrt{\lambda_k}\}_{k \geq 1}$ constitutes an orthonormal basis in $H_0^1(\Omega)$, with respect to the scalar product $(u, v)_{H_0^1(\Omega)} = (\nabla u, \nabla v)_{L^2(\Omega; \mathbb{R}^n)}$.

Actually, the strict inequality $\lambda_1 < \lambda_2$ in (8.22) is a consequence of Theorem 8.9 below, which states that λ_1 is a simple eigenvalue, that is of multiplicity 1.

From Theorem 6.76, p. 414, we deduce the following **variational principle** for λ_1 , the **principal Dirichlet eigenvalue**:

$$\lambda_1 = \min \{ R(u) : u \in H_0^1(\Omega), u \neq 0 \} \quad (8.23)$$

where $R(u)$ is the Rayleigh quotient:

$$R(v) = \frac{\int_{\Omega} |\nabla v|^2 d\mathbf{x}}{\int_{\Omega} v^2 d\mathbf{x}}. \quad (8.24)$$

Moreover, the minimum is attained at any *eigenvector* corresponding to λ_1 . Thus,

$$\|u\|_{L^2(\Omega)} \leq \frac{1}{\sqrt{\lambda_1}} \|\nabla u\|_{L^2(\Omega; \mathbb{R}^n)} \quad (8.25)$$

and equality holds when u is an eigenvector corresponding to λ_1 . An interesting consequence is that $1/\sqrt{\lambda_1}$ is the *best* (i.e. the *smallest*) Poincaré constant for the domain Ω .

More information on λ_1 and its eigenspace are contained in the following theorem.

Theorem 8.9. *Let Ω be a bounded, Lipschitz domain. Then the principal eigenvalue λ_1 is simple, that is the corresponding eigenspace is 1-dimensional. Moreover, every eigenfunction corresponding to λ_1 has constant sign in Ω .*

Proof. Let u_1 be an eigenvector corresponding to λ_1 . First we prove that u_1 has constant sign in Ω . Let $w = |u_1|$. By Proposition 7.68, p. 471, we have $\nabla w = \operatorname{sign}(u_1) \nabla u_1$ and, consequently, $R(w) = R(u_1)$. From Theorem 6.76, p. 414, we deduce that w also is an eigenvector corresponding to λ_1 , that is it satisfies the equation

$$(\nabla w, \nabla v)_{L^2(\Omega; \mathbb{R}^n)} = \lambda_1 (w, v)_{L^2(\Omega)}, \quad \forall v \in H^1(\Omega).$$

Theorems 8.34, p. 543, and 8.27, p. 538, give $w \in C^\infty(\Omega) \cap C(\overline{\Omega})$, so that w is a classical solution of the equation

$$-\Delta w = \lambda_1 w, \quad \text{in } \Omega$$

and $w = 0$ on $\partial\Omega$. Since $w \geq 0$ and $\lambda_1 > 0$, w is superharmonic in Ω and the maximum principle (see Sect. 3.4) gives either $w \equiv 0$ or $w > 0$ in Ω . Thus u_1 never vanishes in Ω and therefore, being continuous, it has constant sign in Ω .

Suppose now by contradiction that the λ_1 -eigenspace has dimension > 1 . Then there should exist another eigenfunction corresponding to λ_1 , orthogonal to u_1 in $L^2(\Omega)$:

$$\int_{\Omega} u_1 u_2 d\mathbf{x} = 0.$$

But this is impossible because every eigenfunction has constant sign in Ω . □

Also the other eigenvalues can be variationally characterized. For instance, denoting by V_1 the eigenspace corresponding to λ_1 , we have (see Problem 8.18):

$$\lambda_2 = \min \{ R(v) : v \neq 0, v \in H_0^1(\Omega) \cap V_1^\perp \}. \quad (8.26)$$

Similar results hold for the other types of boundary value problems as well. For instance, the *Neumann eigenfunctions* for the Laplace operator in Ω are the non-trivial solutions of the problem

$$\begin{cases} -\Delta u = \mu u & \text{in } \Omega \\ \partial_{\nu} u = 0 & \text{on } \partial\Omega. \end{cases} \quad (8.27)$$

A weak solution of (8.27) is a function $u \in H^1(\Omega)$ such that

$$B(u, v) \equiv (\nabla u, \nabla v)_{L^2(\Omega; \mathbb{R}^n)} = \mu(u, v)_{L^2(\Omega)}, \quad \forall v \in H^1(\Omega).$$

If Ω is a bounded Lipschitz domain, $H^1(\Omega)$ is compactly embedded into $L^2(\Omega)$ and the bilinear form B is clearly weakly coercive with constant, say, $\lambda_0 = 1$. Applying Theorem 6.76, we find:

Theorem 8.10. *Let Ω be a bounded Lipschitz domain.*

- a) *There exists in $L^2(\Omega)$ an orthonormal basis $\{u_k\}_{k \geq 1}$ consisting of Neumann eigenfunctions for the Laplace operator.*
- b) *The corresponding eigenvalues form a nondecreasing sequence*

$$0 = \mu_1 < \mu_2 \leq \cdots \leq \mu_k \leq \cdots \quad (8.28)$$

with $\mu_k \rightarrow +\infty$. Each eigenvalue appears in the sequence (8.28) a finite number of times, according to its multiplicity. In particular, the principal eigenvalue $\mu_1 = 0$ is simple.

- c) *The sequence $\{u_k/\sqrt{\mu_k + 1}\}_{k \geq 1}$ constitutes an orthonormal basis in $H^1(\Omega)$, with respect to the scalar product $(u, v)_{H^1(\Omega)} = (u, v)_{L^2(\Omega)} + (\nabla u, \nabla v)_{L^2(\Omega; \mathbb{R}^n)}$.*

8.3.4 An asymptotic stability result

The results in the last subsection may be used sometimes to prove the asymptotic stability of a steady state solution of an evolution equation as time $t \rightarrow +\infty$.

As an example, consider the following problem for the heat equation. Suppose that $u \in C^{2,1}(\overline{\Omega} \times [0, +\infty))$ is the (unique) solution of

$$\begin{cases} u_t - \Delta u = f(\mathbf{x}) & \mathbf{x} \in \Omega, t > 0 \\ u(\mathbf{x}, 0) = u_0(\mathbf{x}) & \mathbf{x} \in \Omega \\ u(\sigma, t) = 0 & \sigma \in \partial\Omega, t > 0, \end{cases}$$

where Ω is a smooth, bounded domain. Denote by $u_{\infty} = u_{\infty}(\mathbf{x})$ the solution of the stationary problem

$$\begin{cases} -\Delta u_{\infty} = f & \text{in } \Omega \\ u_{\infty} = 0 & \text{on } \partial\Omega. \end{cases}$$

Proposition 8.11. *For $t \geq 0$, we have*

$$\|u(\cdot, t) - u_\infty\|_{L^2(\Omega)} \leq e^{-\lambda_1 t} \left\{ \|u_0\|_{L^2(\Omega)} + \lambda_1^{-1} \|f\|_{L^2(\Omega)} \right\} \quad (8.29)$$

where λ_1 is the first Dirichlet eigenvalue for the Laplace operator in Ω .

Proof. Set $g(\mathbf{x}) = u_0(\mathbf{x}) - u_\infty(\mathbf{x})$. The function $w(\mathbf{x}, t) = u(\mathbf{x}, t) - u_\infty(\mathbf{x})$ solves the problem

$$\begin{cases} w_t - \Delta w = 0 & \mathbf{x} \in \Omega, t > 0 \\ w(\mathbf{x}, 0) = g(\mathbf{x}) & \mathbf{x} \in \Omega \\ w(\boldsymbol{\sigma}, t) = 0 & \boldsymbol{\sigma} \in \partial\Omega, t > 0. \end{cases} \quad (8.30)$$

Let us use the method of separation of variables and look for solutions of the form $w(\mathbf{x}, t) = v(\mathbf{x}) z(t)$. We find

$$\frac{z'(t)}{z(t)} = \frac{\Delta v(\mathbf{x})}{v(\mathbf{x})} = -\lambda$$

with λ constant. Thus we are led to the eigenvalue problem

$$\begin{cases} -\Delta v = \lambda v & \text{in } \Omega \\ v = 0 & \text{on } \partial\Omega. \end{cases}$$

From Theorem 8.8, there exists in $L^2(\Omega)$ an orthonormal basis $\{u_k\}_{k \geq 1}$ consisting of eigenvectors, corresponding to a sequence of nondecreasing eigenvalues $\{\lambda_k\}$, with $\lambda_1 > 0$ and $\lambda_k \rightarrow +\infty$. Then, if $g_k = (g, u_k)_{L^2(\Omega)}$, we can write

$$g = \sum_1^\infty g_k u_k \quad \text{and} \quad \|g\|_{L^2(\Omega)}^2 = \sum_{k=1}^\infty g_k^2.$$

As a consequence, we find $z_k(t) = e^{-\lambda_k t}$ and finally

$$w(\mathbf{x}, t) = \sum_1^\infty e^{-\lambda_k t} g_k u_k(\mathbf{x}).$$

Thus,

$$\|u(\cdot, t) - u_\infty\|_{L^2(\Omega)}^2 = \|w(\cdot, t)\|_{L^2(\Omega)}^2 = \sum_{k=1}^\infty e^{-2\lambda_k t} g_k^2$$

and, since $\lambda_k > \lambda_1$ for every k , we deduce that

$$\|u(\cdot, t) - u_\infty\|_{L^2(\Omega)}^2 \leq \sum_{k=1}^\infty e^{-2\lambda_1 t} g_k^2 = e^{-2\lambda_1 t} \|g\|_{L^2(\Omega)}^2.$$

Theorem 8.1, p. 511 yields, in particular, recalling (8.25),

$$\|u_\infty\|_{L^2(\Omega)} \leq \frac{1}{\lambda_1} \|f\|_{L^2(\Omega)},$$

and hence

$$\begin{aligned} \|g\|_{L^2(\Omega)} &\leq \|u_0\|_{L^2(\Omega)} + \|u_\infty\|_{L^2(\Omega)} \\ &\leq \|u_0\|_{L^2(\Omega)} + \frac{1}{\lambda_1} \|f\|_{L^2(\Omega)} \end{aligned}$$

giving (8.29). □

Proposition 8.11 implies that the steady state u_∞ is *asymptotically stable in* $L^2(\Omega)$ -norm as $t \rightarrow +\infty$. The speed of convergence is exponential and it is determined by the first eigenvalue⁴ λ_1 .

8.4 General Equations in Divergence Form

8.4.1 Basic assumptions

In this section we consider boundary value problems for elliptic operators with general diffusion and transport terms. Let $\Omega \subset \mathbb{R}^n$ be a **bounded domain** and set

$$\mathcal{E}u = -\operatorname{div}(\mathbf{A}(\mathbf{x}) \nabla u) - \mathbf{b}(\mathbf{x}) \cdot \nabla u + \mathbf{c}(\mathbf{x}) \cdot \nabla u + a(\mathbf{x}) u \quad (8.31)$$

where

$$\mathbf{A} = (a_{ij})_{i,j=1,\dots,n}, \quad \mathbf{b} = (b_1, \dots, b_n), \quad \mathbf{c} = (c_1, \dots, c_n)$$

and a is a real function.

Throughout this section, we will work under the following assumptions.

1. The differential operator \mathcal{E} is **uniformly elliptic**, i.e. there exist **positive numbers** α and M such that:

$$\alpha |\boldsymbol{\xi}|^2 \leq \sum_{i,j=1}^n a_{ij}(\mathbf{x}) \xi_i \xi_j \quad \text{and} \quad |a_{ij}(\mathbf{x})| \leq M, \quad \forall \boldsymbol{\xi} \in \mathbb{R}^n, \text{ for a.e. } \mathbf{x} \in \Omega. \quad (8.32)$$

2. The coefficients \mathbf{b} , \mathbf{c} and a are all **bounded**:

$$|\mathbf{b}(\mathbf{x})| \leq \beta_0, \quad |\mathbf{c}(\mathbf{x})| \leq \gamma_0, \quad |a(\mathbf{x})| \leq \alpha_0, \quad \text{a.e. in } \Omega. \quad (8.33)$$

The uniform ellipticity condition (8.32) states that \mathbf{A} is *positive*⁵ in Ω . If \mathbf{A} is symmetric its minimum eigenvalue is bounded from below by α , called *ellipticity constant*. We point out that at this level of generality, we allow discontinuities also of the diffusion matrix \mathbf{A} , of the transport coefficients \mathbf{b} and \mathbf{c} , in addition to the reaction coefficient a .

We want to extend to these type of operators the theory developed so far. The uniform ellipticity is a necessary requirement. In this section, we first indicate some sufficient conditions assuring the well-posedness of the usual boundary value problems, based on the use of the Lax-Milgram Theorem.

On the other hand, these conditions may be sometimes rather restrictive. When they are not satisfied, precise information on solvability and well-posedness can be obtained from Fredholm Alternative Theorem 6.66, p. 402.

As in the preceding sections, we start from the Dirichlet problem.

⁴ Compare with Subsect. 2.1.4.

⁵ If \mathbf{A} is only *nonnegative*, the equation is *degenerate elliptic* and things get too complicated for this introductory book.

8.4.2 Dirichlet problem

Consider the problem

$$\begin{cases} \mathcal{E}u = f + \operatorname{div} \mathbf{f} & \text{in } \Omega \\ u = 0 & \text{on } \partial\Omega, \end{cases} \quad (8.34)$$

where $f \in L^2(\Omega)$ and $\mathbf{f} \in L^2(\Omega; \mathbb{R}^n)$.

A comment on the right hand side of (8.34) is in order. We have denoted by $H^{-1}(\Omega)$ the dual of $H_0^1(\Omega)$. We know (Theorem 7.63, p. 468) that every element $F \in H^{-1}(\Omega)$ can be identified with an element in $\mathcal{D}'(\Omega)$ of the form

$$F = f + \operatorname{div} \mathbf{f}.$$

Moreover, since we are using in $H_0^1(\Omega)$ the norm $\|\nabla u\|_{L^2(\Omega)}$,

$$\|F\|_{H^{-1}(\Omega)} \leq C_P \|f\|_{L^2(\Omega)} + \|\mathbf{f}\|_{L^2(\Omega; \mathbb{R}^n)}. \quad (8.35)$$

Thus, the right hand side of (8.34) represents an element of $H^{-1}(\Omega)$.

As in Sect. 8.4, to derive a variational formulation of (8.34), we first assume that all the coefficients and the data f , \mathbf{f} are smooth. Then, we multiply the equation by a test function $v \in C_0^1(\Omega)$ and integrate over Ω :

$$\int_{\Omega} \{-\operatorname{div}(\mathbf{A}\nabla u - \mathbf{b}u) v\} d\mathbf{x} + \int_{\Omega} (\mathbf{c} \cdot \nabla u + au) v d\mathbf{x} = \int_{\Omega} (f + \operatorname{div} \mathbf{f}) v d\mathbf{x}.$$

Integrating by parts, we find, since $v = 0$ on $\partial\Omega$:

$$\int_{\Omega} \{-\operatorname{div}(\mathbf{A}\nabla u - \mathbf{b}u) v\} d\mathbf{x} = \int_{\Omega} \{\mathbf{A}\nabla u \cdot \nabla v - \mathbf{b}u \cdot \nabla v\} d\mathbf{x}$$

and

$$\int_{\Omega} v \operatorname{div} \mathbf{f} d\mathbf{x} = - \int_{\Omega} \mathbf{f} \cdot \nabla v d\mathbf{x}.$$

Thus, the resulting equation is:

$$\int_{\Omega} \{\mathbf{A}\nabla u \cdot \nabla v - \mathbf{b}u \cdot \nabla v + \mathbf{c}v \cdot \nabla u + auv\} d\mathbf{x} = \int_{\Omega} \{fv - \mathbf{f} \cdot \nabla v\} d\mathbf{x} \quad (8.36)$$

for every $v \in C_0^1(\Omega)$.

It is not difficult to check that if the domain and the other data are smooth, *for classical solutions, the two formulations (8.34) and (8.36) are equivalent*.

We now enlarge the space of test functions to $H_0^1(\Omega)$ and introduce the bilinear form

$$B(u, v) = \int_{\Omega} \{\mathbf{A}\nabla u \cdot \nabla v - \mathbf{b}u \cdot \nabla v + \mathbf{c}v \cdot \nabla u + auv\} d\mathbf{x}$$

and the linear functional

$$Fv = \int_{\Omega} \{ fv - \mathbf{f} \cdot \nabla v \} d\mathbf{x}.$$

Then, the **weak formulation** of problem (8.34) is the following:

Determine $u \in H_0^1(\Omega)$ such that

$$B(u, v) = Fv, \quad \forall v \in H_0^1(\Omega). \quad (8.37)$$

A set of hypotheses that ensure the well-posedness of the problem is indicated in the following proposition.

Proposition 8.12. Assume that hypotheses (8.32) and (8.33) hold and that $f \in L^2(\Omega)$, $\mathbf{f} \in L^2(\Omega; \mathbb{R}^n)$. Then, if \mathbf{b} and \mathbf{c} have Lipschitz components and

$$\frac{1}{2} \operatorname{div}(\mathbf{b} - \mathbf{c}) + a \geq 0, \text{ a.e. in } \Omega, \quad (8.38)$$

problem (8.37) has a unique solution. Moreover, the following stability estimate holds:

$$\|\nabla u\|_{L^2(\Omega; \mathbb{R}^n)} \leq \frac{1}{\alpha} \left\{ C_P \|f\|_{L^2(\Omega)} + \|\mathbf{f}\|_{L^2(\Omega; \mathbb{R}^n)} \right\}. \quad (8.39)$$

Proof. We apply the Lax-Milgram Theorem with $V = H_0^1(\Omega)$. The continuity of B in V follows easily. In fact, the Schwarz inequality and the bounds in (8.33) give:

$$\begin{aligned} \left| \int_{\Omega} \mathbf{A} \nabla u \cdot \nabla v \, d\mathbf{x} \right| &\leq \int_{\Omega} \sum_{i,j=1}^n |a_{ij} u_{x_i} v_{x_j}| \, d\mathbf{x} \\ &\leq n^2 M \int_{\Omega} |\nabla u| |\nabla v| \, d\mathbf{x} \leq n^2 M \|\nabla u\|_{L^2(\Omega; \mathbb{R}^n)} \|\nabla v\|_{L^2(\Omega; \mathbb{R}^n)}. \end{aligned}$$

Moreover, using Poincaré's inequality,

$$\left| \int_{\Omega} [\mathbf{b}u \cdot \nabla v - \mathbf{c}v \cdot \nabla u] \, d\mathbf{x} \right| \leq (\beta_0 + \gamma_0) C_P \|\nabla u\|_{L^2(\Omega; \mathbb{R}^n)} \|\nabla v\|_{L^2(\Omega; \mathbb{R}^n)}$$

and

$$\left| \int_{\Omega} a u v \, d\mathbf{x} \right| \leq \alpha_0 \int_{\Omega} |u| |v| \, d\mathbf{x} \leq \alpha_0 C_P^2 \|\nabla u\|_{L^2(\Omega; \mathbb{R}^n)} \|\nabla v\|_{L^2(\Omega; \mathbb{R}^n)}.$$

Thus, we can write

$$|B(u, v)| \leq (n^2 M + (\beta_0 + \gamma_0) C_P + \alpha_0 C_P^2) \|\nabla u\|_{L^2(\Omega; \mathbb{R}^n)} \|\nabla v\|_{L^2(\Omega; \mathbb{R}^n)}$$

which shows the continuity of B . Let us analyze the coercivity of B . We have:

$$B(u, u) = \int_{\Omega} \{ \mathbf{A} \nabla u \cdot \nabla u - (\mathbf{b} - \mathbf{c}) u \cdot \nabla u + au^2 \} \, d\mathbf{x}.$$

Observe that, since $u = 0$ on $\partial\Omega$, integrating by parts we obtain

$$\int_{\Omega} (\mathbf{b} - \mathbf{c}) u \cdot \nabla u \, d\mathbf{x} = \frac{1}{2} \int_{\Omega} (\mathbf{b} - \mathbf{c}) \cdot \nabla u^2 \, d\mathbf{x} = -\frac{1}{2} \int_{\Omega} \operatorname{div}(\mathbf{b} - \mathbf{c}) u^2 \, d\mathbf{x}.$$

Therefore, from (8.32) and (8.38), it follows that

$$B(u, u) \geq \alpha \int_{\Omega} |\nabla u|^2 \, d\mathbf{x} + \int_{\Omega} \left[\frac{1}{2} \operatorname{div}(\mathbf{b} - \mathbf{c}) + a \right] u^2 \, d\mathbf{x} \geq \alpha \|\nabla u\|_{L^2(\Omega; \mathbb{R}^n)}^2$$

so that B is V -coercive. Since we already know that $F \in H^{-1}(\Omega)$, the Lax-Milgram Theorem and (8.35) give existence, uniqueness and the stability estimate (8.39). \square

Remark 8.13. If \mathbf{A} is symmetric and $\mathbf{b} = \mathbf{c} = \mathbf{0}$, the solution u is a minimizer in $H_0^1(\Omega)$ for the “energy” functional

$$E(u) = \int_{\Omega} \{ \mathbf{A} \nabla u \cdot \nabla u + au^2 - 2fu \} \, d\mathbf{x}.$$

As in Remark 8.2, p. 511, the equation (8.37) constitutes the Euler equation for E .

- *Nonhomogeneous Dirichlet conditions.* If the Dirichlet condition is nonhomogeneous, i.e.

$$u = g \quad \text{on } \partial\Omega,$$

with $g \in H^{1/2}(\partial\Omega)$, we consider an extension \tilde{g} of g in $H^1(\Omega)$ and set $w = u - \tilde{g}$. In this case we require that Ω is at least a Lipschitz domain, to ensure the existence of \tilde{g} . Then $w \in H_0^1(\Omega)$ and it solves the equation

$$\mathcal{E}w = f + \operatorname{div}(\mathbf{f} + \mathbf{A} \nabla \tilde{g} - \mathbf{b} \tilde{g}) - \mathbf{c} \cdot \nabla \tilde{g} - a \tilde{g}.$$

From (8.32) and (8.33) we have

$$\mathbf{c} \cdot \nabla \tilde{g} + a \tilde{g} \in L^2(\Omega) \quad \text{and} \quad \mathbf{A} \nabla \tilde{g} - \mathbf{b} \tilde{g} \in L^2(\Omega; \mathbb{R}^n).$$

Therefore, the Lax-Milgram Theorem yields existence, uniqueness and the estimate

$$\|\nabla w\|_{L^2(\Omega; \mathbb{R}^n)} \leq C \left\{ \|f\|_{L^2(\Omega)} + \|\mathbf{f}\|_{L^2(\Omega; \mathbb{R}^n)} + \|\tilde{g}\|_{H^1(\Omega)} \right\}, \quad (8.40)$$

valid for any extension \tilde{g} of g , with $C = C(\alpha, n, M, \beta_0, \gamma_0, \alpha_0, \Omega)$.

Since $\|u\|_{H^1(\Omega)} \leq (1 + C_P) \|\nabla w\|_{L^2(\Omega; \mathbb{R}^n)} + \|\tilde{g}\|_{H^1(\Omega)}$ and recalling that (Sect. 7.9.3)

$$\|g\|_{H^{1/2}(\partial\Omega)} = \inf \left\{ \|\tilde{g}\|_{H^1(\Omega)} : \tilde{g} \in H^1(\Omega), \tilde{g}|_{\partial\Omega} = g \right\},$$

by taking the lowest upper bound with respect to \tilde{g} , from (8.40) we deduce, in terms of u :

$$\|u\|_{H^1(\Omega)} \leq C \left\{ \|f\|_{L^2(\Omega)} + \|\mathbf{f}\|_{L^2(\Omega; \mathbb{R}^n)} + \|g\|_{H^{1/2}(\partial\Omega)} \right\}.$$

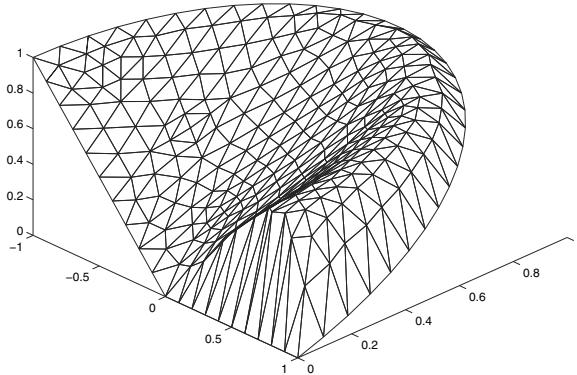


Fig. 8.1 The solution of problem (8.41)

Example 8.14. Figure 8.1 shows the solution of the following Dirichlet problem in the upper half circle $B_1^+(0,0) \subset \mathbb{R}^2$:

$$\begin{cases} -\Delta u - \rho u_\theta = 0 & \rho < 1, 0 < \theta < \pi \\ u(\rho, 1) = \sin(\theta/2) & 0 \leq \theta \leq \pi \\ u(\rho, 0) = 0, u(\rho, \pi) = -\rho & \rho \leq 1 \end{cases} \quad (8.41)$$

where (ρ, θ) denotes polar coordinates. Note that, in rectangular coordinates,

$$-\rho u_\theta = yu_x - xu_y$$

so that it represents a transport term of the type $\mathbf{c} \cdot \nabla u$ with $\mathbf{c} = (y, -x)$. Since $\operatorname{div} \mathbf{c} = 0$ and $a = 0$, Proposition 8.12 ensures the well posedness of the problem.

Alternative for the Dirichlet problem. We will see later that problem (8.34) is actually well posed under the condition

$$\operatorname{div} \mathbf{b} + a \geq 0 \quad \text{a.e. in } \Omega,$$

which does not involve the coefficient \mathbf{c} . In particular, this condition is fulfilled if $a(\mathbf{x}) \geq 0$ and $\mathbf{b}(\mathbf{x}) = \mathbf{0}$ a.e. in Ω . In general however, we cannot prove that the bilinear form B is coercive. What we may affirm is that B is **weakly coercive**, i.e. *there exists* $\lambda_0 \in \mathbb{R}$ such that:

$$\tilde{B}(u, v) = B(u, v) + \lambda_0 (u, v)_{L^2(\Omega)}$$

is coercive. In fact, from the elementary inequality $|ab| \leq \varepsilon a^2 + \frac{1}{4\varepsilon} b^2$, $\varepsilon > 0$, we get

$$\begin{aligned} \left| \int_{\Omega} (\mathbf{b} - \mathbf{c})u \cdot \nabla u \, d\mathbf{x} \right| &\leq (\beta_0 + \gamma_0) \int_{\Omega} |u \cdot \nabla u| \, d\mathbf{x} \\ &\leq \varepsilon \|\nabla u\|_{L^2(\Omega; \mathbb{R}^n)}^2 + \frac{(\beta_0 + \gamma_0)^2}{4\varepsilon} \|u\|_{L^2(\Omega)}^2. \end{aligned}$$

Therefore:

$$\tilde{B}(u, u) \geq (\alpha - \varepsilon) \|\nabla u\|_{L^2(\Omega; \mathbb{R}^n)}^2 + \left(\lambda_0 - \frac{(\beta_0 + \gamma_0)^2}{4\varepsilon} - \alpha_0 \right) \|u\|_{L^2(\Omega)}^2. \quad (8.42)$$

If we choose $\varepsilon = \alpha/2$ and $\lambda_0 = (\beta_0 + \gamma_0)^2/4\varepsilon + \alpha_0$, we obtain

$$\tilde{B}(u, u) \geq \frac{\alpha}{2} \|\nabla u\|_{L^2(\Omega; \mathbb{R}^n)}^2$$

which shows the coercivity of \tilde{B} . Introduce now the Hilbert triplet

$$V = H_0^1(\Omega), \quad H = L^2(\Omega), \quad V^* = H^{-1}(\Omega)$$

and recall that, since Ω is a **bounded** domain, $H_0^1(\Omega)$ is dense and compactly embedded in $L^2(\Omega)$. Finally, define the *adjoint bilinear form of B* by

$$B^*(u, v) = \int_{\Omega} \{(\mathbf{A}^\top \nabla u + \mathbf{c}u) \cdot \nabla v - \mathbf{b}v \cdot \nabla u + auv\} d\mathbf{x} \equiv B(v, u),$$

associated with the *formal adjoint* \mathcal{E}^* of \mathcal{E} , defined by

$$\mathcal{E}^*u = -\operatorname{div}(\mathbf{A}^\top \nabla u + \mathbf{c}u) - \mathbf{b} \cdot \nabla u + au.$$

We are now in position to apply Theorem 6.66, p. 402, to our variational problem. The conclusions are:

1) *The subspaces $\mathcal{N}(B)$ and $\mathcal{N}(B^*)$ of the solutions of the homogeneous problems*

$$B(u, v) = 0, \quad \forall v \in H_0^1(\Omega) \quad \text{and} \quad B^*(w, v) = 0, \quad \forall v \in H_0^1(\Omega),$$

share the same dimension d , $0 \leq d < \infty$.

2) *The problem*

$$B(u, v) = Fv, \quad \forall v \in H_0^1(\Omega)$$

has a solution if and only if $Fw = 0$ for every $w \in \mathcal{N}(B^)$.*

Let us translate the statements 1) and 2) into a less abstract language:

Theorem 8.15. *Let Ω be a bounded, Lipschitz domain, $f \in L^2(\Omega)$ and $\mathbf{f} \in L^2(\Omega; \mathbb{R}^n)$. Assume (8.32) and (8.33) hold. Then, we have the following alternative:*

a) *Either \mathcal{E} is a continuous isomorphism between $H_0^1(\Omega)$ and $H^{-1}(\Omega)$, that is problem (8.34) has a unique weak solution and there exists $C = C(\alpha, n, M, \beta_0, \gamma_0, \Omega)$ such that*

$$\|\nabla u\|_{L^2(\Omega; \mathbb{R}^n)} \leq C \left\{ \|f\|_{L^2(\Omega)} + \|\mathbf{f}\|_{L^2(\Omega; \mathbb{R}^n)} \right\},$$

or the homogeneous and the adjoint homogeneous problems, given, respectively, by

$$\begin{cases} \mathcal{E}u = 0 & \text{in } \Omega \\ u = 0 & \text{on } \partial\Omega, \end{cases} \quad \begin{cases} \mathcal{E}^*w = 0 & \text{in } \Omega \\ w = 0 & \text{on } \partial\Omega \end{cases} \quad (8.43)$$

have each d linearly independent solution, $0 < d < \infty$.

b) Moreover, problem (8.34) has a solution if and only if

$$\int_{\Omega} \{fw - \mathbf{f} \cdot \nabla w\} d\mathbf{x} = 0 \quad (8.44)$$

for every solution w of the adjoint homogeneous problem.

Theorem 8.15 implies that if we can show the uniqueness of the solution of problem (8.34), then automatically we infer both the existence and the stability estimate.

To show uniqueness, the weak maximum principles in Sect. 8.5.5 are quite useful. We will be back to this argument there.

The conditions (8.44) constitute d compatibility conditions that the data have to satisfy in order for a solution to exist.

8.4.3 Neumann problem

Let Ω be a bounded, Lipschitz domain. The Neumann condition for an operator in the divergence form (8.31) assigns on $\partial\Omega$ the flux naturally associated with the operator. This flux is composed by two terms: $\mathbf{A}\nabla u \cdot \boldsymbol{\nu}$, due to the diffusion term $-\operatorname{div}\mathbf{A}\nabla u$, and $-\mathbf{b}u \cdot \boldsymbol{\nu}$, due to the convective term $\operatorname{div}(\mathbf{b}u)$, where $\boldsymbol{\nu}$ is the outward unit normal on $\partial\Omega$. We set

$$\partial_{\boldsymbol{\nu}}^{\mathcal{E}} u \equiv (\mathbf{A}\nabla u - \mathbf{b}u) \cdot \boldsymbol{\nu} = \sum_{i,j=1}^n a_{ij}u_{x_j}\nu_i - u \sum_j b_j\nu_j.$$

We call $\partial_{\boldsymbol{\nu}}^{\mathcal{E}} u$ the conormal derivative of u . Thus, the Neumann problem is:

$$\begin{cases} \mathcal{E}u = f & \text{in } \Omega \\ \partial_{\boldsymbol{\nu}}^{\mathcal{E}} u = g & \text{on } \partial\Omega. \end{cases} \quad (8.45)$$

The variational formulation of problem (8.45) may be obtained by the usual integration by parts technique. It is enough to note, that, multiplying the differential equation $\mathcal{E}u = f$ by a test function $v \in H^1(\Omega)$ and using the Neumann condition, we get, formally:

$$\int_{\Omega} \{(\mathbf{A}\nabla u - \mathbf{b}u)\nabla v + (\mathbf{c} \cdot \nabla u)v + auv\} d\mathbf{x} = \int_{\Omega} fv d\mathbf{x} + \int_{\partial\Omega} gv d\sigma.$$

Introducing the bilinear form

$$B(u, v) = \int_{\Omega} \{(\mathbf{A}\nabla u - \mathbf{b}u)\nabla v + (\mathbf{c} \cdot \nabla u)v + auv\} d\mathbf{x} \quad (8.46)$$

and the linear functional

$$Fv = \int_{\Omega} fv d\mathbf{x} + \int_{\partial\Omega} gv d\sigma,$$

we are led to the following **weak formulation**, that can be easily checked to be equivalent to the original problem, when all the data are smooth:

Determine $u \in H^1(\Omega)$ such that

$$B(u, v) = Fv, \quad \forall v \in H^1(\Omega). \quad (8.47)$$

If the size of $\mathbf{b} - \mathbf{c}$ is small compared to the diffusive and reaction terms, problem (8.47) is well-posed, as the following proposition shows.

Proposition 8.16. Assume that hypotheses (8.32) and (8.33) hold and that $f \in L^2(\Omega)$, $g \in L^2(\partial\Omega)$. If $a(\mathbf{x}) \geq c_0 > 0$ a.e. in Ω and

$$m \equiv \min \{\alpha - (\beta_0 + \gamma_0)/2, c_0 - (\beta_0 + \gamma_0)/2\} > 0, \quad (8.48)$$

then, problem (8.47) has a unique solution. Moreover, the following stability estimate holds:

$$\|u\|_{H^1(\Omega)} \leq \frac{1}{m} \left\{ \|f\|_{L^2(\Omega)} + \overline{C}(n, \Omega) \|g\|_{L^2(\partial\Omega)} \right\}.$$

Proof. We check that the assumptions of Lax-Milgram theorem hold. Since

$$|B(u, v)| \leq (n^2 M + \beta_0 + \gamma_0 + \alpha_0) \|u\|_{H^1(\Omega)} \|v\|_{H^1(\Omega)},$$

B is continuous in $H^1(\Omega)$. Moreover, we may write

$$B(u, u) \geq \alpha \int_{\Omega} |\nabla u|^2 d\mathbf{x} - \left| \int_{\Omega} [(\mathbf{b} - \mathbf{c}) \cdot \nabla u] u d\mathbf{x} \right| + \int_{\Omega} au^2 d\mathbf{x}.$$

From Schwarz's inequality and the inequality $2ab \leq a^2 + b^2$, we obtain

$$\left| \int_{\Omega} [(\mathbf{b} - \mathbf{c}) \cdot \nabla u] u d\mathbf{x} \right| \leq (\beta_0 + \gamma_0) \|\nabla u\|_{L^2(\Omega; \mathbb{R}^n)} \|u\|_{L^2(\Omega)} \leq \frac{(\beta_0 + \gamma_0)}{2} \|u\|_{H^1(\Omega)}^2.$$

Thus, if (8.48) holds, we get $B(u, u) \geq m \|u\|_{H^1(\Omega)}^2$ and therefore B is coercive. Finally, using (8.17), it is not difficult to check that $F \in H^1(\Omega)^*$, with

$$\|F\|_{H^1(\Omega)^*} \leq \|f\|_{L^2(\Omega)} + \overline{C}(n, \Omega) \|g\|_{L^2(\partial\Omega)}. \quad \square$$

Alternative for the Neumann problem. The bilinear form B is coercive also under the conditions

$$(\mathbf{b} - \mathbf{c}) \cdot \boldsymbol{\nu} \leq 0 \quad \text{a.e. on } \partial\Omega \quad \text{and} \quad \frac{1}{2}\operatorname{div}(\mathbf{b} - \mathbf{c}) + a \geq c_0 > 0 \quad \text{a.e. in } \Omega,$$

as it can be checked following the proof of Proposition 8.12, p. 523.

However, in general the bilinear form B is only *weakly coercive*. In fact, choosing in (8.42) $\varepsilon = \alpha/2$ and $\lambda_0 = (\beta_0 + \gamma_0)^2/2\varepsilon + 2\alpha_0$, we easily get

$$\tilde{B}(u, u) = B(u, u) + \lambda_0 \|u\|_{L^2(\Omega)}^2 \geq \frac{\alpha}{2} \|\nabla u\|_{L^2(\Omega)}^2 + \left(\frac{(\beta_0 + \gamma_0)^2}{2\alpha} + \alpha_0 \right) \|u\|_{L^2(\Omega)}^2$$

and therefore \tilde{B} is coercive. Applying Theorem 6.66, p. 402, we obtain the following alternative:

Theorem 8.17. Let Ω be a bounded, Lipschitz domain. Assume that (8.32) and (8.33) hold. Then, if $f \in L^2(\Omega)$ and $g \in L^2(\partial\Omega)$:

- a) Either problem (8.45) has a unique solution $u \in H^1(\Omega)$ and

$$\|u\|_{H^1(\Omega)} \leq C(n, \alpha, M, \beta_0, \gamma_0, \alpha_0, \Omega) \left\{ \|f\|_{L^2(\Omega)} + \|g\|_{L^2(\partial\Omega)} \right\}$$

or the homogeneous and the adjoint homogeneous problems

$$\begin{cases} \mathcal{E}u = 0 & \text{in } \Omega \\ (\mathbf{A}\nabla u - \mathbf{b}u) \cdot \boldsymbol{\nu} = 0 & \text{on } \partial\Omega \end{cases}$$

and

$$\begin{cases} \mathcal{E}^*w = 0 & \text{in } \Omega \\ (\mathbf{A}^\top \nabla w + \mathbf{c}w) \cdot \boldsymbol{\nu} = 0 & \text{on } \partial\Omega \end{cases}$$

have each d linearly independent solutions, $0 < d < \infty$.

- b) Moreover, problem (8.47) has a solution if and only if

$$Fw = \int_\Omega fw \, d\mathbf{x} + \int_{\partial\Omega} gw \, d\sigma = 0 \tag{8.49}$$

for every solution w of the adjoint homogeneous problem.

Remark 8.18. Again, uniqueness implies existence. Note that if $\mathbf{b} = \mathbf{c} = \mathbf{0}$ and $a = 0$, then the solutions of the adjoint homogeneous problem are the constant functions. Therefore $d = 1$ and the compatibility condition (8.49) reduces to the well known equation

$$\int_\Omega f \, d\mathbf{x} + \int_{\partial\Omega} g \, d\sigma = 0.$$

8.4.4 Robin and mixed problems

The variational formulation of the Robin problem

$$\begin{cases} \mathcal{E}u = f & \text{in } \Omega \\ \partial_{\nu}^{\mathcal{E}}u + hu = g & \text{on } \partial\Omega \end{cases} \quad (8.50)$$

is obtained by replacing the bilinear form B in problem (8.47), by

$$\tilde{B}(u, v) = B(u, v) + \int_{\partial\Omega} huv \, d\sigma.$$

If $0 \leq h(\mathbf{x}) \leq h_0$ a.e. on $\partial\Omega$, Proposition 8.16 still holds for problem (8.50).

As for the Neumann problem, in general the bilinear form B is only *weakly coercive* and a theorem perfectly analogous to Theorem 8.17 holds. We leave the details as an exercise.

Let now Γ_D, Γ_N be nonempty, relatively open and regular subsets of $\partial\Omega$ with $\Gamma_N = \partial\Omega \setminus \overline{\Gamma}_D$. Consider the mixed Dirichlet-Neumann problem

$$\begin{cases} \mathcal{E}u = f & \text{in } \Omega \\ u = 0 & \text{on } \Gamma_D \\ \partial_{\nu}^{\mathcal{E}}u = g & \text{on } \Gamma_N. \end{cases}$$

As in Sect. 8.4.2, the correct functional setting is $H_{0,\Gamma_D}^1(\Omega)$ with the norm $\|u\|_{H_{0,\Gamma_D}^1(\Omega)} = \|\nabla u\|_{L^2(\Omega; \mathbb{R}^n)}$. Introducing the linear functional

$$Fv = \int_{\Omega} fv \, d\mathbf{x} + \int_{\Gamma_N} gv \, d\sigma,$$

the **variational formulation** is the following: *Determine $u \in H_{0,\Gamma_D}^1(\Omega)$ such that*

$$B(u, v) = Fv, \quad \forall v \in H_{0,\Gamma_D}^1(\Omega). \quad (8.51)$$

Proceeding as in Proposition 8.12, p. 523, we may prove the following result:

Proposition 8.19. *Assume that hypotheses (8.32) and (8.33) hold and that $f \in L^2(\Omega)$, $g \in L^2(\Gamma_N)$. If \mathbf{b} and \mathbf{c} have Lipschitz components in $\overline{\Omega}$ and*

$$(\mathbf{b} - \mathbf{c}) \cdot \boldsymbol{\nu} \leq 0 \quad \text{a.e. on } \Gamma_N, \quad \frac{1}{2} \operatorname{div}(\mathbf{b} - \mathbf{c}) + a \geq 0, \quad \text{a.e. in } \Omega,$$

then problem (8.51) has a unique solution $u \in H_{0,\Gamma_D}^1(\Omega)$. Moreover, the following stability estimate holds:

$$\|\nabla u\|_{L^2(\Omega; \mathbb{R}^n)} \leq \frac{1}{\alpha} \left\{ C_P \|f\|_{L^2(\Omega)} + \overline{C} \|g\|_{L^2(\partial\Omega)} \right\}.$$

For the mixed problem as well, in general the bilinear form is only weakly coercive and we may resort to the Alternative Theorem, achieving a result similar to Theorem 8.17.

Only note that the compatibility conditions (8.49) take the form

$$Fw = \int_{\Omega} fw \, d\mathbf{x} + \int_{\Gamma_N} gw \, d\sigma = 0$$

for every solution w of the adjoint problem

$$\begin{cases} \mathcal{E}^* w = 0 & \text{in } \Omega \\ w = 0 & \text{on } \Gamma_D \\ (\mathbf{A}^\top \nabla w + \mathbf{c} w) \cdot \boldsymbol{\nu} = 0 & \text{on } \Gamma_N. \end{cases}$$

8.5 Weak Maximum Principles

In Chap. 2 we have given a version of the maximum principle for the Laplace equation. This principle has an extension valid for general divergence form operators.

However, due to the Sobolev functional setting, we need to introduce a notion of “positivity or negativity on $\partial\Omega$ ” which is stronger than the almost everywhere sense on $\partial\Omega$.

Let Ω be a bounded, Lipschitz domain and $u \in H^1(\Omega)$. Since $C^1(\overline{\Omega})$ is dense in $H^1(\Omega)$, we say that $u \geq 0$ on $\partial\Omega$ (in the sense of H^1) if there exists a sequence

$$\{v_k\}_{k \geq 1} \subset C^1(\overline{\Omega})$$

such that $v_k \rightarrow u$ in $H^1(\Omega)$ and $v_k \geq 0$. It is as if the trace $\tau_0 u$ of u on $\partial\Omega$ “inherits” the nonnegativity from the sequence $\{v_k\}_{k \geq 1}$. Clearly, if $u \geq 0$ on $\partial\Omega$ in the above sense, the trace $\tau_0 u$ is nonnegative a.e. on $\partial\Omega$, but the reverse implication is not true in general.⁶

Since $v_k \geq 0$ on $\partial\Omega$ is equivalent to saying that the negative part⁷

$$v_k^- = \max\{-v_k, 0\}$$

has zero trace on $\partial\Omega$, it turns out that $u \geq 0$ on $\partial\Omega$ if and only if $u^- \in H_0^1(\Omega)$. Similarly, $u \leq 0$ on $\partial\Omega$ if and only if $u^+ \in H_0^1(\Omega)$.

Other inequalities follow in a natural way. For instance, we have $u \leq v$ on $\partial\Omega$ if $u - v \leq 0$ on $\partial\Omega$. Thus, we may define:

$$\sup_{\partial\Omega} u = \inf \{k \in \mathbb{R} : u \leq k \text{ on } \partial\Omega\}, \quad \inf_{\partial\Omega} u = \sup \{k \in \mathbb{R} : u \geq k \text{ on } \partial\Omega\}$$

which coincide with the usual *greatest lower bound* and *lowest upper bound* when the trace of u belongs to $C(\partial\Omega)$.

⁶ See [40], Ziemer, 1989.

⁷ Recall from Sect. 7.5.2 that, if $u \in H^1(\Omega)$ then its positive and negative part, $u^+ = \max\{u, 0\}$ and $u^- = \max\{-u, 0\}$, belong to $H^1(\Omega)$ as well.

Assume now that $u \in H^1(\Omega)$ satisfies the inequality

$$\mathcal{E}u = -\operatorname{div}(\mathbf{A}(\mathbf{x}) \nabla u) - \mathbf{b}(\mathbf{x}) \cdot \nabla u + \mathbf{c}(\mathbf{x}) \cdot \nabla u + a(\mathbf{x}) u \leq 0 \quad (\geq 0)$$

in a *weak* sense, that is, for every $v \in H_0^1(\Omega)$, $v \geq 0$ a.e. in Ω ,

$$B(u, v) = \int_{\Omega} \{(\mathbf{A} \nabla u - \mathbf{b} u) \cdot \nabla v + \mathbf{c} v \cdot \nabla u + a u v\} d\mathbf{x} \leq 0 \quad (\geq 0). \quad (8.52)$$

The following weak maximum principle holds.

Theorem 8.20. *Assume that (8.32) and (8.33) hold. Let $u \in H^1(\Omega)$ and $\mathcal{E}u \leq 0$ ($\mathcal{E}u \geq 0$) in a weak sense.*

a) *If $\mathbf{b} = \mathbf{0}$ and $a = 0$ a.e. in Ω , then*

$$\operatorname{ess\,sup}_{\Omega} u \leq \sup_{\partial\Omega} u \quad \left(\operatorname{ess\,inf}_{\Omega} u \geq \inf_{\partial\Omega} u \right). \quad (8.53)$$

b) *If \mathbf{b} be Lipschitz continuous⁸ and $\operatorname{div}\mathbf{b} + a \geq 0$ a.e. in Ω , then,*

$$\operatorname{ess\,sup}_{\Omega} u \leq \sup_{\partial\Omega} u^+ \quad \left(\operatorname{ess\,inf}_{\Omega} u \geq \inf_{\partial\Omega} (-u^-) \right). \quad (8.54)$$

In particular, in both cases, if $u \leq 0$ ($u \geq 0$) on $\partial\Omega$, then $u \leq 0$ ($u \geq 0$) a.e. in Ω .

Proof. We prove only the first of the two inequalities in (8.53) and (8.54). The proof of the others is perfectly analogous.

a) Let $l = \sup_{\partial\Omega} u$. We may assume that $l < \infty$, otherwise there is nothing to prove. Suppose that

$$l < \Lambda = \operatorname{ess\,sup}_{\Omega} u.$$

We want to reach a contradiction. Let λ be such that $l < \lambda < \Lambda$ and select as a test function

$$v_{\lambda} = \max\{u - \lambda, 0\} = (u - \lambda)^+,$$

which clearly belongs to $H_0^1(\Omega)$. Note that $v_{\lambda} > 0$ on a set of positive measure. Let G_{λ} be the support of $|\nabla v_{\lambda}|$. Then $\nabla v_{\lambda} = \nabla u$ a.e. on G_{λ} while $|\nabla v_{\lambda}| = 0$ a.e. outside G_{λ} . In other terms,

$$G_{\lambda} = \operatorname{supp} |\nabla u| \cap \{u > \lambda\}.$$

Inserting v_{λ} into (8.52), we find

$$\int_{\Omega} \mathbf{A} \nabla u \cdot \nabla v_{\lambda} d\mathbf{x} = \int_{G_{\lambda}} \mathbf{A} \nabla v_{\lambda} \cdot \nabla v_{\lambda} d\mathbf{x} \leq - \int_{G_{\lambda}} \mathbf{c} v_{\lambda} \cdot \nabla v_{\lambda} d\mathbf{x}. \quad (8.55)$$

⁸ With a little more effort, condition a) can be replaced by $\mathbf{b} \in L^{\infty}(\Omega; \mathbb{R}^n)$ and $\operatorname{div}\mathbf{b} + a \geq 0$ in $\mathcal{D}'(\Omega)$, that is

$$\int_{\Omega} \{-\mathbf{b} \cdot \nabla \varphi + a \varphi\} \geq 0 \quad \forall \varphi \in \mathcal{D}(\Omega), \varphi \geq 0 \text{ in } \Omega.$$

Using the uniform ellipticity condition (8.32) and the bound (8.33) for \mathbf{c} , we obtain

$$\begin{aligned} \alpha \|\nabla v_\lambda\|_{L^2(G_\lambda; \mathbb{R}^n)}^2 &\leq \gamma_0 \int_{G_\lambda} |v_\lambda \nabla v_\lambda| d\mathbf{x} \\ (\text{Schwarz's inequality}) &\leq \gamma_0 \|v_\lambda\|_{L^2(G_\lambda)} \|\nabla v_\lambda\|_{L^2(G_\lambda; \mathbb{R}^n)}. \end{aligned}$$

Simplifying by $\|\nabla v_\lambda\|_{L^2(G_\lambda; \mathbb{R}^n)}$, we have

$$\|\nabla v_\lambda\|_{L^2(G_\lambda; \mathbb{R}^n)} \leq \frac{\gamma_0}{\alpha} \|v_\lambda\|_{L^2(G_\lambda)}. \quad (8.56)$$

Consider now $n > 2$. Using the Sobolev Embedding Theorem 7.96, p. 492, and Hölder's inequality we get

$$\begin{aligned} \|v_\lambda\|_{L^{2n/(n-2)}(\Omega)} &\leq C_s \|\nabla v_\lambda\|_{L^2(\Omega; \mathbb{R}^n)} = C_s \|\nabla v_\lambda\|_{L^2(G_\lambda; \mathbb{R}^n)} \\ &\leq \frac{C_s \gamma_0}{\alpha} |G_\lambda|^{1/n} \|v_\lambda\|_{L^{2n/(n-2)}(\Omega)} \end{aligned}$$

or

$$|G_\lambda| \geq \alpha^n \gamma_0^{-n} C_s^{-n} \equiv \xi > 0. \quad (8.57)$$

Since ξ is independent of λ , the inequality (8.57) remains valid if we let $\lambda \rightarrow \Lambda$. We infer that $\text{supp}|\nabla u| \cap \{u = \Lambda\}$ is a set of positive measure. But $\nabla u = \mathbf{0}$ a.e. on every set of positive measure where u is constant and we have reached a contradiction.

If $n = 2$ an inequality similar to (8.57) follows by using the Sobolev Embedding Theorem with any $p > 2$. We leave the details to the reader.

b) Let $l = \sup_{\partial\Omega} u^+$, $v_\lambda = (u - \lambda)^+$ with λ such that $l < \lambda < \Lambda = \text{ess sup}_\Omega u$. Since \mathbf{b} is Lipschitz continuous, we can integrate by parts to write

$$\int_\Omega -\mathbf{b}u \cdot \nabla v d\mathbf{x} = \int_\Omega \text{div}(\mathbf{b}u)v d\mathbf{x} = \int_\Omega \{\text{div}\mathbf{b} uv + \mathbf{b}v \cdot \nabla u\} d\mathbf{x}.$$

Inserting into (8.52) we get, after a simple rearrangement of terms:

$$\int_\Omega \mathbf{A} \nabla u \cdot \nabla v d\mathbf{x} \leq - \int_\Omega \{(\mathbf{c} + \mathbf{b})v \cdot \nabla u + (\text{div}\mathbf{b} + a)uv\} d\mathbf{x}. \quad (8.58)$$

Choosing v_λ as a test function in (8.58), we find

$$\begin{aligned} \int_\Omega \mathbf{A} \nabla u \cdot \nabla v_\lambda d\mathbf{x} &= \int_{G_\lambda} \mathbf{A} \nabla v_\lambda \cdot \nabla v_\lambda d\mathbf{x} \\ &\leq - \int_{G_\lambda} (\mathbf{c} + \mathbf{b})v_\lambda \cdot \nabla v_\lambda d\mathbf{x} - \int_{\{u > \lambda\}} (\text{div}\mathbf{b} + a)uv_\lambda d\mathbf{x} \\ &\leq - \int_{G_\lambda} (\mathbf{c} + \mathbf{b})v_\lambda \cdot \nabla v_\lambda d\mathbf{x} \end{aligned}$$

since $\text{div}\mathbf{b} + a \geq 0$ a.e. in Ω and $uv_\lambda > 0$ a.e. on $\{u > \lambda\}$.

From now on the proof proceeds as in a). □

Remark 8.21. It is not possible to substitute $\sup_{\partial\Omega} u^+$ by $\sup_{\partial\Omega} u$ or $\inf_{\partial\Omega} (-u^-)$ with $\inf_{\partial\Omega} u$ in (8.54). A counterexample in dimension one is shown in Fig. 8.2. The solution of $-u'' + u = 0$ in $(0, 1)$, $u(0) = u(1) = -1$, has a negative maximum which is greater than -1 .

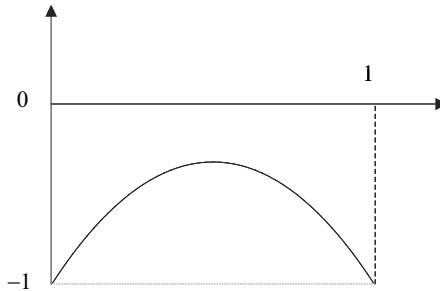


Fig. 8.2 The solution of $-u'' + u = 0$ in $(0, 1)$, $u(0) = u(1) = -1$

Theorem 8.20, in conjunction with the Alternative Theorem 8.15, p. 526, gives the following uniqueness result for the Dirichlet problem.

Corollary 8.22. *Under the hypotheses of Theorem 8.20, for every $f \in L^2(\Omega)$, $\mathbf{f} \in L^2(\Omega; \mathbb{R}^n)$, $g \in H^{1/2}(\partial\Omega)$, the Dirichlet problem*

$$\begin{cases} \mathcal{L}u = f + \operatorname{div} \mathbf{f} & \text{in } \Omega \\ u = g & \text{on } \partial\Omega \end{cases}$$

has a unique solution $u \in H^1(\Omega)$ and

$$\|u\|_{H^1(\Omega)} \leq C(n, \alpha, M, \beta_0, \gamma_0, \alpha_0, \Omega) \left\{ \|f\|_{L^2(\Omega)} + \|\mathbf{f}\|_{L^2(\Omega)} + \|g\|_{H^{1/2}(\partial\Omega)} \right\}.$$

Proof. We use Fredholm's Alternative. The bilinear form B in (8.52) is weakly coercive in $H_0^1(\Omega)$. Therefore, to prove the Corollary it is enough to show uniqueness for the Dirichlet problem. In fact, let u, w be two solutions and set

$$z = u - w \in H_0^1(\Omega).$$

Then z solves the homogeneous problem, that is $B(z, v) = 0$, for every $v \in H_0^1(\Omega)$. From Theorem 8.20 we deduce that

$$0 = \inf_{\partial\Omega} (-z^-) \leq \operatorname{ess\,inf}_{\Omega} z \leq \operatorname{ess\,sup}_{\Omega} z \leq \sup_{\partial\Omega} z^+ = 0.$$

Therefore $z = 0$ a.e. in Ω . □

There are similar maximum principles yielding uniqueness and therefore well-posedness, for problems involving mixed conditions.

To deal with mixed Dirichlet-Neumann problem we, consider two regular and relatively open set $\Gamma_D, \Gamma_N \subset \partial\Omega$, with $\Gamma_D \neq \emptyset$ and $\Gamma_N = \partial\Omega \setminus \overline{\Gamma}_D$. Here we say that $u \geq 0$ (resp. $u \leq 0$) on Γ_D if and only if $u^- \in H_{0,\Gamma_D}^1(\Omega)$ (resp. $u^+ \in H_{0,\Gamma_D}^1(\Omega)$).

With the same technique used for Theorem 8.20, we can prove the following result (compare with Example 8.31, p. 540).

Theorem 8.23. Assume that (8.32) and (8.33) hold. Let $u \in H^1(\Omega)$ satisfy

$$\int_{\Omega} \{(\mathbf{A}\nabla u - \mathbf{b}u) \cdot \nabla v + \mathbf{c}v \cdot \nabla u + auv\} d\mathbf{x} \leq 0 \quad (\geq 0) \quad (8.59)$$

for every $v \in H_{0,\Gamma_D}^1(\Omega)$, $v \geq 0$ a.e. in Ω .

a) If $\mathbf{b} = \mathbf{0}$ and $a = 0$, then

$$ess\sup_{\Omega} u \leq \sup_{\Gamma_D} u \quad \left(ess\inf_{\Omega} u \geq \inf_{\Gamma_D} (-u) \right). \quad (8.60)$$

b) If \mathbf{b} is Lipschitz continuous and

$$\mathbf{b} \cdot \boldsymbol{\nu} \leq 0 \quad \text{a.e. on } \Gamma_N, \quad \operatorname{div} \mathbf{b} + a \geq 0, \quad \text{a.e. in } \Omega, \quad (8.61)$$

then

$$ess\sup_{\Omega} u \leq \sup_{\Gamma_D} u^+ \quad \left(ess\inf_{\Omega} u \geq \inf_{\Gamma_D} (-u^-) \right). \quad (8.62)$$

In particular, in both cases, if $u \leq 0$ (resp. $u \geq 0$) on Γ_D , then $u \leq 0$ (resp. $u \geq 0$) a.e. in Ω .

Proof. We prove only the first of the two inequalities in (8.60) and (8.62). The proof of the others is perfectly analogous.

a) Let $l = \sup_{\Gamma_D} u < \infty$. Assume $l < \Lambda = \operatorname{esssup}_{\Omega} u$ and set:

$$v_{\lambda} = (u - \lambda)^+, \quad \text{with } l < \lambda < \Lambda.$$

Note that v_{λ} belongs to $H_{0,\Gamma_D}^1(\Omega)$, and repeat the proof in Theorem 8.20 a).

b) Let $l = \sup_{\partial\Omega} u^+ < \infty$ and, as in a), $v_{\lambda} = (u - \lambda)^+$, $l < \lambda < \Lambda$. Since \mathbf{b} is Lipschitz continuous, we can integrate by parts to write

$$\begin{aligned} \int_{\Omega} -\mathbf{b}u \cdot \nabla v \, d\mathbf{x} &= \int_{\Omega} \operatorname{div}(\mathbf{b}u)v \, d\mathbf{x} - \int_{\Gamma_N} uv \mathbf{b} \cdot \boldsymbol{\nu} \, d\sigma \\ &= \int_{\Omega} [\operatorname{div} \mathbf{b} \, uv + \mathbf{b}v \cdot \nabla u] \, d\mathbf{x} - \int_{\Gamma_N} uv \mathbf{b} \cdot \boldsymbol{\nu} \, d\sigma. \end{aligned}$$

Inserting into (8.59) we get, after a simple rearrangements of terms:

$$\int_{\Omega} \mathbf{A}\nabla u \cdot \nabla v \, d\mathbf{x} \leq - \int_{\Omega} \{(\mathbf{c} + \mathbf{b})v \cdot \nabla u + (\operatorname{div} \mathbf{b} + a)uv\} \, d\mathbf{x} + \int_{\Gamma_N} uv \mathbf{b} \cdot \boldsymbol{\nu} \, d\sigma.$$

Choosing $v = v_{\lambda}$ as a test function, we find (keeping the notations in the proof of Theorem 8.20)

$$\int_{G_{\lambda}} \mathbf{A}\nabla v_{\lambda} \cdot \nabla v_{\lambda} \, d\mathbf{x} \leq - \int_{G_{\lambda}} (\mathbf{c} + \mathbf{b})v_{\lambda} \cdot \nabla v_{\lambda} \, d\mathbf{x}$$

since $\operatorname{div} \mathbf{b} + a \geq 0$ a.e. in Ω , $\mathbf{b} \cdot \boldsymbol{\nu} \leq 0$ a.e. on Γ_N and $uv_{\lambda} > 0$ a.e. on $\{u > \lambda\}$.

The rest of the proof follows as in Theorem 8.20). \square

Using Theorem 6.66, p. 402, we have:

Theorem 8.24. *Under the hypotheses of Theorem 8.23, for every $f \in L^2(\Omega)$, $g \in L^2(\Gamma_N)$, the mixed Dirichlet-Neumann problem*

$$\begin{cases} \mathcal{E}u = f & \text{in } \Omega \\ u = 0 & \text{on } \Gamma_D \\ \partial_{\nu}^{\mathcal{E}}u = g & \text{on } \Gamma_N \end{cases}$$

has a unique solution $u \in H^1(\Omega)$ and

$$\|u\|_{H^1(\Omega)} \leq C(n, \alpha, M, \beta_0, \gamma_0, \alpha_0, \Omega) \left\{ \|f\|_{L^2(\Omega)} + \|g\|_{L^2(\Gamma_N)} \right\}.$$

Proof. In fact, let u, w be two solutions and set $z = u - w$. Then $z \in H_{0,\Gamma_D}^1(\Omega)$ and solves the homogeneous problem, that is $B(z, v) = 0$, for every $v \in H_{0,\Gamma_D}^1(\Omega)$. From Theorem 8.23 we deduce that

$$0 = \inf_{\Gamma_D} (-z^-) \leq \operatorname{ess\,inf}_{\Omega} z \leq \operatorname{ess\,sup}_{\Omega} z \leq \sup_{\Gamma_D} z^+ = 0.$$

Therefore $z = 0$ a.e. in Ω . The conclusions follow from Fredholm's Alternative. \square

8.6 Regularity

An important task, in general rather technically complicated, is to establish the optimal regularity of a weak solution in relation to the degree of smoothness of the data, namely, *the domain Ω , the boundary data, the coefficients of the operator and the forcing term*. To get a clue of what happens, consider for example the following problem:

$$\begin{cases} -\Delta u + u = F & \text{in } \Omega \\ u = 0 & \text{on } \partial\Omega \end{cases}$$

where $F \in H^{-1}(\Omega)$. Under this hypothesis, the Lax-Milgram Theorem yields a solution $u \in H_0^1(\Omega)$ and we cannot get much more, in terms of smoothness. Indeed, from Sobolev inequalities (see Subsect. 7.10.4) it follows that $u \in L^p(\Omega)$ with $p = \frac{2n}{n-2}$, if $n \geq 3$, or $u \in L^p(\Omega)$, with $2 \leq p < \infty$, if $n = 2$. However, this gain in integrability does not seriously increase the smoothness of u .

Reversing our point of view, we may say that, starting from a function in $H_0^1(\Omega)$ and applying to it a second order operator “two orders of differentiability are lost”: the loss of one order drives from $H_0^1(\Omega)$ into $L^2(\Omega)$ while a further loss leads to $H^{-1}(\Omega)$. It is as if the upper index -1 indicates a “lack” of one order of differentiability.

Nevertheless, consider the case in which $u \in H^1(\mathbb{R}^n)$ is a solution of the equation

$$-\Delta u + u = f \quad \text{in } \mathbb{R}^n. \tag{8.63}$$

We ask: *if $f \in L^2(\mathbb{R}^n)$ what is the optimal regularity of u ?*

Following the above argument, our conclusions would be: it is true that we start from $u \in H^1(\mathbb{R}^n)$, but applying the second order operator $-\Delta + I$, where I denotes the identity operator, we find $f \in L^2(\mathbb{R})$.

Thus we conclude that the starting function should actually be in $H^2(\mathbb{R}^n)$ rather than $H^1(\mathbb{R}^n)$. Indeed this is true and can be easily proved using the Fourier transform. Since

$$\widehat{\partial_{x_i} u}(\xi) = i\xi_i \widehat{u}(\xi), \quad \widehat{\partial_{x_i x_j} u}(\xi) = -\xi_i \xi_j \widehat{u}(\xi),$$

we have

$$-\widehat{\Delta u}(\xi) = |\xi|^2 \widehat{u}(\xi)$$

and equation (8.63) becomes

$$(1 + |\xi|^2) \widehat{u}(\xi) = \widehat{f}(\xi).$$

We deduce, using the H^2 norm in (7.54), p. 473,

$$\|u\|_{H^2(\mathbb{R}^n)}^2 = \int_{\mathbb{R}^n} (1 + |\xi|^2) |\widehat{u}(\xi)|^2 d\xi = \int_{\mathbb{R}^n} |\widehat{f}(\xi)|^2 d\xi = (2\pi)^n \|f\|_{L^2(\mathbb{R}^n)}^2, \quad (8.64)$$

where the last equality follows from formula (7.40), p. 460.

Thus, $u \in H^2(\mathbb{R}^n)$ and moreover, we have obtained the stability estimate (8.64). We may go further. If $f \in H^1(\mathbb{R}^n)$, that is if f has first partial derivatives in $L^2(\mathbb{R}^n)$, a similar computation yields $u \in H^3(\mathbb{R}^n)$. Iterating this argument, we conclude that for every $m \geq 0$,

$$\text{if } f \in H^m(\mathbb{R}^n) \quad \text{then} \quad u \in H^{m+2}(\mathbb{R}^n).$$

Using the Sobolev embedding theorems of Sect. 7.10.4 in (say) any ball of \mathbb{R}^n , we infer that, if m is sufficiently large, u is a *classical solution*. In fact, if $u \in H^{m+2}(\mathbb{R}^n)$ then

$$u \in C^k(\mathbb{R}^n) \text{ for } k < m + 2 - \frac{n}{2},$$

and therefore it is enough that $m > \frac{n}{2}$ to have u at least in $C^2(\mathbb{R}^n)$. An immediate consequence is:

$$\text{if } f \in C^\infty(\mathbb{R}^n) \quad \text{then} \quad u \in C^\infty(\mathbb{R}^n).$$

This kind of results can be extended to any *uniformly elliptic* operators \mathcal{E} in divergence form and to the solutions of the *Dirichlet*, *Neumann* and *Robin* problems. The regularity for mixed problems is more complicated and requires delicate conditions along the border between Γ_D and Γ_N . We will not insist on this subject⁹.

⁹ See Haller-Dintelmann, Meyer, Reberg, Schiela, Holder continuity and optimal control for nonsmooth elliptic domains, Appl. Math. Optim. 60: 397–428, 2009.

There are two kinds of regularity results, concerning *interior regularity* and *global (up to the boundary) regularity*, respectively. Since the proofs are quite technical we only state the main results.¹⁰

In all the theorems below u is a weak solution of

$$\mathcal{E}u = -\operatorname{div}(\mathbf{A}(\mathbf{x}) \nabla u) - \mathbf{b}(\mathbf{x}) u + \mathbf{c}(\mathbf{x}) \cdot \nabla u + a(\mathbf{x}) u = f \quad \text{in } \Omega$$

where Ω is a bounded domain. We keep the hypotheses (8.32) and (8.33).

Interior regularity

The next theorem is a H^2 -interior regularity result. Note that the boundary of the domain does not play any role. We have:

Theorem 8.25. *Let the coefficients a_{ij} and b_j , $i, j = 1, \dots, n$, be Lipschitz in $\overline{\Omega}$, $f \in L^2(\Omega)$. Then $u \in H_{loc}^2(\Omega)$ and is satisfies the differential equation in the pointwise a.e. sense in Ω . Moreover, if $\Omega' \subset\subset \Omega$,*

$$\|u\|_{H^2(\Omega')} \leq C_2 \left\{ \|f\|_{L^2(\Omega)} + \|u\|_{L^2(\Omega)} \right\}. \quad (8.65)$$

Thus, u is a *strong solution* (see Sect. 8.2) in Ω . The constant C_2 depends only on all the relevant parameters $\alpha, \beta_0, \gamma_0, \alpha_0, M$ and also on the distance of Ω' from $\partial\Omega$ and the Lipschitz constant of a_{ij} and b_j , $i, j = 1, \dots, n$.

Remark 8.26. The presence of the norm $\|u\|_{L^2(\Omega)}$ in the right hand side of (8.65) is necessary¹¹ and due to the fact that the bilinear form B associated to \mathcal{E} is only weakly coercive.

If we increase the regularity of the coefficients, the smoothness of u increases according to the following theorem:

Theorem 8.27. *Let $f \in H^m(\Omega)$, $a_{ij}, b_j \in C^{m+1}(\overline{\Omega})$ and $c_j, a_0 \in C^m(\overline{\Omega})$, $m \geq 1$, $i, j = 1, \dots, n$. Then $u \in H_{loc}^{m+2}(\Omega)$ and if $\Omega_0 \subset\subset \Omega$,*

$$\|u\|_{H^{m+2}(\Omega_0)} \leq C_m \left\{ \|f\|_{H^m(\Omega)} + \|u\|_{L^2(\Omega)} \right\}.$$

As a consequence, if $a_{ij}, b_j, c_j, a_0, f \in C^\infty(\Omega)$, then $u \in C^\infty(\Omega)$ as well.

The constant C_m depends only on all the relevant parameters $\alpha, \beta_0, \gamma_0, \alpha_0, M$, on the distance of Ω_0 from $\partial\Omega$ and on the norms of a_{ij}, b_j , in $C^{m+1}(\overline{\Omega})$ and of c_j, a in $C^m(\overline{\Omega})$, $i, j = 1, \dots, n$.

¹⁰ For the proofs, see e.g. [3], *Gilbarg-Trudinger, 1998*.

¹¹ For instance, $u(x) = \sin x$ is a solution of the equation $u'' + u = 0$. Clearly we cannot control any norm of u with the norm of the right hand side alone!

Global regularity

We now focus on the optimal regularity of a solution (nonnecessarily unique!) of the boundary value problems we have considered in the previous sections.

Consider first H^2 -regularity. If $u \in H^2(\Omega)$, its trace on $\partial\Omega$ belongs to $H^{3/2}(\partial\Omega)$ so that a Dirichlet data g_D has to be taken in this space. On the other hand, the trace of the normal derivative belongs to $H^{1/2}(\partial\Omega)$ and hence we have to assign a Neumann or a Robin data g_N in this space. Also, the domain has to be smooth enough, say C^2 , in order to define the traces of u and $\partial_\nu u$.

Thus, assume that u is a solution of $\mathcal{E}u = f$ in Ω , subject to one of the following boundary conditions:

$$u = g_D \in H^{3/2}(\partial\Omega)$$

or

$$\partial_\nu^\mathcal{E} u + hu = g_N \in H^{1/2}(\partial\Omega),$$

with

$$0 \leq h(\sigma) \leq h_0 \quad \text{a.e. on } \partial\Omega.$$

We have:

Theorem 8.28. *Let Ω be a bounded, C^2 -domain. Assume that $f \in L^2(\Omega)$, a_{ij}, b_j , $i, j = 1, \dots, n$, and h are Lipschitz in $\overline{\Omega}$ and on $\partial\Omega$, respectively. Then $u \in H^2(\Omega)$ and*

$$\|u\|_{H^2(\Omega)} \leq C_{2,D} \left\{ \|u\|_{L^2(\Omega)} + \|f\|_{L^2(\Omega)} + \|g_D\|_{H^{3/2}(\partial\Omega)} \right\} \quad (\text{Dirichlet}),$$

$$\|u\|_{H^2(\Omega)} \leq C_{2,N} \left\{ \|u\|_{L^2(\Omega)} + \|f\|_{L^2(\Omega)} + \|g_N\|_{H^{1/2}(\partial\Omega)} \right\} \quad (\text{Neumann/Robin}).$$

The constants $C_{2,D}, C_{2,N}$ depend on all the relevant parameters $\alpha, \beta_0, \gamma_0, \alpha_0, M, (C_{2,N} \text{ also on } h_0)$, on the Lipschitz constants of $a_{ij}, b_j, j = 1, \dots, n$, and on the C^2 character¹² of Ω . If we increase the regularity of the domain, the coefficients and the data, the smoothness of u increases accordingly to the following theorem.

Theorem 8.29. *Let Ω be a bounded C^{m+2} -domain. Assume that $a_{ij}, b_j \in C^{m+1}(\overline{\Omega})$, $c_j, a \in C^m(\overline{\Omega})$, $i, j = 1, \dots, n$, $f \in H^m(\Omega)$. If $g_D \in H^{m+3/2}(\partial\Omega)$ or $g_N \in H^{m+1/2}(\partial\Omega)$ and $h \in C^{m+1}(\partial\Omega)$, then $u \in H^{m+2}(\Omega)$ and moreover,*

$$\|u\|_{H^{m+2}(\Omega)} \leq C \left\{ \|u\|_{L^2(\Omega)} + \|f\|_{H^m(\Omega)} + \|g_D\|_{H^{m+3/2}(\partial\Omega)} \right\} \quad (\text{Dirichlet}),$$

$$\|u\|_{H^{m+2}(\Omega)} \leq C \left\{ \|u\|_{L^2(\Omega)} + \|f\|_{H^m(\Omega)} + \|g_N\|_{H^{m+1/2}(\partial\Omega)} \right\} \quad (\text{Neumann, Robin}).$$

In particular, if Ω is a C^∞ -domain, all the coefficients are in $C^\infty(\overline{\Omega})$ and the boundary data are in $C^\infty(\partial\Omega)$, then $u \in C^\infty(\overline{\Omega})$.

¹² More precisely on the C^2 norms of the local charts that describe locally $\partial\Omega$.

• *A particular case.* Let Ω be a C^2 -domain and $f \in L^2(\Omega)$. The Lax-Milgram Theorem and Theorem 8.28 imply that the solution of the Dirichlet problem

$$\begin{cases} -\Delta u = f & \text{in } \Omega \\ u = 0 & \text{on } \partial\Omega \end{cases}$$

belongs to $H^2(\Omega) \cap H_0^1(\Omega)$ and that

$$\|u\|_{H^2(\Omega)} \leq C \|f\|_{L^2(\Omega)} = C \|\Delta u\|_{L^2(\Omega)}. \quad (8.66)$$

Since clearly we have $\|\Delta u\|_{L^2(\Omega)} \leq \|u\|_{H^2(\Omega)}$, we draw the following important conclusion:

Corollary 8.30. *If $u \in H^2(\Omega) \cap H_0^1(\Omega)$, then*

$$\|\Delta u\|_0 \leq \|u\|_{H^2(\Omega)} \leq C_b \|\Delta u\|_0.$$

In other words, $\|\Delta u\|_0$ and $\|u\|_{H^2(\Omega)}$ are equivalent norms in $H^2(\Omega) \cap H_0^1(\Omega)$.

In Sect. 9.2 we will see an application of Corollary 8.30 to an equilibrium problem for a bent plate.

Mixed problems

As we have already mentioned, the regularity for mixed problems is more complicated and requires delicate conditions along the border between Γ_D and Γ_N . However, the following example gives an idea of what can happen.

Example 8.31. The function $u(r, \theta) = r^{\frac{1}{2}} \sin \frac{\theta}{2}$ is a weak solution in the half circle

$$S_\pi = \{(r, \theta) : 0 < r < 1, 0 < \theta < \pi\}$$

of the mixed problem

$$\begin{cases} \Delta u = 0 & \text{in } S_\pi \\ u(1, \theta) = \sin \frac{\theta}{2} & 0 < \theta < \pi \\ u(r, 0) = 0 \text{ and } \partial_{x_2} u(r, \pi) = 0 & 0 \leq r < 1. \end{cases} \quad (8.67)$$

Namely, $|\nabla u|^2 = \frac{1}{4r}$, so that

$$\int_{S_\pi} |\nabla u|^2 dx_1 dx_2 = \frac{\pi}{4}$$

whence $u \in H^1(S_\pi)$. Moreover, $u(r, 0) = 0$ and

$$\partial_{x_2} u = u_r \sin \theta + \frac{1}{r} u_\theta \cos \theta = \frac{1}{2\sqrt{r}} \cos \frac{\theta}{2}$$

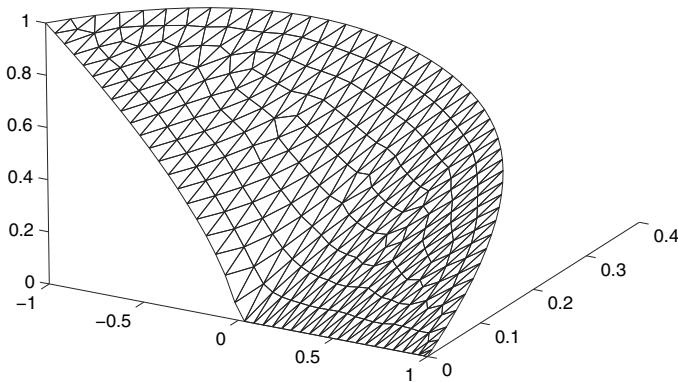


Fig. 8.3 The solution of the mixed problem (8.67)

whence $\partial_{x_2} u(r, \pi) = 0$. However, along the half-line $\theta = \pi/2$, for example, we have $|\partial_{x_i x_j} u| \sim r^{-\frac{3}{2}}$ for $r \sim 0$, so that

$$\int_{S_\alpha} |\partial_{x_i x_j} u|^2 dx_1 dx_2 \sim \int_0^1 r^{-2} dr = \infty$$

and therefore $u \notin H^2(S_\pi)$. Thus, the solution has a low order of regularity near the origin, even though the boundary of S_π is flat there. Note that the origin separates the Dirichlet and Neumann regions (see Fig. 8.3).

Conclusion: *in general, the optimal regularity of the solution of a mixed problem is less than H^2 near the boundary between the Dirichlet and Neumann regions.*

Regularity in Lipschitz domains

The above regularity results hold for smooth domains. However, in several applied situations, Lipschitz domains are the relevant ones. Let us examine the following situation.

Example 8.32. Consider the plane sector:

$$S_\alpha = \{(r, \theta) : 0 < r < 1, -\alpha/2 < \theta < \alpha/2\} \quad (0 < \alpha < 2\pi).$$

The function (see Fig. 8.4)

$$u(r, \theta) = r^{\frac{\pi}{\alpha}} \cos \frac{\pi}{\alpha} \theta$$

is harmonic in S_α , since it is the real part of $f(z) = z^{\frac{\pi}{\alpha}}$, which is holomorphic in S_α . Furthermore,

$$u(r, -\alpha/2) = u(r, \alpha/2) = 0, \quad 0 \leq r \leq 1 \quad (8.68)$$

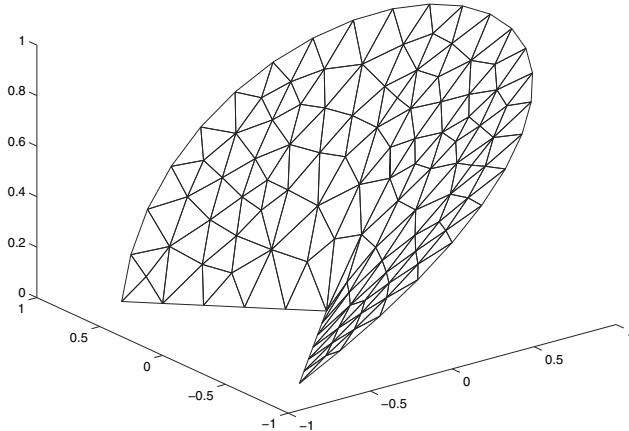


Fig. 8.4 The case $\alpha = \frac{3}{2}\pi$ in Example 8.32

and

$$u(1, \theta) = \cos \frac{\pi}{\alpha} \theta, \quad 0 \leq \theta \leq \alpha. \quad (8.69)$$

We focus on a neighborhood of the origin. If $\alpha = \pi$, S_α is a semicircle and

$$u(r, \theta) = \operatorname{Re} z = x \in C^\infty(\overline{S}_\alpha).$$

Suppose $\alpha \neq \pi$. Since

$$|\nabla u|^2 = u_r^2 + \frac{1}{r^2} u_\theta^2 = \frac{\pi^2}{\alpha^2} r^{2(\frac{\pi}{\alpha} - 1)},$$

we have

$$\int_{S_\alpha} |\nabla u|^2 dx_1 dx_2 = \frac{\pi^2}{\alpha} \int_0^1 r^{2\frac{\pi}{\alpha} - 1} dr = \frac{\pi}{2}$$

so that $u \in H^1(S_\alpha)$ and is the unique weak solution of $\Delta u = 0$ in S_α with the boundary conditions (8.68), (8.69). It is easy to check that for every $i, j = 1, 2$, $|\partial_{x_i x_j} u| \sim r^{\frac{\pi}{\alpha} - 2}$, as $r \rightarrow 0$, whence

$$\int_{S_\alpha} |\partial_{x_i x_j} u|^2 dx_1 dx_2 \simeq \int_0^1 r^{2\frac{\pi}{\alpha} - 3} dr.$$

This integral is convergent only for $2\frac{\pi}{\alpha} - 3 > -1$, i.e. $\alpha < \pi$. The conclusion is that $u \in H^2(S_\alpha)$ if and only if $\alpha \leq \pi$, i.e. if the sector is convex. If $\alpha > \pi$, $u \notin H^2(S_\alpha)$.

Conclusion: in a neighborhood of a nonconvex angle, we expect a low degree of regularity of the solution (less than H^2).

Thus, we may expect H^2 -regularity for bounded, convex domain. Indeed the following theorem holds, where we consider for simplicity only the Laplace operator¹³. Note that a convex domain has a Lipschitz boundary.

Proposition 8.33. *Let Ω be a bounded, convex domain. Then, if $f \in L^2(\Omega)$ and $\lambda > 0$, the weak solution of the problems*

$$-\Delta u = f \text{ in } \Omega, \quad u = 0 \text{ on } \partial\Omega$$

and

$$-\Delta u + \lambda u = f \text{ in } \Omega, \quad \partial_\nu u = 0 \text{ on } \partial\Omega$$

belongs to $H^2(\Omega)$.

Increasing the summability of f we can obtain at least continuity up to the boundary of a Lipschitz domain. Precisely, the following slightly more general result holds for the Dirichlet problem. Note that no regularity is required to the coefficients.

Theorem 8.34. *Let Ω be a bounded, Lipschitz domain and $u \in H_0^1(\Omega)$ be a solution of $\mathcal{E}u = f + \operatorname{div} \mathbf{F}$ in Ω . If $f \in L^p(\Omega)$ and $\mathbf{F} \in L^q(\Omega; \mathbb{R}^n)$ with $p > n/2$ and $q > n$, then $u \in C^{0,\sigma}(\overline{\Omega})$ for some $\sigma \in (0, 1]$, and*

$$\|u\|_{C^{0,\sigma}(\overline{\Omega})} \leq c \left\{ \|u\|_{L^2(\Omega)} + \|f\|_{L^p(\Omega)} + \|\mathbf{F}\|_{L^q(\Omega; \mathbb{R}^n)} \right\}. \quad (8.70)$$

The numbers σ and c depend only all the relevant parameters $\alpha, \beta_0, \gamma_0, \alpha_0, M$ and from the diameter and the Lipschitz constant of Ω .

A similar result hold for the Robin/Neumann conditions¹⁴.

Theorem 8.35. *Let Ω be a bounded, Lipschitz domain and assume that $a(\mathbf{x}) \geq c_0 > 0$ a.e. in Ω . Let $u \in H^1(\Omega)$ be the solution of*

$$\begin{cases} \mathcal{E}u = f & \text{in } \Omega \\ \partial_\nu^\mathcal{E} u + hu = g & \text{on } \partial\Omega. \end{cases}$$

If $f \in L^p(\Omega)$, $g \in L^q(\partial\Omega)$, with $p > n/2$, $q > n - 1$, $n \geq 2$, then $u \in C(\overline{\Omega})$ and

$$\|u\|_{C(\overline{\Omega})} \leq C \left\{ \|f\|_{L^p(\Omega)} + \|g\|_{L^q(\partial\Omega)} \right\}.$$

The constant C depends only all the relevant parameters $\alpha, \beta_0, \gamma_0, \alpha_0, h_0, M$ and from the diameter and the Lipschitz constant of Ω .

¹³ For the proof, see [4], Grisvard, 1985.

¹⁴ For the proof, see J.A. Griebentrop, Linear elliptic boundary value problems with nonsmooth data: Campanato spaces of functionals, Math. Nachr., 243, 19–42, 2002.

Problems

8.1. Write the weak formulation of the following problem:

$$\begin{cases} (x^2 + 1)u'' - xu' = \sin 2\pi x & 0 < x < 1 \\ u(0) = u(1) = 0. \end{cases}$$

Show that there exists a unique solution $u \in H_0^1(0, 1)$ and find an estimate for $\|u'\|_{L^2(0,1)}$.

8.2. Write the weak formulation for the equation

$$(p(x)u')' + b(x)u' + a(x)u = 0 \quad \text{in } (0, 1),$$

with Robin and mixed conditions. Assume p, a, b bounded and $p(x) \geq \alpha > 0$ in $(0, 1)$. Give sufficient conditions on p, b, a which assure existence and uniqueness of the solution and give a stability estimate.

8.3. Write the weak formulation of the following problem:

$$\begin{cases} \cos x u'' - \sin x u' - xu = 1 & 0 < x < \pi/6 \\ u'(0) = -u(0), u(\pi/6) = 0. \end{cases}$$

Discuss existence and uniqueness and derive a stability estimates.

8.4. Legendre equation. Let

$$X = \left\{ v \in L^2(-1, 1) : (1 - x^2)^{1/2}v' \in L^2(-1, 1) \right\}$$

with inner product

$$(u, v)_X = \int_{-1}^1 [uv + (1 - x^2)u'v'] dx.$$

- a) Check that $(u, v)_X$ is indeed an inner product and that X is a Hilbert space.
- b) Study the variational problem

$$(u, v)_X = \int_{-1}^1 fv dx, \quad \text{for every } v \in X, \tag{8.71}$$

where $f \in L^2(-1, 1)$.

- c) Determine the boundary value whose variational formulation is (8.71).

[a] *Hint:* Use Theorem 7.56, p. 461, with

$$V = L^2(-1, 1) \text{ and } Z = L_w^2(-1, 1), w(x) = (1 - x^2)^{1/2}.$$

Check that $Z \hookrightarrow \mathcal{D}'(-1, 1)$.

b) *Hint:* Use the Lax-Milgram Theorem.

c) *Answer:* The boundary value problem is

$$\begin{cases} -[(1-x^2)u']' + u = f & -1 < x < 1 \\ (1-x^2)u'(x) \rightarrow 0 & \text{as } x \rightarrow \pm 1. \end{cases}$$

This is a Legendre equation with the natural Neumann conditions at both end points].

8.5. Let

$$V = H_{per}^1(0, 2\pi) = \{u \in H^1(0, 2\pi) : u(0) = u(2\pi)\}$$

and F be the linear functional

$$F : v \longmapsto \int_0^{2\pi} tv(t) dt.$$

(a) Check that $F \in V^*$.

(b) According to Riesz's Theorem 6.30, p. 378, there is a unique element $u \in V$ such that

$$(u, v)_{H^1}(0, 2\pi) = Fv, \text{ for every } v \in V.$$

Determine explicitly u .

8.6. *Transmission conditions* (I). Consider the problem

$$\begin{cases} (p(x)u')' = f & \text{in } (a, b) \\ u(a) = u(b) = 0 \end{cases} \quad (8.72)$$

where $f \in L^2(a, b)$, $p(x) = p_1 > 0$ in (a, c) and $p(x) = p_2 > 0$ in (c, b) .

Show that problem (8.72) has a unique weak solution in $H^1(a, b)$, satisfying the conditions:

$$\begin{cases} p_1 u'' = f & \text{in } (a, c) \\ p_2 u'' = f & \text{in } (c, b) \\ p_1 u'(c-) = p_2 u'(c+) \end{cases}$$

Observe the jump of the derivative of u at $x = c$ (Fig. 8.5).

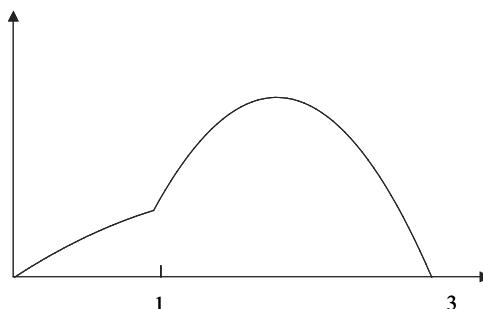


Fig. 8.5 The solution of the transmission problem $(p(x)u')' = -1$, $u(0) = u(3) = 0$, with $p(x) = 3$ in $(0, 1)$ and $p(x) = 1/2$ in $(1, 3)$

8.7. Let $\Omega = (0, 1) \times (0, 1) \subset \mathbb{R}^2$. Prove that the functional

$$E(v) = \frac{1}{2} \int_{\Omega} \{|\nabla v|^2 - 2xv\} dx dy$$

has a unique minimizer $u \in H_0^1(\Omega)$. Write the Euler equation and find an explicit formula for u .

8.8. Consider the following subspace of $H^1(\Omega)$:

$$V = \left\{ u \in H^1(\Omega) : \frac{1}{|\partial\Omega|} \int_{\partial\Omega} u d\sigma = 0 \right\}.$$

a) Show that V is a Hilbert space with inner product $(\cdot, \cdot)_{H^1(\Omega)}$ and find which boundary value problem has the following weak formulation:

$$\int_{\Omega} \{\nabla u \cdot \nabla v + uv\} dx = \int_{\Omega} fv dx, \quad \forall v \in V.$$

b) Show that if $f \in L^2(\Omega)$ there exists a unique solution.

[Answer: a) $-\Delta u + u = f$ in Ω , $\partial_{\nu} u = \text{constant}$ on $\partial\Omega$. b) Use the Riesz Representation Theorem or the Lax-Milgram Theorem].

8.9. Let $\Omega \subset \mathbb{R}^n$ and $g \in H^{1/2}(\partial\Omega)$. Define

$$H_g^1(\Omega) = \{v \in H^1(\Omega) : v = g \text{ on } \partial\Omega\}.$$

Prove the following theorem, known as **Dirichlet principle**: *Among all the functions $v \in H_g^1(\Omega)$, the harmonic one minimizes the Dirichlet integral*

$$D(v) = \int_{\Omega} |\nabla v|^2 dx.$$

[Hint: In $H^1(\Omega)$ use the inner product

$$(u, v)_{1,\partial} = \int_{\partial\Omega} uv d\sigma + \int_{\Omega} \nabla u \cdot \nabla v dx$$

and the norm (see Remark 7.92, p. 489):

$$\|u\|_{1,\partial} = \left(\int_{\partial\Omega} u^2 d\sigma + \int_{\Omega} |\nabla u|^2 dx \right)^{1/2}. \quad (8.73)$$

Then, minimizing $D(v)$ over $H_g^1(\Omega)$ is equivalent to minimizing $\|v\|_{1,\partial}^2$. Let $u \in H_g^1(\Omega)$ be harmonic in Ω . If $v \in H_g^1(\Omega)$, write $v = u + w$, with $w \in H_0^1(\Omega)$. Show that $(u, w)_{1,\partial} = 0$ and conclude that $\|u\|_{1,\partial}^2 \leq \|v\|_{1,\partial}^2$].

8.10. A simple system. Let $\Omega \subset \mathbb{R}^n$ be a bounded Lipschitz domain. Consider the Neumann problem for the following system:

$$\begin{cases} -\Delta u_1 + u_1 - u_2 = f_1 & \text{in } \Omega \\ -\Delta u_2 + u_1 + u_2 = f_2 & \text{in } \Omega \\ \partial_{\nu} u_1 = \partial_{\nu} u_2 = 0 & \text{on } \partial\Omega. \end{cases}$$

Derive a variational formulation and establish a well-posedness theorem.

[Hint: Variational formulation:

$$\int_{\Omega} \{\nabla u_1 \cdot \nabla v_1 + \nabla u_2 \cdot \nabla v_2 + u_1 v_1 - u_2 v_1 + u_1 v_2 + u_2 v_2\} = \int_{\Omega} (f_1 v_1 + f_2 v_2)$$

for every $(v_1, v_2) \in H^1(\Omega) \times H^1(\Omega)$].

8.11. Transmission conditions (II). Let Ω_1 and Ω be bounded, Lipschitz domains in \mathbb{R}^n such that $\Omega_1 \subset\subset \Omega$. Let $\Omega_2 = \Omega \setminus \overline{\Omega}_1$. In Ω_1 and Ω_2 consider the following bilinear forms

$$a_k(u, v) = \int_{\Omega_k} \mathbf{A}^k(\mathbf{x}) \nabla u \cdot \nabla v \, d\mathbf{x} \quad (k = 1, 2)$$

with \mathbf{A}^k uniformly elliptic. Assume that the entries of A^k are continuous in $\overline{\Omega}_k$, but that the matrix

$$\mathbf{A}(\mathbf{x}) = \begin{cases} \mathbf{A}^1(\mathbf{x}) & \text{in } \overline{\Omega}_1 \\ \mathbf{A}^2(\mathbf{x}) & \text{in } \Omega_2 \end{cases}$$

may have a jump across $\Gamma = \partial\Omega_1$. Let $u \in H_0^1(\Omega)$ be the weak solution of the equation

$$a(u, v) = a_1(u, v) + a_2(u, v) = (f, v)_{L^2(\Omega)} \quad \forall v \in H_0^1(\Omega),$$

where $f \in L^2(\Omega)$.

- a) Which boundary value problem does u satisfy?
- b) Which conditions on Γ do express the coupling between u_1 and u_2 ?

[Hint: b) $u_{1|_\Gamma} = u_{2|_\Gamma}$ and

$$\mathbf{A}^1 \nabla u_1 \cdot \boldsymbol{\nu} = \mathbf{A}^2 \nabla u_2 \cdot \boldsymbol{\nu},$$

where $\boldsymbol{\nu}$ points outward with respect to Ω_1].

8.12. Find the mistake in the following argument. Consider the Neumann problem

$$\begin{cases} -\Delta u + \mathbf{c} \cdot \nabla u = f & \text{in } \Omega \\ \partial_{\boldsymbol{\nu}} u = 0 & \text{on } \partial\Omega \end{cases} \quad (8.74)$$

with Ω smooth, $\mathbf{c} \in C^1(\overline{\Omega})$ and $f \in L^2(\Omega)$. Let $V = H^1(\Omega)$ and

$$B(u, v) = \int_{\Omega} \{\nabla u \cdot \nabla v + (\mathbf{c} \cdot \nabla u)v\}.$$

If $\operatorname{div} \mathbf{c} = 0$, we may write

$$\int_{\Omega} (\mathbf{c} \cdot \nabla u) u \, d\mathbf{x} = \frac{1}{2} \int_{\Omega} \mathbf{c} \cdot \nabla(u^2) \, d\mathbf{x} = \frac{1}{2} \int_{\partial\Omega} u^2 \mathbf{c} \cdot \boldsymbol{\nu} \, d\sigma.$$

Thus, if $\mathbf{c} \cdot \boldsymbol{\nu} \geq c_0 > 0$ then, recalling Remark 7.92, p. 489,

$$B(u, u) \geq \|\nabla u\|_{L^2(\Omega)}^2 + c_0 \|u\|_{L^2(\partial\Omega)}^2 \geq C \|u\|_{H^1(\Omega)}^2$$

so that B is V -coercive and problem (8.74) has a unique solution!!

8.13. Let $Q = (0, \pi) \times (0, \pi)$. Study the solvability of the Dirichlet problem

$$\begin{cases} \Delta u + 2u = f & \text{in } Q \\ u = 0 & \text{on } \partial Q. \end{cases}$$

In particular, examine the cases $f(x, y) = 1$ and $f(x, y) = x - \pi/2$.

8.14. Let

$$B_1^+ = \{(x, y) \in \mathbb{R}^2 : x^2 + y^2 < 1, y > 0\}.$$

Examine the solvability of the Robin problem

$$\begin{cases} -\Delta u = f & \text{in } B_1^+ \\ \partial_\nu u + yu = 0 & \text{on } \partial B_1^+. \end{cases}$$

8.15. Let $\Omega = (0, 1) \times (0, 1)$, $a \in \mathbb{R}$. Examine the solvability of the mixed problem

$$\begin{cases} \Delta u + au = 1 & \text{in } \Omega \\ u = 0 & \text{on } \partial\Omega \setminus \{y = 0\} \\ \partial_\nu u = x & \text{on } \{y = 0\}. \end{cases}$$

8.16. Let $\Omega \subset \mathbb{R}^n$ be a bounded, Lipschitz domain and $a, b \in L^\infty(\Omega)$, with $\int_\Omega b(x)dx \neq 0$. Consider the following (nonlocal) Neumann problem:

$$\begin{cases} -\Delta u + a(x) \int_\Omega b(z) u(z) dz = f(x) & \text{in } \Omega \\ \partial_\nu u = 0 & \text{on } \partial\Omega \end{cases} \quad (8.75)$$

where $f \in L^2(\Omega)$.

- (a) Give a weak formulation of the problem.
- (b) Analyze its solvability.

8.17. Let $u \in H^1(\Omega)$ satisfy the *Robin problem*

$$\int_\Omega \{\mathbf{A} \nabla u \cdot \nabla v + auv\} dx + \int_{\partial\Omega} huv d\sigma = \int_\Omega v dx, \quad (8.76)$$

for every $v \in H^1(\Omega)$. Assume that (8.32) and (8.33), p. 521, hold. Moreover, assume that $a \geq 0$ a.e. in Ω , $h \in L^\infty(\partial\Omega)$, $h \geq 0$ a.e. on $\partial\Omega$, and

$$\int_{\partial\Omega} h d\sigma + \int_\Omega a dx = q > 0. \quad (8.77)$$

Prove that there exists a unique solution of (8.76) and derive a stability estimate.

8.18. Let λ_1 be the principal Dirichlet eigenvalue for the Laplace operator and V_1 be the corresponding eigenspace.

- (a) Show that the second Dirichlet eigenvalue λ_2 satisfy the following variational principle, where $R(v)$ is the Rayleigh quotient (8.24), p. 518:

$$\lambda_2 = \min \{R(v) : v \neq 0, v \in H_0^1(\Omega) \cap V_1^\perp, \}.$$

- (b) State and prove a similar variational principle for the $n - th$ Dirichlet eigenvalue.

8.19. Let $\Omega \subset \mathbb{R}^n$ be a bounded, regular domain and $a \in C^\infty(\overline{\Omega})$, $a \geq 0$ in Ω . Consider the following eigenvalue problem:

$$\begin{cases} -\Delta u + a(\mathbf{x})u = \lambda u & \text{in } \Omega \\ u = 0 & \text{on } \partial\Omega. \end{cases}$$

- (a) Check the existence of an orthonormal basis of eigenfunctions $\{w_k\}_{k \geq 1}$ in $L^2(\Omega)$, and of a corresponding nondecreasing sequence of eigenvalues $\{\lambda_k\}_{k \geq 1}$ with $\lambda_k \rightarrow +\infty$. Suitably normalized, $\{w_k\}_{k \geq 1}$ is orthonormal in $H_0^1(\Omega)$ with respect to an appropriate inner product. Which one is it?
- (b) Show that the principal eigenvalue $\lambda_1 = \lambda_1(\Omega, a)$ is positive. Which Rayleigh quotient does λ_1 minimize?
- (c) Deduce the following monotonicity and stability properties of λ_1 with respect to the coefficient a :
1. If $a_1 \leq a_2$ in Ω then $\lambda_1(\Omega, a_1) \leq \lambda_1(\Omega, a_2)$.
 2. $|\lambda_1(\Omega, a_1) - \lambda_1(\Omega, a_2)| \leq \|a_1 - a_2\|_{L^\infty(\Omega)}$.
 3. If $\Omega_1 \subset \Omega_2$ then $\lambda_1(\Omega_1, a) \geq \lambda_1(\Omega_2, a)$.

8.20. Let

$$\mathcal{E}u = -\operatorname{div}(\mathbf{A}(\mathbf{x})\nabla u) + a(\mathbf{x})u,$$

with $\mathbf{A} = (a_{ij})_{ij=1,\dots,n}$ symmetric. Assume that Ω is a regular bounded domain and that a_{ij} and a belong to $C^\infty(\overline{\Omega})$. State and prove the analogue of Theorem 8.8.

8.21. Let $\Omega \subset \mathbb{R}^n$ be a bounded, smooth domain. Show that for every $u \in H^2(\Omega)$ such that $\partial_\nu u = 0$ on $\partial\Omega$, the following inequality holds

$$\int_{\Omega} (\Delta u)^2 \geq \mu_1 \int_{\Omega} |\nabla u|^2 \quad (8.78)$$

where μ_1 is the first positive Neumann eigenvalue for the operator $-\Delta$ in Ω . When does the equality sign hold in (8.78)?

8.22. A) Let V, H be Hilbert spaces with $V \hookrightarrow H$. Denote by $a(u, v)$, $a_1(u, v)$, $a_2(u, v)$ three nonnegative bilinear forms, on V , such that

$$a(u, v) = a_1(u, v) + a_2(u, v), \quad \forall u, v \in V.$$

Define

$$\Lambda = \inf_{u \in V, u \neq 0} \frac{a(u, u)}{\|u\|_H^2} \quad \text{and} \quad \lambda_j = \inf_{u \in V, u \neq 0} \frac{a_j(u, u)}{\|u\|_H^2}, \quad j = 1, 2.$$

Which one of the following inequality is true?

a) $\Lambda \geq \lambda_1 + \lambda_2$ b) $\Lambda \leq \lambda_1 + \lambda_2$.

B) Let $V_{r,s} = H_0^1(r, s)$, $0 < r < s$. Compute

$$\lambda(r, s) = \inf_{u \in V_{r,s}} \frac{\int_r^s (u')^2}{\int_r^s u^2}.$$

C) Let Ω be the sector of elliptic ring defined by $(p, q > 0)$:

$$\Omega = \left\{ (x, y) \in \mathbb{R}^2 : \frac{1}{4} < \frac{x^2}{p^2} + \frac{y^2}{q^2} < 1, \quad p > 0, \quad q > 0 \right\}.$$

Using the results in A) and B), show that, if Λ is the first Dirichlet eigenvalue for the operator $-\Delta$ in Ω , then:

$$\Lambda \geq \frac{4\pi^2}{3} \left(\frac{1}{p^2} + \frac{1}{q^2} \right).$$

8.23. Let $\Omega \subset \mathbb{R}^3$ be a bounded Lipschitz domain, $f \in L^2(\Omega)$. Assume that f has the following expansion:

$$f = \sum_{k=1}^{\infty} f_k u_k$$

where $f_k = (f, u_k)_{L^2(\Omega)}$ and $\{u_k\}_{k \geq 1}$ is a sequence of orthonormal (in $L^2(\Omega)$) Dirichlet eigenvectors for the Laplace operator with corresponding eigenvalues $\{\lambda_k\}_{k \geq 1}$.

a) Show that the solution $u \in H_0^1(\Omega)$ of the Dirichlet problem $-\Delta u = f$ in Ω , $u = 0$ on $\partial\Omega$ has the following expansion:

$$u = - \sum_{k=1}^{\infty} \frac{f_k}{\lambda_k} u_k, \quad (8.79)$$

with convergence in $H_0^1(\Omega)$.

b) Let $G = G(x, y)$ be the Green function for the Laplace operator in Ω . Show that $G(x, y)$ has the following expansion:

$$G(x, y) = - \sum_{k=1}^{\infty} \frac{1}{\lambda_k} u_k(x) u_k(y)$$

with convergence in $L^2(\Omega \times \Omega)$.

[Hint: b) Fix x and write

$$G(x, y) = - \sum_{k=1}^{\infty} g_k(x) u_k(y),$$

where

$$g_k(x) = \int_{\Omega} G(x, z) u_k(z) dz.$$

Recall from formula (3.84), p. 161, that the solution u of the Dirichlet problem in a) has the representation

$$u(x) = \int_{\Omega} G(x, y) f(y) dz.$$

Insert the expansion of G and f , use the orthonormality conditions $(u_k, u_j)_{L^2(\Omega)} = \delta_{kj}$ (the Kronecker symbol) to show that

$$u(x) = - \sum_{k=1}^{\infty} f_k g_k(x).$$

Compare with formula (8.79) and deduce that $g_k(x) = -u_k(x)/\lambda_k$.

Chapter 9

Further Applications

9.1 A Monotone Iteration Scheme for Semilinear Equations

The linear variational theory developed in the last chapter can be applied to a large variety of problems. Here we show how the weak maximum and comparison principles play a major role in constructing an iterative procedure to solve Fisher's type semilinear equation. Precisely, we consider the following problem:

$$\begin{cases} -\Delta u = f(u) & \text{in } \Omega \\ u = g & \text{on } \partial\Omega, \end{cases} \quad (9.1)$$

where $\Omega \subset \mathbb{R}^n$ is a bounded, Lipschitz domain and

$$f \in C^1(\mathbb{R}), g \in H^{1/2}(\partial\Omega).$$

A weak solution of problem (9.1) is a function $u \in H^1(\Omega)$ such that $u = g$ on $\partial\Omega$ and

$$\int_{\Omega} \nabla u \cdot \nabla v \, d\mathbf{x} = \int_{\Omega} f(u) v \, d\mathbf{x}, \quad \forall v \in H_0^1(\Omega). \quad (9.2)$$

We need to introduce *weak sub and super solutions*. We say that $u_* \in H^1(\Omega)$ is a *weak subsolution* of problem (9.1), if $u_* \leq g$ on $\partial\Omega$ and

$$\int_{\Omega} \nabla u_* \cdot \nabla v \, d\mathbf{x} \leq \int_{\Omega} f(u_*) v \, d\mathbf{x}, \quad \forall v \in H_0^1(\Omega), \quad v \geq 0 \text{ a.e. in } \Omega.$$

Similarly, we say that $u^* \in H^1(\Omega)$ is a *weak supersolution* of problem (9.1) if $u^* \geq g$ on $\partial\Omega$ and

$$\int_{\Omega} \nabla u^* \cdot \nabla v \, d\mathbf{x} \geq \int_{\Omega} f(u^*) v \, d\mathbf{x}, \quad \forall v \in H_0^1(\Omega), \quad v \geq 0 \text{ a.e. in } \Omega.$$

We want to prove the following theorem.

Theorem 9.1. Assume that g is bounded on $\partial\Omega$ and that there exist a weak subsolution u_* and a weak supersolution u^* of problem (9.1) such that:

$$a \leq u_* \leq u^* \leq b, \quad \text{a.e. in } \Omega, \quad a, b \in \mathbb{R}.$$

Then, there exists a solution u of problem (9.1) such that

$$u_* \leq u \leq u^*, \quad \text{a.e. in } \Omega.$$

Moreover, if Ω, g and f are of class C^∞ then $u \in C^\infty(\overline{\Omega})$

Proof. The idea is to exploit the linear theory to define recursively a sequence $\{u_k\}_{k \geq 1}$ of functions, monotonically convergent to a solution of problem (9.1). First we slightly modify the equation, in order to have a nondecreasing function of u in the right hand side. Let $M = \max_{[a,b]} |f'|$. Then the function $F(s) = f(s) + Ms$ is nondecreasing in $[a, b]$. Write the differential equation in the form

$$-\Delta u + Mu = F(u).$$

Let now u_1 be the weak solution of the linear problem

$$\begin{cases} -\Delta u_1 + Mu_1 = F(u_*) & \text{in } \Omega \\ u_1 = g & \text{on } \partial\Omega. \end{cases}$$

Given u_k , let u_{k+1} be the weak solution of the linear problem

$$\begin{cases} -\Delta u_{k+1} + Mu_{k+1} = F(u_k) & \text{in } \Omega \\ u_{k+1} = g & \text{on } \partial\Omega. \end{cases} \quad (9.3)$$

We claim that the sequence $\{u_k\}$ is nondecreasing and trapped between u_* and u^* :

$$u_* \leq u_k \leq u_{k+1} \leq u^*, \quad \text{a.e. in } \Omega, \quad k \geq 1.$$

Assuming the claim, we deduce that u_k converges a.e. in Ω (and also in $L^2(\Omega)$) to some bounded function u , as $k \rightarrow +\infty$. Since $F(a) \leq F(u_k) \leq F(b)$, by the Dominated Convergence Theorem, we infer that

$$\int_{\Omega} F(u_k) v d\mathbf{x} \rightarrow \int_{\Omega} F(u) v d\mathbf{x} \quad \text{as } k \rightarrow \infty,$$

for every $v \in H_0^1(\Omega)$. Now it is enough to show that $\{u_k\}$ converges weakly in $H^1(\Omega)$ to u , in order to pass to the limit in the equation

$$\int_{\Omega} (\nabla u_{k+1} \cdot \nabla v + Mu_{k+1}v) d\mathbf{x} = \int_{\Omega} F(u_k)v d\mathbf{x}, \quad \forall v \in H_0^1(\Omega),$$

and obtain (9.2).

We prove the claim. Let us check that $u_* \leq u_1$ a.e. in Ω . Set $w_0 = u_* - u_1$. Then $\sup_{\partial\Omega} w_0^+ = 0$ and

$$\int_{\Omega} (\nabla w_0 \cdot \nabla v + Mw_0 v) d\mathbf{x} \leq 0, \quad \forall v \in H_0^1(\Omega), \quad v \geq 0 \text{ a.e. in } \Omega.$$

From Theorem 8.20, p. 532, we deduce that $w_0 \leq 0$. Similarly, we infer that $u_1 \leq u^*$. Now assume inductively that

$$u_* \leq u_{k-1} \leq u_k \leq u^*, \quad \text{a.e. in } \Omega.$$

We prove that $u_* \leq u_k \leq u_{k+1} \leq u^*$ a.e. in Ω . Let $w_k = u_k - u_{k+1}$. We have $w_k = 0$ on $\partial\Omega$ and

$$\int_{\Omega} (\nabla w_k \cdot \nabla v + M w_k v) \, d\mathbf{x} = \int_{\Omega} [F(u_{k-1}) - F(u_k)]v \, d\mathbf{x}, \quad \forall v \in H_0^1(\Omega).$$

Since F is nondecreasing on $[a, b]$, we deduce that $F(u_{k-1}) - F(u_k) \leq 0$ a.e. in Ω , so that

$$\int_{\Omega} (\nabla w_k \cdot \nabla v + M w_k v) \, d\mathbf{x} \leq 0, \quad \forall v \in H_0^1(\Omega), v \geq 0 \text{ a.e. in } \Omega.$$

Again, Theorem 8.20 yields $w_k \leq 0$ a.e. in Ω . Similarly, we infer that $u_* \leq u_k$ and $u_{k+1} \leq u^*$ and the claim is proved.

To complete the proof, we have to show that $u_k \rightharpoonup u$, weakly in $H^1(\Omega)$. This follows from the estimate for the nonhomogeneous Dirichlet problem (9.3):

$$\begin{aligned} \|u_k\|_{H^1(\Omega)} &\leq C(n, M, \Omega) \left\{ \|F(u_{k-1})\|_{L^2(\Omega)} + \|g\|_{H^{1/2}(\partial\Omega)} \right\} \\ &\leq C_1(n, M, \Omega) \left\{ \max\{|F(a)|, |F(b)|\} + \|g\|_{H^{1/2}(\partial\Omega)} \right\}. \end{aligned}$$

Thus $\{u_k\}$ is bounded in $H^1(\Omega)$ and there exists a subsequence weakly convergent to u . Since the limit of each converging subsequence is unique, the whole sequence weakly converges to u .

If Ω, g and f are smooth, since $F(u) \in L^\infty(\Omega)$, Theorem 8.28, p. 539, gives $u \in H^2(\Omega)$. This implies that $F(u) \in H^2(\Omega)$ and therefore $u \in H^4(\Omega)$, by Theorem 8.29. Iterating¹, we eventually obtain $u \in C^\infty(\overline{\Omega})$. \square

The functions u_* and u^* in the above theorem are called *lower* and *upper* barrier, respectively. Thus, Theorem 9.1 reduces the solvability of problem (9.1) to finding a *lower* and an *upper* barrier. In general we cannot assert that the solution is unique. Here is an example of nonuniqueness for a *stationary Fischer's equation*.

Example 9.2. Consider the problem

$$\begin{cases} -\Delta u = u(1-u) & \text{in } \Omega \\ u = 0 & \text{on } \partial\Omega. \end{cases}$$

Clearly, $u_* \equiv 0$ is a solution. If we assume that the domain Ω is bounded and Lipschitz and that the first Dirichlet eigenvalue for the Laplace operator is $\lambda_1 < 1$, we can show that there exists a solution which is positive in Ω . In fact, $u^* \equiv 1$ is an *upper* barrier. We now exhibit a positive lower barrier. Let w_1 be the nonnegative normalized eigenfunction corresponding to λ_1 . From Theorem 8.9, p. 518, we know that $w_1 > 0$ inside Ω and, from Theorem 8.34 and interior elliptic regularity, that $w_1 \in C^\infty(\Omega) \cap C(\overline{\Omega})$. Let $u_* = \sigma w_1$. We claim that, if σ positive and small

¹ This procedure is called *bootstrapping*.

enough, u_* is a lower barrier. Indeed, since $-\Delta w_1 = \lambda_1 w_1$, we have,

$$-\Delta u_* - u_* (1 - u_*) = \sigma w_1 (\lambda_1 - 1 + \sigma w_1). \quad (9.4)$$

If $m = \max_{\bar{\Omega}} w_1$ and $\sigma < (1 - \lambda_1)/m$, then the right hand side of (9.4) is negative and u_* is a lower barrier. From Theorem 9.1 we infer the existence of a solution u such that $w_1 \leq u \leq 1$. \square

The uniqueness of the solution of problem (9.1) is guaranteed if, for instance, f is nonincreasing:

$$f'(s) \leq 0, \quad s \in \mathbb{R}.$$

Then, if u_1 and u_2 are two solutions of (9.1), we have $w = u_1 - u_2 \in H_0^1(\Omega)$ and we can write, by the Mean Value Theorem,

$$-\Delta w = f(u_1) - f(u_2) = c(\mathbf{x})w$$

where $c(\mathbf{x}) = f'(\bar{u}(\mathbf{x}))$, for a suitable \bar{u} between u_1 and u_2 . Since $c \leq 0$ we get

$$-\int_{\Omega} w \Delta w = \int_{\Omega} |\nabla w|^2 \leq 0$$

from which $w \equiv 0$ or $u_1 = u_2$.

9.2 Equilibrium of a Plate

The range of application of the variational theory is not confined to second order equations. In particular, using the elliptic regularity of Sect. 9.2 we can deal with a fourth order equation (so called *biharmonic* equation), modeling the equilibrium of a plate. More precisely, in this subsection we consider the vertical deflection $u = u(x, y)$ of a bent plate of small thickness (compared with the other dimensions) under the action of a normal load. If $\Omega \subset \mathbb{R}^2$ is a domain representing the transversal section of the plate, it can be shown that u is governed by the fourth order *elliptic* equation²

$$\Delta \Delta u = \Delta^2 u = \frac{q}{D} \equiv f \quad \text{in } \Omega,$$

² It is possible to give the definition of *ellipticity* for an operator of order higher than two. See [15], *Renardy-Rogers, 2004*. For instance, consider the linear operator with constant coefficients $\mathcal{L} = \sum_{|\alpha|=m} a_{\alpha} D^{\alpha}$, $m \geq 2$, where $\alpha = (\alpha_1, \dots, \alpha_n)$ is a multi-index. Associate with \mathcal{L} its **symbol**, given by

$$S_{\mathcal{L}}(\boldsymbol{\xi}) = \sum_{|\alpha|=m} a_{\alpha} (i\boldsymbol{\xi})^{\alpha}.$$

Then \mathcal{L} is said to be **elliptic** if $S_{\mathcal{L}}(\boldsymbol{\xi}) \neq 0$ for every $\boldsymbol{\xi} \in \mathbb{R}^n$, $\boldsymbol{\xi} \neq \mathbf{0}$. The symbol of $\mathcal{L} = \Delta^2$ in 2 dimensions is $-\xi_1^4 - 2\xi_1^2\xi_2^2 - \xi_2^4$, which is negative if $(\xi_1, \xi_2) \neq (0, 0)$. Thus Δ^2 is elliptic. Note that, for $m = 2$, we recover the usual definition of ellipticity.

where q is the external distributed load and D encodes the elastic properties of the material. The operator Δ^2 is called **biharmonic** or **bi-laplacian** and the solutions of $\Delta^2 u = 0$ are called *biharmonic functions*. In two dimensions, the explicit expression of Δ^2 is given by

$$\Delta^2 = \frac{\partial^4}{\partial x^4} + 2 \frac{\partial^4}{\partial x^2 \partial y^2} + \frac{\partial^4}{\partial y^4}.$$

If the plate is rigidly fixed along its boundary (*clamped plate*), then u and its normal derivative must vanish on $\partial\Omega$. Thus, we are led to the following boundary value problem:

$$\begin{cases} \Delta^2 u = f & \text{in } \Omega \\ u = \partial_\nu u = 0 & \text{on } \partial\Omega. \end{cases} \quad (9.5a)$$

To derive a variational formulation, we choose $C_0^2(\Omega)$ as space of test functions, i.e. the set of functions in $C^2(\Omega)$, compactly supported in Ω . This choice takes into account the boundary conditions. Now, we multiply the biharmonic equation by a function $v \in C_0^2(\Omega)$ and integrate over Ω :

$$\int_{\Omega} \Delta^2 u v \, d\mathbf{x} = \int_{\Omega} f v \, d\mathbf{x}. \quad (9.6)$$

Integrating by parts twice and using the conditions $v = \partial_\nu v = 0$ on $\partial\Omega$, we get:

$$\begin{aligned} \int_{\Omega} \Delta^2 u v \, d\mathbf{x} &= \int_{\Omega} (\operatorname{div} \nabla \Delta u) v \, d\mathbf{x} = \int_{\partial\Omega} \partial_\nu (\Delta u) v \, d\sigma - \int_{\Omega} \nabla \Delta u \cdot \nabla v \, d\mathbf{x} \\ &= - \int_{\partial\Omega} \Delta u \partial_\nu v \, d\sigma + \int_{\Omega} \Delta u \Delta v \, d\mathbf{x} = \int_{\Omega} \Delta u \Delta v \, d\mathbf{x}. \end{aligned}$$

Thus, (9.6) becomes

$$\int_{\Omega} \Delta u \Delta v \, d\mathbf{x} = \int_{\Omega} f v \, d\mathbf{x}. \quad (9.7)$$

Now we enlarge the space of test functions by taking the closure of $C_0^2(\Omega)$ in $H^2(\Omega)$, which is $H_0^2(\Omega)$. Note that (see Subsect. 7.9.2) this is precisely the space of functions u such that u and $\partial_\nu u$ have zero trace on $\partial\Omega$.

Since $H_0^2(\Omega) \subset H_0^1(\Omega) \cap H^2(\Omega)$, if Ω is a C^2 domain, from Corollary 8.30, p. 540, we know that in this space we may choose $\|u\|_{H_0^2(\Omega)} = \|\Delta u\|_{L^2(\Omega)}$ as a norm. We are led to the following **variational formulation**:

Determine $u \in H_0^2(\Omega)$ such that

$$\int_{\Omega} \Delta u \Delta v \, d\mathbf{x} = \int_{\Omega} f v \, d\mathbf{x}, \quad \forall v \in H_0^2(\Omega). \quad (9.8)$$

The following result holds:

Proposition 9.3. *Let Ω be a bounded, C^2 domain. If $f \in L^2(\Omega)$, there exists a unique solution $u \in H_0^2(\Omega)$ of (9.8). Moreover,*

$$\|\Delta u\|_0 \leq C_b \|f\|_0.$$

Proof. Note that the bilinear form

$$B(u, v) = \int_{\Omega} \Delta u \cdot \Delta v \, d\mathbf{x}$$

coincides with the inner product in $H_0^2(\Omega)$. On the other hand, setting,

$$Lv = \int_{\Omega} fv \, d\mathbf{x},$$

from Corollary 8.30, we have:

$$|L(v)| = \int_{\Omega} |fv| \, d\mathbf{x} \leq \|f\|_{L^2(\Omega)} \|v\|_{L^2(\Omega)} \leq C_b \|f\|_{L^2(\Omega)} \|\Delta v\|_{L^2(\Omega)}$$

so that $L \in H_0^2(\Omega)^*$. We conclude the proof directly from the Riesz Representation Theorem. \square

Remark 9.4. Let u be the solution of problem (9.8). Setting $w = \Delta u$, we have $\Delta w = f$ with $f \in L^2(\Omega)$. Thus, Corollary 8.30 implies $w \in H_{loc}^2(\Omega)$ which, in turn, yields $u \in H_{loc}^4(\Omega)$.

9.3 The Linear Elastostatic System

The system of equations of linear elastostatics models the small deformations and displacements of a solid body, which at rest occupies a bounded domain $\Omega \subset \mathbb{R}^3$. We assume that Ω is regular (or polyhedral) and write $\partial\Omega = \Gamma_D \cup \Gamma_N$, $\Gamma_D \cap \Gamma_N = \emptyset$, with Γ_D of positive surface measure. Under the action of a body force \mathbf{f} in Ω and of a traction force \mathbf{h} , acting on Γ_N , the body undergoes a deformation and each point moves from position \mathbf{x} to $\mathbf{x} + \mathbf{u}(\mathbf{x})$. Here \mathbf{f} and \mathbf{h} are vector fields in \mathbb{R}^3 . Given \mathbf{f} , \mathbf{h} , our goal is to determine the unknown displacement \mathbf{u} , in an equilibrium situation. As usual we need general and constitutive laws for \mathbf{u} .

To this purpose, introduce the *deformation tensor* given by

$$\varepsilon(\mathbf{u}) = \frac{1}{2} \left(\nabla \mathbf{u} + (\nabla \mathbf{u})^\top \right) = \frac{1}{2} \left(\frac{\partial u_i}{\partial x_j} + \frac{\partial u_j}{\partial x_i} \right)_{i,j=1,2,3}$$

and the stress tensor \mathbf{T} , which takes values on the set of symmetric 3×3 matrices. These tensors are related through *Hooke's law*

$$\mathbf{T} = 2\mu\varepsilon(\mathbf{u}) + \lambda(\operatorname{div} \mathbf{u})\mathbf{I}_3, \quad (9.9)$$

where μ, λ are the so called *Lamé coefficients* and \mathbf{I}_3 is the identity matrix in \mathbb{R}^3 . Assuming that our body is homogeneous and isotropic with respect to deformations, μ, λ are constants verifying the conditions

$$\mu > 0, 2\mu + 3\lambda > 0. \quad (9.10)$$

We add to the two constitutive laws (9.9), (9.10), the law of conservation of the linear momentum and the equation expressing the balance of the traction forces on the part Γ_N of the body surface. Moreover we assume that the part Γ_D of $\partial\Omega$ is kept fixed.

Thus we are led to the mixed problem:

$$\begin{cases} \operatorname{div} \mathbf{T} + \mathbf{f} = \mathbf{0} & \text{in } \Omega \\ \mathbf{u} = \mathbf{0} & \text{on } \Gamma_D \\ \mathbf{T} \cdot \boldsymbol{\nu} = \mathbf{h} & \text{on } \Gamma_N. \end{cases} \quad (9.11)$$

Since $\operatorname{div} \mathbf{T} = \mu \Delta \mathbf{u} + (\mu + \lambda) \operatorname{grad} \operatorname{div} \mathbf{u}$, the first equation in (9.11) can be written in the equivalent form

$$-\mu \Delta \mathbf{u} - (\mu + \lambda) \operatorname{grad} \operatorname{div} \mathbf{u} = \mathbf{f},$$

known as *Navier's equation*.

To solve Problem (9.11), we derive a variational formulation and analyze its well posedness. The boundary conditions suggest that the natural functional setting is given by the Sobolev space

$$V = H_{0,\Gamma_D}^1(\Omega; \mathbb{R}^3) = \{ \mathbf{v} \in H^1(\Omega; \mathbb{R}^3) : \mathbf{v} = \mathbf{0} \text{ on } \Gamma_D \}.$$

Since the Poincarè inequality holds in V , V is a Hilbert space with the norm

$$\| \mathbf{v} \|_V^2 = \int_{\Omega} |\nabla \mathbf{v}|^2 d\mathbf{x} = \sum_{i,j=1}^3 \int_{\Omega} \left(\frac{\partial v_i}{\partial x_j} \right)^2 d\mathbf{x}. \quad (9.12)$$

We now multiply the differential equation in (9.11) by $\mathbf{v} \in V$ (a “virtual displacement”) and integrate by parts the first integral, using the boundary conditions. We find³:

$$\begin{aligned} 0 &= \int_{\Omega} (\operatorname{div} \mathbf{T} + \mathbf{f}) \cdot \mathbf{v} d\mathbf{x} = \int_{\Omega} \sum_{i,j=1}^3 \frac{\partial T_{ij}}{\partial x_j} v_i d\mathbf{x} + \int_{\Omega} \mathbf{f} \cdot \mathbf{v} d\mathbf{x} \\ &= - \int_{\Omega} \sum_{i,j=1}^3 T_{ij} \frac{\partial v_i}{\partial x_j} d\mathbf{x} + \int_{\Gamma_N} \mathbf{h} \cdot \mathbf{v} d\sigma + \int_{\Omega} \mathbf{f} \cdot \mathbf{v} d\mathbf{x}. \end{aligned}$$

³ Recall that $(\operatorname{div} \mathbf{T})_i = \sum_{j=1}^3 \frac{\partial T_{ij}}{\partial x_j}$.

From Hooke's law (9.9), we deduce

$$\sum_{i,j=1}^3 T_{ij} \frac{\partial v_i}{\partial x_j} = 2\mu \sum_{i,j=1}^3 \varepsilon_{ij}(\mathbf{u}) \frac{\partial v_i}{\partial x_j} + \lambda \operatorname{div} \mathbf{u} \operatorname{div} \mathbf{v}.$$

Moreover, given the symmetry of the deformation tensor, it follows that

$$\sum_{i,j=1}^3 \varepsilon_{ij}(\mathbf{u}) \frac{\partial v_i}{\partial x_j} = \sum_{i,j=1}^3 \varepsilon_{ij}(\mathbf{u}) \varepsilon_{ij}(\mathbf{v})$$

and therefore we are lead to the following variational formulation of the *linear elastostatics problem*:

Determine $\mathbf{u} \in V$ such that

$$\int_{\Omega} [2\mu \sum_{i,j=1}^3 \varepsilon_{ij}(\mathbf{u}) \varepsilon_{ij}(\mathbf{v}) + \lambda \operatorname{div} \mathbf{u} \operatorname{div} \mathbf{v}] d\mathbf{x} = \int_{\Gamma_N} \mathbf{h} \cdot \mathbf{v} d\sigma + \int_{\Omega} \mathbf{f} \cdot \mathbf{v} d\mathbf{x} \quad (9.13)$$

for every $\mathbf{v} \in V$.

It is not difficult to check that, under regularity assumptions, the weak formulation is equivalent to (9.11).

We will need the following important inequality⁴, which gives a control of the full L^2 -norm of $\nabla \mathbf{v}$ by the L^2 -norm of its symmetric part $\varepsilon(\mathbf{v})$.

Lemma 9.5 (Korn's inequality). *Let Ω be a regular domain or a polyhedron. There exists a constant $\gamma > 0$ depending on Ω , such that*

$$\int_{\Omega} \left\{ \sum_{i,j=1}^3 \varepsilon_{ij}^2(\mathbf{v}) + |\mathbf{v}|^2 \right\} d\mathbf{x} \geq \gamma \|\mathbf{v}\|_{H^1(\Omega; \mathbb{R}^3)}^2, \quad \forall \mathbf{v} \in H^1(\Omega; \mathbb{R}^3).$$

The following theorem holds.

Theorem 9.6. *Let $\mathbf{f} \in L^2(\Omega; \mathbb{R}^3)$ and $\mathbf{h} \in L^2(\Gamma_N; \mathbb{R}^3)$. If Γ_D has positive surface measure, then the variational elastostatic problem has a unique solution $\mathbf{u} \in V$. Moreover*

$$\|\mathbf{u}\|_V \leq \frac{C}{\mu} \left\{ \|\mathbf{f}\|_{L^2(\Omega; \mathbb{R}^3)} + \|\mathbf{h}\|_{L^2(\Gamma_N; \mathbb{R}^3)} \right\}.$$

Proof. We use Lax-Milgram Theorem 6.39, p. 383. Define

$$B(\mathbf{u}, \mathbf{v}) = \int_{\Omega} [2\mu \sum_{i,j=1}^3 \varepsilon_{ij}(\mathbf{u}) \varepsilon_{ij}(\mathbf{v}) + \lambda \operatorname{div} \mathbf{u} \operatorname{div} \mathbf{v}] d\mathbf{x}$$

⁴ For the proof, see e.g. [24], Dautray, Lions, vol. 2, 1985.

and

$$F\mathbf{v} = \int_{\Gamma_N} \mathbf{h} \cdot \mathbf{v} \, d\sigma + \int_{\Omega} \mathbf{f} \cdot \mathbf{v} \, d\mathbf{x}.$$

Then we want to determine $\mathbf{u} \in V$ such that

$$B(\mathbf{u}, \mathbf{v}) = F\mathbf{v}, \quad \text{for every } \mathbf{v} \in V.$$

Using Schwarz's and Poincaré's inequalities and the trace Theorem 7.85, p. 483, it is easy to check that $F \in V^*$.

Also the continuity of the bilinear form B in $V \times V$ follows from a simple application of Schwarz's and Poincaré's inequalities. We leave the details to the reader. The difficulty is to show the coercivity of B . Since

$$\begin{aligned} B(\mathbf{v}, \mathbf{v}) &= \int_{\Omega} [2\mu \sum_{i,j=1}^3 \varepsilon_{ij}(\mathbf{v}) \varepsilon_{ij}(\mathbf{v}) + \lambda (\operatorname{div} \mathbf{v})^2] d\mathbf{x} \geq \\ &\geq \int_{\Omega} [2\mu \sum_{i,j=1}^3 \varepsilon_{ij}(\mathbf{v}) \varepsilon_{ij}(\mathbf{v})] d\mathbf{x}, \end{aligned}$$

the coercivity of B follows if we prove that there exists $\theta > 0$ such that

$$\sum_{i,j=1}^3 \int_{\Omega} \varepsilon_{ij}^2(\mathbf{v}) \, d\mathbf{x} \geq \theta \|\mathbf{v}\|_V^2, \quad \forall \mathbf{v} \in V. \quad (9.14)$$

We use a contradiction argument. Define

$$\|\varepsilon(\mathbf{v})\|_d^2 = \sum_{i,j=1}^3 \int_{\Omega} \varepsilon_{ij}^2(\mathbf{v}) \, d\mathbf{x}$$

and suppose (9.14) is not true. Then, for every integer $n > 0$, we can find \mathbf{v}_n such that

$$\|\varepsilon(\mathbf{v}_n)\|_d < \frac{1}{n} \|\mathbf{v}_n\|_V.$$

Normalize the sequence $\{\mathbf{v}_n\}$ by setting

$$\mathbf{w}_n = \frac{\mathbf{v}_n}{\|\mathbf{v}_n\|_{L^2(\Omega; \mathbb{R}^3)}}.$$

Still we have

$$\|\varepsilon(\mathbf{w}_n)\|_d < \frac{1}{n} \|\mathbf{w}_n\|_V, \quad \text{for all } n \geq 1.$$

By Lemma 9.5, we deduce

$$\|\mathbf{w}_n\|_V^2 \leq \gamma^{-1} \left(\|\varepsilon(\mathbf{w}_n)\|_d^2 + \|\mathbf{w}_n\|_{L^2(\Omega; \mathbb{R}^3)}^2 \right) \leq \gamma^{-1} \left(\frac{1}{n^2} \|\mathbf{w}_n\|_V^2 + 1 \right)$$

and, for $n^2 > 2/\gamma$, we get

$$\frac{1}{2} \|\mathbf{w}_n\|_V^2 \leq \gamma^{-1}.$$

Thus the sequence $\{\mathbf{w}_n\}$ is bounded in V . Moreover, as $n \rightarrow +\infty$,

$$\|\mathbf{w}_n\|_{L^2(\Omega; \mathbb{R}^3)} = 1 \quad \text{and} \quad \|\varepsilon(\mathbf{w}_n)\|_d \rightarrow 0.$$

By Rellich Theorem 7.90, p. 487, there exists a subsequence, that we still call $\{\mathbf{w}_n\}$, such that

$$\mathbf{w}_n \rightharpoonup \mathbf{w}, \quad \text{in } V,$$

and

$$\mathbf{w}_n \rightarrow \mathbf{w}, \quad \text{in } L^2(\Omega; \mathbb{R}^3).$$

By the weak lower semicontinuity of a norm (see Theorem 6.56, p. 395) we can write

$$\|\varepsilon(\mathbf{w})\|_d \leq \liminf_{n \rightarrow +\infty} \|\varepsilon(\mathbf{w}_n)\|_d = 0,$$

from which $\varepsilon(\mathbf{w}) = \mathbf{0}$ a.e. in Ω . By a classical result in the theory of rigid bodies, $\varepsilon(\mathbf{w}) = \mathbf{0}$ if and only if

$$\mathbf{w}(\mathbf{x}) = \mathbf{a} + \mathbf{M}\mathbf{x}$$

with $\mathbf{a} \in \mathbb{R}^3$ and \mathbf{M} is an antisymmetric matrix, that is $\mathbf{M} = -\mathbf{M}^\top$. Now, $|\Gamma_D| > 0$, and therefore Γ_D contains three linearly independent vectors. Since $\mathbf{w} = \mathbf{0}$ on Γ_D , we infer $\mathbf{a} = \mathbf{0}$, $\mathbf{M} = \mathbf{O}$, contradicting

$$1 = \|\mathbf{w}_n\|_{L^2(\Omega; \mathbb{R}^3)} \rightarrow \|\mathbf{w}\|_{L^2(\Omega; \mathbb{R}^3)}. \quad \square$$

Remark 9.7. As in Remark 8.2, p. 511, the variational formulation (9.13) corresponds to the *principle of virtual work* in Mechanics. In fact it expresses the balance between the works done by the internal elastic forces (given by $B(\mathbf{u}, \mathbf{v})$) and by the external forces (given by $F\mathbf{v}$) due to the *virtual displacement* \mathbf{v} . Moreover, given the symmetry of the bilinear form B , the solution \mathbf{u} minimizes the energy

$$E(\mathbf{v}) = \frac{1}{2} \int_{\Omega} [2\mu \sum_{i,j=1}^3 \varepsilon_{ij}^2(\mathbf{v}) + \lambda(\operatorname{div} \mathbf{v})^2] d\mathbf{x} - \int_{\Gamma_N} \mathbf{h} \cdot \mathbf{v} d\sigma + \int_{\Omega} \mathbf{f} \cdot \mathbf{v} d\mathbf{x}$$

among all the admissible virtual displacements. Accordingly, the variational formulation coincides with the Euler equation of the energy functional E .

Remark 9.8. If $\Gamma_D = \partial\Omega$, that is when $\mathbf{v} \in H_0^1(\Omega; \mathbb{R}^3)$, the proof of Korn's inequality is not difficult. In fact, first observe that we can write

$$\sum_{i,j=1}^3 \varepsilon_{ij}^2(\mathbf{v}) = \sum_{i,j=1}^3 \frac{1}{4} \left(\frac{\partial v_i}{\partial x_j} + \frac{\partial v_j}{\partial x_i} \right)^2 = \frac{1}{2} \sum_{i,j=1}^3 \left[\left(\frac{\partial v_i}{\partial x_j} \right)^2 + \frac{\partial v_i}{\partial x_j} \frac{\partial v_j}{\partial x_i} \right]. \quad (9.15)$$

Integrating by parts, we find

$$\begin{aligned} \int_{\Omega} \frac{\partial v_i}{\partial x_j} \frac{\partial v_j}{\partial x_i} d\mathbf{x} &= \int_{\partial\Omega} \nu_i v_j \frac{\partial v_i}{\partial x_j} d\sigma - \int_{\Omega} \frac{\partial^2 v_i}{\partial x_i \partial x_j} v_j d\mathbf{x} \\ &= \int_{\partial\Omega} \left(\nu_i v_j \frac{\partial v_i}{\partial x_j} - \nu_j v_i \frac{\partial v_i}{\partial x_i} \right) d\sigma + \int_{\Omega} \frac{\partial v_i}{\partial x_i} \frac{\partial v_j}{\partial x_j} d\mathbf{x} \\ &= \int_{\Omega} \frac{\partial v_i}{\partial x_i} \frac{\partial v_j}{\partial x_j} d\mathbf{x} = \int_{\Omega} (\operatorname{div} \mathbf{v})^2 d\mathbf{x} \end{aligned}$$

and therefore, recalling (9.12),

$$\begin{aligned} \int_{\Omega} \sum_{i,j=1}^3 \varepsilon_{ij}^2(\mathbf{v}) d\mathbf{x} &= \frac{1}{2} \int_{\Omega} |\nabla \mathbf{v}|^2 d\mathbf{x} + \frac{1}{2} \int_{\Omega} (\operatorname{div} \mathbf{v})^2 d\mathbf{x} \\ &\geq \frac{1}{2} \int_{\Omega} |\nabla \mathbf{v}|^2 d\mathbf{x} = \frac{1}{2} \|\mathbf{v}\|_V^2, \end{aligned}$$

from which Korn's inequality easily follows. Moreover, observe that

$$|\operatorname{curl} \mathbf{v}|^2 = \sum_{i,j=1, i \neq j}^3 \left[\left(\frac{\partial v_i}{\partial x_j} \right)^2 - \frac{\partial v_i}{\partial x_j} \frac{\partial v_j}{\partial x_i} \right].$$

Therefore, from (9.15), by adding and subtracting

$$\frac{1}{2} \sum_{i,j=1, i \neq j}^3 \frac{\partial v_i}{\partial x_j} \frac{\partial v_j}{\partial x_i},$$

we infer

$$\int_{\Omega} \sum_{i,j=1}^3 \varepsilon_{ij}^2(\mathbf{v}) d\mathbf{x} = \frac{1}{2} \int_{\Omega} |\operatorname{curl} \mathbf{v}|^2 d\mathbf{x} + \int_{\Omega} (\operatorname{div} \mathbf{v})^2 d\mathbf{x}$$

from which the remarkable formula

$$\int_{\Omega} |\nabla \mathbf{v}|^2 d\mathbf{x} = \int_{\Omega} |\operatorname{curl} \mathbf{v}|^2 d\mathbf{x} + \int_{\Omega} (\operatorname{div} \mathbf{v})^2 d\mathbf{x}.$$

9.4 The Stokes System

We have seen, in Subsect. 3.4.3, that the Navier-Stokes equations

$$\frac{\partial \mathbf{u}}{\partial t} + (\mathbf{u} \cdot \nabla) \mathbf{u} = \nu \Delta \mathbf{u} - \frac{1}{\rho} \nabla p + \frac{1}{\rho} \mathbf{f} \quad (9.16)$$

$$\operatorname{div} \mathbf{u} = 0 \quad (9.17)$$

express, respectively, the balance of linear momentum and the incompressibility for the motion of a homogeneous, viscous fluid. Here \mathbf{u} represents the fluid velocity, p its pressure, ρ its density (constant), while \mathbf{f} is a body force per unit volume and $\nu = \mu/\rho$ is the *kinematic viscosity* (constant). Note that the pressure appears in (9.16) only through its gradient so that it is determined up to an additive constant, as it should be.

To the equations (9.16), (9.17) we add an *initial condition*

$$\mathbf{u}(0) = \mathbf{g}$$

and conditions on the boundary of a (bounded) domain $\Omega \subset \mathbb{R}^n$ ($n = 2, 3$), where the fluid is confined. Typically, the observation of real fluids reveals that the normal and tangential components of their velocity vector on a rigid wall must agree with those of the wall itself. Thus if the wall is at rest we have $\mathbf{u} = \mathbf{0}$ (*no slip condition*).

In order to emphasize the role of the convective term $(\mathbf{u} \cdot \nabla)\mathbf{u}$, we pass to dimensionless variables by choosing a reference length L related to Ω , for instance $L = \text{diameter of } \Omega$ or $L = |\Omega|^{1/n}$, and a reference velocity $U = \|\mathbf{g}\|_{L^\infty(\Omega; \mathbb{R}^n)}$.

Then, we rescale the quantities in our problem in the following way⁵:

$$\mathbf{y} = \frac{\mathbf{x}}{L}, \quad \tau = \frac{\nu t}{L^2}, \quad \mathbf{v} = \frac{\mathbf{u}}{U}, \quad P = \frac{pL}{\rho\nu U}, \quad \mathbf{F} = \frac{L^2}{\nu U} \mathbf{f}.$$

After elementary calculations, equation (9.17) remains unaltered while (9.16) takes the following dimensionless form:

$$\frac{\partial \mathbf{v}}{\partial \tau} + \mathcal{R}(\mathbf{v} \cdot \nabla)\mathbf{v} = \Delta \mathbf{v} - \nabla P + \mathbf{F} \quad (9.18)$$

in the rescaled domain $\Omega' = \{\mathbf{y} : L\mathbf{y} \in \Omega\}$, where the only parameter left is

$$\mathcal{R} = \frac{UL}{\nu},$$

known as *Reynolds number*. Low (high) Reynolds number means high viscosity or motion weakly (strongly) affected by the initial data. As $\mathcal{R} \rightarrow 0$ the equation (9.18) linearizes into the diffusion equation for the velocity field. In a stationary regime we are led to the following problem:

$$\begin{cases} -\Delta \mathbf{v} = \mathbf{F} - \nabla P & \text{in } \Omega' \\ \operatorname{div} \mathbf{v} = 0 & \text{in } \Omega' \\ \mathbf{v} = \mathbf{0} & \text{on } \partial\Omega'. \end{cases} \quad (9.19)$$

We want to analyze the well posedness of problem (9.19) by first deriving a variational formulation. Thus, take a test function $\varphi \in C_0^\infty(\Omega'; \mathbb{R}^n)$, multiply by φ the first equation in (9.19) and integrate over Ω' . We find:

$$\int_{\Omega'} -\Delta \mathbf{v} \cdot \varphi \, d\mathbf{x} = \int_{\Omega'} \mathbf{F} \cdot \varphi \, d\mathbf{x} - \int_{\Omega'} \nabla P \cdot \varphi \, d\mathbf{x}.$$

After an integration by parts in the first and in the third term we get:

$$\int_{\Omega'} \nabla \mathbf{v} : \nabla \varphi \, d\mathbf{x} = \int_{\Omega'} \mathbf{F} \cdot \varphi \, d\mathbf{x} + \int_{\Omega'} P \operatorname{div} \varphi \, d\mathbf{x}, \quad \forall \varphi \in C_0^\infty(\Omega'; \mathbb{R}^n) \quad (9.20)$$

⁵ Note that the physical dimensions of ν are $[length]^2 \times [time]^{-1}$.

where

$$\nabla \mathbf{v} : \nabla \varphi = \sum_{i,j=1}^n \frac{\partial u_i}{\partial x_j} \frac{\partial v_i}{\partial x_j}.$$

Now, multiply the equation $\operatorname{div} \mathbf{v} = 0$ by $q \in L^2(\Omega')$ and integrate over Ω' ; we have:

$$\int_{\Omega'} q \operatorname{div} \mathbf{v} d\mathbf{x} = 0, \quad \forall q \in L^2(\Omega'). \quad (9.21)$$

Viceversa, under regularity assumptions, if $\mathbf{v} = \mathbf{0}$ on $\partial\Omega'$ and it satisfies (9.20), going back with the integration by parts we find

$$\int_{\Omega} (-\Delta \mathbf{v} - \mathbf{F} + \nabla P) \cdot \mathbf{v} d\mathbf{x} = 0, \quad \forall \mathbf{v} \in C_0^\infty(\Omega'; \mathbb{R}^n)$$

which implies, by a usual density argument,

$$-\Delta \mathbf{v} - \mathbf{F} + \nabla P = \mathbf{0} \text{ in } \Omega'.$$

Moreover, from (9.21) we deduce

$$\operatorname{div} \mathbf{v} = 0 \text{ in } \Omega'.$$

Thus under regularity assumptions, for functions vanishing on $\partial\Omega$, the Stokes system (9.19) is equivalent to the system of the two equations (9.20), (9.21).

These two equations and the homogeneous boundary condition suggest that the natural functional setting is the space $H_0^1(\Omega'; \mathbb{R}^n)$. Observe that, by density, the equation (9.20) holds for every $\varphi \in H_0^1(\Omega'; \mathbb{R}^n)$. To select uniquely the pressure, introduce the Hilbert space⁶

$$Q = \left\{ q \in L^2(\Omega') : \int_{\Omega} q = 0 \right\}.$$

Finally, let $\mathbf{F} \in L^2(\Omega'; \mathbb{R}^n)$.

Definition 9.9. A variational solution of the problem (9.19) is a pair (\mathbf{v}, P) such that:

$$\begin{cases} \mathbf{v} \in H_0^1(\Omega'; \mathbb{R}^n), P \in Q \\ \int_{\Omega'} \nabla \mathbf{v} : \nabla \varphi d\mathbf{x} = \int_{\Omega'} \mathbf{F} \cdot \varphi d\mathbf{x} + \int_{\Omega'} P \operatorname{div} \varphi d\mathbf{x} & \forall \varphi \in H_0^1(\Omega'; \mathbb{R}^n) \\ \int_{\Omega'} q \operatorname{div} \mathbf{v} d\mathbf{x} = 0 & \forall q \in Q. \end{cases} \quad (9.22)$$

⁶ Other normalization of the pressure are possible and sometimes more convenient from a numerical point of view.

We want to show existence, uniqueness and stability of a weak solution. However, a direct use of the Lax-Milgram Theorem is prevented by the presence of the unknown pressure P . A way to overcome this difficulty is to choose, instead of $H_0^1(\Omega'; \mathbb{R}^n)$, its closed subspace V_{div} , whose elements are the vectors with vanishing divergence. With the norm

$$\|\mathbf{v}\|_{V_{div}}^2 = \int_{\Omega'} |\nabla \mathbf{v}|^2 d\mathbf{x} = \sum_{i,j=1}^n \int_{\Omega'} \nabla \mathbf{v} : \nabla \mathbf{v}$$

is a Hilbert space and if we choose $\varphi \in V_{div}$, the second equation in (9.22) becomes

$$\int_{\Omega'} \nabla \mathbf{v} : \nabla \varphi d\mathbf{x} = \int_{\Omega'} \mathbf{F} \cdot \varphi d\mathbf{x}, \quad \forall \varphi \in V_{div}. \quad (9.23)$$

The bilinear form

$$B(\mathbf{v}, \varphi) = \int_{\Omega'} \nabla \mathbf{v} : \nabla \varphi d\mathbf{x}$$

is clearly bounded and coercive in V_{div} . Moreover, $\mathbf{F} \in (V_{div})^*$ since $\mathbf{F} \in L^2(\Omega'; \mathbb{R}^n)$. The Lax-Milgram Theorem gives a unique solution $\bar{\mathbf{v}} \in V_{div}$, that also incorporates the incompressibility condition $\operatorname{div} \mathbf{v} = 0$. Moreover, using also Poincaré's inequality,

$$\|\bar{\mathbf{v}}\|_{V_{div}} \leq C_P \|\mathbf{F}\|_{L^2(\Omega'; \mathbb{R}^n)}.$$

Still we have to determine the pressure. To this purpose, note that the equation (9.23) asserts that the vector

$$\mathbf{g} = \Delta \bar{\mathbf{v}} + \mathbf{F},$$

which belongs to $H^{-1}(\Omega'; \mathbb{R}^n)$, satisfies the equation

$$\langle \mathbf{g}, \varphi \rangle_{H^{-1}, H_0^1} = 0, \quad \forall \varphi \in V_{div}.$$

In other words, the functional \mathbf{g} vanishes over V_{div} . It turns out that the elements of $H^{-1}(\Omega'; \mathbb{R}^n)$ that vanish over V_{div} have a special form, as indicated in the following lemma.

Lemma 9.10. *let $\Omega' \subset \mathbb{R}^n$ be a bounded Lipschitz domain. A functional $\mathbf{g} \in H^{-1}(\Omega'; \mathbb{R}^n)$ satisfies the condition*

$$\langle \mathbf{g}, \varphi \rangle_{H^{-1}, H_0^1} = 0, \quad \forall \varphi \in V_{div} \quad (9.24)$$

if and only if there exists $P \in L^2(\Omega')$ such that

$$-\nabla P = \mathbf{g}.$$

The function P is unique up to an additive constant.

Justification. The proof is rather complex⁷. We only give a formal argument to support the plausibility of the result. Let us work in dimension $n = 3$, assuming that \mathbf{g} is smooth and that Ω is, say, a cube. First note that if $\mathbf{V} \in \mathcal{D}(\Omega'; \mathbb{R}^3)$,

$$\operatorname{div} \operatorname{curl} \mathbf{V} = \mathbf{0},$$

and therefore (9.24) holds for

$$\boldsymbol{\varphi} = \operatorname{curl} \mathbf{V}.$$

From Gauss formula 8, in Appendix C, we have

$$\begin{aligned} 0 &= \int_{\Omega'} \mathbf{g} \cdot \operatorname{curl} \mathbf{V} \, d\mathbf{x} = \int_{\Omega'} \operatorname{curl} \mathbf{g} \cdot \mathbf{V} \, d\mathbf{x} - \int_{\partial\Omega'} (\mathbf{g} \times \mathbf{V}) \cdot \mathbf{n} \, d\sigma \\ &= \int_{\Omega'} \operatorname{curl} \mathbf{g} \cdot \mathbf{V} \, d\mathbf{x}. \end{aligned}$$

Since \mathbf{V} is arbitrary, we infer that

$$\operatorname{curl} \mathbf{g} = \mathbf{0}.$$

Thus, there exists a scalar potential P such that $-\nabla P = \mathbf{g}$. \square

Thanks to Lemma 9.10, there exists a unique $\overline{P} \in Q$ such that

$$\mathbf{g} = \Delta \overline{\mathbf{v}} + \mathbf{F} = -\nabla \overline{P}.$$

Hence, $(\overline{\mathbf{v}}, \overline{P})$ is the unique solution of problem (9.22). We have proved the following result:

Theorem 9.11. *Let $\Omega' \subset \mathbb{R}^3$ be a bounded, Lipschitz domain and $\mathbf{F} \in L^2(\Omega'; \mathbb{R}^n)$. The problem (9.19) has a unique weak solution $(\overline{\mathbf{v}}, \overline{P})$, with $\overline{\mathbf{v}} \in H_0^1(\Omega'; \mathbb{R}^n)$ and $\overline{P} \in Q$.*

Remark 9.12 (The pressure as a multiplier). Given the symmetry of the bilinear form B , the solution $\overline{\mathbf{v}}$ is a minimizer in V_{div} of the energy functional

$$E(\mathbf{v}) = \frac{1}{2} \int_{\Omega'} \left\{ |\nabla \mathbf{v}|^2 - 2\mathbf{F} \cdot \mathbf{v} \right\} d\mathbf{x}.$$

Equivalently, $\overline{\mathbf{v}}$ minimizes $E(\mathbf{v})$ in all $H_0^1(\Omega'; \mathbb{R}^n)$ under the constraint $\operatorname{div} \mathbf{v} = 0$. Introducing a multiplier $q \in Q$ and the Lagrangian

$$\mathcal{L}(\mathbf{v}, q) = \int_{\Omega'} \left\{ \frac{1}{2} |\nabla \mathbf{v}|^2 - \mathbf{F} \cdot \mathbf{v} - q \operatorname{div} \mathbf{v} \right\} d\mathbf{x},$$

the system (9.22) expresses a necessary optimality condition (Euler-Lagrange equation). Thus the pressure P appears as a multiplier associated to the solenoidality constraint $\operatorname{div} \mathbf{v} = 0$.

⁷ See [9], *Galdi*, 1994.

9.5 The Stationary Navier Stokes Equations

In this section we examine a Dirichlet problem for the Navier-Stokes equations. Using the fixed point theorem of Leray-Schauder in Sect. 6.10, we show the existence of a solution. Then, via the contraction mapping Theorem, we give a uniqueness result.

9.5.1 Weak formulation and existence of a solution

Let $\Omega \subset \mathbb{R}^n$, ($n = 2, 3$) be a bounded domain and consider the stationary Navier-Stokes system, (we assume $\rho = 1$):

$$\begin{cases} -\nu \Delta \mathbf{u} + (\mathbf{u} \cdot \nabla) \mathbf{u} = -\nabla p + \mathbf{f} & \text{in } \Omega \\ \operatorname{div} \mathbf{u} = 0 & \text{in } \Omega \\ \mathbf{u} = \mathbf{0} & \text{on } \partial\Omega. \end{cases} \quad (9.25)$$

Keeping the same notations of the last section, a natural weak formulation of problem (9.25) is the following.

Definition 9.13. A weak (or variational) solution of problem (9.25) is a pair (\mathbf{u}, p) such that $\mathbf{u} \in H_0^1(\Omega; \mathbb{R}^n)$, $p \in Q$ and

$$\begin{aligned} \int_{\Omega} \{\nu \nabla \mathbf{u} : \nabla \mathbf{v} + (\mathbf{u} \cdot \nabla) \mathbf{u} \cdot \mathbf{v}\} d\mathbf{x} &= \int_{\Omega} \{\mathbf{f} \cdot \mathbf{v} + p \operatorname{div} \mathbf{v}\} d\mathbf{x}, \quad \forall \mathbf{v} \in H_0^1(\Omega; \mathbb{R}^n), \\ \int_{\Omega} q \operatorname{div} \mathbf{u} d\mathbf{x} &= 0, \quad \forall q \in Q. \end{aligned}$$

The presence of the nonlinear term introduces nontrivial difficulties. In general we cannot expect uniqueness⁸ and also the proof of the existence of a solution requires some effort.

Using a fixed point technique based on the Leray-Schauder Theorem 6.84, p. 420, we can prove the following result.

Theorem 9.14. Let $\Omega \subset \mathbb{R}^n$ ($n = 2, 3$) be a bounded, Lipschitz domain and $\mathbf{f} \in L^2(\Omega; \mathbb{R}^n)$. Then there exists a weak solution (\mathbf{u}, p) of problem (9.25).

Proof. We split it in several steps.

1. *Pressure elimination.* As in the case of Stokes equation, we incorporate the incompressibility condition $\operatorname{div} \mathbf{u} = 0$, by solving in V_{div} the problem

$$\nu \int_{\Omega} \nabla \mathbf{u} : \nabla \mathbf{v} d\mathbf{x} = \int_{\Omega} \mathbf{f} \cdot \mathbf{v} d\mathbf{x} - \int_{\Omega} (\mathbf{u} \cdot \nabla) \mathbf{u} \cdot \mathbf{v} d\mathbf{x}, \quad \forall \mathbf{v} \in V_{\operatorname{div}}. \quad (9.26)$$

Once \mathbf{u} is determined, we recover the pressure $p \in L^2(\Omega)$ using Lemma 9.10.

⁸ See [9], Galdi, 1994.

2. *Reduction to a fixed point problem.* Introduce the *trilinear form*

$$b(\mathbf{u}, \mathbf{w}, \mathbf{v}) = \int_{\Omega} (\mathbf{u} \cdot \nabla) \mathbf{w} \cdot \mathbf{v} \, d\mathbf{x}, \quad \mathbf{u}, \mathbf{w}, \mathbf{v} \in V_{div},$$

and write (9.26) in the form

$$\nu \int_{\Omega} \nabla \mathbf{u} : \nabla \mathbf{v} \, d\mathbf{x} = \int_{\Omega} \mathbf{f} \cdot \mathbf{v} \, d\mathbf{x} - b(\mathbf{u}, \mathbf{u}, \mathbf{v}).$$

Now fix $\mathbf{w} \in V_{div}$ and consider the *linear* problem

$$a(\mathbf{u}, \mathbf{v}) = F_{\mathbf{w}}(\mathbf{v}), \quad \forall \mathbf{v} \in V_{div}, \quad (9.27)$$

where

$$a(\mathbf{u}, \mathbf{v}) = \nu \int_{\Omega} \nabla \mathbf{u} : \nabla \mathbf{v} \, d\mathbf{x}$$

and

$$F_{\mathbf{w}}(\mathbf{v}) = \int_{\Omega} \mathbf{f} \cdot \mathbf{v} \, d\mathbf{x} - b(\mathbf{w}, \mathbf{w}, \mathbf{v}).$$

We shall use the Lax-Milgram Theorem⁹ to prove that, for every $\mathbf{w} \in V_{div}$, the problem (9.27) has a unique solution

$$\mathbf{u} = T(\mathbf{w}) \in V_{div}.$$

Now, any *fixed point* of the map $T : V_{div} \rightarrow V_{div}$, that is any function \mathbf{u}^* such that $T(\mathbf{u}^*) = \mathbf{u}^*$, is a solution of the equation (9.26). Thus to conclude the proof, we must show that T has a fixed point.

3. *Solution of the linear problem (9.27).* Since

$$a(\mathbf{u}, \mathbf{u}) = \nu \int_{\Omega} \nabla \mathbf{u} : \nabla \mathbf{u} \, d\mathbf{x} = \nu \|\mathbf{u}\|_{V_{div}}^2,$$

the bilinear form $a(\mathbf{u}, \mathbf{v})$ is continuous and coercive in V_{div} . We show that $F_{\mathbf{w}} \in (V_{div})^*$. Since $\mathbf{f} \in L^2(\Omega; \mathbb{R}^n)$ the functional $\mathbf{v} \mapsto \int_{\Omega} \mathbf{f} \cdot \mathbf{v} \, d\mathbf{x}$ clearly belongs to $(V_{div})^*$. Let us examine the trilinear form $b(\mathbf{u}, \mathbf{w}, \mathbf{v})$. Since $\operatorname{div} \mathbf{u} = 0$ and \mathbf{w}, \mathbf{v} vanish on $\partial\Omega$, we have:

$$\begin{aligned} b(\mathbf{u}, \mathbf{w}, \mathbf{v}) + b(\mathbf{u}, \mathbf{v}, \mathbf{w}) &= \int_{\Omega} \sum_{i,j=1}^n [u_i \frac{\partial w_j}{\partial x_i} v_j + u_i \frac{\partial v_j}{\partial x_i} w_j] d\mathbf{x} \\ &= \int_{\Omega} \mathbf{u} \cdot \nabla(\mathbf{w} \cdot \mathbf{v}) \, d\mathbf{x} = \int_{\Omega} (\mathbf{w} \cdot \mathbf{v}) \operatorname{div} \mathbf{u} \, d\mathbf{x} = 0 \end{aligned}$$

so that

$$b(\mathbf{u}, \mathbf{w}, \mathbf{v}) = -b(\mathbf{u}, \mathbf{v}, \mathbf{w}). \quad (9.28)$$

In particular

$$b(\mathbf{u}, \mathbf{w}, \mathbf{w}) = 0. \quad (9.29)$$

Now

$$|(\mathbf{u} \cdot \nabla) \mathbf{w} \cdot \mathbf{v}| \leq \sum_{i,j=1}^n \left| u_i \frac{\partial w_j}{\partial x_i} v_j \right| \leq \|\mathbf{u}\| \|\nabla \mathbf{w}\| \|\mathbf{v}\|$$

⁹ Actually the Riesz Representation Theorem is enough.

and, since $\frac{1}{4} + \frac{1}{2} + \frac{1}{4} = 1$, we can write¹⁰

$$\int_{\Omega} |\mathbf{u}| |\nabla \mathbf{w}| |\mathbf{v}| \, d\mathbf{x} \leq \|\mathbf{u}\|_{L^4(\Omega; \mathbb{R}^n)} \|\mathbf{w}\|_{V_{div}} \|\mathbf{v}\|_{L^4(\Omega; \mathbb{R}^n)}.$$

Thus

$$|b(\mathbf{u}, \mathbf{w}, \mathbf{v})| \leq \|\mathbf{u}\|_{L^4(\Omega; \mathbb{R}^n)} \|\mathbf{w}\|_{V_{div}} \|\mathbf{v}\|_{L^4(\Omega; \mathbb{R}^n)}. \quad (9.30)$$

From the embedding Theorem 7.96, p. 492, we have, if $n = 2$, $H_0^1(\Omega) \subset L^s(\Omega)$ for every $s \geq 2$, with compact embedding. If $n = 3$, $H_0^1(\Omega) \subset L^s(\Omega)$ for every $s \in [2, 6]$, with compact embedding whenever $s < 6$. Moreover the following estimate

$$\|\mathbf{u}\|_{L^s(\Omega; \mathbb{R}^n)} \leq \tilde{C}_s \|\mathbf{u}\|_{V_{div}} \quad (9.31)$$

holds in all cases, with $\tilde{C}_s = \tilde{C}(s, \Omega)$. Then we may also write:

$$|b(\mathbf{u}, \mathbf{w}, \mathbf{v})| \leq \tilde{C}_4^2 \|\mathbf{u}\|_{V_{div}} \|\mathbf{w}\|_{V_{div}} \|\mathbf{v}\|_{V_{div}}. \quad (9.32)$$

Choosing $\mathbf{u} = \mathbf{w}$, we deduce that the linear functional

$$\mathbf{v} \longmapsto b(\mathbf{w}, \mathbf{w}, \mathbf{v})$$

is bounded in V_{div} . Therefore $F_{\mathbf{w}} \in (V_{div})^*$ and (9.27) has a unique solution \mathbf{u} such that

$$\|\nabla \mathbf{u}\|_{L^2(\Omega; \mathbb{R}^n)} \leq \frac{1}{\nu} \left\{ C_P \|\mathbf{f}\|_{L^2(\Omega; \mathbb{R}^n)} + \tilde{C}_4^2 \|\mathbf{w}\|_{V_{div}}^2 \right\}$$

where C_P is a Poincaré constant.

4. T has a fixed point. To use the Leray-Schauder Theorem we must check that:

- a) $T : V_{div} \rightarrow V_{div}$ is compact and continuous.
- b) The set of solutions of the equation

$$\mathbf{u} = sT(\mathbf{u}), \quad 0 \leq s \leq 1,$$

is bounded in $H_0^1(\Omega; \mathbb{R}^n)$.

a) T is compact. Let $\{\mathbf{w}_k\}$ be a bounded sequence in V_{div} , say,

$$\|\mathbf{w}_k\|_{V_{div}} \leq M. \quad (9.33)$$

We want to prove that there exists a convergent subsequence $\{T(\mathbf{w}_{k_j})\}$. By definition, using (9.28),

$$a(T(\mathbf{w}_k) - T(\mathbf{w}_h), \mathbf{v}) = -b(\mathbf{w}_k, \mathbf{w}_k, \mathbf{v}) + b(\mathbf{w}_h, \mathbf{w}_h, \mathbf{v}) = b(\mathbf{w}_k, \mathbf{v}, \mathbf{w}_k) - b(\mathbf{w}_h, \mathbf{v}, \mathbf{w}_h)$$

and, adding and subtracting $b(\mathbf{w}_k, \mathbf{v}, \mathbf{w}_h)$, we find:

$$a(T(\mathbf{w}_k) - T(\mathbf{w}_h), \mathbf{v}) = b(\mathbf{w}_k, \mathbf{v}, \mathbf{w}_k - \mathbf{w}_h) - b(\mathbf{w}_h - \mathbf{w}_k, \mathbf{v}, \mathbf{w}_h).$$

¹⁰ We use the generalized Hölder inequality:

$$\left| \int_{\Omega} fgh \, d\mathbf{x} \right| \leq \|f\|_{L^p} \|g\|_{L^q} \|h\|_{L^r}$$

with $p, q, r > 1$, $\frac{1}{4} + \frac{1}{2} + \frac{1}{4} = 1$.

From (9.30) and (9.33), we get:

$$\begin{aligned} |a(T(\mathbf{w}_k) - T(\mathbf{w}_h), \mathbf{v})| &\leq (\|\mathbf{w}_k\|_{L^4(\Omega; \mathbb{R}^n)} + \|\mathbf{w}_h\|_{L^4(\Omega; \mathbb{R}^n)}) \|\mathbf{v}\|_{V_{div}} \|\mathbf{w}_k - \mathbf{w}_h\|_{L^4(\Omega; \mathbb{R}^n)} \\ &\leq 2\tilde{C}_4 M \|\mathbf{w}_k - \mathbf{w}_h\|_{L^4(\Omega; \mathbb{R}^n)} \|\mathbf{v}\|_{V_{div}}. \end{aligned}$$

Choosing $\mathbf{v} = T(\mathbf{w}_k) - T(\mathbf{w}_h)$ and recalling the definition of a , we obtain, simplifying by $\|T(\mathbf{w}_k) - T(\mathbf{w}_h)\|_{V_{div}}$:

$$\|T(\mathbf{w}_k) - T(\mathbf{w}_h)\|_{V_{div}} \leq \frac{2\tilde{C}_4 M}{\nu} \|\mathbf{w}_k - \mathbf{w}_h\|_{L^4(\Omega; \mathbb{R}^n)}. \quad (9.34)$$

By the compactness of the embedding $H_0^1(\Omega; \mathbb{R}^n) \hookrightarrow L^4(\Omega; \mathbb{R}^n)$, there exists a subsequence $\{\mathbf{w}_{k_j}\}$ convergent in $L^4(\Omega; \mathbb{R}^n)$.

In particular

$$\|\mathbf{w}_{k_j} - \mathbf{w}_{k_m}\|_{L^4(\Omega; \mathbb{R}^n)} \rightarrow 0$$

as $j, m \rightarrow \infty$. From (9.34) with $k = k_j$, $h = k_m$, we infer that $\{T(\mathbf{w}_{k_j})\}$ is a Cauchy sequence, and therefore convergent in V_{div} .

T is continuous. Let $\{\mathbf{w}_k\} \subset V_{div}$, $\mathbf{w}_k \rightarrow \mathbf{w}$, $k \rightarrow +\infty$. In particular $\{\mathbf{w}_k\}$ is bounded, say $\|\mathbf{w}_k\|_{V_{div}} \leq M$. Then $T(\mathbf{w}_k) - T(\mathbf{w})$, solves the equation

$$\begin{aligned} a(T(\mathbf{w}_k) - T(\mathbf{w}), \mathbf{v}) &= -b(\mathbf{w}_k, \mathbf{w}_k, \mathbf{v}) + b(\mathbf{w}, \mathbf{w}, \mathbf{v}) \\ &= b(\mathbf{w}_k, \mathbf{v}, \mathbf{w}_k - \mathbf{w}) - b(\mathbf{w}_k - \mathbf{w}, \mathbf{v}, \mathbf{w}). \end{aligned}$$

From (9.34) and (9.32) we get, choosing $\mathbf{v} = T(\mathbf{w}_k) - T(\mathbf{v})$,

$$\|T(\mathbf{w}_k) - T(\mathbf{w})\|_{V_{div}} \leq \frac{2\tilde{C}_4^2 M}{\nu} \|\mathbf{w}_k - \mathbf{w}\|_{V_{div}}$$

from which the continuity of T follows.

b) Let $\mathbf{u} \in V_{div}$ solve the equation $\mathbf{u} = sT(\mathbf{u})$, for some $s \in (0, 1]$. Then

$$\frac{\nu}{s} \int_{\Omega} \nabla \mathbf{u} : \nabla \mathbf{v} \, d\mathbf{x} = \int_{\Omega} \mathbf{f} \cdot \mathbf{v} \, d\mathbf{x} - b(\mathbf{u}, \mathbf{u}, \mathbf{v}), \quad \forall \mathbf{v} \in V_{div}.$$

In particular, choosing $\mathbf{v} = \mathbf{u}$, we have $b(\mathbf{u}, \mathbf{u}, \mathbf{u}) = 0$ and

$$\begin{aligned} \nu \|\mathbf{u}\|_{V_{div}}^2 &= s \int_{\Omega} \mathbf{f} \cdot \mathbf{u} \, d\mathbf{x} \leq \|\mathbf{f}\|_{L^2(\Omega; \mathbb{R}^n)} \|\mathbf{u}\|_{L^2(\Omega; \mathbb{R}^n)} \\ &\leq C_P \|\mathbf{f}\|_{L^2(\Omega; \mathbb{R}^n)} \|\mathbf{u}\|_{V_{div}} \end{aligned}$$

from which

$$\|\mathbf{u}\|_{V_{div}} \leq \frac{C_P \|\mathbf{f}\|_{L^2(\Omega; \mathbb{R}^n)}}{\nu}. \quad (9.35)$$

This concludes the proof. \square

9.5.2 Uniqueness

We have already observed that a solution of problem (9.25) may not be unique. However, if $\|\mathbf{f}\|_{L^2(\Omega; \mathbb{R}^n)}$ is small or ν is large enough, we can show uniqueness via the *Banach Contraction Theorem* 6.78, p. 417. Precisely:

Theorem 9.15. let $\Omega \subset \mathbb{R}^n$ ($n = 2, 3$) be a bounded, Lipschitz domain and $\mathbf{f} \in L^2(\Omega; \mathbb{R}^n)$. If \tilde{C}_4 is the constant in the embedding inequality (9.31) for $s = 4$ and

$$\frac{C_P \tilde{C}_4^2}{\nu^2} \|\mathbf{f}\|_{L^2(\Omega; \mathbb{R}^n)} < 1, \quad (9.36)$$

then, there exists a unique weak solution $(\mathbf{u}, p) \in V_{div} \times Q$ of problem (9.25).

Proof. We slightly modify the proof of Theorem 9.14. For fixed $\mathbf{w} \in V_{div}$, we consider the linear problem

$$\nu \int_{\Omega} \nabla \mathbf{u} : \nabla \mathbf{v} \, d\mathbf{x} + b(\mathbf{w}, \mathbf{u}, \mathbf{v}) = \int_{\Omega} \mathbf{f} \cdot \mathbf{v}, \quad \forall \mathbf{v} \in V_{div}. \quad (9.37)$$

The bilinear form

$$\tilde{a}(\mathbf{u}, \mathbf{v}) = \nu \int_{\Omega} \nabla \mathbf{u} : \nabla \mathbf{v} \, d\mathbf{x} + b(\mathbf{w}, \mathbf{u}, \mathbf{v})$$

is continuous in V_{div} (from (9.32)) and coercive, since $b(\mathbf{w}, \mathbf{v}, \mathbf{v}) = 0$ and therefore

$$\tilde{a}(\mathbf{v}, \mathbf{v}) = \nu \int_{\Omega} \nabla \mathbf{v} : \nabla \mathbf{v} \, d\mathbf{x} + b(\mathbf{w}, \mathbf{v}, \mathbf{v}) = \nu \|\mathbf{v}\|_{V_{div}}^2.$$

By the Lax-Milgram Theorem, there exists a unique solution $\mathbf{u} = T(\mathbf{w})$ of (9.37) and

$$\|T(\mathbf{w})\|_{V_{div}} \leq \frac{C_P}{\nu} \|\mathbf{f}\|_{L^2(\Omega; \mathbb{R}^n)}.$$

Set $M = \frac{C_P}{\nu} \|\mathbf{f}\|_{L^2(\Omega; \mathbb{R}^n)}$ and define

$$B_M = \left\{ \mathbf{w} \in V_{div} : \|\mathbf{w}\|_{V_{div}} \leq M \right\}.$$

Endowed with the distance $d(\mathbf{w}_1, \mathbf{w}_2) = \|\mathbf{w}_1 - \mathbf{w}_2\|_{V_{div}}$, B_M is a complete metric space and

$$T : B_M \rightarrow B_M.$$

We claim that if M is small enough, T is a strict contraction in B_M , i.e. there exists $\delta < 1$ such that

$$\|T(\mathbf{w}) - T(\mathbf{z})\|_{V_{div}} \leq \delta \|\mathbf{w} - \mathbf{z}\|_{V_{div}}, \quad \forall \mathbf{w}, \mathbf{z} \in B_M.$$

In fact, proceeding as in the proof of Theorem 9.14, we write

$$a(T(\mathbf{w}) - T(\mathbf{z}), \mathbf{v}) = -b(\mathbf{w}, T(\mathbf{w}), \mathbf{v}) + b(\mathbf{z}, T(\mathbf{z}), \mathbf{v}).$$

Adding and subtracting $b(\mathbf{w}, T(\mathbf{z}), \mathbf{v})$, we obtain, using also (9.28):

$$a(T(\mathbf{w}) - T(\mathbf{z}), \mathbf{v}) = -b(\mathbf{w}, T(\mathbf{w}) - T(\mathbf{z}), \mathbf{v}) - b(\mathbf{z} - \mathbf{w}, \mathbf{v}, T(\mathbf{z})),$$

for every $\mathbf{v} \in V_{div}$. Choosing $\mathbf{v} = T(\mathbf{w}) - T(\mathbf{z})$ and remembering (9.29) and (9.32), we find:

$$\|T(\mathbf{w}) - T(\mathbf{z})\|_{V_{div}} \leq \frac{M \tilde{C}_4^2}{\nu} \|\mathbf{w} - \mathbf{z}\|_{V_{div}}.$$

Thus T is a strict contraction if $\frac{M \tilde{C}_4^2}{\nu} < 1$. By the Contraction Theorem 6.78, p. 417, T has a unique fixed point $\mathbf{u} \in B_M$, which is also a solution of problem (9.25). On the other hand, if \mathbf{u} is a solution of problem (9.25), then $\mathbf{u} \in B_M$ and therefore \mathbf{u} is the unique solution. \square

9.6 A Control Problem

Control problems are more and more important in modern technology. We give here an application of the variational theory we have developed so far, to a fairly simple temperature control problem.

9.6.1 Structure of the problem

Suppose that the temperature u of a homogeneous body, occupying a smooth bounded domain $\Omega \subset \mathbb{R}^3$, satisfies the following stationary conditions:

$$\begin{cases} \mathcal{E}u \equiv -\Delta u + \operatorname{div}(\mathbf{b}u) = z & \text{in } \Omega \\ u = 0 & \text{on } \partial\Omega, \end{cases} \quad (9.38)$$

where $\mathbf{b} \in C^1(\overline{\Omega}; \mathbb{R}^3)$ is given, with $\operatorname{div}\mathbf{b} \geq 0$ in Ω .

In (9.38) we distinguish two types of dependent variables: the **control** variable z , that we take in $H = L^2(\Omega)$, and the **state** variable u . Coherently, (9.38) is called the **state system**. Given a control z , from Corollary 8.22, (9.38) has a unique weak solution $u[z] \in V = H_0^1(\Omega)$.

Thus, setting

$$a(u, v) = \int_{\Omega} (\nabla u \cdot \nabla v - u \mathbf{b} \cdot \nabla v) d\mathbf{x},$$

$u[z]$ satisfies the **state** equation

$$a(u[z], v) = (z, v)_H \quad \forall v \in V \quad (9.39)$$

and

$$\|u[z]\|_V \leq C_p \|z\|_H. \quad (9.40)$$

From elliptic regularity (Theorem 8.28, p. 539) it follows that $u \in H^2(\Omega) \cap H_0^1(\Omega)$, so that u is a *strong solution* of the state equation and satisfies it in the a.e. pointwise sense as well.

Our problem is to choose the source term z in order to minimize the “distance” of u from a given target state u_d .

Of course there are many ways to measure the distance of u from u_d . If we are interested in a distance which involves u and u_d over an open subset $\Omega_0 \subseteq \Omega$, a reasonable choice may be

$$J(u, z) = \frac{1}{2} \int_{\Omega_0} (u - u_d)^2 d\mathbf{x} + \frac{\beta}{2} \int_{\Omega} z^2 d\mathbf{x} \quad (9.41)$$

where $\beta > 0$.

$J(u, z)$ is called **cost functional** or **performance index**. The second term in (9.41) is called *penalization term*; its role is, on one hand, to avoid using “too

large" controls in the minimization of J , on the other hand, to assure coercivity for J , as we shall see later on.

Summarizing, we may write our control problem in the following way:

Find $(u^*, z^*) \in H \times V$, such that

$$\begin{cases} J(u^*, z^*) = \min_{(u,z) \in V \times H} J(u, z) \\ \text{under the conditions} \\ \mathcal{E}u = z \text{ in } \Omega, \quad u = 0 \text{ on } \partial\Omega. \end{cases} \quad (9.42)$$

If (u^*, z^*) is a minimizing pair, u^* and z^* are called **optimal state and optimal control**, respectively.

Remark 9.16. When the control z is defined in an open subset Ω_0 of Ω , we say that it is a *distributed control*. In some cases, z may be defined only on $\Gamma \subseteq \partial\Omega$ and then it is called *boundary control*.

Similarly, when the cost functional (9.41) involves the observation of u in $\Omega_0 \subseteq \Omega$, we say that the observation is *distributed*. On the other hand, one may observe u or $\partial_\nu u$ on $\Gamma \subseteq \partial\Omega$. These cases correspond to *boundary observations* and the cost functional has to take an appropriate form. Some examples are given in Problems 9.6, 9.7.

The main questions to face in a control problem are:

- Establish existence and/or uniqueness of an optimal pair (u^*, z^*) .
- Derive necessary and/or sufficient optimality conditions.
- Construct algorithms for the numerical approximation of (u^*, z^*) .

9.6.2 Existence and uniqueness of an optimal pair

Given $z \in H$, we may substitute into J the unique solution $u = u[z]$ of (9.39) to get the functional

$$\tilde{J}(z) = J(u[z], z) = \frac{1}{2} \int_{\Omega_0} (u[z] - u_d)^2 d\mathbf{x} + \frac{\beta}{2} \int_{\Omega} z^2 d\mathbf{x},$$

depending only on z . Thus, our minimization problem (9.42) is reduced to find an optimal control $z^* \in H$ such that

$$\tilde{J}(z^*) = \min_{z \in H} \tilde{J}(z). \quad (9.43)$$

Once z^* is known, the optimal state is given by $u^* = u[z^*]$.

The strategy to prove existence and uniqueness of an optimal control is to use the relationship between minimization of quadratic functionals and abstract variational problems corresponding to symmetric bilinear forms, expressed in The-

orem 6.43, p. 387. The key point is to write $\tilde{J}(z)$ in the following way:

$$\tilde{J}(z) = \frac{1}{2}b(z, z) + Lz + q \quad (9.44)$$

where $q \in \mathbb{R}$ (irrelevant in the optimization) and:

- $b(z, w)$ is a bilinear form in H , *symmetric, continuous and H -coercive*.
- L is a *linear, continuous functional* in H .

Then, by Theorem 6.43, there exists a unique minimizer $z^* \in H$. Moreover z^* is the minimizer if and only if z^* satisfies the Euler equation (see (6.49))

$$\tilde{J}'(z^*)w = b(z^*, w) - Lw = 0, \quad \forall w \in H. \quad (9.45)$$

This procedure yields the following result.

Theorem 9.17. *There exists a unique optimal control $z^* \in H$. Moreover, z^* is optimal if and only if the following Euler equation holds ($u^* = u[z^*]$):*

$$\tilde{J}'(z^*)w = \int_{\Omega_0} (u^* - u_d)u[w] d\mathbf{x} + \beta \int_{\Omega} z^*w = 0, \quad \forall w \in H. \quad (9.46)$$

Proof. According to the above strategy, we write $\tilde{J}(z)$ in the form (9.44).

First note that the map $z \mapsto u[z]$ is *linear*. In fact, if $\alpha_1, \alpha_2 \in \mathbb{R}$, then $u[\alpha_1 z_1 + \alpha_2 z_2]$ is the solution of

$$\mathcal{E}u[\alpha_1 z_1 + \alpha_2 z_2] = \alpha_1 z_1 + \alpha_2 z_2 u_1.$$

Since \mathcal{E} is linear,

$$\mathcal{E}(\alpha_1 u[z_1] + \alpha_2 u[z_2]) = \alpha_1 \mathcal{E}u[z_1] + \alpha_2 \mathcal{E}u[z_2] = \alpha_1 z_1 + \alpha_2 z_2$$

and therefore, by uniqueness, $u[\alpha_1 z_1 + \alpha_2 z_2] = \alpha_1 u[z_1] + \alpha_2 u[z_2]$.

As a consequence,

$$b(z, w) = \int_{\Omega_0} u[z]u[w] d\mathbf{x} + \beta \int_{\Omega} zw \quad (9.47)$$

is a bilinear form and

$$Lw = \int_{\Omega_0} u[w]u_d d\mathbf{x} \quad (9.48)$$

is a linear functional in H .

Moreover, b is symmetric (obvious), continuous and H -coercive. In fact, from (9.40) and the Schwarz and Poincaré inequalities, we have, since $\Omega_0 \subseteq \Omega$,

$$\begin{aligned} |b(z, w)| &\leq \|u[z]\|_{L^2(\Omega_0)} \|u[w]\|_{L^2(\Omega_0)} + \beta \|z\|_H \|w\|_H \\ &\leq (C_P^2 + \beta) \|z\|_H \|w\|_H \end{aligned}$$

which gives the continuity of b . The H -coerciveness of b follows from

$$b(z, z) = \int_{\Omega_0} u^2[z] d\mathbf{x} + \beta \int_{\Omega} z^2 \geq \beta \|z\|_H^2.$$

Finally, from (9.40) and Poincarè's inequality,

$$|Lw| \leq \|u_d\|_{L^2(\Omega_0)} \|u[w]\|_{L^2(\Omega_0)} \leq C_P \|u_d\|_H \|w\|_H,$$

and we deduce that L is continuous in H .

Now, if we set: $q = \int_{\Omega_0} u_d^2 d\mathbf{x}$, it is easy to check that

$$\tilde{J}(z) = \frac{1}{2}b(z, z) - Lz + q.$$

Then, Theorem 6.43 yields existence and uniqueness of the optimal control and Euler equation (9.45) translates into (9.46) after simple computations. \square

9.6.3 Lagrange multipliers and optimality conditions

The Euler equation (9.46) gives a characterization of the optimal control z^* but it is not suitable for its computation. To obtain more manageable optimality conditions, let us change point of view by regarding the state equation $\mathcal{E}u[z] = -\Delta u + \operatorname{div}(\mathbf{b}u) = z$, with $u = 0$ on $\partial\Omega$, as a *constraint* for our minimization problem. Then, the key idea is to introduce a *multiplier* $p \in V$, to be chosen suitably later on, and write $\tilde{J}(z)$ in the augmented form

$$\frac{1}{2} \int_{\Omega_0} (u[z] - u_d)^2 d\mathbf{x} + \frac{\beta}{2} \int_{\Omega} z^2 d\mathbf{x} + \int_{\Omega} p[z - \mathcal{E}u[z]] d\mathbf{x}. \quad (9.49)$$

In fact, we have just added zero. Since clearly,

$$\tilde{L}z = \int_{\Omega} p \mathcal{E}u[z] d\mathbf{x}$$

is a continuous linear functional in H , Theorem 9.17 yields the Euler equation:

$$\tilde{J}'(z^*) w = \int_{\Omega_0} (u^* - u_d) u[w] d\mathbf{x} + \int_{\Omega} (p + \beta z^*) w d\mathbf{x} - \int_{\Omega} p \mathcal{E}u[w] d\mathbf{x} = 0 \quad (9.50)$$

for every $w \in H$. Now we integrate twice by parts the last term, recalling that $u[w] = 0$ on $\partial\Omega$. We find:

$$\begin{aligned} \int_{\Omega} p \mathcal{E}u[w] d\mathbf{x} &= \int_{\partial\Omega} p (-\partial_{\boldsymbol{\nu}} u[w] + (\mathbf{b} \cdot \boldsymbol{\nu}) u[w]) d\sigma + \int_{\Omega} (-\Delta p - \mathbf{b} \cdot \nabla p) u[w] d\mathbf{x} \\ &= - \int_{\partial\Omega} p \partial_{\boldsymbol{\nu}} u[w] d\sigma + \int_{\Omega} \mathcal{E}^* p u[w] d\mathbf{x}, \end{aligned}$$

where the operator $\mathcal{E}^* = -\Delta - \mathbf{b} \cdot \nabla$ is the formal adjoint of \mathcal{E} .

Now we choose the multiplier: let p^* be the solution of the following **adjoint** problem:

$$\begin{cases} \mathcal{E}^* p = (u^* - u_d) \chi_{\Omega_0} & \text{in } \Omega \\ p = 0 & \text{on } \partial\Omega. \end{cases} \quad (9.51)$$

Using (9.51), the Euler equation (9.50) becomes

$$\tilde{J}'(z^*) w = \int_{\Omega} (p^* + \beta z^*) w \, d\mathbf{x} = 0 \quad \forall w \in H, \quad (9.52)$$

equivalent to $p^* + \beta z^* = 0$.

Summarizing, we have proved the following result:

Theorem 9.18. *The control z^* and the state $u^* = u(z^*)$ are optimal if and only if there exists a multiplier $p^* \in V$ such that z^* , u^* and p^* satisfy the following optimality conditions:*

$$\begin{cases} -\Delta u^* + \operatorname{div}(\mathbf{b}u^*) = z^* & \text{in } \Omega, \quad u^* = 0 \text{ on } \partial\Omega \\ -\Delta p^* - \mathbf{b} \cdot \nabla p^* = (u^* - u_d) \chi_{\Omega_0} & \text{in } \Omega, \quad p^* = 0 \text{ on } \partial\Omega \\ p^* + \beta u^* = 0 & \text{a.e. in } \Omega, \text{ Euler equation.} \end{cases}$$

Remark 9.19. The optimal multiplier p^* is also called **adjoint state**. Also note that $\tilde{J}'(z^*) \in H^*$

9.6.4 An iterative algorithm

From Euler equation (9.52) and the Riesz Representation Theorem, we infer that

$$p^* + \beta z^* \text{ is the Riesz element associated with } \tilde{J}'(z^*),$$

called the **gradient of J at z^*** and denoted by the usual symbol $\nabla \tilde{J}(z^*)$ or by $\delta z(z^*, p^*)$. Thus, we have

$$\nabla \tilde{J}(z^*) = p^* + \beta z^*.$$

It turns out that $-\nabla \tilde{J}(z^*)$ plays the role of the *steepest descent direction* for \tilde{J} , as in the finite-dimensional case. This suggests an iterative procedure to compute a sequence of controls $\{z_k\}_{k \geq 0}$, convergent to the optimal one.

Select an initial control z_0 . If z_k is known ($k \geq 0$), then z_{k+1} is computed according to the following scheme.

1. Solve the state equation $a(u_k, v) = (z_k, v)_{L^2(\Omega)}$, $\forall v \in V$.
2. Knowing u_k , solve the adjoint equation

$$a^*(p_k, \varphi) = (u_k - u_d, \varphi)_{L^2(\Omega_0)}, \quad \forall \varphi \in V.$$

3. Set

$$z_{k+1} = z_k - \tau_k \nabla \tilde{J}(z_k) \quad (9.53)$$

and select the *relaxation parameter* τ_k in order to assure that

$$\tilde{J}(z_{k+1}) < \tilde{J}(z_k). \quad (9.54)$$

Clearly, (9.54) implies the convergence of the sequence $\{\tilde{J}(z_k)\}$, though in general not to zero. Concerning the choice of the relaxation parameter, there are several possibilities. For instance, if $\beta \ll 1$, we may chose

$$\tau_k = \tilde{J}(z_k) \left| \nabla \tilde{J}(z_k) \right|^{-2}.$$

With this choice, (9.53) is a Newton type method:

$$z_{k+1} = z_k - \frac{\nabla \tilde{J}(z_k)}{\left| \nabla \tilde{J}(z_k) \right|^2} \tilde{J}(z_k).$$

Problems

9.1. Derive a variational formulation of the following problem

$$\begin{cases} \Delta^2 u = f & \text{in } \Omega \\ u = 0 & \text{on } \partial\Omega \\ \Delta u + \rho \partial_\nu u = 0 & \text{on } \partial\Omega \end{cases}$$

where ρ is a positive constant and Ω is a bounded, smooth domain. Show that the right functional setting is $H^2(\Omega) \cap H_0^1(\Omega)$, i.e. the space of functions in $H^2(\Omega)$ with zero trace. Prove the well-posedness of the resulting problem.

[Hint: The variational formulation is

$$\int_{\Omega} \Delta u \cdot \Delta v \, d\mathbf{x} + \int_{\partial\Omega} \rho \partial_\nu u \cdot \partial_\nu v \, d\sigma = \int_{\Omega} f v \, d\mathbf{x}, \quad \forall v \in H^2(\Omega) \cap H_0^1(\Omega).$$

To show the well-posedness, use the inequality

$$\|\partial_\nu v\|_{L^2(\partial\Omega)} \leq C(n, \Omega) \|\Delta v\|_{L^2(\Omega)},$$

$$\forall v \in H^2(\Omega) \cap H_0^1(\Omega).$$

9.2. Let $\Omega \subset \mathbb{R}^n$ be a bounded, smooth domain and $g : \mathbb{R} \rightarrow \mathbb{R}$ be bounded and Lipschitz continuous, with Lipschitz constant K and $g(0) = 0$. Consider the following problem:

$$\begin{cases} -\Delta u + u = 1 & \text{in } \Omega \\ \partial_\nu u = g(u) & \text{on } \partial\Omega. \end{cases} \quad (9.55)$$

Give a weak formulation and show that there exists a weak solution $\bar{u} \in H^2(\Omega)$.

[Hint: Fix $w \in H^1(\Omega)$. Check that $g(w)$ has a trace in $H^{1/2}(\partial\Omega)$ and

$$\|g(w)\|_{H^{1/2}(\partial\Omega)} \leq C_0(n, \Omega) K \|w\|_{H^1(\Omega)}.$$

Let $u = T(w)$ be the unique solution of the problem

$$\begin{cases} -\Delta u + u = 1 & \text{in } \Omega \\ \partial_\nu u = g(w) & \text{on } \partial\Omega. \end{cases}$$

Use the regularity theory for elliptic equations and Rellich Theorem 7.90, p. 487, to show that $T : H^1(\Omega) \rightarrow H^1(\Omega)$ is compact and continuous. Use Schauder Theorem 6.83, p. 420, to conclude].

9.3. Let $\Omega \subset \mathbb{R}^3$ be a bounded, smooth domain and $f \in L^2(\Omega)$. Consider the following semilinear problem:

$$\begin{cases} -\Delta u + u^3 = f & \text{in } \Omega \\ u = 0 & \text{on } \partial\Omega. \end{cases} \quad (9.56)$$

- a) Give a weak formulation of the problem (recall that $H_0^1(\Omega) \hookrightarrow L^6(\Omega)$, for $n = 3$).
- b) Show that there exists a unique solution $z \in H^2(\Omega) \cap H_0^1(\Omega)$ of the problem

$$\begin{cases} -\Delta z + w^3 = f & \text{in } \Omega \\ z = 0 & \text{on } \partial\Omega, \end{cases}$$

for every fixed $w \in H_0^1(\Omega)$, and derive a stability estimate.

- c) Use the Leray-Schauder Theorem 6.84, p. 420, to show that there exists a solution $u \in H_0^1(\Omega)$ of problem (9.56).
- d) Show that the solution is unique.

9.4. Let $\Omega \subset \mathbb{R}^n$ be a bounded, smooth domain and $f, g \in C^\infty(\overline{\Omega})$. Consider the following Dirichlet problem.

$$\begin{cases} -\Delta u + u = \frac{|v|}{1 + |u| + |v|} + f & \text{in } \Omega \\ -\Delta v + v = \frac{|u|}{1 + |u| + |v|} + g & \text{in } \Omega \\ u = v = 0 & \text{on } \partial\Omega. \end{cases} \quad (9.57)$$

Give a weak formulation of the problem and prove that:

- a) (9.57) has a weak solution $(u, v) \in H_0^1(\Omega) \times H_0^1(\Omega)$.
- b) If (u, v) is a weak solution of (9.57) then $(u, v) \in H^3(\Omega) \times H^3(\Omega)$.
- c) If (u, v) is a weak solution of (9.57) and $f \geq 0, g \geq 0$ in Ω then $u \geq 0, v \geq 0$ in Ω . Moreover $(u, v) \in C^\infty(\overline{\Omega}) \times C^\infty(\overline{\Omega})$, that is (u, v) is a classical solution of (9.57).
- d) If (u, v) is a weak solution of (9.57) and $0 \leq f \leq g$ then $u \leq v$ in Ω . In particular, if $f \equiv g \geq 0$ then $u = v$ and the solution of problem (9.57) is unique.

[Hint: a) Let $X = L^2(\Omega) \times L^2(\Omega)$ with the norm

$$\|(u, v)\|_X^2 = \|u\|_{L^2(\Omega)}^2 + \|v\|_{L^2(\Omega)}^2.$$

For fixed $(u_0, v_0) \in X$, solve the linear uncoupled problem

$$\begin{cases} -\Delta u + u = \frac{|v_0|}{1 + |u_0| + |v_0|} + f & \text{in } \Omega \\ -\Delta v + v = \frac{|u_0|}{1 + |u_0| + |v_0|} + g & \text{in } \Omega \\ u = v = 0 & \text{on } \partial\Omega. \end{cases} \quad (9.58)$$

Let $T : X \rightarrow X$ the solution map so defined. Show that T is compact and continuous. Use Schauder Theorem 6.83, p. 420, to prove that T has a fixed point.

- b) Use the regularity theory for elliptic equations.
- c) Use maximum principle, get rid of the absolute value signs and use bootstrapping].

9.5. Let $\Omega \subset \mathbb{R}^n$ be a bounded, smooth domain. Consider the system:

$$\begin{cases} -\Delta\varphi + u\varphi = 1 & \text{in } \Omega \\ -\operatorname{div}(e^\varphi \nabla u) = 0 & \text{in } \Omega \\ \varphi = 0, u = g & \text{on } \partial\Omega \end{cases} \quad (9.59)$$

where $g \in H^{1/2}(\partial\Omega)$ is the trace of $\tilde{g} \in H^1(\Omega)$. Assume that $0 < \alpha \leq g \leq \beta < \infty$. on $\partial\Omega$. Define

$$X = \{u \in L^2(\Omega) : \alpha \leq u \leq \beta \text{ a.e. in } \Omega\}.$$

- a) Give a weak formulation of problem (9.59).
- b) Prove the existence of a weak solution $(u, \varphi) \in H^1(\Omega) \times H_0^1(\Omega)$, by following the steps below.

1. Fix $\tilde{u} \in X$ and construct a map $T_1 : X \rightarrow H_0^1(\Omega)$, where $\varphi = T_1(\tilde{u})$ is the unique solution of the problem

$$-\Delta\varphi + \tilde{u}\varphi = 1 \text{ in } \Omega, \quad \varphi = 0 \text{ on } \partial\Omega. \quad (9.60)$$

In particular, derive the estimate $\|\varphi\|_{H_0^1(\Omega)} \leq C_1$, with C_1 independent of \tilde{u} .

2. Using that

$$v = \varphi^- = \max\{-\varphi, 0\}$$

is an admissible test function in the weak formulation of (9.60), show that $\varphi \geq 0$ a.e. in Ω . Similarly, using

$$v = (\varphi - 1/\alpha)^+ = \max\{\varphi - 1/\alpha, 0\},$$

show that $\varphi \leq 1/\alpha$ a.e. in Ω .

3. Let $\varphi = T_1(\tilde{u})$ be the the unique solution of problem (9.60). Construct a map $T_2 : H_0^1(\Omega) \rightarrow X$, defining $u = T_2(\varphi)$ as the solution of the problem.

$$-\operatorname{div}(e^\varphi \nabla u) = 0 \text{ in } \Omega, \quad u = g \text{ on } \partial\Omega. \quad (9.61)$$

In particular, prove the following estimates: $\|\nabla u\|_{L^2(\Omega)} \leq C_2$, with C_2 independent of u and $\alpha \leq u \leq \beta$ a.e. in Ω .

4. Prove that the operator

$$T = T_2 \circ T_1 : X \rightarrow X$$

is *continuous in* $L^2(\Omega)$, i.e. if $\{\tilde{u}_m\} \subset X$ and $\tilde{u}_m \rightarrow \tilde{u}$ in $L^2(\Omega)$, then $T(\tilde{u}_m) \rightarrow T(\tilde{u})$ in $L^2(\Omega)$.

5. Show that $\overline{T(X)}$ is compact in X . Deduce that T has a fixed point u_f , that is

$$u_f = T_2(T_1(u_f)).$$

Check that if $\varphi_f = T_1(u_f)$, then (u_f, φ_f) is a solution in $H^1(\Omega) \times H_0^1(\Omega)$ of problem (9.59).

9.6. Distributed observation and control, Neumann conditions. Let $\Omega \subset \mathbb{R}^n$ be a bounded, smooth domain and Ω_0 an *open* (nonempty) subset of Ω . Set

$$V = H^1(\Omega), \quad H = L^2(\Omega)$$

and consider the following control problem:

Minimize the cost functional

$$J(u, z) = \frac{1}{2} \int_{\Omega_0} (u - u_d)^2 d\mathbf{x} + \frac{1}{2} \int_{\Omega} z^2 d\mathbf{x}$$

over $(u, z) \in V \times H$, with state system

$$\begin{cases} \mathcal{E}u = -\Delta u + a_0 u = z & \text{in } \Omega \\ \partial_{\nu} u = g & \text{on } \partial\Omega \end{cases} \quad (9.62)$$

where a_0 is a positive constant, $g \in L^2(\partial\Omega)$ and $z \in H$.

- a) Show that there exists a unique minimizer.
- b) Write the optimality conditions: adjoint problem and Euler equations.

[Hint: a) Follow the proof of Theorem 9.17, observing that, if $u[z]$ is the solution of (9.62) the map

$$z \longmapsto u[z] - u[0]$$

is linear. Then write

$$\tilde{J}(z) = \frac{1}{2} \int_{\Omega_0} (u[z] - u[0] + u[0] - u_d)^2 d\mathbf{x} + \frac{1}{2} \int_{\Omega} z^2 d\mathbf{x}$$

and adjust the bilinear form (9.47) accordingly.

- b) *Answer:* The adjoint problem is ($\mathcal{E} = \mathcal{E}^*$)

$$\begin{cases} -\Delta p + a_0 p = (u - z_d)\chi_{\Omega_0} & \text{in } \Omega \\ \partial_{\nu} p = 0 & \text{on } \partial\Omega. \end{cases}$$

Where χ_{Ω_0} is the characteristic function of Ω_0 . The Euler equation is: $p + z = 0$ in $L^2(\Omega)$.

9.7. Boundary observation and distributed control, Dirichlet conditions. Let $\Omega \subset \mathbb{R}^n$ be a bounded, smooth domain. Consider the following control problem:

Minimize the cost functional

$$J(u, z) = \frac{1}{2} \int_{\partial\Omega} (\partial_\nu u - u_d)^2 d\sigma + \frac{\beta}{2} \int_{\Omega} z^2 d\mathbf{x}$$

over $(u, z) \in (H_0^1(\Omega) \cap H^2(\Omega)) \times L^2(\Omega)$, with state system

$$\begin{cases} -\Delta u + \mathbf{c} \cdot \nabla u = f + z, & \text{in } \Omega \\ u = 0 & \text{on } \partial\Omega \end{cases}$$

where $u_d \in H^1(\Omega)$ \mathbf{c} is a constant vector and $f \in L^2(\Omega)$.

- a) Show that, by elliptic regularity, $J(u, z)$ is well defined and that there exists a unique minimizer.
- b) Write the optimality conditions: adjoint problem and Euler equations.

Chapter 10

Weak Formulation of Evolution Problems

10.1 Parabolic Equations

In Chap. 2 we have considered the diffusion equation and some of its generalizations, as in the reaction-diffusion model (Sect. 2.5) or in the Black-Scholes model (Sect. 2.9). This kind of equations belongs to the class of *parabolic equations*, that we have already classified in spatial dimension 1 (Sect. 2.4.1) and that we are going to define in a more general setting.

Let $\Omega \subset \mathbb{R}^n$ be a *bounded* domain, $T > 0$ and consider the space-time cylinder (see Fig. 10.1) $Q_T = \Omega \times (0, T)$. Let $\mathbf{A} = \mathbf{A}(\mathbf{x}, t)$ be a square matrix of order n , $\mathbf{b} = \mathbf{b}(\mathbf{x}, t)$, $\mathbf{c} = \mathbf{c}(\mathbf{x}, t)$ vectors in \mathbb{R}^n , $a = a(\mathbf{x}, t)$ and $f = f(\mathbf{x}, t)$ real functions. Equations in divergence form of the type

$$u_t - \operatorname{div}(\mathbf{A} \nabla u - \mathbf{b} u) + \mathbf{c} \cdot \nabla u + au = f \quad (10.1)$$

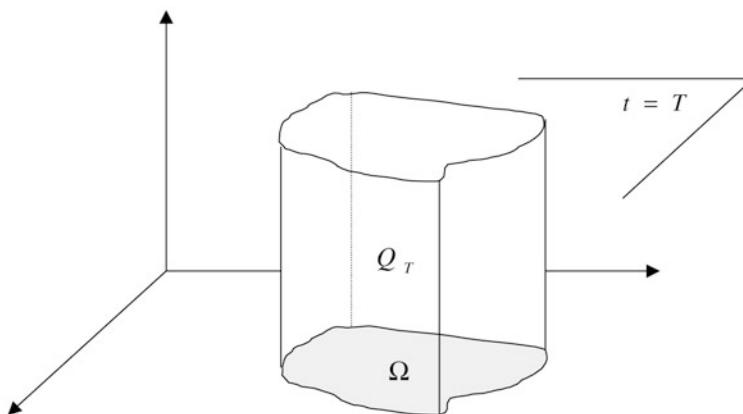


Fig. 10.1 Space-time cylinder

or in *nondivergence form* of the type

$$u_t - \text{Tr}(\mathbf{A} D^2 u) + \mathbf{b} \cdot \nabla u + au = f \quad (10.2)$$

are called **parabolic** in Q_T if

$$A(\mathbf{x}, t) \boldsymbol{\xi} \cdot \boldsymbol{\xi} > 0, \quad \forall (\mathbf{x}, t) \in Q_T, \forall \boldsymbol{\xi} \in \mathbb{R}^n, \boldsymbol{\xi} \neq \mathbf{0}.$$

For parabolic equations we may repeat the considerations concerning elliptic equations made in Sects. 8.1 and 8.2. Also in this case, different notions of solutions may be given, with the obvious corrections due to the evolutionary nature of (10.1) and (10.2). For identical reasons, we develop the basic theory for divergence form equations. Thus, we consider parabolic operators of the type¹

$$\mathcal{P}u = u_t + \mathcal{E}u \equiv u_t - \text{div}(\mathbf{A}(\mathbf{x}, t) \nabla u) + \mathbf{c}(\mathbf{x}, t) \cdot \nabla u + a(\mathbf{x}, t)u.$$

The matrix $\mathbf{A} = (a_{i,j}(\mathbf{x}, t))$ encodes the anisotropy of the medium with respect to diffusion. For instance (see Subsect. 2.6.2) a matrix of the type

$$\begin{pmatrix} \alpha & 0 & 0 \\ 0 & \varepsilon & 0 \\ 0 & 0 & \varepsilon \end{pmatrix}$$

with $\alpha \gg \varepsilon > 0$, denotes higher propensity of the medium towards diffusion along the x_1 -axis, than along the other directions. As in the stationary case, it is important to compare the effects of the drift, reaction and diffusion terms, to control the stability of numerical algorithms. We make the following assumptions:

(a) \mathcal{P} is *uniformly parabolic*: there are positive numbers ν and M such that:

$$\nu |\boldsymbol{\xi}|^2 \leq \sum_{i,j=1}^n a_{ij}(\mathbf{x}, t) \xi_i \xi_j \quad \text{and} \quad |a_{ij}(\mathbf{x}, t)| \leq M, \quad \forall \boldsymbol{\xi} \in \mathbb{R}^n, \text{ a.e. in } Q_T. \quad (10.3)$$

(b) The coefficients \mathbf{c} and a are bounded (i.e. all belong to $L^\infty(Q_T)$), with

$$|\mathbf{c}| \leq \gamma_0, \quad |a| \leq \alpha_0, \quad \text{a.e. in } Q_T. \quad (10.4)$$

We consider initial-boundary value problems of the form:

$$\begin{cases} u_t + \mathcal{E}u = f & \text{in } Q_T \\ u(\mathbf{x}, 0) = g(\mathbf{x}) & \text{in } \Omega \\ \mathcal{B}u(\boldsymbol{\sigma}, t) = 0 & \text{on } S_T, \end{cases} \quad (10.5)$$

where $S_T = \partial\Omega \times (0, T)$ is the lateral part of Q_T and $\mathcal{B}u = 0$ stands for one of the usual *homogeneous* boundary conditions. For instance, $\mathcal{B}u = \mathbf{A} \nabla u \cdot \boldsymbol{\nu} + hu$ for the Neumann/Robin condition, $\mathcal{B}u = u$ for the Dirichlet condition.

We start with the Cauchy-Dirichlet problem for the heat equation to introduce a possible *weak formulation*. This approach requires the use of integrals for functions with values in a Hilbert space and of Sobolev spaces involving time. A brief account of these notions is presented in Sect. 7.11. Then we recast into a

¹ For simplicity we assume $\mathbf{b} = \mathbf{0}$.

general abstract framework all the most common initial-boundary value problems for divergence form equations. As in the elliptic case, the main tool for solving the abstract problem is provided by a sort of evolution version of the Lax Milgram Theorem (Theorem 10.6).

10.2 The Cauchy-Dirichlet Problem for the Heat Equation

Suppose we are given the problem

$$\begin{cases} u_t - \Delta u = f & \text{in } Q_T \\ u(\sigma, t) = 0 & \text{on } S_T \\ u(\mathbf{x}, 0) = g(\mathbf{x}) & \text{in } \Omega. \end{cases} \quad (10.6)$$

Here $\Omega \subset \mathbb{R}^n$ is a *bounded* domain. We want to find a weak formulation. Let us proceed formally, assuming that everything is smooth. Mimicking what we did several times in Chap. 8, we multiply the diffusion equation by a smooth function $v = v(\mathbf{x}, t)$, vanishing on S_T , and integrate over Q_T . We find

$$\int_{Q_T} u_t(\mathbf{x}, t) v(\mathbf{x}, t) d\mathbf{x} dt - \int_{Q_T} \Delta u(\mathbf{x}, t) v(\mathbf{x}, t) d\mathbf{x} dt = \int_{Q_T} f(\mathbf{x}, t) v(\mathbf{x}, t) d\mathbf{x} dt.$$

Integrating by parts the second term with respect to \mathbf{x} , we get, since $v = 0$ on S_T :

$$\int_{Q_T} u_t(\mathbf{x}, t) v(\mathbf{x}, t) d\mathbf{x} dt + \int_{Q_T} \nabla u(\mathbf{x}, t) \cdot \nabla v(\mathbf{x}, t) d\mathbf{x} dt = \int_{Q_T} f(\mathbf{x}, t) v(\mathbf{x}, t) d\mathbf{x} dt. \quad (10.7)$$

This looks like what we did for elliptic equations, except for the presence of u_t . Moreover, here we will have somehow to take into account the initial condition. Which could be a correct functional setting?

First of all, since we are dealing with evolution equations, it is convenient to adopt the point of view of Sect. 7.11, and consider $u = u(\mathbf{x}, t)$ as a function of t with values into a suitable Hilbert space V , that is $u : [0, T] \rightarrow V$.

When we adopt this convention, we write $u(t)$ instead of $u(\mathbf{x}, t)$ and \dot{u} instead of u_t . Accordingly, we write $f(t)$ instead of $f(\mathbf{x}, t)$. Now, the homogeneous Dirichlet condition, i.e. $u(t) = 0$ on $\partial\Omega$ for $t \in [0, T]$, suggests that the natural space for u is $L^2(0, T; V)$, where $V = H_0^1(\Omega)$ equipped with the inner product

$$(w, z)_{H_0^1(\Omega)} = (\nabla w, \nabla z)_{L^2(\Omega; \mathbb{R}^n)}$$

and corresponding norm $\|\nabla w\|_{L^2(\Omega; \mathbb{R}^n)}$. In particular, by Fubini's Theorem, it follows that $u(t) \in V$ for a.e. $t \in (0, T)$.

To shorten the formulas, we **introduce the symbols**

$$(\cdot, \cdot)_0 \quad \text{and} \quad \|\cdot\|_0$$

for the inner product and norm in $L^2(\Omega)$ and $L^2(\Omega; \mathbb{R}^n)$, respectively. With these notations, (10.7) becomes, separating space from time:

$$\int_0^T (\dot{u}(t), v(t))_0 dt + \int_0^T (\nabla u(t), \nabla v(t))_0 dt = \int_0^T (f(t), v(t))_0 dt. \quad (10.8)$$

By looking at the first integral, it would seem appropriate to require that $\dot{u} \in L^2(0, T; L^2(\Omega))$. This, however, is not coherent with the choice $u(t) \in V$, a.e. $t \in (0, T)$, since we have $\Delta u(t) \in V^* = H^{-1}(\Omega)$ and

$$\dot{u}(t) = \Delta u(t) + f(t) \quad (10.9)$$

from the diffusion equation. Thus, we deduce that the natural space for \dot{u} is $L^2(0, T; V^*)$. Consequently, the first term in (10.8) has to be interpreted as

$$\langle \dot{u}(t), v(t) \rangle_*,$$

where $\langle \cdot, \cdot \rangle_*$ denotes the duality between V^* and V .

Similarly, a reasonable assumption on f is $f \in L^2(0, T; V^*)$ and instead of $(f(t), v(t))_0$ we should write $\langle f(t), v(t) \rangle_*$. Now, from Theorem 7.104, p. 498, we know that

$$u \in C([0, T]; L^2(\Omega))$$

so that, if we choose $g \in L^2(\Omega)$, the initial condition $u(0) = g$ makes perfect sense and also

$$\|u(t) - g\|_0 \rightarrow 0 \text{ as } t \rightarrow 0.$$

The above arguments motivate our *first weak formulation*.

Definition 10.1. Let $f \in L^2(0, T; V^*)$ and $g \in L^2(\Omega)$. We say that $u \in L^2(0, T; V)$, with $\dot{u} \in L^2(0, T; V^*)$, is a weak solution of problem (10.6) if:

1. For every $v \in L^2(0, T; V)$,

$$\int_0^T \langle \dot{u}(t), v(t) \rangle_* dt + \int_0^T (\nabla u(t), \nabla v(t))_0 dt = \int_0^T \langle f(t), v(t) \rangle_* dt. \quad (10.10)$$

2. $u(0) = g$.

It is not a difficult task to check that, if u , g and f are smooth functions and u is a weak solution, then u solves problem (10.6) in a classical sense. Thus, under regularity conditions, the weak and the classical formulation are equivalent. The Definition 10.1 is rather satisfactory and works well in many kind of applied situations, like for instance in dealing with so called *optimal control* problems. However, for the numerical treatment of problem (10.6) it is often more convenient an alternative formulation that we give below.

Definition 10.2. A function $u \in L^2(0, T; V)$ is called weak solution of problem (10.6) if $\dot{u} \in L^2(0, T; V^*)$ and:

1. For every $w \in V$ and a. e. $t \in (0, T)$,

$$\langle \dot{u}(t), w \rangle_* + (\nabla u(t), \nabla w)_0 = \langle f(t), w \rangle_*. \quad (10.11)$$

2. $u(0) = g$.

Remark 10.3. Equation (10.11) may be interpreted in the sense of distributions. To see this, observe that, for every $w \in V$, the real function $t \mapsto z(t) = \langle \dot{u}(t), w \rangle_*$ is a distribution in $\mathcal{D}'(0, T)$ and

$$\langle \dot{u}(t), w \rangle_* = \frac{d}{dt} (u(t), w)_0 \quad \text{in } \mathcal{D}'(0, T). \quad (10.12)$$

This means that, for every $\varphi \in \mathcal{D}(0, T)$, we have

$$\int_0^T \langle \dot{u}(t), w \rangle_* \varphi(t) dt = - \int_0^T (u(t), w)_0 \dot{\varphi}(t) dt.$$

In fact, since $|\langle \dot{u}(t), w \rangle_{V^*}| \leq \|\dot{u}(t)\|_* \|w\|_V$, it follows that $z \in L^1_{loc}(0, T) \subset \mathcal{D}'(0, T)$. Moreover, by Bochner's Theorem 7.100, p. 495, and the definition of \dot{u} , we may write

$$\int_0^T \langle \dot{u}(t), w \rangle_* \varphi(t) dt = \left\langle \int_0^T \dot{u}(t) \varphi(t) dt, w \right\rangle_* = \left\langle - \int_0^T u(t) \dot{\varphi}(t) dt, w \right\rangle_*.$$

On the other hand, $\int_0^T u(t) \dot{\varphi}(t) dt \in V$ so that²

$$\left\langle - \int_0^T u(t) \dot{\varphi}(t) dt, w \right\rangle_* = \left\langle - \int_0^T u(t) \dot{\varphi}(t) dt, w \right\rangle_0 = - \int_0^T (u(t), w)_0 \dot{\varphi}(t) dt$$

and (10.12) is true. As a consequence, equation (10.11) may be written in the form

$$\frac{d}{dt} (u(t), w)_0 + a(u(t), w) = (f, w)_0, \quad (10.13)$$

in the sense of distributions in $\mathcal{D}'(0, T)$, for all $w \in V$.

We now prove the equivalence of the two Definitions 10.1 and 10.2.

Theorem 10.4. The two Definitions 10.1 and 10.2 are equivalent.

Proof. Assume that u is a weak solution according to Definition 10.1. We want to show that (10.11) holds for all $v \in V$ and a.e. $t \in (0, T)$. Suppose that this is not true. Then

² Recall from Sect. 6.8 that if $u \in H$ and $w \in V$, $\langle u, w \rangle_* = (u, w)_0$.

there exists $w \in V$ and a set $E \subset (0, T)$ of positive measure, such that

$$\langle \dot{u}(t), w \rangle_* + (\nabla u(t), \nabla w)_0 - \langle f(t), w \rangle_* > 0 \quad \text{for every } t \in E.$$

Then, (10.10) does not hold for $v(t) = \chi_E(t)w$. Contradiction.

Assume now that u is a weak solution according to Definition (10.2). Let $\{w_j\}_{j \geq 1}$ be a countable orthonormal basis in V . For each $w = w_j$, (10.11) holds outside a set F_j of measure zero. Set $E_0 = \bigcup_{j \geq 1} F_j$. Then E_0 has measure zero and (10.11) holds for all w_j , $j \geq 1$, and all $t \in (0, T) \setminus E_0$. Take any $v \in L^2(0, T; V)$ and let $c_j(t) = (v(t), w_j)_V$,

$$v_N(t) \equiv \sum_1^N c_j(t) w_j.$$

Take $w = w_j$ in (10.11), multiply by $c_j(t)$ and sum over j from 1 to N . We find

$$\langle \dot{u}(t), v_N(t) \rangle_* + (\nabla w(t), v_N(t))_0 = \langle f(t), v_N(t) \rangle_*, \quad \forall t \in (0, T) \setminus E_0. \quad (10.14)$$

Since $v_N(t) \rightarrow v(t)$ in V for a.e. $t \in (0, T)$, it follows that

$$\langle \dot{u}(t), v(t) \rangle_* + (\nabla w(t), v(t))_0 = \langle f(t), v(t) \rangle_*, \quad \text{a.e. in } (0, T). \quad (10.15)$$

Integrating (10.15) with respect to t over $(0, T)$ we obtain (10.1). \square

10.3 Abstract Parabolic Problems

10.3.1 Formulation

In this section we formulate an Abstract Parabolic Problem (\mathcal{APP} in the sequel) and give a notion of weak solution. All the most common linear initial-boundary value problems for a wide class of operators can be recast as an \mathcal{APP} . In this general framework we provide an existence, uniqueness and stability result that plays the same role of the Lax-Milgram Theorem for the elliptic case.

Guided by the example in Sect. 10.2, we first single out the main ingredients in the abstract formulation.

- *Functional setting.* The functional setting is constituted by a Hilbert triplet $\{V, H, V^*\}$, where V, H are *separable* Hilbert spaces. Usually $H = L^2(\Omega)$ and $H_0^1(\Omega) \subseteq V \subseteq H^1(\Omega)$. As usual, we denote by $\langle \cdot, \cdot \rangle_*$ the duality between V^* and V .

The choice of V depends on the type of boundary condition we are dealing with. The familiar choices are $V = H_0^1(\Omega)$ for the homogeneous Dirichlet condition, $V = H^1(\Omega)$ for the Neumann or Robin condition, $V = H_{0,\Gamma_D}^1(\Omega)$ in the case of mixed Dirichlet/Neumann or Robin conditions.

- *Spaces for data and solution.* We assume that the initial data g and the source term f belong to H and $L^2(0, T; V^*)$, respectively. Our solution will belong to the Hilbert space

$$H^1(0, T; V, V^*) = \{u : u \in L^2(0, T; V), \dot{u} \in L^2(0, T; V^*)\}$$

with the inner product

$$(u, v)_{H^1(0, T; V, V^*)} = \int_0^T (u(t), v(t))_V dt + \int_0^T (\dot{u}(t), \dot{v}(t))_{V^*} dt$$

and norm

$$\|u\|_{H^1(0, T; V, V^*)}^2 = \int_0^T \|u(t)\|_V^2 dt + \int_0^T \|\dot{u}(t)\|_{V^*}^2 dt.$$

- *Bilinear form.* We are given a bilinear form, possibly time dependent,

$$B(w, z; t) : V \times V \rightarrow \mathbb{R} \text{ for a.e. } t \in (0, T).$$

We shall adopt the following **assumptions**.

- (i) *Continuity:* there exists $\mathcal{M} = \mathcal{M}(T) > 0$, such that

$$|B(w, z; t)| \leq \mathcal{M} \|w\|_V \|z\|_V, \quad \forall w, z \in V, \text{ a.e. in } (0, T). \quad (10.16)$$

- (ii) *$V - H$ weak coercivity:* there exist positive numbers λ and α such that

$$B(w, w; t) + \lambda \|w\|_H^2 \geq \alpha \|w\|_V^2, \quad \forall w \in V, \text{ a.e. in } (0, T). \quad (10.17)$$

- (iii) *t-measurability:* the map $t \mapsto B(w, z; t)$ is measurable for all $w, z \in V$.

Remark 10.5. Note that, under the assumption (i), for every $u, v \in L^2(0, T; V)$, the function $t \mapsto B(u(t), v(t); t)$ belongs to $L^1(0, T)$.

Our **Abstract Parabolic Problem** can be formulated in the following way.

Find $u \in H^1(0, T; V, V^*)$ such that:

\mathcal{AP}_1 . For all $v \in L^2(0, T; V)$,

$$\int_0^T \langle \dot{u}(t), v(t) \rangle_* dt + \int_0^T B(u(t), v(t); t) dt = \int_0^T \langle f(t), v(t) \rangle_* dt. \quad (10.18)$$

\mathcal{AP}_2 . $u(0) = g$.

From Theorem 7.104, p. 498 we know that

$$H^1(0, T; V, V^*) \hookrightarrow C([0, T]; H).$$

Therefore condition \mathcal{AP}_2 makes perfectly sense and moreover $\|u(t) - g\|_H \rightarrow 0$ as $t \rightarrow 0$.

Condition \mathcal{AP}_1 can be stated in the following equivalent form:

\mathcal{AP}'_1 . For all $w \in V$ and a.e. $t \in (0, T)$,

$$\langle \dot{u}(t), w \rangle_* + B(u(t), w; t) = \langle f(t), w \rangle_*. \quad (10.19)$$

The equivalence of the two conditions \mathcal{AP}_1 and \mathcal{AP}'_1 follows from the separability of V and Theorem 6.18, p. 369, by repeating almost verbatim the proof of Theorem 10.4. In particular, from this proof it follows that another equivalent formulation is:

\mathcal{AP}''_1 . For all $v \in L^2(0, T; V)$,

$$\langle \dot{u}(t), v(t) \rangle_* + B(u(t), v(t); t) = \langle f(t), v(t) \rangle_* \quad \text{a.e. on } (0, T). \quad (10.20)$$

Our purpose is to prove the following theorem.

Theorem 10.6. *The Abstract Parabolic Problem has a unique solution u . Moreover, the following energy estimates hold:*

$$\|u(t)\|_H^2, \alpha \int_0^t \|u(s)\|_V^2 ds \leq e^{2\lambda t} \left\{ \|g\|_H^2 + \frac{1}{\alpha} \int_0^t \|f(s)\|_{V^*}^2 ds \right\} \quad (10.21)$$

and

$$\int_0^t \|\dot{u}(s)\|_{V^*}^2 ds \leq \left\{ C_0 \|g\|_H^2 + C_1 \int_0^t \|f(s)\|_{V^*}^2 ds \right\} \quad (10.22)$$

for every $t \in [0, T]$, with $C_0 = 2\alpha^{-1}\mathcal{M}^2 e^{2\lambda t}$, $C_1 = 2\alpha^{-2}\mathcal{M}^2 e^{2\lambda t} + 2$.

Observe how the constants in the estimates (10.21) and (10.22) deteriorates as t becomes large, unless the bilinear form is coercive. Indeed, in this case $\lambda = 0$ and no time dependence appears in those constants. This turns out very useful for studying asymptotic properties of the solution as $t \rightarrow +\infty$.

We split the proof of Theorem 10.6 into the two Lemmas 10.7 and 10.10. In the first one we prove that, assuming that u is a solution of \mathcal{APP} , the *energy estimates* (10.21) and (10.22) hold. These estimates provide a control of the relevant norms

$$\|u\|_{H^1(0, T; V, V^*)} \quad \text{and} \quad \|u\|_{C([0, T]; H)}$$

of the solution, in terms of the norms $\|f\|_{L^2(0, T; V^*)}$ and $\|g\|_H$ of the data. Uniqueness and stability follow immediately.

In Lemma 10.10 we prove the existence of the solution. To this purpose, we approximate our \mathcal{APP} by means of a sequence of Cauchy problems for suitable ordinary differential equations in finite dimensional spaces. Their solutions are called *Faedo-Galerkin approximations*. The task is to show that we can extract from the sequence of Faedo-Galerkin approximations a subsequence converging in some sense to a solution of our \mathcal{APP} . This is a typical compactness problem in Hilbert spaces. The key tool is Theorem 6.57, p. 395: *in a Hilbert space any bounded sequence has a weakly convergent subsequence*. The required bounds are once again provided by the energy estimates.

10.3.2 Energy estimates. Uniqueness and stability.

We prove the following lemma.

Lemma 10.7. *Let u be a solution of the Abstract Parabolic Problem. Then the energy estimates (10.21) and (10.22) hold for u . In particular the Abstract Parabolic Problem has at most one weak solution.*

For the proof we shall use the following elementary but very useful lemma, known as *Gronwall's Lemma*.

Lemma 10.8. *Let Ψ, G be continuous in $[0, T]$, with G nondecreasing and $\gamma > 0$. If*

$$\Psi(t) \leq G(t) + \gamma \int_0^t \Psi(s) ds, \quad \text{for all } t \in [0, T], \quad (10.23)$$

then

$$\Psi(t) \leq G(t) e^{\gamma t}, \quad \text{for all } t \in [0, T].$$

Proof. Let

$$R(s) = \gamma \int_0^s \Psi(r) dr.$$

Then, for all $s \in [0, T]$,

$$R'(s) = \gamma \Psi(s) \leq \gamma \left[G(s) + \gamma \int_0^s \Psi(r) dr \right] = \gamma [G(s) + R(s)].$$

Multiplying both sides by $\exp(-\gamma s)$, we can write the above inequality in the form

$$\frac{d}{ds} [R(s) \exp(-\gamma s)] \leq \gamma G(s) \exp(-\gamma s).$$

An integration over $(0, t)$ yields ($R(0) = 0$):

$$R(t) \leq \gamma \int_0^t G(s) e^{\gamma(t-s)} ds \leq G(t) (e^{\gamma t} - 1), \quad \text{for all } t \in [0, T].$$

and from (10.23) the conclusion follows easily. \square

Proof of Lemma 10.7. Let $u = v$ in (10.20). We get,

$$\langle \dot{u}(s), u(s) \rangle_* + B(u(s), u(s); s) = \langle f(s), u(s) \rangle_*, \quad \text{a.e. in } (0, T). \quad (10.24)$$

Now, from Remark 7.105, p. 498, we have

$$\langle \dot{u}(s), u(s) \rangle_* = \frac{1}{2} \frac{d}{ds} \|u(s)\|_H^2.$$

Also,³

$$|\langle f(s), u(s) \rangle_*| \leq \|f(s)\|_{V^*} \|u(s)\|_V \leq \frac{1}{2\alpha} \|f(s)\|_{V^*}^2 + \frac{\alpha}{2} \|u(s)\|_V^2.$$

³ Recall the elementary inequality $2ab \leq a^2/\alpha + \alpha b^2$, $\forall a, b \in \mathbb{R}, \forall \alpha > 0$.

Thus, using assumption (10.17) on the bilinear form B , we can write, after simple rearrangements,

$$\frac{1}{2} \frac{d}{ds} \|u(s)\|_H^2 + \frac{\alpha}{2} \|u(s)\|_V^2 \leq \frac{1}{2\alpha} \|f(s)\|_{V^*}^2 + \lambda \|u(s)\|_H^2.$$

Integrating over $(0, t)$, and using $u(0) = g$, we obtain, for every $t \in (0, T)$:

$$\|u(t)\|_H^2 + \alpha \int_0^t \|u(s)\|_V^2 ds \leq \|g\|_H^2 + \frac{1}{\alpha} \int_0^t \|f(s)\|_{V^*}^2 ds + 2\lambda \int_0^t \|u(s)\|_H^2 ds. \quad (10.25)$$

In particular, setting

$$\Psi(t) = \|u(t)\|_H^2, \quad G(t) = \|g\|_H^2 + \frac{1}{\alpha} \int_0^t \|f(s)\|_{V^*}^2 ds, \quad \gamma = 2\lambda,$$

(10.25) implies

$$\Psi(t) \leq G(t) + \gamma \int_0^t \Psi(s) ds.$$

Then, the Gronwall Lemma gives $\Psi(t) \leq G(t) e^{\gamma t}$ in $[0, T]$, that is

$$\|u(t)\|_H^2 \leq e^{2\lambda t} \left\{ \|g\|_H^2 + \frac{1}{\alpha} \int_0^t \|f(s)\|_{V^*}^2 ds \right\}. \quad (10.26)$$

Using once more (10.25) and (10.26) we get

$$\begin{aligned} \alpha \int_0^t \|u(s)\|_V^2 ds &\leq \left(1 + \int_0^t 2\lambda e^{2\lambda s} ds \right) \left\{ \|g\|_H^2 + \frac{1}{\alpha} \int_0^t \|f(s)\|_{V^*}^2 ds \right\} \\ &= e^{2\lambda t} \left\{ \|g\|_H^2 + \frac{1}{\alpha} \int_0^t \|f(s)\|_{V^*}^2 ds \right\}. \end{aligned} \quad (10.27)$$

We still have to estimate $\|\dot{u}\|_{L^2(0, t; V^*)} = \int_0^t \|\dot{u}(s)\|_{V^*}^2 ds$. Write (10.19) in the form

$$\langle \dot{u}(s), w \rangle_* = -B(u(s), w; s) + \langle f(s), w \rangle_*$$

and observe that, using assumption (10.16) on B , for all $w \in V$ and a.e. $s \in (0, t)$, we can write:

$$|\langle \dot{u}(s), w \rangle_*| \leq \{ \mathcal{M} \|u(s)\|_V + \|f(s)\|_{V^*} \} \|w\|_V$$

and therefore, by the definition of norm in V^* , we have

$$\|\dot{u}(s)\|_{V^*} \leq \mathcal{M} \|u(s)\|_V + \|f(s)\|_{V^*}.$$

Squaring and integrating over $(0, t)$ we obtain

$$\int_0^t \|\dot{u}(s)\|_{V^*}^2 ds \leq 2\mathcal{M}^2 \int_0^t \|u(s)\|_V^2 ds + 2 \int_0^t \|f(s)\|_{V^*}^2 ds.$$

Finally, using (10.27) we infer

$$\int_0^t \|\dot{u}(s)\|_{V^*}^2 ds \leq \frac{2\mathcal{M}^2 e^{2\lambda t}}{\alpha} \left\{ \|g\|_H^2 + \frac{1}{\alpha} \int_0^t \|f(s)\|_{V^*}^2 ds \right\} + 2 \int_0^t \|f(s)\|_{V^*}^2 ds$$

from which (10.22) follows. \square

10.3.3 The Faedo-Galerkin approximations.

As we have mentioned at the beginning of this section, to prove the existence of a solution, we approximate our \mathcal{APP} by a sequence of Cauchy problems for suitable systems of ODEs in finite dimensional spaces. To realize this program we proceed through the following steps.

- *Selection of a countable basis for V .* Since V is separable, by Proposition 6.18, p. 369, we can select a countable basis $\{w_k\}_{k=1}^{\infty}$ (e.g. an orthonormal basis). In particular, the finite dimensional subspaces $V_m = \text{span}\{w_1, \dots, w_m\}$ satisfy the conditions

$$V_m \subset V_{m+1}, \quad \overline{\cup V_m} = V. \quad (10.28)$$

Since V is dense in H , there exists a sequence $g_m = \sum_{j=1}^m \hat{g}_{jm} w_j \in V_m$ such that $g_m \rightarrow g$ in H .

- *Faedo-Galerkin approximations.* Set now

$$u_m(t) = \sum_{j=1}^m c_{jm}(t) w_j, \quad (10.29)$$

with $c_{jm} \in H^1(0, T)$, and look at the following finite dimensional problem.

Determine $u_m \in H^1(0, T; V)$ such that, for every $s = 1, \dots, m$,

$$\begin{cases} (\dot{u}_m(t), w_s)_H + B(u_m(t), w_s; t) = \langle f(t), w_s \rangle_*, & \text{a.e. in } (0, T) \\ u_m(0) = g_m. \end{cases} \quad (10.30)$$

Note that, since the differential equation in (10.30) is true for each element of the basis w_s , $s = 1, \dots, m$, then it is true for every $v \in V_m$. Moreover, since $\dot{u}_m \in L^2(0, T; V_m)$, we have (see (6.73), p. 401)

$$(\dot{u}_m(t), v)_H = \langle \dot{u}_m(t), v \rangle_*.$$

We call u_m a *Faedo-Galerkin approximation* of the solution u . Inserting (10.29) into (10.30), we see that (10.30) is equivalent to the following system of ODEs for the unknowns c_{jm} : for every $s = 1, \dots, m$,

$$\begin{cases} \sum_{j=1}^m (w_j, w_s)_H \dot{c}_{jm}(t) + \sum_{j=1}^m B(w_j, w_s; t) c_{jm} = \langle f(t), w_s \rangle_*, & \text{a.e. in } (0, T) \\ c_{jm}(0) = \hat{g}_{jm}, \quad j = 1, \dots, m. \end{cases} \quad (10.31)$$

Introduce the $m \times m$ matrices \mathbf{W} and $\mathbf{B}(t)$ whose entries are:

$$w_{sj} = (w_j, w_s)_H \quad \text{and} \quad b_{sj}(t) = B(w_j, w_s; t)$$

and the vectors

$$\begin{aligned}\mathbf{C}_m^\top(t) &= (c_{1m}(t), \dots, c_{mm}(t)), & \mathbf{g}_m^\top &= (\hat{g}_{1m}, \dots, \hat{g}_{mm}), \\ \mathbf{F}_m^\top(t) &= (f_1(t), \dots, f_m(t)),\end{aligned}$$

where $f_s(t) = \langle f(t), w_s \rangle_*$. Note that \mathbf{W} is nonsingular, since the w_j are independent. Then problem (10.31) can be written in the form

$$\dot{\mathbf{C}}_m(t) + \mathbf{W}^{-1} \mathbf{B}(t) \mathbf{C}_m(t) = \mathbf{W}^{-1} \mathbf{F}_m(t), \quad \mathbf{C}_m(0) = \mathbf{g}_m. \quad (10.32)$$

By a Carathéodory Theorem⁴, the system of linear ODEs with bounded measurable coefficients (10.32) has a unique solution $\mathbf{C}_m \in H^1(0, T; \mathbb{R}^m)$. As a consequence, $u_m \in H^1(0, T; V)$ and satisfies the hypotheses of Theorem 10.6. Thus, the following lemma holds.

Lemma 10.9. *Problem 10.30 has a unique solution $u_m \in H^1(0, T; V)$ and u_m satisfies the energy estimates (10.21) and (10.22), with g_m instead of g .*

10.3.4 Existence

We now prove the existence of the solution.

Lemma 10.10. *The sequence of Faedo-Galerkin approximations converges weakly in $H^1(0, T; V, V^*)$ to the solution of our Abstract Parabolic Problem.*

Proof. Let $\{u_m\}$ be the sequence of Faedo-Galerkin approximations. Since $g_m \rightarrow g$ in H , it follows that, for m large, $\|g_m\|_H \leq 1 + \|g\|_H$, say. Then, Lemma 10.9 yields, for a suitable constant $C_0 = C_0(\alpha, \lambda, \mathcal{M}, T)$,

$$\|u_m\|_{L^2(0, T; V)}, \quad \|\dot{u}_m\|_{L^2(0, T; V^*)} \leq C_0 \left\{ 1 + \|g\|_H + \|f\|_{L^2(0, T; V^*)} \right\}.$$

In other words, the sequences $\{u_m\}$ and $\{\dot{u}_m\}$ are bounded in $L^2(0, T; V)$ and $L^2(0, T; V^*)$, respectively. Then, the weak compactness Theorem 6.57, p. 395, implies that there exists a subsequence $\{u_{m_k}\}$ such that⁵

$$u_{m_k} \rightharpoonup u \text{ in } L^2(0, T; V) \quad \text{and} \quad \dot{u}_{m_k} \rightharpoonup \dot{u} \text{ in } L^2(0, T; V^*).$$

We prove that u is the unique solution of our APP.

First of all, to say that $u_{m_k} \rightharpoonup u$, weakly in $L^2(0, T; V)$, as $k \rightarrow \infty$, means that

$$\int_0^T (u_{m_k}(t), v(t))_V dt \rightarrow \int_0^T (u(t), v(t))_V dt$$

for all $v \in L^2(0, T; V)$. Similarly, $\dot{u}_{m_k} \rightharpoonup \dot{u}$ weakly in $L^2(0, T; V^*)$ means that

$$\int_0^T \langle \dot{u}_{m_k}(t), v(t) \rangle_* dt \rightarrow \int_0^T \langle \dot{u}(t), v(t) \rangle_* dt$$

for all $v \in L^2(0, T; V)$.

⁴ See e.g. [33], A.E. Coddington, N. Levinson, 1955.

⁵ More rigorously, $\dot{u}_{m_k} \rightharpoonup z$ in $L^2(0, T; V^*)$ and one shows that $z = \dot{u}$ (see Problem (10.2)).

By the equivalence of conditions (10.19) and (10.20), we can write

$$\int_0^T \langle \dot{u}_{m_k}(t), v(t) \rangle_* dt + \int_0^T B(u_{m_k}(t), v(t); t) dt = \int_0^T \langle f(t), v(t) \rangle_* dt \quad (10.33)$$

for every $v \in L^2(0, T; V_{m_k})$. Let $N \leq m_k$ and $w \in V_N$. Let $\varphi \in C_0^\infty(0, T)$ and insert $v(t) = \varphi(t)w$ into (10.33). Keeping N fixed and letting $m_k \rightarrow +\infty$, thanks to the weak convergence of u_{m_k} and \dot{u}_{m_k} in their respective spaces, we infer

$$\int_0^T \{ \langle \dot{u}(t), w \rangle_* + B(u(t), w; t) - \langle f(t), w \rangle_* \} \varphi(t) dt = 0. \quad (10.34)$$

Letting now $N \rightarrow \infty$, recalling (10.28), we deduce that (10.34) holds for all $w \in V$. Then, the arbitrariness of φ implies

$$\langle \dot{u}(t), w \rangle_* + B(u(t), w; t) dt = \langle f(t), w \rangle_* \quad (10.35)$$

for almost every $t \in (0, T)$ and for all $w \in V$.

It remains to check that u satisfies the initial condition $u(0) = g$. In (10.34), choose $\varphi \in C^1([0, T])$, $\varphi(0) = 1$, $\varphi(T) = 0$. Integrating by parts the first term (see Theorem 7.104, p. 498, c)) we find, since $\langle u(t), w \rangle_* = (u(t), w)_H$,

$$\int_0^T \{ -(u(t), w)_H \dot{\varphi}(t) + B(u(t), w; t) \varphi(t) - \langle f(t), w \rangle_* \varphi(t) \} dt = (u(0), w)_H. \quad (10.36)$$

Similarly, inserting $v(t) = \varphi(t)w$ with $w \in V_{m_k}$ into (10.33), and integrating by parts, we get

$$\int_0^T \{ -(u_{m_k}(t), w)_H \dot{\varphi}(t) + B(u_{m_k}(t), w; t) \varphi(t) - \langle f(t), w \rangle_* \varphi(t) \} dt = (g_{m_k}, w)_H.$$

If $m_k \rightarrow +\infty$, the left hand side of the last equation converges to the left hand side of (10.36), while

$$(g_{m_k}, w)_H \rightarrow (g, w)_H.$$

Thus, we deduce that

$$(u(0), w)_H = (g, w)_H,$$

for every $w \in V$. The density of V in H yields $u(0) = g$.

Therefore u is a solution of our \mathcal{APP} . Finally, note that, due to the uniqueness of the solution of \mathcal{APP} , the whole sequence $\{u_m\}$ converges to u , not only a subsequence. Thus the proof is complete. \square

10.4 Parabolic PDEs

10.4.1 Problems for the heat equation

In this section we examine some examples of applications of the results in Sect. 10.3 to the initial-boundary value problem (10.5). All the results hold in spatial dimension $n \geq 1$. In dimension $n = 1$, Ω in an interval $(a, b) \subset \mathbb{R}$ and, as usual, $\partial_v u$ at the boundary means $-u_x(a, t)$ and $u_x(b, t)$. We start with the diffusion equation.

- *The Cauchy-Dirichlet problem.* We have already introduced the weak formulation of the Cauchy-Dirichlet problem (10.6) (see Definitions 10.1 or 10.2). For this problem the Hilbert triplet is $V = H_0^1(\Omega)$, $H = L^2(\Omega)$, $V^* = H^{-1}(\Omega)$, while the bilinear form is $B(u, v) = (\nabla u, \nabla v)_0$. Referring to the assumptions on B on page 587, (i) holds with $\mathcal{M} = 1$ and (ii) holds with $\alpha = 1$, $\lambda = 0$, since B is coercive. Assumption (iii) is empty since B is independent of t . Theorem 10.6 yields:

Problem (10.6) has a unique weak solution $u \in H^1(0, T; H_0^1(\Omega), H^{-1}(\Omega))$. Moreover, the following estimates hold, for every $t \in [0, T]$:

$$\|u(t)\|_0^2, \int_0^t \|\nabla u(s)\|_0^2 ds \leq \|g\|_0^2 + \int_0^t \|f(s)\|_{H^{-1}(\Omega)}^2 ds$$

and

$$\int_0^t \|\dot{w}(s)\|_{H^{-1}(\Omega)}^2 ds \leq 2\|g\|_0^2 + 4 \int_0^t \|f(s)\|_{H^{-1}(\Omega)}^2 ds.$$

- *The Cauchy-Neumann/Robin problem.* Let Ω be a bounded Lipschitz domain. Consider the following problem⁶:

$$\begin{cases} u_t - \Delta u = f & \text{in } Q_T \\ u(\mathbf{x}, 0) = g(\mathbf{x}) & \text{in } \Omega \\ \partial_\nu u(\boldsymbol{\sigma}, t) + h(\boldsymbol{\sigma}) u(\boldsymbol{\sigma}, t) = 0 & \text{on } S_T. \end{cases} \quad (10.37)$$

Assume that $f \in L^2(0, T; L^2(\Omega))$, $g \in L^2(\Omega)$, $h \in L^\infty(\partial\Omega)$ and $h \geq 0$. Choosing the Hilbert triplet $V = H^1(\Omega)$, $H = L^2(\Omega)$, $V^* = H^1(\Omega)^*$, a weak formulation of problem (10.37) reads:

Find $u \in H^1(0, T; H^1(\Omega), H^1(\Omega)^)$ such that $u(0) = g$ and*

$$\langle \dot{u}(t), v \rangle_* + B(u(t), v) = (f(t), v)_0, \quad \forall v \in H^1(\Omega), \text{ a.e. in } (0, T),$$

where, as in the elliptic case,

$$B(u, v) = (\nabla u, \nabla v)_0 + \int_{\partial\Omega} h u v d\sigma. \quad (10.38)$$

We check assumptions (i), (ii), p. 587. Recall that Theorem 7.82, p. 481, gives

$$\|u\|_{L^2(\partial\Omega)} \leq c_{tr} \|u\|_{H^1(\Omega)}.$$

Then

$$|B(u, v)| \leq (1 + c_{tr}^2 \|h\|_{L^\infty(\partial\Omega)}) \|u\|_{H^1(\Omega)} \|v\|_{H^1(\Omega)},$$

so that B is continuous with $\mathcal{M} = 1 + c_{tr}^2 \|h\|_{L^\infty(\partial\Omega)}$. Furthermore, since $h \geq 0$ a.e. on $\partial\Omega$,

$$B(u, u) \geq \|\nabla u\|_0^2 = \|u\|_{H^1(\Omega)}^2 - \|u\|_0^2.$$

⁶ For nonhomogeneous conditions, see Problem 10.3.

Thus B is weakly coercive with $\alpha = \lambda = 1$. Finally, since $\|f(s)\|_* \leq \|f(s)\|_0$ a.e. in $(0, T)$, we conclude that:

Problem (10.37) has a unique weak solution $u \in H^1(0, T; H^1(\Omega), H^1(\Omega)^)$ and, for every $t \in [0, T]$,*

$$\|u(t)\|_0^2, \int_0^t \|w(s)\|_{H^1(\Omega)}^2 ds \leq e^{2t} \left\{ \|g\|_0^2 + \int_0^t \|f(s)\|_0^2 ds \right\}$$

and

$$\int_0^t \|\dot{w}(s)\|_*^2 ds \leq \left\{ C \|g\|_0^2 + (C+2) \int_0^t \|f(s)\|_0^2 ds \right\}$$

with $C = 2\mathcal{M}^2 e^{2t}$.

- *Mixed Dirichlet-Neumann boundary conditions.* Let Ω be a bounded Lipschitz domain. We consider the problem

$$\begin{cases} u_t - \Delta u = f & \text{in } Q_T \\ u(\mathbf{x}, 0) = g(\mathbf{x}) & \text{in } \Omega \\ \partial_\nu u(\sigma, t) = 0 & \text{on } \Gamma_N \times (0, T) \\ u(\sigma, t) = 0 & \text{on } \Gamma_D \times (0, T), \end{cases} \quad (10.39)$$

where Γ_D is a relatively open subset of $\partial\Omega$ and $\Gamma_N = \partial\Omega \setminus \Gamma_D$. For the weak formulation we choose $H = L^2(\Omega)$ and $V = H_{0,\Gamma_D}^1(\Omega)$, with inner product $(u, v)_V = (\nabla u, \nabla v)_0$. Recall that in $H_{0,\Gamma_D}^1(\Omega)$ a Poincaré's inequality holds:

$$\|v\|_0 \leq C_P \|\nabla v\|_0. \quad (10.40)$$

A weak formulation of problem (10.39) reads:

Find $u \in H^1(0, T; V, V^)$ such that $u(0) = g$ and*

$$\langle \dot{u}(t), v \rangle_* + B(u(t), v) = (f(t), v)_0, \quad \forall v \in V, \text{ a.e. in } (0, T),$$

where $B(u, v) = (\nabla u, \nabla v)_0$.

The bilinear form $B(w, v) = (\nabla w, \nabla v)_0$ is continuous, with $\mathcal{M} = 1$, and coercive, with $\alpha = 1$. Moreover $\|f(t)\|_{V^*} \leq C_P \|f(t)\|_0$ for almost every $t \in (0, T)$. We conclude that:

If $f \in L^2(0, T; L^2(\Omega))$, $g \in L^2(\Omega)$, the problem (10.39) has a unique weak solution $u \in H^1(0, T; V, V^)$. Moreover, the following inequalities hold for all $t \in [0, T]$:*

$$\|u(t)\|_0^2, \int_0^t \|\nabla w(s)\|_0^2 ds \leq \|g\|_0^2 + C_P^2 \int_0^t \|f(s)\|_0^2 ds$$

and

$$\int_0^t \|\dot{w}(s)\|_*^2 ds \leq \left\{ 2 \|g\|_0^2 + 4C_P^2 \int_0^t \|f(s)\|_0^2 ds \right\}.$$

10.4.2 General Equations

The weak formulation of the general problem (10.5) follows the pattern of the previous sections. Choose a Hilbert triplet $\{V, H, V^*\}$, where $H = L^2(\Omega)$ and V is a Sobolev space, $H_0^1(\Omega) \subseteq V \subseteq H^1(\Omega)$. In general Ω is a bounded *Lipschitz* domain. A weak formulation of problem (10.5) reads:

Find $u \in H^1(0, T; V, V^)$ such that $u(0) = g$ and*

$$\langle \dot{u}(t), v \rangle_* + B(u(t), v; t) = \langle f(t), v \rangle_*, \quad \forall v \in V, \text{ a.e. in } (0, T)$$

where

$$B(u, v; t) = \int_{\Omega} \{ \mathbf{A}(\mathbf{x}, t) \nabla u \cdot \nabla v + (\mathbf{c}(\mathbf{x}, t) \cdot \nabla u) v + a(\mathbf{x}, t) uv \} d\mathbf{x}$$

or, in the case of Robin conditions,

$$B(u, v; t) = \int_{\Omega} \{ \mathbf{A}(\mathbf{x}, t) \nabla u \cdot \nabla v + (\mathbf{c}(\mathbf{x}, t) \cdot \nabla u) v + a(\mathbf{x}, t) uv \} d\mathbf{x} + \int_{\partial\Omega} h(\sigma) uv d\sigma$$

with $h \in L^\infty(\partial\Omega)$, $h \geq 0$ a.e. on $\partial\Omega$. Notice that B is time dependent, in general. Under the hypotheses (10.3) and (10.4), p. 582, by following what we did in the elliptic case in Subsects. 8.4.2, 8.4.3 and 8.4.4, it is not difficult to show that

$$|B(u, v; t)| \leq \mathcal{M} \|u\|_V \|v\|_V$$

so that B is *continuous* in V . The constant \mathcal{M} depends only on n, T and on M, γ_0, α_0 , that is on the size of the coefficients a_{ij}, c_j, a (also on Ω and $\|h\|_{L^\infty(\partial\Omega)}$ in the case of Robin condition).

Also, B is *weakly coercive*. In fact from (10.4), we have, for every $\varepsilon > 0$:

$$\int_{\Omega} (\mathbf{c} \cdot \nabla u) u d\mathbf{x} \geq -\gamma_0 \|\nabla u\|_0 \|u\|_0 \geq -\frac{\gamma_0}{2} \left[\varepsilon \|\nabla u\|_0^2 + \frac{1}{\varepsilon} \|u\|_0^2 \right]$$

and

$$\int_{\Omega} au^2 d\mathbf{x} \geq -\alpha_0 \|u\|_0^2,$$

whence, as $h \geq 0$ a.e. on $\partial\Omega$,

$$B(u, u; t) \geq \left(\nu - \frac{\gamma_0 \varepsilon}{2} \right) \|\nabla u\|_0^2 - \left(\frac{\gamma_0}{2\varepsilon} + \alpha_0 \right) \|u\|_0^2. \quad (10.41)$$

Thus, if $\gamma_0 = 0$, i.e. $\mathbf{c} = \mathbf{0}$, assumption (ii), p. 587, holds with any $\lambda > \alpha_0$. If $\gamma_0 > 0$, choose in (10.41)

$$\varepsilon = \frac{\nu}{\gamma_0} \quad \text{and} \quad \lambda = 2 \left(\frac{\gamma_0}{2\varepsilon} + \nu_0 \right) = 2 \left(\frac{\gamma_0^2}{2\nu} + \nu_0 \right).$$

Then

$$B(u, v; t) + \lambda \|u\|_0^2 \geq \frac{\nu}{2} \|\nabla u\|_0^2 + \frac{\lambda}{2} \|u\|_0^2 \geq \min \left\{ \frac{\nu}{2}, \frac{\lambda}{2} \right\} \|u\|_{H^1(\Omega)}^2$$

so that B is *weakly coercive* independently of the choice of V , with $\alpha = \min \left\{ \frac{\nu}{2}, \frac{\lambda}{2} \right\}$. Moreover, for fixed u, v in V , the function

$$t \mapsto B(u, v; t)$$

is measurable by Fubini's Theorem.

Thus, from Theorem 10.6, we can draw the following conclusion.

Theorem 10.11. If $f \in L^2(0, T; V^*)$ and $g \in L^2(\Omega)$, there exists a unique weak solution u of problem (10.5). Moreover

$$\max_{t \in [0, T]} \|u(t)\|_0^2, \quad \int_0^T \|u(t)\|_V^2 dt \leq C \left\{ \int_0^T \|f(t)\|_*^2 dt + \|g\|_0^2 \right\}$$

and

$$\int_0^T \|u(t)\|_*^2 dt \leq C \left\{ \int_0^T \|f(t)\|_*^2 dt + \|g\|_0^2 \right\}$$

where, in general, C depends only on $n, T, \nu, \lambda, M, \gamma_0, \alpha_0$ (and also on $\Omega, \|h\|_{L^\infty(\partial\Omega)}$ for Robin's conditions).

Remark 10.12. The method works with nonhomogeneous boundary conditions as well. For instance, for the initial-Dirichlet problem, if the data is the trace on S_T of a function $\varphi \in L^2(0, T; H^1(\Omega))$ with $\dot{\varphi} \in L^2(0, T; L^2(\Omega))$, the change of variable $w = u - \varphi$ reduces the problem to homogeneous boundary conditions.

Example 10.13. Figure 10.2 shows the graph of the solution of the following Cauchy-Dirichlet problem:

$$\begin{cases} u_t - u_{xx} + 2u_x = 0.2tx & 0 < x < 5, t > 0 \\ u(x, 0) = \max(2 - 2x, 0) & 0 < x < 5 \\ u(0, t) = 2 - t/6, u(5, t) = 0 & t > 0. \end{cases} \quad (10.42)$$

Note the tendency of the drift term $2u_x$, to “transport to the right” the initial data and the effect of the source term $0.2tx$ to increase the solution near $x = 5$, more and more with time.

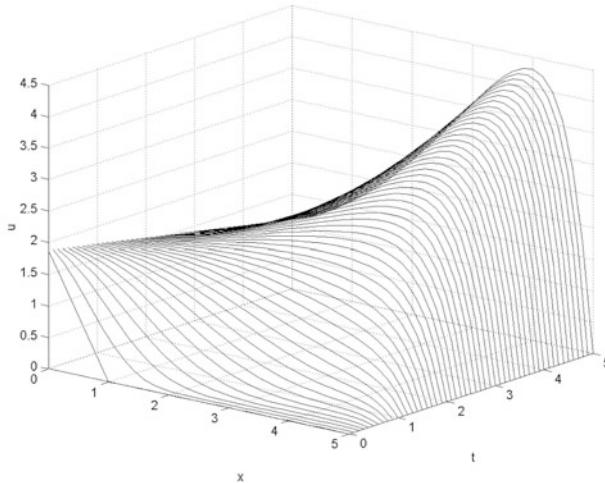


Fig. 10.2 The solution of problem (10.42)

10.4.3 Regularity

As in the elliptic case, the regularity of the solution improves with the regularity of the data. For reasons of brevity, we limit ourselves to give the details in the case of the Cauchy-Dirichlet problem (10.6) for the heat equation. Precisely, we have:

Theorem 10.14. *Let Ω be a bounded, Lipschitz domain and u be the weak solution of problem (10.6). If $g \in H_0^1(\Omega)$ and $f \in L^2(0, T; L^2(\Omega))$, then $u \in L^\infty(0, T; H_0^1(\Omega))$, $\dot{u} \in L^2(0, T; L^2(\Omega))$ and*

$$\max_{s \in [0, T]} \|\nabla u(s)\|_0^2 + \int_0^T \|\dot{u}(s)\|_0^2 ds \leq \|\nabla g\|_0^2 + \int_0^T \|f(s)\|_0^2 ds. \quad (10.43)$$

If in addition, Ω is a C^2 -domain, then $u \in L^2(0, T; H^2(\Omega))$ and

$$\int_0^T \|u(s)\|_{H^2(\Omega)}^2 ds \leq C(n, \Omega) \left\{ \|\nabla g\|_0^2 + \int_0^T \|f(s)\|_0^2 ds \right\}. \quad (10.44)$$

Proof. Let us go back to the weak formulation for the Faedo-Galerkin approximations, that in this case reads

$$\begin{cases} (\dot{u}_m(t), w_s)_0 + (\nabla u_m(t), \nabla w_s)_0 = (f(t), w_s)_0, & \text{a.e } t \in [0, T] \\ u_m(0) = g_m, \end{cases} \quad (10.45)$$

for $s = 1, \dots, m$. Take as a basis $\{w_s\}$ in $H_0^1(\Omega)$ a sequence of Dirichlet eigenfunctions for the Laplace operators in Ω . In particular (see Theorem 8.8, p. 517), we may choose the w_s to be orthonormal in $L^2(\Omega)$ and orthogonal in $H_0^1(\Omega)$. Write $u(t) = \sum_{j=1}^{\infty} c_j(t) w_j$ with convergence in V , and $u_m(t) = \sum_{j=1}^m c_j(t) w_j$. Note that $\dot{u}_m(t) \in L^2(\Omega)$ for almost

every $t \in (0, T)$. Multiplying the differential equation in (10.45) by $\dot{c}_j(t)$ and summing for $j = 1, \dots, m$, we get

$$\|\dot{u}_m(t)\|_0^2 + (\nabla u_m(t), \nabla \dot{u}_m(t))_0 = (f(t), \dot{u}_m(t))_0, \quad (10.46)$$

for a.e. $t \in [0, T]$. Now, observe that

$$(\nabla u_m(t), \nabla \dot{u}_m(t))_0 = \frac{1}{2} \frac{d}{dt} \|\nabla u_m(t)\|_0^2, \quad \text{a.e. in } (0, T)$$

and that, by Schwarz's inequality,

$$(f(t), \dot{u}_m(t))_0 \leq \|f(t)\|_0 \|\dot{u}_m(t)\|_0 \leq \frac{1}{2} \|f(t)\|_0^2 + \frac{1}{2} \|\dot{u}_m(t)\|_0^2.$$

From this inequality and (10.46), we infer

$$\frac{d}{dt} \|\nabla u_m(t)\|_0^2 + \|\dot{u}_m(t)\|_0^2 \leq \|f(t)\|_0^2, \quad \text{a.e. in } (0, T).$$

An integration over $(0, t)$ yields, since $\|\nabla g_m\|_0^2 \leq \|\nabla g\|_0^2$,

$$\|\nabla u_m(t)\|_0^2 + \int_0^t \|\dot{u}_m(s)\|_0^2 ds \leq \int_0^t \|f(s)\|_0^2 ds + \|\nabla g\|_0^2. \quad (10.47)$$

Passing to the limit as $m \rightarrow \infty$, we deduce that the same estimate holds for u and therefore, that $u \in L^\infty(0, T; H_0^1(\Omega))$, $\dot{u} \in L^2(0, T; L^2(\Omega))$ and (10.43) holds. In particular, since $\dot{u}(t) \in L^2(\Omega)$ for almost all $t \in (0, T)$, for every $v \in V$ we may write

$$(\nabla u(t), \nabla v)_0 = (f(t) - \dot{u}(t), v)_0, \quad \text{a.e. in } (0, T).$$

Now, the regularity theory for elliptic equations (Theorem 8.28, p. 539) implies that $u(t) \in H^2(\Omega)$ for almost all $t \in (0, T)$ and that

$$\|u(t)\|_{H^2(\Omega)}^2 \leq C(n, \Omega) \{ \|f(t)\|_0^2 + \|\dot{u}(t)\|_0^2 \}.$$

Integrating over $(0, T)$ and using (10.47) to estimate $\int_0^t \|\dot{u}(s)\|_0^2 ds$, we obtain both $u \in L^2(0, T; H^2(\Omega))$ and the estimate (10.44). \square

Remark 10.15. Theorem 10.14 continues to hold unaltered, if we replace $H_0^1(\Omega)$ by $H^1(\Omega)$. Also the proof is the same, using the Neumann eigenvectors for the Laplace operators instead of the Dirichlet ones. In this way we obtain a regularity result for the homogeneous Neumann problem and the estimates (10.43), (10.44).

Further global regularity requires that the data f and g satisfy suitable compatibility conditions on $\partial\Omega \times \{t = 0\}$. Assume that Ω is smooth, $f \in C^\infty(\overline{Q}_T)$ and $g \in C^\infty(\overline{\Omega})$. Suppose $u \in C^\infty(\overline{Q}_T)$. Since $u = 0$ on the lateral side, we have

$$u = \partial_t u = \dots = \partial_t^j u = \dots = 0, \quad \forall j \geq 0, \text{ on } S_T \quad (10.48)$$

which hold, by continuity, also for $t = 0$ on $\partial\Omega$. On the other hand, the heat equation gives

$$\partial_t u = \Delta u + f, \quad \partial_t^2 u = \Delta(\partial_t u) + \partial_t f = \Delta^2 u + \Delta f + \partial_t f$$

and, in general ($\Delta^0 = \text{Identity}$)

$$\partial_t^j u = \Delta^j u + \sum_{k=0}^{j-1} \Delta^k \partial_t^{j-1-k} f, \quad \forall j \geq 1, \text{ in } Q_T.$$

For $t = 0$ we have $\Delta^j u(0) = \Delta^j g$ and from (10.48) we get the conditions

$$g = 0 \quad \text{and} \quad \Delta^j g + \sum_{k=0}^{j-1} \Delta^k \partial_t^{j-1-k} f(0) = 0, \quad \forall j \geq 1, \text{ on } \partial\Omega. \quad (10.49)$$

Thus, conditions (10.49) are necessary in order to have $u \in C^\infty(\overline{Q}_T)$. It turns out that they are sufficient as well, as stated by the following theorem⁷.

Theorem 10.16. *Let Ω be a bounded smooth domain and u be the weak solution of problem (10.6). If $f \in C^\infty(\overline{Q}_T)$, $g \in C^\infty(\overline{\Omega})$ and the compatibility conditions (10.49) hold, then $u \in C^\infty(\overline{Q}_T)$.*

Remark 10.17. The above regularity results can be extended to general divergence form equations. For instance, assume that \mathbf{A} is symmetric and the coefficients a_{ij} , c_j and a (also h in case of Robin conditions) do not depend on t . Then, if $f \in L^2(0, T; H)$ and $g \in V$, the weak solution u of the homogeneous Cauchy-Dirichlet or Cauchy-Robin/Neumann belongs to $L^\infty(0, T; V)$ while $\dot{u} \in L^2(0, T; H)$.

Moreover, if Ω is of class C^2 and the coefficients a_{ij} are Lipschitz continuous in $\overline{\Omega}$, then $u \in L^2(0, T; H^2(\Omega))$. The proof goes along the lines of that of Theorem 10.14 (see Problem 10.10). Moreover, if all the data are smooth, then $u \in C^\infty(Q_T)$. The regularity up to the parabolic boundary of Q_T requires compatibility conditions generalizing those in (10.49) for which we refer to the more specialized books in the references.

10.5 Weak Maximum Principles

Weak solutions satisfy maximum principles that generalize those in Chap. 2. Consider the operator

$$\mathcal{P}u = u_t - \operatorname{div}(\mathbf{A}(\mathbf{x}, t) \nabla u) + a(\mathbf{x}, t) u,$$

where \mathbf{A} and a satisfies the hypotheses (10.3) and (10.4), and the associated bilinear form

$$B(w, v; t) = \int_{\Omega} \{ \mathbf{A}(\mathbf{x}, t) \nabla w \cdot \nabla v + a(\mathbf{x}, t) wv \} d\mathbf{x}.$$

Let Ω be a bounded Lipschitz domain, $f \in L^2(0, T; H^{-1}(\Omega))$ and $g \in L^2(\Omega)$.

⁷ For the proof, see e.g. [32], Brezis, 2010.

Under this assumptions, we know that there exists a unique weak solution $u \in H^1(0, T; H_0^1(\Omega), H^{-1}(\Omega))$ of the problem

$$\langle \dot{u}(t), v \rangle_* + B(u(t), v; t) = \langle f(t), v \rangle_* \quad (10.50)$$

for almost all $t \in (0, T)$ and all $v \in H_0^1(\Omega)$, with $u(0) = g$.

We say that $f(t) \geq 0$ in $\mathcal{D}'(\Omega)$ for almost every $t \in (0, T)$ if $\langle f(t), \varphi \rangle \geq 0$ a.e. in $(0, T)$, $\forall \varphi \in \mathcal{D}(\Omega)$, $\varphi \geq 0$ in Ω . The following result holds.

Theorem 10.18. *Let u be a weak solution of problem (10.50). If $g \geq 0$ a.e. in Ω and $f(t) \geq 0$ in $\mathcal{D}'(\Omega)$ a.e. in $(0, T)$, then, for every $t \in [0, T]$,*

$$u(t) \geq 0 \text{ a.e. in } \Omega.$$

Proof. For simplicity, we give the proof under the additional assumptions that the coefficients a_{ij}, a are independent of time, $f \in L^2(\Omega)$ and $g \in H_0^1(\Omega)$. Then, from Remark 10.17, we know that $\dot{u} \in L^2(0, T; L^2(\Omega))$ and u satisfies the equation

$$(\dot{u}(t), v)_0 + B(u(t), v) = (f(t), v)_0 \quad (10.51)$$

for all $v \in H_0^1(\Omega)$ and a.e. in $(0, T)$.

Since $u = 0$ on S_T , then $u^-(t) = \max\{-u(t), 0\} \in H_0^1(\Omega)$ for almost every $t \in (0, T)$. Thus, we may choose $v(t) = -u^-(t) \leq 0$ as a test function in (10.51). Recalling that $u(t) = u^+(t) - u^-(t)$, we get, since $|a| \leq \alpha_0$ a.e. in Ω ,

$$\begin{aligned} B(u(t), -u^-(t)) &= \int_{\Omega} \left\{ \mathbf{A}(\mathbf{x}) \nabla u^-(t) \cdot \nabla u^-(t) + a(\mathbf{x}) (u^-(t))^2 \right\} d\mathbf{x} \\ &\geq -\alpha_0 \|u^-(t)\|_0^2 \end{aligned}$$

and, since $f \geq 0$ a.e. in Ω ,

$$(f(t), -u^-(t))_0 \leq 0, \quad \text{a.e. } t \in (0, T).$$

Thus, from (10.51) and $(\dot{u}(t), -u^-(t))_0 = \frac{1}{2} \frac{d}{dt} \|u^-(t)\|_0^2$, we infer that

$$\frac{1}{2} \frac{d}{dt} \|u^-(t)\|_0^2 \leq \alpha_0 \|u^-(t)\|_0^2 \quad (10.52)$$

for almost all $t \in (0, T)$. Now, $g = u(0) \geq 0$ a.e. in Ω entails $\|u^-(0)\|_0^2 = 0$. From (10.52) it follows that $u^-(t) = 0$ and therefore $u(t) = u^+(t) \geq 0$, a.e. in Ω , for all $t \in [0, T]$ since $u^- \in C([0, T]; L^2(\Omega))$. \square

Roughly speaking, Theorem 10.18 says that if u is a weak supersolution for the operator \mathcal{P} , that is $\mathcal{P}u \geq 0$ in Q_T , which is nonnegative on the parabolic boundary $\partial_p Q_T$, then u is nonnegative in all Q_T .

Remark 10.19. Maximum principles for other boundary value problems are also available (see Problem 10.7). For instance, let $u \in H^1(0, T; H^1(\Omega), H^1(\Omega)^*)$ be a weak solution of the Cauchy-Neumann problem

$$\langle \dot{u}(t), v \rangle_* + B(u(t), v; t) = (f(t), v)_0 + (h(t), v)_{L^2(\partial\Omega)} \quad (10.53)$$

a.e. in $(0, T)$ and for all $v \in H^1(\Omega)$, with $u(0) = g$.

Assume that $f \in L^2(Q_T)$, $g \in L^2(\Omega)$ and $h \in L^2(S_T)$. In this case, the maximum principle states that, if $g \geq 0$ a.e. in Ω , $f \geq 0$ a.e. in Q_T and $h \geq 0$ a.e. on S_T , then $u(t) \geq 0$ a.e. in Ω , for all $t \in [0, T]$.

In other words, if

$$\begin{cases} \mathcal{P}u \geq 0 & \text{in } Q_T \\ \mathbf{A}\nabla u \cdot \boldsymbol{\nu} \geq 0 & \text{on } S_T \\ u(\mathbf{x}, 0) \geq 0 & \text{in } \Omega, \end{cases}$$

in a weak sense, then $u \geq 0$ a.e. in Q_T .

10.6 The Wave Equation

10.6.1 Hyperbolic Equations

The wave propagation in a nonhomogeneous and anisotropic medium leads to second order *hyperbolic* equations, generalization of the classical wave equation $u_{tt} - c^2 \Delta u = f$. With the same notations of section 10.1, an equation in *divergence form* of the type

$$u_{tt} - \operatorname{div}(\mathbf{A}(\mathbf{x}, t) \nabla u) + \mathbf{b}(\mathbf{x}, t) \cdot \nabla u + c(\mathbf{x}, t) u = f(\mathbf{x}, t) \quad (10.54)$$

or in *nondivergence form* of the type

$$u_{tt} - \operatorname{tr}(\mathbf{A}(\mathbf{x}, t) D^2 u) + \mathbf{b}(\mathbf{x}, t) \cdot \nabla u + c(\mathbf{x}, t) u = f(\mathbf{x}, t) \quad (10.55)$$

is called **hyperbolic** in $Q_T = \Omega \times (0, T)$ if

$$\mathbf{A}(\mathbf{x}, t) \boldsymbol{\xi} \cdot \boldsymbol{\xi} > 0, \quad \text{a.e. in } Q_T, \forall \boldsymbol{\xi} \in \mathbb{R}^n, \boldsymbol{\xi} \neq \mathbf{0}.$$

The typical problems for hyperbolic equations are those already considered for the wave equation. Given f in Q_T , we want to determine a solution u of (10.54) or (10.55) satisfying the *initial* conditions

$$u(\mathbf{x}, 0) = g(\mathbf{x}), \quad u_t(\mathbf{x}, 0) = h(\mathbf{x}) \quad \text{in } \Omega$$

and one of the usual boundary conditions (*Dirichlet, Neumann, mixed or Robin*) on the lateral boundary $S_T = \partial\Omega \times (0, T]$.

Even if from the phenomenological point of view, the hyperbolic equations display substantial differences from the parabolic ones, for *divergence form* equations it is possible to give a similar weak formulation, which can be analyzed by means of Faedo-Galerkin method. Since for general equations the theory is quite complicated and technical, we will limit ourselves to the Cauchy-Dirichlet problem for the wave equation.

10.6.2 The Cauchy-Dirichlet problem

Consider the problem

$$\begin{cases} u_{tt} - c^2 \Delta u = f & \text{in } Q_T \\ u(\mathbf{x}, 0) = g(\mathbf{x}), \quad u_t(\mathbf{x}, 0) = h(\mathbf{x}) & \text{in } \Omega \\ u(\sigma, t) = 0 & \text{on } S_T. \end{cases} \quad (10.56)$$

As usual, to find a weak formulation, we proceed formally and multiply the wave equation by a smooth function $v = v(\mathbf{x}, t)$, vanishing on S_T . Integrating over Q_T , we find

$$\int_{Q_T} u_{tt}(\mathbf{x}, t) v(\mathbf{x}, t) d\mathbf{x} dt - c^2 \int_{Q_T} \Delta u(\mathbf{x}, t) v(\mathbf{x}, t) d\mathbf{x} dt = \int_{Q_T} f(\mathbf{x}, t) v(\mathbf{x}, t) d\mathbf{x} dt.$$

Integrating by parts the second term with respect to \mathbf{x} , we get, since $v = 0$ on $\partial\Omega$,

$$\int_{Q_T} u_{tt}(\mathbf{x}, t) v(\mathbf{x}, t) d\mathbf{x} dt + c^2 \int_{Q_T} \nabla u(\mathbf{x}, t) \cdot \nabla v(\mathbf{x}, t) d\mathbf{x} dt = \int_{Q_T} f(\mathbf{x}, t) v(\mathbf{x}, t) d\mathbf{x} dt$$

which becomes, in the notations of the previous sections, separating space from time,

$$\int_0^T (\ddot{u}(t), v(t))_0 dt + c^2 \int_0^T (\nabla u(t), \nabla v(t))_0 dt = \int_0^T (f(t), v(t))_0 dt,$$

where \ddot{u} stays for u_{tt} . Again, the natural space for u is $L^2(0, T; H_0^1(\Omega))$. Thus, for almost all $t > 0$, $u(t) \in V = H_0^1(\Omega)$, and $\Delta u(t) \in V^* = H^{-1}(\Omega)$. On the other hand, from the wave equation we have

$$u_{tt} = c^2 \Delta u + f$$

and it is natural to require $\ddot{u} \in L^2(0, T; V^*)$. Accordingly, a reasonable assumption for \dot{u} is $\dot{u} \in L^2(0, T; H)$, $H = L^2(\Omega)$, an intermediate space between $L^2(0, T; V)$ and $L^2(0, T; V^*)$. Thus, we look for solutions u such that

$$u \in L^2(0, T; V), \quad \dot{u} \in L^2(0, T; H), \quad \ddot{u} \in L^2(0, T; V^*). \quad (10.57)$$

For simplicity, we also assume $f \in L^2(0, T; H)$ and

$$u(0) = g \in V, \quad \dot{u}(0) = h \in H. \quad (10.58)$$

Since u satisfies (10.57) and $H \hookrightarrow V^*$, it follows that

$$u \in C([0, T]; H) \quad \text{and} \quad \dot{u} \in C([0, T]; V^*)$$

and therefore the initial conditions (10.58) makes sense.

The above considerations lead to the following definition, analogous to Definition 10.1, p. 584, for the heat equation.

Definition 10.20. Let $f \in L^2(0, T; H)$ and $g \in V$, $h \in H$. We say that $u \in L^2(0, T; V)$ is a weak solution to problem (10.56) if

$$\dot{u} \in L^2(0, T; H), \quad \ddot{u} \in L^2(0, T; V^*)$$

and:

1. For all $v \in L^2(0, T; V)$,

$$\int_0^T \langle \ddot{u}(t), v(t) \rangle_* dt + c^2 \int_0^T (\nabla u(t), \nabla v(t))_0 dt = \int_0^T (f(t), v(t))_0 dt. \quad (10.59)$$

2. $u(0) = g$, $\dot{u}(0) = h$.

As in the case of the heat equation, condition 1 can be stated in the following equivalent form (see Problem 10.15):

- 1'. For all $v \in V$ and a.e. in $(0, T)$,

$$\langle \ddot{u}(t), v \rangle_* + c^2 (\nabla u(t), \nabla v)_0 = (f(t), v)_0. \quad (10.60)$$

Remark 10.21. We leave it to the reader to check that if f , g , h are smooth functions and u is a smooth weak solution of problem (10.56), then u is actually a classical solution.

Remark 10.22. As in Remark 10.3, p. 585, the equation (10.60) may be interpreted in the sense of distributions in $\mathcal{D}'(0, T)$. Indeed (see Problem 10.16) for all $v \in V$, the real function

$$t \mapsto w(t) = \langle \ddot{u}(t), v \rangle_*$$

is a distribution in $\mathcal{D}'(0, T)$ and

$$w(t) = \frac{d^2}{dt^2} (u(t), v)_0, \quad \text{in } \mathcal{D}'(0, T). \quad (10.61)$$

As a consequence, the equation (10.60) may be written in the form

$$\frac{d^2}{dt^2} (u(t), v)_0 + c^2 (\nabla u(t), \nabla v)_0 = (f(t), v)_0$$

for all $v \in V$ and in the sense of distributions in $\mathcal{D}'(0, T)$.

10.6.3 The method of Faedo-Galerkin

We want to show that problem (10.56) has a unique solution, which continuously depends on the data, in appropriate norms. Once more, we are going to use the method of Faedo-Galerkin, but we use it somehow more directly, without reference to an abstract result. Here are the main steps, emphasizing the differences with the parabolic case.

1. Let $\{w_k\}_{k=1}^\infty$ be the sequence of normalized Dirichlet eigenfunctions for the Laplace operator in Ω . Recall that they constitute

an orthogonal basis in V and an orthonormal basis in H .

In particular, we can write

$$g = \sum_{k=1}^{\infty} \hat{g}_k w_k, \quad h = \sum_{k=1}^{\infty} \hat{h}_k w_k,$$

where $\hat{g}_k = (g, w_k)_0$, $\hat{h}_k = (h, w_k)_0$, with the series converging in V and H , respectively.

2. Let $V_m = \text{span}\{w_1, w_2, \dots, w_m\}$ and define

$$u_m(t) = \sum_{k=1}^m r_k(t) w_k, \quad g_m = \sum_{k=1}^m \hat{g}_k w_k, \quad h_m = \sum_{k=1}^m \hat{h}_k w_k.$$

We construct the sequence of Faedo-Galerkin approximations u_m , by solving the following *projected* problem:

Determine $u_m \in H^2(0, T; V)$ such that, for all $s = 1, \dots, m$,

$$\begin{cases} (\ddot{u}_m(t), w_s)_0 + c^2 (\nabla u_m(t), \nabla w_s)_0 = (f(t), w_s)_0, & 0 \leq t \leq T \\ u_m(0) = g_m, \quad \dot{u}_m(0) = h_m. \end{cases} \quad (10.62)$$

Since the differential equation in (10.62) must hold for each element of the basis w_s , $s = 1, \dots, m$, it must hold for every $v \in V_m$. Moreover, since $u_m \in H^2(0, T; V)$ we have $\ddot{u}_m \in L^2(0, T; V)$, so that

$$\langle \ddot{u}_m(t), v \rangle_* = (\ddot{u}_m(t), v)_0.$$

3. We show that the sequences $\{u_m\}$, $\{\dot{u}_m\}$ and $\{\ddot{u}_m\}$ are bounded in $L^2(0, T; V)$, $L^2(0, T; H)$ and $L^2(0, T; V^*)$, respectively (*energy estimates*). Then, once again, the weak compactness Theorem 6.57, p. 395, implies that a subsequence $\{u_{m_k}\}$ converges weakly in $L^2(0, T; V)$ to u , while $\{\dot{u}_{m_k}\}$ and $\{\ddot{u}_{m_k}\}$ converge weakly in $L^2(0, T; H)$ and $L^2(0, T; V^*)$ to \dot{u} and \ddot{u} , respectively.

4. We prove that the function u in step **3** is the unique weak solution of problem (10.56), so that the whole sequences $\{u_m\}$, $\{\dot{u}_m\}$ and $\{\ddot{u}_m\}$ converges weakly to u , \dot{u} , \ddot{u} in their respective spaces.

10.6.4 Solution of the approximate problem

The following lemma holds.

Lemma 10.23. *For all $m \geq 1$, there exists a unique solution to problem (10.62). In particular, since $u_m \in H^2(0, T; V)$, we have $u_m \in C^1([0, T]; V)$.*

Proof. Observe that, since w_1, w_2, \dots, w_m are orthonormal in H ,

$$(\ddot{u}_m(t), w_s)_0 = \sum_{k=1}^m (w_k, w_s)_0 \ddot{r}_k(t) = \ddot{r}_s(t)$$

and since they are orthogonal in V ,

$$c^2 \sum_{k=1}^m (\nabla w_k, \nabla w_s)_0 r_k(t) = c^2 (\nabla w_s, \nabla w_s)_0 r_s(t) = c^2 \|\nabla w_s\|_0^2 r_s(t).$$

Set

$$F_s(t) = (f(t), w_s)_0, \quad \mathbf{F}(t)^\top = (F_1(t), \dots, F_m(t))$$

and

$$\mathbf{R}_m(t)^\top = (r_1(t), \dots, r_m(t)), \quad \mathbf{g}_m = (\hat{g}_1, \dots, \hat{g}_m), \quad \mathbf{h}_m = (\hat{h}_1, \dots, \hat{h}_m).$$

If we introduce the diagonal matrix

$$\mathbf{W} = \text{diag} \{ \|\nabla w_1\|_0^2, \|\nabla w_2\|_0^2, \dots, \|\nabla w_m\|_0^2 \}$$

of order m , problem (10.62) is equivalent to the following system of m uncoupled linear ordinary differential equations, with constant coefficients and right hand side in $L^2(0, T; \mathbb{R}^m)$:

$$\ddot{\mathbf{R}}_m(t) + c^2 \mathbf{W} \mathbf{R}_m(t) = \mathbf{F}_m(t), \quad \text{a.e. in } (0, T) \tag{10.63}$$

with initial conditions

$$\mathbf{R}_m(0) = \mathbf{g}_m, \quad \dot{\mathbf{R}}_m(0) = \mathbf{h}_m.$$

Since $\mathbf{F}_m \in L^2(0, T; \mathbb{R}^m)$, system (10.63) has a unique solution $\mathbf{R}_m(t) \in H^2(0, T; \mathbb{R}^m)$. From

$$u_m(t) = \sum_{k=1}^m r_k(t) w_k,$$

we deduce $u_m \in H^2(0, T; V)$.

□

10.6.5 Energy estimates

We want to show that, from the sequence of Galerkin approximations $\{u_m\}$, it is possible to extract a subsequence converging to the weak solution of the original problem. We are going to prove that the relevant Sobolev norms of u_m can be controlled by the norms of the data, **in a way that does not depend on m** . Moreover, the estimates must be powerful enough in order to pass to the limit as $m \rightarrow +\infty$ in the approximating equation

$$(\ddot{u}_m(t), v)_0 + c^2 (\nabla u_m(t), \nabla v)_0 = (f(t), v)_0.$$

In this case we can estimate the norms of u_m in $L^\infty(0, T; V)$, of \dot{u}_m in $L^\infty(0, T; H)$ and of \ddot{u} in $L^2(0, T; V^*)$, that is the norms

$$\max_{t \in [0, T]} \|\nabla u_m(t)\|_0, \quad \max_{t \in [0, T]} \|\dot{u}_m(t)\|_0 \quad \text{and} \quad \int_0^T \|\ddot{u}_m(t)\|_*^2 dt.$$

Theorem 10.24. *Let u_m be the solution of problem (10.62). Then, for all $t \in [0, T]$,*

$$\|\dot{u}_m(t)\|_0^2 + c^2 \|\nabla u_m(t)\|_0^2 \leq e^t \left\{ c^2 \|\nabla g\|_0^2 + \|h\|_0^2 + \int_0^t \|f(s)\|_0^2 ds \right\}. \quad (10.64)$$

Proof. Since $u_m \in H^2(0, T; V)$, we may choose $v = \dot{u}_m(t)$ as a test function in (10.62). We find

$$(\ddot{u}_m(t), \dot{u}_m(t))_0 + c^2 (\nabla u_m(t), \nabla \dot{u}_m(t))_0 = (f(t), \dot{u}_m(t))_0 \quad (10.65)$$

for a.e. $t \in [0, T]$. Observe that, for almost every $t \in (0, T)$,

$$(\ddot{u}_m(t), \dot{u}_m(t))_0 = \frac{1}{2} \frac{d}{dt} \|\dot{u}_m(t)\|_0^2,$$

and

$$(\nabla u_m(t), \nabla \dot{u}_m(t))_0 = \frac{c^2}{2} \frac{d}{dt} \|\nabla u_m(t)\|_0^2.$$

Moreover, by Schwarz's inequality,

$$(f(t), \dot{u}_m(t))_0 \leq \|f(t)\|_0 \|\dot{u}_m(t)\|_0 \leq \frac{1}{2} \|f(t)\|_0^2 + \frac{1}{2} \|\dot{u}_m(t)\|_0^2,$$

so that, from (10.65), we deduce

$$\frac{d}{dt} \{ \|\dot{u}_m(t)\|_0^2 + c^2 \|\nabla u_m(t)\|_0^2 \} \leq \|f(t)\|_0^2 + \|\dot{u}_m(t)\|_0^2.$$

We now integrate over $(0, t)$, recalling that $u_m(0) = g_m$, $\dot{u}_m(0) = h_m$ and observing that

$$\|\nabla g_m\|_0^2 \leq \|\nabla g\|_0^2, \quad \|h_m\|_0^2 \leq \|h\|_0^2,$$

by the orthogonality of the basis. We find:

$$\begin{aligned} & \|\dot{u}_m(t)\|_0^2 + c^2 \|\nabla u_m(t)\|_0^2 \\ & \leq \|h_m\|_0^2 + c^2 \|\nabla g_m\|_0^2 + \int_0^t \|f(s)\|_0^2 ds + \int_0^t \|\dot{u}_m(s)\|_0^2 ds \\ & \leq \|h\|_0^2 + c^2 \|\nabla g\|_0^2 + \int_0^t \|f(s)\|_0^2 ds + \int_0^t \|\dot{u}_m(s)\|_0^2 ds. \end{aligned}$$

Let

$$\Psi(t) = \|\dot{u}_m(t)\|_0^2 + c^2 \|\nabla u_m(t)\|_0^2, \quad G(t) = \|h\|_0^2 + c^2 \|\nabla g\|_0^2 + \int_0^t \|f(s)\|_0^2 ds.$$

Note that both Ψ and G are continuous in $[0, T]$ and G is nondecreasing. Then we have

$$\Psi(t) \leq G(t) + \int_0^t \Psi(s) ds$$

and Gronwall's Lemma 10.8, p. 589, yields, for every $t \in [0, T]$,

$$\|\dot{u}_m(t)\|_0^2 + c^2 \|\nabla u_m(t)\|_0^2 \leq e^t \left\{ \|h\|_0^2 + c^2 \|\nabla g\|_0^2 + \int_0^t \|f(s)\|_0^2 ds \right\}. \quad \square$$

We now give a control of the norm of \ddot{u}_m in $L^2(0, T; V^*)$.

Theorem 10.25. *Let u_m be the solution of problem (10.62). Then, for all $t \in [0, T]$,*

$$\int_0^t \|\ddot{u}_m(s)\|_*^2 ds \leq C_0 \left\{ \|h\|_0^2 + c^2 \|\nabla g\|_0^2 \right\} + C_1 \int_0^t \|f(s)\|_0^2 ds, \quad (10.66)$$

where $C_0 = 2c^2(e^t - 1)$ and $C_1 = C_0 + 2C_P^2$, where ε_P is the Poincaré constant.

Proof. Let $v \in V$ and write

$$v = w + z$$

with $w \in V_m = \text{span}\{w_1, w_2, \dots, w_m\}$ and $z \in V_m^\perp$. Since w_1, \dots, w_k are orthogonal in V , we have

$$\|\nabla w\|_0 \leq \|\nabla v\|_0.$$

Choosing w as a test function in problem (10.62), we obtain

$$(\ddot{u}_m(t), v)_0 = (\ddot{u}_m(t), w)_0 = -c^2 (\nabla u_m(t), \nabla w)_0 + (f(t), w)_0.$$

Since

$$|(\nabla u_m(t), \nabla w)_0| \leq \|\nabla u_m(t)\|_0 \|\nabla w\|_0, \quad |(f(t), w)_0| \leq C_P \|f(t)\|_0 \|\nabla w\|_0$$

we may write

$$\begin{aligned} |(\ddot{u}_m(t), v)_0| & \leq \{c^2 \|\nabla u_m(t)\|_0 + C_P \|f(t)\|_0\} \|\nabla w\|_0 \\ & \leq \{c^2 \|\nabla u_m(t)\|_0 + C_P \|f(t)\|_0\} \|\nabla v\|_0. \end{aligned}$$

Thus, by the definition of norm in V^* , we infer

$$\|\ddot{u}_m(t)\|_* \leq c^2 \|\nabla u_m(t)\|_0 + C_P \|f(t)\|_0.$$

Squaring and integrating over $(0, t)$ we obtain

$$\int_0^t \|\ddot{u}_m(s)\|_*^2 ds \leq 2c^4 \int_0^t \|\nabla u_m(s)\|_0^2 ds + 2C_P^2 \int_0^t \|f(s)\|_0^2 ds$$

and (10.64) gives (10.66). \square

10.6.6 Existence, uniqueness and stability

Theorems 10.24 and 10.25 imply that the sequences $\{u_m\}$ and $\{\dot{u}_m\}$ are bounded in $L^\infty(0, T; V)$ and $L^\infty(0, T; H)$ respectively, hence, in particular, in $L^2(0, T; V)$ and $L^2(0, T; H)$, while the sequence $\{\ddot{u}_m\}$ is bounded in $L^2(0, T; V^*)$.

The weak compactness Theorem 6.57, p. 395, implies that there exists a subsequence, which for simplicity we still denote by $\{u_m\}$, such that, as $m \rightarrow \infty$,⁸

$$\begin{aligned} u_m &\rightharpoonup u \quad \text{weakly in } L^2(0, T; V) \\ \dot{u}_m &\rightharpoonup \dot{u} \quad \text{weakly in } L^2(0, T; H) \\ \ddot{u}_m &\rightharpoonup \ddot{u} \quad \text{weakly in } L^2(0, T; V^*). \end{aligned}$$

The following theorem holds:

Theorem 10.26. *Let $f \in L^2(0, T; H)$, $g \in V$, $h \in H$. Then u is the unique weak solution of problem (10.56). Moreover,*

$$\|u\|_{L^\infty(0, T; V)}^2, \|\dot{u}\|_{L^\infty(0, T; H)}^2, \|\ddot{u}\|_{L^2(0, T; V^*)}^2 \leq C \left\{ \|f\|_{L^2(0, T; H)}^2 + c^2 \|\nabla g\|_0^2 + \|h\|_0^2 \right\}$$

with $C = C(c, T, \Omega)$.

Proof. **Existence.** We know that:

$$\int_0^T (\nabla u_m(t), \nabla v(t))_0 dt \rightarrow \int_0^T (\nabla u(t), \nabla v(t))_0 dt$$

for all $v \in L^2(0, T; V)$,

$$\int_0^T (\dot{u}_m(t), w(t))_0 dt \rightarrow \int_0^T (\dot{u}(t), w(t))_0 dt$$

for all $w \in L^2(0, T; H)$, and

$$\int_0^T (\ddot{u}_m(t), v(t))_0 = \int_0^T \langle \ddot{u}_m(t), v(t) \rangle_* dt \rightarrow \int_0^T \langle \ddot{u}(t), v(t) \rangle_* dt$$

for all $v \in L^2(0, T; V)$.

⁸ Rigorously, $u_m \rightharpoonup u$ in $L^2(0, T; V)$, $\dot{u}_m \rightharpoonup w$ in $L^2(0, T; H)$, $\ddot{u}_m \rightharpoonup z$ in $L^2(0, T; V^*)$ and one shows that, $w = \dot{u}$, $z = \ddot{u}$.

610 10 Weak Formulation of Evolution Problems

We want to use these properties to pass to the limit as $m \rightarrow +\infty$ in problem (10.62), keeping in mind that the test functions have to be chosen in V_m . Let $N \leq m$ and $w \in V_N$. Let $\varphi \in C_0^\infty(0, T)$ and insert $v(t) = \varphi(t)w$ as a test function into (10.62). Integrating over $(0, T)$, we get

$$\int_0^T \{(\ddot{u}_m(t), w)_0 + c^2 (\nabla u_m(t), \nabla w)_0 - (f(t), w)_0\} \varphi(t) dt = 0. \quad (10.67)$$

Keeping N fixed and letting $m \rightarrow +\infty$, thanks to the weak convergence of u_m and \dot{u}_m in their respective spaces, we obtain

$$\int_0^T \{(\ddot{u}(t), w)_0 + c^2 (\nabla u(t), \nabla w)_0 - (f(t), w)_0\} \varphi(t) dt = 0. \quad (10.68)$$

Letting now $N \rightarrow \infty$, we infer that (10.68) holds for all $w \in V$. Then, the arbitrariness of φ entails

$$\langle \ddot{u}(t), w \rangle_* + c^2 (\nabla u(t), \nabla w)_0 = (f(t), w)_0$$

for all $w \in V$ and almost all $t \in (0, T)$. Therefore u satisfies (10.60) and we know that $u \in C([0, T]; H)$, $\dot{u} \in C([0, T]; V^*)$.

To check the initial conditions, we proceed as in Theorem 10.10, p. 592. Let $v(t) = \varphi(t)w + \psi(t)z$ with $w, z \in V$ and $\varphi, \psi \in C^2([0, T])$ such that

$$\varphi(0) = 1, \dot{\varphi}(0) = \varphi(T) = \dot{\varphi}(T) = 0, \quad \dot{\psi}(0) = 1, \psi(0) = \psi(T) = \dot{\psi}(T) = 0.$$

With this v , integrating by parts twice, using Theorem 7.104, p. 498, c), and the density of H into V^* , we can write, since $\dot{u}(0) \in V^*$,

$$\int_0^T \langle \ddot{u}(t), v(t) \rangle_* dt = \int_0^T (u(t), \ddot{v}(t))_0 dt - \langle \dot{u}(0), z \rangle_* + (u(0), w)_0.$$

Thus, inserting this v into (10.59) we obtain

$$\int_0^T \{(u(t), \ddot{v}(t))_0 + c^2 (\nabla u(t), \nabla v(t))_0 - (f(t), v(t))_0\} = \langle \dot{u}(0), z \rangle_* - (u(0), w)_0. \quad (10.69)$$

On the other hand, let $v_N(t) = \varphi(t)w_N + \psi(t)z_N$, where w_N, z_N are the projections of w, z , respectively, into V_N . Insert this v_N as a test function into (10.62), integrate twice by parts and let first $m \rightarrow +\infty$ and then $N \rightarrow \infty$. We deduce

$$\int_0^T \{\langle u(t), \ddot{v}(t) \rangle_* + c^2 (\nabla u(t), \nabla v(t))_0 - (f(t), v(t))_0\} = (h, z)_0 - (g, w)_0. \quad (10.70)$$

From (10.69) and (10.70) we deduce

$$\langle \dot{u}(0), z \rangle_* - (u(0), w)_0 = (h, z)_0 - (g, w)_0 = \langle h, z \rangle_* - (g, w)_0,$$

for every $w, z \in V$. Since V is dense in H , the arbitrariness of w and z gives

$$\dot{u}(0) = h \quad \text{and} \quad u(0) = g.$$

Uniqueness. Assume $g = h \equiv 0$ and $f \equiv 0$. We want to show that $u \equiv 0$. The proof would be easy if we could choose \dot{u} as a test function in (10.60), but $\dot{u}(t)$ does not belong

to V . For fixed s , set

$$v(t) = \begin{cases} \int_t^s u(r) dr & \text{if } 0 \leq t \leq s \\ 0 & \text{if } s \leq t \leq T. \end{cases}$$

We have $v(t) \in V$ for all $t \in [0, T]$, so that we may insert it into (10.60). After an integration over $(0, T)$, we deduce

$$\int_0^s \{ \langle \ddot{u}(t), v(t) \rangle_* + c^2 (\nabla u(t), \nabla v(t))_0 \} dt = 0. \quad (10.71)$$

An integration by parts yields,

$$\int_0^s \langle \ddot{u}(t), v(t) \rangle_* dt = - \int_0^s (\dot{u}(t), \dot{v}(t))_0 dt = \int_0^s (\dot{u}(t), u(t))_0 dt = \frac{1}{2} \int_0^s \frac{d}{dt} \|u(t)\|_0^2 dt,$$

since $v(s) = \dot{u}(0) = 0$ and $\dot{v}(t) = -u(t)$, if $0 < t < s$. On the other hand,

$$\int_0^s (\nabla u(t), \nabla v(t))_0 dt = - \int_0^s (\nabla \dot{v}(t), \nabla v(t))_0 dt = - \frac{1}{2} \int_0^s \frac{d}{dt} \|\nabla v(t)\|_0^2 dt.$$

Hence, from (10.71),

$$\int_0^s \frac{d}{dt} \{ \|u(t)\|_0^2 - c^2 \|\nabla v(t)\|_0^2 \} dt = 0$$

or

$$\|u(s)\|_0^2 + c^2 \|\nabla v(0)\|_0^2 = 0$$

which entails $u(s) \equiv 0$.

Stability. To prove the estimates for u and \dot{u} , use Proposition 7.102, p. 496, to pass to the limit as $m \rightarrow \infty$ in (10.64). The estimate for \ddot{u} follows from the weak lower semicontinuity of the norm in $L^2(0, T; V^*)$. \square

Problems

10.1. Consider the Abstract Parabolic Problem of Sect. 10.3, under the assumptions on p. 587. Show that $w(t) = e^{-\lambda_0 t} u(t)$, with $\lambda_0 > \lambda$, satisfies a similar problem with a coercive bilinear form.

10.2. Let $\{V, H, V^*\}$ be a Hilbert triplet. Let $\{u_m\}_{m \geq 1} \subset L^2(0, T; V)$ satisfy the following conditions:

i) $u_m \rightharpoonup u$ in $L^2(0, T; V)$.

ii) $\dot{u}_m \rightharpoonup z$ in $L^2(0, T; V^*)$.

Prove that $z = \dot{u}$.

[Hint: note that $\int_0^T \langle \dot{u}_m(t), v \rangle_* \varphi(t) dt = - \int_0^T (u_m(t), v)_H \dot{\varphi}(t) dt$ for all $\varphi \in \mathcal{D}(0, T)$ and all $v \in V$].

10.3. Consider the problem

$$\begin{cases} u_t - (a(x)u_x)_x + b(x)u_x + c(x)u = f(x, t) & 0 < x < 1, 0 < t < T \\ u(x, 0) = g(x), & 0 \leq x \leq 1 \\ u(0, t) = 0, u(1, t) = k(t). & 0 \leq t \leq T. \end{cases}$$

- 1) Reduce the problem to homogeneous Dirichlet conditions.
- 2) Write a weak formulation for the new problem.
- 3) Prove the well-posedness of the problem, under suitable hypotheses on the coefficients a, b, c and the data f, g, k . Write a stability estimate for the original u .

10.4. Consider the Neumann/Robin problem (10.37), p. 594, with non-homogeneous boundary condition $\partial_{\nu}u + hu = q$, with $q \in L^2(S_T)$.

- a) Give a weak formulation of the problem and derive the main energy estimates.
- b) Deduce existence and uniqueness of the solution.

10.5. Let

$$\Omega = \{\mathbf{x} \in \mathbb{R}^2 : x_1 + 4x_2^2 < 4\}, \Gamma_D = \partial\Omega \cap \{x_1 \geq 0\}, \Gamma_N = \partial\Omega \setminus \Gamma_D.$$

Consider the mixed problem

$$\begin{cases} u_t - \operatorname{div}(A_{\alpha}(\mathbf{x})\nabla u) + \mathbf{b}(\mathbf{x}) \cdot \nabla u - \alpha u = x_2 & \text{in } \Omega \times (0, T) \\ u(\mathbf{x}, 0) = \mathcal{H}(x_1) & \text{in } \Omega \\ u(\boldsymbol{\sigma}, t) = 0 & \text{on } \Gamma_D \times [0, T] \\ A_{\alpha}(\boldsymbol{\sigma})\nabla u(\boldsymbol{\sigma}, t) \cdot \boldsymbol{\nu}(\boldsymbol{\sigma}) = -\sigma_1 & \text{on } \Gamma_N \times [0, T] \end{cases}$$

where \mathcal{H} is the Heaviside function and

$$A_{\alpha}(\mathbf{x}) = \begin{pmatrix} 1 & 0 \\ 0 & \alpha e^{|\mathbf{x}|^2} \end{pmatrix}, \quad \mathbf{b}(\mathbf{x}) = \begin{pmatrix} x_2 \\ x_2/|x_2| \end{pmatrix}.$$

Determine for which values of the real parameter α , the problem is uniformly parabolic. Give a weak formulation and analyze the well-posedness.

10.6. The potassium concentration $c = c(\mathbf{x}, t)$, $\mathbf{x} = (x, y, z)$, in a cell of spherical shape Ω and radius R , satisfies the following evolution problem:

$$\begin{cases} c_t - \operatorname{div}(\mu \nabla c) - \sigma c = 0 & \text{in } \Omega \times (0, T) \\ \mu \nabla c \cdot \boldsymbol{\nu} + \chi c = c_{ext}, & \text{on } S_T \\ c(\mathbf{x}, 0) = c_0(\mathbf{x}, 0). & \text{in } \Omega \end{cases}$$

where: c_{ext} is a given external concentration, σ and χ are positive scalars and μ is strictly positive. Write a weak formulation and analyze the well-posedness, providing suitable assumptions on μ, c_{ext} and c_0 .

10.7. State and prove an analogue of Theorem 10.14, p. 598, for the heat equation and homogeneous Robin/Neumann boundary conditions.

10.8. State and prove an analogue of the weak maximum principle in Theorem 10.18, p. 601, for the initial Robin/Neumann boundary value problem.

[Hint: Mimick the proof of Theorem 10.18. Observe that if $v \in H^1(\Omega)$ and $v \geq 0$ a.e. in Ω , then the trace of v on $\partial\Omega$ is nonnegative on $\partial\Omega$.]

10.9. Let Ω be a bounded Lipschitz domain and u be the weak solution of

$$\begin{cases} u_t - \Delta u = 0 & \text{in } Q_T \\ u = 0 & \text{on } S_T \\ u(\mathbf{x}, 0) = g(\mathbf{x}) & \text{in } \Omega. \end{cases}$$

a) Prove that, if $g \in L^\infty(\Omega)$,

$$\min\{0, \operatorname{ess\,inf}_\Omega u\} = u(\mathbf{x}, t) \leq \max\{0, \operatorname{ess\,sup}_\Omega g\} \text{ in } Q_T.$$

b) Prove that, if $g \in C(\overline{\Omega})$ with $g = 0$ on $\partial\Omega$, then $u \in C^\infty(Q_T) \cap C(\overline{Q}_T)$.

[Hint: b) Let $\{g_m\} \subset C_0^\infty(\Omega)$ such that $\|g_m - g\|_{L^\infty(\Omega)} \rightarrow 0$ as $m \rightarrow +\infty$. Denote by u_m the weak solution corresponding to the initial data g_m . Check that $u_m \in C^\infty(\overline{Q}_T)$. Show that

$$\|u_m(t) - u(t)\|_{L^\infty(\Omega)} \leq \|g_m - g\|_{L^\infty(\Omega)}$$

for all $t \in [0, T]$.

10.10. Let Ω be a bounded, Lipschitz domain and u be the weak solution of

$$\begin{cases} u_t - \operatorname{div}(\mathbf{A}(\mathbf{x}) \nabla u) + \mathbf{c}(x) \cdot \nabla u + a(\mathbf{x})u = f(\mathbf{x}, t) & \text{in } Q_T \\ u(\mathbf{x}, 0) = g(\mathbf{x}) & \text{in } \Omega \\ u = 0 & \text{on } S_T. \end{cases}$$

Assume that $f \in L^2(0, T; L^2(\Omega))$, $g \in H_0^1(\Omega)$ and that \mathbf{A} is symmetric.

- a) Show that $u \in L^\infty(0, T; H_0^1(\Omega))$, $\dot{u} \in L^2(0, T; L^2(\Omega))$ and prove a stability estimate similar to (10.43).
- b) If Ω is of class C^2 and the coefficients a_{ij} are Lipschitz continuous in $\overline{\Omega}$, show that $u \in L^2(0, T; H^2(\Omega))$ and prove a stability estimate similar to (10.44).

10.11. Let $\Omega \subset \mathbb{R}^n$ be a bounded, Lipschitz domain, $Q_T = \Omega \times (0, T)$. For $k \geq 1$, let u_k be the unique weak solution of the following Cauchy-Dirichlet problem:

$$\begin{cases} \mathcal{P}_k u_k = \partial_t u_k - \operatorname{div}(\mathbf{A}_k(\mathbf{x}, t) \nabla u_k) + \mathbf{c}_k(\mathbf{x}, t) \cdot \nabla u_k + a_k(\mathbf{x}, t)u_k = f & \text{in } Q_T \\ u_k = 0 & \text{on } S_T \\ u_k(\mathbf{x}, 0) = g(\mathbf{x}) & \text{in } \Omega \end{cases}$$

where $f \in L^2(0, T; H^{-1}(\Omega))$, $g \in L^2(\Omega)$ and \mathcal{P}_k is a uniformly parabolic operator satisfying conditions (10.3) and (10.4), p. 582. Assume that, as $k \rightarrow +\infty$,

$$\mathbf{A}_k \rightarrow \mathbf{A}_0 \text{ in } L^\infty(Q_T; \mathbb{R}^{n^2}), \quad \mathbf{c}_k \rightarrow \mathbf{c}_0 \text{ in } L^\infty(Q_T; \mathbb{R}^n), \quad a_k \rightarrow a_0 \text{ in } L^\infty(Q_T).$$

Denote by u_0 the unique weak solution of the same problem for the limit operator \mathcal{P}_0 . Show that $u_0 \rightarrow u$ in $L^2(0, T; H_0^1(\Omega))$ and in $C([0, T; L^2(\Omega)])$, and $\dot{u}^k \rightarrow \dot{u}_0$ in $L^2(0, T; H^{-1}(\Omega))$.

10.12. Consider the Cauchy-Neumann problem

$$\begin{cases} u_{tt} - c^2 \Delta u = f & \text{in } Q_T \\ \partial_\nu u(x) = 0 & \text{on } S_T \\ u(\mathbf{x}, 0) = g(\mathbf{x}), u_t(\mathbf{x}, 0) = h(\mathbf{x}) & \text{in } \Omega. \end{cases}$$

Write a weak formulation of the problem and prove the analogues of Theorems 10.24, p. 607, 10.25, p. 608, and 10.26, p. 609.

10.13. *Concentrated reaction.* Consider the problem

$$\begin{cases} u_{tt} - u_{xx} + u(x, t) \delta(x) = 0 & -1 < x < 1, 0 < t < T \\ u(x, 0) = g(x), u_t(x, 0) = h(x) & -1 \leq x \leq 1 \\ u(-1, t) = u(1, t) = 0 & 0 \leq t \leq T, \end{cases}$$

where $\delta(x)$ denotes the Dirac δ at the origin.

- a) Write a weak formulation for the problem.
- b) Prove the well-posedness of the problem, under suitable hypotheses on g and h .

[Hint: a) Let $V = H_0^1(-1, 1)$ and $H = L^2(-1, 1)$. A weak formulation is: find $u \in L^2(-1, 1; V)$, with $\dot{u} \in L^2(-1, 1; H)$ and $\ddot{u} \in L^2(-1, 1; V^*)$, such that, for every $v \in V$,

$$\langle \ddot{u}(t), v \rangle_* + (u_x(t), v_x) + u(0, t)v(0) = 0 \quad \text{a.e. in } (0, T)$$

and

$$\|u(t) - g\|_H \rightarrow 0, \|\dot{u}(t) - h\|_{V^*} \rightarrow 0$$

as $t \rightarrow 0$].

10.14. Show that the two conditions (10.60) and (10.59), p. 604, are equivalent.

10.15. Let $u \in L^2(0, T; V)$, with $\dot{u} \in L^2(0, T; H)$ and $\ddot{u} \in L^2(0, T; V^*)$. Show that the function $t \mapsto w(t) = \langle \ddot{u}(t), v \rangle_*$ is a distribution in $\mathcal{D}'(0, T)$ and

$$w(t) = \frac{d^2}{dt^2} (u(t), v)_0 \quad \text{in } \mathcal{D}'(0, T).$$

Chapter 11

Systems of Conservation Laws

11.1 Introduction

The motion of a compressible fluid with negligible viscosity (gas dynamics) or the propagation of waves in shallow waters are typical phenomena leading to systems of first order conservation laws. This chapter constitutes an introduction to the basic concepts in this important area of nonlinear PDEs, still lacking of a completely satisfactory theory. Let us introduce the following notations:

$$\mathbf{u} : \mathbb{R}^n \times [0, T] \rightarrow \mathbb{R}^m, \quad \mathbf{F} : U \subseteq \mathbb{R}^m \rightarrow \mathcal{M}_{m,n}$$

where $\mathcal{M}_{m,n}$ is the set of $m \times n$ matrices. In this context, \mathbf{u} is a *state variable*, and we write it as a column vector or as a point in \mathbb{R}^m . If $\Omega \subset \mathbb{R}^n$ is a bounded domain, the equation

$$\frac{d}{dt} \int_{\Omega} \mathbf{u} \, d\mathbf{x} = - \int_{\partial\Omega} \mathbf{F}(\mathbf{u}) \cdot \boldsymbol{\nu} \, d\sigma \quad (11.1)$$

expresses the balance between the rate of variation of $\int_{\Omega} \mathbf{u} \, d\mathbf{x}$ and the *inward* flux of \mathbf{u} through $\partial\Omega$, governed by the *flux function* \mathbf{F} . As usual, $\boldsymbol{\nu}$ denotes the outward unit normal to $\partial\Omega$. Using Gauss formula, we rewrite (11.1) in the form

$$\int_{\Omega} [\mathbf{u}_t + \operatorname{div}\mathbf{F}(\mathbf{u})] \, d\mathbf{x} = \mathbf{0}. \quad (11.2a)$$

If (11.2a) holds in an arbitrary region Ω , we deduce the *conservation law*

$$\mathbf{u}_t + \operatorname{div}\mathbf{F}(\mathbf{u}) = \mathbf{0}. \quad (11.3)$$

We shall limit ourselves to present the very basic concepts and results in the case $n = 1$. For the sake of simplicity, we develop the theory for $U = \mathbb{R}^m$. If $n = 1$, (11.3) becomes

$$\mathbf{u}_t + \mathbf{F}(\mathbf{u})_x = \mathbf{0}. \quad (11.4)$$

Under smoothness condition for \mathbf{F} , we can take the derivative with respect to x in the second term and write (11.4) into the *nonconservative form*

$$\mathbf{u}_t + D\mathbf{F}(\mathbf{u}) \mathbf{u}_x = \mathbf{0}, \quad (11.5)$$

where $D\mathbf{F}$ denotes the Jacobian matrix of \mathbf{F} . System (11.5) is a particular case of *first order quasilinear system*, of the form

$$\mathbf{u}_t + \mathbf{A}(x, t, \mathbf{u}) \mathbf{u}_x = \mathbf{f}(x, t, \mathbf{u}), \quad (11.6)$$

where \mathbf{A} is a $m \times m$ matrix and

$$\mathbf{f} : \mathbb{R} \times \mathbb{R} \times \mathbb{R}^m \rightarrow \mathbb{R}^m.$$

If \mathbf{A} does not depend on \mathbf{u} the system is *semilinear*; if, moreover,

$$\mathbf{f}(x, t, \mathbf{u}) = \mathbf{B}(x, t) \mathbf{u} + \mathbf{c}(x, t),$$

the system is *linear*.

For example, the change of variables

$$u_x = w_1 \text{ and } u_t = w_2$$

transforms the wave equation $u_{tt} - c^2 u_{xx} = f$ into the linear system

$$\mathbf{w}_t + \mathbf{A}\mathbf{w}_x = \mathbf{f}, \quad (11.7)$$

where $\mathbf{w} = \begin{pmatrix} w_1 \\ w_2 \end{pmatrix}$, $\mathbf{f} = \begin{pmatrix} 0 \\ f \end{pmatrix}$ and

$$\mathbf{A} = \begin{pmatrix} 0 & -1 \\ -c^2 & 0 \end{pmatrix}.$$

Note that the matrix \mathbf{A} has the two real distinct eigenvalues $\lambda_{\pm} = \pm c$, with eigenvectors

$$\mathbf{v}_+ = \begin{pmatrix} 1 \\ -c \end{pmatrix} \text{ and } \mathbf{v}_- = \begin{pmatrix} 1 \\ c \end{pmatrix}$$

normal to the two families of characteristics $x - ct = k$, $x + ct = k$, reflecting the *hyperbolic* nature of the wave equation. By analogy, we give the following definition¹.

Definition 11.1. We say that the system (11.6), or simply the matrix \mathbf{A} , is strictly hyperbolic if for every $(x, t, \mathbf{u}) \in \mathbb{R} \times \mathbb{R} \times \mathbb{R}^m$, the matrix \mathbf{A} has m real and distinct eigenvalues:

$$\lambda_1(x, t, \mathbf{u}) < \lambda_2(x, t, \mathbf{u}) < \dots < \lambda_m(x, t, \mathbf{u}).$$

If \mathbf{A} is strictly hyperbolic, there exist m linearly independent corresponding eigenvectors

$$\mathbf{r}_1(x, t, \mathbf{u}), \mathbf{r}_2(x, t, \mathbf{u}), \dots, \mathbf{r}_m(x, t, \mathbf{u})$$

¹ The notion of hyperbolicity is deeply analyzed e.g. in [15], M. Renardy and R.C. Rogers, 1993.

and if $\mathbf{\Gamma} = (\mathbf{r}_1 \mid \mathbf{r}_2 \mid \dots \mid \mathbf{r}_m)$ is the matrix of the (column) vectors $\mathbf{r}_1, \mathbf{r}_2, \dots, \mathbf{r}_m$, then we can diagonalize \mathbf{A} , obtaining

$$\mathbf{\Gamma}^{-1} \mathbf{A} \mathbf{\Gamma} = \mathbf{\Lambda} = \text{diag}(\lambda_1, \lambda_2, \dots, \lambda_m). \quad (11.8)$$

The rows of $\mathbf{\Gamma}^{-1}$, denoted by

$$\mathbf{l}_1^\top(x, t, \mathbf{u}), \mathbf{l}_2^\top(x, t, \mathbf{u}), \dots, \mathbf{l}_m^\top(x, t, \mathbf{u}),$$

are *left eigenvectors of \mathbf{A}* , since $\mathbf{\Gamma}^{-1} \mathbf{A} = \mathbf{\Lambda} \mathbf{\Gamma}^{-1}$. Clearly

$$\mathbf{l}_j^\top \mathbf{r}_k = \delta_{jk}$$

where δ_{jk} denotes the Kronecker symbol.

The following two important examples will help us to illustrate the various concepts during the development of the theory.

Example 11.2. Gas dynamics. Consider a gas in motion along a one-dimensional pipe. Using Eulerian variables, denote by x the coordinate along the axis of the pipe and by $\rho(x, t), v(x, t), p(x, t), E(x, t)$, density, velocity, pressure and specific energy² of the gas, at the point x and time t , respectively. The motion is governed by the laws of conservation of mass, momentum and energy that give for the above four thermodynamic variables the following system of three equations:

$$\begin{cases} \rho_t + (\rho v)_x = 0 & \text{(mass)} \\ (\rho v)_t + (\rho v^2 + p)_x = 0 & \text{(momentum)} \\ (\rho E)_t + ([\rho E + p]v)_x = 0 & \text{(energy).} \end{cases} \quad (11.9)$$

To close the model, we need an equation of state, for instance of the form $p = p(\rho, e)$ where

$$e = E - \frac{v^2}{2}$$

is the specific internal energy. A special case is the *ideal gas law* $p = R\rho e/c_v$, where c_v is the specific heat at constant volume.

Setting

$$\mathbf{q} = \begin{pmatrix} \rho \\ \rho v \\ \rho E \end{pmatrix} \quad \text{and} \quad \mathbf{F}(\rho, q, e) = \begin{pmatrix} \rho v \\ \rho v^2 + p \\ (\rho E + p)v \end{pmatrix},$$

we may rewrite (11.9) in the form

$$\mathbf{q}_t + \mathbf{F}(\mathbf{q})_x = \mathbf{0}.$$

For smooth flows, after simple computations, system (11.9) can be rewritten in the nonconservative form

$$\begin{cases} \rho_t + v\rho_x + \rho v_x = 0 \\ v_t + vv_x + \rho^{-1}p_x = 0 \\ E_t + vE_x + p\rho^{-1}v_x = 0. \end{cases} \quad (11.10)$$

² Energy per unit mass.

The third equation can be further simplified by introducing the specific entropy s and taking into account the first law of Thermodynamics:

$$T \frac{Ds}{Dt} = \frac{DE}{Dt} - \frac{p}{\rho^2} \frac{D\rho}{Dt} \quad (11.11)$$

where T is the absolute temperature. In fact, (11.11) reads more explicitly as

$$T(s_t + vs_x) = E_t + vE_x - p\rho^{-2}(\rho_t + v\rho_x)$$

and, from the first and third equation in (11.10), after some manipulations, we deduce

$$s_t + vs_x = 0.$$

Thus the system (11.10) can be written in the form :

$$\begin{cases} \rho_t + v\rho_x + \rho v_x = 0 \\ v_t + vv_x + \rho^{-1}p_x = 0 \\ s_t + vs_x = 0. \end{cases} \quad (11.12)$$

In this case, the equation of state is given in the form $p = p(\rho, s)$ with $p_\rho > 0$, $p_s > 0$ and $p_{\rho\rho} + \frac{2}{\rho}p_\rho > 0$. For an ideal polytropic gas, $p(\rho, s) = (\gamma - 1)e^{s/c_v}\rho^{-\gamma}$, with $\gamma = c_p/c_v$, where c_p is the specific heat at constant pressure. Observing that

$$p_x = p_\rho\rho_x + p_s s_x$$

and setting

$$\mathbf{z} = \begin{pmatrix} \rho \\ v \\ s \end{pmatrix}, \quad \mathbf{A}(\mathbf{z}) = \mathbf{A}(\rho, v, s) = \begin{pmatrix} v & \rho & 0 \\ p_\rho/\rho & v & p_s/\rho \\ 0 & 0 & v \end{pmatrix},$$

(11.12) is equivalent to

$$\mathbf{z}_t + \mathbf{A}(\mathbf{z})\mathbf{z}_x = \mathbf{0}.$$

The eigenvalues of \mathbf{A} are, putting $p_\rho = c^2$,

$$\lambda_1(\rho, v, s) = v - c, \quad \lambda_2(\rho, v, s) = v, \quad \lambda_3(\rho, v, s) = v + c, \quad (11.13)$$

with corresponding eigenvectors

$$\mathbf{r}_1 = \begin{pmatrix} \rho \\ -c \\ 0 \end{pmatrix}, \quad \mathbf{r}_2 = \begin{pmatrix} p_s \\ 0 \\ -c^2 \end{pmatrix}, \quad \mathbf{r}_3 = \begin{pmatrix} \rho \\ c \\ 0 \end{pmatrix}. \quad (11.14)$$

Thus, the system of gas dynamics (11.12) is strictly hyperbolic.

Example 11.3 (The p-system). Another model for the motion of a gas along a pipe can be derived using Lagrangian coordinates. Under the hypothesis of isoentropic flow³, the equation of state is of the form $p = p(w)$ and the equations of motion

³ No heat exchange occurs among the fluid particles and entropy is at thermodynamical equilibrium.

can be reduced to the system

$$\begin{cases} w_t - v_x = 0 \\ v_t + p(w)_x = 0, \end{cases} \quad (11.15)$$

also known as *p-system*. Here $w = \rho^{-1}$ is the specific volume (thus $w > 0$) and x denotes a spatial coordinate which, in the Lagrangian description, represents a given gas particle in motion. Natural hypotheses on p are:

$$a(w) \equiv -p'(w) > 0 \quad \text{and} \quad a'(w) = -p''(w) < 0$$

for all $w > 0$. Moreover, we assume that

$$\lim_{w \rightarrow 0^+} p(w) = +\infty.$$

The above assumptions are satisfied for instance by an ideal polytropic gas, for which $p(w) = kw^{-\gamma}$, where $k > 0$ and $\gamma \geq 1$.

The *p*-system can be written in the form (11.4) with

$$\mathbf{u} = \begin{pmatrix} w \\ v \end{pmatrix} \quad \text{and} \quad \mathbf{F}(\mathbf{u}) = \begin{pmatrix} -v \\ p(w) \end{pmatrix}.$$

Since

$$\mathbf{A}(\mathbf{u}) = D\mathbf{F}(\mathbf{u}) = \begin{pmatrix} 0 & -1 \\ -a(w) & 0 \end{pmatrix}$$

we see that \mathbf{A} is strictly hyperbolic with eigenvalues

$$\lambda_1 = -\sqrt{a(w)} \quad \text{and} \quad \lambda_2 = \sqrt{a(w)} \quad (11.16)$$

and corresponding eigenvectors

$$\mathbf{r}_1(w, v) = \begin{pmatrix} 1 \\ \sqrt{a(w)} \end{pmatrix} \quad \text{and} \quad \mathbf{r}_2(w, v) = \begin{pmatrix} 1 \\ -\sqrt{a(w)} \end{pmatrix}. \quad (11.17)$$

Our main goal is the analysis of the Riemann problem (see Sect. 11.4) for the system (11.4), due to its importance as a model problem and as a key tool in the numerical approximation methods. After some basic facts about linear hyperbolic system, we first solve the Riemann problem in the case of linear homogeneous system with constant coefficients. In Sect. 11.3 we start the analysis of the quasilinear system (11.4) under the strict hyperbolicity assumption, introducing the notion of characteristic and of Riemann invariants. Then, as in the scalar case $m = 1$, treated in Chap. 4, we construct special solutions under the form of rarefaction waves, contact discontinuities and shocks, introducing an entropy condition to select admissible discontinuities. The last section is devoted to a complete analysis of the Riemann problem for the *p*-System.

11.2 Linear Hyperbolic Systems

In this section we extend the method of characteristics to linear hyperbolic systems.

11.2.1 Characteristics

We have seen in Chap. 4 that the method of *characteristics* plays a key role in the analysis of scalar first order equations ($m = 1$). The method can be easily extended to strictly hyperbolic systems. We start by examining the following Cauchy problem in the linear case:

$$\begin{cases} \mathbf{u}_t + \mathbf{A}(x, t) \mathbf{u}_x = \mathbf{B}(x, t) \mathbf{u} + \mathbf{c}(x, t) & x \in \mathbb{R}, t > 0 \\ \mathbf{u}(x, 0) = \mathbf{g}(x) & x \in \mathbb{R}. \end{cases} \quad (11.18)$$

Let us briefly review the scalar case

$$u_t + a(x, t) u_x = b(x, t) u + c(x, t) \quad (11.19)$$

with initial data

$$u(x, 0) = g(x), \quad x \in \mathbb{R}.$$

For this equation, a characteristic is a line γ of equation $x = x(t)$, obtained as a solution of the ODE

$$\dot{x} = a(x, t).$$

Evaluating u along this characteristic, that is setting $z(t) = u(x(t), t)$, and differentiating this expression, we get

$$\dot{z} = u_t + \dot{x} u_x = u_t + a(x, t) u_x.$$

Using (11.19) and the initial condition, we end up with the following Cauchy problem:

$$\begin{cases} \dot{z} = b(x(t), t) z + c(x(t), t) \\ z(0) = u(x(0), 0) = g(x(0)). \end{cases} \quad (11.20)$$

Solving (11.20), we obtain the values of u along γ .

For hyperbolic systems we can reproduce something similar. Indeed, let us choose m linearly independent eigenvectors

$$\mathbf{r}_1(x, t), \mathbf{r}_2(x, t), \dots, \mathbf{r}_m(x, t)$$

and form the matrix $\mathbf{\Gamma} = (\mathbf{r}_1 \mid \mathbf{r}_2 \mid \dots \mid \mathbf{r}_m)$. Set $\mathbf{v} = \mathbf{\Gamma}^{-1} \mathbf{u}$. Using (11.8), the system (11.18) for \mathbf{v} becomes

$$\begin{cases} \mathbf{v}_t + \mathbf{\Lambda} \mathbf{v}_x = \mathbf{B}^* \mathbf{v} + \mathbf{c}^* & x \in \mathbb{R}, t > 0 \\ \mathbf{v}(x, 0) = \mathbf{\Gamma}^{-1}(x, 0) \mathbf{g}(x) = \mathbf{g}^*(x) & x \in \mathbb{R}, \end{cases} \quad (11.21)$$

where $\mathbf{B}^* = \mathbf{\Gamma}^{-1} \mathbf{B} \mathbf{\Gamma} - \mathbf{\Gamma}^{-1} (\mathbf{A} \mathbf{\Gamma}_x + \mathbf{\Gamma}_t)$ and $\mathbf{c}^* = \mathbf{\Gamma}^{-1} \mathbf{c}$.

In the left hand side of system (11.21), the unknowns are *uncoupled* and the equation for the component v_k of \mathbf{v} takes the following form:

$$\frac{\partial v_k}{\partial t} + \lambda_k \frac{\partial v_k}{\partial x} = \sum_{j=1}^m b_{kj}^* v_j + c_k^*. \quad (11.22)$$

Note that if $b_{kj}^* = 0$ for $j \neq k$, then the right hand side is uncoupled as well and each equation is of the form (11.19). Thus, it is natural to call *characteristics* the solutions of the equations

$$\frac{dx}{dt} = \lambda_k(x, t), \quad k = 1, \dots, m.$$

11.2.2 Classical solutions of the Cauchy problem

We know examine problem (11.18). Under suitable hypotheses on the matrices \mathbf{A}, \mathbf{B} and on the vectors \mathbf{c} and \mathbf{g} , there exists a unique classical solution, i.e. of class C^1 in the strip $\mathbb{R} \times [0, T]$. To prove it we use the Contraction Theorem 6.78, p. 417. Precisely, we prove the following theorem.

Theorem 11.4. *Let $S = \mathbb{R} \times [0, T]$. Assume that \mathbf{A} is strictly hyperbolic and moreover:*

- i) \mathbf{A} and \mathbf{B} have entries of class $C^1(S)$, bounded with bounded derivatives.
- ii) \mathbf{c} and \mathbf{g} are of class $C^1(S)$ and $C^1(\mathbb{R})$, respectively, bounded with bounded derivatives.

Then problem (11.18) has a unique solution $\mathbf{u} \in C^1(S)$.

Proof. It is enough to solve problem (11.21). First note that, since each eigenvalue $\lambda_k = \lambda_k(x, t)$ of \mathbf{A} is simple, then $\lambda_k \in C^1(S)$ and it is bounded with bounded derivatives⁴. In particular, there exists a number $L > 0$ such that

$$|\lambda_k(x, t)| \leq L \quad x \in \mathbb{R}, 0 \leq t \leq T$$

for all $k = 1, \dots, m$. Moreover, from these bounds and the hypotheses i) and ii) we can find β, γ, η such that:

$$\sup_S |b_{ij}^*(x, t)| \leq \beta, \quad \sup_S |\mathbf{c}^*(x, t)| \leq \gamma, \quad \sup_{\mathbb{R}} |\mathbf{g}^*(x)| \leq \eta.$$

Consider now a point (ξ, τ) , $\tau \leq T_1 \leq T$, with T_1 to be suitably chosen later, and the m characteristics Γ_i issued from (ξ, τ) , of equation

$$x_i(t) = x_i(t; \xi, \tau),$$

where $x_i = x_i(t)$ solves the equation

$$\frac{dx_i}{dt} = \lambda_i(x_i, t).$$

⁴ It is a simple application of the Implicit Function Theorem.

Note that x_i is continuously differentiable⁵ with respect to ξ, τ . Since each λ_i is continuously differentiable and $|\lambda_i(x, t)| \leq L$, each Γ_i is well defined in a common interval $0 \leq t \leq \tau$. Define $w_i(t) = v_i(x_i(t; \xi, \tau), t)$. Then

$$\dot{w}_i = \lambda_i \partial_x v_i + \partial_t v_i$$

and from (11.21) we deduce

$$\dot{w}_i(t) = c_i^*(x_i(t; \xi, \tau)) + \sum_{j=1}^m b_{ij}^*(x_i(t; \xi, \tau), t) w_j(t).$$

Integrating over $(0, \tau)$ we find, recalling that $w_i(\tau) = v_i(\xi, \tau)$ and $w_i(0) = g_i^*(x_i(0; \xi, \tau))$,

$$v_i(\xi, \tau) = g_i^*(x_i(0; \xi, \tau)) + \int_0^\tau [c_i^*(x_i(s; \xi, \tau), s) + \sum_{j=1}^m b_{ij}^*(x_i(s; \xi, \tau), s) v_j(x_i(s; \xi, \tau), s)] ds \quad (11.23)$$

for every $i = 1, \dots, m$. Let $\mathbf{G}(\xi, \tau)$ denote the vector whose components are

$$g_i^*(x_i(0; \xi, \tau)) + \int_0^\tau c_i^*(x_i(s; \xi, \tau), s) ds, \quad i = 1, \dots, m.$$

Then (11.23) can be written in the fixed point form

$$\mathbf{v} = \mathbf{G} + \mathcal{P}(\mathbf{v}), \quad (11.24)$$

where $\mathbf{z} = \mathcal{P}(\mathbf{v})$ has components

$$z_i(\xi, \tau) = \int_0^\tau \sum_{j=1}^m b_{ij}^*(x_i(s; \xi, \tau), s) v_j(x_i(s; \xi, \tau), s) ds.$$

We now introduce the Banach space

$$X = C_b(\mathbb{R} \times [0, T_1]; \mathbb{R}^m)$$

of the bounded and continuous functions in $\mathbb{R} \times [0, T_1]$, equipped with the norm

$$\|\mathbf{z}\|_X = \sup_{i=1, \dots, m} \{|z_i(\xi, \tau)|; \xi \in \mathbb{R}, 0 \leq \tau \leq T_1\}.$$

We have $\mathcal{P}: X \rightarrow X$ and moreover

$$\|\mathbf{z}\|_X \leq T_1 m \beta \|\mathbf{v}\|_X.$$

Then, if $T_1 m \beta < 1$, \mathcal{P} is a strict contraction and therefore, by Theorem 6.78, equation (11.24) has a unique solution $\mathbf{v} \in X$. Moreover, the sequence

$$\mathbf{v}^{n+1} = \mathbf{G} + \mathcal{P}(\mathbf{v}^n), \quad \mathbf{v}_0 = \mathbf{0} \quad (11.25)$$

converges to \mathbf{v} in X . Thus, \mathbf{v} solves the integral system (11.23). Still, to infer that \mathbf{v} solves the problem (11.21), we need to show that \mathbf{v} possesses continuous derivatives. To

⁵ We use here the theorems on dependence from initial data for solutions of systems of ODEs. See e.g. [33], E.A. Coddington, N. Levinson, 1955.

this purpose, we introduce the narrower Banach space

$$X^1 = \{\mathbf{w} \in X : \mathbf{w}_\xi \in X\},$$

equipped with the norm $\|\mathbf{w}\|_{X^1} = \max \{\|\mathbf{w}\|_X, \|\mathbf{w}_\xi\|_X\}$. Let \mathcal{P}^1 be the restriction of \mathcal{P} to X^1 . Then $\mathcal{P}^1 : X^1 \rightarrow X^1$ and

$$\|\mathcal{P}^1(\mathbf{w})\|_{X^1} \leq \rho \|\mathbf{w}\|_{X^1},$$

where

$$\rho = T_1 m \max \{(\beta + \sup |\partial_x b_{ij}^*|) \sup |\partial_\xi x_i(t; \xi, \tau)|, \beta\}.$$

If $\rho < 1$, i.e. if T_1 is sufficiently small, depending *only* on the bounds of the coefficients of the original equation, then \mathcal{P}^1 is a strict contraction. Therefore also $\partial_\xi \mathbf{v}^n$ converges in X and the limit is $\partial_\xi \mathbf{v}$. Now, since $\mathbf{v}^n \in X^1$, we can differentiate $\mathcal{P}(\mathbf{v}^n)$ with respect to τ . We deduce that also $\partial_\tau \mathbf{v}^n$ converges in X and the limit is $\partial_\tau \mathbf{v}$. Thus \mathbf{v} solves problem (11.21) and it is clearly unique, due to the way it has been constructed.

The procedure can be iterated, by solving the Cauchy problem with initial time $t = T_1$ and initial data $\mathbf{g}(x) = \mathbf{v}(x, T_1)$. In this way we obtain an extension of the solution to the interval $0 \leq t \leq 2T_1$. After a finite number of steps we get the solution in the whole strip $\mathbb{R} \times [0, T]$. \square

11.2.3 Homogeneous systems with constant coefficients. The Riemann problem

In the particular case of *homogeneous systems*, with constant coefficients, that is when

$$\mathbf{u}_t + \mathbf{A}\mathbf{u}_x = \mathbf{0}, \quad x \in \mathbb{R}, t > 0, \quad (11.26)$$

equation (11.22) becomes

$$\frac{\partial v_k}{\partial t} + \lambda_k \frac{\partial v_k}{\partial x} = 0, \quad x \in \mathbb{R}, t > 0 \quad (11.27)$$

and its general solution is the travelling wave $v_k(x, t) = w_k(x - \lambda_k t)$, where w_k is an arbitrary and differentiable function. Then, since $\mathbf{u} = \Gamma \mathbf{v}$, the general solution of (11.26) is given by the following linear combination of travelling waves:

$$\mathbf{u}(x, t) = \sum_{k=1}^m v_k(x, t) \mathbf{r}_k = \sum_{k=1}^m w_k(x - \lambda_k t) \mathbf{r}_k. \quad (11.28)$$

Choosing $w_k = g_k^*$, we obtain the unique solution satisfying $\mathbf{u}(x, 0) = \mathbf{g}(x)$.

We can explicitly compute (11.28) in the particularly important case (Riemann problem):

$$\mathbf{g}(x) = \begin{cases} \mathbf{u}_l & x < 0 \\ \mathbf{u}_r & x > 0. \end{cases}$$

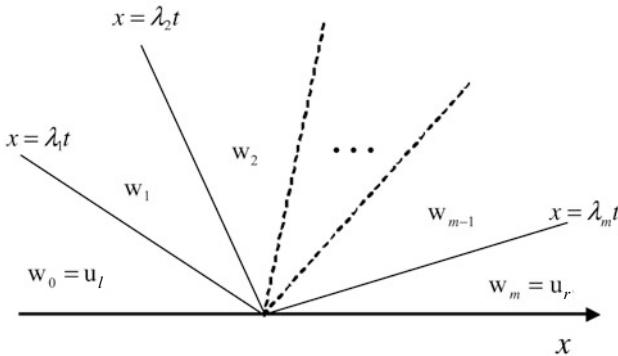


Fig. 11.1 Characteristics and solution of the Riemann problem

Since the eigenvectors $\mathbf{r}_1, \dots, \mathbf{r}_k$ constitute a basis in \mathbb{R}^m , we can write:

$$\mathbf{u}_l = \sum_{k=1}^m \alpha_k \mathbf{r}_k \quad \text{and} \quad \mathbf{u}_r = \sum_{k=1}^m \beta_k \mathbf{r}_k. \quad (11.29)$$

Then, from (11.28), each scalar v_k solves the equation (11.27) with initial data

$$v_k(x, 0) = \begin{cases} \alpha_k & x < 0 \\ \beta_k & x > 0. \end{cases}$$

Thus

$$v_k(x, t) = \begin{cases} \alpha_k & x < \lambda_k t \\ \beta_k & x > \lambda_k t. \end{cases}$$

To write the analytic expression of \mathbf{u} , we divide the half plane x, t into $m + 1$ sectors, delimited by the characteristics as in Fig. 11.1. Recall that

$$\lambda_1 < \lambda_2 < \dots < \lambda_{m-1} < \lambda_m.$$

Inside the sector $S_0 = \{x < \lambda_1 t, t > 0\}$ we have $x < \lambda_k t$ for every $k = 2, \dots, m$. Therefore, in S_0 , we have $v_k(x, t) = \alpha_k$ for every $k = 1, \dots, m$. From (11.28), (11.29) we deduce:

$$\mathbf{u}(x, t) = \sum_{k=1}^m v_k(x, t) \mathbf{r}_k = \sum_{k=1}^m \alpha_k \mathbf{r}_k = \mathbf{u}_l.$$

Similarly, inside the sector $S_m = \{x > \lambda_m t, t > 0\}$ we have $v_k(x, t) = \beta_k$ for every $k = 1, \dots, m$ and therefore $\mathbf{u} = \mathbf{u}_r$.

Let now $1 \leq j \leq m - 1$. Inside the sector

$$S_j = \{\lambda_j t < x < \lambda_{j+1} t, t > 0\}$$

we have

$$\cdots < \lambda_{j-1} < \lambda_j < \frac{x}{t} < \lambda_{j+1} < \lambda_{j+2} < \cdots, \quad (t > 0)$$

and \mathbf{u} assumes the constant value

$$\mathbf{w}_j = \sum_{k=1}^j \beta_k \mathbf{r}_k + \sum_{k=j+1}^m \alpha_k \mathbf{r}_k \quad 1 \leq j \leq m-1. \quad (11.30)$$

Summarizing, the solution of the Riemann problem is a self-similar solution of the form (Fig. 11.1):

$$\mathbf{u}(x, t) = \mathbf{w}\left(\frac{x}{t}; \mathbf{u}_l, \mathbf{u}_r\right) = \begin{cases} \mathbf{w}_0 = \mathbf{u}_l & \frac{x}{t} < \lambda_1 \\ \mathbf{w}_1 & \lambda_1 < \frac{x}{t} < \lambda_2 \\ \vdots & \vdots \\ \mathbf{w}_{m-1} & \lambda_{m-1} < \frac{x}{t} < \lambda_m \\ \mathbf{w}_m = \mathbf{u}_r & \lambda_m < \frac{x}{t}. \end{cases} \quad (11.31)$$

Example 11.5. Let $m = 3$, with initial states

$$\mathbf{u}_L = \sum_{k=1}^3 \alpha_k \mathbf{r}_k \quad \text{and} \quad \mathbf{u}_R = \sum_{k=1}^3 \beta_k \mathbf{r}_k.$$

The construction of the solution is described in Fig. 11.2. We see how the contribution of the initial data to the value of \mathbf{u} at (x, t) is carried along the characteristics through (x, t) .

Going back to the solution (11.31), if we define

$$J^-(x, t) = \{k : x < \lambda_k t\} \quad \text{and} \quad J^+(x, t) = \{k : x > \lambda_k t\},$$

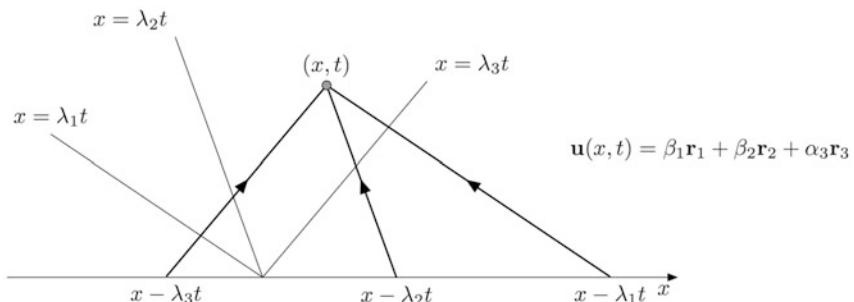


Fig. 11.2 Computation of the solution of the Riemann problem of Example 11.5, at the point (x, t)

we can write the analytic expression of \mathbf{u} in the following way:

$$\mathbf{u}(x, t) = \mathbf{u}_l + \sum_{k \in J^+(x, t)} (\beta_k - \alpha_k) \mathbf{r}_k = \mathbf{u}_r - \sum_{k \in J^-(x, t)} (\beta_k - \alpha_k) \mathbf{r}_k. \quad (11.32)$$

Now, for every $k = 1, \dots, m$, along the characteristic $x = \lambda_k t$, the solution \mathbf{u} undergoes a jump discontinuity from left to right given by

$$[\mathbf{u}(\cdot, t)]_k = \mathbf{w}_k - \mathbf{w}_{k-1} = (\beta_k - \alpha_k) \mathbf{r}_k, \quad (t > 0) \quad (11.33)$$

and hence, (11.32) expresses the value of \mathbf{u} as a superposition of these jumps. Thus, the initial discontinuity breaks into m *discontinuities*, each one propagating with its own speed λ_k , $k = 1, 2, \dots, m$. In this case we say that the states \mathbf{u}_l and \mathbf{u}_r are connected by m *contact discontinuities*⁶.

Since $\mathbf{A}\mathbf{r}_k = \lambda_k \mathbf{r}_k$, we infer that, for $t > 0$,

$$\mathbf{A}[\mathbf{u}(\cdot, t)]_k = \lambda_k (\beta_k - \alpha_k) \mathbf{r}_k = \lambda_k [\mathbf{u}(\cdot, t)]_k. \quad (11.34)$$

We shall see that the equation (11.34) corresponds to the Rankine-Hugoniot condition for the *k-contact discontinuity*.

- *Hugoniot straight lines.* The relation (11.34) holds in general only for $t > 0$, unless the initial jump $\mathbf{u}_r - \mathbf{u}_l$ is an eigenvector of \mathbf{A} . When this happens, that is if

$$\mathbf{u}_r - \mathbf{u}_l = (\beta_j - \alpha_j) \mathbf{r}_j$$

for some index j , then it must be $\beta_k = \alpha_k$ for $k \neq j$, and from (11.32) we infer that

$$\mathbf{u}(x, t) = \begin{cases} \mathbf{u}_l & \text{for } x < \lambda_j t \\ \mathbf{u}_r & \text{for } x > \lambda_j t. \end{cases}$$

Thus the solution takes the initial discontinuity and propagates it with speed λ_j . The other characteristics do not carry any jump. In this situation the states \mathbf{u}_l and \mathbf{u}_r are connected by the single j -contact discontinuity.

Changing perspective, let us fix \mathbf{u}_l and ask which states \mathbf{u} can be connected to \mathbf{u}_l (*on the right of \mathbf{u}_l*) by a single j -*contact discontinuity*, for some $j = 1, \dots, m$. From what we have just seen, this is possible if (and only if) the vectors $\mathbf{u} - \mathbf{u}_l$ are parallel to \mathbf{r}_j , that is if

$$\mathbf{A}(\mathbf{u} - \mathbf{u}_l) = \lambda_j (\mathbf{u} - \mathbf{u}_l). \quad (11.35)$$

The set of the states \mathbf{u} satisfying (11.35) coincides with the straight line

$$\mathbf{u}(s) = \mathbf{u}_l + s \mathbf{r}_j, \quad s \in \mathbb{R}$$

called the *jth-Hugoniot line issued from \mathbf{u}_l* .

⁶ According to the terminology of Subsect. 4.5.4.

Summarizing: a state \mathbf{u} can be connected to \mathbf{u}_l (*on the right of* \mathbf{u}_l) by a single contact discontinuity if and only if it belongs to one of the m Hugoniot lines issued from \mathbf{u}_l .

Remark 11.6. When the relevant domain is a quadrant, say $x > 0, t > 0$, or a half-strip $(a, b) \times (0, +\infty)$, some caution is necessary to get a well posed problem. For instance, consider the problem

$$\mathbf{u}_t + \mathbf{A}\mathbf{u}_x = \mathbf{0}, \quad x \in (0, R), t > 0 \quad (11.36)$$

with the initial condition

$$\mathbf{u}(x, 0) = \mathbf{g}(x), \quad x \in [0, R].$$

Which kind of data (and where) should be assigned to uniquely determine \mathbf{u} ? Look at the k -th equation of the uncoupled problem, that is

$$\frac{\partial v_k}{\partial t} + \lambda_k \frac{\partial v_k}{\partial x} = 0.$$

Suppose that $\lambda_k > 0$, so that the corresponding characteristic γ_k is *inflow on* $x = 0$ and *outflow on* $x = R$. Guided by the scalar case (Subsect. 4.2.4), we must assign the value of v_k *only on* $x = 0$. On the contrary, if $\lambda_k < 0$, the value of v_k has to be assigned on $x = R$. Thus, we can draw the following conclusion: suppose that r eigenvalues (say $\lambda_1, \lambda_2, \dots, \lambda_r$) are positive and the other $m - r$ eigenvalues are negative. *Then the values of* v_1, \dots, v_r *have to be assigned on* $x = 0$ *and the values of* v_{r+1}, \dots, v_m *on* $x = R$. In terms of the original unknown \mathbf{u} , this amounts to assign, on $x = 0$, r independent linear combinations of the \mathbf{u} components:

$$(\mathbf{\Gamma}^{-1}\mathbf{u})_k = \sum_{j=1}^m c^{jk} u_j \quad k = 1, 2, \dots, r,$$

while other $m - r$ have to be assigned on $x = R$.

11.3 Quasilinear Conservation Laws

11.3.1 Characteristics and Riemann invariants

In this section we consider quasilinear systems of the form

$$\mathbf{u}_t + \mathbf{F}(\mathbf{u})_x = \mathbf{0} \quad (11.37)$$

in $\mathbb{R} \times (0, +\infty)$, with $\mathbf{F} : \mathbb{R}^m \rightarrow \mathbb{R}^m$, smooth. We assume that the matrix $\mathbf{A} = D\mathbf{F}$ is strictly hyperbolic with eigenvalues

$$\lambda_1(\mathbf{u}) < \lambda_2(\mathbf{u}) < \dots < \lambda_m(\mathbf{u})$$

and corresponding eigenvectors $\mathbf{r}_1(\mathbf{u}), \dots, \mathbf{r}_m(\mathbf{u})$. Under these hypotheses, the eigenvalues and the eigenvectors are smooth functions⁷ of \mathbf{u} .

To each eigenvalue $\lambda_k(\mathbf{u})$, $k = 1, \dots, m$, we can associate a family of characteristic lines, defined by the ODE

$$\dot{x} = \lambda_k(\mathbf{u}(x, t)).$$

In the scalar case, the solution u is constant along a characteristic so that this line carries the initial value of u . If $m > 1$ this is not true anymore and we may ask if there are scalar functions $R(\mathbf{u})$ that remain constant on the characteristics. We distinguish the cases $m = 2$ and $m > 2$.

- **Case $m = 2$.** Write $\mathbf{u} = (w, v)$ and let Γ_1 be a characteristic of equation $\dot{x} = \lambda_1(\mathbf{u}(x, t))$. If $R(\mathbf{u}(x, t))$ remains constant along Γ_1 it must be

$$\frac{d}{dt}R(\mathbf{u}(x(t), t)) = \nabla R(\mathbf{u}) \cdot (\dot{x}\mathbf{u}_x + \mathbf{u}_t) = \nabla R(\mathbf{u}) \cdot (\lambda_1(\mathbf{u})\mathbf{u}_x + \mathbf{u}_t) \equiv 0. \quad (11.38)$$

Since $\mathbf{u}_t = -\mathbf{A}(\mathbf{u})\mathbf{u}_x$, (11.38) is equivalent to

$$[\nabla R(\mathbf{u}) \lambda_1(\mathbf{u}) - \nabla R(\mathbf{u}) \mathbf{A}(\mathbf{u})] \cdot \mathbf{u}_x \equiv 0. \quad (11.39)$$

Now (11.39) is satisfied if $\nabla R(\mathbf{u})$ is a *left* (row) eigenvector $\mathbf{l}_1^\top(\mathbf{u})$ of $\mathbf{A}(\mathbf{u})$. Since we are in dimension $m = 2$, this is equivalent to say that

$$\nabla R(\mathbf{u}) \cdot \mathbf{r}_2(\mathbf{u}) \equiv 0. \quad (11.40)$$

Since (11.40) involves the eigenvector \mathbf{r}_2 , R is called a *2-Riemann invariant*. Letting

$$\mathbf{r}_2(w, v) = \begin{pmatrix} r_{21}(w, v) \\ r_{22}(w, v) \end{pmatrix}$$

and $R = R(w, v)$, (11.40) may be written in the more explicit form

$$r_{21} \frac{\partial R}{\partial w} + r_{22} \frac{\partial R}{\partial v} = 0,$$

which is a first order scalar equation for R .

Similarly, a *1-Riemann invariant* is a scalar function S that remains constant along a characteristic of equation $\dot{x} = \lambda_2(\mathbf{u}(x, t))$ and hence satisfies the condition

$$\nabla S(\mathbf{u}) \cdot \mathbf{r}_1(\mathbf{u}) \equiv 0. \quad (11.41)$$

Letting

$$\mathbf{r}_1(w, v) = \begin{pmatrix} r_{11}(w, v) \\ r_{12}(w, v) \end{pmatrix}$$

⁷ See e.g. [7], F. John, 1982.

and $S = (w, v)$, we can write (11.41) in the more explicit form

$$r_{11} \frac{\partial S}{\partial w} + r_{12} \frac{\partial S}{\partial v} = 0.$$

- **Case $m > 2$.** Changing slightly point of view, the equations (11.40) and (11.41) express the fact that the Riemann invariants S and R remain constant *along the integral curves of the eigenvectors \mathbf{r}_1 and \mathbf{r}_2* , respectively. In this form, Riemann invariants can be defined also for $m > 2$. Thus, in general, we can give the following definition.

Definition 11.7. A smooth function $w_k : \mathbb{R}^m \rightarrow \mathbb{R}$ is called a k -Riemann invariant if it satisfies the following first order PDE:

$$\nabla w_k(\mathbf{u}) \cdot \mathbf{r}_k(\mathbf{u}) \equiv 0. \quad (11.42)$$

If $\xi \mapsto \mathbf{v}(\xi)$ is an integral curve of \mathbf{r}_k , that is

$$\frac{d}{d\xi} \mathbf{v}(\xi) = \mathbf{r}_k(\mathbf{v}(\xi)),$$

then

$$\frac{d}{d\xi} w_k(\mathbf{v}(\xi)) = \nabla w_k(\mathbf{v}(\xi)) \cdot \frac{d}{d\xi} \mathbf{v}(\xi) = \nabla w_k(\mathbf{v}(\xi)) \cdot \mathbf{r}_k(\mathbf{v}(\xi)) \equiv 0$$

and therefore a k -Riemann invariant is constant along the trajectories of \mathbf{r}_k .

In general, the Riemann invariants exist only locally. In fact, we have⁸:

Proposition 11.8. For every $k = 1, \dots, m$, there exist, locally, $m - 1$ independent k -Riemann invariants⁹.

Example 11.9. p -system. It is easy to check that

$$R(w, v) = v + \int_{w_0}^w \sqrt{a(s)} ds, \quad S(w, v) = v - \int_{w_0}^w \sqrt{a(s)} ds$$

($w_0 > 0$, arbitrary) are Riemann invariants.

Example 11.10. Gas dynamics. Since $m = 3$, we have 3 pairs of invariants.

The pair of 1-invariants $w_{1j} = w_{1j}(\rho, v, s)$, $j = 1, 2$, can be found solving the first order PDE

$$\nabla w \cdot \mathbf{r}_1 = \rho w_\rho - c w_v = 0.$$

⁸ See e.g. [18], J. Smoller, 1983.

⁹ That is whose gradients are linearly independent.

The characteristic system for this equation can be written, formally, as (see Chap. 4, Remark 4.22, p. 240)

$$\frac{d\rho}{\rho} = -\frac{dv}{c} = \frac{ds}{0}. \quad (11.43)$$

From the last equation we have $ds = 0$. Hence $w_{11}(\rho, v, s) = s$ is a first integral of the system (11.43) and therefore $w_{11} = s$ is a 1-Riemann invariant. From the first equation we get ($c^2 = p_\rho$)

$$-dv = \frac{\sqrt{p_\rho}}{\rho} d\rho.$$

Then, another first integral of (11.43) is

$$w_{12}(\rho, v, s) = v + l(\rho, s)$$

where $l(\rho, s) = \int \frac{\sqrt{p_\rho(\rho, s)}}{\rho} d\rho$. Thus $w_{12} = v + l(\rho, s)$ is the second 1-Riemann invariant. Similar computations give $w_{31}(\rho, v, s) = s$ and $w_{32}(\rho, v, s) = v - l(\rho, s)$ for the pair of 3-invariants.

To compute the 2-invariants $w_{2j} = w_{2j}(\rho, v, s)$, $j = 1, 2$, we solve the PDE

$$\nabla w \cdot \mathbf{r}_2 = p_s w_\rho - c^2 w_s = 0$$

whose characteristic system reads

$$\frac{d\rho}{p_s} = \frac{dv}{0} = -\frac{ds}{p_\rho}.$$

Thus $dv = 0$ and $p_\rho d\rho + p_s ds = dp = 0$ from which we find the two 2-invariants

$$w_{21}(\rho, v, s) = v \quad \text{and} \quad w_{22}(\rho, v, s) = p.$$

Summarizing, the three pairs of invariants are:

$$\{s, v + l\}, \quad \{v, p\}, \quad \{s, v - l\}.$$

The Riemann invariants can be employed to find the explicit solution in problems of practical interest in the context of classical fluid dynamics, like for instance the Riemann problem in gas dynamics, for which we refer e.g. to [45], *Godlewski-Raviart, 1996*.

11.3.2 Weak (or integral) solutions and the Rankine-Hugoniot condition

The definition of *weak* or *integral* solution of the Cauchy problem

$$\begin{cases} \mathbf{u}_t + \mathbf{F}(\mathbf{u})_x = \mathbf{0} & \text{in } \mathbb{R} \times (0, +\infty) \\ \mathbf{u}(x, 0) = \mathbf{g}(x) & \text{in } \mathbb{R} \end{cases} \quad (11.44)$$

parallels the analogous definition for the scalar case.

Definition 11.11. We say that a locally bounded function $\mathbf{u} : \mathbb{R} \times [0, +\infty) \rightarrow \mathbb{R}^m$ is a weak solution of problem (11.44) if the equation

$$\int_0^\infty dt \int_{\mathbb{R}} [\mathbf{v}_t \cdot \mathbf{u} + \mathbf{v}_x \cdot \mathbf{F}(\mathbf{u})] dx + \int_{\mathbb{R}} \mathbf{v}(x, 0) \cdot \mathbf{g}(x) dx = 0 \quad (11.45)$$

holds for every smooth $\mathbf{v} : \mathbb{R} \times [0, +\infty) \rightarrow \mathbb{R}^m$, vanishing outside a compact subset of the halfplane $\mathbb{R} \times [0, +\infty)$.

As in the scalar case, a weak solution of class $C^1(\mathbb{R} \times [0, +\infty))$ satisfies the Cauchy problem in classical sense. Moreover, in the class of weak piecewise C^1 solutions (see Subsect. 4.4.3), the only admissible discontinuities in the half plane $t > 0$ are those prescribed by the following vectorial Rankine-Hugoniot condition valid along a discontinuity line Γ of equation $x = s(t)$:

$$\sigma [\mathbf{u}_+ - \mathbf{u}_-] = [\mathbf{F}(\mathbf{u}_+) - \mathbf{F}(\mathbf{u}_-)] . \quad (11.46)$$

Condition (11.46) consists of m scalar equations, where \mathbf{u}_+ and \mathbf{u}_- are the values that \mathbf{u} attains on the right and left sides of Γ and $\sigma = \dot{s}(t)$ is the speed of the discontinuity. The proof is identical to the proof of Theorem 4.7, p. 205.

11.4 The Riemann Problem

Our purpose is to construct a solution of the system

$$\mathbf{u}_t + \mathbf{F}(\mathbf{u})_x = \mathbf{0} \quad (11.47)$$

in $\mathbb{R} \times (0, +\infty)$, satisfying the initial condition

$$\mathbf{g}(x, 0) = \begin{cases} \mathbf{u}_l & \text{for } x < 0 \\ \mathbf{u}_r & \text{for } x > 0, \end{cases} \quad (11.48)$$

where \mathbf{u}_l and \mathbf{u}_r are constant states. We always assume that \mathbf{F} is a smooth function in \mathbb{R}^m . Let us briefly review the scalar case $m = 1$,

$$u_t + q(u)_x = 0, \quad x \in \mathbb{R}, t > 0. \quad (11.49)$$

We have seen in Sect. 4.5 that the Riemann problem for equation (11.49) has a unique *entropic* solution, for which we can write an explicit formula. Suppose

$$q \in C^2(\mathbb{R}), \quad q''(u) > 0,$$

so that the equation is *genuinely nonlinear*. The case q concave is similar.

If $u_l < u_r$, the solution u is constructed connecting the two states u_l and u_r through a *rarefaction wave*, that is:

$$u(x, t) = \begin{cases} u_l & x < q'(u_l) t \\ f\left(\frac{x}{t}\right) & q'(u_l) t < x < q'(u_r) t \\ u_r & x > q'(u_r) t \end{cases}$$

where $f = (q')^{-1}$, i.e. the inverse of q' .

If $u_l > u_r$, then u is constructed connecting the two states u_l and u_r through a *shock wave*, that is

$$u(x, t) = \begin{cases} u_l & x < \sigma(u_l, u_r) t \\ u_r & x > \sigma(u_l, u_r) t \end{cases} \quad (11.50)$$

where $\sigma = \sigma(u_l, u_r)$ is the shock speed, given by the *Rankine-Hugoniot* condition

$$\sigma(u_l, u_r) = \frac{q(u_r) - q(u_l)}{u_r - u_l}.$$

Recall that the *entropy* condition, imposed to get rid of non-physically acceptable solutions, implies the inequalities

$$q'(u_r) < \sigma(u_l, u_r) < q'(u_l). \quad (11.51)$$

Geometrically, (11.51) means that the characteristics “impinge” into the shock curve from both sides.

On the other hand, in the linear case $u_t + au_x = 0$, the solution is given by

$$u(x, t) = \begin{cases} u_l & x < at \\ u_r & x > at \end{cases}$$

exhibiting a discontinuity line parallel to the characteristics (*contact discontinuity* (Fig. 11.3)).

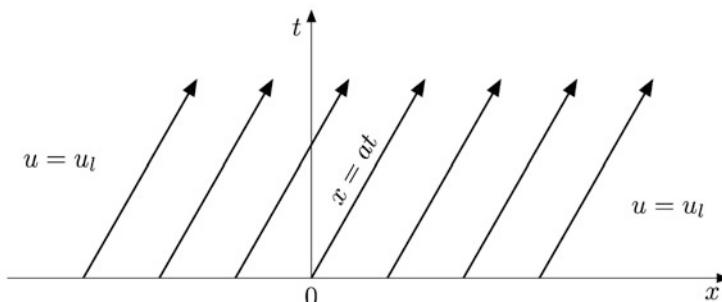


Fig. 11.3 Contact discontinuity

Consider now the m -dimensional case, $m > 1$. Guided by the case of linear systems with constant coefficients, we expect that the initial discontinuity breaks down, in general, into m waves, each one propagating with its own speed. In the nonlinear case, however, these waves are of different kinds, namely rarefaction waves, shocks and contact discontinuities. Thus, we first introduce and construct these special solutions. Then, to solve the Riemann problem, we first examine which pairs of constant states \mathbf{u}_l and \mathbf{u}_r can be connected by a single rarefaction wave or by a single shock wave or by a single contact discontinuity. Finally, we obtain a solution in the general case by a mix of those waves.

11.4.1 Rarefaction curves and waves. Genuinely nonlinear systems

We start by examining the existence of solutions to (11.47) of the type

$$\mathbf{u}(x, t) = \mathbf{v}(\theta(x, t)), \quad (11.52)$$

where θ is a scalar function. These solutions are called *simple waves*. Inserting (11.52) into (11.47), we find the equation

$$\mathbf{v}'(\theta)\theta_t + \mathbf{A}(\mathbf{v}(\theta))\mathbf{v}'(\theta)\theta_x = \mathbf{0}$$

that, assuming $\theta_x \neq 0$, we rewrite as

$$\mathbf{A}(\mathbf{v}(\theta))\mathbf{v}'(\theta) = -\frac{\theta_t}{\theta_x}\mathbf{v}'(\theta). \quad (11.53)$$

Disregarding the trivial case $\mathbf{v}'(\theta) = \mathbf{0}$, it is apparent that (11.53) is satisfied if

$$\mathbf{A}(\mathbf{v}(\theta))\mathbf{v}'(\theta) = \lambda(\mathbf{v}(\theta))\mathbf{v}'(\theta),$$

where $\theta = \theta(x, t)$ is a solution of the scalar equation

$$\theta_t + \lambda(\mathbf{v}(\theta))\theta_x = 0. \quad (11.54)$$

In other terms, (11.53) is satisfied if $\mathbf{v}'(\theta)$ coincides with an *eigenvector* $\mathbf{r}_k(\mathbf{v}(\theta))$ of $\mathbf{A}(\mathbf{v}(\theta))$, with corresponding *eigenvalue* $\lambda_k(\mathbf{v}(\theta))$. Note that if $\mathbf{v}'(\theta) \neq \mathbf{0}$ in a given interval (θ_0, θ_1) , the index k does not depend on θ , since the eigenvalues are distinct and $\theta \mapsto \lambda_k(\mathbf{v}(\theta))$ is a regular function. Thus we can use the notation $\mathbf{v}_k(\theta)$ and write

$$\mathbf{v}'_k(\theta) = \mathbf{r}_k(\mathbf{v}_k(\theta)). \quad (11.55)$$

The equation (11.55) means that $\mathbf{v}_k = \mathbf{v}_k(\theta)$ is an integral curve of the vector field \mathbf{r}_k . Given a state $\mathbf{u}_0 \in \mathbb{R}^m$, we denote by $R_k(\mathbf{u}_0)$ the integral curve of \mathbf{r}_k satisfying the Cauchy condition $\mathbf{v}_k(\theta_0) = \mathbf{u}_0$, for some θ_0 .

Due to the smoothness of \mathbf{r}_k , $\mathbf{v}_k(\theta)$ is uniquely defined, at least in a neighborhood of θ_0 , and we can solve for $\theta = \theta(x, t)$ the equation (11.54), which becomes

$$\theta_t + \lambda_k(\mathbf{v}_k(\theta)) \theta_x = 0. \quad (11.56)$$

Setting

$$q_k(\theta) = \int_{\theta_0}^{\theta} \lambda_k(\mathbf{v}_k(s)) ds,$$

equation (11.56) can be written in the form

$$\theta_t + q_k(\theta)_x = 0 \quad (11.57)$$

which is a scalar conservation law. Since

$$q'_k(\theta) = \lambda_k(\mathbf{v}_k(\theta))$$

and

$$q''_k(\theta) = \nabla \lambda_k(\mathbf{v}_k(\theta)) \cdot \mathbf{v}'_k(\theta) = \nabla \lambda_k(\mathbf{v}_k(\theta)) \cdot \mathbf{r}_k(\mathbf{v}(\theta)),$$

we distinguish some cases according to the following definition, where by k^{th} -characteristic field we mean the eigenvalue-eigenvector pair $\lambda_k(\mathbf{u}), \mathbf{r}_k(\mathbf{u})$.

Definition 11.12. We say that:

a) The k^{th} -characteristic field $\lambda_k(\mathbf{u}), \mathbf{r}_k(\mathbf{u})$ is genuinely nonlinear if

$$\nabla \lambda_k(\mathbf{u}) \cdot \mathbf{r}_k(\mathbf{u}) \neq 0, \quad \text{for all } \mathbf{u} \in \mathbb{R}^m. \quad (11.58)$$

b) The system is genuinely nonlinear if

$$\nabla \lambda_k(\mathbf{u}) \cdot \mathbf{r}_k(\mathbf{u}) \neq 0, \quad \text{for all } \mathbf{u} \in \mathbb{R}^m \text{ and } k = 1, \dots, m.$$

c) The k^{th} -characteristic field $\lambda_k(\mathbf{u}), \mathbf{r}_k(\mathbf{u})$ is linearly degenerate if

$$\nabla \lambda_k(\mathbf{u}) \cdot \mathbf{r}_k(\mathbf{u}) = 0, \quad \text{for all } \mathbf{u} \in \mathbb{R}^m. \quad (11.59)$$

Going back to the construction of a simple wave, assume that the pair $\lambda_k(\mathbf{u}), \mathbf{r}_k(\mathbf{u})$ is genuinely nonlinear. Then the function

$$\theta \mapsto q'_k(\theta) = \lambda_k(\mathbf{v}_k(\theta)) \quad (11.60)$$

is invertible. Setting $f_k = (q'_k)^{-1}$, a solution of (11.57) is given by

$$\theta_k(x, t) = f_k\left(\frac{x}{t}\right).$$

Thus, from (11.52) we obtain *the simple wave*

$$\mathbf{u}_k(x, t) = \mathbf{v}_k\left(f_k\left(\frac{x}{t}\right)\right). \quad (11.61)$$

Note that if $\mathbf{v}_k(\theta) = \mathbf{u}$ then $\mathbf{u}_k = \mathbf{u}$ on the half straight line $x = \lambda_k(\mathbf{u})t$, $t > 0$.

Definition 11.13. *The simple wave \mathbf{u}_k is called k -rarefaction wave centered at $(0, 0)$, through \mathbf{u}_0 .*

Summarizing, given a genuinely nonlinear characteristic field $\lambda_k(\mathbf{v})$, $\mathbf{r}_k(\mathbf{v})$, we may construct a k -rarefaction wave centered at the origin and equal to \mathbf{u}_0 on the half straight line $x = \lambda_k(\mathbf{u}_0)t$, $t > 0$, by first solving the Cauchy problem

$$\begin{cases} \mathbf{v}'_k(\theta) = \mathbf{r}_k(\mathbf{v}_k(\theta)) \\ \mathbf{v}_k(\theta_0) = \mathbf{u}_0. \end{cases} \quad (11.62)$$

Once $\mathbf{v}_k = \mathbf{v}_k(\theta)$ is known, we compute f_k , i.e. the inverse function of

$$\theta \mapsto \lambda_k(\mathbf{v}_k(\theta)),$$

and then (11.61), obtaining

$$\mathbf{u}_k(x, t) = \mathbf{v}_k\left(f_k\left(\frac{x}{t}\right)\right).$$

Remark 11.14. Let $(\lambda_k(\mathbf{u}), \mathbf{r}_k(\mathbf{u}))$ be linearly degenerate. Then, $w = \lambda_k(\mathbf{u})$ is a k -Riemann invariant by definition.

Example 11.15. The p -system. From (11.16) and (11.17), we have:

$$\nabla \lambda_1(\mathbf{u}) \cdot \mathbf{r}_1(\mathbf{u}) = \frac{-a'(w)}{2\sqrt{a(w)}} > 0, \quad \nabla \lambda_2(\mathbf{u}) \cdot \mathbf{r}_2(\mathbf{u}) = \frac{a'(w)}{2\sqrt{a(w)}} < 0$$

for all (w, v) , $w > 0$. Hence the p -system is genuinely nonlinear.

Example 11.16. Gas dynamics. From (11.13) and (11.14), we have:

$$\nabla \lambda_1 = \begin{pmatrix} -c_\rho \\ 1 \\ -c_s \end{pmatrix}, \quad \nabla \lambda_2 = \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix}, \quad \nabla \lambda_3 = \begin{pmatrix} c_\rho \\ 1 \\ c_s \end{pmatrix}$$

whence

$$\begin{aligned} \nabla \lambda_1 \cdot \mathbf{r}_1 &= -(c + \rho c_\rho) \\ \nabla \lambda_2 \cdot \mathbf{r}_2 &\equiv 0 \\ \nabla \lambda_3 \cdot \mathbf{r}_3 &= (c + \rho c_\rho). \end{aligned}$$

The pairs λ_1, \mathbf{r}_1 and λ_3, \mathbf{r}_3 are *genuinely nonlinear*, while λ_2, \mathbf{r}_2 is linearly degenerate.

11.4.2 Solution of the Riemann problem by a single rarefaction wave

We are now ready to explore the possibility to solve problem (11.47), (11.48) by means of a single rarefaction wave. In other words, we look for the pairs of states \mathbf{u}_l , \mathbf{u}_r that can be connected by a rarefaction wave. To this purpose we examine the structure of the integral curves $R_k(\mathbf{u}_l)$, $k = 1, \dots, m$. If $\lambda_k(\mathbf{u}_l), \mathbf{r}_k(\mathbf{u}_l)$ is genuinely nonlinear, $R_k(\mathbf{u}_l)$ takes the name of k -rarefaction curve issued from \mathbf{u}_l .

Theorem 11.17. *Let $\lambda_k(\mathbf{u}), \mathbf{r}_k(\mathbf{u})$ be genuinely nonlinear, $1 \leq k \leq m$. Near a given a state \mathbf{u}_0 , there exists a parametrization of the curve $R_k(\mathbf{u}_0)$ given by*

$$\varepsilon \mapsto \varphi_k(\varepsilon; \mathbf{u}_0), \quad |\varepsilon| \leq \varepsilon_0,$$

for some $\varepsilon_0 > 0$, such that

$$\varphi_k(\varepsilon; \mathbf{u}_0) = \mathbf{u}_0 + \varepsilon \mathbf{r}_k(\mathbf{u}_0) + \frac{\varepsilon^2}{2} D\mathbf{r}_k(\mathbf{u}_0) \cdot \mathbf{r}_k(\mathbf{u}_0) + o(\varepsilon^2). \quad (11.63)$$

Moreover, we can decompose $R_k(\mathbf{u}_0)$ into the following three disjoint sets:

$$R_k(\mathbf{u}_0) = R_k^+(\mathbf{u}_0) \cup \{\mathbf{u}_0\} \cup R_k^-(\mathbf{u}_0) \quad (11.64)$$

where

$$R_k^+(\mathbf{u}_0) = \{\mathbf{u} \in R_k(\mathbf{u}_0) : \lambda_k(\mathbf{u}) > \lambda_k(\mathbf{u}_0)\} \quad (11.65)$$

and

$$R_k^-(\mathbf{u}_0) = \{\mathbf{u} \in R_k(\mathbf{u}_0) : \lambda_k(\mathbf{u}) < \lambda_k(\mathbf{u}_0)\}. \quad (11.66)$$

Proof. Normalize $\mathbf{r}_k(\mathbf{u})$ to have

$$\nabla \lambda_k(\mathbf{u}) \cdot \mathbf{r}_k(\mathbf{u}) \equiv 1. \quad (11.67)$$

Let $\theta \mapsto \mathbf{v}_k(\theta)$ be the solution of the Cauchy problem (11.62) with $\theta_0 = \lambda_k(\mathbf{u}_0)$:

$$\begin{cases} \mathbf{v}'_k(\theta) = \mathbf{r}_k(\mathbf{v}_k(\theta)) \\ \mathbf{v}_k(\lambda_k(\mathbf{u}_0)) = \mathbf{u}_0. \end{cases}$$

The function $\mathbf{v}_k = \mathbf{v}_k(\theta)$ exists at least in some interval $\lambda_k(\mathbf{u}_0) - \varepsilon_0 \leq \theta \leq \lambda_k(\mathbf{u}_0) + \varepsilon_0$, $\varepsilon_0 > 0$. Using (11.67), we write:

$$\frac{d}{d\theta} \lambda_k(\mathbf{v}_k(\theta)) = \nabla \lambda_k(\mathbf{v}_k(\theta)) \cdot \mathbf{r}_k(\mathbf{v}_k(\theta)) \equiv 1.$$

In particular, note that $\theta \mapsto \lambda_k(\mathbf{v}_k(\theta))$ is strictly increasing. Integrating from $\lambda_k(\mathbf{u}_0)$ to θ and using $\mathbf{v}_k(\lambda_k(\mathbf{u}_0)) = \mathbf{u}_0$, we find

$$\lambda_k(\mathbf{v}_k(\theta)) = \theta.$$

Therefore, near \mathbf{u}_0 , we may write

$$R_k(\mathbf{u}_0) = \{\mathbf{v}_k(\theta) : \lambda_k(\mathbf{u}_0) - \varepsilon_0 \leq \theta \leq \lambda_k(\mathbf{u}_0) + \varepsilon_0\}.$$

Put $\theta = \lambda_k(\mathbf{u}_0) + \varepsilon$ and

$$\varphi_k(\varepsilon; \mathbf{u}_0) = \mathbf{v}_k(\lambda_k(\mathbf{u}_0) + \varepsilon), \quad |\varepsilon| \leq \varepsilon_0.$$

We have $\varphi_k(0; \mathbf{u}_0) = \mathbf{v}_k(\lambda_k(\mathbf{u}_0)) = \mathbf{u}_0$ and

$$\frac{d}{d\varepsilon}\varphi_k(\varepsilon; \mathbf{u}_0)|_{\varepsilon=0} = \mathbf{r}_k(\mathbf{v}_k(\lambda_k(\mathbf{u}_0))) = \mathbf{r}_k(\mathbf{u}_0). \quad (11.68)$$

Moreover, since

$$\frac{d^2}{d\theta^2}\mathbf{v}_k(\theta) = D\mathbf{r}_k(\mathbf{v}_k(\theta)) \frac{d}{d\theta}\mathbf{v}_k(\theta) = D\mathbf{r}_k(\mathbf{v}_k(\theta)) \mathbf{r}_k(\mathbf{v}_k(\theta)),$$

we deduce that

$$\frac{d^2}{d\varepsilon^2}\varphi_k(\varepsilon; \mathbf{u}_0)|_{\varepsilon=0} = D\mathbf{r}_k(\mathbf{u}_0) \mathbf{r}_k(\mathbf{u}_0). \quad (11.69)$$

From (11.68) and (11.69), (11.63) follows. Finally, since $\lambda_k(\mathbf{v}_k(\theta)) = \theta$, we can write

$$R_k(\mathbf{u}_0) = \{\mathbf{v}_k(\theta) : \lambda_k(\mathbf{u}_0) - \varepsilon_0 \leq \lambda_k(\mathbf{v}_k(\theta)) \leq \lambda_k(\mathbf{u}_0) + \varepsilon_0\}. \quad (11.70)$$

To get the decomposition (11.64) define

$$R_k^-(\mathbf{u}_0) = \{\mathbf{v}(\theta) : \lambda_k(\mathbf{u}_0) - \varepsilon_0 \leq \lambda_k(\mathbf{v}_k(\theta)) < \lambda_k(\mathbf{u}_0)\}$$

and

$$R_k^+(\mathbf{u}_0) \{\mathbf{v}(\theta) : \lambda_k(\mathbf{u}_0) < \lambda_k(\mathbf{v}_k(\theta)) \leq \lambda_k(\mathbf{u}_0) + \varepsilon_0\}. \quad \square$$

Remark 11.18. We emphasize that, under the normalization condition (11.67), the part $R_k^+(\mathbf{u}_0)$ of $R_k(\mathbf{u}_0)$ corresponds to the positive values of ε , in the parametrization (11.63).

Let now $\mathbf{u}_0 = \mathbf{u}_l$ be the left initial state in the Riemann problem, i.e. assigned on $x < 0$. By the structure Theorem 11.17 we can write

$$R_k(\mathbf{u}_l) = R_k^+(\mathbf{u}_l) \cup \{\mathbf{u}_l\} \cup R_k^-(\mathbf{u}_l)$$

with

$$R_k^+(\mathbf{u}_l) = \{\mathbf{u} \in R_k(\mathbf{u}_l) : \lambda_k(\mathbf{u}) > \lambda_k(\mathbf{u}_l)\}. \quad (11.71)$$

The next theorem shows that if $\mathbf{u}_r \in R_k^+(\mathbf{u}_l)$, then the Riemann problem can be solved by a single k -rarefaction wave.

Theorem 11.19. Assume that for some k , $1 \leq k \leq m$:

- i) The pair $\lambda_k(\mathbf{u}), \mathbf{r}_k(\mathbf{u})$ is genuinely nonlinear.
- ii) $\mathbf{u}_r \in R_k^+(\mathbf{u}_l)$.

Then there exists a weak solution of the Riemann problem with initial states \mathbf{u}_l , \mathbf{u}_r , given by a rarefaction wave.

Proof. Since $\mathbf{u}_r \in R_k^+(\mathbf{u}_l)$, there exist two numbers θ_r and θ_l such that

$$\mathbf{u}_l = \mathbf{v}_k(\theta_l) \quad \text{and} \quad \mathbf{u}_r = \mathbf{v}_k(\theta_r).$$

Assume that $\theta_l < \theta_r$. The opposite case is similar. From ii) we deduce $\lambda_k(\mathbf{u}_r) > \lambda_k(\mathbf{u}_l)$. From (11.60), p. 634, we get

$$q'_k(\theta_r) = \lambda_k(\mathbf{u}_r) \quad \text{and} \quad q'_k(\theta_l) = \lambda_k(\mathbf{u}_l),$$

so that $q'_k(\theta_l) < q'_k(\theta_r)$. For identical reasons, q'_k is strictly increasing in the interval $[\theta_l, \theta_r]$ and therefore q_k is strictly convex. Then we can solve the scalar equation (11.57), that is

$$\theta_t + (q_k(\theta))_x = 0,$$

with initial data

$$g(x, 0) = \begin{cases} \theta_l & x < 0 \\ \theta_r & x > 0. \end{cases}$$

The solution is given by the formula

$$\theta_k(x, t) = \begin{cases} \theta_l & x < \lambda_k(\mathbf{u}_l)t \\ f_k\left(\frac{x}{t}\right) & \lambda_k(\mathbf{u}_l)t < x < \lambda_k(\mathbf{u}_r)t \\ \theta_r & x > \lambda_k(\mathbf{u}_r)t \end{cases}$$

where $f_k = (q'_k)^{-1}$. Then, the k -rarefaction wave

$$\mathbf{u}_k(x, t) = \mathbf{v}_k(\theta_k(x, t)) = \begin{cases} \mathbf{u}_l & x < \lambda_k(\mathbf{u}_l)t \\ \mathbf{v}_k\left(f_k\left(\frac{x}{t}\right)\right) & \lambda_k(\mathbf{u}_l)t < x < \lambda_k(\mathbf{u}_r)t \\ \mathbf{u}_r & x > \lambda_k(\mathbf{u}_r)t \end{cases} \quad (11.72)$$

is a solution of the Riemann problem with initial states $\mathbf{u}_l, \mathbf{u}_r$. □

11.4.3 Lax entropy condition. Shock waves and contact discontinuities

We have seen in Subsect. 11.3.2 that, along a jump discontinuity Γ , the conservation law for a piecewise continuous solution translates into the Rankine-Hugoniot conditions

$$\sigma [\mathbf{u}_r - \mathbf{u}_l] = [\mathbf{F}(\mathbf{u}_r) - \mathbf{F}(\mathbf{u}_l)] \quad \text{on } \Gamma, \quad (11.73)$$

where σ is the speed along Γ and \mathbf{u}_l and \mathbf{u}_r denote the values that \mathbf{u} attains on the shock line Γ from the left and the right sides, respectively. We also know from the scalar case, that these conditions are not enough to select a physically meaningful integral solution. Here an entropy criterion comes into play. We will limit ourselves to the so called *Lax entropy condition*, direct generalization of (4.62), p. 210. We will say that a shock wave is *admissible* or *entropic* if it satisfies this condition.

To introduce it, let us consider the situation in the *scalar* case from a different (heuristic) perspective. In order to uniquely determine a shock, we need to know the values u_l, u_r and σ . The Rankine-Hugoniot condition gives an equation for

these three unknowns. Now, in the case of strictly convex/concave flux function q , the entropy condition requires that

$$q'(u_r) < \sigma(u_l, u_r) < q'(u_l)$$

prescribing that the characteristics *impinge into* Γ , both from the left and the right sides. In other words, the characteristics carry into Γ the values of u_l and u_r from the initial data, respectively. Thus, in principle, we have the correct number of information to select in a unique way u_l , u_r , σ on Γ .

In the vectorial case, the Rankine-Hugoniot conditions (11.73) provide m scalar equations for the speed σ and the $2m$ components of \mathbf{u}_l and \mathbf{u}_r , that is, $2m + 1$ scalars. Somehow we are missing $m + 1$ information. Where do we find them? Assume that, for some k , $1 \leq k \leq m$, the k^{th} -characteristic field $\lambda_k(\mathbf{u}_r)$, $r_k(\mathbf{u})$ is genuinely nonlinear and that,

$$\lambda_k(\mathbf{u}_r) < \sigma < \lambda_{k+1}(\mathbf{u}_r).$$

This means that k characteristics impinge Γ from the *right*, thus providing k information on the values of \mathbf{u}_r .

Similarly, assume that, for some j , $1 \leq j \leq m$,

$$\lambda_j(\mathbf{u}_l) < \sigma < \lambda_{j+1}(\mathbf{u}_l).$$

This means that $m - j$ characteristics impinge Γ from the *left*, thus providing $m - j$ information on the values of \mathbf{u}_l . We need that

$$k + (m - j) = m + 1$$

from which

$$j = k - 1.$$

In conclusion, a possible generalization of the entropy condition is the following:
for some index k it holds

$$\lambda_k(\mathbf{u}_r) < \sigma < \lambda_{k+1}(\mathbf{u}_r) \quad (11.74)$$

and

$$\lambda_{k-1}(\mathbf{u}_l) < \sigma < \lambda_k(\mathbf{u}_l). \quad (11.75)$$

We may rewrite the last two relations in the following form:

$$\lambda_k(\mathbf{u}_r) < \sigma < \lambda_k(\mathbf{u}_l) \quad (11.76)$$

and

$$\lambda_{k-1}(\mathbf{u}_l) < \sigma < \lambda_{k+1}(\mathbf{u}_r). \quad (11.77)$$

Condition (11.76) implies that the shock speed σ is intermediate between the speeds of the k^{th} characteristics coming from the right and the left sides of Γ . From (11.77) we infer that this happens for the index k **only**.

Observe that if $m = 1$, (11.77) is empty and (11.76) reduces to the (11.51). If $k = 1$ or $k = m$, (11.77) becomes, respectively

$$\sigma < \lambda_2(\mathbf{u}_r) \quad \text{or} \quad \lambda_{m-1}(\mathbf{u}_l) < \sigma.$$

Definition 11.20. *If the k^{th} characteristic field $\lambda_k(\mathbf{u})$, $r_k(\mathbf{u})$ is genuinely nonlinear and the entropy conditions (11.76), (11.77) hold, we say that the discontinuity is a k -shock.*

We will see that, if the pair $\lambda_k(\mathbf{u})$, $r_k(\mathbf{u})$ is linearly degenerate, then

$$\lambda_k(\mathbf{u}_r) = \sigma = \lambda_k(\mathbf{u}_l)$$

and the entropy inequality is satisfied in a weaker sense. We shall call this type of discontinuity a *k-contact discontinuity*.

11.4.4 Solution of the Riemann problem by a single k -shock

We now explore the possibility to solve problem (11.47), (11.48) by means of a single shock wave. In other words, we look for the pairs of states \mathbf{u}_l , \mathbf{u}_r that can be connected by a k -shock. Guided by the linear case in Subsect. 11.2.3, we introduce the following set.

Definition 11.21. *Given $\mathbf{u}_0 \in \mathbb{R}^m$, the Rankine-Hugoniot locus of \mathbf{u}_0 , denoted by $S(\mathbf{u}_0)$, is the set of the states $\mathbf{u} \in \mathbb{R}^m$ for which there exist a scalar $\sigma = \sigma(\mathbf{u}, \mathbf{u}_0)$ such that*

$$\mathbf{F}(\mathbf{u}) - \mathbf{F}(\mathbf{u}_0) = \sigma(\mathbf{u}, \mathbf{u}_0)(\mathbf{u} - \mathbf{u}_0). \quad (11.78)$$

For \mathbf{u} close to \mathbf{u}_0 , we can write

$$\mathbf{F}(\mathbf{u}) - \mathbf{F}(\mathbf{u}_0) \sim D\mathbf{F}(\mathbf{u}_0)(\mathbf{u} - \mathbf{u}_0),$$

so that $\sigma(\mathbf{u}, \mathbf{u}_0) \sim \lambda(\mathbf{u}_0)$, with $\lambda(\mathbf{u}_0)$ eigenvalue of $D\mathbf{F}(\mathbf{u}_0)$. On the other hand, in the linear case, $S(\mathbf{u}_0)$ coincides with the union of m Hugoniot straight lines issued from \mathbf{u}_0 , each one directed as one of the eigenvectors \mathbf{r}_k . Thus, in the non-linear case, we expect that $S(\mathbf{u}_0)$ is given by the union of m regular curves, each one tangent at \mathbf{u}_0 to an eigenvector of $D\mathbf{F}(\mathbf{u}_0)$. Precisely, the following theorem holds¹⁰.

Theorem 11.22. *Let $\mathbf{u}_0 \in \mathbb{R}^m$. Near \mathbf{u}_0 , $S(\mathbf{u}_0)$ consists of the union of m curves $S_k(\mathbf{u}_0)$ of class C^2 . Moreover, for each $k = 1, 2, \dots, m$, there exists a parametrization of $S_k(\mathbf{u}_0)$ given by*

$$\varepsilon \mapsto \psi_k(\varepsilon; \mathbf{u}_0), \quad |\varepsilon| \leq \varepsilon_0$$

¹⁰ For the proof, see e.g. [45], Godlewsky-Raviart, 1996.

for some $\varepsilon_0 > 0$, with the following properties:

- i) $\psi_k(\varepsilon; \mathbf{u}_0) = \mathbf{u}_0 + \varepsilon \mathbf{r}_k(\mathbf{u}_0) + \frac{\varepsilon^2}{2} D\mathbf{r}_k(\mathbf{u}_0) \mathbf{r}_k(\mathbf{u}_0) + o(\varepsilon^2)$.
- ii) $\sigma(\psi_k(\varepsilon; \mathbf{u}_0), \mathbf{u}_0) = \lambda_k(\mathbf{u}_0) + \frac{\varepsilon}{2} \nabla \lambda_k(\mathbf{u}_0) \cdot \mathbf{r}_k(\mathbf{u}_0) + o(\varepsilon)$.

Remark 11.23. Theorems 11.17 and 11.22 imply that the curves $S_k(\mathbf{u}_0)$ and $R_k(\mathbf{u}_0)$ have a second order contact at \mathbf{u}_0 .

When the k^{th} -characteristic field $\lambda_k(\mathbf{u}), \mathbf{r}_k(\mathbf{u})$ is genuinely nonlinear, $S_k(\mathbf{u}_0)$ takes the name of *k-shock curve issued from \mathbf{u}_0* .

Claim. In this case, near a given a *left* state \mathbf{u}_l , we can decompose $S_k(\mathbf{u}_l)$ into the following three disjoint sets:

$$S_k(\mathbf{u}_l) = S_k^+(\mathbf{u}_l) \cup \{\mathbf{u}_l\} \cup S_k^-(\mathbf{u}_l)$$

where

$$S_k^-(\mathbf{u}_l) = \{\mathbf{u} \in S_k(\mathbf{u}_l) : \lambda_k(\mathbf{u}) < \sigma(\mathbf{u}, \mathbf{u}_l) < \lambda_k(\mathbf{u}_l)\} \quad (11.79)$$

and

$$S_k^+(\mathbf{u}_l) = \{\mathbf{u} \in S_k(\mathbf{u}_l) : \lambda_k(\mathbf{u}_l) < \sigma(\mathbf{u}, \mathbf{u}_l) < \lambda_k(\mathbf{u})\}. \quad (11.80)$$

Proof of the claim. Let us use the normalization

$$\nabla \lambda_k(\mathbf{u}) \cdot \mathbf{r}_k(\mathbf{u}) \equiv 1. \quad (11.81)$$

Then, by Theorem 11.22, we can write

$$\begin{aligned} \sigma_k(\varepsilon) &\equiv \sigma(\psi_k(\varepsilon; \mathbf{u}_l), \mathbf{u}_l) = \lambda_k(\mathbf{u}_l) + \frac{\varepsilon}{2} + o(\varepsilon) \\ \mathbf{u}_k(\varepsilon) &\equiv \psi_k(\varepsilon; \mathbf{u}_l) = \mathbf{u}_l + \varepsilon \mathbf{r}_k(\mathbf{u}_l) + o(\varepsilon). \end{aligned}$$

Hence

$$\begin{aligned} \lambda_k(\mathbf{u}_k(\varepsilon)) &= \lambda_k(\mathbf{u}_l) + \varepsilon \nabla \lambda_k(\mathbf{u}_l) \cdot \mathbf{r}_k(\mathbf{u}_l) + o(\varepsilon) = \lambda_k(\mathbf{u}_l) + \varepsilon + o(\varepsilon) \\ &= \sigma_k(\varepsilon) + \frac{\varepsilon}{2} + o(\varepsilon). \end{aligned}$$

Thinking of $\mathbf{u}_k(\varepsilon)$ as a possible *initial* state for $x > 0$, in the Riemann problem, we select which states satisfy the entropy conditions (11.74), (11.75), that is:

$$\begin{aligned} \lambda_k(\mathbf{u}_k(\varepsilon)) &< \sigma_k(\varepsilon) < \lambda_{k+1}(\mathbf{u}_k(\varepsilon)) \\ \lambda_{k-1}(\mathbf{u}_l) &< \sigma_k(\varepsilon) < \lambda_k(\mathbf{u}_l). \end{aligned}$$

First of all, we have $\lambda_k(\mathbf{u}_k(\varepsilon)) < \sigma_k(\varepsilon)$ if and only if $\varepsilon < 0$ and $|\varepsilon|$ small enough. Also, since $\sigma_k(\varepsilon) \rightarrow \lambda_k(\mathbf{u}_l)$ and $\lambda_{k+1}(\mathbf{u}_k(\varepsilon)) \rightarrow \lambda_{k+1}(\mathbf{u}_l)$ as $\varepsilon \rightarrow 0$, we get

$$\sigma_k(\varepsilon) < \lambda_{k+1}(\mathbf{u}_k(\varepsilon)),$$

for $|\varepsilon|$ small enough. On the other hand, $\sigma_k(\varepsilon) < \lambda_k(\mathbf{u}_l)$ if and only if $\varepsilon < 0$ and $|\varepsilon|$ is small enough. Since $\lambda_{k-1}(\mathbf{u}_l) < \lambda_k(\mathbf{u}_l)$ we deduce

$$\lambda_{k-1}(\mathbf{u}_l) < \sigma_k(\varepsilon)$$

for $|\varepsilon|$ small enough.

Thus we can distinguish on the curve $S_k(\mathbf{u}_l)$, the *admissible (entropic)* part given by

$$S_k^-(\mathbf{u}_l) = \{\mathbf{u}_k(\varepsilon) : \lambda_k(\mathbf{u}_k(\varepsilon)) < \sigma_k(\varepsilon) < \lambda_k(\mathbf{u}_l)\}$$

corresponding to $\varepsilon < 0$, and the other part, given by

$$S_k^+(\mathbf{u}_l) = \{\mathbf{u}_k(\varepsilon) : \lambda_k(\mathbf{u}_l) < \sigma_k(\varepsilon) < \lambda_k(\mathbf{u}_k(\varepsilon))\}.$$

These two sets correspond precisely to (11.79) and (11.80), respectively. \square

At this point, the proof of the following theorem is immediate.

Theorem 11.24. *Assume that, for some k , $1 \leq k \leq m$:*

- i) *The pair $\lambda_k(\mathbf{u}), r_k(\mathbf{u})$ is genuinely nonlinear.*
- ii) $\mathbf{u}_r \in S_k^-(\mathbf{u}_l)$.

Then there exists a k -shock solution of the Riemann problem with initial data \mathbf{u}_l , \mathbf{u}_r , given by

$$\mathbf{u}_k(x, t) = \begin{cases} \mathbf{u}_l & \text{for } x < \sigma(\mathbf{u}_l, \mathbf{u}_r)t \\ \mathbf{u}_r & \text{for } x > \sigma(\mathbf{u}_l, \mathbf{u}_r)t. \end{cases} \quad (11.82)$$

11.4.5 The linearly degenerate case

Let us now examine the linearly degenerate case. Once more we first prove a structure lemma.

Lemma 11.25. *Let the pair $\lambda_k(\mathbf{u}), r_k(\mathbf{u})$ be linearly degenerate. Then, for every $\mathbf{u}_0 \in \mathbb{R}^m$:*

- i) $S_k(\mathbf{u}_0) = R_k(\mathbf{u}_0)$.
- ii) $\sigma(\psi_k(\varepsilon; \mathbf{u}_0), \mathbf{u}_0) = \lambda_k(\psi_k(\varepsilon; \mathbf{u}_0)) = \lambda_k(\mathbf{u}_0)$.

Proof. Let $\mathbf{v}_k = \mathbf{v}_k(\theta)$ be the equation for $R_k(\mathbf{u}_0)$, with $\mathbf{v}_k(0) = \mathbf{u}_0$. Thus $\mathbf{v}'_k(\theta) = \mathbf{r}_k(\mathbf{v}(\theta))$. Since

$$\nabla \lambda_k(\mathbf{u}) \cdot \mathbf{r}_k(\mathbf{u}) \equiv 0,$$

the function $\theta \mapsto \lambda_k(\mathbf{v}_k(\theta))$ is constant and equal to $\lambda_k(\mathbf{u}_0)$. Thus, we can write:

$$\begin{aligned} \mathbf{F}(\mathbf{v}_k(\theta)) - \mathbf{F}(\mathbf{u}_0) &= \int_0^\theta \frac{d}{ds} \mathbf{F}(\mathbf{v}_k(s)) ds \\ &= \int_0^\theta D\mathbf{F}(\mathbf{v}_k(s)) \frac{d}{ds} \mathbf{v}_k(s) ds = \int_0^\theta D\mathbf{F}(\mathbf{v}_k(s)) \mathbf{r}_k(\mathbf{v}_k(s)) ds \\ &= \int_0^\theta \lambda_k(\mathbf{v}_k(s)) \mathbf{r}_k(\mathbf{v}_k(s)) ds = \lambda_k(\mathbf{u}_0) \int_0^\theta \frac{d}{ds} \mathbf{v}_k(s) ds \\ &= \lambda_k(\mathbf{u}_0) (\mathbf{v}_k(\theta) - \mathbf{u}_0). \end{aligned}$$

It follows that $\mathbf{v}_k = \mathbf{v}_k(\theta)$ defines both $S_k(\mathbf{u}_0)$ and $R_k(\mathbf{u}_0)$ and that ii) holds. \square

As an immediate consequence we have the following theorem

Theorem 11.26. *Let the pair $\lambda_k(\mathbf{u}), r_k(\mathbf{u})$ be linearly degenerate and $\mathbf{u}_r \in S_k(\mathbf{u}_l)$. Then a solution of the Riemann problem with initial data $\mathbf{u}_l, \mathbf{u}_r$ is given by*

$$\mathbf{u}(x, t) = \begin{cases} \mathbf{u}_l & \text{for } x < \sigma t \\ \mathbf{u}_r & \text{for } x > \sigma t \end{cases} \quad (11.83)$$

with

$$\sigma = \sigma(\mathbf{u}_r, \mathbf{u}_l) = \lambda_k(\mathbf{u}_l) = \lambda_k(\mathbf{u}_r).$$

The solution (11.83) is called a *k-contact discontinuity*. Indeed, the characteristics on both sides are parallel to the discontinuity line (see Fig. 11.3, p. 632).

11.4.6 Local solution of the Riemann problem

Let us summarize the consequences of the theory we have developed so far. Select $k \in \{1, \dots, m\}$. If the pair $\lambda_k(\mathbf{u}), r_k(\mathbf{u})$ is genuinely nonlinear, define

$$T_k(\mathbf{u}_l) = S_k^-(\mathbf{u}_l) \cup \{\mathbf{u}_l\} \cup R_k^+(\mathbf{u}_l).$$

By Theorems 11.17 and 11.22, $T_k(\mathbf{u}_l)$ is a curve of class C^2 in a neighborhood of \mathbf{u}_l and we can use for $T_k(\mathbf{u}_l)$ the following parametrization:

$$\mathbf{u}_k(\varepsilon; \mathbf{u}_l) = \begin{cases} \varphi_k(\varepsilon; \mathbf{u}_l) & \varepsilon > 0 \\ \psi_k(\varepsilon; \mathbf{u}_l) & \varepsilon < 0 \end{cases}$$

with $|\varepsilon| < \varepsilon_0$. We have seen that, if \mathbf{u}_r coincides with one of the states on $T_k(\mathbf{u}_l)$, for some $k \in \{1, \dots, m\}$, then the Riemann problem can be solved by a single k -rarefaction wave or by a k -shock.

If $\lambda_k(\mathbf{u}), r_k(\mathbf{u})$ is linearly degenerate, define

$$T_k(\mathbf{u}_l) = S_k(\mathbf{u}_l) = R_k(\mathbf{u}_l)$$

and we can choose the following parametrization:

$$\mathbf{u}_k(\varepsilon; \mathbf{u}_l) = \psi_k(\varepsilon; \mathbf{u}_l), \quad \text{for } |\varepsilon| < \varepsilon_0.$$

In this case, if \mathbf{u}_r coincides with one of the states on $T_k(\mathbf{u}_l)$, then the Riemann problem can be solved by a single k -contact discontinuity.

When \mathbf{u}_r does not belong to any $T_k(\mathbf{u}_l)$, still we can solve the Riemann problem if \mathbf{u}_l and \mathbf{u}_r are close enough. Precisely, the following result holds¹¹. We denote by \mathcal{S} the class of integral solution, consisting of at most $m + 1$ constant states, connected by admissible shocks, rarefaction waves or contact discontinuities.

¹¹ For the proof, see e.g. [45], Godlewski-Raviart, 1996.

Theorem 11.27. Assume that for all k , $1 \leq k \leq m$, the k^{th} -characteristic field $\lambda_k(\mathbf{u})$, $r_k(\mathbf{u})$ is genuinely nonlinear or linearly degenerate. Given $\mathbf{u}_l \in \mathbb{R}^m$, there exists a neighborhood $N(\mathbf{u}_l)$ of \mathbf{u}_l such that, if $\mathbf{u}_r \in N(\mathbf{u}_l)$, the Riemann problem has a unique solution belonging to \mathcal{S} .

11.5 The Riemann Problem for the p -system

In this section we provide a complete analysis of the Riemann problem for the p -system

$$\begin{cases} w_t - v_x = 0 \\ v_t + p(w)_x = 0. \end{cases}$$

We already know that this system is genuinely nonlinear, with eigenvalues

$$\lambda_1(w, v) = -\sqrt{a(w)} \quad \text{and} \quad \lambda_2(v, u) = \sqrt{a(w)}$$

and corresponding eigenvectors given, respectively, by

$$\mathbf{r}_1(w, v) = \begin{pmatrix} 1 \\ \sqrt{a(w)} \end{pmatrix} \quad \mathbf{r}_2(w, v) = \begin{pmatrix} 1 \\ -\sqrt{a(w)} \end{pmatrix}.$$

We recall that p is decreasing and convex, that is

$$a(w) = -p'(w) > 0 \quad \text{and} \quad a'(w) = -p''(w) < 0 \quad (11.84)$$

and moreover

$$\lim_{w \rightarrow 0^+} p(w) = +\infty. \quad (11.85)$$

Given the two states

$$\mathbf{u}_l = (w_l, v_l) \quad \text{and} \quad \mathbf{u}_r = (w_r, v_r)$$

with $w_l, w_r > 0$, we first look for shock solutions.

11.5.1 Shock waves

Being $m = 2$, we have two types of admissible shock waves. To determine them we first write the Rankine-Hugoniot locus of \mathbf{u}_l . Since

$$F(\mathbf{u}) = \begin{pmatrix} -v \\ p(w) \end{pmatrix},$$

$S(\mathbf{u}_l)$ is the set of the states $\mathbf{u} = (w, v)$ satisfying the system

$$\begin{cases} (w - w_l)\sigma = (v_l - v) \\ (v - v_l)\sigma = (p(w) - p(w_l)) \end{cases}, \quad (11.86)$$

with $\sigma = \sigma(\mathbf{u}, \mathbf{u}_L)$. Solving for v , we find

$$v - v_l = \pm \sqrt{(p(w) - p(w_l))(w_l - w)}. \quad (11.87)$$

- **1-shock.** Consider λ_1, \mathbf{r}_1 . The states on the admissible part $S_1^- (\mathbf{u}_l)$, defined in (11.79) for $k = 1$, satisfy the entropy condition

$$\lambda_1(w, v) < \sigma < \lambda_1(w_l, v_l)$$

(see (11.76)), that is

$$-\sqrt{a(w)} < \sigma < -\sqrt{a(w_l)}, \quad (11.88)$$

which implies $\sigma < 0$ (thus it is *back shock*) and $w_l > w$, since $a' < 0$. Then, from the first equation in (11.86) it follows that $v_l > v$. Thus, for $S_1^- (\mathbf{u}_l)$ we must pick the minus sign in the right hand side of (11.87), i.e.

$$v = v_l - \sqrt{(p(w) - p(w_l))(w_l - w)}, \quad w_l > w. \quad (11.89)$$

Since

$$\frac{dv}{dw} = -\frac{p'(w)(w_l - w) - (p(w) - p(w_l))}{2\sqrt{(p(w) - p(w_l))(w_l - w)}} > 0, \quad (11.90)$$

v is increasing. Moreover, one can check that $d^2v/dw^2 < 0$.

Summarizing, we have: *the initial states \mathbf{u}_r that can be connected by a 1-shock to the state \mathbf{u}_l belong to a curve $S_1^- = S_1^- (\mathbf{u}_l)$ increasing and concave (see Fig. 11.4). The solution of the corresponding Riemann problem is given by*

$$\mathbf{u}(x, t) = \begin{cases} \mathbf{u}_l & \text{for } x < \sigma t \\ \mathbf{u}_r & \text{for } x < \sigma t, \end{cases} \quad (11.91)$$

where $\sigma = \frac{v_r - v_l}{w_l - w_r}$.

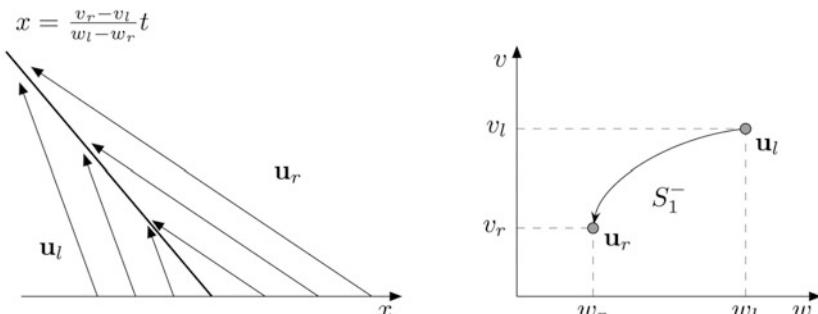


Fig. 11.4 States connected by a 1-shock

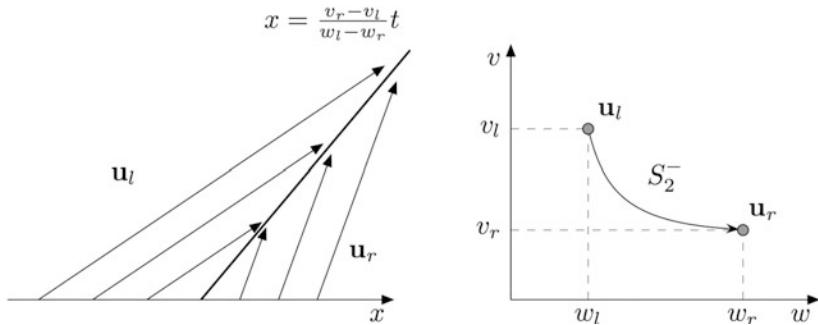


Fig. 11.5 States connected by a 2-shock

- **2-shock.** Consider now λ_2, \mathbf{r}_2 . Let $S_2^- (\mathbf{u}_l)$ be defined by (11.79), for $k = 2$. The states on $S_2^- (\mathbf{u}_l)$ satisfy the entropy condition

$$\lambda_2 (w, v) < \sigma < \lambda_2 (w_l, v_l)$$

that is

$$\sqrt{a(w)} < \sigma < \sqrt{a(w_l)}.$$

In particular, we infer that $\sigma > 0$ (thus it is a *front shock*) and that $w_l < w$. Then, from the first equation in (11.86) it follows that $v_l > v$. Thus the curve $S_2^- (\mathbf{u}_l)$ is again defined by the equation with the minus sign in (11.87), i.e.

$$v = v_l - \sqrt{(p(w) - p(w_l))(w_l - w)}, \quad \text{with } w_l < w. \quad (11.92)$$

From (11.90) it follows that $dv/dw < 0$ and it is easy to check that $d^2v/dw^2 > 0$.

Summarizing, we have: *the initial states \mathbf{u}_r that can be connected by a 2-shock to the state \mathbf{u}_l belong to a curve $S_2^- = S_2^- (\mathbf{u}_l)$ decreasing and convex (see Fig. 11.5).* The solution of the corresponding Riemann problem is still given by (11.91).

11.5.2 Rarefaction waves

Since the system is genuinely nonlinear we have two types of rarefaction waves. We first construct the rarefaction curves $R_1 (\mathbf{u}_l)$ and $R_2 (\mathbf{u}_l)$.

- **1-rarefaction wave.** Consider the 1^{th} -characteristic field

$$\lambda_1 (w, v) = -\sqrt{a(w)}, \quad \mathbf{r}_1 (w, v) = \left(\frac{1}{\sqrt{a(w)}} \right).$$

Since $\lambda_1 < 0$, the corresponding solution is called a *back rarefaction wave*.

To construct the curve $R_1(\mathbf{u}_l)$ we solve system (11.55), p. 633. Writing $\mathbf{v}(\theta) = (w(\theta), v(\theta))$, we find

$$\frac{dw}{d\theta} = 1, \quad \frac{dv}{d\theta} = \sqrt{a(w(\theta))}. \quad (11.93)$$

The choice of $\theta = w$ as a parameter yields the equation

$$\frac{dv}{dw} = \sqrt{a(w)}. \quad (11.94)$$

The solution of (11.94) with initial data \mathbf{u}_l is given by

$$v(w) = v_l + \int_{w_l}^w \sqrt{a(s)} ds. \quad (11.95)$$

The states on the admissible part $R_1^+(\mathbf{u}_l)$ of $R_1(\mathbf{u}_l)$ must satisfy the relation $\lambda_1(w, v) > \lambda_1(w_l, v_l)$, i.e. $\sqrt{a(w)} < \sqrt{a(w_l)}$, which implies $w > w_l$. Thus, the curve $R_1^+(\mathbf{u}_l)$ has equation

$$v(w) = v_l + \int_{w_l}^w \sqrt{a(s)} ds, \quad \text{with } w > w_l.$$

Observe that $w \mapsto v(w)$ is increasing and $d^2v/dw^2 = a'(w)/2\sqrt{a(w)} < 0$. In particular, imposing $v_r = v(w_r)$, $w_r > w_l$, we get $v_l < v_r$.

Finally we solve for $w = w(x, t)$ the scalar conservation law

$$w_t + \lambda_1(w) w_x = w_t - \sqrt{a(w)} w_x = 0.$$

We find $w(x, t) = f(-\frac{x}{t})$, where f is the inverse of the function $w \mapsto \sqrt{a(w)}$.

Summarizing, we have: the initial states \mathbf{u}_r that can be connected by a 1-rarefaction wave to the state \mathbf{u}_l belong a curve $R_1^+ = R_1^+(\mathbf{u}_l)$ increasing and concave (see Fig. 11.6). The solution of the corresponding Riemann problem is given by

$$\mathbf{u}_1(x, t) = \begin{cases} \mathbf{u}_l & x \leq -\sqrt{a(w_l)}t \\ \left(w(x, t), v_l + \int_{w_l}^{w(x, t)} \sqrt{a(s)} ds \right) & -\sqrt{a(w_l)}t < x < -\sqrt{a(w_r)}t \\ \mathbf{u}_r & x \geq -\sqrt{a(w_r)}t, \end{cases}$$

where $w(x, t) = f(-\frac{x}{t})$ and f is the inverse function of $w \mapsto \sqrt{a(w)}$.

- **2-rarefaction wave.** Consider now the 2th-characteristic field

$$\lambda_2(w, v) = \sqrt{a(w)}, \quad \mathbf{r}_2(w, v) = \begin{pmatrix} \frac{1}{-\sqrt{a(w)}} \\ -\sqrt{a(w)} \end{pmatrix}.$$

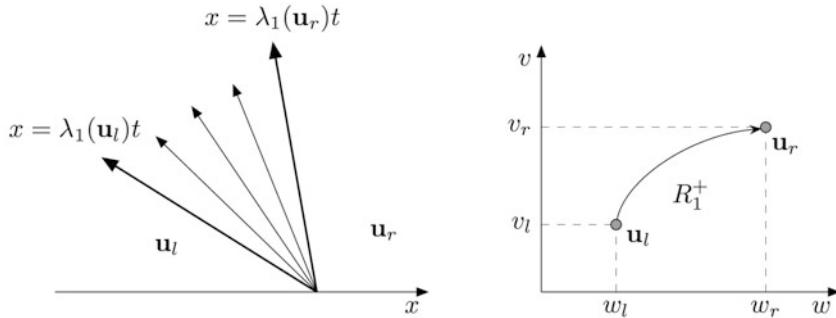


Fig. 11.6 States connected by a 1-rarefaction wave

Since $\lambda_2 > 0$, the corresponding solution is called a *front rarefaction wave*. To construct the curve $R_2(\mathbf{u}_l)$ we can proceed as before. The system (11.55) is given by

$$\frac{dw}{d\theta} = 1, \quad \frac{dv}{d\theta} = -\sqrt{a(w(\theta))}$$

which yields the equation

$$\frac{dv}{dw} = -\sqrt{a(w)}. \quad (11.96)$$

The solution of (11.96) with initial data \mathbf{u}_l is given by

$$v(w) = v_l - \int_{w_l}^w \sqrt{a(s)} ds.$$

The states on the admissible part $R_2^+(\mathbf{u}_l)$ of $R_2(\mathbf{u}_l)$ must satisfy the relation $\lambda_2(w, v) > \lambda_2(w_l, v_l)$, i.e. $\sqrt{a(w)} > \sqrt{a(w_l)}$, which implies $w < w_l$. Thus, the curve $R_2^+(\mathbf{u}_l)$ has equation

$$v(w) = v_l - \int_{w_l}^w \sqrt{a(s)} ds \quad \text{with } w < w_l.$$

Observe that $w \mapsto v(w)$ is decreasing and $d^2v/dw^2 = dw^2 = -a'(w)/2\sqrt{a(w)} > 0$. In particular, imposing $v_r = v(w_r)$, with $w_r < w_l$, we infer $v_l < v_r$. Finally solving

$$w_t + \lambda_2(w) w_x = w_t + \sqrt{a(w)} w_x = 0,$$

we find $w(x, t) = f(\frac{x}{t})$.

Summarizing, we have: the initial states \mathbf{u}_r that can be connected by a 2-rarefaction wave to the state \mathbf{u}_l belong to a curve $R_2^+ = R_2^-(\mathbf{u}_l)$ decreasing and convex

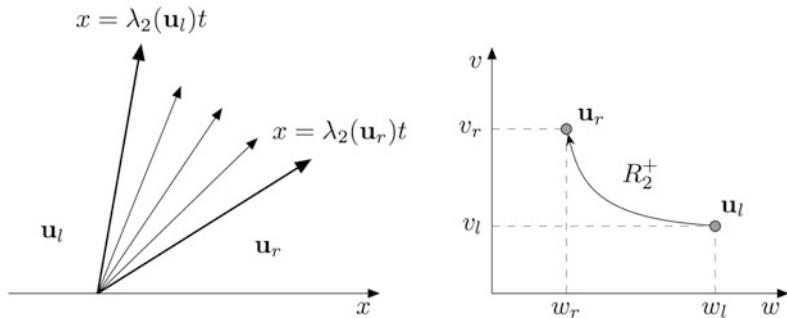


Fig. 11.7 States connected by a 2-rarefaction wave

(Fig. 11.7). The solution of the corresponding Riemann problem is given by

$$\mathbf{u}_2(x, t) = \begin{cases} \left(w(x, t), v_l - \int_{w_l}^{w(x, t)} \sqrt{a(s)} ds \right) & x \leq \sqrt{a(w_l)}t \\ \mathbf{u}_r & \sqrt{a(w_l)}t < x < \sqrt{a(w_r)}t \\ \left(w(x, t), v_l \right) & x \geq \sqrt{a(w_r)}t \end{cases}$$

where $w(x, t) = f\left(\frac{x}{t}\right)$ and f is the inverse function of $w \mapsto \sqrt{a(w)}$.

11.5.3 The solution in the general case

Based on the results in the previous subsections, given a *left* initial datum $\mathbf{u}_l = (v_l, w_l)$, the two C^2 -curves

$$W_1(\mathbf{u}_l) = R_1^+(\mathbf{u}_l) \cup \{\mathbf{u}_l\} \cup S_1^-(\mathbf{u}_l)$$

and

$$W_2(\mathbf{u}_l) = R_2^+(\mathbf{u}_l) \cup \{\mathbf{u}_l\} \cup S_2^-(\mathbf{u}_l)$$

partition the plane v, w in four regions *I*, *II*, *III*, *IV*, as it is shown in Fig. 11.8

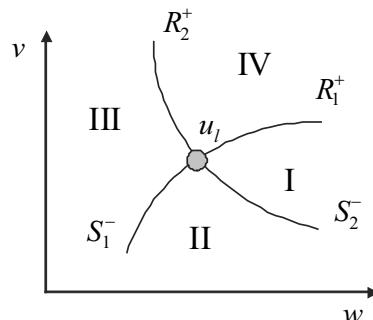


Fig. 11.8 Partition of the plane v, w induced by the curves $W_1(\mathbf{u}_l)$ and $W_2(\mathbf{u}_l)$

If \mathbf{u}_r belongs to one of these two curves, the solution of the Riemann problem is given by a single wave, rarefaction or shock. In general, by Theorem 11.27, p. 643, there exists a neighborhood $\mathcal{N}(\mathbf{u}_l)$ such that, if

$$\mathbf{u}_r \in \mathcal{N}(\mathbf{u}_l),$$

the Riemann problem has a unique solution belonging to the class \mathcal{S} .

For the p -system we have precise information on $\mathcal{N}(\mathbf{u}_l)$. Indeed, we show that if \mathbf{u}_r belongs to one of the regions I , II , III the Riemann problem is always solvable, while \mathbf{u}_r must be sufficiently close to \mathbf{u}_l if \mathbf{u}_r belongs to the region IV . We start with the following result.

Theorem 11.28. *If \mathbf{u}_r belongs to one of the regions I , II , III , there exists a unique weak solution of the Riemann problem, belonging to \mathcal{S} .*

Proof. Let \mathcal{F} be the family of curves W_2 starting from points on $W_1(\mathbf{u}_l)$. Precisely, let us define

$$\mathcal{F} = \{W_2(\mathbf{u}_0) : \mathbf{u}_0 \in W_1(\mathbf{u}_l)\}.$$

To prove the theorem, it is enough to show that every point \mathbf{u}_r in one of the regions I , II , III belongs to one and only one curve $W_2(\mathbf{u}_0)$ of the family \mathcal{F} . In fact, if this is true, we can construct the unique entropic solution by first connecting \mathbf{u}_l to \mathbf{u}_0 by a 1-rarefaction wave or a 1-shock and then by connecting \mathbf{u}_0 to \mathbf{u}_r by a 2-rarefaction wave or a 2-shock. Clearly, the location of \mathbf{u}_r determine the type of waves to be chosen.

To be specific, assume that \mathbf{u}_r belongs to region I . With reference to Fig. 11.9, consider the points $\mathbf{p} \in W_1(\mathbf{u}_l)$ within the vertical lines $w = w_l$ and $w = w_r$. In particular, being $\mathbf{p} \in R_1^+(\mathbf{u}_l)$, the coordinates of \mathbf{p} are $(w, v(w))$, where (see (11.95))

$$v(w) = v_l + \int_{w_l}^w \sqrt{a(s)} ds.$$

Since the slope of $S_2^-(\mathbf{p})$ is negative and bounded, all these curves intersect the straight-line $w = w_r$ at some point $\varphi(\mathbf{p}(w))$ of coordinates $(w_r, V(w))$ where, since $\varphi(\mathbf{p}) \in$

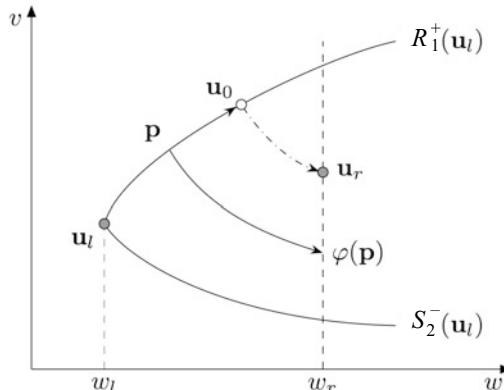


Fig. 11.9 The map φ in the proof of Theorem 11.27

$S_2^- (\mathbf{p})$, (see (11.92)):

$$\begin{aligned} V(w) &= v(w) - \sqrt{(p(w_r) - p(w))(w - w_r)} \\ &= v_l + \int_{w_l}^w \sqrt{a(s)} ds - \sqrt{(p(w_r) - p(w))(w - w_r)}. \end{aligned}$$

Hence the map $w \mapsto \varphi(\mathbf{p}(w))$ is continuous and if w is close to w_r , the point $\varphi(\mathbf{p}(w))$ is above \mathbf{u}_r . Since $\varphi(\mathbf{u}_l)$ is below \mathbf{u}_r , by continuity, there exists a point \mathbf{u}_0 such that $\varphi(\mathbf{u}_0) = \mathbf{u}_r$.

This shows that every point \mathbf{u}_r in the region I belongs to one of the curves of the family \mathcal{F} . To prove uniqueness of this curve, it is enough to check that dV/dw is strictly positive. In fact, we have

$$\frac{dV}{dw} = \sqrt{a(w)} - \frac{p(w_r) - p(w) + p'(w)(w_r - w)}{2\sqrt{(p(w_r) - p(w))(w - w_r)}} > 0,$$

since $p'(w) < 0$ and $w_r > w$. Thus, when \mathbf{u}_r belongs to the region I the proof is complete.

If \mathbf{u}_r belongs to the region III the argument is similar. For \mathbf{u}_r in the region II , we show that the horizontal line $v = v_r$ intersects $S_1^- (\mathbf{u}_l)$ and $S_2^- (\mathbf{u}_l)$ at two points $\mathbf{p}_1 = (w_1, v_r)$ and $\mathbf{p}_2 = (w_2, v_r)$, uniquely determined. To find w_2 , we solve the equation (see (11.92))

$$v_r = v_l - \sqrt{(p(w_l) - p(w))(w - w_l)}, \quad w > w_l. \quad (11.97)$$

Now, the function

$$w \mapsto v_l - \sqrt{(p(w_l) - p(w))(w - w_l)}$$

is one-to-one from $[w_l, +\infty)$ onto $[-\infty, v_l]$. Since $v_r < v_l$, there exists exactly one solution w_2 of (11.97).

Similarly, using the fact that

$$p(w) \rightarrow +\infty \text{ as } w \rightarrow 0^+,$$

it follows that there exists a unique solution w_1 of the equation (see (11.89))

$$v_r = v_l - \sqrt{(p(w_l) - p(w))(w - w_l)}, \quad w < w_l.$$

The rest of the proof follows closely the argument we used for the region I . We leave the details to the reader. \square

The construction of the solution when \mathbf{u}_r is in the region I is described in Fig. 11.10: \mathbf{u}_l is connected to \mathbf{u}_0 by a *back rarefaction wave* and then \mathbf{u}_0 is connected to \mathbf{u}_r by a *front-shock*.

When \mathbf{u}_r is in the region IV , *sufficiently close to \mathbf{u}_l* , the solution can be constructed as in the other cases, as shown in Fig. 11.11: \mathbf{u}_l is connected to \mathbf{u}_0 by a *back rarefaction wave* and then \mathbf{u}_0 is connected to \mathbf{u}_r by a *front rarefaction wave*.

We now show that, when \mathbf{u}_r belongs to the region IV , it is not always possible to construct a solution in the class we are considering.

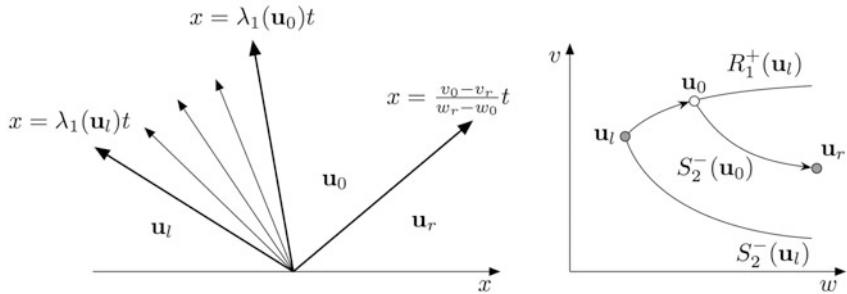


Fig. 11.10 Solution of the Riemann problem when \mathbf{u}_r belongs to region I

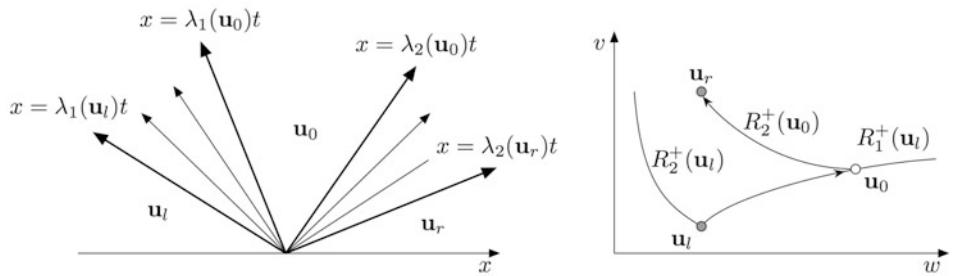


Fig. 11.11 The solution of the Riemann problem when \mathbf{u}_r belongs to region IV

Proposition 11.29. Let \mathbf{u}_r belong to the region IV and assume that

$$v_\infty \equiv \int_{w_l}^{+\infty} \sqrt{a(w)} dw < \infty. \quad (11.98)$$

If $v_r > v_l + 2v_\infty$, the Riemann problem cannot have a solution belonging to the class \mathcal{S} .

Proof. If (11.98) holds, the rarefaction curve $R_1^+(\mathbf{u}_l)$, of equation

$$v = v_l + \int_{w_l}^w \sqrt{a(s)} ds$$

has a horizontal asymptote, given by the straight line $v = v_\infty + v_l$. With reference to Fig. 11.12, consider the state $\mathbf{u}_r = (w_r, v_r)$ with $v_r > 2v_\infty + v_l$. We show that no 2-rarefaction curve issued from a point on $R_1^+(\mathbf{u}_l)$ can reach \mathbf{u}_r .

In fact, for every point $\mathbf{u}_0 = (w_0, v_0) \in R_1^+(\mathbf{u}_l)$ we have

$$v_0 = v_l + \int_{w_l}^{w_0} \sqrt{a(s)} ds < v_l + v_\infty.$$

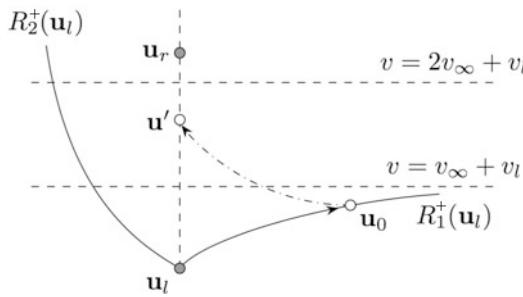


Fig. 11.12 Non-existence for the Riemann problem

Therefore, if $\mathbf{u}' = (w_l, v') \in R_2^+(\mathbf{u}_l) \cap \{w = w_l\}$, then

$$v' = v_0 - \int_{w_0}^{w_l} \sqrt{a(s)} ds = v_0 + \int_{w_l}^{w_0} \sqrt{a(s)} ds < v_l + 2v_\infty < v_r.$$

Thus, if $v_r > v_l + 2v_\infty$, there is no way to connect \mathbf{u}_l with \mathbf{u}_r through an intermediate state $\mathbf{u}_0 \in R_1^+(\mathbf{u}_l)$, and this is clearly the only way to construct a solution in the class \mathcal{S} , with initial data $\mathbf{u}_l, \mathbf{u}_r$. \square

Remark 11.30. Note that (11.98) is true in the important case of a polytropic gas, for which

$$a(w) = \gamma k w^{-\gamma-1} \quad \gamma > 1.$$

To get a physical interpretation of what happens, consider the extreme situation $v_r = 2v_\infty + v_l$. Then $\mathbf{u}_r = (w_r, v_r)$ formally belongs to a 2- rarefaction curve coming from $\mathbf{u}_0 = (+\infty, v_\infty + v_l)$. Since $\lambda_1(\mathbf{u}_0) = \lambda_2(\mathbf{u}_0) = 0$, the solution of the Riemann problem is described in Fig. 11.11, but with the two rarefaction waves separated by the vertical line $x = 0$. On this line $w = 1/\rho = +\infty$ or $\rho = 0$, which, in the case of the motion of a gas along a tube, corresponds to a formation of a vacuum zone.

Problems

11.1. The telegrapher system. The following system

$$\begin{cases} LI_t + V_x + RI = 0, \\ CV_t + I_x + GV = 0 \end{cases} \quad (x \in \mathbb{R}, t > 0)$$

describes the flow of electricity in a line, such as a coaxial cable. The variable x is a coordinate along the cable. $I = I(x, t)$ and $V(x, t)$ represent the current in the inner wire and the voltage across the cable, respectively. The electrical properties of the line are encoded into the constants C , capacitance to ground, R , resistance, L , inductance and G , conductance to ground, all per unit length. We assign initial conditions $I(x, 0) = I_0(x)$, $V(x, 0) = V_0(x)$.

654 11 Systems of Conservation Laws

- a) Show that the system is strictly hyperbolic.
- b) Show that in the *distortionless* case $RC = GL$, the full system is uncoupled and reduces to the equations

$$w_t^\pm \pm \frac{1}{\sqrt{LC}} w_x^\pm = -\frac{R}{L} w^\pm$$

with initial conditions

$$w^\pm(x, 0) = \frac{1}{2} \left[\frac{I_0(x)}{\sqrt{C}} \pm \frac{V_0(x)}{\sqrt{L}} \right] \equiv w_0^\pm(x).$$

- c) In the *distortionless* case $RC = GL$, find an explicit formula for the original solution $(U(x, t), V(x, t))$.

[Answer: c) The solution is given by the following superposition of damped travelling waves:

$$\begin{pmatrix} U(x, t) \\ V(x, t) \end{pmatrix} = \left\{ w_0^+(x + t/\sqrt{LC}) \begin{pmatrix} \sqrt{C} \\ \sqrt{L} \end{pmatrix} + w_0^-(x - t/\sqrt{LC}) \begin{pmatrix} \sqrt{C} \\ -\sqrt{L} \end{pmatrix} \right\} e^{-\frac{R}{L}t}.$$

11.2. Show that a k -Riemann invariant is constant along a k -rarefaction wave.

11.3. Consider the p -system. Is it possible to have a shock with only one discontinuous component?

11.4. Consider the solution of the Riemann problem for p -system when \mathbf{u}_r is in the region I. Instead of proceeding as in Fig. 11.10, can we alternatively connect \mathbf{u}_l to an intermediate state \mathbf{u}_0 on S_2^- (\mathbf{u}_l) by a *front shock* and then connect \mathbf{u}_0 to \mathbf{u}_r by a *back rarefaction wave*?

11.5. Describe as in Figs. 11.10 and 11.11 the structure of the solution of the Riemann problem for the p -system, when \mathbf{u}_r belongs to one of the regions II and III.

11.6. In the Riemann problem for the gas dynamic system one assigns the two states (ρ_l, v_l, p_l) and (ρ_r, v_r, p_r) . Assume that we are in presence of a discontinuity, propagating at speed σ . Let $U = v - \sigma$, which represents the flow velocity relative to the discontinuity.

- a) Using the system in the conservation form (11.9), show that the Rankine-Hugoniot jump conditions can be written in the following form:

$$\begin{cases} [\rho U] = 0 \\ [\rho U^2 + p] = 0 \\ [(\rho(E + \frac{1}{2}U^2) + p)] = 0. \end{cases}$$

- b) Set $M = \rho U$. What does M represent? Show that if $M = 0$ the discontinuity must be a contact discontinuity.

[Answer: b) M represent the mass flux across the discontinuity curve. If $M = 0$, no mass goes through the discontinuity and we necessarily have $U_l = U_r = 0$. Hence $v_r = v_l = \sigma$ and, from the second Rankine-Hugoniot condition, we infer $p_r = p_l$. Since we have a discontinuity, it must be $\rho_l \neq \rho_r$ and therefore we have a 2-contact discontinuity with $\sigma = \lambda_2 = v$. It can be shown that, if $M \neq 0$, we get a 1 or 2-shock].

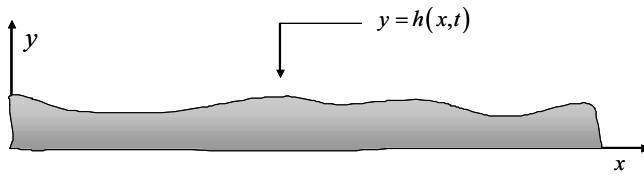


Fig. 11.13 Reference frame for the shallow water system

11.7. Shallow water waves (I): The so called *shallow water system* governs the motion of water, with constant density, under the condition that the ratio between the water depth and the typical wave length of the free surface is small. Other simplifying assumptions are:

1. The motion is essentially bidimensional and we use a reference frame where x and y are the horizontal and vertical coordinates, respectively. In this coordinate system, the free surface is described by a graph $y = h(x, t)$, while the bottom is flat, at level $y = 0$ (see Fig. 11.13).
2. The vertical acceleration is small compared to g , the gravitational acceleration, and the horizontal component u of the fluid velocity has no relevant variations along the vertical direction. Thus, $u = u(x, t)$.

Under the above conditions, the governing system reduces to mass conservation and momentum balance along the x direction, and takes the following form:

$$\begin{cases} ht + hu_x + uh_x = 0 \\ ut + uu_x + gh_x = 0. \end{cases} \quad (11.99)$$

- a) Show that (11.99) is strictly hyperbolic and genuinely nonlinear.
- b) Compute the Riemann invariants.

[Answer: a) The eigenvalues are $\lambda_1(h, u) = u - \sqrt{gh}$, $\lambda_2(h, u) = u + \sqrt{gh}$, with eigenvectors

$$\mathbf{r}_1(h, u) = \begin{pmatrix} -\sqrt{h} \\ \sqrt{g} \end{pmatrix}, \quad \mathbf{r}_2(h, u) = \begin{pmatrix} \sqrt{h} \\ \sqrt{g} \end{pmatrix}.$$

Moreover $\nabla \lambda_1 \cdot \mathbf{r}_1 = \nabla \lambda_2 \cdot \mathbf{r}_2 = \frac{3}{2}\sqrt{g}$.

b) Riemann invariants are $R(h, u) = u - 2\sqrt{hg}$, $S(h, u) = u + 2\sqrt{hg}$.

11.8. Shallow water waves (II): the dam break problem. Assume that a quantity of shallow water of height h_0 is held motionless in the domain $x < 0$ by a dam located at $x = 0$ (see Fig. 11.14). Ahead of the dam there is no water. Suppose that the dam suddenly breaks at time $t = 0$ and that we want to determine the water flow u and the profile h of the free surface for $t > 0$. Note that $0 \leq h \leq h_0$ and $u \geq 0$. According to the shallow water model in Problem 11.7, we have to solve the Riemann problem for system (11.99), with data $(h_l, u_l) = (h_0, 0)$, $(h_r, u_r) = (0, 0)$. We ask the reader to provide an explicit analytic solution by filling in the details in the following steps, along the guideline of the analysis of the p -system¹². We set

$$c_0 = \sqrt{gh_0}.$$

¹² The solution can be also found by a direct use of the Riemann invariants. See e.g. [22], Billingham-King, 2000.

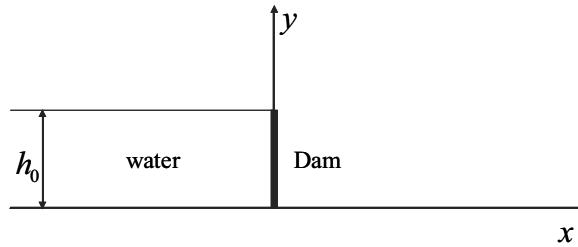


Fig. 11.14 Schematic representation for the dam break problem

1. Examine which states (h, u) can be connected to the right of $(h_0, 0)$. In particular, show that these states belong to the rarefaction curve $R_1^+(h_0, 0)$ given by the parametric equations

$$h(\theta) = \left(\sqrt{h_0} - \frac{\theta}{2} \right)^2, \quad u(\theta) = \sqrt{g}\theta, \quad \text{with } \theta \geq 0,$$

or, eliminating θ , by the equation

$$u = 2 \left(c_0 - \sqrt{gh} \right), \quad (11.100)$$

in the first quadrant of the state plane $h \geq 0, u \geq 0$. In particular, check that, if $(h, u) \in R_1^+(h_0, 0)$, then

$$\lambda_1(h, u) > -c_0 = \lambda_1(h_0, 0). \quad (11.101)$$

2. Deduce from (11.100) and (11.101), that the states $(h_0, 0)$ and $(0, 2c_0)$ are connected by a 1-rarefaction wave defined in the sector

$$-c_0 t < x < 2c_0 t$$

and that the state $(0, 2c_0)$ is connected to $(0, 0)$ by a 2-shock, for $x > 2c_0 t$.

3. Compute the inverse function of

$$\theta \longmapsto \lambda_1(h(\theta), u(\theta)) = \frac{3}{2} \sqrt{g}\theta - c_0,$$

to derive the following analytical expression for the 1-rarefaction wave (see Fig. 11.15 for the h profile):

$$h(x, t) = \frac{1}{9g} \left(2c_0 - \frac{x}{t} \right)^2, \quad u(x, t) = \frac{2}{3} \left(c_0 + \frac{x}{t} \right).$$

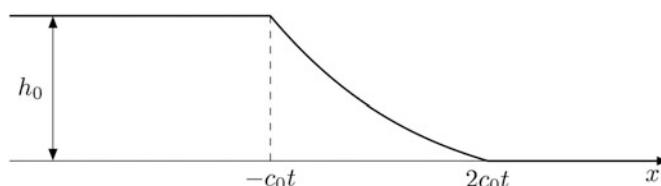


Fig. 11.15 The h profile in the dam break problem at time t

Appendix A

Fourier Series

A.1 Fourier Coefficients

Let u be a $2T$ -periodic function in \mathbb{R} and assume that u can be expanded in a trigonometric series as follows:

$$u(x) = U + \sum_{k=1}^{\infty} \{a_k \cos k\omega x + b_k \sin k\omega x\} \quad (\text{A.1})$$

where $\omega = \pi/T$.

First question: how u and the coefficients U , a_k and b_k are related to each other? To answer, we use the following so called *orthogonality relations*, whose proof is elementary:

$$\int_{-T}^T \cos k\omega x \cos m\omega x \, dx = \int_{-T}^T \sin k\omega x \sin m\omega x \, dx = 0 \quad \text{if } k \neq m$$

$$\int_{-T}^T \cos k\omega x \sin m\omega x \, dx = 0 \quad \text{for all } k, m \geq 0.$$

Moreover

$$\int_{-T}^T \cos^2 k\omega x \, dx = \int_{-T}^T \sin^2 k\omega x \, dx = T. \quad (\text{A.2})$$

Now, suppose that the series (A.1) converges *uniformly* in \mathbb{R} . Multiplying (A.1) by $\cos n\omega x$ and integrating term by term over $(-T, T)$, the orthogonality relations and (A.2) yield, for $n \geq 1$,

$$\int_{-T}^T u(x) \cos n\omega x \, dx = Ta_n$$

or

$$a_n = \frac{1}{T} \int_{-T}^T u(x) \cos n\omega x \, dx. \quad (\text{A.3})$$

For $n = 0$ we get

$$\int_{-T}^T u(x) \, dx = 2UT$$

or, setting $U = a_0/2$,

$$a_0 = \frac{1}{T} \int_{-T}^T u(x) \, dx \quad (\text{A.4})$$

which is coherent with (A.3) as $n = 0$. Similarly, we find

$$b_n = \frac{1}{T} \int_{-T}^T u(x) \sin n\omega x \, dx. \quad (\text{A.5})$$

Thus, if u has the uniformly convergent expansion (A.1), the coefficients a_n, b_n (with $a_0 = 2U$) must be given by the formulas (A.3) and (A.5). In this case we say that the trigonometric series

$$\frac{a_0}{2} + \sum_{k=1}^{\infty} \{a_k \cos k\omega x + b_k \sin k\omega x\} \quad (\text{A.6})$$

is the *Fourier series* of u and the coefficients (A.3), (A.4) and (A.5) are called the *Fourier coefficients* of u .

- *Odd and even functions.* If u is an *odd* function, i.e. $u(-x) = -u(x)$, we have $a_k = 0$ for every $k \geq 0$, while

$$b_k = \frac{2}{T} \int_0^T u(x) \sin k\omega x \, dx.$$

Thus, if u is odd, its Fourier series is a *sine* Fourier series:

$$u(x) = \sum_{k=1}^{\infty} b_k \sin k\omega x.$$

Similarly, if u is *even*, i.e. $u(-x) = u(x)$, we have $b_k = 0$ for every $k \geq 1$, while

$$a_k = \frac{2}{T} \int_0^T u(x) \cos k\omega x \, dx.$$

Thus, if u is even, its Fourier series is a *cosine* Fourier series:

$$u(x) = \frac{a_0}{2} + \sum_{k=1}^{\infty} a_k \cos k\omega x.$$

- *Fourier coefficients of a derivative.* Let $u \in C^1(\mathbb{R})$ be $2T$ -periodic. Then we may compute the Fourier coefficients a'_k and b'_k of u' . We have, integrating by parts, for $k \geq 1$:

$$\begin{aligned} a'_k &= \frac{1}{T} \int_{-T}^T u'(x) \cos k\omega x \, dx \\ &= \frac{1}{T} [u(x) \cos k\omega x]_{-T}^T + \frac{k\omega}{T} \int_{-T}^T u(x) \sin k\omega x \, dx \\ &= \frac{k\omega}{T} \int_{-T}^T u(x) \sin k\omega x \, dx = k\omega b_k \end{aligned}$$

and

$$\begin{aligned} b'_k &= \frac{1}{T} \int_{-T}^T u'(x) \sin k\omega x \, dx \\ &= \frac{1}{T} [u(x) \sin k\omega x]_{-T}^T - \frac{k\omega}{T} \int_{-T}^T u(x) \cos k\omega x \, dx \\ &= -\frac{k\omega}{T} \int_{-T}^T u(x) \cos k\omega x \, dx = -k\omega a_k. \end{aligned}$$

Thus, the Fourier coefficients a'_k and b'_k are related to a_k and b_k by the following formulas:

$$a'_k = k\omega b_k, \quad b'_k = -k\omega a_k. \quad (\text{A.7})$$

- *Complex form of a Fourier series.* Using the Euler identities

$$e^{\pm ik\omega x} = \cos k\omega x \pm i \sin k\omega x$$

the Fourier series (A.6) can be expressed in the complex form

$$\sum_{k=-\infty}^{\infty} c_k e^{ik\omega x},$$

where the complex Fourier coefficients c_k are given by

$$c_k = \frac{1}{2T} \int_{-T}^T u(z) e^{-ik\omega z} dz.$$

The relations among the real and the complex Fourier coefficients are:

$$c_0 = \frac{1}{2} a_0$$

and

$$c_k = \frac{1}{2} (a_k - b_k), \quad c_{-k} = \bar{c}_k \quad \text{for } k > 0.$$

A.2 Expansion in Fourier Series

In the above computations we started from a function u admitting a uniform convergent expansion in Fourier series. Adopting a different point of view, let u be a $2T$ -periodic function and assume we can compute its Fourier coefficients, given by formulas (A.3) and (A.5). Thus, we can *associate* with u its Fourier series and write

$$u \sim \frac{a_0}{2} + \sum_{k=1}^{\infty} \{a_k \cos k\omega x + b_k \sin k\omega x\}.$$

The main questions are now the following:

1. Which conditions on u do assure “the convergence” of its Fourier series? Of course there are several notions of convergence (e.g pointwise, uniform, least squares).
2. If the Fourier series is convergent in some sense, does it always have sum u ?

A complete answer to the above questions is not elementary. The convergence of a Fourier series is a rather delicate matter. We indicate some basic results (for the proofs, see e.g. [36] *Rudin, 1976* or [41], *Zygmund and Wheeden, 1977*).

- *Least squares or L^2 convergence.* This is perhaps the most natural type of convergence for Fourier series (see Subsect. 6.4.2). Let

$$S_N(x) = \frac{a_0}{2} + \sum_{k=1}^N \{a_k \cos k\omega x + b_k \sin k\omega x\}$$

be the N -partial sum of the Fourier series of u . We have

Theorem A.1. *Let u be a square integrable function¹ on $(-T, T)$. Then*

$$\lim_{N \rightarrow +\infty} \int_{-T}^T [S_N(x) - u(x)]^2 dx = 0.$$

Moreover, the following Parseval relation holds:

$$\frac{1}{T} \int_{-T}^T u^2 = \frac{a_0^2}{2} + \sum_{k=1}^{\infty} (a_k^2 + b_k^2). \tag{A.8}$$

¹ That is $\int_{-T}^T u^2 < \infty$.

Since the numerical series in the right hand side of (A.8) is convergent, we deduce the following important consequence:

Corollary A.2 (Riemann-Lebesgue).

$$\lim_{k \rightarrow +\infty} a_k = \lim_{k \rightarrow +\infty} b_k = 0.$$

- *Pointwise convergence.* We say that u satisfies the *Dirichlet conditions* in $[-T, T]$ if it is continuous in $[-T, T]$ except possibly at a finite number of points of jump discontinuity and moreover if the interval $[-T, T]$ can be partitioned in a finite numbers of subintervals such that u is monotone in each one of them.

The following theorem holds.

Theorem A.3. *If u satisfies the Dirichlet conditions in $[-T, T]$ then the Fourier series of u converges at each point of $[-T, T]$. Moreover²:*

$$\frac{a_0}{2} + \sum_{k=1}^{\infty} \{a_k \cos k\omega x + b_k \sin k\omega x\} = \begin{cases} \frac{u(x+) + u(x-)}{2} & x \in (-T, T) \\ \frac{u(T-) + u(-T+)}{2} & x = \pm T \end{cases}$$

In particular, under the hypotheses of Theorem A.3, at every point x of continuity of u , the Fourier series converges to $u(x)$.

- *Uniform convergence.* A simple criterion of uniform convergence is provided by the Weierstrass test (see Sect.1.4). Since

$$|a_k \cos k\omega x + b_k \sin k\omega x| \leq |a_k| + |b_k|$$

we deduce: *If the numerical series*

$$\sum_{k=1}^{\infty} |a_k| \quad \text{and} \quad \sum_{k=1}^{\infty} |b_k|$$

are convergent, then the Fourier series of u is uniformly convergent in R , with sum u .

This is the case, for instance, if $u \in C^1(\mathbb{R})$ and is $2T$ periodic. In fact, from (A.7) we have for every $k \geq 1$,

$$a_k = -\frac{1}{\omega k} b'_k \quad \text{and} \quad b_k = \frac{1}{\omega k} a'_k.$$

Therefore

$$|a_k| \leq \frac{1}{\omega k^2} + (b'_k)^2$$

² We set $f(x\pm) = \lim_{y \rightarrow \pm x} f(y)$.

and

$$|b_k| \leq \frac{1}{\omega k^2} + (a'_k)^2.$$

Now, the series $\sum \frac{1}{k^2}$ is convergent. On the other hand, also the series

$$\sum_{k=1}^{\infty} (a'_k)^2 \quad \text{and} \quad \sum_{k=1}^{\infty} (b'_k)^2$$

are convergent, by Parseval's relation (A.8) applied to u' in place of u . The conclusion is that *if $u \in C^1(\mathbb{R})$ and $2T$ periodic, its Fourier series is uniformly convergent in \mathbb{R} with sum u .*

Another useful result is a refinement of Theorem A.2.

Theorem A.4. Assume u satisfies the Dirichlet conditions in $[-T, T]$. Then:

- a) If u is continuous in $[a, b] \subset (-T, T)$, then its Fourier series converges uniformly in $[a, b]$.
- b) If u is continuous in $[-T, T]$ and $u(-T) = u(T)$, then its Fourier series converges uniformly in $[-T, T]$ (and therefore in \mathbb{R}).

Appendix B

Measures and Integrals

B.1 Lebesgue Measure and Integral

B.1.1 A counting problem

Two persons, that we denote by \mathcal{R} and \mathcal{L} , must compute the total value of M coins, ranging from 1 to 50 cents. \mathcal{R} decides to group the coins arbitrarily in piles of, say, 10 coins each, then to compute the value of each pile and finally to sum all these values. \mathcal{L} , instead, decides to partition the coins according to their value, forming piles of 1-cent coins, of 5-cents coins and so on. Then he computes the value of each pile and finally sums all their values.

In more analytical terms, let

$$V : M \rightarrow \mathbb{N}$$

be a *value function* that associates to each element of M (i.e. each coin) its value. \mathcal{R} partitions the **domain** of V in disjoint subsets, sums the values of V in such subsets and then sums everything. \mathcal{L} considers each point p in the **image** of V (the value of a single coin), considers the inverse image $V^{-1}(p)$ (the pile of coins with the same value p), computes the corresponding value and finally sums over every p .

These two ways of counting correspond to the strategy behind the definitions of the integrals of Riemann and Lebesgue, respectively. Since V is defined on a discrete set and is integer valued, in both cases there is no problem in summing its values and the choice is determined by an efficiency criterion. Usually, the method of \mathcal{L} is more efficient.

In the case of a real (or complex) function f , the “sums of its values” corresponds to an integration of f . While the construction of \mathcal{R} remains rather elementary, the one of \mathcal{L} requires new tools.

Let us examine the particular case of a *bounded* and *positive* function, defined on an interval $[a, b] \subset \mathbb{R}$. Thus, let

$$f : [a, b] \rightarrow [\inf f, \sup f].$$

To construct the Riemann integral, we partition $[a, b]$ in subintervals I_1, \dots, I_N (the piles of \mathcal{R}), then we choose in each interval I_k a point ξ_k and we compute $f(\xi_k)l(I_k)$, where $l(I_k)$ is the length of I_k , (i.e. the value of the k -th pile). Now we sum the values $f(\xi_k)l(I_k)$ and set

$$(\mathcal{R}) \int_a^b f = \lim_{\delta \rightarrow 0} \sum_{k=1}^N f(\xi_k)l(I_k),$$

where $\delta = \max\{l(I_1), \dots, l(I_N)\}$. If the limit is finite and moreover is independent of the choice of the points ξ_k , then this limit defines the Riemann integral of f in $[a, b]$.

Now, let us examine the Lebesgue strategy. This time we partition the interval $[\inf f, \sup f]$ in subintervals $[y_{k-1}, y_k]$ (the values of each coin for \mathcal{L}) with

$$\inf f = y_0 < y_1 < \dots < y_{N-1} < y_N = \sup f.$$

Then we consider the inverse images $E_k = f^{-1}([y_{k-1}, y_k])$ (the piles of homogeneous coins) and we would like to compute their *length*. However, in general E_k is *not* an interval or a union of intervals and, in principle, it could be a very irregular set so that it is not clear what is the “length” of E_k .

Thus, the need arises to associate with every E_k a *measure*, which replaces the length when E_k is an irregular set. This leads to the introduction of the *Lebesgue measure* of (practically every) set $E \subseteq \mathbb{R}$, denoted by $|E|$.

Once we know how to measure E_k (the number of coins in the k -th pile), we choose an arbitrary point $\bar{\alpha}_k \in [y_{k-1}, y_k]$ and we compute $\bar{\alpha}_k |E_k|$ (the value of the k -th pile). Then, we sum all the values $\bar{\alpha}_k |E_k|$ and set

$$(\mathcal{L}) \int_a^b f = \lim_{\rho \rightarrow 0} \sum_{k=1}^N \bar{\alpha}_k |E_k|$$

where ρ is the maximum among the lengths of the intervals $[y_{k-1}, y_k]$. It can be seen that under our hypotheses, the limit exists, is finite and is independent of the choice of $\bar{\alpha}_k$. Thus, we may always choose $\bar{\alpha}_k = y_{k-1}$. This remark leads to the definition of the Lebesgue integral in Sect. B.3: the number $\sum_{k=1}^N y_{k-1} |E_k|$ is nothing else than the integral of a *simple function*, which approximates f from below and whose range is the finite set $y_0 < \dots < y_{N-1}$. The integral of f is the supremum of these numbers.

The resulting theory has several advantages with respect to that of Riemann. For instance, the class of integrable functions is much wider and there is no need to distinguish among bounded or unbounded functions or integration domains.

Especially important are the convergence theorems presented in Sect. B.1.4, which allow the possibility of interchanging the operation of limit and integration, under rather mild conditions.

Finally, the construction of the Lebesgue measure and integral can be greatly generalized as we will mention in Sect. B.1.5.

For the proofs of the theorems stated in this Appendix, the interested reader can consult, e.g. [36], *Rudin, 1976*, or [41], *Zygmund and Wheeden, 1977*.

B.1.2 Measures and measurable functions

A measure in a set Ω is a *set function*, defined on a particular class of subsets of Ω called *measurable set* which “behaves well” with respect to union, intersection and complementation. Precisely:

Definition B.1. A collection \mathcal{F} of subsets of Ω is called σ -algebra if:

- (i) $\emptyset, \Omega \in \mathcal{F}$.
- (ii) $A \in \mathcal{F}$ implies $\Omega \setminus A \in \mathcal{F}$.
- (iii) If $\{A_k\}_{k \in \mathbb{N}} \subset \mathcal{F}$ then also $\cup A_k$ and $\cap A_k$ belong to \mathcal{F} .

Example B.2. If $\Omega = \mathbb{R}^n$, we can define the smallest σ -algebra containing all the open subsets of \mathbb{R}^n , called the *Borel σ -algebra*. Its elements are called *Borel sets*, typically obtained by countable unions and/or intersections of open sets.

Definition B.3. Given a σ -algebra \mathcal{F} in a set Ω , a measure on \mathcal{F} is a function $\mu : \mathcal{F} \rightarrow \mathbb{R}$ such that:

- (i) $\mu(A) \geq 0$ for every $A \in \mathcal{F}$.
- (ii) If A_1, A_2, \dots are pairwise disjoint sets in \mathcal{F} , then

$$\mu(\cup_{k \geq 1} A_k) = \sum_{k \geq 1} \mu(A_k) \quad (\sigma - \text{additivity}).$$

The elements of \mathcal{F} are called *measurable sets*.

The Lebesgue measure in \mathbb{R}^n is defined on a σ -algebra \mathcal{M} containing the Borel σ -algebra, through the following theorem.

Theorem B.4. There exists in \mathbb{R}^n a σ -algebra \mathcal{M} and a measure

$$|\cdot|_n : \mathcal{M} \rightarrow [0, +\infty]$$

with the following properties:

1. Each open and closed set belongs to \mathcal{M} .
2. If $A \in \mathcal{M}$ and A has measure zero, every subset of A belongs to \mathcal{M} and has measure zero.
3. If

$$A = \{\mathbf{x} \in \mathbb{R}^n : a_j < x_j < b_j; j = 1, \dots, n\}$$

then $|A| = \prod_{j=1}^n (b_j - a_j)$.

The elements of \mathcal{M} are called *Lebesgue measurable sets* and $|\cdot|_n$ (or simply $|\cdot|$ if no confusion arises) is called the *n-dimensional Lebesgue measure*. Unless explicitly said, from now on, *measurable* means *Lebesgue measurable* and the measure is the Lebesgue measure.

Not every subset of \mathbb{R}^n is measurable. However, the nonmeasurable ones are quite . . . pathological¹ !

The sets of measure zero are quite important. Here are some examples: all countable sets, e.g. the set \mathbb{Q} of rational numbers; straight lines or smooth curves in \mathbb{R}^2 ; straight lines, hyperplanes, smooth curves and surfaces in \mathbb{R}^3 .

Notice that a straight line segment has measure zero in \mathbb{R}^2 but, of course not in \mathbb{R} .

We say that a *property holds almost everywhere in $A \in \mathcal{M}$* (in short, a.e. in A) if it holds at every point of A except that in a subset of measure zero.

For instance, the sequence $f_k(x) = \exp(-n \sin^2 x)$ converges to zero a.e. in \mathbb{R} , a Lipschitz function is differentiable a.e. in its domain (Rademacher's Theorem 1.3, p. 14).

The Lebesgue integral is defined for *measurable* functions, characterized by the fact that the inverse image of every closed set is measurable.

Definition B.5. Let $A \subseteq \mathbb{R}^n$ be measurable, and $f : A \rightarrow \mathbb{R}$. We say that f is measurable if

$$f^{-1}(C) \in \mathcal{F}$$

for any closed set $C \subseteq \mathbb{R}$.

If f is continuous, is measurable. The sum and the product of a finite number of measurable functions is measurable. The pointwise limit of a sequence of measurable functions is measurable.

If $f : A \rightarrow \mathbb{R}$ is measurable, we define its *essential supremum* or *least upper bound* by the formula:

$$\text{ess sup } f = \inf \{K : f \leq K \text{ a.e. in } A\}.$$

Note that, if $f = \chi_{\mathbb{Q}}$, the characteristic functions of the rational numbers, we have $\sup f = 1$, but $\text{esssup } f = 0$, since $|\mathbb{Q}| = 0$.

Every measurable function may be approximated by **simple functions**. A function $s : A \subseteq \mathbb{R}^n \rightarrow \mathbb{R}$ is said to be **simple** if its range is constituted by a *finite number* of values s_1, \dots, s_N , attained respectively on measurable sets A_1, \dots, A_N , contained in A . Introducing the characteristic functions χ_{A_j} , we may write

$$s = \sum_{j=1}^N s_j \chi_{A_j}.$$

¹ See e.g. [36], Rudin, 1976.

We have:

Theorem B.6. Let $f : A \rightarrow \mathbb{R}$, be measurable. There exists a sequence $\{s_k\}$ of simple functions converging pointwise to f in A . Moreover, if $f \geq 0$, we may choose $\{s_k\}$ increasing.

B.1.3 The Lebesgue integral

We define the Lebesgue integral of a measurable function on a measurable set A . For a simple function

$$s = \sum_{j=1}^N s_j \chi_{A_j}$$

we set:

$$\int_A s = \sum_{j=1}^N s_j |A_j|,$$

with the agreement that, if $s_j = 0$ and $|A_j| = +\infty$, then $s_j |A_j| = 0$.

If $f \geq 0$ is measurable, we define

$$\int_A f = \sup \int_A s,$$

where the supremum is computed over the set of all simple functions s such that $s \leq f$ in A .

In general, if f is measurable, we write $f = f^+ - f^-$, where $f^+ = \max\{f, 0\}$ and $f^- = \max\{-f, 0\}$ are the positive and negative parts of f , respectively. Then we set:

$$\int_A f = \int_A f^+ - \int_A f^-,$$

under the condition that at least one of the two integrals in the right hand side is finite.

If both these integrals are finite, the function f is said to be **integrable** or **summable** in A . From the definition, it follows immediately that a measurable function f is *integrable in A if and only if $|f|$ is integrable in A* .

All the functions Riemann integrable in a set A are Lebesgue integrable as well. An interesting example of non Lebesgue integrable function in $(0, +\infty)$ is given by $h(x) = \sin x/x$. In fact²

$$\int_0^{+\infty} \frac{|\sin x|}{x} dx = +\infty.$$

² We may write

$$\int_0^{+\infty} \frac{|\sin x|}{x} dx = \sum_{k=1}^{\infty} \int_{(k-1)\pi}^{k\pi} \frac{|\sin x|}{x} dx \geq \sum_{k=1}^{\infty} \frac{1}{k\pi} \int_{(k-1)\pi}^{k\pi} |\sin x| dx = \sum_{k=1}^{\infty} \frac{2}{k\pi} = +\infty.$$

On the contrary, it may be proved that

$$\lim_{N \rightarrow +\infty} \int_0^N \frac{\sin x}{x} dx = \frac{\pi}{2}$$

and therefore the *improper* Riemann integral of h is finite.

The set of the integrable functions in A is denoted by $L^1(A)$. If we identify two functions when they agree a.e. in A , $L^1(A)$ becomes a Banach space with the norm³

$$\|f\|_{L^1(A)} = \int_A |f| .$$

We denote by $L^1_{loc}(A)$ the set of *locally summable functions*, i.e. of the functions which are summable in every compact subset of A .

B.1.4 Some fundamental theorems

The following theorems are among the most important and useful in the theory of integration.

Theorem B.7 (Dominated Convergence Theorem). *Let $\{f_k\}$ be a sequence of summable functions in A such that $f_k \rightarrow f$ a.e. in A . If there exists $g \geq 0$, summable in A and such that $|f_k| \leq g$ a.e. in A , then f is summable and*

$$\|f_k - f\|_{L^1(A)} \rightarrow 0 \quad \text{as } k \rightarrow +\infty.$$

In particular

$$\lim_{k \rightarrow \infty} \int_A f_k = \int_A f .$$

Theorem B.8. *Let $\{f_k\}$ be a sequence of summable functions in A such that*

$$\|f_k - f\|_{L^1(A)} \rightarrow 0 \quad \text{as } k \rightarrow +\infty.$$

Then there exists a subsequence $\{f_{k_j}\}$ such that $f_{k_j} \rightarrow f$ a.e. as $j \rightarrow +\infty$.

Theorem B.9 (Monotone Convergence Theorem). *Let $\{f_k\}$ be a sequence of non-negative, measurable functions in A such that*

$$f_1 \leq f_2 \leq \dots \leq f_k \leq f_{k+1} \leq \dots .$$

Then

$$\lim_{k \rightarrow \infty} \int_A f_k = \int_A \lim_{k \rightarrow \infty} f_k .$$

³ See Chap. 6.

Example B.10. A typical situation we often encounter in this book is the following. Let $f \in L^1(A)$ and, for $\varepsilon > 0$, set $A_\varepsilon = \{|f| > \varepsilon\}$. Then, we have

$$\int_{A_\varepsilon} f \rightarrow \int_A f \quad \text{as } \varepsilon \rightarrow 0.$$

This follows from Theorem B.7 since, for every sequence $\varepsilon_j \rightarrow 0$, we have $|f| \chi_{A_{\varepsilon_j}} \leq |f|$ and therefore

$$\int_{A_{\varepsilon_j}} f = \int_A f \chi_{A_{\varepsilon_j}} \rightarrow \int_A f \quad \text{as } \varepsilon \rightarrow 0.$$

Let $C_0(A)$ be the set of continuous functions in A , compactly supported in A . An important fact is that any summable function may be approximated by a function in $C_0(A)$.

Theorem B.11. *Let $f \in L^1(A)$. Then, for every $\delta > 0$, there exists a continuous function $g \in C_0(A)$ such that*

$$\|f - g\|_{L^1(A)} < \delta.$$

The fundamental theorem of calculus extends to the Lebesgue integral in the following form:

Theorem B.12 (Differentiation). *Let $f \in L^1_{loc}(\mathbb{R})$. Then*

$$\frac{d}{dx} \int_a^x f(t) dt = f(x) \quad \text{a.e. } x \in \mathbb{R}.$$

Finally, the integral of a summable function can be computed via iterated integrals in any order. Precisely, let

$$I_1 = \{\mathbf{x} \in \mathbb{R}^n : -\infty \leq a_i < x_i < b_i \leq \infty; i = 1, \dots, n\}$$

and

$$I_2 = \{\mathbf{y} \in \mathbb{R}^m : -\infty \leq a_j < y_j < b_j \leq \infty; j = 1, \dots, m\}.$$

Theorem B.13 (Fubini). *Let f be summable in $I = I_1 \times I_2 \subset \mathbb{R}^n \times \mathbb{R}^m$. Then*

1. $f(\mathbf{x}, \cdot) \in L^1(I_2)$ for a.e. $\mathbf{x} \in I_1$, and $f(\cdot, \mathbf{y}) \in L^1(I_1)$ for a.e. $\mathbf{y} \in I_2$.
2. $\int_{I_2} f(\cdot, \mathbf{y}) d\mathbf{y} \in L^1(I_1)$ and $\int_{I_1} f(\mathbf{x}, \cdot) d\mathbf{x} \in L^1(I_2)$.
3. The following formulas hold:

$$\int_I f(\mathbf{x}, \mathbf{y}) d\mathbf{x} d\mathbf{y} = \int_{I_1} d\mathbf{x} \int_{I_2} f(\mathbf{x}, \mathbf{y}) d\mathbf{y} = \int_{I_2} d\mathbf{y} \int_{I_1} f(\mathbf{x}, \mathbf{y}) d\mathbf{x}.$$

Let $f \in L^1(B_R(\mathbf{p}))$. The following formula can be considered a version of Fubini's Theorem in spherical coordinates:

$$\int_{B_R(\mathbf{p})} f(\mathbf{x}) d\mathbf{x} = \int_0^R ds \int_{\partial B_s(\mathbf{p})} f(\boldsymbol{\sigma}) d\sigma.$$

B.1.5 Probability spaces, random variables and their integrals

Let \mathcal{F} be a σ -algebra in a set Ω . A *probability measure* P on \mathcal{F} is a measure in the sense of definition B.2, such that $P(\Omega) = 1$ and

$$P : \mathcal{F} \rightarrow [0, 1].$$

The triplet (Ω, \mathcal{F}, P) is called a *probability space*. In this setting, the elements ω of Ω are *sample points*, while a set $A \in \mathcal{F}$ has to be interpreted as an *event*. $P(A)$ is the probability of (occurrence of) A .

A typical example is given by the triplet

$$\Omega = [0, 1], \mathcal{F} = \mathcal{M} \cap [0, 1], P(A) = |A|$$

which models a *uniform random choice* of a point in $[0, 1]$.

A *1-dimensional random variable* in (Ω, \mathcal{F}, P) is a function

$$X : \Omega \rightarrow \mathbb{R}$$

such that X is \mathcal{F} -measurable, that is

$$X^{-1}(C) \in \mathcal{F}$$

for each closed set $C \subseteq \mathbb{R}$.

Example B.14. The number k of steps to the right after N steps in the random walk of Sect. 2.4 is a random variable. Here Ω is the set of walks of N steps.

By the same procedure used to define the Lebesgue integral we can define the integral of a random variable with respect to a probability measure. We sketch the main steps.

If X is *simple*, i.e. $X = \sum_{j=1}^N s_j \chi_{A_j}$, we define

$$\int_{\Omega} X dP = \sum_{j=1}^N s_j P(A_j).$$

If $X \geq 0$ we set

$$\int_{\Omega} X dP = \sup \left\{ \int_{\Omega} Y dP : Y \leq X, Y \text{ simple} \right\}.$$

Finally, if $X = X^+ - X^-$ we define

$$\int_{\Omega} X \, dP = \int_{\Omega} X^+ \, dP - \int_{\Omega} X^- \, dP$$

provided at least one of the integral on the right hand side is finite.

In particular, if

$$\int_{\Omega} |X| \, dP < \infty,$$

then

$$E(X) = \langle X \rangle = \int_{\Omega} X \, dP$$

is called *the expected value (or mean value or expectation)* of X , while

$$\text{Var}(X) = \int_{\Omega} (X - E(X))^2 \, dP$$

is called the *variance of X* .

Analogous definitions can be given componentwise for n -dimensional random variables

$$\mathbf{X} : \Omega \rightarrow \mathbb{R}^n.$$

Appendix C

Identities and Formulas

C.1 Gradient, Divergence, Curl, Laplacian

Let \mathbf{F} be a smooth vector field and f a smooth real function, in \mathbb{R}^3 .

Orthogonal cartesian coordinates

1. *gradient:*

$$\nabla f = \frac{\partial f}{\partial x}\mathbf{i} + \frac{\partial f}{\partial y}\mathbf{j} + \frac{\partial f}{\partial z}\mathbf{k}.$$

2. *divergence* ($\mathbf{F} = F_1\mathbf{i} + F_2\mathbf{j} + F_3\mathbf{k}$):

$$\operatorname{div} \mathbf{F} = \frac{\partial}{\partial x}F_1 + \frac{\partial}{\partial y}F_2 + \frac{\partial}{\partial z}F_3.$$

3. *Laplacian:*

$$\Delta f = \frac{\partial^2 f}{\partial x^2} + \frac{\partial^2 f}{\partial y^2} + \frac{\partial^2 f}{\partial z^2}.$$

4. *curl:*

$$\operatorname{curl} \mathbf{F} = \begin{vmatrix} \mathbf{i} & \mathbf{j} & \mathbf{k} \\ \frac{\partial}{\partial x} & \frac{\partial}{\partial y} & \frac{\partial}{\partial z} \\ F_1 & F_2 & F_3 \end{vmatrix}.$$

Cylindrical coordinates

$$x = r \cos \theta, \quad y = r \sin \theta, \quad z = z \quad (r > 0, \quad 0 \leq \theta \leq 2\pi)$$

$$\mathbf{e}_r = \cos \theta \mathbf{i} + \sin \theta \mathbf{j}, \quad \mathbf{e}_\theta = -\sin \theta \mathbf{i} + \cos \theta \mathbf{j}, \quad \mathbf{e}_z = \mathbf{k}.$$

1. *gradient:*

$$\nabla f = \frac{\partial f}{\partial r}\mathbf{e}_r + \frac{1}{r} \frac{\partial f}{\partial \theta}\mathbf{e}_\theta + \frac{\partial f}{\partial z}\mathbf{e}_z.$$

674 Appendix C Identities and Formulas

2. *divergence* ($\mathbf{F} = F_r \mathbf{e}_r + F_\theta \mathbf{e}_\theta + F_z \mathbf{k}$):

$$\operatorname{div} \mathbf{F} = \frac{1}{r} \frac{\partial}{\partial r} (r F_r) + \frac{1}{r} \frac{\partial}{\partial \theta} F_\theta + \frac{\partial}{\partial z} F_z.$$

3. *Laplacian*:

$$\Delta f = \frac{1}{r} \frac{\partial}{\partial r} \left(r \frac{\partial f}{\partial r} \right) + \frac{1}{r^2} \frac{\partial^2 f}{\partial \theta^2} + \frac{\partial^2 f}{\partial z^2} = \frac{\partial^2 f}{\partial r^2} + \frac{1}{r} \frac{\partial f}{\partial r} + \frac{1}{r^2} \frac{\partial^2 f}{\partial \theta^2} + \frac{\partial^2 f}{\partial z^2}.$$

4. *curl*:

$$\operatorname{curl} \mathbf{F} = \frac{1}{r} \begin{vmatrix} \mathbf{e}_r & r \mathbf{e}_\theta & \mathbf{e}_z \\ \partial_r & \partial_\theta & \partial_z \\ F_r & r F_\theta & F_z \end{vmatrix}.$$

Spherical coordinates

$$x = r \cos \theta \sin \psi, \quad y = r \sin \theta \sin \psi, \quad z = r \cos \psi \quad (r > 0, 0 \leq \theta \leq 2\pi, 0 \leq \psi \leq \pi)$$

$$\mathbf{e}_r = \cos \theta \sin \psi \mathbf{i} + \sin \theta \sin \psi \mathbf{j} + \cos \psi \mathbf{k}$$

$$\mathbf{e}_\theta = -\sin \theta \mathbf{i} + \cos \theta \mathbf{j}$$

$$\mathbf{e}_\psi = \cos \theta \cos \psi \mathbf{i} + \sin \theta \cos \psi \mathbf{j} - \sin \psi \mathbf{k}.$$

1. *gradient*:

$$\nabla f = \frac{\partial f}{\partial r} \mathbf{e}_r + \frac{1}{r \sin \psi} \frac{\partial f}{\partial \theta} \mathbf{e}_\theta + \frac{1}{r} \frac{\partial f}{\partial \psi} \mathbf{e}_\psi.$$

2. *divergence* ($\mathbf{F} = F_r \mathbf{e}_r + F_\theta \mathbf{e}_\theta + F_\psi \mathbf{e}_\psi$):

$$\operatorname{div} \mathbf{F} = \underbrace{\frac{\partial}{\partial r} F_r + \frac{2}{r} F_r}_{\text{radial part}} + \underbrace{\frac{1}{r} \left[\frac{1}{\sin \psi} \frac{\partial}{\partial \theta} F_\theta + \frac{\partial}{\partial \psi} F_\psi + \cot \psi F_\psi \right]}_{\text{spherical part}}.$$

3. *Laplacian*:

$$\Delta f = \underbrace{\frac{\partial^2 f}{\partial r^2} + \frac{2}{r} \frac{\partial f}{\partial r}}_{\text{radial part}} + \underbrace{\frac{1}{r^2} \left\{ \frac{1}{(\sin \psi)^2} \frac{\partial^2 f}{\partial \theta^2} + \frac{\partial^2 f}{\partial \psi^2} + \cot \psi \frac{\partial f}{\partial \psi} \right\}}_{\text{spherical part (Laplace-Beltrami operator)}}.$$

4. *curl*:

$$\operatorname{rot} \mathbf{F} = \frac{1}{r^2 \sin \psi} \begin{vmatrix} \mathbf{e}_r & r \mathbf{e}_\psi & r \sin \psi \mathbf{e}_\theta \\ \partial_r & \partial_\psi & \partial_\theta \\ F_r & r F_\psi & r \sin \psi F_z \end{vmatrix}.$$

C.2 Formulas

Gauss' formulas

In \mathbb{R}^n , $n \geq 2$, let:

- Ω be a bounded smooth domain and ν the outward unit normal on $\partial\Omega$.
- \mathbf{u}, \mathbf{v} be vector fields of class $C^1(\overline{\Omega})$.
- φ, ψ be real functions of class $C^1(\overline{\Omega})$.
- $d\sigma$ be the area element on $\partial\Omega$.

1. $\int_{\Omega} \operatorname{div} \mathbf{u} d\mathbf{x} = \int_{\partial\Omega} \mathbf{u} \cdot \nu d\sigma$ (Divergence Theorem).
2. $\int_{\Omega} \nabla \varphi d\mathbf{x} = \int_{\partial\Omega} \varphi \nu d\sigma.$
3. $\int_{\Omega} \Delta \varphi d\mathbf{x} = \int_{\partial\Omega} \nabla \varphi \cdot \nu d\sigma = \int_{\partial\Omega} \partial_{\nu} \varphi d\sigma.$
4. $\int_{\Omega} \psi \operatorname{div} \mathbf{F} d\mathbf{x} = \int_{\partial\Omega} \psi \mathbf{F} \cdot \nu d\sigma - \int_{\Omega} \nabla \psi \cdot \mathbf{F} d\mathbf{x}$ (Integration by parts).
5. $\int_{\Omega} \psi \Delta \varphi d\mathbf{x} = \int_{\partial\Omega} \psi \partial_{\nu} \varphi d\sigma - \int_{\Omega} \nabla \varphi \cdot \nabla \psi d\mathbf{x}$ (Green's identity I).
6. $\int_{\Omega} (\psi \Delta \varphi - \varphi \Delta \psi) d\mathbf{x} = \int_{\partial\Omega} (\psi \partial_{\nu} \varphi - \varphi \partial_{\nu} \psi) d\sigma$ (Green's identity II).
7. $\int_{\Omega} \operatorname{curl} \mathbf{u} d\mathbf{x} = - \int_{\partial\Omega} \mathbf{u} \times \nu d\sigma.$
8. $\int_{\Omega} \mathbf{u} \cdot \operatorname{curl} \mathbf{v} d\mathbf{x} = \int_{\Omega} \mathbf{v} \cdot \operatorname{curl} \mathbf{u} d\mathbf{x} - \int_{\partial\Omega} (\mathbf{u} \times \mathbf{v}) \cdot \nu d\sigma.$

Identities

1. $\operatorname{div} \operatorname{curl} \mathbf{u} = 0.$
2. $\operatorname{curl} \nabla \varphi = \mathbf{0}.$
3. $\operatorname{div} (\varphi \mathbf{u}) = \varphi \operatorname{div} \mathbf{u} + \nabla \varphi \cdot \mathbf{u}.$
4. $\operatorname{curl} (\varphi \mathbf{u}) = \varphi \operatorname{curl} \mathbf{u} + \nabla \varphi \times \mathbf{u}.$
5. $\operatorname{curl} (\mathbf{u} \times \mathbf{v}) = (\mathbf{v} \cdot \nabla) \mathbf{u} - (\mathbf{u} \cdot \nabla) \mathbf{v} + (\operatorname{div} \mathbf{v}) \mathbf{u} - (\operatorname{div} \mathbf{u}) \mathbf{v}.$
6. $\operatorname{div} (\mathbf{u} \times \mathbf{v}) = \operatorname{curl} \mathbf{u} \cdot \mathbf{v} - \operatorname{curl} \mathbf{v} \cdot \mathbf{u}.$
7. $\nabla (\mathbf{u} \cdot \mathbf{v}) = \mathbf{u} \times \operatorname{curl} \mathbf{v} + \mathbf{v} \times \operatorname{curl} \mathbf{u} + (\mathbf{u} \cdot \nabla) \mathbf{v} + (\mathbf{v} \cdot \nabla) \mathbf{u}.$
8. $(\mathbf{u} \cdot \nabla) \mathbf{u} = \operatorname{curl} \mathbf{u} \times \mathbf{u} + \frac{1}{2} \nabla |\mathbf{u}|^2.$
9. $\operatorname{curl} \operatorname{curl} \mathbf{u} = \nabla (\operatorname{div} \mathbf{u}) - \Delta \mathbf{u}.$

References

Partial Differential Equations

- [1] *DiBenedetto, E.*: Partial Differential Equations, Birkhäuser, Boston, 1995.
- [2] *Friedman, A.*: Partial Differential Equations of parabolic Type, Prentice-Hall, Englewood Cliffs, 1964.
- [3] *Gilbarg, D. and Trudinger, N.*: Elliptic Partial Differential Equations of Second Order, 2nd ed., Springer-Verlag, Berlin Heidelberg, 1998.
- [4] *Grisvard, P.*: Elliptic Problems in nonsmooth domains, Pitman, Boston, 1985.
- [5] *Guenther, R.B. and Lee, J.W.*: Partial Differential Equations of Mathematical Physics and Integral Equations, Dover Publications, Inc., New York, 1998.
- [6] *Helms, O.*: Introduction to Potential Theory, Krieger Publishing Company, New York, 1975.
- [7] *John, F.*: Partial Differential Equations, 4th ed., Springer-Verlag, New York, 1982.
- [8] *Kellogg, O.*: Foundations of Potential Theory, Dover, New York, 1954.
- [9] *Galdi, G.*: Introduction to the Mathematical Theory of Navier-Stokes Equations, vols. I and II, Springer-Verlag, New York, 1994.
- [10] *Lieberman, G.M.*: Second Order Parabolic Partial Differential Equations, World Scientific, Singapore, 1996.
- [11] *Lions, J.L., Magenes, E.*: Nonhomogeneous Boundary Value Problems and Applications, vols. 1, 2, Springer-Verlag, New York, 1972.
- [12] *McOwen, R.*: Partial Differential Equations: Methods and Applications, Prentice-Hall, New Jersey, 1996.
- [13] *Olver, P.J.*: Introduction to Partial Differential Equations, Springer International Publishing Switzerland, 2014.
- [14] *Protter, M. and Weinberger, H.*: Maximum Principles in Differential Equations, Prentice-Hall, Englewood Cliffs, 1984.
- [15] *Renardy, M. and Rogers, R.C.*: An Introduction to Partial Differential Equations, Springer-Verlag, New York, 1993.
- [16] *Rauch, J.*: Partial Differential Equations, Springer-Verlag, Heidelberg, 1992.

- [17] *Salsa, S. and Verzini, G.*: Partial Differential Equation in Action. Complements and Exercises, Springer International Publishing Switzerland, 2015.
- [18] *Smoller, J.*: Shock Waves and Reaction-Diffusion Equations, Springer-Verlag, New York, 1983.
- [19] *Strauss, W.*: Partial Differential Equation: An Introduction, Wiley, 1992.
- [20] *Widder, D.V.*: The Heat Equation, Academic Press, New York, 1975.

Mathematical Modelling

- [21] *Acheson, A.J.*: Elementary Fluid Dynamics, Clarendon Press, Oxford, 1990.
- [22] *Billingham, J. and King, A.C.*: Wave Motion, Cambridge University Press, 2000.
- [23] *Courant, R. and Hilbert, D.*: Methods of Mathematical Physics, vols. 1 and 2, Wiley, New York, 1953.
- [24] *Dautray, R. and Lions, J.L.*: Mathematical Analysis and Numerical Methods for Science and Technology, vols. 1–5, Springer-Verlag, Berlin Heidelberg, 1985.
- [25] *Lin, C.C. and Segel, L.A.*: Mathematics Applied to Deterministic Problems in the Natural Sciences, SIAM Classics in Applied Mathematics, 4th ed., 1995.
- [26] *Murray, J.D.*: Mathematical Biology, vols. 1 and 2, Springer-Verlag, Berlin Heidelberg, 2001.
- [27] *Rhee, H., Aris, R., and Amundson, N.*: First Order Partial Differential Equations, vols. 1 and 2, Dover, New York, 1986.
- [28] *Scherzer, O., Grasmair, M., Grossauer, H.; Haltmeier, M., and Lenzen, F.*: Variational Methods in Imaging, Applied Mathematical Sciences 167, Springer, New York, 2008.
- [29] *Segel, L.A.*: Mathematics Applied to Continuum Mechanics, Dover Publications, Inc., New York, 1987.
- [30] *Whitham, G.B.*: Linear and Nonlinear Waves, Wiley-Interscience, 1974.

ODEs, Analysis and Functional Analysis

- [31] *Adams, R.*: Sobolev Spaces, Academic Press, New York, 1975.
- [32] *Brezis, H.*: Functional Analysis, Sobolev Spaces and Partial Differential Equations, Springer, New York, 2010.
- [33] *Coddington, E.A. and Levinson, N.*: Theory of Ordinary Differential Equations, McGraw-Hill, New York, 1955.
- [34] *Gelfand, I.M. and Shilov, E.*: Generalized Functions, vol. 1: Properties and Operations, Academic Press, 1964.
- [35] *Maz'ya, V.G.*: Sobolev Spaces, Springer-Verlag, Berlin Heidelberg, 1985.
- [36] *Rudin, W.*: Principles of Mathematical Analysis, 3rd ed., McGraw-Hill, 1976.
- [37] *Schwartz, L.*: Théorie des Distributions, Hermann, Paris, 1966.
- [38] *Taylor, A.E.*: Introduction to Functional Analysis, John Wiley & Sons, 1958.
- [39] *Yoshida, K.*: Functional Analysis, 3rd ed., Springer-Verlag, New York, 1971.

- [40] *Ziemer, W.*: Weakly Differentiable Functions, Springer-Verlag, Berlin Heidelberg, 1989.
- [41] *Zygmund, R. and Wheeden, R.*: Measure and Integral, Marcel Dekker, 1977.

Numerical Analysis

- [42] *Dautray, R. and Lions, J.L.*: Mathematical Analysis and Numerical Methods for Science and Technology, vols. 4 and 6, Springer-Verlag, Berlin Heidelberg, 1985.
- [43] *Quarteroni, A.*: Numerical Models for Differential Problems, MS&A, Springer-Verlag Italia, Milan, 2014.
- [44] *Quarteroni, A. and Valli, A.*: Numerical Approximation of Partial Differential Equations, Springer-Verlag, Berlin Heidelberg, 1994.
- [45] *Godlewski, E. and Raviart, P.A.*: Numerical Approximation of Hyperbolic Systems of Conservation Laws, Springer-Verlag, New York, 1996.

Stochastic Processes and Finance

- [46] *Baxter, M. and Rennie, A.*: Financial Calculus: An Introduction to Derivative Pricing, Cambridge University Press, 1996.
- [47] *Øksendal, B.K.*: Stochastic Differential Equations: An Introduction with Applications, 4th ed., Springer-Verlag, Berlin Heidelberg, 1995.
- [48] *Wilmott, P., Howison, S., and Dewinne, J.*: The Mathematics of Financial Derivatives. A Student Introduction, Cambridge University Press, 1996.

Index

A

Absorbing barriers 111
 Adjoint of a bilinear form 402
 Adjoint problem 574
 Advection 180
 Alternative
 – for the Dirichlet problem 526
 – for the Neumann problem 529
 Angular frequency 260
 Arbitrage 92

B

Barenblatt solutions 103
 Barrier 143
 Bernoulli's equation 330
 Bessel function 73, 297
 Bilinear form 382
 Bond number 333
 Boundary conditions 21
 – Dirichlet 21, 33
 – mixed 22, 33
 – Neumann 22, 33
 – Robin 22, 33
 Breaking time 201
 Brownian motion 55
 Brownian path 55
 Burgers, viscous 216

C

Canonical form 293, 295
 Canonical isometry 379
 Characteristic 181, 230, 620, 621
 – parallelogram 277
 – strip 246
 – system 245
 Chebyshev polynomials 370

Closure

8
 Comparison 38
 Compatibility conditions 403, 405
 Condition
 – compatibility 118
 – E 220
 – entropy 638
 – Rankine-Hugoniot 631
 Conjugate exponent 11, 358
 Conormal derivative 527
 Contact discontinuity 626, 632, 640
 Continuous isomorphism 385
 Convection 61
 Convergence
 – least squares 660
 – uniform 11
 – weak 394
 Convolution 430, 449
 Cost functional 571
 Critical mass 75
 Critical survival value 66
 Curve
 – rarefaction 633
 – shock 640

D

d'Alembert formula 276
 d-harmonic function 121
 Darcy's law 102
 Diffusion 18
 Diffusion coefficient 54
 Dirac comb 436
 Dirac measure 44
 Direct product 452
 Direct sum 364
 Dirichlet eigenfunctions 517

- Dirichlet Principle 546
- Dispersion 287
 - relation 261, 287, 335
- Dissipation
 - external/internal 286
- Distribution 434
 - composition 445
 - division 448
- Distributional derivative 438
- Domain 8
 - C^1, C^k 12
 - Lipschitz 14
 - of dependence 278, 315
 - smooth 12
- Drift 60, 89
- Duhamel method 285

- E**
- Eigenfunction 370
 - of a bilinear form 411
- Eigenspace 407, 409, 411
- Eigenvalue 370, 409
 - of a bilinear form 411
- Eigenvector 409
- Elastic restoring force 111
- Elliptic equation 505
- Entropy condition 209, 210
- Equal area rule 202
- Equation
 - backward 94
 - backward heat 39
 - Bessel 73
 - Bessel's 372
 - biharmonic 555
 - Black-Scholes 3, 93
 - Buckley-Leverett 243
 - Burgers 4
 - Chebyshev 370
 - diffusion 2, 17
 - Eiconal 5
 - eikonal 249
 - elastostatics 557
 - elliptic 289
 - Fisher 4
 - fully nonlinear 2
 - Hermite's 371
 - hyperbolic 289
 - Klein-Gordon 287
 - Laplace 3
 - Legendre's 371
 - linear elasticity 5
 - linear, nonlinear 2
 - Maxwell 5
 - minimal surface 4
 - Navier 557
 - Navier Stokes 5
 - Navier-Stokes 153, 561
 - Navier-Stokes, stationary 566
 - parabolic 289, 581
 - parametric Bessel's (of order p) 372
 - partial differential 2
 - Poisson 3, 115
 - porous media 103
 - Porous medium 4
 - quasilinear 2
 - reduced wave 178
 - Schrodinger 4
 - semilinear 2
 - stationary Fisher 553
 - stochastic differential 89
 - Sturm-Liouville 370
 - transport 2
 - Tricomi 289
 - uniformly parabolic 582
 - vibrating plate 3
 - wave 3
- Equicontinuity 392
- Equipartition of energy 342
- Escape probability 137
- Essential support 429
- Essential supremum 358
- Euler equation 388
- European options 88
- Expectation 58, 70
- Expiry date 88
- Extension operator 477
- Exterior Dirichlet problem 163
- Exterior domain 164
- Exterior Robin problem 165, 177

- F**
- Fick's law 61
- Final payoff 94
- First exit time 135
- First integral 238, 240
- First variation 388
- Flux function 179
- Focussing effect 345
- Forward cone 313
- Fourier coefficients 367
- Fourier law 20
- Fourier series 28
- Fourier transform 454, 473
- Fourier-Bessel series 74, 373
- Frequency 260
- Froude number 333

Function

- Bessel's of first kind and order p 373
- characteristic 10
- compactly supported 10
- complementary error 220
- continuous 10
- d-harmonic 119
- essentially bounded 358
- Green's 157
- Hölder continuous 357
- harmonic 18, 115
- Heaviside 44
- piecewise continuous 205
- summable 357
- test 48, 429
- weight 370

Functional 377

Fundamental solution 43, 48, 148, 282, 311

G

- Gas dynamics 617
- Gaussian law 56, 68
- Genuinely nonlinear 634
- Global Cauchy problem 23, 34, 76
 - nonhomogeneous 80
- Gram-Schmidt process 369
- Greatest lower bound 9
- Group velocity 261

H

- Harmonic lifting 141
- Harmonic measure 138
- Harnack's inequality 131
- Heisenberg Uncertainty Principle
 - for the first eigenvalue 421
- Helmholtz decomposition formula 151
- Hermite polynomials 371
- Hilbert triplet 401
- Hooke's law 556
- Hopf's maximum principle 126
- Hopf-Cole transformation 218
- Hugoniot line 626

I

- Identity
 - Green's (first and second) 15
 - strong Parseval's 460
 - weak Parseval's 458
- Inequality
 - Hölder 358
- Infimum 9
- Inflow/outflow boundary 239

Inflow/outflow characteristics 185

- Inner product space 359
- Integral surface 230
- integration by parts 15
- Interior shere condition 126
- Invasion problem 113
- Inward heat flux 33
- Isometry
 - isometric 360
- Ito's formula 90

K

- Kernel 374
- Kinematic condition 331
- Kinetic energy 266

L

- Lagrange multiplier 565
- Lattice 66, 118
- Least squares 28
- Least upper bound 9
- Lebesgue spine 145
- Legendre polynomials 371
- Light cone 249
- Linearly degenerate 642
- Liouville Theorem 132
- little o 11
- Local chart 12
- Local wave speed 190
- Localization 477
- Logarithmic potential 150
- Logistic growth 105
- Lognormal density 91

M

- Mach number 308
- Markov properties 57, 69
- Mass conservation 60
- Material derivative 154
- Maximum principle 83, 120
 - weak 36, 532, 601
- Mean value property 123
- Method 23
 - Duhamel 81
 - electrostatic images 157
 - Galerkin's 388
 - of characteristics 189
 - of descent 316
 - of Faedo-Galerkin 591, 605
 - of stationary phase 263
 - reflection 477
 - separation of variables 23, 26, 269, 304, 407, 520

- time reversal 325
- vanishing viscosity 214
- Metric space 353
- Minimax property (of the eigenvalues)
 - for the first eigenvalue 416, 426
- Mollifier 430
- Monotone iteration scheme 551
- Multidimensional symmetric random walk 66
- Multiplicity (of an eigenvalue) 409

N

- Neumann eigenfunctions 519
- Neumann function 163
- Norm
 - Integral of order p 357
 - least squares 355
 - maximum 355
 - maximum of order k 356
- Normal probability density 43
- Normed space 353
- Numerical sets 7

O

- Omeomorphism 418
- Open covering 477
- Operator
 - adjoint 380
 - bounded,continuous 374
 - compact 397
 - discrete Laplace 119
 - linear 373
 - mean value 118
- Optimal control 572
- Optimal state 572
- Orthonormal basis 367

P

- Parabolic
 - boundary 23, 34
- Parabolic dilations 40
- Parallelogram law 360
- Partition of unity 478
- Perron method 142
- Phase speed 260
- Poincaré's inequality 466, 488
- Point 7
 - boundary 8
 - interior 7
 - limit 8
- Point source solution
 - two dimensional 345
- Poisson formula 131

- Potential 115
 - double layer 166
 - energy 267
 - Newtonian 149
 - retarded 318, 345
 - single layer 170

Principal Dirichlet eigenvalue 518

Principle of virtual work 560

Probability

- measure 670
- space 670

Problem

- abstract parabolic 586
- abstract variational 382
- Characteristic Cauchy 343
- eigenvalue 27
- Goursat 343
- ill posed (heat equation) 343
- inverse 325
- Riemann 623
- well posed 7, 21

Projected characteristics 238

Projection

- on closed convex sets 424
- Put-call parity 97

Q

- Quantum mechanics harmonic oscillator 422

R

- Random variable 54
- Random walk 49
 - with drift 58
- Range 374
 - of influence 278, 313
- Rankine-Hugoniot condition 198, 205
- Rarefaction/simple waves 194
- Rayleigh quotient 414, 518
- Reaction 63
- Reflecting barriers 111
- Regular point 144
- Resolvent 407
 - of a bilinear form 411
 - of a bounded operator 408
- Retarded potential 318
- retrocone 301
- Reynolds number 562
- Riemann invariant 628
- Riemann problem 212
- Rodrigues' formula 371

S

- Schwarz inequality 360
- Schwarz reflection principle 175
- Self-financing portfolio 92, 100
- Selfadjoint operator 381
- Sequence
 - Cauchy 354
 - fundamental 354
- Set
 - bounded 8
 - closed 8
 - compact 8
 - compactly contained 8
 - connected 8
 - convex 8
 - dense 8
 - open 8
 - precompact 391
 - sequentially closed 8
 - sequentially compact 8, 391
- Shock
 - curve 198
 - speed 198
 - wave 198
- Similarity, self-similar solutions 41
- Sobolev exponent 491
- Solution
 - classical 507
 - distributional 507
 - integral 205
 - self-similar 103
 - steady state 25
 - strong 507
 - unit source 46
 - variational 507
 - viscosity 507
 - weak 205
- Sommerfeld condition 178
- Space
 - separable 367
- Space-like curve 250
- Spectral decomposition
 - of a matrix 407
 - of an operator 411
- Spectrum 407
 - continuous 409
 - of a bilinear form 411
 - of a bounded operator 408
 - point 409
 - residual 409
- Spherical waves 261
- Stability estimate 386
- Standing wave 271

- Stationary phase (method of) 340
- Steepest descent 575
- Stiffness matrix 389
- Stochastic process 55, 68
- Stokes System
 - equazione biarmonica 562
- Stopping time 57, 135
- Strike price 88
- Strip condition 247
- Strong Huygens' principle 313, 315
- Sub/superharmonic function 141
- Sub/supersolution 36
- Superposition principle 17, 77, 268
- Support 10
 - of a distribution 437
- Surface
 - of the unit sphere (ω_n) 7
 - tension 328, 330
- Symbol $o(h)$ 53
- Symbol “big O” 63
- System
 - hyperbolic 616
 - p-system 619

T

- Tempered distribution 456
- Tensor
 - deformation 556
 - stress 153, 556
- Tensor product 452
- Term by term
 - differentiation 12
 - integration 12
- Theorem
 - Ascoli-Arzelà 392
 - Contraction mapping 417
 - Dominated Convergence 668
 - Fubini 669
 - Lax-Milgram 383
 - Leray-Shauder 420
 - Monotone Convergence 668
 - projection 364
 - Rellich 487
 - Riesz's representation 378
 - Riesz-Fréchet-Kolmogoroff 393
 - Schauder 419
- time-like curve 250
- Topology 7, 354
 - euclidean 8
 - relative 9
- Trace 479
 - inequality 486
- Traffic in a tunnel 253

Transition function 69
Transition layer 216
Transition probability 57, 121
Transmission conditions 547
Travelling wave 182, 190, 215
trivial extension 477
Tychonov class 83

U

Uniform ellipticity 521
Unit impulse 45
Upper,lower limit 9

V

Value function 88
Variational formulation
– Dirichlet problem 510, 523
– Mixed problem 516, 530
– Neumann problem 513, 528
– Robin problem 515
Variational inequality
– on closed convex sets 424
Variational principle
– for the first eigenvalue 414
– for the k-th eigenvalue 415
Volatility 89

W

Wave
– capillarity 337
– cylindrical 296
– gravity 336
– harmonic 259
– incoming/outgoing 298
– linear 328
– linear gravity 346
– monochromatic/harmonic 296
– number 260
– packet 262
– plane 261, 296
– rarefaction, p-system 646, 647
– shock 640
– shock, p-system 645, 646
– simple 633
– spherical 297
– standing 260
– travelling 259

Weak coerciveness 525

Weak formulation

– Cauchy-Dirichlet problem 584
– Cauchy-Robin/Neumann problem 594–
596
– Initial-Dirichlet problem (wave eq.) 604
Weakly coercive (bilinear form) 402, 596
Weierstrass test 11, 29

Y

Young modulus 342