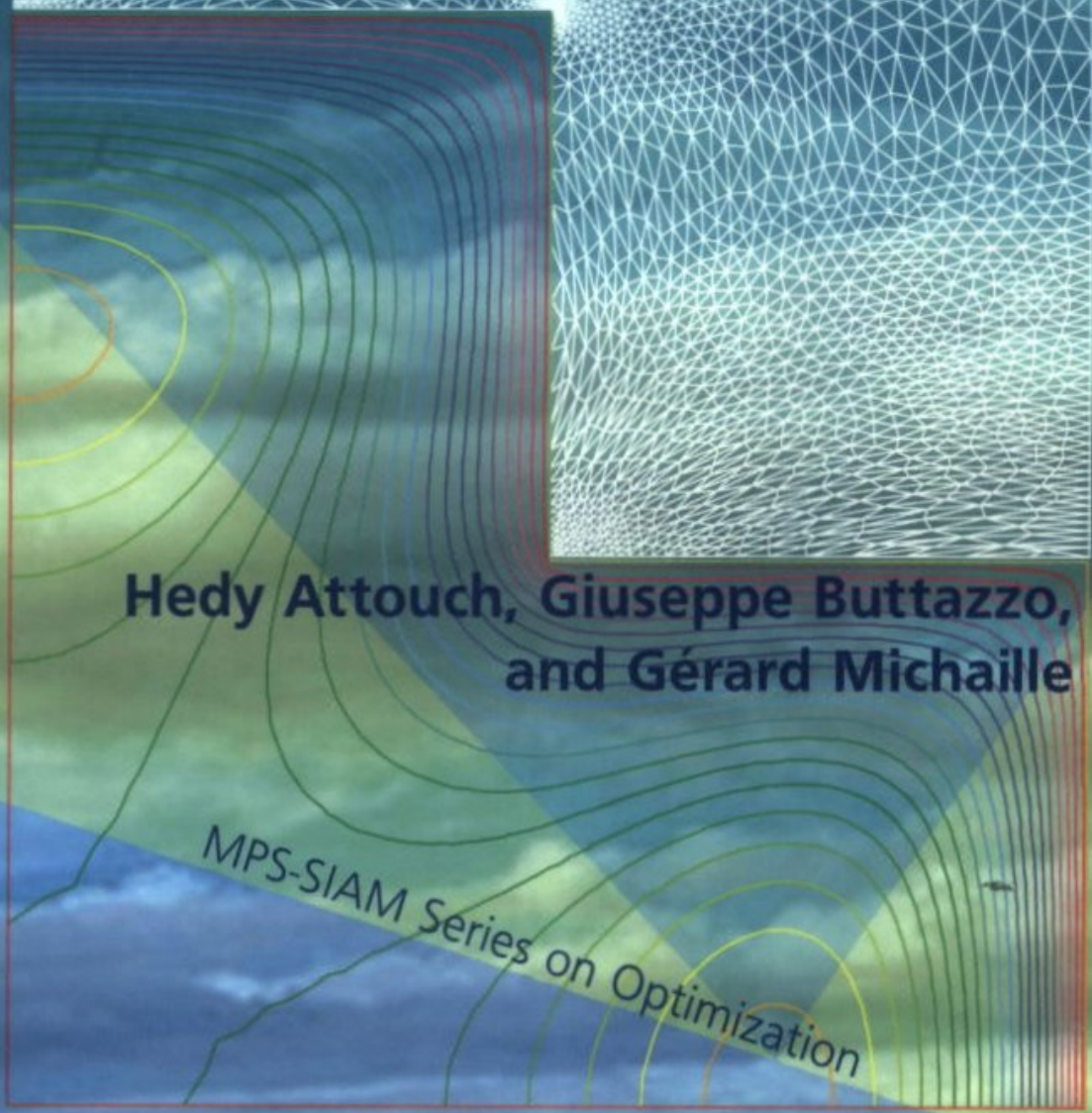
Variational Analysis in Sobolev and *BV* Spaces

Applications to PDEs and Optimization



Hedy Attouch, Giuseppe Buttazzo, and Gérard Michaille

MPS-SIAM Series on Optimization

Variational Analysis in Sobolev and *BV* Spaces

MPS-SIAM Series on Optimization

This series is published jointly by the Mathematical Programming Society and the Society for Industrial and Applied Mathematics. It includes research monographs, books on applications, textbooks at all levels, and tutorials. Besides being of high scientific quality, books in the series must advance the understanding and practice of optimization. They must also be written clearly and at an appropriate level.

Series Volumes

Attouch, Hedy, Buttazzo, Giuseppe, and Michaille, Gérard, *Variational Analysis in Sobolev and* BV *Spaces: Applications to PDEs and Optimization*

Wallace, Stein W. and Ziemba, William T., editors, *Applications of Stochastic Programming*

Grötschel, Martin, editor, *The Sharpest Cut: The Impact of Manfred Padberg and His Work*

Renegar, James, *A Mathematical View of Interior-Point Methods in Convex Optimization*

Ben-Tal, Aharon and Nemirovski, Arkadi, *Lectures on Modern Convex Optimization: Analysis, Algorithms, and Engineering Applications*

Conn, Andrew R., Gould, Nicholas I. M., and Toint, Phillippe L., *Trust-Region Methods*

Variational Analysis in Sobolev and *BV* Spaces

Applications to PDEs and Optimization

Hedy Attouch
Université Montpellier II
Montpellier, France

Giuseppe Buttazzo
Università di Pisa
Pisa, Italy

Gérard Michaille
Université Montpellier II
Montpellier, France

siam
Society for Industrial and Applied Mathematics
Philadelphia

MPS
Mathematical Programming Society
Philadelphia

10 9 8 7 6 5 4 3 2 1

Library of Congress Cataloging-in-Publication Data

Attouch, H.
Variational analysis in Sobolev and BV spaces : applications to PDEs and optimization / Hedy Attouch, Giuseppe Buttazzo, Gérard Michaille.
p. cm. – (MPS-SIAM series on optimization)
Includes bibliographical references and index.
ISBN 0-89871-600-4 (pbk.)
Mathematical optimization. 2. Calculus of variations. 3. Sobolev spaces. 4. Functions of bounded variation. 5. Differential equations, Partial. I. Buttazzo, Giuseppe. II. Michaille, Gérard. III. Title. IV. Series.

QA402.5.A84 2005
519.6—dc22

2005051592

Contents

Preface

Most of the material in this book comes from graduate-level courses on variational analysis, PDEs, and optimization which have been given during the last decades by the authors, H. Attouch and G. Michaille at the University of Montpellier (France) and G. Buttazzo at the University of Pisa (Italy). Our objective is twofold.

The first objective is to provide to students the basic tools and methods of variational analysis and optimization in infinite dimensional spaces together with applications to classical PDE problems. This corresponds to the first part of the book, chapters 1 through 9, and takes place in classical Sobolev spaces. We have made an effort to provide, as much as possible, a self-contained exposition, and we try to introduce each new development from various perspectives (historical, numerical, ...).

The second objective, which is oriented more toward research, is to present new trends in variational analysis and some of the most recent developments and applications. This corresponds to the second part of the book, chapters 10 through 16, where in particular are introduced the $BV(\Omega)$ spaces.

This organization is intended to make the book accessible to a large audience, from students to researchers, with various backgrounds in mathematics, as well as physicists, engineers, and others. As a guideline, we try to portray direct methods in modern variational analysis—one century after D. Hilbert delineated them in his famous lecture at Collège de France, Paris, 1900. The extraordinary success of these methods is intimately linked with the development, throughout the 20th century, of new branches in mathematics: functional analysis, measure theory, numerical analysis, (nonlinear) PDEs, and optimization.

We try to show in this book the interplay among all these theories and also between theory and applications. Variational methods have proved to be very flexible. In recent years, they have been developed to study a number of advanced problems of modern technology, like composite materials, phase transitions, thin structures, large deformations, fissures, and shape optimization.

To grasp these often involved phenomena, the classical framework of variational analysis, which is presented in the first part, must be enlarged. This is the motivation for the introduction in the second part of the book of some advanced techniques, like BV and SBV spaces, Young measures, Γ-convergence, recession analysis, and relaxation methods.

Finally, we wish to stress that variational analysis is a remarkable example of international collaboration. All mathematical schools have contributed to its success, and it is a modest symbol that this book has been written in collaboration between mathematicians of two schools, French and Italian. This book owes much to the support of the Universities

of Montpellier (France) and Pisa (Italy) and of their mathematical departments, and the convention of cooperation that connects them.

Acknowledgments. We would like to express our sincere thanks to all the students and colleagues whose comments and encouragement helped us in writing the final manuscript. Year by year, the redaction of the book profited much from their comments.

We are grateful to our colleagues in the continuous optimization community, who strongly influenced the contents of the book and encouraged us from the very beginning in writing this book. The chapter on convex analysis benefitted much from the careful reading of L. Thibault and M. Valadier.

We would like to thank SIAM and the editorial board of the MPS-SIAM Series on Optimization for the quality of the editing process. We address special thanks to B. Lacan in Montpellier, who helped us when we started the project.

Finally, we take this opportunity to express our consideration and gratitude to H. Brezis and E. De Giorgi, who were our first guides in the discovery of this fascinating world of variational methods and their applications.

Chapter 1
Introduction

Let us detail the contents of each of the two parts of the book.

Part I: Basic Variational Principles. In Part I, we follow as a guideline the variational treatment of the celebrated Dirichlet problem. We show how the program of D. Hilbert, which was first delineated in his famous lecture at Collège de France in 1900 [154], has been progressively solved throughout the 20^{th} century. We introduce the basic elements of variational analysis which allow one to solve this classical problem and closely related ones, like the Neumann problem and the Stokes problem.

Chapter 2 contains an extensive exposition of weak solutions methods in variational analysis and of the accompanying notions: test functions, the distribution theory of L. Schwartz, weak convergences, and topologies.

Chapter 3 provides an exposition of the basic abstract variational principles. We enhance the importance of the direct method for solving minimization problems and put to the fore some of its basic topological ingredients: lower semicontinuity, coercivity, and inf-compactness. We show how weak topologies, reflexivity, and convexity properties come naturally into play. We insist on the modern approach to optimization theory where the concept of epigraph of a function plays a central role; see the recent monograph of Rockafellar and Wets [205] on variational analysis where the epigraphical analysis is systematically developed.

Chapter 4 contains some complements on geometric measure theory. We introduce in a self-contained way the notion of Hausdorff measure, which allows us to recover, as special cases, both the Lebesgue measure on an open set of $\mathbf{R}^N$ and surface measures (which play an important role, for example, in the definition of the space trace of Sobolev spaces).

These two basic ingredients, roughly speaking, the generalized differential calculus of distribution theory and the generalized integration theory of Lebesgue, allow us to introduce in chapter 5 the classical Sobolev spaces which provide the right functional setting for the variational approach to the studied problems.

All the ingredients of the variational approach to the model examples are now available: in chapter 6 we describe some of them, including Dirichlet, Neumann, and mixed problems. Let us say that we are in the classical favorable situation: we have to minimize a convex

coercive lower semicontinuous function on a reflexive Banach space and, in such a situation, the direct method of Hilbert and Tonelli does apply.

Chapters 7 and 8 complete this classical portrayal of variational methods by introducing two of the most powerful numerical tools which allow one to compute approximate solutions of variational problems: finite element methods and spectral analysis methods. Each of these two methods corresponds to a very specific type of Galerkin approximation of an infinite dimensional problem by a sequence of finite dimensional ones. Each method has its own advantages; for example, finite element methods allow one to treat engineering problems involving general domains, like the wing of a plane, which explains their great success.

Around 1970, the study of constrained problems and variational inequalities led Stampacchia, Browder, Brezis, Moreau, Rockafellar et al. to develop the elements of a unilateral variational analysis. In particular, convex variational analysis has known considerable success and has familiarized mathematicians with the idea that sets play a decisive role in analysis. The Fenchel duality, the subdifferential calculus of convex functions, and the extension of the Fermat rule are striking examples of this new approach. The role of the epigraph has progressively emerged as essential in the geometrical understanding of these concepts. Chapter 9 provides a thorough exposition of these elements of convex variational analysis in infinite dimensional spaces. We stress the importance of the Fenchel duality, which allows us to associate to each convex variational problem a dual one, whose solutions have in general a deep physical (or numerical or economical) interpretation as multipliers.

Part II: Advanced Variational Analysis. This second part corresponds to chapters 10 through 16 and deals with our second objective, which is to present new trends in variational analysis. Indeed, in recent years, variational methods have proved to be very flexible. They have been developed to study a number of advanced problems of modern technology, like composite materials, image processing, and shape optimization. To grasp these phenomena, the classical framework of variational analysis, which was studied in Part I, must be enlarged. Let us describe some of these extensions:

1. The modelization of a large number of problems in physics, image processing, requires the introduction of new functional spaces permitting *discontinuities* of the solution. In phase transitions, image segmentation, plasticity theory and the study of cracks and fissures, in the study of the wake in fluid dynamics and the shock theory in mechanics, the solution of the problem presents discontinuities along one-codimensional manifolds. Its first distributional derivatives are now measures which may charge zero Lebesgue measure sets, and the solution of these problems cannot be found in classical Sobolev spaces.

   The classical theory of Sobolev spaces, which was developed in chapter 5, is completed in chapter 10 by a self-contained and detailed presentation of these spaces, $BV(\Omega)$, $SBV(\Omega)$, $BD(\Omega)$. The space $BV(\Omega)$, for example, is the space of functions with bounded variations, and a function u belongs to $BV(\Omega)$ iff its first distributional derivatives are bounded measures. The $SBV(\Omega)$ space is the subspace of $BV(\Omega)$ which consists of functions whose first distributional derivatives are bounded measures with no Cantor part.

2. In chapter 12, we introduce the concept of Γ-convergence, which provides a parametrized version of the direct method in variational analysis.

   Following Stampacchia's work, Mosco [187], [188] and Joly [161] introduced the Mosco epiconvergence (1970) of sequences of convex functions to study approximation and perturbation schemes in variational analysis and potential theory.

   The general topological concept, without any convexity assumption, has progressively emerged, and De Giorgi in 1975 introduced the notion of Γ-convergence for sequences of functions. It corresponds to the topological set convergence of the epigraphs, whence the equivalent terminology "epi-convergence." This concept has been successfully applied to a large variety of approximation and perturbation problems in calculus of variations and mechanics: homogenization of composite materials, materials with many small holes and porous media, thin structures and reinforcement problems, and so forth.

   We illustrate the concept by describing some recent applications to thin structures, composite material, phase transitions, and image segmentation.

3. Chapters 11 and 13 deal with the question of lower semicontinuity and relaxation of functionals of calculus of variations. Indeed, as a general rule, when applying the direct method to a functional F which is not lower semicontinuous, one obtains that minimizing sequences converge to solutions of the relaxed problem, which is the minimization of the lower semicontinuous envelope cl F of F.

   In the vectorial case, i.e., when functionals are defined on Sobolev spaces $W^{1,p}(\Omega, \mathbf{R}^m)$, $\Omega \subset \mathbf{R}^N$, relaxation with respect to the weak topology of $W^{1,p}(\Omega, \mathbf{R}^m)$ (or strong topology of $L^p(\Omega, \mathbf{R}^m)$) leads to the important concepts of quasiconvexity (in the sense of Morrey), polyconvexity, and rank-one convexity. We consider as well the case of functionals with linear growth and the corresponding lower semicontinuity and relaxation problems on BV and SBV spaces. All these notions play an important role in the modeling of large deformations in mechanics and plasticity, as described in chapter 14. Following the microstructure school of Ball and James, in the modeling of the solid/solid phase transformations, the density energy possesses a multiwell structure. An alternative and appropriate procedure consists in relaxing the corresponding free energy functional in the space of Young measures generated by gradients.

4. Another important aspect of the direct method concerns the coercivity property. In chapter 15, we examine how the method works when the variational problem has a lack of coercivity. In that case, existence of solutions relies on compatibility conditions, whose general formulation involves recession functions.

5. The last topic considered, in chapter 16, is shape optimization, which is a good illustration of the powerfulness of direct methods in variational analysis and also of their limitations.

Part I

Basic Variational Principles

Chapter 2

Weak solution methods in variational analysis

2.1 The Dirichlet problem: Historical presentation

Throughout this book, we adopt the following notation: Ω is an open subset of $\mathbf{R}^N$ ($N \leq 3$ for applications in classical mechanics), and $x = (x_1, x_2, \ldots, x_N)$ is a generic point in Ω. The topological boundary of Ω is denoted by $\partial\Omega$.

Let $g : \partial\Omega \longrightarrow \mathbf{R}$ be a given function which is defined on the boundary of Ω. The Dirichlet problem consists in finding a function $u : \overline{\Omega} \longrightarrow \mathbf{R}$ which satisfies

$$\Delta u = 0 \quad \text{on} \quad \Omega, \tag{2.1}$$

$$u = g \quad \text{on} \quad \partial\Omega. \tag{2.2}$$

The operator Δ is the Laplacian

$$\Delta u = \sum_{i=1}^{N} \frac{\partial^2 u}{\partial x_i^2} = \frac{\partial^2 u}{\partial x_1^2} + \cdots + \frac{\partial^2 u}{\partial x_N^2};$$

it is equal to the sum of the second partial derivatives of u with respect to each variable $x_1, x_2, \ldots, x_N$. Equation (2.1) is the Laplace equation, and a solution of this equation is said to be harmonic on Ω. Clearly, there are many harmonic functions.

The following examples of harmonic functions illustrate how rich this family is:

- $u(x) = \sum_{i=1}^{N} a_i x_i + b$ (affine function) is harmonic on $\mathbf{R}^N$;
- $u(x) = ax_1^2 + bx_2^2 + cx_3^2$ is harmonic on $\mathbf{R}^3$ iff $a + b + c = 0$;
- $u(x) = e^{x_1} \cos x_2$ and $v(x) = e^{x_1} \sin x_2$ are harmonic on $\mathbf{R}^2$ (note that $u(x_1, x_2) =$ Re $e^{(x_1+ix_2)}$ and $v(x_1, x_2) =$ Im $e^{(x_1+ix_2)}$);
- $u(x) = (x_1^2 + x_2^2 + x_3^2)^{-1/2}$ is harmonic on $\mathbf{R}^3 \backslash \{0\}$ (Newtonian potential).

The study of harmonic functions is a central topic of the so-called potential theory and of harmonic analysis. We will see further, as the above examples suggest, the close connections between this theory, the theory of the complex variable, and the potential theory.

Thus, we can reformulate the Dirichlet problem by saying that we are looking for a harmonic function on Ω which satisfies the boundary data $u = g$ on $\partial\Omega$. It is called a boundary value problem. The condition $u = g$ on $\partial\Omega$, which consists in prescribing the value of the function u on the boundary of Ω, is called the Dirichlet boundary condition and gives rise to the name of the problem.

Let us consider, for illustration, the elementary case $N = 1$. Take $\Omega =]a, b[$ an open bounded interval. Given two real numbers g_1 and g_2, the Dirichlet problem reads as follows: find $u : [a, b] \longrightarrow \mathbf{R}$ such that

$$\begin{cases} u'' = 0 \text{ on }]a, b[, \\ u(a) = g_1, \quad u(b) = g_2. \end{cases}$$

Clearly, this problem has a unique solution, whose graph in $\mathbf{R}^2$ is the line segment joining point (a, g_1) to point (b, g_2).

We will see that for an open bounded subset Ω and under some regularity hypotheses on Ω and g covering most practical situations, one can prove existence and uniqueness of a solution of the Dirichlet problem. Indeed, this is a long story whose important steps are summarized below.

It is in 1782 that the Laplace equation appears for the first time. When studying the orbits of the planets, Laplace discovered that the Newtonian gravitational potential of a distribution of mass of density ρ on a domain $\Omega \subset \mathbf{R}^3$, which is given by the formula

$$u(x) = \frac{1}{4\pi} \int_\Omega \frac{\rho(y)}{|x - y|} dy, \tag{2.3}$$

satisfies the equation

$$\Delta u = 0 \quad \text{on} \quad \mathbf{R}^3 \backslash \bar{\Omega}. \tag{2.4}$$

Indeed, it is a good exercise to establish this formula. One first verifies that the Newtonian potential

$$v(x_1, x_2, x_3) = (x_1^2 + x_2^2 + x_3^2)^{-1/2}$$

satisfies $\Delta v = 0$ on $\mathbf{R}^3 \backslash \{0\}$. Then a direct derivation under the integral sign yields that u is harmonic on $\mathbf{R}^3 \backslash \bar{\Omega}$.

In 1813, Poisson establishes that on Ω the potential u satisfies

$$-\Delta u = \rho \quad \text{on } \Omega, \tag{2.5}$$

which is the so-called Poisson equation.

The central role played by the Laplace and Poisson equations in mathematical physics appeared with more and more evidence, especially because of the work of Gauss. In 1813, Gauss established the following formula (which is often called the Gauss formula or the divergence theorem). Given a vector field $\overrightarrow{V} : \Omega \subset \mathbf{R}^3 \longrightarrow \mathbf{R}^3$,

$$\int\int\int_\Omega \operatorname{div} \overrightarrow{V}(x)\, dx = \int\int_{\partial\Omega} \overrightarrow{V}(x) \cdot \overrightarrow{n}(x)\, ds(x), \tag{2.6}$$

which states that the volume integral of the divergence of the vector field $\overrightarrow{V}$ is equal to the global outward flux of $\overrightarrow{V}$ through the boundary of Ω. In the above formula, if we denote

$$\overrightarrow{V}(x) = (v_1(x), v_2(x), v_3(x)),$$

$$\operatorname{div} \overrightarrow{V}(x) = \sum_i \frac{\partial v_i}{\partial x_i}(x) = \frac{\partial v_1}{\partial x_1}(x) + \frac{\partial v_2}{\partial x_2}(x) + \frac{\partial v_3}{\partial x_3}(x)$$

is the divergence of $\overrightarrow{V}$. The vector $\overrightarrow{n}(x)$ is the unit vector which is orthogonal to $\partial\Omega$ at x and which is oriented toward the outside of Ω. The measure ds is the two-dimensional Hausdorff measure on $\partial\Omega$.

Let us briefly explain how the mathematical formulation of conservation laws in physics leads to the Laplace equation. Suppose that the vector field $\overrightarrow{V}$ derives from a potential u, that is,

$$\overrightarrow{V}(x) = Du(x) = \operatorname{grad} u(x) = \nabla u(x) = \left(\frac{\partial u}{\partial x_i}(x)\right)_{i=1,\dots,N} \tag{2.7}$$

(these are the most commonly used notations). Suppose, moreover, that the vector field $\overrightarrow{V}(x)$ is such that $\int\int_{\partial G} \overrightarrow{V} \cdot \overrightarrow{n}\, ds = 0$ for all closed surfaces $\partial G \subset \Omega$.

By the Gauss theorem, it follows that

$$\int\int\int_G \operatorname{div} \overrightarrow{V}\, dx = 0 \quad \forall\, G \subset \Omega$$

and hence

$$\operatorname{div} \overrightarrow{V} = 0 \quad \text{on } \Omega. \tag{2.8}$$

From (2.7) and (2.8), noticing that

$$\operatorname{div}(\operatorname{grad} u) = \Delta u,$$

we obtain

$$\Delta u = 0,$$

that is, u is harmonic.

The above argument is valid both in the case of the gravitational vector field of Newton and in the case of the electrostatic field of Coulomb in the regions where there is no mass (respectively, charges). With the help of this formula, Gauss was able to prove a number of important properties of harmonic functions, like the mean value property; this was the beginning of the potential theory.

Riemann, who was successively a student of Gauss (1846–1847 in Göttingen) and of Dirichlet (1847–1849 in Berlin), established the foundations of the theory of the complex variable and made the link (when $N = 2$) with the Laplace equation.

Let us recall that for any function $z \in \mathbf{C} \longmapsto f(z)$, which is assumed to be derivable as a function of the complex variable z (f is then said to be holomorphic), its real and imaginary parts P and Q ($f(z) = P(x, y) + iQ(x, y)$, where $z = x + iy$) satisfy the so-called Cauchy–Riemann equations

$$\begin{cases} \dfrac{\partial P}{\partial x} = \dfrac{\partial Q}{\partial y}, \\ \dfrac{\partial P}{\partial y} = -\dfrac{\partial Q}{\partial x}. \end{cases} \tag{2.9}$$

It follows that

$$\Delta P = \frac{\partial^2 Q}{\partial y \partial x} - \frac{\partial^2 Q}{\partial x \partial y} = 0, \qquad \Delta Q = -\frac{\partial^2 P}{\partial x \partial y} - \frac{\partial^2 P}{\partial y \partial x} = 0.$$

Thus, the real part and the imaginary part of a holomorphic function are harmonic. This approach allows, for example, solution in an elegant way of the Dirichlet problem in a disc. Take $\Omega = D(0,1) = \{z \in \mathbf{C} : |z| < 1\}$, the unit disc centered at the origin in $\mathbf{R}^2$. Given $g : \partial D \longrightarrow \mathbf{R}$ a continuous function, we want to solve the Dirichlet problem

$$\begin{cases} \Delta u = 0 \text{ on } D, \\ u = g \text{ on } \partial D. \end{cases} \tag{2.10}$$

Let us start with the Fourier expansion of the 2π-periodic function

$$g(e^{i\theta}) = \sum_{n \in \mathbf{Z}} c_n(g) e^{in\theta}, \tag{2.11}$$

where $c_n(g) = \frac{1}{2\pi} \int_{-\pi}^{+\pi} g(e^{it}) e^{-int} dt$. Note that the above Fourier series converges in $L^2(-\pi, +\pi)$ norm sense (Dirichlet theorem). Indeed, when starting with the Fourier expansion of the boundary data g, one can give an explicit formula for the solution u of the corresponding Dirichlet problem:

$$u(re^{i\theta}) := \sum_{n \in \mathbf{Z}} c_n(g) r^{|n|} e^{in\theta}. \tag{2.12}$$

Clearly, when taking $r = 1$, one obtains $u = g$ on ∂D. Thus, the only point one has to verify is that u is harmonic on D. Take $r < 1$ and replace $c_n(g)$ by its integral expression in (2.12). By a standard argument based on the uniform convergence of the series, one can exchange the symbols $\sum_n$ and $\int_{-\pi}^{+\pi}$ to obtain

$$u(re^{i\theta}) = \frac{1}{2\pi} \int_{-\pi}^{+\pi} g(e^{it}) \sum_{n \in \mathbf{Z}} r^{|n|} e^{in(\theta - t)} dt.$$

Let us introduce

$$P_r(\theta) = \sum_{n \in \mathbf{Z}} r^{|n|} e^{in\theta}, \tag{2.13}$$

the so-called Poisson kernel, and observe that (we denote $z = re^{i\theta}$)

$$P_r(\theta - t) = \operatorname{Re}\left(\frac{e^{it} + z}{e^{it} - z}\right) = \frac{1 - r^2}{1 - 2r\cos(\theta - t) + r^2}. \tag{2.14}$$

Hence,

$$u(re^{i\theta}) = \operatorname{Re} \frac{1}{2\pi} \int_{-\pi}^{+\pi} \frac{e^{it} + z}{e^{it} - z} g(e^{it}) dt, \tag{2.15}$$

and u appears as the real part of a holomorphic function, which says that u is harmonic on D. Using (2.14) one obtains the Poisson formula

$$u(re^{i\theta}) = \frac{1}{2\pi} \int_{-\pi}^{+\pi} \frac{1 - r^2}{1 - 2r\cos(\theta - t) + r^2} g(e^{it})\, dt. \tag{2.16}$$

By similar arguments, one can explicitely solve the Dirichlet problem on a square $]0, a[\times]0, a[$ of $\mathbf{R}^2$. Unfortunately, these methods can be used only in the two-dimensional case.

The modern general treatment of the Dirichlet problem starts with the Dirichlet principle, whose formulation goes back to Gauss (1839), Lord Kelvin, and Dirichlet. It can be formulated as follows.

The solution of the Dirichlet problem is the solution of the following minimization problem:

$$\min\left\{\int_\Omega \sum_{i=1}^N \left(\frac{\partial v}{\partial x_i}\right)^2 dx : v = g \text{ on } \partial\Omega\right\}. \tag{2.17}$$

The functional

$$v \mapsto J(v) := \int_\Omega \sum_{i=1}^N \left(\frac{\partial v}{\partial x_i}\right)^2 dx$$

is called the Dirichlet integral or Dirichlet energy. Thus, the Dirichlet principle states that the solution of the Dirichlet problem minimizes, over all functions v satisfying the boundary data $v = g$ on $\partial\Omega$, the Dirichlet energy. Equivalently,

$$\begin{cases} J(u) \le J(v)\ \forall v, \quad v = g \text{ on } \partial\Omega, \\ u = g \text{ on } \partial\Omega, \end{cases}$$

that is, u is characterized by a minimization of the energy principle.

One can easily verify, at least heuristically, that the solution u of the minimization problem (2.17) is a solution of the Dirichlet problem (2.1), (2.2). Indeed, the Laplace equation (2.1) can be seen as the optimality condition satisfied by the solution of the minimization problem (2.17). The celebrated Fermat's rule which asserts that the derivative of a function is equal to zero at any point u where f achieves a minimum (respectively, maximum) can be developed in our situation by using the notion of directional derivative. When dealing with problems coming from variational analysis, the so-obtained optimality condition is called the Euler equation.

Thus let us take $v : \Omega \longrightarrow \mathbf{R}$, which satisfies $v = 0$ on $\partial\Omega$. Then, for any $t \in \mathbf{R}$, $u + tv$ still satisfies $u + tv = g$ on $\partial\Omega$ and, since u minimizes J on the set $\{w = g \text{ on } \partial\Omega\}$, we have

$$J(u) \le J(u + tv).$$

Let us compute for any $t \in \mathbf{R}$, $t \neq 0$

$$\begin{aligned} \frac{1}{t}\big[J(u + tv) - J(u)\big] &= \frac{1}{t}\int_\Omega |Du + tDv|^2 - |Du|^2 \\ &= 2\int_\Omega Du \cdot Dv\, dx + t\int_\Omega |Dv|^2. \end{aligned}$$

For any $t > 0$, this is a nonnegative quantity, and thus by letting $t \longrightarrow 0^+$

$$\int_\Omega Du \cdot Dv\, dx \ge 0.$$

By taking $t < 0$ and letting $t \longrightarrow 0^-$, we obtain the reverse inequality

$$\int_\Omega Du \cdot Dv \, dx \leq 0.$$

Finally

$$\int_\Omega Du \cdot Dv \, dx = 0 \quad \text{for any } v = 0 \text{ on } \partial\Omega.$$

Taking v regular and after integration by parts, we obtain

$$\int_\Omega (\Delta u) v \, dx = 0 \quad \text{for any } v \text{ regular, } v = 0 \text{ on } \partial\Omega,$$

that is, $\Delta u = 0$ on $\partial\Omega$.

Riemann recognized the importance of this principle but he did not discuss its validity. In 1870, Weierstrass, who was a very systematic and rigorous mathematician, discovered when studying some results of his friend Riemann that the Dirichlet principle raises some difficulties. Indeed, Weierstrass proposed the following example (apparently close to the Dirichlet problem!):

$$\min \left\{ \int_{-1}^{+1} x^2 \left(\frac{dv}{dx} \right)^2 dx : v(-1) = a, v(+1) = b \right\}, \tag{2.18}$$

which fails to have a solution. A minimizing sequence for (2.18) can be obtained by considering the viscosity approximation problem

$$\min \left\{ \int_{-1}^{+1} \left(x^2 + \frac{1}{n^2} \right) \left(\frac{dv}{dx} \right)^2 dx : v(-1) = a, v(+1) = b \right\}, \tag{2.19}$$

which now has a unique solution u_n given by

$$u_n(x) = \frac{a+b}{2} - \frac{a-b}{2} \frac{\arctan nx}{\arctan n}, \quad n = 1, 2, \ldots.$$

One can directly verify that u_n satisfies the boundary data and that

$$\int_{-1}^{+1} x^2 \left(\frac{du_n}{dx} \right)^2 dx \longrightarrow 0 \text{ as } n \longrightarrow +\infty.$$

Thus the value of the infinimum of (2.18) is zero. But there is no regular function (continuous, piecewise $\mathbf{C}^1$) which satisfies the boundary conditions and such that

$$\int_{-1}^{+1} x^2 \left(\frac{du}{dx} \right)^2 dx = 0.$$

Such a function would satisfy $u =$ constant on $]-1, +1[$ which is incompatible with the boundary data when $a \neq b$.

As we will see, the pathology of the Weierstrass example comes from the coefficient x^2 in front of $\left(\frac{dv}{dx}\right)^2$ which vanishes (at zero) on the domain $\Omega = (-1, +1)$. As a result,

there is a lack of uniform ellipticity or coercivity, which is not the case in the Dirichlet problem. Thus, the Weierstrass example is not a counterexample to the Dirichlet principle; its merit is to underline the shortcomings of this principle. Moreover, it raises a decisive question, which is to understand in which class of functions one has to look for the solution of the Dirichlet principle. Until that date, it was commonly admitted that the functions to deal with have to be regular $\mathbf{C}^1$ or $\mathbf{C}^2$ (differentiation being taken in the classical sense), depending on the situation.

It is only in 1900 with a famous conference at the Collège de France in Paris that Hilbert formulates the foundations of the modern variational approach to the Dirichlet principle and hence to the Dirichlet problem. These ideas, which have been worked out in the classical book of mathematical physics of Courant and Hilbert (1937) [115], can be summarized as follows.

The basic idea of Hilbert is to enlarge the class of functions in which one looks for a solution of the Dirichlet principle and simultaneously to generalize the notion of solution. More precisely, Hilbert proposed the following general method for solving the Dirichlet principle:

1. First, construct a minimizing sequence of functions $(u_n)_{n\in\mathbf{N}}$.

2. Then, extract from this minimizing sequence a convergent subsequence, say, $u_{n_k} \longrightarrow \bar{u}$. The so-obtained function $\bar{u}$ is the (generalized) solution of the original problem.

This is what in modern terminology is called a compactness argument. Let us first notice that, even when starting with a sequence $(u_n)_{n\in\mathbf{N}}$ of smooth functions, its limit $\bar{u}$ may be no more differentiable in the classical sense. Such construction of a generalized solution, which corresponds to finding a space obtained by a completion procedure, is very similar to the one which consists of passing from the set of rational numbers to the set of real numbers or from the Riemann integrable functions to Lebesgue integrable functions.

This was a decisive step, and this program was developed throughout the 20th century by a number of mathematicians from different countries.

The functional space in which to find the generalized solution was only in an implicit form in the work of Courant and Hilbert (as a completion of piecewise $\mathbf{C}^1$ functions). The celebrated Sobolev spaces were gradually introduced in the work of Friedrichs (1934) [140] and, for the Soviet mathematical school, Sobolev (1936) [208] and Kondrakov.

The compactness argument, that is, the compact embedding of the Sobolev space $H^1(\Omega)$ into $L^2(\Omega)$ when Ω is bounded, was proved by Rellich (1930).

The modern language of distributions which provides a generalized notion of derivatives for nonsmooth functions (and much more) was systematically developed by L. Schwartz (1950), who was teaching at the Ecole Polytechnique in Paris. This has proved to be a very flexible tool for handling generalized solutions for PDEs.

The ideas of the compactness method introduced by Hilbert to solve the Dirichlet principle were developed in a systematic way by the Italian school. Tonelli (1921) had the intuition to put together the semicontinuity notion of Baire and the Ascoli-Arzela compactness theorem. So doing, he was able to transfer from real functions to functionals of the calculus of variations (like the Dirichlet integral) the classical compactness argument. He developed the so-called *direct methods in the calculus of variations* whose basic topological ingredients are the lower semicontinuity of the functional and the compactness of the lower

level sets of the functional. In the line of the Hilbert approach, he founded the basis of the topological method for minimization problems in infinite dimensional spaces.

Thus, modern tools in variational analysis provide a general and quite simple approach to existence results of generalized solutions for a large number of boundary value problems from mathematical physics. The natural question which then arises is to study the regularity of such solutions and to establish under which conditions on the data and the domain we have a classical solution. A large number of contributions have been devoted to this difficult question. Let us say that in the case of the Dirichlet problem, if the domain Ω and g are sufficiently smooth, then there exists a classical solution $u \in \mathbf{C}^2(\bar{\Omega})$. For a detailed bibliography on the regularity problem, see Brezis [90, Ch. IX].

So far, we have considered the Dirichlet problem as it has been introduced historically. Indeed, one can reformulate it in an equivalent form which is more suitable for a variational treatment.

Let us introduce $\tilde{g} : \Omega \longrightarrow \mathbf{R}$ a function defined on the whole of Ω and whose restriction on $\partial\Omega$ is equal to g:

$$\tilde{g}|_{\partial\Omega} = g. \tag{2.20}$$

One usually prescribes $\tilde{g}$ to preserve the regularity properties on g and $\partial\Omega$ (for example, continuity or Lipschitz continuity) and, in most practical situations, this is quite easy to achieve. Take as a new unknown function

$$v := u - \tilde{g}. \tag{2.21}$$

Clearly it is equivalent to find v or u. The boundary value problem satisfied by v is

$$\begin{cases} -\Delta v = f \ \text{ on } \Omega, \\ \quad v = 0 \ \text{ on } \partial\Omega \end{cases} \tag{2.22}$$

with $f = \Delta\tilde{g}$. The Dirichlet boundary data $v = 0$ on $\partial\Omega$ is then said to be homogeneous, and problem (2.22) is often called the homogeneous Dirichlet problem.

Note that a number of important physical situations lead to (2.22).

For example, when describing the electrostatic potential u in a domain Ω with a density of charge f and whose boundary is connected with the earth, then

$$\begin{cases} -\Delta u = f \ \text{ on } \Omega, \\ \quad u = 0 \ \text{ on } \partial\Omega. \end{cases}$$

Let us consider an elastic membrane in the horizontal plane $x_3 = 0$ occuping a domain Ω in the (x_1, x_2) plane. Suppose that at each point $x \in \Omega$ a vertical force of intensity $f(x)$ is exerted and that the membrane is fixed on its boundary. Let us denote by $u(x)$ the vertical displacement of the point x of the membrane when the equilibrium is attained. Then

$$\begin{cases} -c\,\Delta u = f \ \text{ on } \Omega, \\ \quad u = 0 \ \text{ on } \partial\Omega, \end{cases}$$

where $c > 0$ is the elasticity coefficient of the membrane.

2.2 Test functions and distribution theory

2.2.1 Definition of distributions

The concept of distribution is quite natural if we start from some simple physical observations.

Let us first consider a function $f \in L^1_{loc}(\Omega)$, where Ω is an open subset of $\mathbf{R}^N$. One cannot, for an arbitrary $x \in \Omega$, give a meaning to $f(x)$. But, from a physical point of view, it is meaningful to consider the average of f on a small ball with center x and radius $\varepsilon > 0$ and let ε go to zero. Indeed, it follows from the Lebesgue theory that for almost every $x \in \Omega$,

$$\lim_{\varepsilon \longrightarrow 0} \frac{1}{|B(x,\varepsilon)|} \int_{B(x,\varepsilon)} f(\xi) d\xi$$

exists and the limit is a representative of f. (Such points x are called Lebesgue points of f.) Let us notice that

$$\frac{1}{|B(x,\varepsilon)|} \int_{B(x,\varepsilon)} f(\xi) d\xi = \int_\Omega f(\xi) v_{x,\varepsilon}(\xi) d\xi,$$

where

$$v_{x,\varepsilon}(\xi) = \begin{cases} \frac{1}{|B(x,\varepsilon)|} & \text{if } \xi \in B(x,\varepsilon), \\ 0 & \text{elsewhere.} \end{cases}$$

Thus, it is equivalent to know f as an element of $L^1_{loc}(\Omega)$ or to know the value of the integrals $\int_\Omega f(x)v(x)dx$ for v belonging to a sufficiently large class of functions. This is the starting point of the notion of distribution. Functions v will be called test functions. It is equivalent to know f as a function or as a distribution, the distribution being viewed as the mapping which to a test function v associated $\int_\Omega f v dx$:

$$f : v \mapsto \int_\Omega f v d\xi.$$

If one is concerned only with L^1_{loc} functions, there are many possibilities for the choice of the class of test functions.

Let us go further and suppose we want to model the concept of a Dirac mass, that is, of a unit mass concentrated at a point. This is an important physical notion which can be viewed as the limiting case of the unit mass concentrated in a ball of radius $\varepsilon > 0$ with $\varepsilon \longrightarrow 0$. For example consider the Dirac mass at the origin $0 \in \Omega$ and the functions

$$f_\varepsilon(x) = \begin{cases} \frac{1}{|B(0,\varepsilon)|} & \text{if } x \in B(0,\varepsilon), \\ 0 & \text{elsewhere.} \end{cases}$$

Then, the distribution attached to f_ε is the mapping

$$v \mapsto \int_\Omega f_\varepsilon(\xi) v(\xi) d\xi = \frac{1}{|B(0,\varepsilon)|} \int_{B(0,\varepsilon)} v(\xi) d\xi.$$

When passing to the limit as $\varepsilon \longrightarrow 0$, we need to take test functions at least continuous (at the origin) in order for the above limit to exist. The limiting distribution is the mapping

$$v \in \mathbf{C}_c(\Omega) \mapsto v(0),$$

where $\mathbf{C}_c(\Omega)$ is the set of continuous real-valued functions, with compact support in Ω. This is the modern way to consider a Dirac mass (at the origin) as the linear mapping which to a regular test function v associates its value at the origin. Note that this distribution is no longer attached to a function.

A similar device can be developed to attach a distribution to more general mathematical objects, such as the derivative of a L^1_{loc} function. Take $f \in L^1_{loc}(\Omega)$ and try to define the distribution attached to $\frac{\partial f}{\partial x_i}$. Let us approximate f by a sequence f_n of smooth functions f_n. Then the distribution of $\frac{\partial f_n}{\partial x_i}$ is the mapping

$$v \mapsto \int_\Omega \frac{\partial f_n}{\partial x_i}(x)v(x)dx.$$

But we cannot pass to the limit on this quantity just by taking v continuous, like in the previous step. So let us assume v to be of class $\mathbf{C}^1$ on Ω with a compact support. Then, let us integrate by parts

$$\int_\Omega \frac{\partial f_n}{\partial x_i} v dx = -\int_\Omega f_n \frac{\partial v}{\partial x_i} dx;$$

we can now pass to the limit on this last expression. Finally, the distribution attached to $\frac{\partial f}{\partial x_i}$ is the mapping

$$v \in \mathbf{C}^1_c(\Omega) \mapsto -\int_\Omega f \frac{\partial v}{\partial x_i} dx,$$

where $\mathbf{C}^1_c(\Omega)$ is the set of real-valued functions of class $\mathbf{C}^1$ with compact support in Ω.

We are now ready to define the concept of distribution. We consider the space of test functions $\mathcal{D}(\Omega)$, which is the vector space of real or complex valued functions on Ω which are indefinitely derivable and with compact support in Ω. (This allows us to cover all the previous situations and much more!) For $v \in \mathcal{D}(\Omega)$, we say that the support of v is contained in a compact subset $K \subset \Omega$, and we write $\mathrm{spt} v \subset K$ if $v = 0$ on $\Omega \setminus K$ (equivalently, $\{v \neq 0\} \subset K$).

We use the following notation. An element $p \in \mathbf{N}^N$, $p = (p_1, p_2, \ldots, p_N)$, where N is the dimension of the space ($\Omega \subset \mathbf{R}^N$), is called a multi-index. The integer $|p| = p_1 + p_2 + \cdots + p_N$ is called the lenght of the multi-index p.

For $v \in \mathcal{D}(\Omega)$, we write

$$D^p v := \frac{\partial^{|p|} v}{\partial x_1^{p_1} \partial x_2^{p_2} \ldots \partial x_N^{p_N}}.$$

The operator D^p can be viewed as the composition of elementary partial derivation operators

$$D^p = \left(\frac{\partial}{\partial x_1}\right)^{p_1} \circ \cdots \circ \left(\frac{\partial}{\partial x_N}\right)^{p_N},$$

where $(\frac{\partial}{\partial x_i})^{p_i} = \frac{\partial}{\partial x_i} \circ \frac{\partial}{\partial x_i} \circ \cdots \circ \frac{\partial}{\partial x_i}$, p_i times.

Let us introduce the notion of sequential convergence on $\mathcal{D}(\Omega)$. It is the only topological notion on $\mathcal{D}(\Omega)$ that we use.

Definition 2.2.1. *A sequence $(v_n)_{n\in\mathbf{N}}$ of functions converges in the sense of the space $\mathcal{D}(\Omega)$ to a function $v \in \mathcal{D}(\Omega)$ if the two following conditions are satisfied:*

(i) *There exists a compact subset K in Ω such that* $\operatorname{spt} v_n \subset K$ *for all $n \in \mathbf{N}$ and* $\operatorname{spt} v \subset K$.

(ii) *For all multi-index $p \in \mathbf{N}^N$, $D^p v_n \longrightarrow D^p v$ uniformly on K.*

One can prove the existence of a locally convex topology on the space $\mathcal{D}(\Omega)$ with respect to which a linear functional F is continuous iff it is sequentially continuous, that is, $F(v_n) \longrightarrow F(v)$ whenever $v_n \longrightarrow v$ in the sense of $\mathcal{D}(\Omega)$. But this topology is not easy to handle (it is not metrizable); we don't really need to use it, so we will use only the notion of convergent sequence in $\mathcal{D}(\Omega)$ as defined above.

Definition 2.2.2. *A distribution T on Ω is a continuous linear form on $\mathcal{D}(\Omega)$. Equivalently, a linear form $T : \mathcal{D}(\Omega) \longrightarrow \mathbf{R}$ is a distribution on Ω if for any sequence $(v_n)_{n\in\mathbf{N}}$ in $\mathcal{D}(\Omega)$, the following implication holds:*

$$v_n \longrightarrow 0 \quad \textit{in the sense of} \quad \mathcal{D}(\Omega) \Longrightarrow T(v_n) \longrightarrow 0.$$

The space of distributions on Ω is denoted by $\mathcal{D}'(\Omega)$. It is the topological dual space of $\mathcal{D}(\Omega)$ and we will write $\langle T, v\rangle_{(\mathcal{D}'(\Omega),\mathcal{D}(\Omega))} := T(v)$ the duality pairing between $T \in \mathcal{D}'(\Omega)$ and $v \in \mathcal{D}(\Omega)$.

Let us now give a practical criterion which allows us to verify that a linear form on $\mathcal{D}(\Omega)$ is continuous (and hence is a distribution).

Proposition 2.2.1. *Let T be a linear form on $\mathcal{D}(\Omega)$. Then T is a distribution on Ω iff for all compact K in Ω, there exists $n \in \mathbf{N}$ and $C \geq 0$, possibly depending on K, such that*

$$\forall v \in \mathcal{D}(\Omega) \textit{ with } \operatorname{spt} v \subset K, \quad |T(v)| \leq C \sum_{|p|\leq n} \|D^p v\|_\infty.$$

Proof. Clearly, the above condition implies that T is continuous on $\mathcal{D}(\Omega)$. To prove the converse statement, let us argue by contradiction. Thus, given $T \in \mathcal{D}'(\Omega)$, let us assume that there exists a compact K in Ω and a sequence $(v_n)_{n\in\mathbf{N}}$ in $\mathcal{D}(\Omega)$ such that for each $n \in \mathbf{N}$

$$\operatorname{spt} v_n \subset K \text{ and } |T(v_n)| > n \sum_{|p|\leq n} \|D^p v_n\|_\infty.$$

Let us define

$$w_n := \frac{1}{n \sum_{|p|\leq n} \|D^p v_n\|_\infty} v_n.$$

Then, $w_n \in \mathcal{D}(\Omega)$, spt $w_n \subset K$ and for each $m \in \mathbf{N}$

$$D^m w_n = \frac{1}{n \sum\limits_{|p| \leq n} \|D^p v_n\|_\infty} D^m v_n,$$

so that

$$\forall\, n > m \quad \|D^m w_n\|_\infty \leq \frac{1}{n},$$

and $w_n \longrightarrow 0$ in $\mathcal{D}(\Omega)$. By linearity of T

$$|T(w_n)| > 1,$$

so that $T(w_n)$ does not tend to zero, a contradiction with the fact that $T \in \mathcal{D}'(\Omega)$. □

Proposition 2.2.1 allows us to naturally introduce the notion of distribution with finite order.

Definition 2.2.3. *A distribution $T \in \mathcal{D}'(\Omega)$ has a finite order if there exists an integer $n \in \mathbf{N}$ such that for each compact subset $K \subset \Omega$, there exists a constant $C(K)$ such that*

$$\forall\, v \in \mathcal{D}(\Omega) \quad \textit{with } \operatorname{spt} v \subset K, \ \ |T(v)| \leq C(K) \sup_{|p| \leq n} \|D^p v\|_\infty.$$

If T has a finite order, the order of T is the smallest integer n for which the above inequality holds.

In Proposition 2.2.1, the integer n a priori depends on the compact set K. A distribution has a finite order if n can be taken independent of K.

Let us describe some first examples of distributions.

2.2.2 Locally integrable functions as distributions: Regularization by convolution and mollifiers

Take $f \in L^1_{loc}(\Omega)$, which means that for each compact subset K of Ω, $\int_K |f(x)|\,dx < +\infty$. One can associate to f the linear mapping

$$T_f : v \in \mathcal{D}(\Omega) \longmapsto \int_\Omega f(x)v(x)\,dx.$$

For any compact subset $K \subset \Omega$, for any $v \in \mathcal{D}(\Omega)$ with spt $v \subset K$, the following inequality holds:

$$|T_f(v)| \leq C(K)\|v\|_\infty$$

with $C(K) = \int_K |f(x)|\,dx < +\infty$. By Proposition 2.2.1 T_f is a distribution of order zero.

Indeed, a function $f \in L^1_{loc}(\Omega)$ is uniquely determined by its corresponding distribution T_f, as stated in the following.

Theorem 2.2.1. *Let $f \in L^1_{loc}(\Omega)$, $g \in L^1_{loc}(\Omega)$ be such that*

$$\forall v \in \mathcal{D}(\Omega) \quad \int_\Omega f(x)v(x)\,dx = \int_\Omega g(x)v(x)\,dx.$$

Then $f = g$ almost everywhere (a.e.) on Ω.

The above result allows us to identify $f \in L^1_{loc}$ with the corresponding distribution T_f, which gives the injection $L^1_{loc}(\Omega) \hookrightarrow \mathcal{D}'(\Omega)$.

The proof of Theorem 2.2.1 is a direct consequence of the density of the space of test functions $\mathcal{D}(\Omega)$ in the space $\mathbf{C}_c(\Omega)$.

Proposition 2.2.2. *$\mathcal{D}(\Omega)$ is dense in $\mathbf{C}_c(\Omega)$ for the topology of the uniform convergence. More precisely, for every $v \in \mathbf{C}_c(\Omega)$, there exists a sequence $(v_n)_{n\in\mathbf{N}}$, $v_n \in \mathcal{D}(\Omega)$, and a compact set $K \subset \Omega$ such that $v_n \to v$ uniformly and* spt $v_n \subset K$.

Proof. Take $v \in \mathbf{C}_c(\Omega)$ and extend v by zero outside of Ω. We so obtain an element, which we still denote by v, which is continuous on $\mathbf{R}^N$ and with compact support in Ω. Let us use the *regularization method by convolution* and introduce a smoothing kernel ρ:

$$\begin{cases} \rho \in \mathcal{D}(\mathbf{R}^N), \ \rho \geq 0, \\ \text{spt}\, \rho \subset B(0,1), \\ \displaystyle\int_{\mathbf{R}^N} \rho(x)dx = 1. \end{cases}$$

Take, for example,

$$\rho(x) = \begin{cases} m\, e^{-1/(1-|x|^2)} & \text{if } |x| \leq 1, \\ 0 & \text{elsewhere,} \end{cases}$$

m being choosen in order to have $\int_{\mathbf{R}^N} \rho(x)dx = 1$.

Then let us define for each integer $n = 1, 2, \ldots,$

$$\rho_n(x) := n^N \rho(n\,x),$$

which satisfies

$$\begin{cases} \rho_n \in \mathcal{D}(\mathbf{R}^N), \ \rho_n \geq 0, \\ \text{spt}\, \rho_n \subset B(0,1/n), \\ \displaystyle\int_{\mathbf{R}^N} \rho_n(x)dx = 1. \end{cases}$$

The sequence $(\rho_n)_{n\in\mathbf{N}}$ is said to be a *mollifier*. Given $v \in \mathbf{C}_c(\mathbf{R}^N)$, let us define $v_n = v \star \rho_n$, that is,

$$v_n(x) = \int_{\mathbf{R}^N} v(y)\rho_n(x-y)dy.$$

We have

$$\begin{aligned} \text{spt}\, v_n &\subset \text{spt}\, v + \text{spt}\, \rho_n \\ &\subset \text{spt}\, v + B(0,1/n) \end{aligned}$$

and v_n has a compact support in Ω for n large enough. The classical derivation theorem under the integral sign yields

$$\forall\, \alpha \in \mathbf{N}^N \qquad D^\alpha v_n = v \star D^\alpha \rho_n$$

and v_n belongs to $\mathbf{C}^\infty(\Omega)$. Thus v_n belongs to $\mathcal{D}(\Omega)$.

Let us now prove that v_n converges uniformly to v. To that end, we use the other equivalent formulation of v_n

$$v_n(x) = \int_{\mathbf{R}^N} v(x-y)\rho_n(y)dy$$

and the fact that $\int_{\mathbf{R}^N} \rho_n = 1$ to obtain

$$v_n(x) - v(x) = \int_{\mathbf{R}^N} [v(x-y) - v(x)]\rho_n(y)dy.$$

Using that spt $\rho_n \in B(0,1)$, we have

$$\sup_{x\in\mathbf{R}^N} |v_n(x) - v(x)| \leq \sup_{\substack{x,z\in\mathbf{R}^N \\ \|x-z\|\leq 1/n}} |v(z) - v(x)|.$$

This last quantity goes to zero as $n \to +\infty$; this is a consequence of the uniform continuity of v on $\mathbf{R}^N$ (recall that v is continuous with compact support). □

Let us now complete the proof.

PROOF OF THEOREM 2.2.1. Take $h = f - g$, $h \in L^1_{loc}$ which satisfies

$$\forall v \in \mathcal{D}(\Omega) \quad \int_\Omega h(x)v(x)\,dx = 0.$$

By density of $\mathcal{D}(\Omega)$ in $\mathbf{C}_c(\Omega)$ (see Proposition 2.2.2) for any $v \in \mathbf{C}_c(\Omega)$ there exists a sequence $(v_n)_{n\in\mathbf{N}}$ in $\mathcal{D}(\Omega)$ such that

$$\begin{cases} \text{spt } v_n \subset K & \text{for some fixed compact } K \text{ in } \Omega, \\ v_n \longrightarrow v & \text{uniformly on } K. \end{cases}$$

Since $\int_\Omega h(x)v_n(x)\,dx = \int_K h(x)v_n(x)\,dx = 0$, and $h \in L^1(K)$, by passing to the limit as $n \longrightarrow \infty$, we obtain

$$\forall\, v \in \mathbf{C}_c(\Omega) \quad \int_\Omega h(x)v(x)dx = 0.$$

The conclusion $h = 0$ follows by the Riesz–Alexandrov theorem (see Theorem 2.4.6) and the uniqueness of the representation. Let us give a direct independant proof of the fact that $h = 0$. It will mostly rely on the Tietze–Urysohn separation lemma.

One can first reduce to consider the case $h \in L^1(\Omega)$ with $|\Omega| < +\infty$ (write $\Omega = \bigcup_{n\in\mathbf{N}} \Omega_n$, with Ω_n open, $\overline{\Omega}_n$ compact, and take $h|_{\Omega_n}$). By density of $\mathbf{C}_c(\Omega)$ in $L^1(\Omega)$, for each $\varepsilon > 0$ there exists some $h_\varepsilon \in \mathbf{C}_c(\Omega)$ such that $\|h - h_\varepsilon\|_{L^1(\Omega)} < \varepsilon$. Hence

$$\left| \int_\Omega h_\varepsilon(x)v(x)\,dx \right| \leq \varepsilon \|v\|_{L^\infty(\Omega)} \quad \forall v \in \mathbf{C}_c(\Omega). \tag{2.23}$$

Consider

$$K_1 = \{x \in \Omega : h_\varepsilon(x) \geq \varepsilon\}, \qquad K_2 = \{x \in \Omega : h_\varepsilon(x) \leq -\varepsilon\}.$$

These two sets are disjoint and compact. By the Tietze–Urysohn separation lemma, there exists a function $\varphi \in \mathbf{C}_c(\Omega)$ such that

$$\begin{cases} \varphi(x) = 1 & \text{on } K_1, \\ \varphi(x) = -1 & \text{on } K_2, \\ -1 \le \varphi(x) \le 1 & \forall\, x \in \Omega. \end{cases}$$

Taking $K = K_1 \cup K_2$,

$$\int_\Omega |h_\varepsilon| dx = \int_K |h_\varepsilon| dx + \int_{\Omega\setminus K} |h_\varepsilon|\, dx.$$

On K, we have $|h_\varepsilon| = h_\varepsilon\, \varphi$, so that

$$\int_K |h_\varepsilon|\, dx = \int_K h_\varepsilon \varphi = \int_\Omega h_\varepsilon\, \varphi\, dx - \int_{\Omega\setminus K} h_\varepsilon\, \varphi\, dx.$$

By (2.23), $\int_\Omega h_\varepsilon \varphi\, dx \le \varepsilon \|\varphi\|_{L^\infty} \le \varepsilon$ and, since $|h_\varepsilon| \le \varepsilon$ on $\Omega \setminus K$,

$$\left| \int_{\Omega\setminus K} h_\varepsilon\, \varphi\, dx \right| \le \varepsilon |\Omega|.$$

So,

$$\int_K |h_\varepsilon|\, dx \le \varepsilon(1 + |\Omega|).$$

Finally,

$$\int_\Omega |h_\varepsilon|\, dx \le \varepsilon(1 + |\Omega|) + \varepsilon|\Omega| = \varepsilon + 2\varepsilon|\Omega|$$

and

$$\int_\Omega |h|\, dx \le \int_\Omega |h - h_\varepsilon|\, dx + \int_\Omega |h_\varepsilon|\, dx \le 2\varepsilon(1 + |\Omega|).$$

This being true for any $\varepsilon > 0$, we conclude that $h = 0$. □

Noticing that $L^p(\Omega) \hookrightarrow L^1_{loc}(\Omega)$ for any $1 \le p \le +\infty$, as a direct consequence of Proposition 2.2.2, we obtain the following corollary.

Corollary 2.2.1. *Given $1 \le p \le +\infty$, let us suppose that $f \in L^p(\Omega)$, $g \in L^p(\Omega)$ satisfy*

$$\int_\Omega f(x) v(x)\, dx = \int_\Omega g(x) v(x)\, dx \quad \forall v \in \mathcal{D}(\Omega);$$

then $f = g$ a.e. on Ω.

Let us notice that when taking $1 < p < +\infty$, we can obtain this result in a more direct way, by using the density of $\mathcal{D}(\Omega)$ in $L^q(\Omega)$, $1 \le q < +\infty$. Then take $q = p'$ the Hölder conjugate exponent of p, $\frac{1}{p} + \frac{1}{p'} = 1$. Clearly the density of $\mathcal{D}(\Omega)$ in $L^q(\Omega)$ is a consequence of Proposition 2.2.2. Because of the importance of this result, let us give another proof of it, of independent interest, which relies only on L^p techniques.

Proposition 2.2.3. *Let Ω be an arbitrary open subset of $\mathbf{R}^N$. Then $\mathcal{D}(\Omega)$ is dense in $L^p(\Omega)$ for $1 \le p < +\infty$.*

This will result from the following.

Proposition 2.2.4. *Let $f \in L^p(\mathbf{R}^N)$ with $1 \le p < +\infty$. Then, for any mollifier $(\rho_n)_{n\in\mathbf{N}}$, the following properties hold:*

(i) $f \star \rho_n \in L^p(\mathbf{R}^N)$,

(ii) $\|f \star \rho_n\|_{L^p(\mathbf{R}^N)} \le \|f\|_{L^p(\mathbf{R}^N)}$,

(iii) $f \star \rho_n \longrightarrow f$ in $L^p(\mathbf{R}^N)$ as $n \to +\infty$.

Proof. To prove (i) and (ii) we omit the subscript $n \in \mathbf{N}$. Let us consider the case $1 < p < \infty$ and introduce p' with $\frac{1}{p} + \frac{1}{p'} = 1$.

$$\begin{aligned} |(f \star \rho)(x)| &\le \int_{\mathbf{R}^N} |f(x-y)|\rho(y)dy \\ &\le \int_{\mathbf{R}^N} |f(x-y)|\, \rho(y)^{1/p}\, \rho(y)^{1/p'}\, dy. \end{aligned}$$

Let us apply the Hölder inequality

$$|(f \star \rho)(x)| \le \left(\int_{\mathbf{R}^N} |f(x-y)|^p \rho(y)dy\right)^{1/p} \left(\int_{\mathbf{R}^N} \rho(y)\, dy\right)^{1/p'}.$$

Since $\int_{\mathbf{R}^N} \rho(y)dy = 1$, we obtain

$$|(f \star \rho)(x)|^p \le \int_{\mathbf{R}^N} |f(x-y)|^p \rho(y)dy.$$

Let us integrate with respect to $x \in \mathbf{R}^N$ and apply the Fubini–Tonelli theorem

$$\begin{aligned} \int_{\mathbf{R}^N} |(f \star \rho)(x)|^p dx &\le \int_{\mathbf{R}^N} \left(\int_{\mathbf{R}^N} |f(x-y)|^p \rho(y)\, dy\right) dx \\ &\le \int_{\mathbf{R}^N} \rho(y) \left(\int_{\mathbf{R}^N} |f(x-y)|^p dx\right) dy \\ &\le \int_{\mathbf{R}^N} |f(x)|^p\, dx, \end{aligned}$$

where we have used again that $\int \rho\, dx = 1$ and the fact that the Lebesgue measure on $\mathbf{R}^N$ is invariant by translation. Thus, $f \star \rho \in L^p$ and

$$\|f \star \rho\|_{L^p} \le \|f\|_{L^p}.$$

The convergence of $f \star \rho_n$ to f in $L^p(\mathbf{R}^N)$ relies on a quite similar computation. Using that $\int \rho_n\, dx = 1$, we can write

$$f(x) - (f \star \rho_n)(x) = \int_{\mathbf{R}^N} [f(x) - f(x-y)]\, \rho_n(y)\, dy$$

and

$$|f(x) - (f \star \rho_n)(x)| \leq \int_{\mathbf{R}^N} |f(x) - f(x-y)|\, \rho_n(y)\, dy.$$

Let us rewrite this last inequality as

$$|f(x) - (f \star \rho_n)(x)| \leq \int_{\mathbf{R}^N} |f(x) - f(x-y)|\, \rho_n(y)^{1/p}\, \rho_n(y)^{1/p'}\, dy$$

and apply the Höder inequality to obtain

$$|f(x) - (f \star \rho_n)(x)|^p \leq \int_{\mathbf{R}^N} |f(x) - f(x-y)|^p\, \rho_n(y)\, dy.$$

Integrating with respect to x on $\mathbf{R}^N$ and applying the Fubini–Tonelli theorem, we obtain

$$\|f - (f \star \rho_n)\|_{L^p}^p \leq \int_{\mathbf{R}^N} \rho_n(y)\, \|f - \tau_y f\|_{L^p}^p\, dy.$$

Let us introduce $\varphi(y) := \|f - \tau_y\, f\|_{L^p}^p$. Since $f \in L^p(\mathbf{R}^N)$, φ is a continuous function on $\mathbf{R}^N$ such that $\varphi(0) = 0$. We conclude thanks to the following property:

$$\forall \varphi : \mathbf{R}^N \longrightarrow \mathbf{R}^N \text{ continuous with } \varphi(0) = 0, \text{ we have}$$

$$\lim_{n \to +\infty} \int_{\mathbf{R}^N} \varphi(y)\, \rho_n(y)\, dy = 0.$$

This results from the inequality

$$\begin{aligned} \left| \int_{\mathbf{R}^N} \varphi(y)\, \rho_n(y)\, dy - \varphi(0) \right| &= \left| \int_{\mathbf{R}^N} (\varphi(y) - \varphi(0))\, \rho_n(y)\, dy \right| \\ &\leq \int_{B(0,1/n)} |\varphi(y) - \varphi(0)|\, \rho_n(y)\, dy \\ &\leq \sup_{|y| \leq 1/n} |\varphi(y) - \varphi(0)|, \end{aligned}$$

which tends to zero as $n \to +\infty$. Indeed, we will interpret this last result as a convergence in $\mathcal{D}'(\Omega)$ of the sequence (ρ_n) to δ_0, the Dirac mass at the origin. □

We can now complete the proof.

Proof of proposition 2.2.3. Take $f \in L^p(\Omega)$, $\varepsilon > 0$, and $g \in \mathbf{C}_c(\Omega)$ such that

$$\|f - g\|_{L^p(\Omega)} < \varepsilon.$$

Then let us extend g outside of Ω by zero to obtain a function that we still denote by g which belongs to $\mathbf{C}_c(\mathbf{R}^N)$. Take $f_n = g \star \rho_n$. Then, $f_n \in \mathcal{D}(\mathbf{R}^N)$, and in fact $f_n \in \mathcal{D}(\Omega)$ for n large enough, because spt $f_n \subset$ spt $g + B(0, 1/n)$. Moreover, $f_n \longrightarrow g$ in $L^p(\Omega)$, so that

$$\|f - f_n\|_{L^p} \leq \varepsilon$$

for n large enough. □

2.2.3 Radon measures

Let us recall that a Radon measure μ is a linear form on $\mathbf{C}_c(\Omega)$ such that for each compact $K \subset \Omega$, the restriction of μ to $\mathbf{C}_K(\Omega)$ is continuous, that is, for each $K \subset \Omega$, K compact, there exists some $C(K) \geq 0$ such that

$$\forall\, v \in \mathbf{C}_c(\Omega) \quad \text{with spt}\, v \subset K, \quad |\mu(v)| \leq C(K)\|v\|_\infty.$$

To such a Radon measure, one can associate its restriction to $D(\Omega)$,

$$T_\mu : v \in \mathcal{D}(\Omega) \longrightarrow \int_\Omega v(x)\, d\mu(x),$$

which by the definition of μ is a distribution of order zero.

Conversely, μ is completely determined by the corresponding distribution T_μ. This is a consequence of the density of $\mathcal{D}(\Omega)$ in $\mathbf{C}_c(\Omega)$; see Proposition 2.2.2. As a consequence, we can identify any Radon measure with its corresponding distribution and $\mathcal{M} \hookrightarrow \mathcal{D}'(\Omega)$.

As a typical example of a distribution measure which is not in $L^1_{loc}(\Omega)$, if $0 \in \Omega$, take $\mu = \delta_0$ the Dirac mass at the origin, with

$$\langle \mu, v \rangle_{(\mathcal{D}',\mathcal{D})} := v(0).$$

To describe further examples of great importance in applications, we need to introduce further notions, namely, the derivation of distributions and weak limits of distributions.

2.2.4 Derivation of distributions, introduction to Sobolev spaces

Definition 2.2.4. *Let $T \in \mathcal{D}'(\Omega)$ be a distribution on Ω. Then $\frac{\partial T}{\partial x_i}$ is defined as the linear mapping on $\mathcal{D}(\Omega)$,*

$$\frac{\partial T}{\partial x_i} : v \in \mathcal{D}(\Omega) \longmapsto -\left\langle T, \frac{\partial v}{\partial x_i} \right\rangle_{(\mathcal{D}',\mathcal{D})}.$$

More generally, for any multi-index $p = (p_1, \ldots, p_N)$, we define

$$D^p T : v \in \mathcal{D}(\Omega) \longmapsto (-1)^{|p|} \langle T, D^p v \rangle_{(\mathcal{D}',\mathcal{D})}.$$

Proposition 2.2.5. *For any distribution T on Ω, for any multi-index $p \in \mathbf{N}^N$, we have that $D^p T$ is still a distribution on Ω. Therefore, for any $v \in \mathcal{D}(\Omega)$*

$$\langle D^p T, v \rangle_{(\mathcal{D}'(\Omega),\mathcal{D}(\Omega))} = (-1)^{|p|} \langle T, D^p v \rangle_{(\mathcal{D}'(\Omega),\mathcal{D}(\Omega))}.$$

Proof. One just needs to notice that for $p \in \mathbf{N}^N$ fixed, the mapping $v \longmapsto D^p v$ is continuous from $\mathcal{D}(\Omega)$ into $\mathcal{D}(\Omega)$. This is an immediate consequence of the definition of the sequential convergence in $\mathcal{D}(\Omega)$, which, we recall, involves a compact support condition and the uniform convergence of the derivatives of arbitrary order. These two properties are clearly preserved by the operations D^p. □

Therefore, every distribution in $\mathcal{D}'(\Omega)$ possesses derivatives of arbitrary orders in $\mathcal{D}'(\Omega)$. Indeed, the notion of derivative $D^p T$ of a distribution has been defined so as to

extend the classical notion of derivative for a smooth function. Let us recall the identification we make between $f \in L^1_{loc}$ and the corresponding distribution T_f.

Proposition 2.2.6. *Let f be some function in the set $\mathbf{C}^m(\Omega)$ of real-valued functions of class $\mathbf{C}^m$ in Ω. Then, for any $p \in \mathbf{N}^N$ with $|p| \leq m$, the distribution derivative $D^p f$ coincides with the classical derivative $D^p f$ of functions.*

PROOF. It is a direct consequence of the integration by parts formula. If $f \in \mathbf{C}^1(\Omega)$ and $v \in \mathcal{D}(\Omega)$,

$$\int_\Omega \frac{\partial f}{\partial x_i}(x)\, v(x)\, dx = -\int_\Omega f(x) \frac{\partial v}{\partial x_i}(x)\, dx.$$

Similarly, integration by parts $|p|$ times gives the following formula:

$$\int_\Omega (D^p f)(x)\, v(x)\, dx = (-1)^{|p|} \int_\Omega f(x)\, D^p v(x)\, dx;$$

this formula is valid for $f \in \mathbf{C}^{|p|}(\Omega)$ and $v \in \mathcal{D}(\Omega)$. □

The fact that the test function v has a *compact support* is essential to make, in the integration by parts formula, the integral term on $\partial\Omega$ equal to zero. We stress the fact that the notion of derivative $D^p T$ of a distribution $T \in \mathcal{D}'(\Omega)$ takes as a definition the integration by parts formula, the derivation operation being transferred, by this operation, on the test functions. This can be done at an arbitrary order since the test functions have been taken *indefinitely differentiable*. The two previous remarks justify the *choice of test functions* $v \in \mathcal{D}(\Omega)$.

We can now describe a fundamental example of distribution coming from the theory of Sobolev spaces. This theory, which plays a central role in the variational approach to a large number of boundary value problems (like the Dirichlet problem) will be developed in detail in chapter 5. Here we give some definitions and elementary examples. For any $m \in \mathbf{R}$, $p \in [1, +\infty]$,

$$W^{m,p}(\Omega) = \{f \in L^p(\Omega) : D^j f \in L^p(\Omega) \quad \forall\, j,\ |j| \leq m\}.$$

One of the most important Sobolev spaces is the space

$$W^{1,2}(\Omega) = H^1(\Omega) = \left\{ f \in L^2(\Omega) : \frac{\partial f}{\partial x_i} \in L^2(\Omega),\ i = 1, 2, \ldots, N \right\}.$$

In the above definition, the derivation $\frac{\partial f}{\partial x_i}$ (or more generally $D^j f$) is taken in the distribution sense. We will see that the choice of this notion of derivation is fundamental to obtain the desirable properties for the corresponding spaces.

As an elementary example, let us consider $\Omega = (-1, 1)$ and $f(x) = |x|$. Clearly, f is not differentiable in the classical sense at the origin. The function f is continuous on Ω, it belongs to any $L^p(\Omega)$, $1 \leq p \leq +\infty$, and thus it defines a distribution and we can compute its first distribution derivative Df

$$\langle Df, v\rangle_{(\mathcal{D}'(\Omega), \mathcal{D}(\Omega))} := -\int_{-1}^{1} f(x) v'(x) dx.$$

Let us write

$$\int_{-1}^{1} f(x)v'(x)dx = \int_{-1}^{0} f(x)v'(x)dx + \int_{0}^{1} f(x)v'(x)dx$$

and let us integrate by parts on each interval $(-1, 0)$ and $(0, 1)$. This is possible since now $f \in \mathbf{C}^1([-1, 0])$ and $f \in \mathbf{C}^1([0, 1])$. We have

$$\int_{-1}^{0} f(x)v'(x)dx = f(0)v(0) - f(-1)v(-1) - \int_{-1}^{0} f'(x)v(x)dx,$$

$$\int_{0}^{1} f(x)v'(x)dx = f(1)v(1) - f(0)v(0) - \int_{0}^{1} f'(x)v(x)dx.$$

Note that since $v \in \mathcal{D}(-1, 1)$, we have $v(-1) = v(1) = 0$, but for a general $v \in \mathcal{D}(-1, 1)$, $v(0) \neq 0$. By adding the two above equalities, the terms containing $v(0)$ cancel and we obtain

$$\forall v \in \mathcal{D}(-1, 1) \quad \int_{-1}^{1} f(x)v'(x)\,dx = -\left[\int_{-1}^{0} -v(x)\,dx + \int_{0}^{1} v(x)\,dx\right],$$

that is,

$$\int_{-1}^{1} f(x)v'(x)dx = -\int_{-1}^{1} g(x)v(x)dx,$$

where

$$g(x) = \begin{cases} -1 & \text{if} \quad -1 < x < 0, \\ 1 & \text{if} \quad 0 < x < 1. \end{cases}$$

The above function g is then the distributional derivative of $f(x) = |x|$ on $\Omega = (-1, 1)$. It belongs to $L^p(\Omega)$ for any $1 \leq p \leq +\infty$, so that $f \in W^{1,p}(\Omega)$ for any $1 \leq p \leq +\infty$.

The parameters $m \in \mathbf{N}$ and $p \in [1, +\infty]$ yield a scale of spaces which allow us to distinguish, for example, in our situation the different behavior of the functions $f_\alpha(x) = |x|^\alpha$ and of their derivatives at zero. A similar computation as above yields

$$Df_\alpha = \alpha|x|^{\alpha-2}x \quad \text{in } \mathcal{D}'(-1, 1).$$

Hence, $f_\alpha \in W^{1,p}(-1, 1)$ iff $\int_{-1}^{1} |x|^{p(\alpha-1)}dx < +\infty$, that is, $p < \frac{1}{1-\alpha}$. When $p = 2$, we have that $f_\alpha(x) = |x|^\alpha$ belongs to $H^1(-1, 1)$ iff $\alpha > \frac{1}{2}$: this expresses that in some sense the derivative of f_α at x does not tend to $+\infty$ too rapidly when x goes to zero.

Let us now examine the other parameter m, which is relative to the order of derivation. Take again $f(x) = |x|$ and compute the second order derivative of f on $(-1, 1)$. This amounts to computing the first order derivative of $g(x) = \operatorname{sign} x$. Thus

$$\langle D^2 f, v\rangle_{(\mathcal{D}',\mathcal{D})} = \langle Dg, v\rangle_{(\mathcal{D}',\mathcal{D})} = -\int_{-1}^{1} g(x)v'(x)dx.$$

As before, let us split the integral over $(-1, 1)$ into two parts,

$$\begin{aligned} \int_{-1}^{1} g(x)v'(x)\,dx &= -\int_{-1}^{0} v'(x)\,dx + \int_{0}^{1} v'(x)\,dx \\ &= -[v(0) - v(-1)] + [v(1) - v(0)] \\ &= -2v(0), \end{aligned}$$

since $v \in \mathcal{D}(-1,1)$ and $v(-1) = v(1) = 0$. Thus

$$\langle Dg, v\rangle_{(\mathcal{D}'(-1,1),\mathcal{D}(-1,1))} = 2v(0)$$

and $Dg = 2\delta_0$, where δ_0 is the Dirac mass at the origin.

But the distribution δ_0 is a measure which is not representable by a function: suppose that there exists a function $h \in L^1_{loc}$ such that

$$\forall v \in \mathcal{D}(-1,1) \quad v(0) = \int_{-1}^{1} h(x)v(x)\,dx.$$

Then, taking successively $v \in \mathcal{D}(-1,0)$ and $v \in \mathcal{D}(0,1)$, we conclude by Theorem 2.2.1 that $h = 0$ a.e. on $(-1,0)$ and on $(0,1)$. Hence $h = 0$ a.e. on $(-1,1)$, which would imply $v(0) = 0$ for every $v \in \mathcal{D}(-1,1)$, a clear contradiction. Therefore $f \in W^{1,2}(-1,1)$ but $f \notin W^{2,1}(-1,1)$.

The above computation of the distributional derivative of a function g which has a discontinuity is very important in a number of applications (phase transitions, plasticity, image segmentation, etc.). This situation will be considered in detail in chapter 10 and will lead us to the introduction of the functional space $BV(\Omega)$, the space of functions with bounded variation, which can be characterized as the space of integrable functions whose first distributional derivatives are bounded measures.

The next operation on distributions which is very useful for applications is the notion of limit of a sequence of distributions.

2.2.5 Convergence of sequences of distributions

Definition 2.2.5. *Let $T_n \in \mathcal{D}'(\Omega)$ for all $n \in \mathbf{N}$ and $T \in \mathcal{D}'(\Omega)$. The sequence $(T_n)_{n\in\mathbf{N}}$ is said to converge to T in $\mathcal{D}'(\Omega)$ if*

$$\forall v \in \mathcal{D}(\Omega) \quad \lim_{n\to+\infty} T_n(v) = T(v).$$

We will write $T_n \longrightarrow T$ in $\mathcal{D}'(\Omega)$ or $\lim_{n\to+\infty} T_n = T$ in $\mathcal{D}'(\Omega)$.

In Section 2.4, we will interpret this convergence as a weak* convergence in the dual space $\mathcal{D}'(\Omega)$ of $\mathcal{D}(\Omega)$.

Let us recall that any distribution $T \in \mathcal{D}'(\Omega)$ possesses derivatives or arbitrary orders. The following result expresses that for any multi-index $p \in \mathbf{N}^N$ the mapping $T \longrightarrow D^pT$ is continuous.

Proposition 2.2.7. *Let $p \in \mathbf{N}^N$. The mapping*

$$T \in \mathcal{D}'(\Omega) \longmapsto D^pT \in \mathcal{D}'(\Omega)$$

is continuous, which means that for any sequence $(T_n)_{n\in\mathbf{N}}$, T in $\mathcal{D}'(\Omega)$, the following implication holds:

$$T_n \longrightarrow T \text{ in } \mathcal{D}'(\Omega) \implies D^pT_n \longrightarrow D^pT \text{ in } \mathcal{D}'(\Omega).$$

PROOF. The proof is a direct consequence of the definitions. Let us assume that $T_n \longrightarrow T$ in $\mathcal{D}'(\Omega)$ and take $v \in \mathcal{D}(\Omega)$.

By definition of D^p

$$\langle D^p T_n, v\rangle = (-1)^{|p|}\langle T_n, D^p v\rangle \quad \forall\, v \in \mathcal{D}(\Omega).$$

Since $v \in \mathcal{D}(\Omega)$, $D^p v$ still belongs to $\mathcal{D}(\Omega)$, and the convergence of T_n to T in $\mathcal{D}'(\Omega)$ implies

$$\lim_{n\to+\infty}\langle T_n, D^p v\rangle = \langle T, D^p v\rangle.$$

Thus, again by the definition of D^p,

$$\lim_{n\to+\infty}\langle D^p T_n, v\rangle = (-1)^{|p|}\langle T, D^p v\rangle = \langle D^p T, v\rangle,$$

which expresses that $D^p T_n \longrightarrow D^p T$ in $\mathcal{D}'(\Omega)$. □

The above proposition is one of the reasons that explains the success of the theory of distributions. It makes this theory a very flexible tool for the study of PDEs. We will often use this type of argument—for example, in the chapter on Sobolev spaces. Suppose $(v_n)_{n\in\mathbf{N}}$ is a sequence in $H^1(\Omega)$ such that

$$\begin{aligned} v_n &\longrightarrow v \text{ in } L^2(\Omega),\\ \frac{\partial v_n}{\partial x_i} &\longrightarrow g_i \text{ in } L^2(\Omega), \qquad i = 1, 2, \ldots, N. \end{aligned}$$

Then, $v \in H^1(\Omega)$ and $g_i = \frac{\partial v}{\partial x_i}$, $i = 1, 2, \ldots, N$. This can be justified with the language of distribution as follows. Since $v_n \longrightarrow v$ in $L^2(\Omega)$, $v_n \longrightarrow v$ in $\mathcal{D}'(\Omega)$ and hence $\frac{\partial v_n}{\partial x_i} \longrightarrow \frac{\partial v}{\partial x_i}$ in $\mathcal{D}'(\Omega)$. On the other hand, $\frac{\partial v_n}{\partial x_i} \longrightarrow g_i$ in $L^2(\Omega)$ and hence in $\mathcal{D}'(\Omega)$. The uniqueness of the limit in $\mathcal{D}'(\Omega)$ implies that $\frac{\partial v}{\partial x_i} = g_i$ for all $i = 1, 2, \ldots, N$ and $v \in H^1(\Omega)$.

Let us give another illustration of the above tools and compute the fundamental solution of the Laplacian in $\mathbf{R}^3$.

Proposition 2.2.8. *Take $N = 3$ and consider the Newtonian potential*

$$f(x) = \frac{1}{\sqrt{x_1^2 + x_2^2 + x_3^2}}.$$

Then, $-\Delta\left(\frac{1}{4\pi} f\right) = \delta$.

PROOF. Let us denote $r(x) = \sqrt{x_1^2 + x_2^2 + x_3^2}$ the Euclidean distance of $x = (x_1, x_2, x_3)$ from the origin, and notice that $f(x) = \frac{1}{r(x)}$ belongs to $L^1_{loc}(\mathbf{R}^3)$ and thus defines a distribution on $\mathbf{R}^3$. Let us compute Δf in $\mathcal{D}'(\Omega)$. A standard approach consists in approximating f by a sequence f_ε of smooth functions, then computing Δf_ε in a classical sense by Proposition 2.2.6 and passing to the limit in $\mathcal{D}'(\Omega)$ as $\varepsilon \to 0$. Clearly, the difficulty is at the origin where f has a singularity. Thus, the parameter ε is intended to isolate the origin and

regularize f at the origin. At this point, there are two possibilities which lead to different computations and that we examine now.

First, take for $\varepsilon > 0$

$$f_\varepsilon(x) = \begin{cases} 1/\varepsilon & \text{if } r(x) \le \varepsilon, \\ 1/r(x) & \text{if } r(x) \ge \varepsilon. \end{cases}$$

Clearly, f_ε is now a continuous, piecewise smooth function (it is not $\mathbf{C}^1$) on $\mathbf{R}^3$ and a standard computation yields that $\frac{\partial f_\varepsilon}{\partial x_i}$ belongs to $L^2(\mathbf{R}^3)$ with

$$\frac{\partial f_\varepsilon}{\partial x_i} = \begin{cases} 0 & \text{if } r < \varepsilon, \\ -x_i/r^3 & \text{if } r > \varepsilon. \end{cases}$$

Note that $\frac{\partial r}{\partial x_i} = \frac{x_i}{r}$ for $i = 1, 2, 3$.

Let us now compute $-\Delta f_\varepsilon$. By definition,

$$\begin{aligned} \langle -\Delta f_\varepsilon, v\rangle_{(\mathcal{D}'(\mathbf{R}^3),\mathcal{D}(\mathbf{R}^3))} &= \langle f_\varepsilon, -\Delta v\rangle_{(\mathcal{D}',\mathcal{D})} \\ &= \sum_{i=1}^{3} \left\langle \frac{\partial f_\varepsilon}{\partial x_i}, \frac{\partial v}{\partial x_i} \right\rangle_{(\mathcal{D}',\mathcal{D})}. \end{aligned}$$

Since $\frac{\partial f_\varepsilon}{\partial x_i}$ belongs to L^1_{loc},

$$\begin{aligned} \langle -\Delta f_\varepsilon, v\rangle &= \sum_{i=1}^{3} \int_{\mathbf{R}^3} \frac{\partial f_\varepsilon}{\partial x_i} \frac{\partial v}{\partial x_i} dx \\ &= -\sum_{i=1}^{3} \int_{r\ge\varepsilon} \frac{x_i}{r^3} \cdot \frac{\partial v}{\partial x_i}\, dx. \end{aligned}$$

Let us now integrate by parts this last expression. Noticing that on $\mathbf{R}^3 \setminus \{0\}$,

$$\begin{aligned} \sum_i \frac{\partial}{\partial x_i}\left(\frac{x_i}{r^3}\right) &= \frac{3}{r^3} + \sum x_i \cdot \frac{-3}{r^4} \cdot \frac{x_i}{r} \\ &= \frac{3}{r^3} - 3\sum \frac{x_i^2}{r^5} = \frac{3}{r^3} - \frac{3}{r^3} = 0 \end{aligned}$$

(which means that $\Delta f = 0$ on $\mathbf{R}^3 \setminus \{0\}$), we obtain

$$\langle -\Delta f_\varepsilon, v\rangle = -\int_{S_\varepsilon} \sum_{i=1}^{3} \frac{x_i}{r^3}\left(-\frac{x_i}{r}\right) v\, dx,$$

where $S_\varepsilon = \{x \in \mathbf{R}^3 : r(x) = \varepsilon\}$ is the sphere of radius ε centered at the origin. Note that the unit normal to S_ε at x which is oriented toward the outside of $\{r \ge \varepsilon\}$ is equal to $-\frac{x}{r}$. Hence

$$\begin{aligned} \langle -\Delta f_\varepsilon, v\rangle_{(\mathcal{D}',\mathcal{D})} &= \int_{S_\varepsilon} \frac{1}{r^2}(x) v(x)\, dx \\ &= \langle \mu_\varepsilon, v\rangle_{(\mathcal{D}',\mathcal{D})}, \end{aligned}$$

where $\mu_\varepsilon = \varepsilon^{-2}\mathcal{H}^2 \lfloor S_\varepsilon$ and $\mathcal{H}^2 \lfloor S_\varepsilon$ is the two-dimensional Hausdorff measure supported by S_ε. An elementary calculus yields that

$$\lim_{\varepsilon \to 0} \frac{1}{4\pi} \mu_\varepsilon = \delta \text{ in } \mathcal{D}'(\mathbf{R}^3).$$

By definition of the convergence in $\mathcal{D}'(\mathbf{R}^3)$ we have

$$-\Delta \left(\frac{1}{4\pi} f_\varepsilon \right) \longrightarrow \delta \text{ in } \mathcal{D}'(\mathbf{R}^3).$$

On the other hand, f_ε converges to f in $L^1(\mathbf{R}^3)$ (for example, by the dominated convergence theorem) and hence in $\mathcal{D}'(\mathbf{R}^3)$. By the continuity of the differential operator Δ in $\mathcal{D}'(\mathbf{R}^3)$ (cf. Proposition 2.2.7) we finally obtain that $-\Delta \left(\frac{1}{4\pi} f \right) = \delta$.

Another regularization consists of building f_ε a $\mathbf{C}^1(\mathbf{R}^3)$ function which approximates f. Take, for example,

$$f_\varepsilon(x) = \begin{cases} a_\varepsilon r^2(x) + b_\varepsilon & \text{if } r(x) \leq \varepsilon, \\ 1/r(x) & \text{if } r(x) \geq \varepsilon, \end{cases}$$

a_ε and b_ε being choosen in order to have $f_\varepsilon \in \mathbf{C}^1(\mathbf{R}^3)$. This is equivalent to the system

$$\begin{cases} a_\varepsilon \varepsilon^2 + b_\varepsilon = 1/\varepsilon, \\ a_\varepsilon = -1/(2\varepsilon^3), \end{cases}$$

which gives $b_\varepsilon = 3/(2\varepsilon)$. Noticing that

$$\begin{aligned} a_\varepsilon r^2 + b_\varepsilon &= -\frac{r^2}{2\varepsilon^3} + \frac{3}{2\varepsilon} \\ &= \frac{1}{2\varepsilon}\Big[3 - \frac{r^2}{\varepsilon^2}\Big] \\ &\leq \frac{3}{2\varepsilon} \quad \text{for } r(x) \leq \varepsilon, \end{aligned}$$

we have that

$$0 \leq f_\varepsilon(x) \leq \frac{3}{2r(x)} = \frac{3}{2} f(x) \text{ on } \mathbf{R}^3.$$

Hence, by the Lebesgue dominated convergence, $f_\varepsilon \longrightarrow f$ in $L^1(\mathbf{R}^3)$ as $\varepsilon \longrightarrow 0$. So $f_\varepsilon \longrightarrow f$ in $\mathcal{D}'(\mathbf{R}^3)$ and

$$\Delta f_\varepsilon \longrightarrow \Delta f \text{ in } \mathcal{D}'(\mathbf{R}^3).$$

An elementary computation yields

$$\begin{aligned} -\Delta f_\varepsilon &= -6a_\varepsilon 1_{B(0,\varepsilon)} \\ &= +\frac{3}{\varepsilon^3} 1_{B(0,\varepsilon)}. \end{aligned}$$

Hence

$$-\Delta\left(\frac{1}{4\pi}f_\varepsilon\right) = \frac{1}{\frac{4}{3}\pi\varepsilon^3}1_{B(0,\varepsilon)}.$$

Noticing that $\frac{4}{3}\pi\varepsilon^3$ is precisely the volume of the ball $B(0,\varepsilon)$, we have

$$\lim_{\varepsilon\to 0}\frac{1}{\frac{4}{3}\pi\varepsilon^3}1_{B(0,\varepsilon)} = \delta,$$

which finally implies

$$-\Delta\left(\frac{1}{4\pi}f\right) = \delta$$

and completes the proof. □

2.3 Weak solutions

2.3.1 Weak formulation of the model examples

The Dirichlet problem. Let Ω be an open subset of $\mathbf{R}^N$ and $f : \Omega \longrightarrow \mathbf{R}$ a given function; take $f \in L^2(\Omega)$, for example. We recall that the Dirichlet problem is to find a function $u : \overline{\Omega} \longrightarrow \mathbf{R}$ which solves

$$-\Delta u = f \quad \text{on } \Omega, \qquad u = 0 \quad \text{on } \partial\Omega. \tag{2.24}$$

In Section 2.1, we explained that it is difficult to prove directly the existence of a classical solution to this problem. By classical solution, we mean a function u which is continuous on $\overline{\Omega}$ and of class $\mathbf{C}^2$ on Ω. So, the idea is to allow u to be less regular (at least in a first stage) and to interpret Δu in a weak sense, namely, in a distribution sense.

Taking test functions $v \in \mathcal{D}(\Omega)$, (2.24) is equivalent to

$$\langle -\Delta u, v\rangle_{(\mathcal{D}'(\Omega),\mathcal{D}(\Omega))} = \int_\Omega f v dx \quad \forall\, v \in \mathcal{D}(\Omega).$$

The definition of the derivation of distributions (see Definition 2.2.4) is precisely based on the integration by parts formula and allows us to transfer the derivation operation from u onto the test functions $v \in \mathcal{D}(\Omega)$. At this point, we have two possibilities. The two equalities

$$\langle -\Delta u, v\rangle = \sum_{i=1}^N \left\langle \frac{\partial u}{\partial x_i}, \frac{\partial v}{\partial x_i}\right\rangle_{(\mathcal{D}'(\Omega),\mathcal{D}(\Omega))}, \tag{2.25}$$

$$\langle -\Delta u, v\rangle = \langle u, -\Delta v\rangle_{(\mathcal{D}'(\Omega),\mathcal{D}(\Omega))} \tag{2.26}$$

correspond, respectively, to a partial transfer and a global transfer of the derivatives on the test functions. They give rise to two distinct weak formulations of the initial problem, which indeed depend on the regularity properties which we expect the solution u to satisfy.

If we expect the solution u to have first distribution derivatives which are integrable, then (2.25) gives rise to

$$\begin{cases} \displaystyle\int_\Omega \sum_{i=1}^N \frac{\partial u}{\partial x_i}\frac{\partial v}{\partial x_i} dx = \int_\Omega f v\, dx \quad \forall\, v \in \mathcal{D}(\Omega), \\ u = 0 \text{ on } \partial\,\Omega. \end{cases} \tag{2.27}$$

If we don't expect u to have first derivatives which are integrable and just expect u to be $L^1(\Omega)$ or in $\mathbf{C}(\overline{\Omega})$, then by using (2.26), we obtain the (very) weak formulation

$$\begin{cases} -\displaystyle\int_\Omega u\Delta v\, dx = \int_\Omega f v\, dx \quad \forall\, v \in \mathcal{D}(\Omega), \\ u = 0 \text{ on } \partial\Omega. \end{cases} \tag{2.28}$$

For many reasons, the weak formulation (2.27) is the one which is well adapted to our situation:

(a) As a general rule, we will see that it is preferable to perform the integrations by parts which are necessary and no more. Otherwise, the solution which is obtained satisfies the equation in a very weak sense, we have only poor informations on this solution, and the study of its uniqueness and regularity becomes quite involved.

(b) When trying to give a sense to the boundary condition $u = 0$ on $\partial\Omega$, it is useful to have some information on the derivatives of u on Ω. To find a weak solution u for which we know just that u belongs to some $L^p(\Omega)$ space is not sufficient to give meaning to the trace of u on $\partial\Omega$ (recall that $\partial\Omega$ has a zero Lebesgue measure).

Since we will be able to find a weak solution of the Dirichlet problem in the space $H^1(\Omega)$ (that is, with first order distribution derivatives in $L^2(\Omega)$), we will use (2.27) as a variational formulation of the Dirichlet problem. Note, too, that (2.27) has another advantage over (2.28): the left member of the equation, which is the important part and which involves the partial differential operator governing the equation, is symmetric with respect to u and v in (2.27), while it is not in (2.28). This has important consequences on the variational formulation of the problem; see Section 2.3.2 and chapter 3.

Let us summarize the above comments and give a first definition of the notion of a weak solution for the Dirichlet problem. It will be made precise later and solved in chapter 6.

Given $f \in L^2(\Omega)$, a weak solution u of the Dirichlet problem (2.24) is a function $u \in H^1(\Omega)$ which satisfies

$$\begin{cases} \displaystyle\int_\Omega \sum_{i=1}^N \frac{\partial u}{\partial x_i}\frac{\partial v}{\partial x_i} dx = \int_\Omega f v\, dx \quad \forall\, v \in \mathcal{D}(\Omega), \\ u = 0 \text{ on } \partial\Omega. \end{cases} \tag{2.29}$$

Indeed, we will justify the choice of the functional space $H^1(\Omega)$ and explain how to interpret the trace of such functions on $\partial\Omega$. We will reformulate (2.29) by introducting the subspace $H_0^1(\Omega)$ of $H^1(\Omega)$:

$$H_0^1(\Omega) = \{v \in H^1(\Omega) : v = 0 \text{ on } \partial\Omega\}.$$

Indeed, $H_0^1(\Omega)$ is equal to the closure of $\mathcal{D}(\Omega)$ in $H^1(\Omega)$. As a consequence, the equality (2.29) can be extended by a density and continuity argument to $H_0^1(\Omega)$. In this way we obtain the classical weak formulation of the Dirichlet problem.

Definition 2.3.1. *A weak solution of the Dirichlet problem is a solution of the following system:*

$$\begin{cases} \displaystyle\int_\Omega \sum_{i=1}^{N} \frac{\partial u}{\partial x_i}\frac{\partial v}{\partial x_i} dx = \int_\Omega f v\, dx \quad \forall\, v \in H_0^1(\Omega), \\ u \in H_0^1(\Omega). \end{cases} \tag{2.30}$$

Note that (2.30) can be written in the following abstract form: find $u \in V$ such that $a(u, v) = L(v)$ for all $v \in V$, where $a : V \times V \longrightarrow \mathbf{R}$ is a bilinear form which is symmetric and positive ($a(v, v) \geq 0$ for every $v \in V$) and L is a linear form on V.

In chapter 3, an existence result for such an abstract problem will be proved (Lax–Milgram theorem); in chapter 5 the basic ingredients of the theory of Sobolev spaces will be developed, for example, to treat the case $V = H_0^1(\Omega)$. Thus, in chapter 6 we will be able to prove the existence of a weak solution to the Dirichlet problem.

The Neumann problem. We recall that the Neumann problem consists of finding a solution u to the boundary problem

$$u - \Delta u = f \text{ on } \Omega, \qquad \frac{\partial u}{\partial n} = 0 \text{ on } \partial\Omega, \tag{2.31}$$

where $\frac{\partial u}{\partial n} = Du \cdot n$ is the outward normal derivative of u on $\partial\Omega$. A major difference between the Dirichlet and the Neumann problem is that in the Neumann problem, the value of u on the boundary is not prescribed (it is $\frac{\partial u}{\partial n}$ which is prescribed). As a consequence, we have to test u on Ω and on $\partial\Omega$; it is not sufficient to take test functions $v \in \mathcal{D}(\Omega)$.

We will take test functions $v \in \mathbf{C}^1(\overline{\Omega})$. We are no longer in the setting of the distribution theory, but we can follow the lines of this theory. Let us first assume that u is regular and let us multiply (2.31) by $v \in \mathbf{C}^1(\overline{\Omega})$ and integrate by parts. Recall that from the divergence theorem,

$$\int_\Omega \operatorname{div}(v\, Du)\, dx = \int_{\partial\Omega} v\, Du \cdot n\, d\sigma.$$

Thus

$$\int_\Omega (v\, \Delta u + Du \cdot Dv)\, dx = \int_{\partial\Omega} v \frac{\partial u}{\partial n}\, d\sigma. \tag{2.32}$$

By using (2.31) and (2.32) we obtain

$$\int_\Omega (uv + Du \cdot Dv)\, dx = \int_\Omega f v\, dx \quad \forall v \in \mathbf{C}^1(\overline{\Omega}). \tag{2.33}$$

Now (2.33) makes sense even for a function u for which we are only able to define first generalized derivatives as functions. So, we will take (2.33) in a first step as a notion of weak solution. Precisely, given $f \in L^2(\Omega)$, a weak solution of the Neumann problem is a function $u \in H^1(\Omega)$ such that

$$\int_\Omega (uv + Du \cdot Dv)\, dx = \int_\Omega f v\, dx \quad \forall v \in \mathbf{C}^1(\overline{\Omega}). \tag{2.34}$$

The striking feature is that in this weak formulation (2.34), the Neumann boundary condition has disappeared! It is important to verify that, so doing, we have not lost any information. In other words, we need to show that, conversely, if $u \in H^1(\Omega)$ verifies (2.34), then u satisfies (2.31). The second condition in (2.31) is the so-called Neumann boundary condition.

First take $v \in \mathcal{D}(\Omega)$. Clearly $\mathcal{D}(\Omega)$ is a subspace of $\mathbf{C}^1(\overline{\Omega})$ and so we obtain

$$u - \Delta u = f \text{ in } \mathcal{D}'(\Omega). \tag{2.35}$$

To recover the Neumann boundary condition, we have to perform the integration by parts in a reverse way. To do so, we assume that we have been able to prove that the weak solution u of (2.34) is in fact a regular function. So, by using (2.32) and (2.34),

$$\int_\Omega (u - \Delta u) v dx + \int_{\partial\Omega} v\, \frac{\partial u}{\partial n} d\sigma = \int_\Omega f v\, dx \quad \forall v \in \mathbf{C}^1(\overline{\Omega}). \tag{2.36}$$

By (2.35), $u - \Delta u = f$, so we can simplify (2.36) to obtain

$$\int_{\partial\Omega} v\, \frac{\partial u}{\partial n}\, d\sigma = 0 \quad \forall v \in \mathbf{C}^1(\overline{\Omega}),$$

which implies $\frac{\partial u}{\partial n} = 0$.

Thus, the Neumann boundary condition is implicitely contained in the weak variational formulation (2.34). Indeed, just like for the Dirichlet problem, we will prove a density result, namely, "$\mathbf{C}^1(\overline{\Omega})$ is dense in $H^1(\Omega)$." As a consequence, the equality (2.34) can be extended to all $v \in H^1(\Omega)$ and the final variational formulation of the Neumann problem will be the following.

Definition 2.3.2. *A weak solution of the Neumann problem is a solution u of the following system:*

$$\begin{cases} \displaystyle\int_\Omega (uv + Du \cdot Dv)\, dx = \int_\Omega f v\, dx \quad \forall\, v \in H^1(\Omega), \\ u \in H^1(\Omega). \end{cases} \tag{2.37}$$

Note again that the above problem can be written as

$$\begin{cases} \text{find } u \in V = H^1(\Omega) \text{ such that} \\ a(u, v) = L(v)\ \forall\, v \in V, \end{cases}$$

where $a(u, v) = \int_\Omega (uv + Du \cdot Dv)\, dx$ is a bilinear form, symmetric, and positive and

$$L(v) = \int_\Omega f v \, dx \quad \text{is a linear form on } V.$$

The basic difference between the weak variational formulations of the Dirichlet and Neumann problems is in the choice of the space V which reflects the choice of the test functions:

$V = H_0^1(\Omega)$ in the Dirichlet problem;

$V = H^1(\Omega)$ in the Neumann problem.

The Stokes system. Given $\vec{f} = (f_1, f_1, \ldots, f_N) \in L^2(\Omega)^N$ and $\mu > 0$, we are looking for the velocity vector field of the fluid $\vec{u} = (u_1, u_2, \ldots, u_N)$ and the pressure $p : \Omega \longrightarrow \mathbf{R}$ of the fluid which satisfy

$$-\mu \Delta u_i + \frac{\partial p}{\partial x_i} = f_i \text{ on } \Omega, \quad i = 1, \ldots, N, \tag{2.38}$$

$$\operatorname{div} \vec{u} = 0 \qquad \text{on } \Omega, \tag{2.39}$$

$$u_i = 0 \qquad \text{on } \partial\Omega, \quad i = 1, \ldots, N. \tag{2.40}$$

The condition $\operatorname{div} \vec{u} = \sum_{i=1}^N \frac{\partial u_i}{\partial x_i} = 0$ expresses that the fluid is incompressible.

The choice of the test functions is not as immediate as in the two previous situations. A guideline is to choose the test functions smooth enough to perform the integration by parts and which looks like the function or vector field we want to test. A clever choice (J. Leray developed this method) is to take test fields $\vec{v} \in \mathcal{V}$, where

$$\mathcal{V} = \{\vec{v} = (v_1, \ldots, v_n),\ v_i \in \mathcal{D}(\Omega), i = 1, \ldots, N, \text{ and } \operatorname{div} \vec{v} = 0\}.$$

One may require the v_i to be $\mathbf{C}^1$ function with compact support as well. The important point is to assume that the divergence of $\vec{v}$ is equal to zero.

Let us interpret (2.38) in the sense of distributions. If we expect to find u_i with first partial derivatives in $L^2(\Omega)$ (i.e., $u_i \in H^1(\Omega)$) and $p \in L^2(\Omega)$, this is equivalent to writing for each $i = 1, 2, \ldots, N$

$$\mu \int_\Omega Du_i \cdot Dv_i \, dx - \int_\Omega p \cdot \frac{\partial v_i}{\partial x_i} \, dx = \int_\Omega f_i v_i \, dx \quad \forall v_i \in \mathcal{D}(\Omega). \tag{2.41}$$

The trick is now to add these equalities ($i = 1, 2, \ldots, N$). Since the test functions $v_1, \ldots, v_N$, by definition of $\mathcal{V}$, verify $\sum \frac{\partial v_i}{\partial x_i} = 0$, we obtain

$$\mu \sum_{i=1}^N \int_\Omega Du_i \cdot Dv_i \, dx = \sum_{i=1}^N \int_\Omega f_i v_i \, dx \quad \forall \vec{v} \in \mathcal{V}. \tag{2.42}$$

Conversely, it is easy to verify that if u is regular and satisfies (2.42), then

$$\sum_{i=1}^N \int_\Omega (-\mu \Delta u_i - f_i) v_i \, dx = 0 \quad \forall \vec{v} \in \mathcal{V}.$$

In other words, the vector $(\mu\Delta u_i + f_i)_{i=1,\dots,N}$ is orthogonal to $\mathcal{V}$ in $L^2(\Omega)^N$. One can prove—indeed, this is quite an involved result (see chapter 6)—that this property implies the existence of $p \in L^2(\Omega)$ such that

$$\mu\ \Delta u_i + f_i = \frac{\partial p}{\partial x_i}, \quad i = 1, 2, \dots, N.$$

Indeed, as in the previous examples, the equality (2.42) can be extended by a density and continuity argument to

$$V = \{\vec{v} \in H_0^1(\Omega)^N : \operatorname{div} \vec{v} = 0\}.$$

Finally, the variational formulation of the Stokes system is given below.

Definition 2.3.3. *A weak solution of the Stokes system is a solution* $\vec{u} = (u_1, u_2, \dots, u_N)$ *of the system*

$$\begin{cases} \mu \displaystyle\int_\Omega \sum_{i=1}^N Du_i \cdot Dv_i\, dx = \sum_{i=1}^N \int_\Omega f_i\, v_i\, dx \quad \forall \vec{v} \in V, \\ \vec{u} \in V, \end{cases} \tag{2.43}$$

where $V = \{\vec{v} \in H_0^1(\Omega)^N : \operatorname{div} \vec{v} = 0\}$.

The choice of the functional space V (which is obtained by a completion of the space $\mathcal{V}$ of test functions) is of fundamental importance. The pressure p has apparently disappeared in this formulation. It is contained implicitly in it, since p can be interpreted as a Lagrange multiplier of the constraint $\operatorname{div} v = 0$.

Notice that, once more, the weak formulation we have obtained can be written in the following form: find $u \in V$ such that

$$a(u, v) = L(v) \quad \forall v \in V,$$

where $a(u, v) = \mu \int_\Omega \sum_{i=1}^N Du_i\, Dv_i dx$ and $L(v) = \int_\Omega \sum f_i v_i$ are, respectively, a bilinear form and a linear form on V.

2.3.2 Positive quadratic forms and convex minimization

The weak formulations of the model examples studied in the previous section have very similar structures. Indeed, they can be viewed as particular cases of the following abstract problem.

Given V a linear vector space, $a : V \times V \longrightarrow \mathbf{R}$ a bilinear form, and $L : V \longrightarrow \mathbf{R}$ a linear form, find $u \in V$ such that

$$a(u, v) = L(v) \quad \forall v \in V. \tag{2.44}$$

In chapter 3, we will study in detail the existence of solutions to such problems. This will require some topological assumptions on the data V, a, L.

For the moment, we will examine algebraic properties of such problems and make the link, when $a(\cdot, \cdot)$ is symmetric and positive, with convex minimization problems.

Let us first make precise these notions concerning bilinear and quadratic forms.

Definition 2.3.4. *Let V be a linear vector space and $a : V \times V \longrightarrow \mathbf{R}$ a bilinear form, i.e.,*

$$\begin{cases} \forall\, u \in V \quad v \longmapsto a(u, v) \text{ is a linear form,} \\ \forall\, v \in V \quad u \longmapsto a(u, v) \text{ is a linear form.} \end{cases}$$

The bilinear form is said to be symmetric if

$$\forall u, v \in V \quad a(u, v) = a(v, u).$$

When a is symmetric, one can associate to $a(\cdot, \cdot)$ the quadratic form $q : V \longrightarrow \mathbf{R}$ which is equal to $q(v) = a(v, v)$.

The bilinear form a is said to be positive (one can say as well that the associated quadratic form q is positive) if

$$\forall\, v \in V \quad a(v, v) \geq 0.$$

We say that $a(\cdot, \cdot)$ (or $q(\cdot)$) is positive definite if

$$\forall\, v \in V \quad a(v, v) \geq 0 \quad \textit{and} \quad a(v, v) = 0 \Longrightarrow v = 0.$$

We can now make the link between problem (2.44) and a minimization problem. All the notions used in the following statement are algebraic.

Proposition 2.3.1. *Let V be a linear vector space, $L : V \longrightarrow \mathbf{R}$ a linear form, and $a : V \times V \longrightarrow \mathbf{R}$ a bilinear, symmetric, positive form. Then the two following statements are equivalent:*

(i) $u \in V$, $a(u, v) = L(v)\ \forall v \in V$;

(ii) $u \in V$, $J(u) \leq J(v)\ \forall\, v \in V$, *where*

$$J(v) := \frac{1}{2} a(v, v) - L(v).$$

Proof. Let us first prove (i) $\Longrightarrow$ (ii). Since V is a linear space, it is equivalent to prove that

$$J(u) \leq J(u + v) \quad \forall v \in V.$$

A simple computation gives

$$J(u + v) - J(u) = \left[\frac{1}{2} a(u + v, u + v) - L(u + v)\right] - \left[\frac{1}{2} a(u, u) - L(u)\right].$$

Note that because of the symmetry assumption on the bilinear form $a(\cdot, \cdot)$,

$$a(u + v, u + v) = a(u, u) + 2a(u, v) + a(v, v).$$

Thus

$$\begin{aligned} J(u + v) - J(u) &= \left[\frac{1}{2} a(u, u) + a(u, v) + \frac{1}{2} a(v, v)\right] \\ &\quad - \frac{1}{2} a(u, u) - [L(u) + L(v)] + L(u) \\ &= [a(u, v) - L(v)] + \frac{1}{2} a(v, v). \end{aligned}$$

Since, by assumption, u is a solution of (i), $a(u, v) = L(v)$ and

$$J(u + v) - J(u) = \frac{1}{2}a(v, v),$$

which is nonnegative, since a has been assumed to be positive.

Let us now prove (ii) $\Longrightarrow$ (i). We know that u is a solution of the minimization problem, i.e., u minimizes $J(\cdot)$. One is naturally tempted to write an optimality condition which expresses that some derivative of J at u is equal to zero. Since V was only assumed to be a linear vector space, the only derivation notion we can use is the directional derivative which always makes sense since it relies only on the topological structure of the real line. Since u minimizes J, for any $t \in \mathbf{R}$, for any $v \in V$,

$$J(u + tv) - J(u) \geq 0.$$

Dividing by $t > 0$, we have

$$\frac{1}{t}[J(u + tv) - J(u)] \geq 0.$$

Before letting t go to zero, let us compute this last expression,

$$\begin{aligned}\frac{1}{t}[J(u + tv) - J(u)] &= \frac{1}{t}\left[\frac{1}{2}a(u + tv, u + tv) - L(u + tv) - \frac{1}{2}a(u, u) + L(u)\right] \\ &= \frac{1}{t}\left[ta(u, v) + \frac{t^2}{2}a(v, v) - tL(v)\right] \\ &= a(u, v) + \frac{t}{2}a(v, v) - L(v).\end{aligned}$$

Thus, by letting t go to zero, we obtain

$$\lim_{t \to 0^+} \frac{1}{t}[J(u + tv) - J(u)] = a(u, v) - L(v) \geq 0.$$

Then, one can either make the same argument by using $t < 0$ or replace v by $-v$ in the above inequality to obtain the opposite inequality and conclude that

$$a(u, v) = L(v) \quad \forall v \in V. \qquad \square$$

Let us return to the model examples studied in Section 2.3.1 and use their weak formulations together with Proposition 2.3.1 to obtain the results below.

Corollary 2.3.1. *With the notation of Section* 2.3.1*, the following facts hold.*

(a) *The weak solution u of the Dirichlet problem*

$$\begin{cases} -\Delta u = f & \text{on } \Omega, \\ u = 0 & \text{on } \partial\Omega \end{cases}$$

is a solution of the minimization problem

$$\begin{cases} J(u) \leq J(v) & \forall\, v \in H_0^1(\Omega), \\ u \in H_0^1(\Omega), \end{cases}$$

where $J(v) := \frac{1}{2}\int_\Omega |Dv|^2 dx - \int_\Omega fv dx$. This is the Dirichlet variational principle.

(b) *The weak solution u of the Neumann problem*

$$\begin{cases} u - \Delta u = f & \text{on } \Omega, \\ \frac{\partial u}{\partial n} = 0 & \text{on } \partial\Omega \end{cases}$$

is a solution of the minimization problem

$$\begin{cases} J(u) \leq J(v) \quad \forall\, v \in H^1(\Omega), \\ u \in H^1(\Omega), \end{cases}$$

where $J(v) := \frac{1}{2}\int_\Omega (|Dv|^2 + v^2)dx - \int_\Omega f v\, dx$.

(c) *The weak solution u of the Stokes system*

$$\begin{cases} -\mu \Delta u_i + \frac{\partial p}{\partial x_i} = f_i & i = 1, 2, \ldots, N \text{ on } \Omega, \\ \operatorname{div} \vec{u} = 0 & \text{on } \Omega, \\ \vec{u} = \vec{0} & \text{on } \partial\Omega \end{cases}$$

is a solution of the minimization problem

$$\begin{cases} J(u) \leq J(v) \quad \forall\, v \in V = \{v \in H_0^1(\Omega)^N : \operatorname{div} v = 0\}, \\ u \in V, \end{cases}$$

where $J(v) = \frac{1}{2}\sum_{i=1}^N \int_\Omega |Dv_i|^2 dx - \sum_{i=1}^N \int_\Omega f_i v_i dx$.

Note that in all these examples, the bilinear form $a(\cdot, \cdot)$ is symmetric and positive. We now come to the question of the nature of the minimization problem and the properties of the functional J given by

$$J(v) = \frac{1}{2} a(v, v) - L(v).$$

Note that J is the sum of a quadratic form and of a linear form.

We are ready to introduce a property of fundamental importance in the study of the minimization problems—convexity.

Recall that a function $J : V \longrightarrow \mathbf{R}$, where V is a linear vector space, is convex if

$$\forall\, u, v \in V \quad \forall\, \lambda \in [0, 1] \quad J(\lambda u + (1 - \lambda)v) \leq \lambda J(u) + (1 - \lambda) J(v).$$

The role of convexity in minimization problems will be examined in detail in chapters 3, 9, 13, and 15. The class of convex functionals is stable with respect to the sum, it contains the linear forms, and we are going to see that it contains the positive quadratic forms. Thus, functionals of the form $J(v) = \frac{1}{2} a(v, v) - L(v)$ with a and L as above will be convex.

Let us now formulate the convexity property for the positive quadratic forms.

Proposition 2.3.2. *Let V be a linear vector space and $a : V \times V \longrightarrow \mathbf{R}$ a bilinear form which is symmetric and positive. Then, the quadratic form $q : V \longrightarrow \mathbf{R}$ which is associated with a, i.e., $q(v) = a(v, v)$ is a convex function.*

PROOF. This is just an algebraic computation. For any $u, v \in V$ and $\lambda \in [0, 1]$,

$$\begin{aligned} q(\lambda u + (1-\lambda)v) &= a(\lambda u + (1-\lambda)v, \lambda u + (1-\lambda)v) \\ &= \lambda^2 a(u,u) + (1-\lambda)^2 a(v,v) + 2\lambda(1-\lambda)a(u,v). \end{aligned}$$

Thus,

$$\begin{aligned} \lambda q(u) + (1-\lambda)q(v) - q(\lambda u + (1-\lambda)v) &= (\lambda - \lambda^2)a(u,u) - 2\lambda(1-\lambda)a(u,v) \\ &\quad + [(1-\lambda) - (1-\lambda)^2]a(v,v) \\ &= \lambda(1-\lambda)[a(u,u) - 2a(u,v) + a(v,v)] \\ &= \lambda(1-\lambda)a(u-v, u-v), \end{aligned}$$

which is nonnegative, because $\lambda \in [0, 1]$ and a is positive. □

When examinating the question of the uniqueness of the solution of the previous problems, the notion which plays a central role is the strict convexity. Recall that $J : V \longrightarrow \mathbf{R}$ is strictly convex if J is convex and the convexity inequality

$$J(\lambda u + (1-\lambda)v) < \lambda J(u) + (1-\lambda)J(v)$$

is strict whenever $u \neq v$ and $\lambda \in]0, 1[$.

The importance of this notion is justified by the following elementary result.

Proposition 2.3.3. *Let V be a linear space and $J : V \longrightarrow \mathbf{R}$ a strictly convex function. Then there exists at most one solution u to the minimization problem*

$$\begin{cases} J(u) \leq J(v) \; \forall \, v \in V, \\ u \in V. \end{cases}$$

PROOF. Suppose that we have two distinct solutions u_1 and u_2 to the above minimization problem. Then

$$J\left(\frac{u_1 + u_2}{2}\right) < \frac{1}{2}[J(u_1) + J(u_2)] = \inf J(u),$$

a clear contradiction. Hence $u_1 = u_2$. □

Proposition 2.3.4. *Let V be a linear vector space and $a : V \times V \longrightarrow \mathbf{R}$ a bilinear form which is symmetric and positive definite. Then, the quadratic form $q : V \longrightarrow \mathbf{R}$ which is associated with a, i.e., $q(v) = a(v, v)$ is strictly convex.*

PROOF. The proof is the same computation as in the proof of Proposition 2.3.2: for any $u, v \in V$, for any $\lambda \in [0, 1]$,

$$\lambda q(u) + (1-\lambda)q(v) - q(\lambda u + (1-\lambda)v) = \lambda(1-\lambda)a(u-v, u-v).$$

When taking $\lambda \in]0, 1[$ we have $\lambda(1-\lambda) > 0$, and when taking $u \neq v$ we have $a(u-v, u-v) > 0$ because a is positive definite. So, for $\lambda \in]0, 1[$ and $u \neq v$, $\lambda q(u) + (1-\lambda)q(v) > q(\lambda u + (1-\lambda)v)$ and q is strictly convex. □

Proposition 2.3.5. *The sum of a convex function and of a strictly convex function is strictly convex.*

PROOF. The proof is a direct consequence of the fact that when adding an inequality and a strict inequality, one obtains a strict inequality. □

Let us return to the model examples and their variational formulations as minimization problems as given in Corollary 2.3.1. Noticing that in all these situations, the quadratic form $q(v) = a(v, v)$ is positive definite, we obtain that the corresponding functional J is strictly convex. So, the weak solution of the problems under consideration, when it exists, is characterized as the unique solution of the associated minimization problems. This makes a natural transition to chapter 3, where the existence question will be examined.

2.4 Weak topologies and weak convergences

In recent decades, weak topologies have proved useful as a basic tool in variational analysis in the study of PDEs, and more generally in all fields using tools from functional analysis. Let us explain some of the reasons for the success of weak convergence methods.

(a) Distributions are defined as continuous linear forms on $\mathcal{D}(\Omega)$. In other words, a distribution $T \in \mathcal{D}'(\Omega)$ is viewed via its action on test functions $v \in \mathcal{D}(\Omega)$:

$$T \in \mathcal{D}'(\Omega) : v \in \mathcal{D}(\Omega) \longmapsto \langle T, v\rangle_{(\mathcal{D}'(\Omega),\mathcal{D}(\Omega))}.$$

Given a sequence $T_1, T_2, \ldots, T_n, \ldots$ of distributions (for example, functions, measures), a natural mode of convergence for such sequences is to assume that

$$\forall\, v \in \mathcal{D}(\Omega) \quad \lim_{n\to\infty} \langle T_n, v\rangle_{(\mathcal{D}',\mathcal{D})} = \langle T, v\rangle.$$

This is a typical example of weak convergence.

(b) A celebrated theorem from Riesz asserts that the closed unit ball of a normed linear space is compact iff the space has a finite dimension. Thus, when looking for topologies making bounded sets relatively compact in infinite dimensional spaces, one is naturally led to introduce new topologies which are weaker than the topology of the norm. This is why weak topologies play a decisive role.

(c) Besides the importance of weak topologies from a theoretical point of view, we will see that weak convergences naturally occur when describing concrete situations. For example, weak convergences allow us to describe high oscillations of a sequence of functions, as well as concentration phenomena on zero Lebesgue measure sets.

Before introducing weak toplologies on normed linear spaces, let us recall some basic facts from general topology.

2.4.1 Topologies induced by functions in general topological spaces

First we need to fix the notation. Recall that a topology on a space X is a family θ of subsets of X, called the family of the *open sets* of X, satisfying the axioms of the open sets, namely,

(i) X and $\emptyset$ belong to θ;

(ii) $\forall (G_i)_{i\in I}\ G_i \in \theta$, I arbitrary, $\cup_{i\in I} G_i \in \theta$;

(iii) $\forall (G_i)_{i\in I}\ G_i \in \theta$, I finite, $\cap_{i\in I} G_i \in \theta$.

In other words, the open sets of X for a given topology is a family of subsets of X which is *stable with respect to arbitrary unions and finite intersection*.

We will often denote by τ a topology on a space X and by θ_τ the family of the τ-open sets. A topology can be seen as a subset of $P(X)$, where $P(X)$ is the family of all subsets of X. There is a natural partial ordering on the topologies on a given space X, which is induced by the inclusion ordering on the subsets of $P(X)$: we will say that a topology τ_1 *is coarser or weaker* than a topology τ_2 and we write $\tau_1 < \tau_2$ if $\theta_{\tau_1} \subset \theta_{\tau_2}$, that is, if any element G of θ_{τ_1} also belongs to θ_{τ_2}. Conversely, we will say that τ_2 is *stronger or finer than* τ_1.

Proposition 2.4.1. *The family of the topologies on a set X forms a complete lattice for the relation $\tau_1 < \tau_2$ (τ_1 weaker than τ_2), that is, given an arbitrary collection of topologies $(\tau_i)_{i\in I}$ on X,*

(a) *There exists a lower bound, that is a topology which is the largest among all the topologies weaker than the τ_i, $i \in I$. We denote $\tau = \wedge_{i\in I}\tau_i$ the lower bound (or infimum) of the τ_i. Clearly $\theta_\tau = \cap_{i\in I}\theta_{\tau_i}$, that is $G \in \theta_\tau$ iff G belongs to θ_{τ_i} for all $i \in I$.*

(b) *There exists an upper bound (or supremum) that is a topology which is the smallest among the topologies which are stronger than all the τ_i. We denote $\tau = \vee_{i\in I}\tau_i$ the upper bound of the τ_i. We have that θ_τ is generated by $\cup_{i\in I}\theta_{\tau_i}$ in the sense of Proposition* 2.4.2.

Proof. (a) Clearly if $(\theta_{\tau_i})_{i\in I}$ is a family of topologies on X, then $\cap_{i\in I}\theta_{\tau_i} = \{G \in P(X) : G \in \theta_{\tau_i}\ \forall i \in I\}$ still satisfies the axioms of the open sets, and it is a topology. The topology τ attached to the family $\theta = \cap_{i\in I}\theta_{\tau_i}$ is weaker than all the topologies τ_i, $i \in I$, and clearly it is the largest among the weaker ones.

(b) In contradiction to the previous case, if a topology τ is stronger than all the topologies τ_i, then θ_τ must contain all the families θ_{τ_i}, that is,

$$\theta_\tau \supset \cup_{i\in I}\theta_{\tau_i}.$$

But now $\mathcal{A} = \cup_{i\in I}\theta_{\tau_i}$ does not satisfy (in general) the axioms of the open sets. So, one is naturally led to address the following question: given a class $\mathcal{A}$ of subsets of an abstract space X, does there exist a smallest topology θ_τ on X which contains $\mathcal{A}$?

Clearly by (a), the answer is yes, one has to take for θ_τ the intersection (or, with an equivalent terminology, the infimum) of all the topologies containing $\mathcal{A} = \cup_{i\in I}\theta_{\tau_i}$. This is made precise in Proposition 2.4.2. □

Proposition 2.4.2. *Let X be an abstract space and let $\mathcal{A}$ be any class of subsets of X. Then there exists a smallest (weakest) topology on X containing $\mathcal{A}$, denoted by $\tau_{\mathcal{A}}$, called the topology generated by $\mathcal{A}$. It is equal to the intersection of all the topologies containing $\mathcal{A}$. It can be obtained via the following two-step procedure:*

1. *First, take the finite intersections of elements of $\mathcal{A}$. One so obtains a family of sets which we call $\mathcal{B}_{\mathcal{A}}$.*

2. *Then $\mathcal{B}_{\mathcal{A}}$ is a base for the topology τ_A generated by $\mathcal{A}$, that is, any member of $\tau_{\mathcal{A}}$ can be obtained as the union of a family of members of $\mathcal{B}_{\mathcal{A}}$.*

We stress that in this construction, one has first to take finite intersections of elements of $\mathcal{A}$, then arbitrary unions of the so-obtained sets. When reversing the two operations, one obtains a family which is not stable by union. For a proof, see one of several books on general topology (for instance, Bourbaki [79]).

We now come to the situation which is of interest when considering weak topologies. Suppose X is an abstract space and $(Y_i, \tau_i)_{i\in I}$ is a family of topological spaces. Suppose that for each $i \in I$, a function $f_i : X \to Y_i$ is given. We want to investigate the topologies on X with respect to which all the functions f_i are continuous, and, among these topologies, examine the question of the existence of a smallest (weakest) one.

Noticing that for each $i \in I$, $f_i^{-1}(\theta_{\tau_i}) := \{f_i^{-1}(G_i) : G_i \in \theta_{\tau_i}\}$ still satisfies the axioms of the open sets, we denote $f_i^{-1}(\tau_i)$ the corresponding topology on X, which is the weakest making f_i (for i fixed) continuous. It follows from Propositions 2.4.1 and 2.4.2 that the answer to the previous question is given by $\tau = \vee_{i\in I} f_i^{-1}(\tau_i)$, whose precise description is given in the following.

Theorem 2.4.1. *Let X be an abstract space and let $(Y_i, \tau_i)_{i\in I}$ be an arbitrary collection of topological spaces with for each $i \in I$, $f_i : X \to Y_i$ a given function. Then, there exists a weakest topology τ on X making all the functions $(f_i)_{i\in I}$ continuous, $f_i : (X, \tau) \longrightarrow (Y_i, \tau_i)$ $i \in I$. This topology τ is equal to $\vee_{i\in I} f_i^{-1}(\tau_i)$, that is, τ is generated by the class*

$$\mathcal{A} = \{f_i^{-1}(G_i), G_i \in \theta_{\tau_i}, i \in I\}.$$

The class

$$\mathcal{B}_\tau = \left\{\bigcap_{i\in J} f_i^{-1}(G_i),\ G_i \in \theta_{\tau_i},\ J \subset I,\ J \text{ finite}\right\}$$

is a base for this topology, that is, each element of θ_τ can be written as a union of elements of $\mathcal{B}_\tau$.

We say that the topology τ is *induced by the family* $(f_i)_{i\in I}$. The following properties are quite elementary.

Proposition 2.4.3. *Let $f_i : X \longrightarrow (Y_i, \tau_i)$, $i \in I$, be given, and let $\tau = \vee f_i^{-1}(\tau_i)$ be the topology on X induced by the $(f_i)_{i\in I}$. For any sequence $(x_n)_{n\in\mathbf{N}}$ of elements of X, the two conditions are equivalent:*

(i) $x_n \xrightarrow{\tau} x$ *as* $n \to \infty$;

(ii) $\forall i \in I \quad f_i(x_n) \xrightarrow{\tau_i} f_i(x)$ *as* $n \to \infty$.

PROOF. Since the topology τ makes each f_i continuous, we have clearly (i) $\Longrightarrow$ (ii).

Conversely, let us assume (ii) and prove (i). When considering a neighborhood of x, it is equivalent to take an element of the base $\mathcal{B}_\tau$ which contains x. So let, $x \in \bigcap_{i\in J} f_i^{-1}(G_i)$, $G_i \in \theta_{\tau_i}$, $J \subset I$, J finite. For each $i \in J$, since $f_i(x_n) \xrightarrow{\tau_i} f_i(x)$ we have that $x_n \in f_i^{-1}(G_i)$ for $n \geq n_i$. Take $N = \max_{i\in J} n_i$, since J is finite, N is a finite integer, and $x_n \in \bigcap_{i\in J} f_i^{-1}(G_i)$ for $n \geq N$, which expresses that $x_n \xrightarrow{\tau} x$. □

A similar type argument yields the following result.

Proposition 2.4.4. *Let $(Z, \mathcal{T})$ be a topological space and let $g : (Z, \mathcal{T}) \longrightarrow (X, \tau)$ be a given function, where τ is the topology induced by the family $f_i : X \longrightarrow (Y_i, \tau_i)$. Then g is continuous iff $f_i \circ g$ is continuous from $(Z, \mathcal{T})$ into (Y_i, τ_i) for each $i \in I$.*

2.4.2 The weak topology $\sigma(V, V^*)$

We now assume that X is a vector space. To enhance this property, we denote it by V (like vector) and assume that V is a normed linear space, the norm of $v \in V$ being denoted by $\|v\|_V$ or $\|v\|$ when no confusion is possible.

We denote by V^* the topological dual of V, which is the set of all linear continuous forms on V. To avoid confusion, generic elements of V and V^* are denoted, respectively, by $v \in V$ and $v^* \in V^*$. We will write $\langle v^*, v\rangle = v^*(v)$ for the canonical pairing between V^* and V, which is just the evaluation of $v^* \in V^*$ at $v \in V$. Recall that V^* is a normed linear space (indeed, it is a Banach space) when equipped with the dual norm

$$\|v^*\|_{V^*} = \sup\{|v^*(v)| : \|v\|_V \leq 1\}.$$

With this definition, we have

$$\forall v \in V, \forall v^* \in V^* \quad |\langle v^*, v\rangle| \leq \|v^*\| \, \|v\|,$$

and $\|v^*\|$ is precisely the smallest constant for which the above inequality holds.

Definition 2.4.1. *Let $(V, \|\cdot\|)$ be a normed linear space with topological dual V^*. The topology $\sigma(V, V^*)$, called the weak topology on V, is the weakest topology on V making continuous all the elements of V^*.*

Let us first comment on this definition. By definition, each element $v^* \in V^*$ is a function from V into $\mathbf{R}$, that is,

$$v^* : V \longrightarrow \mathbf{R}, \qquad v \longmapsto \langle v^*, v\rangle_{(V^*, V)}.$$

The weak topology on V is defined as the weakest topology on V making all these functions $\{v^* : v^* \in V^*\}$ continuous. By Theorem 2.4.1, such a topology exists, and it is weaker than the norm topology (since by definition all the elements v^* of V^* are continuous for the norm topology). We collect below some first results on the topology $\sigma(V, V^*)$ which are direct consequences of its definition.

Proposition 2.4.5. *Let V be a normed space and $\sigma(V, V^*)$ the weak topology on V.*

(i) *A local base of neighborhoods of $v_0 \in V$ for $\sigma(V, V^*)$ consists of all sets of the form*

$$N(v_0) = \{v \in V : |\langle v_i^*, v - v_0\rangle| < \varepsilon \ \forall i \in I\},$$

where I is a finite index set, $v_i^* \in V^*$ for each $i \in I$, and $\varepsilon > 0$.

(ii) *$(V, \sigma(V, V^*))$ is a Hausdorff topological space.*

(iii) *The topology $\sigma(V, V^*)$ is coarser than the topology of the norm on V.*

(iv) *When V is finite dimensional, the weak topology and the norm topology coincide.*

(v) *When V is infinite dimensional, the weak topology $\sigma(V, V^*)$ is strictly coarser than the norm topology.*

PROOF. (i) We have

$$N(v_0) = \bigcap_{i \in I} (v_i^*)^{-1}(]\alpha_i - \varepsilon, \alpha_i + \varepsilon[)$$

where $\alpha_i = \langle v_i^*, v_0 \rangle$. By definition of the weak topology $\sigma(V, V^*)$, $N(v_0)$ is an open set for this topology.

Let us prove that such sets form a local base of open neighborhoods of v_0 for $\sigma(V, V^*)$. Take A an open set for $\sigma(V, V^*)$ containing v_0. By Theorem 2.4.1, there exists some open set B for $\sigma(V, V^*)$ such that

$$v_0 \in B \subset A$$

with $B = \bigcap_{i \in I} (v_i^*)^{-1}(G_i)$, $v_i^* \in V^*$, G_i open in $\mathbf{R}$, I finite.

Since $v_i^*(v_0) \in G_i$ and G_i is open in $\mathbf{R}$, there exists some $\varepsilon > 0$ such that $|v_i^*(v) - v_i^*(v_0)| < \varepsilon$ for all $i \in I$ implies $v_i^*(v) \in G_i$ for all $i \in I$. Hence

$$v_0 \in N(v_0) \subset B \subset A$$

with

$$N(v_0) = \bigcap_{i \in I} (v_i^*)^{-1}(]\alpha_i - \varepsilon, \alpha_i + \varepsilon[), \ \alpha_i = \langle v_i^*, v_0 \rangle.$$

(ii) Let us prove that the topology $\sigma(V, V^*)$ is Hausdorff. Take v_1 and v_2 two distinct elements of V and prove that there exist A_1 and A_2 two open sets for $\sigma(V, V^*)$ such that $v_1 \in A_1$, $v_2 \in A_2$, and $A_1 \cap A_2 = \emptyset$. This is a direct consequence of the Hahn–Banach separation theorem. There exists a closed hyperplane which strictly separates v_1 and v_2, that is, there exists some $v^* \in V^*$ and $\alpha \in \mathbf{R}$ such that

$$\langle v^*, v_1 \rangle < \alpha < \langle v^*, v_2 \rangle.$$

Take

$$A_1 = \{v \in V : \langle v^*, v \rangle < \alpha\},$$
$$A_2 = \{v \in V : \langle v^*, v \rangle > \alpha\}.$$

They are open for the topology $\sigma(V, V^*)$ and separate v_1 and v_2.

Assertion (iii) is obvious since all the elements v^* of V^* are continuous for the norm topology.

(iv) Since the weak topology $\sigma(V, V^*)$ is coarser than the norm topology, it has fewer open sets. Let us prove that when V is a finite dimensional space, the opposite inclusion is true, that is, any open set A for the norm topology is also an open set for the weak topology.

Take $v_0 \in B(v_0, \varepsilon) \subset A$, where $B(v_0, \varepsilon)$ is an open ball in $(V, \| \cdot \|)$, and prove that there exists some open set U for $\sigma(V, V^*)$ such that

$$v_0 \in U \subset B(v_0, \varepsilon) \subset A.$$

Let us choose a base $e_1, \dots, e_N$ of V with $\|e_i\| = 1$, $i = 1, \dots, N$. Each element v of V can be uniquely written as $v = \sum x_i e_i$ and the mappings $v \xrightarrow{e_i^*} x_i$ are linear continuous forms on V, i.e., they are elements of V^*. We have (with $v_0 = \sum x_{oi} e_i$)

$$\begin{aligned}\|v - v_0\| &= \|\sum_i (x_i - x_{oi}) e_i\| \\ &\leq \sum_i |x_i - x_{oi}| \\ &\leq \sum_i |\langle e_i^*, v - v_o\rangle|.\end{aligned}$$

Therefore, $v \in B(v_0, \varepsilon)$ as soon as $v \in U := \bigcap_{i=1,\dots,n} (e_i^*)^{-1}(]\alpha_i - \frac{\varepsilon}{N}, \alpha_i + \frac{\varepsilon}{N}[)$, where $\alpha_i = \langle e_i^*, v_0\rangle = x_{oi}$. Then notice that U is open for $\sigma(V, V^*)$, which concludes the proof of (iv).

(v) There are different ways to prove that in infinite dimensional spaces the weak topology is strictly coarser than the norm topology. One of them consists of proving that the unit sphere $S = \{v \in V : \|v\| = 1\}$ is never closed in infinite dimensional spaces for the topology $\sigma(V, V^*)$. Indeed $\bar{S}^{\sigma(V,V^*)} = \{v \in V : \|v\| \leq 1\}$; see, for instance, [90, Proposition III.6] and related comments for a detailed proof.

Remark 2.4.1. It is quite convenient when formulating topological properties to express them with the help of sequences. This can be done without loss of generality when the topology under consideration is metrizable. But *the weak topology* $\sigma(V, V^*)$ when V is an infinite dimensional normed space is a locally convex topology (the basic operations on V, vectorial sum and multiplication by a scalar, are continuous for the topology $\sigma(V, V^*)$) which is *not metrizable*.

Therefore, it is important to state the properties of this topology with general topological arguments as we have done up to now.

Nevertheless, in most pratical situations, one can just use weakly convergent sequences. This will follow from deep results like the Eberlein–Smulian compactness theorem or from simpler observations like the following one: if V^* is separable, the weak topology $\sigma(V, V^*)$ is metrizable on each bounded set of V.

Consequently we now focus on properties of sequences which are $\sigma(V, V^*)$ convergent. Let us start with the following elementary results, which are direct consequences of the definition and of Proposition 2.4.3.

Proposition 2.4.6. *Let V be a normed linear space and $\sigma(V, V^*)$ the weak topology on V. For any sequence $(v_n)_{n\in\mathbf{N}}$ in V the following properties hold:*

(i) $v_n \xrightarrow{\sigma(V,V^*)} v \iff \forall v^* \in V^* \ \langle v^*, v_n\rangle \longrightarrow \langle v^*, v\rangle$;

(ii) $v_n \xrightarrow{\|\cdot\|} v \implies v_n \xrightarrow{\sigma(V,V^*)} v$;

(iii) $v_n \xrightarrow{\sigma(V,V^*)} v \implies$ *the sequence* $(v_n)_{n\in\mathbf{N}}$ *is bounded and* $\|v\| \leq \liminf_n \|v_n\|$;

(iv) $v_n \xrightarrow{\sigma(V,V^*)} v$ *and* $v_n^* \xrightarrow{\|\cdot\|_*} v^* \implies \langle v_n^*, v_n\rangle \longrightarrow \langle v^*, v\rangle$.

PROOF. (iii) The fact that the sequence $(\|v_n\|)_{n\in\mathbf{N}}$ is bounded is a consequence of the Banach–Steinhaus theorem: consider the family of linear operators from the Banach space V^* into $\mathbf{R}$

$$T_n : V^* \longrightarrow \mathbf{R} \quad n \in \mathbf{N}, \qquad v^* \longmapsto \langle v^*, v_n\rangle.$$

For each $n \in \mathbf{N}$, T_n is a linear continuous operator with norm

$$\|T_n\|_{\mathcal{L}(V^*,\mathbf{R})} = \sup_{\|v^*\|_*\leq 1} |\langle v^*, v_n\rangle| = \|v_n\|.$$

(This last equality is a consequence of the Hahn–Banach theorem.) For each $v^* \in V^*$, the sequence $(T_n(v^*))n \in \mathbf{N}$ is bounded in $\mathbf{R}$. This is a direct consequence of the equality $T_n(v^*) = \langle v^*, v_n\rangle$ and of the weak convergence of the sequence $(v_n)_{n\in\mathbf{N}}$. By the Banach–Steinhaus theorem, $\sup_{n\in\mathbf{N}} \|T_n\|_{\mathcal{L}(V^*,\mathbf{R})} < +\infty$, which is equivalent to $\sup_{n\in\mathbf{N}} \|v_n\| < \infty$.

Let us now prove the inequality $\|v\| \leq \liminf_n \|v_n\|$. By assumption, for each $v^* \in V^*$

$$\langle v^*, v\rangle = \lim_{n\to+\infty} \langle v^*, v_n\rangle.$$

By using the inequality $|\langle v^*, v_n\rangle| \leq \|v^*\|_*\|v_n\|$, we infer

$$\forall v^* \in V^* \qquad |\langle v^*, v\rangle| \leq (\liminf_{n\to+\infty} \|v_n\|)\|v^*\|_*.$$

By using the Hahn–Banach theorem, we obtain

$$\|v\| = \sup_{\|v^*\|_*\leq 1} |\langle v^*, v\rangle| \leq \liminf_n \|v_n\|.$$

(iv) This is just a triangulation argument. Write

$$\langle v_n^*, v_n\rangle - \langle v^*, v\rangle = \langle v_n^* - v^*, v_n\rangle + \langle v^*, v_n - v\rangle.$$

Hence

$$|\langle v_n^*, v_n\rangle - \langle v^*, v\rangle| \leq \|v_n^* - v^*\|_*\|v_n\| + |\langle v^*, v_n - v\rangle|.$$

The previous result (iii) tells us that there exists some constant $C \in \mathbf{R}^+$ such that $\|v_n\| \leq C$ for all $n \in \mathbf{N}$. So,

$$|\langle v_n^*, v_n\rangle - \langle v^*, v\rangle| \leq C\|v_n^* - v^*\| + |\langle v^*, v_n - v\rangle|,$$

which clearly implies the result. □

Remark 2.4.2. We will interpret in Section 3.2.3 the property

$$v_n \overset{\sigma(V,V^*)}{\longrightarrow} v \Longrightarrow \|v\| \leq \liminf_n \|v_n\|$$

as a lower semicontinuity property of the norm $\|\cdot\|_V$ for the topology $\sigma(V, V^*)$. Indeed, more generally, this can be viewed as a consequence of the fact that $\|\cdot\|_V$ is convex and continuous on V (and hence lower semicontinuous for the topology $\sigma(V, V^*)$).

Because of its importance, let us say a few words about the weak convergence in Hilbert spaces (which we denote by H). The Riesz representation theorem tells us that any element of the topological dual space can be represented as

$$H \ni v \longmapsto \langle f, v\rangle,$$

where f is a given element of H and $\langle\cdot,\cdot\rangle$ is the scalar product in H. Let us complete this observation by a few elementary results.

Proposition 2.4.7. *Let H be a Hilbert space. A sequence $(v_n)_{n\in\mathbf{N}}$ is weakly convergent in H iff*

$$\forall z \in H \quad \langle v_n, z\rangle \underset{n\to+\infty}{\longrightarrow} \langle v, z\rangle.$$

Moreover, we have the following implication:

$$v_n \overset{\sigma(H,H)}{\longrightarrow} v \quad \textit{and} \quad \|v_n\| \longrightarrow \|v\| \Longrightarrow v_n \overset{\|\cdot\|}{\longrightarrow} v.$$

PROOF. We just need to prove the last statement. We have

$$\|v_n - v\|^2 = \|v_n\|^2 + \|v\|^2 - 2\langle v_n, v\rangle.$$

Hence

$$\lim_{n\to+\infty} \|v_n - v\|^2 = \|v\|^2 + \|v\|^2 - 2\langle v, v\rangle = 0,$$

that is, $v_n \overset{\|\cdot\|}{\longrightarrow} v$. □

In the next section we will prove that this last property, which is quite important in the applications, is valid in a much larger class than the Hilbert spaces, namely, the uniformly convex Banach spaces. For the moment, we pause in these theoretical developments to give some examples of sequences which are $\sigma(V, V^*)$ convergent but not $\|\cdot\|_V$ convergent.

Example 2.4.1. Take $V = l^2$. An element v of V is a sequence of real numbers, $v = (v_k)_{k\in\mathbf{N}}$, such that $\sum_{k\in\mathbf{N}} |v_k|^2 < +\infty$. The scalar product $\langle u, v\rangle := \sum_{k\in\mathbf{N}} u_k v_k$ and the corresponding norm $\|v\| = (\sum |v_k|^2)^{1/2}$ give to V a Hilbert space structure.

Consider the sequence $e_1, e_2, \ldots, e_n, \ldots$ with

$$e_n = (\delta_{n,k})_{k\in\mathbf{N}},$$

where $\delta_{n,k}$ (the Kronecker symbol) takes the value 1 if $k = n$ and 0 elsewhere. The family $(e_n)_{n\in\mathbf{N}}$ is called the canonical basis of l^2 (it is a Hilbertian basis).

Let us show that $(e_n)_{n\in\mathbf{N}}$ weakly converges to 0 in V, that is,

$$\forall v \in V \quad \langle e_n, v\rangle \underset{n\to+\infty}{\longrightarrow} 0.$$

Observe that $\langle e_n, v\rangle = v_n$, where $v = (v_n)_{n\in\mathbf{N}}$. Since $\Sigma_{n\in\mathbf{N}}|v_n|^2 < +\infty$, the general term of this convergent series, that is, v_n, tends to zero as n goes to $+\infty$, which proves the result.

The sequence $(e_n)_{n\in\mathbf{N}}$ is not norm convergent in V; otherwise it would necessarly norm-converge to zero (recall that the norm convergence implies the weak convergence). This is impossible since for each $n \in \mathbf{N}$, $\|e_n\| = 1$. This can be equivalently obtained when observing that for all $n \neq m$

$$\|e_n - e_m\| = \sqrt{2},$$

the sequence $(e_n)_{n\in\mathbf{N}}$ is not a Cauchy sequence and hence it is not norm convergent.

Example 2.4.2 (weak convergence in $L^p(\Omega)$, $1 \leq p < \infty$). Take Ω a bounded open set in $\mathbf{R}^N$, $1 \leq p < \infty$, and

$$V = L^p(\Omega) = \left\{ v : \Omega \longrightarrow \mathbf{R} \text{ Lebesgue measurable: } \int_\Omega |v(x)|^p dx < +\infty \right\}.$$

V equipped with the norm $\|v\| = \left(\int_\Omega |v(x)|^p dx\right)^{1/p}$ is a Banach space with dual $V^* = L^{p'}(\Omega)$, $\frac{1}{p} + \frac{1}{p'} = 1$ (with the convention that the conjugate exponent of 1 is $+\infty$, i.e., $L^1(\Omega)^* = L^\infty(\Omega)$).

The weak convergence in $V = L^p(\Omega)$ can be formulated as follows:

$$v_n \overset{\sigma(L^p,L^{p'})}{\longrightarrow} v \iff \forall z \in L^{p'}(\Omega) \quad \int_\Omega v_n(x)z(x)dx \underset{n\to+\infty}{\longrightarrow} \int_\Omega v(x)z(x)dx.$$

The weak convergence in L^p allows us to model two different types of phenomena (they may occur simultaneously):

1. *Oscillations*. We describe the simplest situation of wild oscillations. Take $\Omega = (a, b)$ a bounded open interval of the real line, dx the Lebesgue mesure on Ω, and $v_n(x) = \sin(nx)$, $n \in \mathbf{N}$. Clearly v_n oscillates between -1 and $+1$ with period equal to $T_n = 2\pi/n$. When n goes to $+\infty$, $T_n \longrightarrow 0$ and simultaneously its frequency goes to $+\infty$.

Let us prove that $v_n \overset{\sigma(L^p,L^{p'})}{\longrightarrow} 0$ for any $1 \leq p < \infty$. Indeed, we can state a slightly more precise result: for any $z \in L^1(\Omega)$

$$\int_a^b z(x) \sin nx dx \longrightarrow 0 \quad \text{as } n \to +\infty.$$

(Note that for all $1 \leq p < +\infty$, $L^p(a, b) \subset L^1(a, b)$.)

Indeed, that is exactly the Riemann's theorem which states that the Fourier coefficients of an integrable function tend to zero as $n \to +\infty$. For convenience, we give the proof, which is a nice illustration of a density argument. We also emphasize that this result is a particular case of an ergodic theorem (see Section 13.2, Proposition 13.2.1, Remark 13.2.5, and Theorem 14.2.2).

For an arbitrary $z \in L^1$, it is difficult to compute or get information on the integral $\int_a^b z(x) \sin nx\, dx$. So, let us first consider the case where z belongs to some dense subspace Z of $L^1(\Omega)$, Z being chosen to make the computation easier.

For example, take $Z = \mathbf{C}_c^1(a, b)$, the subspace of $\mathbf{C}^1$ functions with compact support in (a, b). By integration by parts, for any $z \in \mathbf{C}_c^1(a, b)$,

$$\int_a^b z(x) \sin nx dx = \int_a^b z'(x) \frac{\cos nx}{n} dx,$$

which implies

$$\left|\int_a^b z(x)\sin nx dx\right| \leq \frac{1}{n}\int_a^b |z'(x)|dx \underset{n\to+\infty}{\longrightarrow} 0.$$

Another choice would consist of taking for Z the subspace of the step functions on (a, b). In that case a direct computation yields a similar result. So, for any z belonging to a dense subspace Z of $L^1(\Omega)$, we have that

$$\int_a^b z(x)\sin nx dx \underset{n\to+\infty}{\longrightarrow} 0.$$

We complete the proof by a density argument.

Take z an arbitrary element of $L^1(\Omega)$. By the density of Z in $L^1(\Omega)$, for any $\varepsilon > 0$, there exists some element $z_\varepsilon \in Z$ such that $\|z - z_\varepsilon\|_1 < \varepsilon$. Let us write

$$\int_a^b z(x)\sin nx\, dx = \int_a^b z_\varepsilon(x)\sin nx\, dx + \int_a^b (z(x) - z_\varepsilon(x))\sin nx\, dx.$$

Thus,

$$\begin{aligned}\left|\int_a^b z(x)\sin nx dx\right| &\leq \left|\int_a^b z_\varepsilon(x)\sin nx dx\right| + \int_a^b |z(x) - z_\varepsilon(x)|dx \\ &\leq \left|\int_a^b z_\varepsilon(x)\sin nx dx\right| + \varepsilon.\end{aligned}$$

Since $z_\varepsilon \in Z$, we obtain

$$\limsup_{n\to+\infty}\left|\int_a^b z(x)\sin nx dx\right| \leq \varepsilon \quad \forall\, \varepsilon > 0,$$

which implies $\lim_{n\to+\infty}\int_a^b z(x)\sin nx dx = 0$.

We now observe that the sequence $(v_n)_{n\in\mathbf{N}}$, $v_n(x) = \sin nx$ does not norm-converge in $V = L^p(a, b)$. Otherwise, it would be norm convergent to zero, but this is impossible since, for example, with $p = 2$,

$$\begin{aligned}\|v_n\|_2 = \int_a^b (\sin nx)^2 dx &= \int_a^b \frac{1}{2}(1 - \cos 2nx)dx \\ &= \frac{b-a}{2} - \frac{1}{4n}(\sin 2nb - \sin 2na) \\ &\underset{n\to+\infty}{\longrightarrow} \frac{b-a}{2}, \quad \text{which is different from zero!} \qquad \square\end{aligned}$$

2. *Concentration.* Take $\Omega = (0, 1)$ and $(v_n)_{n\in\mathbf{N}}$ a sequence of step functions which is described as follows:

let $A_n = \bigcup_{k=1,\ldots,n}\left[\frac{k}{n+1} - \frac{1}{2n^2}, \frac{k}{n+1} + \frac{1}{2n^2}\right]$ and take $v_n = \sqrt{n}$ on A_n and $v_n = 0$ elsewhere.

Let us examine the mode of convergence of the sequence $(v_n)_{n\in\mathbf{N}}$ in $V = L^2(0, 1)$. One can first observe that

(a) $\int_0^1 v_n^2(x)\,dx = n \cdot \frac{1}{n^2} \cdot n = 1$ for all $n \in \mathbf{N}$;
(b) the sequence $(v_n)_{n\in\mathbf{N}}$ converges to zero in measure, that is,

$$\forall \delta > 0 \quad \text{meas}\{x \in (0,1) : |v_n(x)| > \delta\} \longrightarrow 0.$$

In fact, one just needs to observe that $\{x \in (0,1) : |v_n(x)| > \delta\} = A_n$ and $\text{meas}(A_n) = n \cdot \frac{1}{n^2} = \frac{1}{n} \underset{n\to+\infty}{\longrightarrow} 0$.

Therefore, the sequence $(v_n)_{n\in\mathbf{N}}$ does not norm-converge in $L^2(0,1)$; otherwise it would converge to zero (recall that the norm convergence in L^2 implies the convergence in measure), which is impossible since $\|v_n\|_{L^2} = 1$.

Let us now prove that the sequence $(v_n)_{n\in\mathbf{N}}$ weakly converges to zero in $V = L^2(0,1)$. Indeed, by using the same density argument as in the previous oscillation example, we just need to prove that for any step function $z : (0,1) \longrightarrow \mathbf{R}$,

$$\int_0^1 v_n(x)z(x)dx \underset{n\to+\infty}{\longrightarrow} 0.$$

By linearity of the integral, we just need to compute for any $0 < a < b < 1$ the integral $\int_a^b v_n(x)dx$. Let us now observe that as n goes to $+\infty$,

$$\int_a^b v_n(x)dx \simeq n(b-a) \cdot \frac{1}{n^2} \cdot \sqrt{n} = \frac{b-a}{\sqrt{n}} \longrightarrow 0$$

($\simeq$ stands for equivalent). □

We stress the fact that in the concentration example, the weak convergence occurs simultaneously with the pointwise convergence. What happens in this situation is that the mass of $|v_n|^2$ is concentrated in a set of small Lebesgue measure. This is the concentration phenomenon. Let us notice, too, that the sequence $(v_n)_{n\in\mathbf{N}}$, in the above example norm-converges to zero in any $L^p(0,1)$, $1 \le p < 2$! To see this, just compute

$$\int_0^1 |v_n(x)|^p\,dx = n \cdot \frac{1}{n^2} \cdot n^{p/2} = n^{(p/2)-1} \longrightarrow 0$$

as $n \to +\infty$ as soon as $(p/2) - 1 < 0$, that is, $p < 2$.

Thus $p = 2$, in this situation, is a *critical exponent*, for which we pass from strong convergence of the sequence $(v_n)_{n\in\mathbf{N}}$ in L^p, $p < 2$, to weak convergence in L^2. As we will see, weak convergences related to concentration effect, often occurs in situations where some critical exponent is involved (like the critical Sobolev exponent).

The two previous examples illustrate the utility of the density arguments when proving weak convergence. Let us state it in an abstract setting.

Proposition 2.4.8. *Let V be a normed linear space and Z a dense subset of V^*. For any bounded sequence $(v_n)_{n\in\mathbf{N}}$ in V, the following assertions are equivalent:*

(i) $v_n \overset{\sigma(V,V^*)}{\longrightarrow} v$.

(ii) $\forall z^* \in Z, \langle z^*, v_n\rangle \longrightarrow \langle z^*, v\rangle$ *as* $n \to +\infty$.

PROOF. Clearly (i) $\Longrightarrow$ (ii). So let us assume (ii) and prove that for any $v^* \in V^*$, we have

$$\lim_{n\to+\infty} \langle v^*, v_n\rangle = \langle v^*, v\rangle.$$

By the density of Z in V^* for any $\varepsilon > 0$, there exists some element $z_\varepsilon^* \in Z$ such that $\|v^* - z_\varepsilon^*\|_* < \varepsilon$. Let us write

$$\langle v^*, v_n - v\rangle = \langle z_\varepsilon^*, v_n - v\rangle + \langle v^* - z_\varepsilon^*, v_n - v\rangle,$$

which by the triangle inequality and the definition of the dual norm $\|\cdot\|_*$ yields

$$|\langle v^*, v_n - v\rangle| \le |\langle z_\varepsilon^*, v_n - v\rangle + \|v^* - z_\varepsilon^*\|_* \cdot \|v_n - v\|.$$

Using the assumption that the sequence $(v_n)_{n\in\mathbf{N}}$ is bounded in V and that $\|v^* - z_\varepsilon^*\|_* < \varepsilon$, we obtain that for some constant $C \in \mathbf{R}^+$,

$$|\langle v^*, v_n - v\rangle| \le |\langle z_\varepsilon^*, v_n - v\rangle| + C\varepsilon.$$

Now let n tend to $+\infty$, and use assumption (ii) together with $z_\varepsilon^* \in Z$ to get

$$\limsup_{n\to+\infty} |\langle v^*, v_n - v\rangle| \le C\varepsilon.$$

This inequality being true for any $\varepsilon > 0$, we finally infer

$$\forall v^* \in V^* \quad \lim_{n\to+\infty} \langle v^*, v_n - v\rangle = 0,$$

that is, $v = \sigma(V, V^*) \lim_{n\to+\infty} v_n$. □

2.4.3 Weak convergence and geometry of uniformly convex spaces

In this section we pay attention to a particular class of Banach spaces, namely, the uniformly convex Banach spaces, where we will be able to extend the result of Proposition 2.4.7, that is, weak convergence and convergence of the norms imply the strong convergence.

Definition 2.4.2. *A Banach space $(V, \|\cdot\|)$ is said to be uniformly convex if for any sequences $(u_n)_{n\in\mathbf{N}}$, $(v_n)_{n\in\mathbf{N}}$ in V with $\|u_n\| = \|v_n\| = 1$ for all $n \in \mathbf{N}$, the following implication holds:*

$$\left\|\frac{u_n + v_n}{2}\right\| \underset{n\to+\infty}{\longrightarrow} 1 \Longrightarrow \|u_n - v_n\| \longrightarrow 0.$$

This result reflects a geometrical property of the unit ball which has to be well rotund. Note that this definition is not stable when replacing a norm by an equivalent one. As an elementary example, one can observe that $V = \mathbf{R}^N$ equipped with the norm $\|x\|_2 = \left(\sum_{i=1}^n x_i^2\right)^{1/2}$ is uniformly convex, whereas the norms $\|x\|_1 = \sum_{i=1}^N |x_i|$ and $\|x\|_\infty = \max_{1\le i\le N} |x_i|$ are not uniformly convex.

The uniform convexity of the norm expresses that if u and v are on the unit sphere, the fact that $\frac{u+v}{2}$ is close to the sphere forces u and v to be close one to the other.

Proposition 2.4.9. (a) *The Hilbert spaces are uniformly convex.*
(b) *The L^p spaces, $1 < p < \infty$ are uniformly convex.*

PROOF. (a) The uniform convexity of Hilbert spaces is a direct consequence of the parallelogram equality:

$$\forall u, v \in V, \quad \|u+v\|^2 + \|u-v\|^2 = 2(\|u\|^2 + \|v\|^2).$$

Notice that this property characterizes Hilbert spaces among general Banach spaces. So, let us take u_n, v_n in V such that $\|u_n\| = \|v_n\| = 1$. We have

$$\|u_n - v_n\|^2 = 4\left(1 - \left\|\frac{u_n+v_n}{2}\right\|^2\right).$$

So $\|\frac{u_n+v_n}{2}\| \longrightarrow 1$ forces $\|u_n - v_n\|$ to converge to zero as $n \to +\infty$.

(b) The same type of argument works in L^p spaces, $1 < p < \infty$ when replacing the parallelogram identity by the so-called Clarkson's inequalities; if $\|\cdot\|$ is the L^p norm, one has to distinguish two cases:

$$\left\|\frac{u-v}{2}\right\|^p + \left\|\frac{u+v}{2}\right\|^p \le \frac{1}{2}(\|u\|^p + \|v\|^p) \qquad \text{if } 2 \le p < \infty,$$

$$\left\|\frac{u-v}{2}\right\|^{p'} + \left\|\frac{u+v}{2}\right\|^{p'} \le \left[\frac{1}{2}(\|u\|^{p'} + \|v\|^{p'})\right]^{p-1} \qquad \text{if } 1 < p \le 2$$

with $\frac{1}{p} + \frac{1}{p'} = 1$. □

The following result justifies the introduction of the notion of uniform convexity in this section devoted to the weak convergence.

Proposition 2.4.10. *Let V be a uniformly convex Banach space. Then for any sequence $(v_n)_{n\in\mathbf{N}}$ in V the following implication holds:*

$$v_n \overset{\sigma(V,V^*)}{\longrightarrow} v \text{ and } \|v_n\| \longrightarrow \|v\| \Longrightarrow v_n \overset{\|\cdot\|}{\longrightarrow} v.$$

PROOF. Let us reduce ourselves to the case $\|v_n\| = \|v\| = 1$. To that end, let us consider $w_n := \frac{v_n}{\|v_n\|}$ and $w := \frac{v}{\|v\|}$ (the case $v = 0$ is obvious, so, we can assume $v \neq 0$).

One can notice that $\|w_n\| = 1$ and $w_n \overset{\sigma(V,V^*)}{\longrightarrow} w$: indeed, for any $v^* \in V^*$,

$$\begin{aligned}\langle v^*, w_n - w\rangle &= \left\langle v^*, \frac{v_n}{\|v_n\|} - \frac{v}{\|v\|}\right\rangle \\ &= \frac{1}{\|v_n\|}\langle v^*, v_n - v\rangle + \left(\frac{1}{\|v_n\|} - \frac{1}{\|v\|}\right)\langle v^*, v\rangle,\end{aligned}$$

which goes to zero as $n \to +\infty$, since $v_n \overset{\sigma(V,V^*)}{\longrightarrow} v$ and $\|v_n\| \longrightarrow \|v\| \neq 0$.

Let us show that $\left\|\frac{w_n+w}{2}\right\| \to 1$ as $n \to +\infty$. Since $\frac{w_n+w}{2} \overset{\sigma(V,V^*)}{\longrightarrow} w$, by the lower semicontinuity of the norm for the weak topology (Proposition 2.4.6(iii)),

$$1 = \|w\| \le \liminf_n \left\|\frac{w_n+w}{2}\right\| \le \limsup_n \left\|\frac{w_n+w}{2}\right\| \le 1.$$

The last inequality follows from the triangle inequality and the fact that $\|w_n\| = \|w\| = 1$. So we have $\|w_n\| = \|w\| = 1$ and $\left\|\frac{w_n+w}{2}\right\| \longrightarrow 1$. It follows from the uniform convexity property that $w_n \xrightarrow{\|\cdot\|} w$. Using once more that $\|v_n\| \longrightarrow \|v\|$ we derive that $v_n = \|v_n\| w_n$ norm converges to $v = \|v\| w$. □

Remark 2.4.3. It is a quite useful method, when proving that a sequence $(u_n)_{n\in\mathbf{N}}$ is norm converging in a Hilbert space, or more generally in a uniformly convex Banach space, to prove first that the weak convergence holds and then to prove that the norms converge, too. For example, when minimizing the norm over a closed convex bounded subset of a uniformly convex Banach space, one automatically obtains that any minimizing sequence is norm convergent.

The property for a Banach space to verify for any sequence $(v_n)_{n\in\mathbf{N}}$ in V the implication $v_n \xrightarrow{\sigma(V,V^*)} v$ and $\|v_n\| \longrightarrow \|v\| \Longrightarrow v_n \xrightarrow{\|\cdot\|} v$ is often called the *Kadek property*.

Remark 2.4.4. Let us observe that the Kadek property fails to be true in general Banach spaces. For example, it is false in the space $L^1(\Omega)$, $\Omega \subset \mathbf{R}^N$ equipped with the Lebesgue measure: indeed, take

$$\Omega = (0, \pi), \quad v_n(x) = 1 + \sin nx, \quad n = 1, 2, \ldots .$$

Then $v_n \xrightarrow{\sigma(L^1,L^\infty)} v \equiv 1$, $\|v_n\|_1 = \int_0^\pi v_n(x)dx = \pi + \frac{1}{n}(1 - \cos n\pi)$, so that $\|v_n\|_1 \underset{n\to\infty}{\longrightarrow} \pi = \|v\|_1$. But $\|v_n - v\|_1 = \int_0^\pi |\sin nx|dx$ does not converge to zero as $n \to +\infty$. Indeed,

$$\int_0^\pi |\sin nx|dx = n \cdot \int_0^{\pi/n} \sin nx dx = 2.$$

2.4.4 Weak compactness theorems in reflexive Banach spaces

We have already observed that L^p spaces enjoy quite different properties with respect to the weak convergence, depending on the two situations $1 < p < \infty$ or $p = 1$, $p = \infty$. One can distinguish them by introducing the concept of uniform convexity, or local uniform convexity of the space as done in Section 2.4.3. But this is a geometrical concept related to the choice of the norm, and when dealing with topological concepts like compactness, one is naturally led to consider notions which are of topological nature (i.e., invariant by the choice of an equivalent norm). This is where the notion of reflexive Banach space plays a fundamental role. Let us first recall its definition.

Let V be a Banach space, V^* its topological dual, and V^{**} its topological bidual equipped, respectively, with the norms

$$\|\cdot\|_V = \|\cdot\|, \quad \|v^*\|_* = \sup_{\|v\|\le 1} |\langle v^*, v\rangle|, \quad \|v^{**}\|_{**} = \sup_{\|v^*\|_*\le 1} |\langle v^{**}, v^*\rangle|.$$

There exists a canonical embedding of V into V^{**} denoted by $J : V \longrightarrow V^{**}$ which is defined as follows:

$$\forall v \in V, \forall v^* \in V^* \quad \langle Jv, v^*\rangle_{(V^{**},V^*)} = \langle v^*, v\rangle_{(V^*,V)}.$$

Let us comment on this definition. For any $v \in V$, the mapping

$$v^* \in V^* \longmapsto \langle v^*, v\rangle_{(V^*,V)} \in \mathbf{R}$$

is linear and continuous on V^*, so it defines uniquely an element of V^{**} which is denoted by Jv. Let us observe that

$$|\langle v^*, v\rangle| \leq \|v^*\|_* \cdot \|v\| \quad \forall\, v \in V,$$

so that $\|Jv\|_{**} \leq \|v\|$. Indeed, as a consequence of the Hahn–Banach theorem we have

$$\|Jv\|_{**} = \sup_{\|v^*\|_* \leq 1} |\langle v^*, v\rangle| = \|v\|,$$

so that J is a *linear isometry* from V into V^{**}. As a consequence, J is an embedding of V into V^{**}.

Definition 2.4.3. *A Banach space V is said to be reflexive if $J(V) = V^{**}$. When V is reflexive one can identify V and V^{**} with the help of J.*

Remark 2.4.5. J is a linear isometry. Thus it preserves the linear and the normed structures and allows us to identify V and V^{**} when J is onto, that is, in the case of reflexive Banach spaces. One has to pay attention to the following fact: the definition of reflexive Banach spaces says that the map J realizes an isometrical isomorphism between V and V^{**}. It is essential to use J in the definition since one can exhibit a nonreflexive Banach space V such that there exists an isometry from V onto V^{**}!

Proposition 2.4.11. *Let V be a uniformly convex Banach space. Then V is reflexive.*

For a proof of this result, see, for instance, [225], [90]. When considering L^p spaces, this result is in accordance with the results concerning the dual of L^p spaces. When $1 < p < \infty$, L^p is uniformly convex, $(L^p)^* = L^{p'}$, where $\frac{1}{p} + \frac{1}{p'} = 1$, so that $(L^p)^{**} = (L^{p'})^* = L^p$ (equalities above mean isometric isomorphisms).

Remark 2.4.6. Note that there exist reflexive Banach spaces which do not admit an equivalent norm which makes the space uniformly convex. However, one can always renorm a reflexive Banach space with a norm (equivalent) which is locally uniformly convex both with its dual norm. With this renorming, it will satisfy the Kadek property (as well as its dual); see Section 2.4.3.

The importance of reflexive Banach spaces is justified by the following theorem.

Theorem 2.4.2. (a) *In a reflexive Banach space $(V, \|\cdot\|)$ the closed unit ball*

$$B = \{v \in V : \|v\|_V \leq 1\}$$

is compact for the topology $\sigma(V, V^*)$. As a consequence, the bounded subsets of V are relatively compact for the topology $\sigma(V, V^*)$.

(b) *The above property characterizes the reflexive Banach spaces: a Banach space is reflexive iff the closed unit ball is compact for the topology* $\sigma(V, V^*)$.

PROOF. The proof is a direct consequence of the Banach–Alaoglu–Bourbaki theorem, Theorem 2.4.7. It makes use of the weak* topology on the dual of a Banach space. □

Let us now state a theorem from Eberlein and Smulian which states that from every bounded sequence in a reflexive Banach space one can extract a sequence which converges for the topology $\sigma(V, V^*)$. This is an important and quite suprising result, since the weak topology is not metrizable. One does not expect to have such a sequential compactness result!

Theorem 2.4.3. *Let* V *be a reflexive Banach space. Then, from each bounded sequence* $(u_n)_{n\in\mathbf{N}}$ *in* V*, one can extract a subsequence* $(u_{n_k})_{k\in\mathbf{N}}$ *which converges for the topology* $\sigma(V, V^*)$.

PROOF. (a) Let us first assume that V^* is separable, that is, there exists a countable set $D = (v_k^*)_{k\in\mathbf{N}}$ which is dense in V^*. The proof relies on a diagonalization argument. First, let us notice that for all $v^* \in V^*$,

$$|\langle v^*, u_n\rangle| \leq \|v^*\|_* \cdot \|u_n\| \leq C\|v^*\|_*,$$

where $C = \sup_{n\in\mathbf{N}} \|u_n\| < +\infty$. So the sequence $\{\langle v^*, u_n\rangle : n \in \mathbf{N}\}$ is bounded in $\mathbf{R}$. For each $v^* \in V^*$, one can extract from the sequence $\{\langle v^*, u_n\rangle : n \in \mathbf{N}\}$ a convergent subsequence. The difficult point is that without any further argument, the so extracted sequence depends on v^*. This is where the separability of V^* and the diagonalization argument take place.

Let us start with $v_1^* \in D$ (D dense countable subset of V^*) and extract a convergent subsequence $\{\langle v_1^*, u_{\sigma_1(n)}\rangle : n \in \mathbf{N}\}$, where $\sigma_1 : \mathbf{N} \to \mathbf{N}$ is a strictly increasing mapping. In a similar way, the sequence $\{\langle v_2^*, u_{\sigma_1(n)}\rangle : n \in \mathbf{N}\}$ is bounded in $\mathbf{R}$, so there exists a convergent subsequence $\{\langle v_2^*, u_{\sigma_1\circ\sigma_2(n)}\rangle : n \in \mathbf{N}\}$. Let us iterate this argument by induction. We can so construct for each $n \in \mathbf{N}$ an increasing mapping $\sigma_n : \mathbf{N} \longrightarrow \mathbf{N}$ such that the sequence

$$\{\langle v_n^*, u_{\sigma_1\circ\sigma_2\circ\ldots\circ\sigma_n(k)}\rangle : k \in \mathbf{N}\}$$

is convergent in $\mathbf{R}$. Note that it is important to have the composition of the mapping σ_i in the precise order $\sigma_1 \circ \ldots \circ \sigma_n$ to have this subsequence extracted from all the previous ones, $\sigma_1, \sigma_1 \circ \sigma_2, \ldots, \sigma_1 \circ \ldots \circ \sigma_{n-1}$.

Now the diagonalization argument consists of taking $\tau : \mathbf{N} \longrightarrow \mathbf{N}$ defined by

$$\tau(n) = (\sigma_1 \circ \sigma_2 \circ \ldots \circ \sigma_n)(n).$$

In other words, $\tau(n)$ is the element of rank n of the subsequence which has been extracted at step n. Clearly τ is strictly increasing.

It is important to notice that for each $n \in \mathbf{N}$, the sequence $(u_{\tau(k)})_{k\geq n}$ is extracted from the sequence $(u_{\sigma_1\circ\ldots\circ\sigma_n(k)})_{k\geq n}$: indeed, for $k \geq n$,

$$u_{\tau(k)} = u_{\sigma_1\circ\sigma_2\circ\ldots\circ\sigma_k(k)} = u_{(\sigma_1\circ\ldots\circ\sigma_n)(p)},$$

where $p = (\sigma_{n+1} \circ \cdots \circ \sigma_k)(k)$. Since $k \geq n$ and σ_i are strictly increasing, we also have $p \geq n$.

It follows that for each $k \in \mathbf{N}$, the sequence $(\langle v_k^*, u_{\tau(n)} \rangle : n \in \mathbf{N})$, as a subsequence of a convergent sequence, is convergent.

By the density of D in V^*, the result is easily extented to an arbitrary element v^* of V^*: for each $\varepsilon > 0$ take $v_{k_\varepsilon}^* \in \mathcal{D}$ such that $\|v^* - v_{k_\varepsilon}^*\| < \varepsilon$. Then for each $n, m \in \mathbf{N}$,

$$\begin{aligned} |\langle v^*, u_{\tau(n)} \rangle - \langle v^*, u_{\tau(m)} \rangle| &\leq |\langle v_{k_\varepsilon}^*, u(\tau_n) - u(\tau_m) \rangle| \\ &\quad + |\langle v^* - v_{k_\varepsilon}^*, u(\tau_n) - u(\tau_m) \rangle| \\ &\leq 2C\varepsilon + |\langle v_{k_\varepsilon}^*, u(\tau_n) - u(\tau_m) \rangle|. \end{aligned}$$

Hence,

$$\limsup_{n,m \to \infty} |\langle v^*, u_{\tau(n)} - u_{\tau(m)} \rangle| \leq 2C\varepsilon.$$

This being true for any $\varepsilon > 0$, we infer that the sequence $\{\langle v^*, u_{\tau(n)} \rangle : n \in \mathbf{N}\}$ satisfies the Cauchy criteria and is thus convergent in $\mathbf{R}$.

Let us denote for all $v^* \in V^*$, $L(v^*) := \lim_{n \to +\infty} \langle v^*, u_{\tau(n)} \rangle$. Clearly, L is a linear continuous form on V^*; note that by passing to the limit in the inequality $|\langle v^*, u_n \rangle| \leq C\|v^*\|_*$, we also have $|L(v^*)| \leq C\|v^*\|_*$. Hence $L \in V^{**}$.

Let us stress the fact that up to this point, we have not used the reflexivity hypothesis. We now use that V is reflexive to assert that $L = J(u)$ for some $u \in V$, where J is the canonical embedding of V into V^{**}. So for all $v^* \in V^*$ we have

$$\lim_{n \to +\infty} \langle v^*, u_{\tau(n)} \rangle = \langle v^*, u \rangle,$$

that is, $u = \sigma(V, V^*) \lim_{n \to \infty} u_{\tau(n)}$.

(b) Now take V a reflexive Banach space and do not make any separability assumptions. Take E the subspace of V generated by the $(u_n)_{n \in \mathbf{N}}$ and define $W = \bar{E}$, the closure of E in $(V, \|\cdot\|)$.

It is easy to verify that W is reflexive and separable. This in turn implies that W^* is reflexive and separable (see [90, Corollary III.24]). We are now in the situation studied in the first part of the proof. One can extract a subsequence $(u_{\tau(n)})_{n \in \mathbf{N}}$ which converges in W for the topology $\sigma(W, W^*)$. Since $V^* \subset W^*$ (by restriction of the linear continuous forms on V to W) we derive that $(u_{\tau(n)})_{n \in \mathbf{N}}$ is convergent for the topology $\sigma(V, V^*)$. □

2.4.5 The Dunford–Pettis weak compactness theorem in $L^1(\Omega)$

When $1 < p < \infty$, the L^p spaces are reflexive Banach spaces, and it follows from Theorem 2.4.3 that the relatively compact subsets of L^p for the topology $\sigma(L^p, L^{p'})$ are exactly the bounded subsets of L^p. The situation in the case $p = 1$ is very different (L^1 is not a reflexive Banach space), and the comprehension of the weak convergence properties of bounded sequences in L^1 is a subject of great importance, quite involved.

Let us first examine the following example. Take $\Omega = (-1, 1)$ equipped with the Lebesgue measure and take

$$v_n(x) = \begin{cases} n \text{ if } -\frac{1}{2n} \leq x \leq +\frac{1}{2n}, \\ 0 \text{ elsewhere.} \end{cases}$$

Clearly, the sequence $(v_n)_{n\in\mathbf{N}}$ satisfies

$$\begin{cases} v_n \geq 0, \displaystyle\int_\Omega v_n(x)dx = 1, \\ v_n(x) \longrightarrow 0 \text{ for almost every } x \in \Omega, \\ \displaystyle\int_\Omega v_n(x)z(x)dx \underset{n\to+\infty}{\longrightarrow} z(0) \text{ for any } z \in \mathbf{C}(\Omega). \end{cases}$$

We can observe that the sequence $(v_n)_{n\in\mathbf{N}}$ is bounded in $L^1(\Omega)$, but one cannot extract a weakly convergent subsequence in the sense $\sigma(L^1, L^\infty)$. By contrast, we will see in the next section that one can interpret the convergence of the sequence $(v_n)_{n\in\mathbf{N}}$ with the help of a weak topology of dual $\sigma(V^*, V)$, for example, $\sigma(\mathcal{M}_b(\Omega), \mathbf{C}_0(\Omega))$.

For the moment, we just retain from this example that to obtain $\sigma(L^1, L^\infty)$ compactness of a sequence of functions, it is not sufficient to assume that the sequence is bounded in L^1. This is where the notion of uniform integrability plays a central role.

Definition 2.4.4. *Let $(\Omega, \mathcal{A}, \mu)$ be a measured space with μ a positive and finite measure $(\mu(\Omega) < +\infty)$. Let $\mathbf{K}$ be a subset of $L^1(\Omega, \mathcal{A}, \mu)$. We say that $\mathbf{K}$ is uniformly integrable if* (a) *and* (b) *hold:*

(a) $\mathbf{K}$ *is bounded in $L^1(\Omega, \mathcal{A}, \mu)$;*

(b) *for every $\varepsilon > 0$ there exists some $\delta(\varepsilon) > 0$ such that*

$$A \in \mathcal{A},\ \mu(A) < \delta(\varepsilon) \Longrightarrow \sup_{v\in\mathbf{K}} \int_A |v(x)|d\mu(x) < \varepsilon.$$

A comprehensive characterization of this property is given by the De La Vallée–Poussin theorem.

Theorem 2.4.4. *Let $(\Omega, \mathcal{A}, \mu)$ be a measure space with μ a positive and finite measure and $\mathbf{K}$ a subset of $L^1(\Omega, \mathcal{A}, \mu)$. The following properties are equivalent:*

(i) $\mathbf{K}$ *is uniformly integrable;*

(ii) *there exists a function $\theta : [0, +\infty[\to [0, +\infty[$ (θ can be taken convex and increasing) such that $\lim_{s\to+\infty} \frac{\theta(s)}{s} = +\infty$ and*

$$\sup_{v\in\mathbf{K}} \int_\Omega \theta(|v(x)|)d\mu(x) < +\infty.$$

Proof. The implication (ii) $\Longrightarrow$ (i) is important for applications. Let us prove it.

First, one can observe that since θ has a superlinear growth, for each $M \in \mathbf{R}^+$ there exists some $C(M) \in \mathbf{R}^+$ such that

$$\forall s \in \mathbf{R}^+ \quad 0 \leq s \leq \frac{1}{M}\theta(s) + C(M).$$

Let us fix $M_0 > 0$. We have for each $v \in \mathbf{K}$

$$\int_\Omega |v(x)|d\mu(x) \leq \frac{1}{M_0} \int_\Omega \theta(|v(x)|d\mu(x) + C(M_0)\mu(\Omega)$$

and hence $\sup_{v\in\mathbf{K}} \int |v|d\mu < +\infty$, which proves (a) of Definition 2.4.4.

Let us now prove (b). Fix $\varepsilon > 0$; for any $v \in \mathbf{K}$, $A \in \mathcal{A}$, and $M \in \mathbf{R}^+$,

$$\begin{aligned}\int_A |v(x)| d\mu(x) &\leq \frac{1}{M} \int_A \theta(|v(x)|) d\mu(x) + C(M)\mu(A) \\ &\leq \frac{1}{M} \sup_{v \in \mathbf{K}} \int_\Omega \theta(|v|) d\mu + C(M)\mu(A).\end{aligned}$$

Take

$$M(\varepsilon) := \frac{2}{\varepsilon} \sup_{v \in \mathbf{K}} \int_\Omega \theta(|v|)\, d\mu, \qquad \delta(\varepsilon) := \frac{\varepsilon}{2C(M(\varepsilon))}.$$

Then if $\mu(A) \leq \delta(\varepsilon)$ we have

$$\sup_{v \in \mathbf{K}} \int_A |v(x)| d\mu(x) \leq \frac{\varepsilon}{2} + \frac{\varepsilon}{2} = \varepsilon,$$

which proves (b).

For the proof of the reverse implication (i) $\Longrightarrow$ (ii), see, for instance, [125, Theorem 22] or [102, Theorem 2.12]. □

We are now ready to state the Dunford–Pettis theorem, which gives a characterization of the weak compactness property in L^1.

Theorem 2.4.5 (Dunford–Pettis theorem). *Let $(\Omega, \mathcal{A}, \mu)$ be a measure space with μ a positive and finite measure. Let $\mathbf{K}$ be a subset of $L^1(\Omega, \mathcal{A}, \mu)$. The following properties are equivalent:*

(i) $\mathbf{K}$ *is relatively compact for the weak topology $\sigma(L^1, L^\infty)$.*

(ii) $\mathbf{K}$ *is uniformly integrable.*

(iii) *From each sequence $(v_n)_{n \in \mathbf{N}}$ contained in $\mathbf{K}$, one can extract a subsequence converging for the topology $\sigma(L^1, L^\infty)$.*

Proof. See [125, Theorem 25], [127, Theorem IV.8.9, Corollary IV.8.11], and [102]. When Ω is an open subset of $\mathbf{R}^N$ and μ the Lebesgue measure on Ω, one can find a proof of implication (ii) $\Longrightarrow$ (iii) in Proposition 4.3.7 by using the notion of Young measures. □

Remark 2.4.7. As an illustration of the Dunford–Pettis theorem, let us consider $(v_n)_{n \in \mathbf{N}}$ a sequence in $L^1(\Omega, \mathcal{A}, \mu)$, $\mu(\Omega) < +\infty$ such that $\sup_{n \in \mathbf{N}} \int_\Omega |v_n| \ln |v_n| d\mu < +\infty$. We claim that the sequence (v_n) is $\sigma(L^1, L^\infty)$ relatively compact. To obtain this result, just use the De La Vallée–Poussin theorem with $\theta(r) = r \ln r$ (which is superlinear) and then use the Dunford–Pettis theorem.

2.4.6 The weak* topology $\sigma(V^*, V)$

Let $(V, \|\cdot\|)$ be a normed linear space with topological dual V^*. On V^* we have already defined two topologies:

- The norm topology associated to the dual norm

$$\|v^*\|_* = \sup\{|\langle v^*, v\rangle| : v \in V, \|v\| \leq 1\},$$

  which makes V^* a Banach space;

- The weak topology $\sigma(V^*, V^{**})$, where V^{**} is the topological bidual of V. But this topology is often difficult to handle because the space V^{**} may have a rather involved structure (think, for example, $V = L^1$, $V^* = L^\infty$, and $V^{**} = (L^\infty)^*$), and moreover it may be too strong to enjoy desirable compactness properties.

The idea is to introduce a topology weaker than $\sigma(V^*, V^{**})$ by considering a weak topology on V^* induced not by all the linear continuous forms on V^* but only by a subfamily. At this point, there is a natural candidate which consists in taking $J(V) \subset V^{**}$ where $J : V \longrightarrow V^{**}$ is the canonical embedding from V into its bidual V^{**}. Recall that (see Section 2.4.4)

$$\forall v \in V, \forall v^* \in V^* \quad \langle J(v), v^*\rangle_{(V^{**},V^*)} := \langle v^*, v\rangle_{(V^*,V)}.$$

Definition 2.4.5. *Let V be a normed space with topological dual V^*. The weak* topology $\sigma(V^*, V)$ on V^* is the weakest topology on V^* making continuous all the mappings $(J(v))_{v\in V}$, where J is the canonical embedding from V into V^{**}:*

$$\begin{aligned} J(v) : V^* &\longrightarrow \mathbf{R}, \\ v^* &\longmapsto \langle J(v), v^*\rangle_{(V^{**},V^*)} = \langle v^*, v\rangle. \end{aligned}$$

Let us collect in the following proposition some first elementary properties of the topology $\sigma(V^*, V)$.

Proposition 2.4.12. *Let V be a normed space and $\sigma(V^*, V)$ the weak* topology on the topological dual V^*. Then*

(i) *A local base of neighborhoods of $v_0^* \in V^*$ for the topology $\sigma(V^*, V)$ consists of all sets of the form*

$$N(v_0^*) = \{v^* \in V^* : |\langle v^* - v_0^*, v_i\rangle| < \varepsilon \ \ \forall i \in I\},$$

where I is a finite index set, $v_i \in V$ for each $i \in I$, and $\varepsilon > 0$.

(ii) *For any sequence $(v_n^*)_{n\in\mathbf{N}}$ in V^*, we have*

(a) $v_n^* \overset{\sigma(V^*,V)}{\longrightarrow} v^* \Longleftrightarrow \forall\, v \in V \quad \langle v_n^*, v\rangle \longrightarrow \langle v^*, v\rangle$;

(b) $v_n^* \overset{\|\cdot\|}{\longrightarrow} v^* \Longrightarrow v_n^* \overset{\sigma(V^*,V)}{\longrightarrow} v^*$.

Assume now that V is a Banach space. Then

(c) $v_n^* \overset{\sigma(V^*,V)}{\longrightarrow} v^* \Longrightarrow$ *the sequence (v_n^*) is bounded and*

$$\|v^*\|_* \leq \liminf_n \|v_n^*\|_*;$$

(d) $v_n^* \overset{\sigma(V^*,V)}{\longrightarrow} v^*$ *and* $v_n \overset{\|\cdot\|}{\longrightarrow} v \Longrightarrow \langle v_n^*, v_n\rangle \longrightarrow \langle v^*, v\rangle$.

PROOF. (i) and (ii)(a) are direct consequences of the general properties of topologies induced by functions; see, respectively, Theorem 2.4.1 and Proposition 2.4.3. (ii)(b), (ii)(c), (ii)(d) are obtained in a similar way as in the proof of Proposition 2.4.6. Just notice that to apply the uniform boundedness theorem, one needs to assume that V is a Banach space. (In Proposition 2.4.6 one works on V^*, which is always a Banach space!) □

An important example: $\mathbf{V = C_0(\Omega)}$, $\mathbf{V^* = \mathcal{M}_b(\Omega)}$. Let Ω be a locally compact topological space which is σ-compact (i.e., Ω can be written as $\Omega = \bigcup_{n\in\mathbf{N}} K_n$ with K_n compact). For example, we may take $\Omega = \mathbf{R}^N$, or Ω an open subset of $\mathbf{R}^N$, or Ω an arbitrary topological compact set.

Take $V = \mathbf{C}_0(\Omega)$ the linear space of real continuous functions on Ω which tend to zero at infinity; more precisely, we say that a continuous function u is in $\mathbf{C}_0(\Omega)$ if for every $\varepsilon > 0$ there exists a compact set K_ε such that $|u(x)| < \varepsilon$ on $\Omega \setminus K_\varepsilon$. Notice that $\mathbf{C}_0(\Omega)$ reduces to $\mathbf{C}(\Omega)$ when Ω is compact. We may endow the space $\mathbf{C}_0(\Omega)$ with the norm

$$\|v\|_V = \sup_{x\in\Omega} |v(x)|.$$

Then V is a Banach space whose topological dual V^* can be described thanks to the celebrated result below.

Theorem 2.4.6 (Riesz–Alexandroff representation theorem). *The topological dual of $\mathbf{C}_0(\Omega)$ can be isometrically identified with the space of bounded Borel measures. More precisely, to each bounded linear functional Φ on $\mathbf{C}_0(\Omega)$ there is a unique Borel measure μ on Ω such that for all $f \in \mathbf{C}_0(\Omega)$,*

$$\Phi(f) = \int_\Omega f(x)d\mu(x).$$

Moreover, $\|\phi\| = |\mu|(\Omega)$.

For a proof of this theorem, see, for instance, [206, Theorem 6.19] or [94, Theorem 1.4.22]. Notice that this theorem holds true also when Ω is not supposed to be σ-compact, but in that case one has to consider measures μ which are regular. We recall that a Borel measure $\mu \geq 0$ is said to be regular if

$$\begin{cases} \forall B \in \mathcal{B}(\Omega) & \mu(B) = \inf\{\mu(V) : V \supset B,\ V \text{ open}\}, \\ \forall B \in \mathcal{B}(\Omega) & \mu(B) = \sup\{\mu(K) : K \subset B,\ K \text{ compact}\}, \end{cases}$$

and a signed measure μ is said to be regular if its *total variation measure* $|\mu|$ is regular. When Ω is σ-compact, this property is automatically satisfied by Borel measures which are bounded. Note also that since $\mathbf{C}_c(\Omega)$ is dense in $\mathbf{C}_0(\Omega)$, these two spaces have the same topological dual. We prefer to consider $\mathbf{C}_0(\Omega)$ in this construction because it is a Banach space for the sup norm.

Thus a bounded Borel measure μ can be as well considered as a σ-additive set function on the Borel σ-algebra (that's the probabilistic approach) or as a continuous linear form on $\mathbf{C}_0(\Omega)$ or $\mathbf{C}_c(\Omega)$ (that's the functional analysis approach). Given a sequence $(\mu_n)_{n\in\mathbf{N}}$ of

bounded Borel measures, we can consider these measures as elements of the topological dual space $V^* = \mathcal{M}_b(\Omega)$ of $V = \mathbf{C}_0(\Omega)$. This leads to the following definition.

Definition 2.4.6.

(i) *A sequence $(\mu_n)_{n\in\mathbf{N}} \subset \mathcal{M}_b(\Omega)$ converges weakly to $\mu \in \mathcal{M}_b(\Omega)$, and we write*

$$\mu_n \longrightarrow \mu \text{ in } \mathcal{M}_b(\Omega)$$

provided

$$\int_\Omega \varphi \, d\mu_n \longrightarrow \int_\Omega \varphi \, d\mu \text{ as } n \to \infty$$

for each $\varphi \in \mathbf{C}_c(\Omega)$.

(ii) *A sequence $(\mu_n) \subset \mathcal{M}_b(\Omega)$ converges $\sigma(\mathcal{M}_b, \mathbf{C}_0)$ to $\mu \in \mathcal{M}_b(\Omega)$, and we write*

$$\mu_n \overset{\sigma(\mathcal{M}_b,\mathbf{C}_0)}{\longrightarrow} \mu$$

provided

$$\int_\Omega \varphi \, d\mu_n \longrightarrow \int_\Omega \varphi \, d\mu$$

for each $\varphi \in \mathbf{C}_0(\Omega)$.

The relation between these two close concepts is given by the following result.

Proposition 2.4.13. *Given $(\mu_n)_{n\in\mathbf{N}} \subset \mathcal{M}_b(\Omega)$, $\mu \in \mathcal{M}_b(\Omega)$ one has the equivalence*

$$\mu_n \overset{\sigma(\mathcal{M}_b,\mathbf{C}_0)}{\longrightarrow} \mu \Longleftrightarrow \mu_n \longrightarrow \mu \text{ and } \sup_{n\in\mathbf{N}} |\mu_n|(\Omega) < +\infty.$$

Proof. Since $\mathbf{C}_c(\Omega) \subset \mathbf{C}_0(\Omega)$ the implication $\mu_n \overset{\sigma(\mathcal{M}_b,\mathbf{C}_0)}{\longrightarrow} \mu \Longrightarrow \mu_n \longrightarrow \mu$ is clear. Moreover, since $\mathcal{M}_b(\Omega) = V^*$ with $V = \mathbf{C}_0(\Omega)$ which is a Banach space, the uniform boundedness theorem implies (see Proposition 2.4.12(ii)(c))

$$\mu_n \overset{\sigma(\mathcal{M}_b,\mathbf{C}_0)}{\longrightarrow} \mu \Longrightarrow \sup \|\mu_n\| < +\infty,$$

that is, $\sup_{n\in\mathbf{N}} |\mu_n|(\Omega) < +\infty$. The converse statement follows from a density argument which is similar to the one developed in Proposition 2.4.8. (Note that $\mathbf{C}_c(\Omega)$ is dense in $\mathbf{C}_0(\Omega)$.) □

Corollary 2.4.1. *On any bounded subset of $\mathcal{M}_b(\Omega)$ there is the equivalence*

$$\mu_n \overset{\sigma(\mathcal{M}_b,\mathbf{C}_0)}{\longrightarrow} \mu \Longleftrightarrow \mu_n \longrightarrow \mu.$$

Let us now return to the general abstract properties of the weak* topologies and state the following compactness theorem, which explains the importance of these topologies.

Theorem 2.4.7 (Banach–Alaoglu–Bourbaki). *Let V be a normed linear space. Then the unit ball $B_{V^*} = \{v^* \in V^* : \|v^*\|_* \leq 1\}$ of the topological dual V^* is compact for the topology $\sigma(V^*, V)$.*

PROOF. An element $v^* \in V^*$ is a function from V into $\mathbf{R}$. Let us write briefly $\mathbf{R}^V$ for the set of all functions from V into $\mathbf{R}$ and denote by i the canonical embedding

$$\begin{aligned} i : V^* &\longrightarrow \mathbf{R}^V, \\ v^* &\longmapsto \{\langle v^*, v\rangle\}_{v\in V}. \end{aligned}$$

When $\|v^*\|_* \leq 1$, we have indeed $|\langle v^*, v\rangle| \leq \|v\|$, so

$$\begin{aligned} i : B_{V^*} &\longmapsto \prod_{v\in V} [-\|v\|, +\|v\|] := Y, \\ i(v^*) &= \{\langle v^*, v\rangle\}_{v\in V}. \end{aligned}$$

Let us endow Y with the product topology, which is the weakest topology on Y making all the projections continuous. This topology induces on $i(B_{V^*})$ the weak* topology $\sigma(V^*, V)$; this is exactly the way it has been defined. The topological space Y, which is a product of compact spaces and which is equipped with the product topology, is compact; this is the compactness Tikhonov theorem. So, we have just to verify that $i(B_{V^*})$ is closed in Y. This is clear since for all generalized sequence $(v_\nu^*)_{\nu\in I}$, the convergence of $i(v_\nu^*)$ for the product topology

$$\langle v_\nu^*, v\rangle \longrightarrow \Phi(v) \in Y$$

implies that Φ is still linear and $|\Phi(v)| \leq \|v\|$, so that

$$\Phi(v) = \langle v^*, v\rangle$$

for some $v^* \in B_{V^*}$ and $i(v_\nu^*) \longrightarrow i(v^*)$. □

To obtain a weak* sequential compactness result on B_{V^*} we need to assume a separability condition on V.

Let us recall that a topological space V is said to be separable if there exists a dense countable subset of V. Typically that is the case of spaces $\mathbf{C}_0(\Omega)$, $L^p(\Omega)$ for $1 \leq p < +\infty$ but not $L^\infty(\Omega)$.

Theorem 2.4.8. *Let V be a separable normed space. Then the unit ball B_{V^*} of V^* is metrizable for the topology $\sigma(V^*, V)$.*

Before proving Theorem 2.4.8 let us formulate the following important result, which is a direct consequence of Theorems 2.4.7 and 2.4.8.

Corollary 2.4.2. *Let V be a separable normed linear space and $(v_n^*)_{n\in\mathbf{N}}$ a bounded sequence in V^*. Then one can extract a subsequence $(v_{n_k}^*)_{k\in\mathbf{N}}$ which converges for the topology $\sigma(V^*, V)$.*

PROOF OF THEOREM 2.4.8. Let $(v_n)_{n\geq 1}$ be a dense countable subset of the unit ball B_V of V (which exists since V is assumed to be separable; note that separability is a hereditary property).

Then define on the unit ball B_{V^*} of V^* the following distance d:

$$\forall\, u^*, v^* \in B_{V^*}\ \ d(u^*, v^*) = \sum_{n=1}^{\infty} \frac{1}{2^n} |\langle u^* - v^*, v_n \rangle|.$$

Let us verify that the topology associated to the distance d coincides with the weak* topology $\sigma(V^*, V)$ on B_{V^*}. This can be done with the help of neighborhoods or by using generalized sequences (nets): one has to verify that for an arbitrary net $(v_i^*)_{i\in I}$ contained in B_{V^*},

$$v_i^* \overset{\sigma(V^*,V)}{\longrightarrow} v^* \Longleftrightarrow d(v_i^*, v) \underset{I}{\longrightarrow} 0.$$

It is easy to verify that since $\sup_{i\in I} \|v_i^*\| \leq 1$,

$$d(v_i^*, v) \underset{I}{\longrightarrow} 0 \Longleftrightarrow \forall k \in \mathbf{N} \quad \langle v_i^*, v_k \rangle \underset{I}{\longrightarrow} \langle v^*, v_k \rangle.$$

So we have to verify that

$$\forall v \in B_V \langle v_i^*, v \rangle \longrightarrow \langle v^*, v \rangle \Longleftrightarrow \forall k \in \mathbf{N}\ \langle v_i^*, v_k \rangle \longrightarrow \langle v^*, v_k \rangle.$$

This is exactly the same argument as in Proposition 2.4.8, where one just uses nets instead of sequences. □

Back to weak* convergence of measures. Let us denote by Ω a locally compact topological metrizable space which is σ-compact. Recall that $V = \mathbf{C}_0(\Omega)$ is separable: one can first notice that $\mathbf{C}_c(\Omega)$ is dense in $\mathbf{C}_0(\Omega)$, $\mathbf{C}_c(\Omega) = \bigcup_n \mathbf{C}(K_n)$, where K_n are compact, because of the σ-compactness assumption. Then observe that $\mathbf{C}(K)$ is separable; this can be obtained as a consequence of the Stone–Weierstrass theorem. Note that the metrizability of Ω is equivalent to the separability of $\mathbf{C}(K)$ (see [94, Theorem 2.3.29]).

We can now reformulate Corollary 2.4.2 in the case of sequences of measures.

Proposition 2.4.14. *Let Ω be a locally compact, metrizable, σ-compact topological space. Then, from any bounded sequence of Borel measures $(\mu_n)_{n\in\mathbf{N}}$ on Ω, i.e., verifying*

$$\sup_{n\in\mathbf{N}} |\mu_n|(\Omega) < +\infty,$$

one can extract a subsequence $(\mu_{n_k})_{k\in\mathbf{N}}$ which is $\sigma(\mathcal{M}_b, \mathbf{C}_0)$ convergent to some bounded Borel measure μ:

$$\forall \varphi \in \mathbf{C}_0(\Omega) \quad \int_\Omega \varphi\, d\mu_{n_k} \underset{k\to+\infty}{\longrightarrow} \int_\Omega \varphi\, d\mu.$$

As a particular important situation where the result above can be directly applied let us mention the following.

Corollary 2.4.3. *Let Ω be an open subset of $\mathbf{R}^N$ equipped with the Lebesgue measure dx and $(f_n)_{n\in\mathbf{N}}$ a sequence of functions which is bounded in $L^1(\Omega)$, i.e., $\sup_{n\in\mathbf{N}} \int_\Omega |f_n(x)|dx < +\infty$. Then there exists a subsequence $(f_{n_k})_{k\in\mathbf{N}}$ and a bounded Borel measure μ on Ω such that*

$$\forall \varphi \in \mathcal{C}_0(\Omega) \quad \int_\Omega \varphi(x) f_{n_k}(x) dx \underset{k\to+\infty}{\longrightarrow} \int_\Omega \varphi(x) d\mu(x).$$

PROOF. Apply Proposition 2.4.14 to the sequence $\mu_n = f_n\,dx$ and note that $|\mu_n|(\Omega) = \int_\Omega |f_n|\,dx$. □

Remark 2.4.8. (a) The above result, which allows us to extract from any bounded sequence $(f_n)_{n\in\mathbf{N}}$ in $L^1(\Omega)$ a subsequence which is $\sigma(\mathcal{M}_b, \mathbf{C}_0)$ convergent to a bounded Borel measure μ, relies on the embedding of L^1 into $\mathcal{M}_b(\Omega)$ which is a dual, namely, of $\mathbf{C}_0$. This is a general method which consists, when one has some estimations on a sequence $(v_n)_{n\in\mathbf{N}}$ in some space X, to embed $X \hookrightarrow Y^*$, where Y^* is the topological dual of some (separable) normed space Y. Then one can extract a subsequence $v_{n_k} \overset{\sigma(Y^*,Y)}{\longrightarrow} y^*$, the limit y^* belonging to Y^*.

(b) As an example of application of Corollary 2.4.3, take the sequence $(v_n)_{n\in\mathbf{N}}$ defined at the beginning of Section 2.4.5; one has

$$v_n \overset{\sigma(\mathcal{M}_b,\mathbf{C}_0)}{\longrightarrow} \delta_0,$$

where δ_0 is the Dirac measure at the origin.

(c) As a counterpart of its generality, the information given by a weak* convergence $f_n dx \overset{\sigma(\mathcal{M}_b,\mathbf{C}_0)}{\longrightarrow} \mu$ or even $f_n \overset{\sigma(L^1,L^\infty)}{\longrightarrow} f$ is often not sufficient to analyze some situations, for example, occuring in the study of some nonlinear PDEs. To treat such situations, we will introduce the concept of Young measures in chapter 4.

Chapter 3

Abstract variational principles

The introduction in chapter 2 of a weak formulation of the model examples (Dirichlet problem, Stokes system) leads to study of the following problem.

Given a linear vector space V, a bilinear symmetrical form $a : V \times V \longrightarrow \mathbf{R}$, and a linear form $L : V \longrightarrow \mathbf{R}$, find $u \in V$ such that

$$a(u, v) = L(v) \quad \forall\, v \in V. \tag{3.1}$$

When a is positive, this turns out to be equivalent to the following minimization problem: find $u \in V$ such that

$$J(u) \leq J(v), \tag{3.2}$$

where $J(v) := \frac{1}{2}a(v, v) - L(v)$.

In this chapter, we introduce the topological and geometrical concepts which allows us to solve this kind of problem and much more.

3.1 The Lax–Milgram theorem and the Galerkin method

3.1.1 The Lax–Milgram theorem

In this section, V is a Hilbert space equipped with the scalar product $\langle \cdot, \cdot \rangle$ and the associated norm:

$$\forall v \in V \quad \|v\|^2 = \langle v, v \rangle.$$

Let us recall the celebrated Riesz theorem.

Theorem 3.1.1 (Riesz). *Let V be a Hilbert space and $L \in V^*$ a linear continuous form on V. Then there exists a unique $f \in V$ such that*

$$\forall\, v \in V \quad L(v) = \langle f, v \rangle.$$

Notice that given $f \in V$, the linear form L_f defined by

$$L_f(v) = \langle f, v \rangle$$

satisfies (by application of the Cauchy–Schwarz inequality)

$$|L_f(v)| \leq \|f\| \, \|v\|$$

and hence

$$\|L_f\|_* \leq \|f\|,$$

where $\|L_f\|_*$ is the dual norm of the continuous linear form L_f. On the other hand, by taking $v = \frac{1}{\|f\|} f$ (if $f \neq 0$) we obtain

$$\|L_f\|_* \geq \|f\|.$$

The Riesz theorem tells us that the linear isometrical embedding $f \mapsto L_f$ from V into V^* is onto. So V and V^* can be identified both as vector spaces and as Hilbert spaces.

Note that it is not completely correct to say that the topological dual of V is V itself! The Riesz theorem tells us that the topological dual of V, that is, V^*, is isometric to V and describes how any element of V^* can be uniquely represented with the help of an element of V: the mapping

$$f \in V \mapsto L_f \in V^*$$

is an isometrical isomorphism from V onto V^*.

So we will often identify V^* with V. But one may imagine other representations of V^*. We will illustrate this when describing the dual of the Sobolev space $H_0^1(\Omega)$.

An important situation where one has to be careful with such identifications is when we have an embedding (i linear continuous, V dense in H)

$$V \overset{i}{\hookrightarrow} H$$

of two Hilbert spaces, $(V, (\cdot, \cdot))$ and $(H, \langle \cdot, \cdot \rangle)$. Clearly, any linear continuous form L on H when "restricted" to V (indeed, $L_{|V} = L \circ i$) defines a linear continuous form on V. So $H^* \subset V^*$ and the mapping $L \in H^* \mapsto L_{|V} \in V^*$ is one to one because V is dense in H.

When identifying H and H^* we have the usual "triplet"

$$V \hookrightarrow H \hookrightarrow V^*.$$

But now we cannot also identify V and V^* because we should end with the conclusion that $H \hookrightarrow V$!

Thus one has to choose one identification. One cannot identify both (H and H^*) and (V and V^*). A typical situation is

$$V = H_0^1(\Omega) \hookrightarrow L^2(\Omega) \hookrightarrow H^{-1}(\Omega) = V^*.$$

This leads to a representation of the dual of the Hilbert space $V = H_0^1(\Omega)$ which is different from the Riesz representation. The Riesz representation theorem will play a key role when establishing the following theorem.

Theorem 3.1.2 (Lax–Milgram). *Let V be a Hilbert space with the scalar product $\langle \cdot, \cdot \rangle$ and $\| \cdot \| = \sqrt{\langle \cdot, \cdot \rangle}$ the associated norm.*

Let $a : V \times V \longrightarrow \mathbf{R}$ be a bilinear form which satisfies (i) *and* (ii)*:*

(i) *a is continuous, that is, there exists a constant $M \in \mathbf{R}^+$ such that*

$$\forall u, v \in V \quad |a(u, v)| \leq M\|u\| \cdot \|v\|;$$

(ii) *a is coercive, that is, there exists a constant $\alpha > 0$ such that*

$$\forall v \in V \quad a(v, v) \geq \alpha \|v\|^2.$$

Then for any $L \in V^*$ (L is a linear continuous form on V) there exists a unique $u \in V$ such that

$$a(u, v) = L(v) \quad \forall v \in V.$$

Remark 3.1.1. Let us make some complements to the discussions above.

(a) Before proving the Lax–Milgram theorem, let us notice that it contains as a particular case the Riesz representation theorem. Take $a(u, v) = \langle u, v \rangle$ and verify (i) and (ii). By the Cauchy–Schwarz inequality we have

$$|a(u, v)| \leq \|u\| \, \|v\|.$$

Hence a is continuous (take $M = 1$). Moreover, $a(v, v) = \langle v, v \rangle = \|v\|^2$ and a is coercive (take $\alpha = 1$). So, for any $L \in V^*$ there exists a unique $u \in V$ such that

$$L(v) = \langle u, v \rangle \quad \forall \, v \in V,$$

and this is the Riesz representation theorem.

(b) One can easily verify that for a bilinear form, the continuity property is equivalent to the existence of some constant $M \geq 0$ such that

$$|a(u, v)| \leq M\|u\| \, \|v\|. \tag{3.3}$$

Let us first verify that if (3.3) is satisfied, then a is continuous: take $u_n \longrightarrow u$ and $v_n \longrightarrow v$. Then

$$\begin{aligned} a(u_n, v_n) - a(u, v) &= a(u_n, v_n) - a(u_n, v) + a(u_n, v) - a(u, v) \\ &= a(u_n, v_n - v) + a(u_n - u, v). \end{aligned}$$

It follows that

$$|a(u_n, v_n) - a(u, v)| \leq M\|u_n\| \cdot \|v_n - v\| + M\|u_n - u\| \cdot \|v\|. \tag{3.4}$$

The sequence $(u_n)_{n \in \mathbf{N}}$ being norm convergent is bounded in V and there exists a constant $C \geq 0$ such that $\sup_n \|u_n\| \leq C$. Returning to (3.4),

$$|a(u_n, v_n) - a(u, v)| \leq M\big[C\|v_n - v\| + \|v\| \cdot \|u_n - u\|\big],$$

which implies that $\lim_n a(u_n, v_n) = a(u, v)$.

Conversely, let us assume that a is a bilinear continuous form on $V \times V$. Since $a(0,0) = 0$ and a is continuous at $(0,0)$, for any $\varepsilon > 0$ there exists some $\eta(\varepsilon) > 0$ such that

$$\|u\| \leq \eta(\varepsilon) \text{ and } \|v\| \leq \eta(\varepsilon) \Longrightarrow a(u,v)| \leq \varepsilon.$$

Take u, v arbitrary elements of V, $u \neq 0$, $v \neq 0$. Then

$$\left\| \frac{\eta(\varepsilon)}{\|u\|} u \right\| \leq \eta(\varepsilon) \quad \text{and} \quad \left\| \frac{\eta(\varepsilon)}{\|v\|} v \right\| \leq \eta(\varepsilon).$$

Hence

$$\left| a\left(\frac{\eta(\varepsilon)}{\|u\|} u, \frac{\eta(\varepsilon)}{\|v\|} v \right) \right| \leq \varepsilon,$$

which implies

$$\forall\, u, v \neq 0 \quad |a(u,v)| \leq \frac{\varepsilon}{\eta(\varepsilon)^2} \|u\| \cdot \|v\|.$$

This is still true if u or v is the zero element of V. So, one can take $M = \varepsilon/\eta^2(\varepsilon)$.

Proof of the Lax–Milgram theorem. When establishing a weak formulation for some partial differential equations or systems

$$Au = f \tag{3.5}$$

(for example, $Au = -\Delta u$ for the Dirichlet problem with prescribed boundary data contained in the domain of A), we have been led to study problems of the form

$$a(u,v) = L(v) \quad \forall\, v \in V. \tag{3.6}$$

Indeed, we are going to reconstruct an abstract equation (3.5) from (3.6). The major interest of this reverse operation is that now we are able to formulate precisely the topological and geometrical properties of the operator A.

Let us first apply the Riesz theorem to L: there exists some $f \in V$ such that

$$L(v) = \langle f, v \rangle \quad \forall v \in V. \tag{3.7}$$

For any fixed $u \in V$, the mapping $v \mapsto a(u,v)$ is a continuous linear form on V; note that

$$|a(u,v)| \leq M\|u\| \, \|v\| \quad \forall v \in V.$$

Applying once more the Riesz theorem, there exists a unique element, which we denote $A(u) \in V$, such that

$$\forall v \in V \quad a(u,v) = \langle A(u), v \rangle.$$

The mapping $u \mapsto A(u)$ from V into V is linear: given u_1, u_2 belonging to V

$$\begin{aligned} \langle A(u_1 + u_2), v \rangle &= a(u_1 + u_2, v) \\ &= a(u_1, v) + a(u_2, v) \\ &= \langle A(u_1), v \rangle + \langle A(u_2), v \rangle \\ &= \langle A(u_1) + A(u_2), v \rangle \qquad \forall\, v \in V. \end{aligned}$$

Hence,

$$A(u_1 + u_2) = A(u_1) + A(u_2).$$

Similarly,

$$\forall\, \lambda \in \mathbf{R},\ \forall\, u \in V \quad A(\lambda u) = \lambda A(u).$$

So, our problem can be reformulated as follows: find $u \in V$ such that

$$\langle A(u), v\rangle = \langle f, v\rangle \quad \forall v \in V,$$

that is,

$$Au = f. \tag{3.8}$$

Let us reformulate in terms of A the properties of the bilinear form $a(\cdot, \cdot)$:

(a) since a is bilinear we have that

$$A : V \longrightarrow V \text{ is a linear mapping;} \tag{3.9}$$

(b) A is continuous: for any $u, v \in V$

$$\begin{aligned}\langle Au, v\rangle &= a(u, v)\\ &\leq M\|u\|\ \|v\|.\end{aligned}$$

Taking $v = Au$ we obtain

$$\|Au\|^2 \leq M\|u\|\ \|Au\|,$$

which implies

$$\|Au\| \leq M\ \|u\|. \tag{3.10}$$

This expresses that A is a linear continuous operator from V into V with $\|A\|_{L(V,V)} \leq M$.

(c) A is coercive in the following sense: there exists some $\alpha > 0$ such that

$$\forall v \in V \qquad \langle Av, v\rangle \geq \alpha\|v\|^2. \tag{3.11}$$

To solve (3.8) we formulate it as a fixed point problem. Let λ be some strictly positive parameter. Clearly, to solve (3.8) is equivalent to finding $u \in V$ such that

$$u - \lambda(Au - f) = u. \tag{3.12}$$

In other words, we are looking for a fixed point $u \in V$ of the mapping $g_\lambda : V \longrightarrow V$ given by

$$g_\lambda(v) = v - \lambda(Av - f). \tag{3.13}$$

Let us prove that with λ adequately chosen, the mapping g_λ satisfies the condition of the Banach fixed point theorem, which we recall now.

Theorem 3.1.3 (Banach fixed point theorem—Picard iterative method). *Let (X, d) be a complete metric space and $g : X \longrightarrow X$ be a Lipschitz continuous mapping with a Lipschitz constant k strictly less than one, i.e.,*

$$\forall\, x, y \in X \quad d(g(x), g(y)) \leq k\ d(x, y).$$

Then, there exists a unique $\bar{x} \in X$ such that $g(\bar{x}) = \bar{x}$. Moreover, for any $x_0 \in X$, the sequence (x_n) starting from x_0 with $x_{n+1} = g(x_n)$ for all $n \in \mathbf{N}$ converges to $\bar{x}$ as n goes to $+\infty$.

PROOF OF LAX–MILGRAM THEOREM CONTINUED. Take $v_1, v_2 \in V$. Then

$$\begin{aligned} g_\lambda(v_2) - g_\lambda(v_1) &= [v_2 - \lambda(Av_2 - f)] - [v_1 - \lambda(Av_1 - f)] \\ &= (v_2 - v_1) - \lambda A(v_2 - v_1). \end{aligned}$$

Let us denote $v = v_2 - v_1$. So

$$\|g_\lambda(v_2) - g_\lambda(v_1)\| = \|v - \lambda Av\|.$$

To majorize this quantity, we consider its square and take advantage of the Hilbertian structure of V:

$$\begin{aligned} \|g_\lambda(v_2) - g_\lambda(v_1)\|^2 &= \|v - \lambda Av\|^2 \\ &= \|v\|^2 - 2\lambda\langle Av, v\rangle + \lambda^2\|Av\|^2. \end{aligned}$$

By using (3.10) and (3.11) (note that we have assumed $\lambda > 0$), we obtain

$$\|g_\lambda(v_2 - g_\lambda(v_1)\|^2 \leq (1 - 2\lambda\alpha + \lambda^2 M^2)\ \|v\|^2. \tag{3.14}$$

So, the question is to find some $\lambda > 0$ such that $1 - 2\lambda\alpha + \lambda^2 M^2 < 1$. Take $\bar{\lambda}$ for which the quantity $1 - 2\alpha\lambda + \lambda^2 M^2$ is minimal, that is, $\bar{\lambda} = \alpha/M^2$, in which case

$$1 - 2\bar{\lambda}\alpha + \bar{\lambda}^2 M^2 = 1 - \frac{\alpha^2}{M^2} < 1.$$

Hence

$$\|g_{\bar{\lambda}}(v_2) - g_{\bar{\lambda}}(v_1)\| \leq \sqrt{1 - \frac{\alpha^2}{M^2}}\ \|v_2 - v_1\|, \tag{3.15}$$

and $k_{\bar{\lambda}} = \sqrt{1 - \alpha^2/M^2}$ is strictly less than one (note that $\alpha > 0$). So $g_{\bar{\lambda}}$ has a unique fixed point $\bar{u}$; equivalently, the equation $Au = f$ has a unique solution $\bar{u}$. □

Remark 3.1.2. (a) One of the main advantages of the proof above is that since it relies on the Banach fixed point theorem, it is a constructive proof.

(b) A second advantage is that it can be easily extended to nonlinear equations: solve

$$Au = f,$$

where $A : V \longrightarrow V$ satisfies

$$\exists M \geq 0 \ \text{ such that } \ \forall u, v \in V \quad \|Au - Av\| \leq M\|u - v\|;$$

$$\exists \alpha > 0 \ \text{ such that } \ \forall u, v \in V \quad \langle Au - Av, u - v\rangle \geq \alpha\|u - v\|^2.$$

(c) Another approach consists of proving that A is onto, that is, $R(A) = V$. To that end, one first establishes that

(i) $R(A)$ is closed: for this one can first notice that

$$\forall v \in V \quad \alpha\|v\|^2 \leq \langle Av, v\rangle \leq \|Av\| \cdot \|v\|$$

and hence

$$\alpha\|v\| \leq \|Av\|.$$

If $Av_n \longrightarrow z$, we have

$$\alpha\|v_n - v_m\| \leq \|A(v_n - v_m)\| = \|Av_n - Av_m\|$$

and (v_n) is a Cauchy sequence in V. Hence $v_n \longrightarrow v$ for some $v \in V$ and by continuity of A, $Av_n \longrightarrow Av$. Consequently $z = Av$.

(ii) $R(A)$ is dense: if $z \in R(A)^{\perp}$, then

$$\langle Av, z\rangle = 0 \quad \forall\, v \in V.$$

Take $v = z$ to conclude $z = 0$. Hence $R(A)^{\perp} = \{0\}$, that is, $\overline{R(A)} = V$.

3.1.2 The Galerkin method

A quite natural idea when considering an infinite dimensional (variational) problem is to approximate it by finite dimensional problems. This has important consequences both from the theoretical (existence, etc.) and the numerical point of view. In this section, we consider the situation corresponding to the Lax–Milgram theorem, and by using the Galerkin method we will both provide another proof of the existence of a solution and describe a corresponding approximation numerical schemes. We stress the fact that this type of finite dimensional approximation method is very flexible and can be applied (we will illustrate it further in various situations) to a large number of linear or nonlinear problems.

Definition 3.1.1. *Let V be a Banach space. A Galerkin approximation scheme is a sequence $(V_n)_{n\in\mathbf{N}}$ of finite dimensional subspaces of V such that for all $v \in V$, there exists some sequence $(v_n)_{n\in\mathbf{N}}$ with $v_n \in V_n$ for all $n \in \mathbf{N}$ and $(v_n)_{n\in\mathbf{N}}$ norm-converging to v.*

This approximation property can be reformulated as

$$\forall v \in V \quad \lim_{n\to+\infty} \operatorname{dist}(v, V_n) = 0,$$

where $\operatorname{dist}(v, V_n) = \inf_{w\in V_n} \|v - w\|$.

Proposition 3.1.1. *Let V be a separable Banach space. Then one can construct a Galerkin approximation scheme (V_n) by the following method:*

(i) *take $(u_n)_{n\in\mathbf{N}}$ a countable dense subset of V (the separability of V just expresses that such set exists);*

(ii) *let $V_n = \operatorname{span}\{u_1, u_2, \ldots, u_n\}$.*

Then $(V_n)_{n\in\mathbf{N}}$ is a Galerkin scheme.

PROOF. Let $v \in V$. By (i), there exists a mapping $k \mapsto n(k)$ from $\mathbf{N}$ into $\mathbf{N}$ such that $\|u_{n(k)} - v\| \leq \frac{1}{k}$ for all $k \in \mathbf{N}^*$. For k fixed, $u_{n(k)} \in V_{n(k)}$ and hence

$$\operatorname{dist}(v, V_{n(k)}) \leq \frac{1}{k}.$$

Since $V_n \supset V_{n(k)}$ for $n \geq n(k)$, we obtain

$$\forall n \geq n(k) \quad \operatorname{dist}(v, V_n) \leq \operatorname{dist}(v, V_{n(k)}) \leq \frac{1}{k},$$

that is, $\operatorname{dist}(v, V_n) \longrightarrow 0$ as $n \to +\infty$. □

Remark 3.1.3. (a) The vectors $u_1, \ldots, u_n$ need not to be linearly independent. By a classical linear algebra argument, one can replace in Proposition 3.1.1 the sequence $(u_n)_{n\in\mathbf{N}}$ by a sequence $(w_n)_{n\in\mathbf{N}}$, $w_n \in V$ made by linearly independent vectors.

(b) In Proposition 3.1.1 the sequence of subspaces $(V_n)_{n\in\mathbf{N}}$ satisfies

• $V_1 \subset V_2 \subset V_3 \subset \cdots \subset V_n \subset \cdots$ is an increasing sequence of finite dimensional subspaces;

• $\overline{U_{n\in\mathbf{N}} V_n} = V$.

The Galerkin approach to the Lax–Milgram theorem. We now suppose that V is a separable Hilbert space, $a : V \times V \longrightarrow \mathbf{R}$ is a bilinear, continuous, coercive form, $L : V \longrightarrow \mathbf{R}$ is a linear, continuous form.

We want to study the following problem: find $u \in V$ such that

$$a(u, v) = L(v) \quad \forall\, v \in V. \tag{3.16}$$

Since V is separable, by Proposition 3.1.1 there exists a Galerkin scheme $(V_n)_{n\in\mathbf{N}}$ with V_n increasing with $n \in \mathbf{N}$. Consider the approximated problems

$$\begin{cases} \text{find } u_n \in V_n \text{ such that} \\ a(u_n, v) = L(v) \quad \forall\, v \in V_n. \end{cases} \tag{3.17}$$

Problem (3.17) can be equivalently reformulated as

$$A_n u_n = f_n, \tag{3.18}$$

where A_n is the linear operator from $V_n \longrightarrow V_n$ such that

$$a(u, v) = \langle A_n u, v\rangle \qquad \forall\, u, v \in V_n$$

and

$$L(v) = \langle f_n, v\rangle \qquad \forall\, v \in V_n.$$

This is exactly the same argument as in the proof of the Lax–Milgram theorem except that now we work on finite dimensional spaces, which makes the existence of $(u_n)_{n\in\mathbf{N}}$ very easy: since $\ker A_n = 0$ (which follows from the coercivity of A_n), then A_n is onto. Note the basic difference with the infinite dimensional situation where such an argument is false!

The question now is to study the convergence of the sequence $(u_n)_{n\in\mathbf{N}}$. In (3.17) take $v = u_n$ so that

$$\alpha\|u_n\|^2 \le a(u_n, u_n) = L(u_n) \le \|L\|_*\|u_n\|$$

and

$$\|u_n\| \le \frac{\|L\|_*}{\alpha}. \tag{3.19}$$

The sequence $(u_n)_{n\in\mathbf{N}}$ is bounded and hence weakly relatively compact, that is, there exists a subsequence (u_{n_k}) and some $u \in V$ such that

$$\langle u_{n_k}, v\rangle \longrightarrow \langle u, v\rangle \ \forall v \in V. \tag{3.20}$$

We write $u_{n_k} \overset{w-V}{\longrightarrow} u$. (See Theorem 2.4.3 with a direct independent proof of this result in separable Hilbert spaces.) Given $v \in V_m$, we have $V_{n_k} \supset V_m$ for k sufficiently large and hence

$$a(u_{n_k}, v) = L(v) \tag{3.21}$$

for all k sufficiently large. Then notice that

$$u_{n_k} \overset{w-V}{\longrightarrow} u \Longrightarrow a(u_{n_k}, v) \longrightarrow a(u, v) \text{ as } k \longrightarrow +\infty.$$

This follows from the representation of the linear continuous form $u \mapsto a(u, v)$ on V, where $a(u, v) = \langle u, A^t v\rangle$ (A^t denotes the adjoint of A). Hence

$$u_{n_k} \overset{w-V}{\longrightarrow} u \Longrightarrow a(u_{n_k}, v) = \langle u_{n_k}, A^t v\rangle \longrightarrow \langle u, A^t v\rangle = a(u, v).$$

So, when passing to the limit in (3.21), we obtain that, given $v \in V_m$,

$$a(u, v) = L(v).$$

Hence $a(u, v) = L(v)$ for every $v \in \bigcup_{m\in\mathbf{N}} V_m$. Since $\overline{\bigcup_{m\in\mathbf{N}} V_m} = V$ by continuity of $a(u, \cdot)$ and $L(\cdot)$ we finally infer

$$a(u, v) = L(v) \quad \forall\, v \in V.$$

Since u is the unique solution of this problem, by a classical compactness argument, the whole sequence $(u_n)_{n\in\mathbf{N}}$ weakly converges to u. Indeed, one can prove that the sequence $(u_n)_{n\in\mathbf{N}}$ norm converges to u. This is what is explained below with an explicit bound on $\|u_n - u\|$.

Proposition 3.1.2 (Cea lemma). *Let V be a separable Hilbert space and $(V_n)_{n\in\mathbf{N}}$ a Galerkin scheme. Suppose*

$$\begin{cases} a(u_n, v) = L(v) & \forall\, v \in V_n, \\ u_n \in V_n, \end{cases}$$

and

$$\begin{cases} a(u, v) = L(v) & \forall\, v \in V, \\ u \in V, \end{cases}$$

where a and L satisfy the assumptions of the Lax–Milgram theorem. Then,

$$\|u_n - u\| \leq \frac{M}{\alpha} \operatorname{dist}(u, V_n).$$

PROOF. Substracting (3.16) from (3.17), we obtain

$$a(u - u_n, v) = 0 \quad \forall v \in V_n,$$

and in particular

$$a(u - u_n, u_n) = 0.$$

It follows, for every $v \in V_n$,

$$\begin{aligned} a(u - u_n, u - u_n) &= a(u - u_n, u - v) + a(u - u_n, v - u_n) \\ &= a(u - u_n, u - v). \end{aligned}$$

Let us now use the continuity and coercivity property of a,

$$\alpha \ \|u - u_n\|^2 \leq M \|u - u_n\| \cdot \|u - v\|,$$

to obtain

$$\alpha \|u - u_n\| \leq M \ \|u - v\| \quad \forall \ v \in V_n.$$

Hence

$$\|u - u_n\| \leq \frac{M}{\alpha} \operatorname{dist}\ (u, V_n).$$

Since $(V_n)_{n \in \mathbf{N}}$ is a Galerkin scheme, $\operatorname{dist}(u, V_n) \longrightarrow 0$ as $n \longrightarrow +\infty$, and $(u_n)_{n \in \mathbf{N}}$ norm-converges to u. □

3.2 Minimization problems: The topological approach

As a corollary of the Lax–Milgram theorem, we have obtained that if $a : V \times V \longrightarrow \mathbf{R}$ is bilinear, continuous, coercive, and symmetric, then, for any $L \in V^*$, there exists a unique solution to the minimization problem: find $u \in V$ such that

$$J(u) \leq J(v) \quad \forall v \in V,$$

where $J(v) = \frac{1}{2}\, a(v, v) - L(v)$.

Let us observe that $J : V \longrightarrow \mathbf{R}$ is convex, continuous, and coercive. By coercive, we mean that $\lim_{\|v\| \longrightarrow +\infty} \ J(v) = +\infty$, which follows easily from the inequality

$$J(v) \geq \frac{\alpha}{2} \ \|v\|^2 - \|L\|_* \cdot \|v\|.$$

Indeed, we will prove in Section 3.3.2 the following general result, which contains as a particular case the above situation (Theorem 3.3.4).

Let $J : V \longrightarrow \mathbf{R} \cup \{+\infty\}$ be a real-extended valued function on a reflexive Banach space V, which is convex, lower semicontinuous, and coercive. Then there exists at least one $u \in V$ such that

$$J(u) \leq J(v) \quad \forall \ v \in V.$$

Before proving this theorem, in this section we will successively examine its basic ingredients. We will first justify the introduction of extended real-valued functions. Then, we will state the Weierstrass minimization theorem, which is purely topological, and in the process we will study the notions of lower semicontinuity, inf-compactness, and the interplay between inf-compactness, coercivity, and the role of the weak topology. So doing, we will be able to explain why convexity plays an important role in such questions.

3.2.1 Extended real-valued functions

A main reason for introducing extended real-valued functions is that they provide a natural and flexible modelization of minimization (or maximization) problems with constraints. Since in this chapter we consider minimization problems, we just need to consider functions

$$f : X \longrightarrow \mathbf{R} \cup \{+\infty\}.$$

On the other hand, if one considers maximization problems, one needs to introduce functions $f : X \longrightarrow \mathbf{R} \cup \{-\infty\}$. If maximization and minimization are both involved, just like in saddle value problems, one needs to work with functions $f : X \longrightarrow \bar{\mathbf{R}}$.

The (effective) *domain* of a function $f : X \longrightarrow \mathbf{R} \cup \{+\infty\}$ is the set

$$\operatorname{dom} f = \{x \in X \ : \ f(x) < +\infty\}.$$

The function f is said to be proper if $\operatorname{dom} f \neq \emptyset$.

Let us briefly justify the introduction of extended real-valued functions. Most minimization problems can be written as

$$\min\{f_0(x) \ : \ x \in C\}, \tag{3.22}$$

where $f_0 : X \longrightarrow \mathbf{R}$ is a real-valued function and $C \subset X$ is the set of constraints. In economics, C describes the available resources, the possible productions of a firm, or a set of decisions, and f_0 is the corresponding cost or economical criteria. In physics, the configurations x of the system are subject to constraints (unilateral or bilateral) and f_0, for example, is the corresponding energy.

A natural way to solve such problems is to approach them by penalization. For example, let us introduce a distance d on X and, for any positive real number k, consider the minimization problem

$$\min\{f_0(x) + k\, d(x, C) \ : \ x \in X\}, \tag{3.23}$$

where

$$d(x, C) = \inf\{d(x, y) \ : \ y \in C\} \tag{3.24}$$

is the distance function from x to C. Note that the penalization term is equal to zero if $x \in C$ (that is, if the constraint is fulfilled), and when $x \notin C$ (that is, if the constraint is violated) it takes larger and larger values which are increasing to $+\infty$ with k.

Let us also notice that the approximated problem (3.23) can be written as

$$\min\{f_k(x) \ : \ x \in X\},$$

where

$$f_k(x) = f_0(x) + k\, d(x, C)$$

is a real-valued function. Then the approximated problems are unconstrained problems, which makes the method interesting. As $k \to +\infty$, the sequence of functions $\{f_k : k \in \mathbf{N}\}$ increases to the function $f : X \longrightarrow \mathbf{R} \cup \{+\infty\}$, which is equal to

$$f(x) = \begin{cases} f_0(x) & \text{if } x \in C, \\ +\infty & \text{otherwise.} \end{cases} \tag{3.25}$$

The function f is an extended real-valued function, so, if we want to treat in a unified way problems (3.22) and (3.23) we are naturally led to introduce extended real-valued functions. The minimization problem (3.22) can be equivalently formulated as

$$\min\{f(x) : x \in X\},$$

where f is given by (3.25). Note that in this formulation the constraint is equal to the domain of f. A particularly useful function in this unilateral framework is the *indicator function* δ_C of the set C:

$$\delta_C(x) = \begin{cases} 0 & \text{if } x \in C, \\ +\infty & \text{otherwise.} \end{cases} \tag{3.26}$$

With this notation we have $f = f_0 + \delta_C$.

More generally, in variational analysis and optimization, one is often faced with expressions of the form

$$f = \sup_{i \in I} f_i$$

and one should notice that the class of extended real-valued functions is stable under such supremum operation. As a further illustration of these considerations, the convex duality theory establishes a one-to-one correspondence between a convex $C \subset X$ of a normed linear space and its support function

$$\sigma_C : X^* \longrightarrow \mathbf{R} \cup \{+\infty\},$$

which is defined by

$$\sigma_C(x^*) = \sup\{x^*(x) : x \in C\}$$

with X^* the topological dual of X. One cannot avoid the value $+\infty$ for σ_C as soon as the set C is not bounded (which is often the case!) and general duality statements must consider extended real-valued functions.

So, from now on, unless explicitly specified, we will consider functions $f : X \longrightarrow \mathbf{R} \cup \{+\infty\}$ possibly taking the value $+\infty$. We adopt the conventions that $\lambda \times (+\infty) = +\infty$ if $\lambda > 0$ and $0 \times (+\infty) = 0$.

3.2.2 The interplay between functions and sets: The role of the epigraph

The analysis of unilateral problems (like minimization) naturally leads to the introduction of mathematical concepts which have a unilateral character. The classical approach of analysis

does not provide the appropriate tools to deal with the mathematical objects and operations that are intrinsically unilateral (constraints, minimization). In the previous section, we justified the introduction of extended real-valued functions $f : X \longrightarrow \mathbf{R} \cup \{+\infty\}$ which is the appropriate concept for dealing with minimization problems. Although in classical analysis, the properties of the graph of a function play a fundamental role, in variational analysis it is the epigraph that will take over this role. The set

$$epi\ f = \{(x, \lambda) \in X \times \mathbf{R} \ : \ \lambda \geq f(x)\} \tag{3.27}$$

is the epigraph of the function $f : X \longrightarrow \mathbf{R} \cup \{+\infty\}$.

For any $\gamma \in \mathbf{R}$, the lower γ-level set of f is

$$lev_\gamma\ f = \{x \in X \ : \ f(x) \leq \gamma\}. \tag{3.28}$$

When considering the minimization problem of a given function $f : X \longrightarrow \mathbf{R} \cup \{+\infty\}$, the solution set is

$$\arg\min f = \{\bar{x} \in X \ : \ f(\bar{x}) = \inf_X f(x)\}. \tag{3.29}$$

Note that $\arg\min f$ can be possibly empty and

$$\arg\min f = \bigcap_{\gamma > \inf_X f} lev_\gamma f. \tag{3.30}$$

Thus, to an extended real-valued function, we have associated many different sets, its epigraph, its lower level sets, its minimum set. Conversely, to a set we have associated extended real-valued functions, for example, the indicator function, the support function, the distance function (this last one is real-valued if the set is nonempty). In the following section, we describe how some basic topological properties of functions for minimization problems can be naturally formulated with the help of the attached geometrical sets, epigraphs, and lower level sets. Note that the lower level sets can be obtained by cutting operations on the epigraph:

$$lev_\gamma f \times \{\gamma\} = epi\ f \cap (X \times \{\gamma\}).$$

Most basic operations in variational analysis can be naturally formulated with the help of the epigraphs. They give rise to the so-called epigraphical calculus; see [39], [47], [71]. The following result is valid for an arbitrary family of functions on an abstract space X.

Proposition 3.2.1. *Let X be an abstract space and $(f_i)_{i \in I}$ be a family of extended real-valued functions $f_i : X \longrightarrow \mathbf{R} \cup \{+\infty\}$ indexed by an arbitrary set I. Then*

$$epi(\sup_{i \in I}\ f_i) = \bigcap_{i \in I} epi\ f_i,$$

$$epi(\inf_{i \in I}\ f_i) = \bigcup_{i \in I} epi\ f_i.$$

3.2.3 Lower semicontinuous functions

Let (X, τ) be a topological space. For any $x \in X$, we denote by $\mathcal{V}_\tau(x)$ the family of the neighborhoods of x for the topology τ.

We recall the classical definition of continuity for a function $f : (X, \tau) \longrightarrow \mathbf{R}$. The function f is said to be continuous at $x \in X$ for the topology τ if

$$\forall\, \varepsilon > 0 \quad \exists\, V_\varepsilon \in \mathcal{V}_\tau(x) \text{ such that } \forall\, y \in V_\varepsilon \quad |f(y) - f(x)| < \varepsilon.$$

This can be viewed as the conjunction of the two following properties:

1. $\forall \varepsilon > 0 \quad \exists\, V_\varepsilon \in \mathcal{V}_\tau(x)$ such that $\forall\, y \in V_\varepsilon \quad f(y) > f(x) - \varepsilon$;
2. $\forall \varepsilon > 0 \quad \exists\, W_\varepsilon \in \mathcal{V}_\tau(x)$ such that $\forall\, y \in W_\varepsilon \quad f(y) < f(x) + \varepsilon$.

Then, take $V_\varepsilon \cap W_\varepsilon$ which still belongs to $\mathcal{V}_\tau(x)$ to obtain the continuity result. These properties are, respectively, called the lower semicontinuity and the upper semicontinuity of f at x for the topology τ. To deal with possibly extended real-valued functions, definition 1 of lower semicontinuity has to be formulated slightly differently.

Definition 3.2.1. *Let (X, τ) be a topological space and $f : X \longrightarrow \mathbf{R} \cup \{+\infty\}$. The function f is said to be τ-lower semicontinuous (τ-lsc) at x if*

$$\forall\, \lambda < f(x) \quad \exists\, V_\lambda \in \mathcal{V}_\tau(x) \quad \text{such that } f(y) > \lambda \quad \forall\, y \in V_\lambda.$$

(We write V_λ to stress the dependance of the set V upon the choice of λ!)

If f is τ-lsc at every point of X, then f is said to be τ-lower semicontinuous on X.

Proposition 3.2.2. *Let (X, τ) be a topological space and $f : X \longrightarrow \mathbf{R} \cup \{+\infty\}$ an extended real-valued function. The following statements are equivalent:*

(i) *f is τ-lower semicontinuous;*

(ii) *$epi\ f$ is closed in $X \times \mathbf{R}$ (where $X \times \mathbf{R}$ is equipped with the product topology of τ on X and of the usual topology on $\mathbf{R}$);*

(iii) $\forall \gamma \in \mathbf{R}$, *$lev_\gamma f$ is closed in (X, τ);*

(iv) $\forall \gamma \in \mathbf{R}$, $\{x \in X \,:\, f(x) > \gamma\}$ *is open in (X, τ);*

(v) $\forall x \in X \quad f(x) \leq \liminf_{y \to x} f(y) := \sup \inf_{V \in \mathcal{V}_\tau(x)\ y \in V} f(y)$.

Proof. We are going to prove (i) $\Longrightarrow$ (ii) $\Longrightarrow$ (iii) $\Longrightarrow$ (iv) $\Longrightarrow$ (v) $\Longrightarrow$ (i).

Assume that f is τ-lsc and prove that $epi f$ is closed. Equivalently, let us prove that the complement of $epi\ f$ in $X \times \mathbf{R}$ is open. Take $(x, ł) \notin epi\ f$. By definition of $epi\ f$, $ł < f(x)$. Take $ł < \gamma < f$. Since f is τ-lsc, there exists $V_\gamma \in \mathcal{V}_\tau(x)$ such that

$$f(y) > \gamma\ \forall y \in V_\gamma.$$

Equivalently, $(y, \gamma) \notin epi\ f$ for all $y \in V_\gamma$. It follows that

$$(V_\gamma \times]-\infty, \gamma[) \cap epi\ f = \emptyset.$$

Noticing that $V_\gamma \times]-\infty, \gamma[$ is a neighborhood of (x, λ), the conclusion follows.

(ii) $\Longrightarrow$ (iii). The implication follows directly from the relation

$$lev_\gamma f \times \{\gamma\} = epi\ f \cap (X \times \{\gamma\}).$$

Assuming that $epi\ f$ is closed, we infer that $lev_\gamma f \times \{\gamma\}$ is closed, and hence $lev_\gamma f$ is closed. (Note that $x \mapsto (x, \gamma)$ from X onto $X \times \{\gamma\}$ is a homeomorphism.)

(iii) $\Longrightarrow$ (iv) is obvious just by taking the complement of $lev_\gamma f$.

(iv) $\Longrightarrow$ (v). Let $\gamma < f(x)$. Since, by assumption, $\{y \in X : f(y) > \gamma\}$ is open, there exists some $V \in \mathcal{V}_\tau(x)$ such that $V \subset \{y \in X : f(y) > \gamma\}$. Equivalently,

$$\forall\ y \in V \quad f(y) > \gamma,$$

which implies

$$\inf_{y \in V} \quad f(y) \geq \gamma.$$

Hence

$$\sup_{V \in \mathcal{V}_\tau(x)} \inf_{y \in V} f(y) \geq \gamma,$$

and this being true for any $\gamma < f(x)$, it follows that

$$\sup_{V \in \mathcal{V}_\tau(x)} \inf_{y \in V} f(y) \geq f(x),$$

that is,

$$f(x) \leq \liminf_{y \longrightarrow x} f(y).$$

(v) $\Longrightarrow$ (i). Let $\lambda < f(x)$. By assumption (v)

$$\lambda < \sup_{V \in \mathcal{V}_\tau(x)} \inf_{y \in V} f(y),$$

which implies the existence of some $V_\lambda \in \mathcal{V}_\tau(x)$ such that

$$\inf_{y \in V_\lambda} \quad f(y) > \lambda,$$

i.e., $f(y) > \lambda$ for every $y \in V_\lambda$. This is exactly the lower semicontinuity property. □

The class of lower semicontinuous functions enjoys remarkable stability properties. Since closedness is preserved under arbitrary intersections and finite union of sets, we derive from Proposition 3.2.1 (epigraphical interpretation of sup and inf) and Proposition 3.2.2 (equivalence between f lsc and $epi f$ closed) the following important result.

Proposition 3.2.3. *Let (X, τ) be a topological space and $(f_i)_{i \in I}$, $f_i : X \longrightarrow \mathbf{R} \cup \{+\infty\}$ be an arbitrary collection of τ-lower semicontinuous functions. Then, $\sup_{i \in I} f_i$ is still τ-lower semicontinuous. When I is a finite set of indices, $\inf_{i \in I} f_i$ is still τ-lower semicontinuous.*

As a consequence, the supremum of a family of continuous functions is lower semicontinuous. One can prove that if (X, τ) is metrizable, the converse is true: if f is τ-lsc,

there exists an increasing sequence $(f_n)_{n\in\mathbf{N}}$ of τ-continuous functions which is pointwise convergent to f. We will establish this important approximation result in Theorem 9.2.1 by using the epigraphical regularization. See also [60, Theorem 1.3.7].

Proposition 3.2.4. *Let $f, g : (X,\tau) \longrightarrow \mathbf{R}\cup\{+\infty\}$ be two lower semicontinuous functions. Then, $f+g$ is still lower semicontinuous.*

PROOF. Take $x_\nu \longrightarrow x$ a τ-converging net. Then

$$\begin{aligned}\liminf_\nu (f+g)(x_\nu) &= \liminf_\nu [f(x_\nu)+g(x_\nu)]\\ &\geq \liminf_\nu f(x_\nu) + \liminf_\nu g(x_\nu)\\ &\geq f(x)+g(x). \quad \square\end{aligned}$$

3.2.4 The lower closure of a function and the relaxation problem

In some important situations, the function f to minimize fails to be lower semicontinuous for a topology τ which makes a minimizing sequence τ-relatively compact (see Section 3.2.5). In that case, the analysis of the behavior of the minimizing sequences requires the introduction of the lower closure of f and leads to the introduction of the relaxed problem.

Definition 3.2.2. *Given (X,τ) a topological space and $f : X \longrightarrow \mathbf{R}\cup\{+\infty\}$, the τ-lower envelope of f is defined as*

$$cl_\tau\ f = \sup\{g : X \to \mathbf{R}\cup\{+\infty\}\ :\ g\ \tau - lsc,\ g\leq f\}.$$

Proposition 3.2.5. *Let (X,τ) be a topological space and $f : X \longrightarrow \mathbf{R}\cup\{+\infty\}$ an extended real-valued function. Then $cl_\tau f$ is τ-lsc; it is the greatest τ-lsc function which minorizes f. We have the following properties:*

(a) *$epi(cl_\tau f)$ is the closure of $epi\ f$ in $X\times\mathbf{R}$ equipped with the product topology of τ with the usual topology of $\mathbf{R}$:*

$$epi(cl_\tau f) = cl(epi\ f);$$

(b) $cl_\tau f = \sup_{V\in\mathcal{V}_\tau(x)} \inf_{y\in V}\ f(y) = \liminf_{y\longrightarrow x} f(y)$;

(c) *f is τ-lsc at x iff $f(x)\leq cl_\tau\ f(x)$;*

(d) *f is τ-lsc at x iff $f(x) = cl_\tau\ f(x)$.*

PROOF. Let us first notice that a set C in $X\times\mathbf{R}$ is an epigraph iff

(i) C recedes in the vertical direction:

$$(x,ł)\in C \text{ and } \mu > ł \Longrightarrow (x,\mu)\in C;$$

(ii) C is vertically closed:

$$\text{for every } x\in X \text{ the set } \{ł\in\mathbf{R}\ :\ (x,ł)\in C\} \text{ is closed.}$$

(a) This implies that the closure of an epigraph is an epigraph. Set $cl(epi\ f) = epi\ g$. Since $epi\ g$ is closed, this implies that g is τ-lsc. Moreover, since $epi\ g \supset epi\ f$, we have $g \leq f$. Hence g is a τ-lsc minorant of f. We claim that g is the lower envelope of f, that is, g is the greatest of such τ-lsc minorants of f.

Take $h \leq f$ and h τ-lsc. Hence

$$epi\ h \supset epi\ f,$$

which implies

$$cl(epi\ h) = epi\ h \supset cl(epi\ f) = epi\ g.$$

Hence, $epi\ h \supset epi\ g$ and $h \leq g$. So

$$g = \sup\{h : X \to \mathbf{R} \cup \{+\infty\}\ :\ h\ \tau - lsc,\ h \leq f\},$$

that is, $g = cl_\tau f$. We have proved that

$$cl(epi\ f) = epi\ g = epi\ (cl_\tau\ f).$$

(b) Since $cl_\tau f$ is τ-lsc, it follows from Proposition 3.2.2(v) that

$$(cl_\tau f)(x) \leq \liminf_{y \longrightarrow x} (cl_\tau\ f)(y);$$

and since $cl_\tau\ f \leq f$,

$$(cl_\tau\ f)(x) \leq \liminf_{y \longrightarrow x}\ f(y) \quad \forall\, x \in X. \tag{3.31}$$

Then, notice that the function $h(x) = \liminf_{y \longrightarrow x}\ f(y)$ is less than or equal to f and is τ-lsc. To verify this last point, take

$$\lambda < \liminf_{y \longrightarrow x}\ f(y) = \sup_{V \in \mathcal{V}_\tau(x)}\ \inf_{y \in V}\ f(y).$$

Then there exists some open set $V_\lambda \in \mathcal{V}_\tau(x)$ such that $\inf_{y \in V_\lambda} f(y) > \lambda$. It follows that for all $\xi \in V_\lambda$, $V_\lambda \in \mathcal{V}_\tau(\xi)$ and hence

$$\sup_{V \in \mathcal{V}_\tau(\xi)}\ \inf_{y \in V}\ f(y) > \lambda,$$

that is,

$$\liminf_{y \longrightarrow \xi}\ f(y) > \lambda.$$

Thus, $x \mapsto \liminf_{y \longrightarrow x}\ f(x)$ is τ-lsc and minorizes f. By definition of $cl_\tau f$ we have

$$\liminf_{y \longrightarrow x} f(y) \leq (cl_\tau f)(x). \tag{3.32}$$

Then compare (3.31) and (3.32) to obtain

$$\forall x \in X \quad (cl_\tau f)(x) = \liminf_{y \longrightarrow x}\ f(y).$$

(c) Proposition 3.2.2 expresses that f is τ-lsc at x iff

$$f(x) \leq \liminf_{y \longrightarrow x} f(y).$$

This is equivalent to saying that

$$f(x) \leq cl_\tau f(x),$$

and since $cl_\tau f \leq f$ is always true, it is equivalent to $f(x) = cl_\tau f(x)$. □

We have the following "sequential" formulation of $cl_\tau f$.

Proposition 3.2.6. *Let (X, τ) be a topological space and $f : X \longrightarrow \mathbf{R} \cup \{+\infty\}$. Then, for any $x \in X$*

$$\begin{aligned}(cl_\tau f)(x) &= \liminf_{y \longrightarrow x} f(y) \\ &= \min\{\liminf_{\nu} f(x_\nu) \ : \ x_\nu \text{ is a net }, x_\nu \longrightarrow x\}.\end{aligned}$$

When (X, τ) is metrizable

$$\begin{aligned}(cl_\tau f)(x) &= \liminf_{y \longrightarrow x} f(y) \\ &= \min\{\liminf_{n} f(x_n) \ : \ (x_n) \text{ sequence, } x_n \longrightarrow x\}.\end{aligned}$$

Corollary 3.2.1. *Let (X, τ) be a metrizable space and let $f : X \longrightarrow \mathbf{R} \cup \{+\infty\}$ be an extended real-valued function. Then f is τ-lower semicontinuous at $x \in X$ iff*

$$\forall x_n \longrightarrow x \quad f(x) \leq \liminf_{n} f(x_n).$$

Proof of Proposition 3.2.6. We give for simplicity the proof only in the metrizable case. We have

$$\liminf_{y \longrightarrow x} f(y) = \sup_{\varepsilon > 0} \inf_{y \in B_\tau(x, \varepsilon)} f(y),$$

where $B_\tau(x, \varepsilon) = \{y \in X \ : \ d_\tau(y, x) < \varepsilon\}$, d_τ being a distance inducing the topology τ. Then, for any $x_n \longrightarrow x$, for any $\varepsilon > 0$, x_n belongs to $B_\tau(x, \varepsilon)$ for n sufficiently large. Hence

$$\inf_{y \in B_\tau(x, \varepsilon)} f(y) \leq f(x_n) \quad \forall n \geq N(\varepsilon).$$

Passing to the limit as $n \to +\infty$ gives

$$\inf_{y \in B_\tau(x, \varepsilon)} f(y) \leq \liminf_{n} f(x_n).$$

This being true for any $\varepsilon > 0$ and any $x_n \longrightarrow x$, leads to the inequality

$$\liminf_{y \longrightarrow x} f(y) \leq \inf\{\liminf_{n} f(x_n) \ : \ x_n \longrightarrow x\}. \tag{3.33}$$

On the other hand, for each $n \in \mathbf{N}$, there exists some $x_n \in B_\tau(x, 1/n)$ such that

$$\inf_{y\in B_\tau(x,1/n)} f(y) \geq f(x_n) - \frac{1}{n} \quad \text{if} \quad \inf_{y\in B_\tau(x,1/n)} f(y) > -\infty$$
$$-n \geq f(x_n) \quad \text{if} \quad \inf_{y\in B_\tau(x,1/n)} f(y) = -\infty.$$

In both cases,

$$\begin{aligned}\liminf_{y\longrightarrow x} f(y) = \lim_n \inf_{y\in B_\tau(x,1/n)} f(y) &\geq \limsup_n f(x_n) \\ &\geq \liminf_n f(x_n).\end{aligned} \tag{3.34}$$

Then compare (3.33) and (3.34) to obtain

$$\begin{aligned}\liminf_{y\longrightarrow x} f(y) &= \min\{\liminf_n \ f(x_n) \ : \ x_n \longrightarrow x\} \\ &= \min\{\limsup_n \ f(x_n) \ : \ x_n \longrightarrow x\}. \quad \square\end{aligned}$$

Proposition 3.2.7. *Let $f : (X,\tau) \longrightarrow \mathbf{R}\cup\{+\infty\}$ be an extended real-valued function. Then*

$$\inf_X \ f = \inf_X \ cl_\tau f.$$

More generally, for any τ-open subset G of X

$$\inf_G f = \inf_G cl_\tau f.$$

Moreover,

$$\arg\min f \subset \arg\min cl_\tau f.$$

Proof. (a) Since $f \geq cl_\tau f$, we just need to prove that

$$\inf_G \ cl_\tau f \geq \inf_G f.$$

For any $x \in G$, since G is τ-open, we have $G \in \mathcal{V}_\tau(x)$. By Proposition 3.2.5,

$$(cl_\tau f)(x) = \sup_{V\in\mathcal{V}_\tau(x)} \inf_{y\in V} f(y).$$

Hence, for every $x \in G$

$$(cl_\tau f)(x) \geq \inf_{y\in G} f(y).$$

This being true for any $x \in G$ gives

$$\inf_G (cl_\tau f) \geq \inf_G f.$$

(b) Let $x \in \arg\min f$. We have

$$cl_\tau f(x) \leq f(x) \leq \inf_X f = \inf_X \ cl_\tau f,$$

which implies $x \in \arg\min cl_\tau f$. $\square$

Remark 3.2.1. The function $cl_\tau f$, which we call the lower envelope of f, is often called the lower semicontinuous regularization of f or the *relaxed function* or the τ-closure. The problem $\min\{cl_\tau f(x) \ : \ x \in X\}$ is called the *relaxed problem*.

3.2.5 Inf-compactness functions, coercivity

Besides lower semicontinuity, the second basic ingredient in minimization problems is the inf-compactness property.

Definition 3.2.3. *Let (X, τ) be a topological space and let $f : X \longrightarrow \mathbf{R} \cup \{+\infty\}$ be an extended real-valued function. The function f is said to be τ-inf-compact if for any $\gamma \in \mathbf{R}$*

$$lev_\gamma\ f = \{x \in X\ :\ f(x) \leq \gamma\}$$

is relatively compact in X for the topology τ.

When f is τ-lsc, its lower level sets are closed for τ, and τ-compactness is equivalent to saying that the lower level sets of f are τ-compact.

Definition 3.2.4. *Let X be a normed linear space. A function $f : X \longrightarrow \mathbf{R} \cup \{+\infty\}$ is said to be coercive if $\lim_{\|x\| \longrightarrow +\infty}\ f(x) = +\infty$.*

The relation between the two concepts is made clear in the following.

Proposition 3.2.8. *Let X be a normed space and $f : X \longrightarrow \mathbf{R} \cup \{+\infty\}$. The following conditions are equivalent:*

(i) *f is coercive;*

(ii) *for any $\gamma \in \mathbf{R}$, $lev_\gamma f$ is bounded.*

PROOF. (i) $\Longrightarrow$ (ii). Assume f is coercive. If for some $\gamma_0 \in \mathbf{R}$, $lev_{\gamma_0} f$ is not bounded, then there exists a sequence $(x_n)_{n \in \mathbf{N}}$ such that $\|x_n\| \longrightarrow +\infty$ as n goes to $+\infty$ and $f(x_n) \leq \gamma_0$ for all $n \in \mathbf{N}$. But f coercive implies that $f(x_n) \longrightarrow +\infty$, a clear contradiction.

(ii) $\Longrightarrow$ (i). Assume that for any $\gamma \in \mathbf{R}$, $lev_\gamma f$ is bounded. If f is not coercive, we can construct a sequence $(x_n)_{n \in \mathbf{N}}$ such that $\|x_n\| \longrightarrow +\infty$ as n goes to $+\infty$ and such that $f(x_n) \leq \gamma_0$ for some $\gamma_0 \in \mathbf{R}$. This contradicts the fact that $lev_{\gamma_0} f$ is bounded. □

Let us recall the Riesz theorem: the bounded sets in a normed space are relatively compact iff the space has a finite dimension.

Corollary 3.2.2. *Let $X = \mathbf{R}^n$ equipped with the usual topology and let $f : X \longrightarrow \mathbf{R} \cup \{+\infty\}$. The following conditions are equivalent:*

(i) *f is coercive;*

(ii) *f is* inf*-compact.*

In infinite dimensional spaces, the topologies which are directly related to coercivity are the weak topologies (see Section 2.4).

3.2.6 Topological minimization theorems

In this section, unless otherwise specified (X, τ) is a general topological space.

Theorem 3.2.1. *Let (X, τ) be a topological space and let $f : X \longrightarrow \mathbf{R} \cup \{+\infty\}$ be an extended real-valued function which is τ-lower semicontinuous and τ-inf compact. Then, $\inf_X f > -\infty$ and there exists some $\bar{x} \in X$ which minimizes f on X:*

$$f(\bar{x}) \leq f(x) \quad \forall\, x \in X.$$

Because of the importance of this theorem, which is often referred to as the Weierstrass theorem, we give two different proofs of it, each of independent interest. Without any restriction we may assume that f is proper, that is, $f \not\equiv +\infty$.

FIRST PROOF. We want to prove that $\arg\min f \neq \emptyset$. We use formula (3.30), which relates $\arg\min f$ to the lower level sets of f.

$$\begin{aligned} \arg\min f &= \textstyle\bigcap_{\gamma > \inf_X f} lev_\gamma f \\ &= \textstyle\bigcap_{\gamma_0 > \gamma > \inf_X f} lev_\gamma f, \end{aligned}$$

where $\gamma_0 \in \mathbf{R}$ is taken arbitrary with $\gamma_0 > \inf_X f$. This comes from the fact that the sets $lev_\gamma f$ are decreasing with γ. The τ-lower semicontinuity of f implies that the sets $lev_\gamma f$ are closed for the topology τ. Moreover, for $\gamma_0 > \gamma > \inf_X f$ the sets $lev_\gamma f$ are nonempty and contained in $lev_{\gamma_0} f$ which is compact by the τ-inf compact property of f.

Therefore, we have a family $\{lev_\gamma f : \gamma_0 < \gamma < \inf_X f\}$ of nonempty closed subsets, contained in a fixed compact set, and which is decreasing with γ. For any finite subfamily $\{lev_{\gamma_i} f : i = 1, 2, \ldots, m\}$

$$\bigcap_{i=1,\ldots,m} lev_{\gamma_i} f = lev_\gamma f$$

with $\gamma = \inf\{\gamma_1, \ldots, \gamma_m\} > \inf_X f$, and hence

$$\bigcap_{i=1,\ldots,m} lev_{\gamma_i} f \neq \emptyset.$$

From the finite intersection property (which characterizes topological compact sets; it is obtained from the Heine–Borel property just by passing to the complement, and so replacing open sets by closed sets), we conclude that $\bigcap_{\gamma_0 > \gamma > \inf_X f} lev_\gamma f \neq \emptyset$.

SECOND PROOF. The proof we present now illustrates the direct method in the calculus of variations. It is the proof initiated by Hilbert and furter developed by Tonelli, which first introduces a minimizing sequence.

Let us observe that given a function $f : X \longrightarrow \mathbf{R} \cup \{+\infty\}$, one can always construct a minimizing sequence, that is, a sequence $(x_n)_{n \in \mathbf{N}}$ such that $f(x_n) \longrightarrow \inf_X f$ as $n \longrightarrow +\infty$. To do so, we just rely on the definition of the infimum of a family of real numbers:

if $\inf_X f > -\infty$, take $\inf_X f \leq f(x_n) \leq \inf_X f + 1/n$;

if $\inf_X f = -\infty$, take $f(x_n) \leq -n$.

Since f is proper, $\inf_X f < +\infty$ and for $n \geq 1$

$$\begin{aligned} f(x_n) &\leq \max\{\inf_X f + 1/n, -n\} \\ &\leq \max\{\inf_X f + 1, -1\} := \gamma_0. \end{aligned}$$

Note that $\gamma_0 > \inf_X f$, which implies $lev_{\gamma_0} f \neq \emptyset$ and

$$x_n \in lev_{\gamma_0} f \quad \forall\, n \geq 1.$$

Thus, the sequence $(x_n)_{n\in\mathbf{N}}$ is trapped in a lower level set of f which is compact for the topology τ (f is τ-inf compact). We follow the argument and assume that τ is metrizable. In the general topological case, one has to replace sequences by nets. So, we can extract a subsequence τ-converging to some $\bar{x} \in X$,

$$x_{n_k} \longrightarrow \bar{x}.$$

We have

$$\lim_k f(x_{n_k}) = \lim_n f(x_n) = \inf_X f.$$

By the τ-lower semicontinuity of f,

$$f(\bar{x}) \leq \lim_k \ f(x_{n_k}).$$

Hence,

$$f(\bar{x}) \leq \inf_X f,$$

which says both that $\inf_X f > -\infty$ since $f : X \longrightarrow \mathbf{R} \cup \{+\infty\}$ and

$$f(\bar{x}) \leq f(x) \quad \forall\, x \in X,$$

and the proof is complete. □

Remark 3.2.2. Indeed, the inf-compactness assumption can be slightly weakened by noticing that in the proof of Theorem 3.2.1, we just need to know that some lower level set of f is relatively compact. Notice that it is equivalent to assume that $lev_{\gamma_0} f$ is relatively compact or to assume that $lev_\gamma f$ is relatively compact for all $\gamma \leq \gamma_0$. So, we can formulate the following result.

Theorem 3.2.2. *Let (X, τ) be a topological space and $f : X \longrightarrow \mathbf{R} \cup \{+\infty\}$ an extended real-valued function which is τ-lower semicontinuous and such that for some $\gamma_0 \in \mathbf{R}$, $lev_{\gamma_0} f$ is τ-compact.*

Then $\inf_X f > -\infty$ and there exists some $\bar{x} \in X$ which minimizes f on X:

$$f(\bar{x}) \leq f(x) \quad \forall\, x \in X.$$

To illustrate the difference between Theorems 3.2.1 and 3.2.2, take $X = \mathbf{R}$ and $f(x) = \frac{x^2}{1+x^2}$. Then, $lev_\gamma f$ is compact for $\gamma < 1$, but $lev_1 f = \mathbf{R}$. Thus we can apply

Theorem 3.2.2 to conclude the existence of a minimizer (which is zero!), but Theorem 3.2.1 does not apply!

Corollary 3.2.3. *Let (X, τ) be a topological space and assume that $f : X \longrightarrow \mathbf{R} \cup \{+\infty\}$ is τ-lower semicontinuous. Then, for any compact subset K of (X, τ), there exists some $\bar{x} \in K$ such that*

$$f(\bar{x}) \leq f(x) \quad \forall x \in K.$$

PROOF. Take $g := f + \delta_K$. Since K is closed, δ_K is lower semicontinuous. So, g as a sum of two lower semicontinuous functions is still lower semicontinuous. The sublevel sets of g are contained in K, so g is τ-inf compact. Therefore, by Theorem 3.2.1 there exists some $\bar{x} \in X$ such that $g(\bar{x}) \leq g(x)$ for every $x \in X$, that is,

$$\begin{cases} f(\bar{x}) \leq f(x) \quad \forall\, x \in K, \\ \bar{x} \in K, \end{cases}$$

which completes the proof. □

The above statement is the unilateral version of the classical theorem which says that a continuous function achieves on any compact its minimum value and maximum value. If one is concerned only with the minimization problem, one just needs to consider lower semicontinuous functions.

Corollary 3.2.4. *Take $X = \mathbf{R}^N$ with the usual topology. Take $f : \mathbf{R}^N \longrightarrow \mathbf{R} \cup \{+\infty\}$ which is lower semicontinuous and coercive. Then, there exists some $\bar{x} \in \mathbf{R}^N$ such that*

$$f(\bar{x}) \leq f(x) \quad \forall x \in \mathbf{R}^N.$$

PROOF. Since $f : \mathbf{R}^N \longrightarrow \mathbf{R} \cup \{+\infty\}$ is coercive, it is inf-compact for the usual topology (Corollary 3.2.2). This combined with the lower semicontinuity of f implies the existence of a minimizer. □

Comments on the direct methods of the calculus of variations. Theorems 3.2.1 and 3.2.2 provide both a general existence result for the global minimization problem of an extended real-valued function and a method for solving such problems, originally introduced by Hilbert and further developed by Tonelli: when considering a minimization problem for a function $f : X \longrightarrow \mathbf{R} \cup \{+\infty\}$

$$\min\{f(x) \ : \ x \in X\},$$

one first constructs a minimizing sequence, which is a sequence $(x_n)_{n\in\mathbf{N}}$ such that

$$f(x_n) \longrightarrow \inf_X f \quad \text{as } n \longrightarrow +\infty.$$

This is always possible; at this point, we don't need any structure on X. Then, one has to establish that the sequence $(x_n)_{n\in\mathbf{N}}$ is relatively compact for some topology τ on X, and this is how the topology τ appears. This usually comes from some estimations on $(x_n)_{n\in\mathbf{N}}$

which follow from a coercivity property of f with respect to some norm on X. Then, one may use weak topologies or some compact embeddings to find the topology τ.

We stress that there is a great flexibility in this method. One may consider special minimizing sequences enjoying compactness properties which are not shared by the whole lower level sets. We will return to this important point. We just say here that it is the skill of the mathematician to find a minimizing sequence which is relatively compact for a topology τ which is as strong as possible.

Indeed, and this is the second point of the direct methods, one then has to verify that f is τ-lower semicontinuous. Clearly, the stronger the topology τ, the easier it is to verify the lower semicontinuity property.

As a general rule, inf-compactness and lower semicontinuity are two properties which are antagonist: if $\tau_1 > \tau_2$, then

$$f \ \tau_1 \quad \text{inf-compact} \implies f \ \tau_2 \quad \text{inf-compact}$$

while

$$f \ \tau_2\text{−lsc} \implies f \ \tau_1\text{−lsc}.$$

Thus, a balance with respect to these two properties determines the choice of a "good" topology τ (if it exists!).

As we will see, in some important situations, the function f fails to be τ-lower semicontinuous and there is no solution to the minimization problem of f. In such situations, it is still interesting to understand the behavior of the minimizing sequences. The following "relaxation result" gives a first general answer to this question.

Theorem 3.2.3. *Let (X, τ) be a topological space and let $f : X \longrightarrow \mathbf{R} \cup \{+\infty\}$ be an extended real-valued function. Let $(x_n)_{n\in\mathbf{N}}$ be a minimizing sequence for f, and suppose that a subsequence x_{n_k} τ-converges to some $\bar{x} \in X$. Then*

$$(cl_\tau f)(\bar{x}) \leq (cl_\tau f)(x) \quad \forall\, x \in X,$$

that is, $\bar{x}$ is a minimum point for $cl_\tau f$.

PROOF. Since $(x_n)_{n\in\mathbf{N}}$ is a minimizing sequence,

$$\lim_k \ f(x_{n_k}) = \lim_n f(x_n) = \inf_X f.$$

By Proposition 3.2.6, since $\bar{x} = \tau - \lim x_{n_k}$,

$$cl_\tau f(\bar{x}) \leq \lim_k \ f(x_{n_k}).$$

By Proposition 3.2.7,

$$\inf_X f = \inf_X cl_\tau f.$$

Hence

$$cl_\tau f(\bar{x}) \leq \lim f(x_{n_k}) = \lim f(x_n) = \inf_X f = \inf_X cl_\tau f,$$

that is,

$$(cl_\tau f)(\bar{x}) \leq (cl_\tau f)(x) \quad \forall x \in X. \qquad \square$$

We say that $\min\{(cl_\tau f)(x) : x \in X\}$ is the *relaxed problem* of the initial minimization problem of f over X.

3.2.7 Weak topologies and minimization of weakly lower semicontinuous functions

Until now, the basic ingredients used in the direct approach to minimization problems have been purely topological notions. We now assume that the underlying space (X, τ) is a vector space. To stress that fact, we denote it by V (like vector) and assume that V is a normed space, the norm of $v \in V$ being denoted by $\|v\|_V$ or $\|v\|$ when no confusion is possible.

The consideration of coercive functions $f : V \longrightarrow \mathbf{R} \cup \{+\infty\}$ leads naturally to study of the topological properties of the bounded subsets of V, and this is a basic reason for studying weak topologies on topological vector spaces.

We recall from Theorems 2.4.2 and 2.4.3 the following result.

Theorem 3.2.4. *In a reflexive Banach space V, the bounded sets are weakly relatively compact. Moreover, from any bounded sequence $(u_n)_{n\in\mathbf{N}}$ in V one can extract a weakly convergent subsequence.*

As a direct application of Proposition 3.2.8 and of the previous compactness result, we obtain that a coercive function on a reflexive Banach space is weakly inf-compact. By using the Weierstrass minimization Theorem 3.2.1 we obtain the following existence result.

Theorem 3.2.5. *Let V be a reflexive Banach space and $f : V \longrightarrow \mathbf{R} \cup \{+\infty\}$ an extended real-valued function which is coercive and weakly lower semicontinuous. Then there exists some $u \in V$ such that*

$$f(u) \leq f(v) \quad \forall v \in V.$$

The question that now naturally arises is to describe the class of functions which are weakly lower semicontinuous. This is where convexity plays a central role.

3.3 Convex minimization theorems

Let us first recall some definitions and elementary properties of extended real-valued convex functions.

3.3.1 Extended real-valued convex functions and weak lower semicontinuity

Definition 3.3.1. *Let V be a linear space and let $f : V \longrightarrow \mathbf{R} \cup \{+\infty\}$. Then, f is said to be convex if for each $u, v \in V$ and each $\lambda \in [0, 1]$ we have*

$$f(\lambda u + (1-\lambda)v) \leq \lambda f(u) + (1-\lambda) f(v).$$

Proposition 3.3.1. *Let V be a linear space and $f : V \longrightarrow \mathbf{R} \cup \{+\infty\}$. Then, f is convex if and only if its epigraph is a convex subset of $V \times \mathbf{R}$.*

PROOF. Let us first assume that f is convex. Fix (u, α) and (v, β) in epi f and $\lambda \in [0, 1]$. Since $\alpha \geq f(u)$ and $\beta \geq f(v)$, then $f(u)$ and $f(v)$ are finite and we have

$$\begin{aligned}\lambda\alpha + (1-\lambda)\beta &\geq \lambda f(u) + (1-\lambda) f(v) \\ &\geq f(\lambda u + (1-\lambda)v).\end{aligned}$$

This is equivalent to saying that $(\lambda u + (1-\lambda)v, \lambda\alpha + (1-\lambda)\beta) \in epif$, i.e.,

$$\lambda(u, \alpha) + (1-\lambda)(v, \beta) \in epi\ f,$$

and so $epi\ f$ is convex.

Conversely, let us assume that $epi\ f$ is convex. Fix $\lambda \in [0, 1]$. If either $f(u) = +\infty$ or $f(v) = +\infty$, since $0 \times (+\infty) = 0$, the inequality is clearly valid. So, let us assume that $f(u) < +\infty$ and $f(v) < +\infty$. The two points $(u, f(u))$ and $(v, f(v))$ are in $epi\ f$ and so is the segment joining this two points. In particular,

$$\lambda(u, f(u)) + (1-\lambda)(v, f(v)) \in epi\ f,$$

which is equivalent to saying that $f(\lambda u + (1-ł)v) \leq \lambda f(u) + (1-\lambda) f(v)$. □

So, it is equivalent to study convex sets or convex functions. Let us now recall the geometrical version of the Hahn–Banach theorem, which plays a basic role in convex analysis (see chapter 9).

Theorem 3.3.1 (Hahn–Banach separation theorem). *Let $(V, \|\cdot\|)$ be a normed linear space and suppose that C is a nonempty closed convex subset of V. Then, each point $u \notin C$ can be strongly separated from C by a closed hyperplane, which means*

$$\exists u^* \in V^*,\ \exists \alpha \in \mathbf{R} \text{ such that } \forall v \in C \quad u^*(v) \leq \alpha \text{ and } u^*(u) > \alpha.$$

This is equivalent to saying that C is contained in the closed half-space $\mathcal{H}_{\alpha,u^*} = \{v \in V\ :\ u^*(v) \leq \alpha\}$, whereas u is in its complement.

Corollary 3.3.1. *Let $(V, \|\cdot\|)$ be a normed linear space and let C be a nonempty closed convex subset of V. Then C is equal to the intersection of the closed half-spaces that contain it.*

Theorem 3.3.1 and Corollary 3.3.1 have important consequences with respect to topological (closedness) properties of convex sets: let us notice that by definition of the weak topology on V, any linear strongly continuous form is continuous for the weak topology which implies that any closed half-space

$$\mathcal{H}_{\alpha,u^*} = \{v \in V\ :\ u^*(v) \leq \alpha\}$$

is closed for the weak topology. So, any closed convex set, which is equal to an intersection of closed half spaces, is closed for the weak topology. Since for an arbitrary set, the reverse implication "closed for the weak topology" $\Longrightarrow$ "closed for the strong topology" is always true, we finally obtain the following result.

Theorem 3.3.2. *Let $(V, \|\cdot\|)$ be a normed linear space and C a nonempty convex subset of V. Then, the following statements are equivalent:*
(i) *C is closed for the norm topology of V;*
(ii) *C is closed for the weak topology of V.*

When translating the above theorem from sets to functions via the correspondence $f \longrightarrow epi\ f$ and recalling that

$$f \text{ convex} \iff epi\ f \text{ convex in } X \times \mathbf{R},$$
$$f\ \tau - lsc \iff epi\ f \text{ closed in } X \times \mathbf{R},$$

we obtain the following result.

Theorem 3.3.3. *Let V be a normed linear space and $f : V \longrightarrow \mathbf{R} \cup \{+\infty\}$ a convex proper function. The following statements are equivalent:*
(i) *f is lower semicontinuous for the norm topology on V;*
(ii) *f is lower semicontinuous for the weak topology on V.*

Of course it is the implication (i) $\Longrightarrow$ (ii) which is important. It tells us that a convex lower semicontinuous function is automatically weakly lower semicontinuous.

As a particular case, a convex continuous function is weakly lower semicontinuous.

3.3.2 Convex minimization in reflexive Banach spaces

In this section $(V, \|\cdot\|)$ is a reflexive Banach space. We can now state the following important result.

Theorem 3.3.4. *Let $(V, \|\cdot\|)$ be a reflexive Banach space and $f : V \longrightarrow \mathbf{R} \cup \{+\infty\}$ a convex, lower semicontinuous, and coercive function. Then, there exists $u \in V$ which minimizes f on V:*

$$f(u) \leq f(v) \quad \forall\, v \in V.$$

FIRST PROOF. Since f is coercive, its lower level sets are bounded in V and hence weakly relatively compact. So f is weakly inf-compact. Since f is convex lower semicontinuous, it is weakly lower semicontinuous. Then apply the Weierstrass minimization Theorem 3.2.1 to f with τ equal to the weak topology of V.

SECOND PROOF. We use the direct methods of the calculus of variations. Take $(u_n)_{n\in\mathbf{N}}$, a minimizing sequence for f, that is, $\lim_n f(u_n) = \inf_V f$. For n sufficiently large, $(u_n)_{n\in\mathbf{N}}$ remains in a fixed sublevel set of f, which is bounded by coercivity of f. Because the space V is reflexive, one can extract a weakly convergent subsequence $u_{n_k} \rightharpoonup u$. We have

$$\lim_k f(u_{n_k}) = \lim_n f(u_n) = \inf_V f.$$

The function f is convex and lower semicontinuous, so it is weakly lower semicontinuous and

$$f(u) \leq \liminf_k f(u_{n_k}).$$

It follows that

$$f(u) \leq \liminf_k f(u_{n_k}) = \lim_n f(u_n) = \inf_V f,$$

that is, $f(u) \leq f(v)\ \forall v \in V$. □

Remark 3.3.1. Concerning the question of uniqueness, we need to recall the notion of strict convexity: the function $f \longrightarrow \mathbf{R} \cup \{+\infty\}$ is said to be strictly convex if

$$\forall u \neq v \quad \forall \lambda \in]0,1[\quad f(\lambda u + (1-\lambda)v) < \lambda f(u) + (1-\lambda) f(v).$$

It is easily seen that the conclusion of Proposition 2.3.3 still holds for functions with values in $\mathbf{R} \cup \{+\infty\}$.

Example 3.3.1. Take for V a Hilbert space with norm $\|\cdot\|^2 = \langle \cdot, \cdot \rangle$. Then $\|\cdot\|^2$ is strictly convex: indeed, for $v_1, v_2 \in V$, $\lambda \in]0,1[$ we have

$$\begin{aligned} &\|\lambda v_1 + (1-\lambda)v_2\|^2 - \lambda\|v_1\|^2 - (1-\lambda)\|v_1\|^2 \\ &\quad = -\lambda(1-\lambda)\big[\|v_2\|^2 + \|v_1\|^2 - 2\langle v_1, v_2\rangle\big] \\ &\quad = -\lambda(1-\lambda)\|v_1 - v_2\|^2 \\ &\quad \leq 0. \end{aligned}$$

The above inequality becomes an equality iff $v_1 = v_2$.

As an application of the results above, we consider the problem of the best approximation in Hilbert spaces.

Theorem 3.3.5. *Let $(V, \|\cdot\|)$, be a Hilbert space with norm $\|\cdot\| = \sqrt{\langle \cdot, \cdot \rangle}$. Given $C \subset V$ a closed convex nonempty subset of V and $u_0 \in V$, there exists a unique element $\bar{u} \in C$ such that*

$$\|u_0 - \bar{u}\| \leq \inf_{v \in C} \|u_0 - v\|.$$

We have $\|u_0 - \bar{u}\| = d(u_0, C)$, that is, $\bar{u} \in C$ realizes the minimum of the distance between u_0 and C. We say that $\bar{u}$ is the projection of u_0 on C and we write

$$\bar{u} = proj_C\, u_0.$$

Moreover, $\bar{u}$ is characterized by the following property:

$$\begin{cases} \bar{u} \in C, \\ \langle u_0 - \bar{u}, v - \bar{u} \rangle \leq 0 \quad \forall\, v \in C. \end{cases}$$

Because of the importance of this result, we give two proofs of independent interest. The first relies on Theorem 3.3.4 and is straightforward, but recall that we have used the weak topology to prove Theorem 3.3.4. The second is a direct one and completely elementary (it does not use the weak topology) and can be the starting point for developing a theory of Hilbert spaces at a more elementary level without using the weak topologies.

First proof. First notice that it is equivalent to have

$$\|u_0 - \bar{u}\| \leq \|u_0 - v\| \qquad \forall\, v \in C$$

or

$$\|u_0 - \bar{u}\|^2 \leq \|u_0 - v\|^2 \qquad \forall\, v \in C.$$

So, we may say that our problem is equivalent to minimizing

$$f(v) = \|u_0 - v\|^2 + \delta_C(v)$$

over V. Clearly f is a convex function, as a sum of convex functions. It is strictly convex, because $\|\cdot\|^2$ is strictly convex and the sum of a convex and of a strictly convex function is still strictly convex. Since C is closed, δ_C is lower semicontinuous, and since $\|\cdot\|^2$ is continuous it is also lower semicontinuous, and f as a sum of two lower semicontinuous functions is still lower semicontinuous.

Finally $f(v) \geq \|u_0 - v\|^2$, which is clearly coercive, and so is f. So, f is strictly convex, lower semicontinuous, and coercive. It achieves its minimum at a unique point $\bar{u} \in C$.

SECOND PROOF. Let $(u_n)_{n\in\mathbf{N}}$ be a minimizing sequence, that is,

$$\begin{cases} u_n \in C, \\ \|u_0 - u_n\|^2 \longrightarrow \inf\{\|u_0 - v\|^2 \ :\ v \in C\} = d(u_0, C)^2. \end{cases}$$

Then, use the parallelogram equality: given $n, m \in \mathbf{N}$,

$$2\|u_0 - u_n\|^2 + 2\|u_0 - u_m\|^2 = \|u_n - u_m\|^2 + 4\left\|u_0 - \frac{(u_n + u_m)}{2}\right\|^2.$$

Since C is convex, $(u_n + u_m)/2$ belongs to C and

$$\left\|u_0 - \left(\frac{u_n + u_m}{2}\right)\right\|^2 \geq d(u_0, C)^2.$$

Hence

$$\|u_n - u_m\|^2 \leq 2\|u_0 - u_n\|^2 + 2\|u_0 - u_m\|^2 - 4d(u_0, C)^2.$$

It follows that

$$\limsup_{n,m\longrightarrow+\infty} \|u_n - u_m\|^2 \leq 0,$$

that is, the sequence $(u_n)_{n\in\mathbf{N}}$ is a Cauchy sequence. Since V is a Hilbert space, the sequence $(u_n)_{n\in\mathbf{N}}$ norm-converges to some element $\bar{u}$ which still belongs to C, because C is closed. Moreover,

$$\|u_0 - \bar{u}\|^2 = \lim_n \|u_0 - u_n\|^2 = d(u_0, C)^2.$$

Let us now prove the optimality condition for $\bar{u}$, that is,

$$\begin{cases} \bar{u} \in C, \\ \langle u_0 - \bar{u}, v - \bar{u}\rangle \leq 0 \quad \forall v \in C. \end{cases}$$

This property says that $\bar{u}$ is characterized by the following geometrical property: for any $v \in C$, the angle between the two vectors $u_0 - \bar{u}$ and $v - \bar{u}$ is greater than or equal to $\pi/2$. We will later derive this property from general subdifferential calculus. At the moment, we give a direct elementary proof of it.

For any $v \in C$, by convexity of C, the line segment $[\bar{u}, v]$ still belongs to C and hence, for all $t \in [0, 1]$,

$$w_t = tv + (1-t)\bar{u} \text{ belongs to } C.$$

By definition of $\bar{u}$

$$\begin{aligned}\|u_0 - \bar{u}\|^2 &\leq \|u_0 - tv - (1-t)\bar{u}\|^2 \\ &\leq \|(u_0 - \bar{u}) - t(v - \bar{u})\|^2.\end{aligned}$$

By developing this last expression, we obtain

$$2t\langle u_0 - \bar{u}, v - \bar{u}\rangle \ \leq \ t^2\|v - \bar{u}\|^2.$$

Divide by $t > 0$, and then let t go to zero to obtain

$$\langle u_0 - \bar{u}, v - \bar{u}\rangle \ \leq 0.$$

Conversely, let us prove that if $\overline{u}$ satisfies the optimality condition above, then

$$\|u_0 - \overline{u}\| \leq \inf_{v \in C} \|u_0 - v\|.$$

First notice that the optimality condition implies

$$\forall v \in C, \ \langle u_0 - v, v - \bar{u}\rangle \leq 0. \tag{3.35}$$

Indeed,

$$\begin{aligned}\langle u_0 - v, v - \bar{u}\rangle &= \langle u_0 - \bar{u} + \bar{u} - v, v - \bar{u}\rangle \\ &= \langle u_0 - \bar{u}, v - \bar{u}\rangle - \|\bar{u} - v\|^2 \\ &\leq 0.\end{aligned}$$

We then have

$$\begin{aligned}\|u_0 - \bar{u}\|^2 &= \langle u_0 - \bar{u}, u_0 - v + v - \bar{u}\rangle \\ &\leq \langle u_0 - \bar{u}, u_0 - v\rangle \\ &= \langle u_0 - v + v - \bar{u}, u_0 - v\rangle \\ &= \|u_0 - v\|^2 + \langle u_0 - v, v - \bar{u}\rangle \\ &\leq \|u_0 - v\|^2,\end{aligned}$$

where we have used the optimality condition in the first inequality and relation (3.35) in the last one. □

Corollary 3.3.2. *When $C = W$ is a closed subspace, then $\bar{u} = proj_W u_0$ is characterized by*

$$\begin{cases} \bar{u} \in W, \\ u_0 - \bar{u} \in W^{\perp}, \end{cases}$$

that is,

$$u_0 = (u_0 - \bar{u}) + \bar{u} \in W^{\perp} + W.$$

This is the orthogonal decomposition of $V = W \oplus W^{\perp}$ as the sum of two orthogonal subspaces. Moreover, the projection operator $proj : V \longrightarrow W$ is linear.

Proposition 3.3.2. *When V is a Hilbert space and C is a closed convex nonempty subset of V, the projection operator $V \longrightarrow C$ which associates to each $u \in V$ its projection $proj_C u$ on C is a contraction:*

$$\forall\, u, v \in V \quad \|proj_C u - proj_C v\| \leq \|u - v\|.$$

PROOF. We have

$$\langle u - proj_C u, z - proj_C u\rangle \leq 0 \quad \forall\, z \in C,$$
$$\langle v - proj_C v, z - proj_C v\rangle \leq 0 \quad \forall\, z \in C.$$

Take $z = proj_C v$ in the first inequality and $z = proj_C u$ in the second one. Summing up, we obtain

$$\langle proj_C v - proj_C u, u - proj_C u - v + proj_C v\rangle \leq 0,$$

that is,

$$\|proj_C v - proj_C u\|^2 \leq \langle proj_C v - proj_C u, v - u\rangle.$$

By the Cauchy–Schwarz inequality, it follows that $\|proj_C v - proj_C u\| \leq \|v - u\|$. □

We end this section by remarking that the Hilbertian structure plays a fundamental role in the previous results for the best approximation. The existence of the best approximation $\bar{u}$ still holds true when $(V, \|\cdot\|)$ is a reflexive Banach space. But when the space $(V, \|\cdot\|)$ is no longer reflexive, even the existence of $\bar{u}$ may fail to be true, as shown by the following example.

Example 3.3.2. Take $V = \mathbf{C}([0, 1]; \mathbf{R})$ equipped with the sup norm

$$\forall v \in V \quad \|v\|_\infty = \sup\{|v(t)| \,:\, t \in [0, 1]\}.$$

Then $(V, \|\cdot\|_\infty)$ is a Banach space which is not reflexive. Indeed, this is a consequence of the fact that the existence of a projection may fail to be true in this space: take

$$C = \left\{ v \in V \,:\, \int_0^{1/2} v(t)dt - \int_{1/2}^1 v(t)dt = 1 \right\}.$$

Clearly C is a closed convex nonempty subset of V (it is in fact a closed hyperplane). One can easily verify that $d(0, C) = 1$, but there is no element $v \in C$ such that $\|v\|_\infty = 1$. Indeed, if $v \in C$,

$$1 = \int_0^{1/2} v(t)dt - \int_{1/2}^1 v(t)dt \leq \frac{1}{2} \int_0^{1/2} |v(t)|dt + \frac{1}{2} \int_{1/2}^1 |v(t)|dt \leq \|v\|_\infty.$$

Hence $d(0, C) = \inf\{\|v\|_\infty \,:\, v \in C\} \geq 1$, and it is not difficult to show that $d(0, C) = 1$. Suppose now that for some $v \in C$, $\|v\|_\infty = 1$. Since

$$\left| \int_0^{1/2} v(t)\, dt \right| \leq \frac{1}{2}, \qquad \left| \int_{1/2}^1 v(t)\, dt \right| \leq \frac{1}{2},$$

we necessarily have $\int_0^{1/2} v(t)\, dt = \frac{1}{2}$ and $\int_{1/2}^1 v(t)\, dt = -\frac{1}{2}$.

So $\int_0^{1/2}(1-v(t))dt = 0$. Since $1-v(x) \geq 0$, this implies $v(t) \equiv 1$ on $[0, \frac{1}{2}]$. Similarly $v(t) \equiv -1$ on $[\frac{1}{2}, 1]$, a clear contradiction with the fact that v has to be continuous.

3.4 Ekeland's ε-variational principle

The so-called ε-variational principle was introduced by Ekeland in 1972 [128], [129], [130]. It is a general powerful tool in variational analysis and optimization which can be traced back to a maximality result for a partial ordering introduced by Bishop and Phelps in 1962 [70].

Ekeland's ε-variational principle asserts the existence of minimizing sequences of a particular kind. Not only do they approach the infimal value of the minimization problem, but they also simultaneously satisfy the first order necessary conditions up to any desired approximation. In many instances, this makes this variational principle play a key role in the application of the direct method. Indeed, Ekeland's ε-variational principle has known a considerable success with applications in a wide variety of topics in nonlinear analysis and optimization (critical point theory, geometry of Banach spaces, etc.).

In the last two decades, there has been increasing evidence that Ekeland's ε-variational principle bears close connections with dissipative dynamical systems (dynamical systems with entropy). Indeed, solutions provided by the ε-variational principle can be seen as stable equilibria of such dynamics [92], [46]. In particular, the recent model in dynamical decision theory introduced by Attouch and Soubeyran [37] will serve as a guideline in the proof and interpretation of the results.

3.4.1 Ekeland's ε-variational principle and the direct method

Let us start with the following formulation of the Ekeland's ε-variational principle.

Theorem 3.4.1 (Ekeland). *Let (X, d) be a complete metric space and $f : X \longrightarrow \mathbf{R} \cup \{+\infty\}$ an extended real-valued function which is lower semicontinuous and bounded below ($\inf_X f > -\infty$). Then, for each $\varepsilon > 0$, there exists some $x_\varepsilon \in X$ which satisfies the two following properties:*

$$\begin{cases} \text{(i)} & \inf\limits_X f \leq f(x_\varepsilon) \leq \inf\limits_X f + \varepsilon, \\ \text{(ii)} & f(x) \geq f(x_\varepsilon) - \varepsilon d(x, x_\varepsilon) \ \forall\, x \in X. \end{cases}$$

Let us first comment on this result and show some direct consequences of it. (Its proof is postponed to the next section.) Condition (ii) has a clear interpretation when $f : X \longrightarrow \mathbf{R}$ is Gateaux differentiable on a Banach space $(X, \|.\|)$. It can be seen as a unilateral nonsmooth version of the condition $\|Df(x_\varepsilon)\|_* \leq \varepsilon$. Let us start with the definition of the Gateaux differentiability property of f at x_ε. For any $\xi \in X$, with $\|\xi\| = 1$ and any $t > 0$,

$$f(x_\varepsilon + t\xi) = f(x_\varepsilon) + t\langle Df(x_\varepsilon), \xi\rangle + o(t).$$

By taking $x = x_\varepsilon + t\xi$ in (ii) and using the above equality, we obtain

$$f(x_\varepsilon) + t\langle Df(x_\varepsilon), \xi\rangle + o(t) \geq f(x_\varepsilon) - \varepsilon t.$$

Let us simplify, divide by $t > 0$, and let $t \to 0^+$. We obtain

$$\langle Df(x_\varepsilon), \xi \rangle \geq -\varepsilon.$$

Changing ξ into $-\xi$ yields

$$|\langle Df(x_\varepsilon), \xi \rangle| \leq \varepsilon.$$

This being true for any $\xi \in X$ with $\|\xi\| \leq 1$, we finally obtain $\|Df(x_\varepsilon)\|_* \leq \varepsilon$.

We can summarize this result in the following

Corollary 3.4.1. *Let $(X, \|.\|)$ be a Banach space and $f : X \longrightarrow \mathbf{R}$ a real-valued function which is lower semicontinuous, Gateaux differentiable, and bounded below. Then, for each $\varepsilon > 0$, there exists some $x_\varepsilon \in X$ such that*

$$\begin{cases} \inf\limits_X f \leq f(x_\varepsilon) \leq \inf\limits_X f + \varepsilon, \\ \|Df(x_\varepsilon)\|_* \leq \varepsilon. \end{cases}$$

The above result asserts the existence of minimizing sequences $(x_n)_{n\in\mathbf{N}}$ of particular type: take $x_n = x_{\varepsilon_n}$ with $\varepsilon_n \to 0^+$; then

$$\begin{cases} f(x_n) \to \inf_X f \text{ as } n \to +\infty, \\ Df(x_n) \to 0 \text{ in } X^* \text{ as } n \to +\infty. \end{cases}$$

Application of the direct method when dealing with such particular minimizing sequences leads us naturally to introduce the so-called Palais–Smale compactness condition for a functional f.

Definition 3.4.1. *Let $(X, \|.\|)$ be a Banach space. We say that a $\mathbf{C}^1$ function $f : X \longrightarrow \mathbf{R}$ satisfies the Palais–Smale condition if every sequence $(x_n)_{n\in\mathbf{N}}$ in X which satisfies*

$$\sup_n |f(x_n)| < +\infty \text{ and } Df(x_n) \to 0 \text{ in } X^* \text{ as } n \to +\infty$$

possesses a convergent subsequence (for the topology of the norm of X).

As an immediate consequence of Corollary 3.4.1, we obtain the next theorem.

Theorem 3.4.2. *Let $(X, \|.\|)$ be a Banach space and $f : X \longrightarrow \mathbf{R}$ a $\mathbf{C}^1$ function which satisfies the Palais–Smale condition and which is bounded below. Then the infimum of f on X is achieved at some point $\bar{x} \in X$ and $\bar{x}$ is a critical point of f, i.e., $Df(\bar{x}) = 0$.*

Proof. Using Corollary 3.4.1 of the Ekeland's ε-variational principle, we have the existence of a sequence $(x_n)_{n\in\mathbf{N}}$ which satisfies

$$f(x_n) \to \inf_X f, \ Df(x_n) \to 0.$$

Since $\inf_X f \in \mathbf{R}$, we have $\sup_n |f(x_n| < +\infty$, and the sequence $(x_n)_{n\in\mathbf{N}}$ satisfies the hypotheses of the Palais–Smale condition. Hence, one can extract a convergent subsequence

$x_{n_k} \to \bar{x}$. By using the continuity properties of f and Df, one gets at the limit $f(\bar{x}) = \inf_X f$ and $Df(\bar{x}) = 0$. □

Judicious applications of this kind of result (based on the Palais–Smale compactness condition) provide existence results for critical points, not only local minima or maxima but also saddle points. One of the most celebrated of these results is the mountain pass theorem of Ambrosetti and Rabinowitz [12]. For further results in this direction, see [46] or [121].

Indeed, it turns out that when f is not necessarily smooth, condition (ii) is a convenient formulation of an ε-approximate optimality condition. The key is the following observation: Property (ii) of Theorem 3.4.1 just expresses that x_ε is an exact solution of the perturbed minimization problem $(\mathcal{P}_\varepsilon)$

$$(\mathcal{P}_\varepsilon) \qquad \inf \big\{ f(x) + \varepsilon d(x, x_\varepsilon) : x \in X \big\}.$$

In the particular and important case where f is convex and lower semicontinuous on a Banach space, one gets the following result.

Corollary 3.4.2. *Let $(X, \|.\|)$ be a Banach space and $f : X \longrightarrow \mathbf{R} \cup \{+\infty\}$ an extended real-valued function which is convex, lower semicontinuous, proper ($f \not\equiv +\infty$), and bounded below. Then, for each $\varepsilon > 0$ there exist $x_\varepsilon \in X$ and $x_\varepsilon^* \in X^*$ such that*

$$\begin{cases} \inf\limits_X f \le f(x_\varepsilon) \le \inf\limits_X f + \varepsilon, \\ x_\varepsilon^* \in \partial f(x_\varepsilon), \ \|x_\varepsilon\|_* \le \varepsilon, \end{cases}$$

where $\partial f(x_\varepsilon)$ is the subdifferential of f at x_ε.

Proof. We use standard tools from convex subdifferential calculus (see chapter 9). Since x_ε minimizes the closed convex proper function $x \mapsto \varphi(x) := f(x) + \varepsilon\|x - x_\varepsilon\|$, we have $\partial\varphi(x_\varepsilon) \ni 0$. The norm being a continuous function in X, the additivity rule for the subdifferential calculus holds (Theorem 9.5.4) and we have

$$\partial f(x_\varepsilon) + \varepsilon \mathbf{B}(0, 1) \ni 0.$$

Equivalently, there exists some $x_\varepsilon^* \in \partial f(x_\varepsilon)$ with $\|x_\varepsilon^*\|_* \le \varepsilon$. □

3.4.2 A dynamical approach and proof of Ekeland's ε-variational principle

Ekeland's ε-variational principle bears close connection with dissipative dynamical systems. This fact was recognized by Brezis and Browder [92]: "A general ordering principle"; Aubin and Ekeland [46]: "walking in complete metric spaces"; and Zeidler [227]: "The abstract entropy principle." More recently, the importance of this principle in the modelization of dynamical decision with bounded rationality was put to the fore by Attouch and Soubeyran [37]. This cognitive interpretation will serve as a guideline throughout this section.

The central concept in the dynamical approach to Ekeland's ε-variational principle is the following partial ordering relation.

Definition 3.4.2. *Let (X, d) be a metric space and $f : X \longrightarrow \mathbf{R} \cup \{+\infty\}$ an extended real-valued function which is proper ($f \not\equiv +\infty$). Let us introduce the following partial ordering on X:*

$$y \succeq_s x \iff f(y) + d(x, y) \leq f(x).$$

We call it the marginal satisficing relation. We write

$$\begin{aligned} S(x) &= \left\{ y \in X : y \succeq_s x \right\} \\ &= \left\{ y \in X : f(y) + d(x, y) \leq f(x) \right\} \end{aligned}$$

the set of elements of X which satisfy this ordering with respect to x.

Let us introduce some elements of decision theory that allow us to interpret this relation in a natural and intuitive way. Space X is the decision or performance space (the state space in physics). It is supposed that to each element $x \in X$ the agent is able to attribute a value or valence $f(x) \in \mathbf{R} \cup \{+\infty\}$ which measures the quality of the decision or performance x (the value $+\infty$ allows us to take account of the constraints). For example, when performing x, $f(x)$ measures how far the agent is from a given goal. In our context, $f(x)$ measures the dissatisfaction of the agent who, making $x \in X$, is faced with a problem which is not completely solved. Thus the agent is willing to reduce its dissatisfaction and make $f(x)$ as small as possible. The connection with the traditional formulation in decision sciences is obtained by taking $f(x) = \overline{g} - g(x)$, where g is a classical utility or gain function and $\overline{g}$ is a desirable level of resolution of the problem (for example $\overline{g} = \sup_{x \in X} g(x)$). We choose this presentation to fit well with the classical formulation of variational principles in mathematics and physics and, in our situation, with the usual formulation of Ekeland's ε-variational principle.

The classical decision theory deals with perfectly rational agents who have immediate and free access to a global knowledge of their environment, and correspondingly minimize their value function f on X.

Modelization of decision processes in a complex real world requires us to introduce some further notions. Following Simon's [207] pioneering work in decision theory and bounded rationality, one needs to modelize the ability and difficulty of the agent to move and decide in a complex environment. A major difficulty for the agent is that it needs to explore its environment and get enough information to make further decisions. In this context, making decision becomes a dynamical process which at each step $k = 1, 2, \ldots$ is based on the following question: *Is it worthwhile for the agent to pass from a given state $x_k \in X$ (performance, decision, allocation) at time t_k into a further state $x_{k+1} \in X$ at time t_{k+1}?*

A key ingredient of the modelization of this balance between the advantage for the agent to pass from x to y and the possibility and difficulty of realizing it is the notion of cost to change. Following Attouch and Soubeyran [37], one introduces for any x and y in X, $c(x, y) \geq 0$, which is the cost to pass (change, move) from x to y. In our context, we assume that $c(x, y) \geq \theta d(x, y)$, where d is a metric on X and $\theta > 0$, is a unitary cost to move. This expresses that the cost to move is high for small deplacements (by contrast with $c(x, y) = d(x, y)^2$, for example!). This metric d modelizes the difficulty for the agent to pass from a state x to a further state y. This is where the metric d in the cognitive interpretation of Ekeland's ε-variational principle appears! This is a rich concept

which covers several aspects: there are costs to explore and get information (this comes from limited time and energy available for the agent), physical costs to move, and also costs with psychological and cognitive interpretation (dissimilarity costs, cost to quit a routine and enter into an other one, excitation and inhibition costs). From now on, we consider the particular situation $c(x, y) = d(x, y)$, which is enough for our purpose.

We now have the two terms of the balance: on one hand, the marginal gain $f(x)-f(y)$, and on the other hand, the cost to change $d(x, y)$. Precisely, the marginal satisficing relation $y \succeq_s x$ says that it is worthwhile for the agent to pass from x to y if the expected marginal gain ("reduction of dissatisfaction") $f(x) - f(y)$ is greater than or equal to the cost to pass from x to y:

$$f(x) - f(y) \geq d(x, y).$$

Let us examine some properties of the marginal satisficing relation $\succeq_s$.

Proposition 3.4.1. *The marginal satisficing relation $\succeq_s$ is a partial ordering relation on $\mathrm{dom}\, f$. An element $\bar{x} \in X$ is maximal with respect to this order iff for all $x \in X$, $x \neq \bar{x}$ one has*

$$f(\bar{x}) < f(x) + d(\bar{x}, x).$$

Proof. Clearly $\succeq_s$ is reflexive; we have $x \succeq_s x$ for all $x \in X$, because $d(x, x) = 0$.

Let us verify that $\succeq_s$ is antisymetric. Suppose $y \succeq_s x$ and $x \succeq_s y$. We thus have

$$f(x) \geq f(y) + d(x, y) \text{ and } f(y) \geq f(x) + d(x, y).$$

Let us add these two inequalities and simplify the resulting expression. (At this point, note that it is important to consider x and y in $\mathrm{dom}\, f$.) One obtains $2d(x, y) \leq 0$, and hence $x = y$.

Let us now verify that $\succeq_s$ is transitive. Suppose that $z \succeq_s y$ and $y \succeq_s x$. We have

$$\begin{aligned} f(y) &\geq f(z) + d(y, z), \\ f(x) &\geq f(y) + d(x, y). \end{aligned}$$

By adding the two above inequalities, and using again that $x, y, z \in \mathrm{dom}\, f$, we obtain

$$\begin{aligned} f(x) &\geq f(z) + d(x, y) + d(y, z) \\ &\geq f(z) + d(x, z). \end{aligned}$$

In the above inequality we used the triangle inequality property satisfied by the metric d. Hence, $z \succeq_s x$ and $\succeq_s$ is transitive.

Let us now express that an element $\bar{x}$ of X is maximal with respect to the partial ordering relation $\succeq_s$. This means that for any $x \in X$, the following implication holds:

$$x \succeq_s \bar{x} \Longrightarrow x = \bar{x}, \quad \text{i.e.,}$$

$$\forall x \in X,\ x \neq \bar{x}, \quad f(\bar{x}) < f(x) + d(\bar{x}, x),$$

which completes the proof. □

Cognitive interpretation of maximal elements for $\succeq_s$. Let us show that maximal elements of the satisficing relation $\succeq_s$ can be interpreted as stable routines of the corresponding "worthwhile to move" (marginal satisficing) dynamical system. In our cognitive version, $\bar{x} \in X$ is said to be a stable routine if, starting from $\bar{x}$, the agent prefers to stay at $\bar{x}$ than to move from $\bar{x}$ to x for all x different from $\bar{x}$. Let us make this precise and consider, when starting from $\bar{x}$, the two following possibilities:

(a) if the agent chooses to stay at $\bar{x}$, his gain (dissatisfaction in our case) will be

$$f(\bar{x}) + d(\bar{x}, \bar{x}) = f(\bar{x});$$

(b) if the agent considers to move from $\bar{x}$ to x, his gain (dissatisfaction) is after moving (one adds two unsatisfactions, the cost to move $d(\bar{x}, x)$ and the dissatisfaction attached to x):

$$f(x) + d(\bar{x}, x).$$

Thus, the agent is willing to stay at $\bar{x}$ and rejects any move from $\bar{x}$ to x for all $x \neq \bar{x}$ iff

$$f(\bar{x}) < f(x) + d(\bar{x}, x),$$

which, by Proposition 3.4.1, just expresses that $\bar{x}$ is maximal for $\succeq_s$.

Therefore, Ekeland's ε-variational principle can be reformulated as an existence result of a maximal element for the partial ordering $\succeq_s$. Let us make this precise in the following statement.

Theorem 3.4.3. *Let us assume that (X, d) is a complete metric space and $f : X \longrightarrow \mathbf{R} \cup \{+\infty\}$ is an extended real-valued function which is lower semicontinuous and bounded below. Then for any $x_0 \in \operatorname{dom} f$ there exists some $\bar{x} \in X$ which satisfies the two following properties:*

$$\begin{cases} \text{(i)} & \bar{x} \succeq_s x_0, \\ \text{(ii)} & \bar{x} \text{ is maximal with respect to the partial ordering } \succeq_s . \end{cases}$$

Before proving Theorem 3.4.3, let us show how Ekeland's ε-variational principle can be derived from it: given $\varepsilon > 0$, take $x_0 \in \operatorname{dom} f$ such that

$$\inf_X f \leq f(x_0) \leq \inf_X f + \varepsilon.$$

Then, let us apply Theorem 3.4.3 with the metric εd and the corresponding satisficing relation

$$y \succeq_s x \iff f(y) + \varepsilon d(x, y) \leq f(x).$$

Theorem 3.4.3 asserts the existence of $\bar{x}_\varepsilon$ such that $\bar{x}_\varepsilon \succeq_s x_0$ and $\bar{x}_\varepsilon$ maximal with respect to $\succeq_s$. The property $\bar{x}_\varepsilon \succeq_s x_0$ implies

$$\begin{aligned} f(\bar{x}_\varepsilon) &\leq f(x_0) - \varepsilon d(\bar{x}_\varepsilon, x_0) \\ &\leq f(x_0) \\ &\leq \inf_X f + \varepsilon. \end{aligned}$$

On the other hand, by Proposition 3.4.1 and the maximality property of $\bar{x}_\varepsilon$, we have

$$\forall x \neq \bar{x}_\varepsilon \quad f(\bar{x}_\varepsilon) < f(x) + \varepsilon d(\bar{x}_\varepsilon, x).$$

Thus, $\bar{x}_\varepsilon$ satisfies the two desired properties (i) and (ii) of Theorem 3.4.1.

We are going to prove Theorem 3.4.3 (and hence Ekeland's ε-variational principle 3.4.1) by using the dynamical system, which is naturally associated to the marginal satisficing relation.

Definition 3.4.3. *A trajectory $(x_k)_{k\in\mathbf{N}}$ of the marginal satisficing dynamics is a sequence of elements x_k of X such that*

$$(S) \qquad x_{k+1} \in S(x_k) \quad \forall\, k = 0, 1, 2, \ldots,$$

where S is the marginal satisficing relation. Equivalently, we have

$$x_0 \preceq_s x_1 \preceq_s x_2 \preceq_s \cdots \preceq_s x_k \preceq_s x_{k+1} \preceq_s \cdots,$$

that is,

$$f(x_{k+1}) + d(x_k, x_{k+1}) \leq f(x_k) \;\forall\, k = 0, 1, 2, \ldots.$$

Let us establish some general properties of the trajectories of the above dynamical system (S). We are mostly concerned with the asymptotical behavior as $k \to +\infty$ of these trajectories.

Proposition 3.4.2. *Let (X, d) be a metric space and $f : X \longrightarrow \mathbf{R} \cup \{+\infty\}$ an extended real-valued function which is proper and bounded below. Take any trajectory $(x_k)_{k\in\mathbf{N}}$ of (S) starting from some $x_0 \in \operatorname{dom} f$*

$$x_0 \preceq_s x_1 \preceq_s x_2 \preceq_s \cdots \preceq_s x_k \preceq_s x_{k+1} \preceq_s \cdots.$$

Then, the following properties hold:

(i) *$(f(x_k))_{k\in\mathbf{N}}$ decreases with k, and $f(x_k) \to \inf_k f(x_k) \in \mathbf{R}$ when $k \to +\infty$;*

(ii) *The sequence $(x_k)_{k\in\mathbf{N}}$ satisfies $\sum_{k=0}^{+\infty} d(x_k, x_{k+1}) < +\infty$. Hence, it is a Cauchy sequence in (X, d). When (X, d) is a complete metric space, the sequence $(x_k)_{k\in\mathbf{N}}$ converges in (X, d) to some $\bar{x} \in X$. Moreover, when f is lower semicontinuous, we have $\bar{x} \succeq_s x_k$ for all $k \in \mathbf{N}$.*

Proof. (i) For any $k \in \mathbf{N}$, by definition of $\preceq_s$

$$\begin{aligned} f(x_{k+1}) &\leq f(x_{k+1}) + d(x_k, x_{k+1}) \\ &\leq f(x_k). \end{aligned}$$

We have used $d(x_k, x_{k+1}) \geq 0$, which expresses that changes are costly. Therefore, the sequence $(f(x_k))_{k\in\mathbf{N}}$ is decreasing. Since

$$-\infty < \inf_X f \leq f(x_k) \leq f(x_0) < +\infty,$$

we have $f(x_k) \downarrow \inf_k f(x_k)$, which is a finite real number.

(ii) Let us write the inequality $f(x_{k+1})+d(x_k, x_{k+1}) \leq f(x_k)$ for $k = 0, 1, \ldots, n-1$,

$$\begin{aligned} f(x_1) + d(x_0, x_1) &\leq f(x_0) \\ &\cdot \\ &\cdot \\ &\cdot \\ f(x_n) + d(x_{n-1}, x_n) &\leq f(x_{n-1}). \end{aligned}$$

Then, we sum these inequalities and simplify the resulting expression (note that $f(x_k) \in \mathbf{R}$ for all $k \in \mathbf{N}$). We obtain

$$f(x_n) + \sum_{k=0}^{n-1} d(x_k, x_{k+1}) \leq f(x_0).$$

Let us now use the minorization $f(x_n) \geq \inf_X f$ and the assumption $\inf_X f > -\infty$. We thus have

$$\sum_{k=0}^{n-1} d(x_k, x_{k+1}) \leq f(x_0) - \inf_X f < +\infty.$$

This being true for any $n \in \mathbf{N}$, we deduce

$$\sum_{k=0}^{+\infty} d(x_k, x_{k+1}) \leq f(x_0) - \inf_X f < +\infty.$$

Note that this holds true, just by assuming that (X, d) is a metric space. Then, by a classical argument, when (X, d) is a complete metric space, this implies the convergence of the sequence $(x_k)_{k\in\mathbf{N}}$ in (X, d). To see this, write the triangle inequality

$$\begin{aligned} d(x_n, x_{n+p}) &\leq \sum_{k=n}^{n+p-1} d(x_k, x_{k+1}) \\ &\leq \sum_{k=n}^{+\infty} d(x_k, x_{k+1}), \end{aligned}$$

which tends to zero as $n \to +\infty$. Hence $(x_k)_{k\in\mathbf{N}}$ is a Cauchy sequence in (X, d) which implies its convergence when (X, d) is a complete metric space. Let

$$x_k \to \bar{x} \text{ in } (X, d) \text{ as } k \to +\infty.$$

Let us prove that $\bar{x} \succeq_s x_n$ for all $n \in \mathbf{N}$. We have $x_k \succeq_s x_n$ for all $k \geq n$ (by transitivity of $\succeq_s$), i.e.,

$$f(x_k) + d(x_k, x_n) \leq f(x_n) \quad \forall\, k \geq n.$$

Let us fix $n \in \mathbf{N}$ and let $k \to +\infty$ in this inequality. Since $x_k \to \bar{x}$ in (X, d), by using the lower semicontinuity property of f (up to now we have not used it!), we obtain

$$f(\bar{x}) + d(\bar{x}, x_n) \leq f(x_n),$$

that is, $\bar{x} \succeq_s x_n$. □

Our objective is to prove the existence of a trajectory $(x_k)_{k\in\mathbf{N}}$ of the marginal satisficing dynamics (S) which converges to a maximal element $\bar{x}$ for $\succeq_s$. So doing, we will have $\bar{x} \succeq_s x_k$ for all $k \in \mathbf{N}$ and hence $\bar{x} \succeq_s x_0$, which combined with the maximality of $\bar{x}$ is precisely the claim of Theorem 3.4.3.

In this perspective, to consider an arbitrary trajectory of (S) does not provide enough information: note that $x_k \equiv x_0$ for all $k \in \mathbf{N}$ is a trajectory of (S)! The dynamical system (S) modelizes a general rejection decision mechanism. We are now going to consider some trajectory of (S) which describes the decision process of a motivated agent. This means that at each step, the agent is willing to substantially improve his performance. This is a rich modelization subject involving some optimization aspects. In this perspective, the notion of aspiration index $m(x)$, which is defined in the next statement, plays an important role.

Lemma 3.4.1. *For any $x \in X$,*

$$\textit{diam } S(x) \leq 2(f(x) - m(x)),$$

where

$$m(x) = \inf\left\{f(y) : y \in S(x)\right\} = \inf\left\{f(y) : y \succeq_s x\right\}$$

is called the aspiration index of the agent at x.

Proof. The proof is an immediate consequence of the definition of $y \in S(x)$:

$$y \in S(x) \iff f(y) + d(x, y) \leq f(x).$$

Noticing that for $y \in S(x)$ we have $f(y) \geq m(x)$, we deduce

$$\forall y \in S(x) \qquad d(x, y) \leq f(x) - m(x).$$

As a consequence, for any $y, z \in S(x)$,

$$\begin{aligned} d(y, z) &\leq d(y, x) + d(x, z) \\ &\leq 2(f(x) - m(x)), \end{aligned}$$

and diam $S(x) \leq 2(f(x) - m(x))$. □

The aspiration index of the agent at x, say, $m(x)$, measures the gap between its present level of satisfaction at x and the maximum level of satisfaction that it can hope to obtain at a further step. Note that $m(x)$ is, in general, not known by the agent who is not able to explore all of $S(x)$. The cognitive model says that if the agent is motivated and is willing to explore enough at each step (and pay corresponding exploration costs!), then it knows a sufficiently good approximation of $m(x)$ and the process converges to a stable routine.

We are going to consider trajectories $(x_k)_{k\in\mathbf{N}}$ corresponding to a motivated agent who satisfies enough at each step. As an example, we consider that at each step k, the agent satisfies and fills a given fraction $\lambda \in]0,1[$ of the gap between $f(x_k)$ and $m(x_k)$: thus $x_{k+1} \succeq_s x_k$ and

$$f(x_{k+1}) \leq \lambda m(x_{k+1}) + (1-\lambda) f(x_k) = f(x_k) - \lambda[f(x_k) - m(x_k)].$$

We now have all the ingredients to state a dynamical, cognitive version and proof of Ekeland's ε-variational principle [37].

Theorem 3.4.4 (Attouch and Soubeyran). *Let (X,d) be a complete metric space and $f : X \longrightarrow \mathbf{R}\cup\{+\infty\}$ an extended real-valued function which is lower semicontinuous and bounded below.*

(a) *Then, for any $x_0 \in \operatorname{dom} f$, there exists a trajectory $(x_k)_{k\in\mathbf{N}}$ of the marginal satisficing dynamical system (S),*

$$x_0 \preceq_s x_1 \preceq_s x_2 \preceq_s \ldots \preceq_s x_k \preceq_s x_{k+1} \preceq_s \ldots,$$

which converges in (X,d) to some $\bar{x} \in X$ which is a maximal element for the partial ordering $\succeq_s$.

(b) *Such trajectory can be obtained by satisficing enough at each step, for example, given some positive parameter $0 < \lambda < 1$, by taking at each step*

$$\begin{cases} x_{k+1} \succeq_s x_k \text{ and} \\ f(x_{k+1}) \leq f(x_k) - \lambda[f(x_k) - m(x_k)], \end{cases}$$

where $m(.)$ is the aspiration index : $m(x) = \inf\left\{f(y) : y \succeq_s x\right\}$.

Proof. Take a trajectory $(x_k)_{k\in\mathbf{N}}$ of (S) which satisfies enough at each step. One can always construct such a trajectory just by using the definition of $m(x)$ as an infimum. Then, observe that the sequence $(S(x_k))_{k\in\mathbf{N}}$ is nested. Since $\succeq_s$ is transitive and $x_{k+1} \succeq_s x_k$, we have the following implication:

$$y \in S(x_{k+1}) \iff y \succeq_s x_{k+1} \Longrightarrow y \succeq_s x_k \iff y \in S(x_k),$$

i.e., $S(x_{k+1}) \subset S(x_k)$ for all $k \in \mathbf{N}$.

Let us prove that $\operatorname{diam} S(x_k) \to 0$ as $k \to +\infty$. By using Lemma 3.4.1, it is enough to prove that $f(x_k) - m(x_k) \to 0$ as $k \to +\infty$.

Since $S(x_{k+1}) \subset S(x_k)$ we have

$$\begin{aligned} m(x_{k+1}) &= \inf\left\{f(y) : y \in S(x_{k+1})\right\} \\ &\geq \inf\left\{f(y) : y \in S(x_k)\right\} = m(x_k). \end{aligned}$$

We now use that this agent satisfies enough, i.e.,

$$f(x_{k+1}) \leq f(x_k) - \lambda[f(x_k) - m(x_k)],$$

and the inequality $m(x_{k+1}) \geq m(x_k)$ to obtain

$$\begin{aligned} f(x_{k+1}) - m(x_{k+1}) &\leq f(x_k) - \lambda[f(x_k) - m(x_k)] - m(x_k) \\ &\leq (1-\lambda)[f(x_k) - m(x_k)]. \end{aligned}$$

Hence

$$f(x_k) - m(x_k) \leq (1-\lambda)^k [f(x_0) - m(x_0)]$$

and

$$\text{diam } S(x_k) \leq 2(1-\lambda)^k [f(x_0) - m(x_0)].$$

Since $0 < \lambda < 1$ we have diam $S(x_k) \to 0$ as $k \to +\infty$. The sequence $(S(x_k))_{k\in\mathbf{N}}$ is a decreasing sequence of closed nonempty sets (closedness follows from the lower semicontinuity of f) whose diameter tends to zero. Since (X, d) is complete, we have, by a classical result, that $\bigcap_{k\in\mathbf{N}} S(x_k) = \{\bar{x}\}$ is nonvoid and is reduced to a single element $\bar{x} \in X$. For any $k \in \mathbf{N}$, we have x_k and $\bar{x}$, which belong to $S(x_k)$, hence $d(x_k, \bar{x}) \leq$ diam $S(x_k)$ which tends to zero. Thus, x_k converges to $\bar{x}$ in (X, d) as $k \to +\infty$.

The maximality of $\bar{x}$ with respect to $\succeq_s$ follows from the following observation: suppose that $y \succeq_s \bar{x}$. Since $\bar{x} \in S(x_k)$ for every $k \in \mathbf{N}$, we have $y \succeq_s x_k$ for all $k \in \mathbf{N}$, i.e., $y \in \bigcap_{k\in\mathbf{N}} S(x_k) = \{\bar{x}\}$. □

Indeed, when proving Theorem 3.4.3 and its dynamical version (Theorem 3.4.4), we have obtained a stronger version of Ekeland's variational principle, which is formulated below.

Theorem 3.4.5. *Let (X, d) be a complete metric space and $f : X \longrightarrow \mathbf{R} \cup \{+\infty\}$ a proper lower semicontinuous function which is bounded below. Let $\varepsilon > 0$ and $x_0 \in X$ be given such that*

$$f(x_0) \leq \inf_X f + \varepsilon,$$

and let $\lambda > 0$. Then, there exists some $\bar{x}_{\varepsilon,\lambda} \in X$ such that

$$\begin{aligned} &f(\bar{x}_{\varepsilon,\lambda}) \leq f(x_0) \leq \inf_X f + \varepsilon; \\ &d(\bar{x}_{\varepsilon,\lambda}, x_0) \leq \lambda; \\ &f(\bar{x}_{\varepsilon,\lambda}) < f(x) + \frac{\varepsilon}{\lambda} d(\bar{x}_{\varepsilon,\lambda}, x) \ \forall\, x \neq \bar{x}_{\varepsilon,\lambda}. \end{aligned}$$

Proof. Let us apply Theorem 3.4.3 with $\frac{\varepsilon}{\lambda} d$ instead of d. One obtains the existence of some $\bar{x}_{\varepsilon,\lambda} \in X$ which satisfies

$$\bar{x}_{\varepsilon,\lambda} \succeq_s x_0, \text{ i.e., } f(\bar{x}_{\varepsilon,\lambda}) + \frac{\varepsilon}{\lambda} d(\bar{x}_{\varepsilon,\lambda}, x_0) \leq f(x_0),$$

and $\bar{x}_{\varepsilon,\lambda}$ is maximal with respect to $\succeq_s$. We thus have $f(\bar{x}_{\varepsilon,\lambda}) \leq f(x_0)$ and

$$\begin{aligned} \inf_X f + \frac{\varepsilon}{\lambda} d(\bar{x}_{\varepsilon,\lambda}, x_0) &\leq f(\bar{x}_{\varepsilon,\lambda}) + \frac{\varepsilon}{\lambda} d(\bar{x}_{\varepsilon,\lambda}, x_0) \\ &\leq f(x_0) \\ &\leq \inf_X f + \varepsilon, \end{aligned}$$

which implies $d(\bar{x}_{\varepsilon,\lambda}, x_0) \leq \lambda$. The last property expresses that $\bar{x}_{\varepsilon,\lambda}$ is maximal with respect to $\succeq_s$. □

Chapter 4

Complements on measure theory

4.1 Hausdorff measures and Hausdorff dimension

This section aims to define intrinsically the intuitive notion of length, area, and volume. More precisely, we would like to provide a nonnegative measure for any subset of $\mathbf{R}^N$, which agrees with the well known k-dimensional measure for regular k-dimensional surfaces when k is an integer. Hausdorff's construction, as shown below, is particularly well suited to the geometry of sets and does not require any local parametrization on these sets; therefore, no assumption of regularity is needed. For instance, the Hausdorff measure offers the possibility of measuring fractal sets, as well as defining a new notion of dimension for any set, thereby extending the classical topological dimension. Note that the process described in the first subsection is the Carathéodory general approach to construct a measure from a σ-subadditive set function (or outer measure).

4.1.1 Outer Hausdorff measures and Hausdorff measures

We denote the collection of all subsets of $\mathbf{R}^N$ by $\mathcal{P}(\mathbf{R}^N)$ and, for any nonempty set E of $\mathcal{P}(\mathbf{R}^N)$, we set $\operatorname{diam}(E) = \sup\{d(x, y) : (x, y) \in E\}$, the diameter of E, where d is the Euclidean distance in $\mathbf{R}^N$. When s is a positive integer we denote the volume of the unit ball of $\mathbf{R}^s$ by ω_s; in the general case $s \geq 0$, we set

$$\omega_s = \frac{\pi^{s/2}}{\Gamma(1 + s/2)},$$

where Γ is the well-known Euler function

$$\Gamma(t) = \int_0^{+\infty} x^{t-1} e^{-x}\, dx.$$

We also set $c_s = 2^{-s}\omega_s$. For instance, we have

$$c_0 = 1, \quad c_1 = 1, \quad c_2 = \pi/4, \quad c_3 = \pi/6.$$

Let E be any set of $\mathcal{P}(\mathbf{R}^N)$ and $\delta > 0$. A finite or countable family $(A_i)_{i\in\mathbf{N}}$ of sets in $\mathcal{P}(\mathbf{R}^N)$ satisfying $0 < \operatorname{diam}(A_i) \leq \delta$ and $E \subset \bigcup_{i\in\mathbf{N}} A_i$ will be called a δ-covering of E.

Definition 4.1.1. *For each $s \geq 0$, $\delta > 0$, and $E \subset \mathbf{R}^N$, let us set*

$$\mathcal{H}^s_\delta(E) := \inf\left\{ c_s \sum_{i\in\mathbf{N}} \operatorname{diam}(A_i)^s : (A_i)_{i\in\mathbf{N}} \text{ } \delta\text{-covering of } E \right\}.$$

The s-dimensional outer Hausdorff measure is the set mapping $\mathcal{H}^s$ taking its values in $[0, +\infty]$ defined by

$$\begin{aligned}\mathcal{H}^s(E) &:= \sup_{\delta>0} \mathcal{H}^s_\delta(E)\\ &= \lim_{\delta\to 0} \mathcal{H}^s_\delta(E).\end{aligned}$$

Remark 4.1.1. The constant c_s is a normalization parameter so that when $s \in \mathbf{N}$ and E is an s-dimensional regular hypersurface of $\mathbf{R}^N$, $\mathcal{H}^s(E)$ is the s-dimensional area of E (see Proposition 4.1.6 or Theorem 4.1.1). The positive number δ, intended to tend to zero, forces the sets A_i of the δ-covering to follow the geometry of E.

Proposition 4.1.1. *The set function $\mathcal{H}^s : \mathcal{P}(\mathbf{R}^N) \longrightarrow [0, +\infty]$ is an outer measure, i.e., satisfies*

(i) $\mathcal{H}^s(\emptyset) = 0$;

(ii) *(σ-subadditivity) for all sequences $(E_i)_{i\in\mathbf{N}}$ of subsets of $\mathbf{R}^N$ such that $E \subset \cup_{i\in\mathbf{N}} E_i$,*

$$\mathcal{H}^s(E) \leq \sum_{i\in\mathbf{N}} \mathcal{H}^s(E_i);$$

(iii) *$\mathcal{H}^s$ is a nondecreasing set function, that is, $\mathcal{H}^s(A) \leq \mathcal{H}^s(B)$ whenever $A \subset B$.*

Proof. For all A in $\mathcal{P}(\mathbf{R}^N)$ such that $\operatorname{diam}(A) \leq \delta$, one has $\mathcal{H}^s_\delta(\emptyset) \leq \operatorname{diam}(A)^s \leq \delta^s$ and (i) follows. The monotonicity property (iii) follows straightforwardly from the definition of $\mathcal{H}^s$.

Let now $\varepsilon > 0$ and $(A_{i,j})_{j\in\mathbf{N}}$ be a δ-covering of E_i satisfying

$$c_s \sum_{j\in\mathbf{N}} \operatorname{diam}(A_{i,j})^s \leq \frac{\varepsilon}{2^i} + \mathcal{H}^s_\delta(E_i).$$

Obviously, $(A_{i,j})_{(i,j)\in\mathbf{N}^2}$ is a δ-covering of $\bigcup_{i\in\mathbf{N}} E_i$ so that

$$\mathcal{H}^s_\delta\left(\bigcup_{i\in\mathbf{N}} E_i\right) \leq \sum_{i\in\mathbf{N}} \mathcal{H}^s_\delta(E_i) + 2\varepsilon.$$

Letting $\delta \to 0$, conclusion (ii) follows from the fact that $\mathcal{H}^s$ is a nondecreasing set function and ε is arbitrary. □

Following the classical construction of a measure from an outer measure, we define the subset $\mathcal{M}_s$ of $\mathcal{H}^s$-measurable sets in the sense of Carathéodory:

$$A \in \mathcal{M}_s \Longleftrightarrow \forall X \in \mathcal{P}(\mathbf{R}^N),\ \mathcal{H}^s(X) = \mathcal{H}^s(X \cap A) + \mathcal{H}^s(X \setminus A).$$

Note that $\emptyset$ and $\mathbf{R}^N$ belong to $\mathcal{M}_s$.

Proposition 4.1.2. *The set $\mathcal{M}_s$ is a σ-algebra and $\mathcal{H}^s$ is σ-additive on $\mathcal{M}_s$.*

Proof. Obviously $\mathbf{R}^N$ belongs to $\mathcal{M}_s$. The proof of stability of $\mathcal{M}_s$ with respect to the complementary operation is easily established from a straightforward calculation. The proof of stability for countable union and the σ-additivity of $\mathcal{H}^s$ is divided into three steps.

First step. Stability for finite unions and finite intersections. Let $A_1,\ A_2$ be two sets in $\mathcal{M}_s$ and X some set in $\mathcal{P}(\mathbf{R}^N)$. Since A_1 and A_2 are two measurable sets, an elementary calculation on sets gives

$$\begin{aligned}\mathcal{H}^s(X) &= \mathcal{H}^s(A_1 \cap X) + \mathcal{H}^s(X \setminus A_1)\\ &= \mathcal{H}^s(A_1 \cap X) + \mathcal{H}^s((X \setminus A_1) \cap A_2) + \mathcal{H}^s((X \setminus A_1) \setminus A_2)\\ &= \mathcal{H}^s(A_1 \cap X) + \mathcal{H}^s(X \cap A_2 \setminus A_1) + \mathcal{H}^s(X \setminus (A_1 \cup A_2)).\end{aligned}$$

According to the identities

$$X \cap A_1 = X \cap (A_1 \cup A_2) \cap A_1 \text{ and } X \cap A_2 \setminus A_1 = X \cap (A_1 \cup A_2) \setminus A_1,$$

we derive

$$\mathcal{H}^s(X) = \mathcal{H}^s(X \cap (A_1 \cup A_2)) + \mathcal{H}^s(X \setminus (A_1 \cup A_2))$$

so that $A_1 \cup A_2 \in \mathcal{M}_s$. Stability for finite intersection is then obtained thanks to the stability with respect to the complementary operation.

Note that substituting X by $X \cap (A_1 \cup A_2)$ in $\mathcal{H}^s(X) = \mathcal{H}^s(A_1 \cap X) + \mathcal{H}^s(X \setminus A_1)$ we obtain the following important equality used in the step below:

$$\mathcal{H}^s(X \cap (A_1 \cup A_2) = \mathcal{H}^s(X \cap A_1) + \mathcal{H}^s(X \cap A_2) \tag{4.1}$$

whenever A_1 and A_2 are disjoint sets in $\mathcal{M}_s$.

Second step. Stability for disjoint countable unions and σ-additivity of $\mathcal{H}^s$. Let $(A_i)_{i\in\mathbf{N}}$ be a family of pairwise disjoint sets in $\mathcal{M}_s$ and A their union. According to the subadditivity of $\mathcal{H}^s$ for all X in $\mathcal{P}(\mathbf{R}^N)$ we have

$$\begin{aligned}\mathcal{H}^s(X) &\le \mathcal{H}^s(X \setminus A) + \mathcal{H}^s(X \cap A)\\ &\le \mathcal{H}^s(X \setminus A) + \sum_{i=0}^{\infty} \mathcal{H}^s(X \cap A_i)\\ &= \mathcal{H}^s(X \setminus A) + \lim_{n\to+\infty} \mathcal{H}^s\left(\bigcup_{i=0}^{n} X \cap A_i\right)\\ &\le \liminf_{n\to+\infty} \left(\mathcal{H}^s\left(X \setminus \bigcup_{i=0}^{n} A_i\right) + \mathcal{H}^s\left(\bigcup_{i=0}^{n} X \cap A_i\right)\right)\\ &= \mathcal{H}^s(X),\end{aligned}$$

where we have used (4.1) in the first equality, the nondecreasing of $\mathcal{H}^s$ in the inequality, and the stability for finite union in the last equality. This proves that A belongs to $\mathcal{M}_s$. The σ-additivity of $\mathcal{H}^s$ is obtained by taking $X = A$.

Last step. Stability for countable union. Let $(A_i)_{i\in\mathbf{N}}$ be a family of sets in $\mathcal{M}_s$. Set

$$B_0 = A_0 \text{ and for } n \geq 1,\ B_n = A_n \setminus \bigcup_{i=0}^{n-1} A_i.$$

According to the first step, the family $(B_i)_{i\in\mathbf{N}}$ is made up of pairwise disjoint sets of $\mathcal{M}_s$ so that, from step 2, $\bigcup_{i\in\mathbf{N}} B_i = \bigcup_{i\in\mathbf{N}} A_i$ belongs to $\mathcal{M}_s$. □

Definition 4.1.2. *The restriction to $\mathcal{M}_s$ of the set function $\mathcal{H}^s$ is called the s-dimensional Hausdorff measure.*

The s-dimensional Hausdorff measure $\mathcal{H}^s$ is a $[0, +\infty]$-valued Borel measure in the following sense:

Proposition 4.1.3. *The σ-algebra $\mathcal{M}_s$ contains the σ-algebra of all the Borel sets of $\mathbf{R}^N$.*

Proof. The proof is based on the following Carathéodory criterion. An outer measure satisfying this criterion is sometimes called an outer metric measure.

Lemma 4.1.1. *Let μ be an outer measure on a metric space E equipped with its Borel σ-algebra $\mathcal{B}(E)$ and $\mathcal{M}_\mu$ be the σ-algebra of the measurable sets in the Carathéodory sense. Then $\mathcal{B}(E) \subset \mathcal{M}_\mu$ iff μ is additive on any pair $\{A, B\}$ of sets of E satisfying $d(A, B) > 0$.*

Proof of Lemma 4.1.1. Let us assume $\mathcal{B}(E) \subset \mathcal{M}_\mu$ and let A, B be two subsets of E such that $d(A, B) > 0$. Since $\overline{A} \in \mathcal{M}_\mu$ the two decompositions

$$A = (A \cup B) \cap \overline{A},$$
$$B = (A \cup B) \setminus \overline{A}$$

imply $\mu(A \cup B) = \mu(A) + \mu(B)$.

Conversely, let A be an open subset of E and X any fixed subset of E such that $\mu(X) < +\infty$. According to subadditivity, it is enough to establish the inequality

$$\mu(X) \geq \mu(X \cap A) + \mu(X \setminus A).$$

One may assume $\mu(X) < +\infty$. For all $k \in \mathbf{N}^*$, let us define the sets

$$A_k := \left\{ x \in A : d(x, E \setminus A) > \frac{1}{k} \right\},$$
$$B_k := A_{k+1} \setminus A_k.$$

Noticing that for $k - l \geq 2$, $d(B_k, B_l) \geq \frac{1}{l+1} - \frac{1}{k} > 0$, we have for all $n \in \mathbf{N}$

$$\sum_{i=1}^{n} \mu(X \cap B_{2i}) = \mu\left(\bigcup_{i=1}^{n} X \cap B_{2i} \right) \leq \mu(X),$$

hence

$$\sum_{k \text{ even}} \mu(X \cap B_k) \leq \mu(X).$$

The same calculation gives

$$\sum_{k \text{ odd}} \mu(X \cap B_k) \leq \mu(X)$$

so that

$$\sum_{k \in \mathbf{N}} \mu(X \cap B_k) \leq 2\mu(X)$$

and

$$\lim_{k \to +\infty} \sum_{i \geq k} \mu(X \cap B_i) = 0.$$

We infer

$$\begin{aligned} \lim_{k \to +\infty} \mu(X \cap (A \setminus A_k)) &= \lim_{k \to +\infty} \mu\big(X \cap (\cup_{i \geq k} B_i)\big) \\ &\leq \lim_{k \to +\infty} \sum_{i \geq k} \mu(X \cap B_i) = 0, \end{aligned}$$

and, by subadditivity, we deduce

$$\begin{aligned} \mu(X \cap A) &\leq \mu(X \cap A_k) + \mu(X \cap (A \setminus A_k)) \\ &\leq \liminf_{k \to +\infty} \mu(X \cap A_k). \end{aligned}$$

Since obviously $\limsup_{n \to +\infty} \mu(X \cap A_k) \leq \mu(X \cap A)$, one has

$$\mu(X \cap A) = \lim_{k \to +\infty} \mu(X \cap A_k).$$

From $d(A_k, X \setminus A) > \frac{1}{k} > 0$ we now obtain

$$\mu(X) \geq \mu((X \cap A_k) \cup (X \setminus A)) = \mu(X \cap A_k) + \mu(X \setminus A),$$

and we complete the proof of Lemma 4.1.1 by letting $k \to +\infty$. □

Proof of Proposition 4.1.3 continued. According to Lemma 4.1.1, it is enough to prove that $\mathcal{H}^s$ is an outer metric measure. Let A and B be two subsets of $\mathbf{R}^N$ such that $d(A, B) > 0$ and let $\mathcal{C} = (C_i)_{i \in \mathbf{N}}$ be a covering of $A \cup B$ with $\operatorname{diam}(C_i) \leq \delta < d(A, B)$. We can write this covering as the union of two disjoint δ-coverings: take $\mathcal{A} = \{C \in \mathcal{C} : C \cap A \neq \emptyset\}$ and $\mathcal{B} = \{C \in \mathcal{C} : C \cap B \neq \emptyset\}$. Therefore

$$\sum_{i=1}^{\infty} c_s \operatorname{diam}(C_i)^s = \sum_{C \in \mathcal{A}} c_s \operatorname{diam}(C)^s + \sum_{C \in \mathcal{B}} c_s \operatorname{diam}(C)^s$$

and $\mathcal{H}^s_\delta(A \cup B) \geq \mathcal{H}^s_\delta(A) + \mathcal{H}^s_\delta(B)$. We end the proof by letting $\delta \to 0$. □

Remark 4.1.2. The following process is often used to construct an outer measure and thus a measure on Cantor-type sets C in $\mathbf{R}^N$. Let $\mathcal{E}_0 = \{C\}$ and for $n \in \mathbf{N}^*$ let $\mathcal{E}_n$ be a finite collection of disjoint subsets E of C, such that each $E \in \mathcal{E}_n$ is contained in one of the sets of $\mathcal{E}_{n-1}$. We assume moreover that $\lim_{n\to+\infty}\{\mathrm{diam}(E) : E \in \mathcal{E}_n\} = 0$. We set now $\mu(C) = a$, where a is an arbitrary number satisfying $0 < a < +\infty$ and, for all the sets $E_i,\ i = 1, \ldots, m_1$, of $\mathcal{E}_1$, we define the masses $\mu(E_i)$ such that

$$\sum_{i=1}^{m_1} \mu(E_i) = \mu(C).$$

Similarly, we assign masses $\mu(E_i)$ to the sets of $\mathcal{E}_n$ such that if $E \in \mathcal{E}_{n-1}$, $E = \cup_{i=1}^{m_n} E_i$, $E_i \in \mathcal{E}_n$,

$$\sum_{i=1}^{m_n} \mu(E_i) = \mu(E).$$

We finally set

$$\mu\left(\mathbf{R}^N \setminus \bigcup_{E\in\mathcal{E}_n} E\right) = 0.$$

Let us denote the collection of all the complementary sets of $\mathcal{E}_n$ by $\overline{\mathcal{E}}_n$ and set $\mathcal{E} = \cup_{n\in\mathbf{N}}(\mathcal{E}_n \cup \overline{\mathcal{E}}_n)$. For each set $A \in \mathcal{P}(\mathbf{R}^N)$, we now extend μ by setting

$$\mu(A) = \inf\left\{\sum_{i\in\mathbf{N}} \mu(E_i) : A \subset \bigcup_{i\in\mathbf{N}} E_i,\ E_i \in \mathcal{E}\right\},$$

which defines an outer measure. One can prove that μ is a measure whose σ-algebra of measurable sets contains the σ-algebra of Borel sets of $\mathbf{R}^N$. Moreover, the support of μ, that is, the smallest closed set X of $\mathbf{R}^N$ such that $\mu(\mathbf{R}^N \setminus X) = 0$, is contained in $\cap_{n\in\mathbf{N}}\overline{\cup_{E\in\mathcal{E}_n} E}$.

Remark 4.1.3. The construction of Hausdorff measures $\mathcal{H}^s$ can be made in a general metric space X and indeed most of the properties continue to hold in this larger framework. For a detailed presentation of Hausdorff measures in metric spaces, see Ambrosio and Tilli [23].

Theorem 4.1.1. *For all Lebesgue measurable sets E in $\mathbf{R}^N$, we have $\mathcal{H}^N(E) = \mathcal{L}^N(E)$, where $\mathcal{L}^N$ denotes the Lebesgue measure on $\mathbf{R}^N$. Moreover, $\mathcal{H}^s(E) = 0$ for $s > N$, whereas $\mathcal{H}^0$ is the counting measure.*

Proof. For all Lebesgue measurable sets E in $\mathbf{R}^N$, let us recall the so-called isodiametric inequality, which asserts that the Lebesgue measure of E is less than the Lebesgue measure of any ball having the same diameter:

$$\mathcal{L}^N(E) \le c_N(\mathrm{diam}(E))^N.$$

For a proof see, for instance, Evans and Gariepy [132]. For all covering $(A_i)_{i\in\mathbf{N}}$ of E we then have

$$\sum_{i\in\mathbf{N}} c_N(\mathrm{diam}(A_i))^N \ge \sum_{i\in\mathbf{N}} \mathcal{L}^N(A_i) \ge \mathcal{L}^N(E),$$

hence $\mathcal{H}^N_\delta(E) \ge \mathcal{L}^N(E)$. Letting $\delta \to 0$ gives $\mathcal{H}^N(E) \ge \mathcal{L}^N(E)$.

The converse inequality is more involved. We will use the so-called Vitali's covering lemma. Let us firstly define the notion of fine covering.

Definition 4.1.3. *A family $\mathcal{F}$ of closed balls in $\mathbf{R}^N$ is said to cover a set $E \subset \mathbf{R}^N$ finely if, for each $x \in E$ and each $\varepsilon > 0$, there exists $\overline{B}_r(x) \in \mathcal{F}$ with $r < \varepsilon$, where $B_r(x)$ denotes the open ball with radius r and centered at x.*

Lemma 4.1.2. *(Vitali's covering theorem). Let $E \subset \mathbf{R}^N$ with $\mathcal{L}^N(E) < +\infty$ and finely covered by a family of closed balls $\mathcal{F}$. Then there exists a countable subfamily $\mathcal{G}$ of pairwise disjoint elements of $\mathcal{F}$ such that*

$$\mathcal{L}^N\left(E \setminus \bigcup_{B \in \mathcal{G}} B\right) = 0.$$

PROOF. For the proof see, for instance, Ziemer [229]. □

Remark 4.1.4. In the definition of a family finely covering a subset E of $\mathbf{R}^N$ one may replace the family of closed balls by a regular family of closed subsets of $\mathbf{R}^N$ (see [229]).

Vitali's covering theorem is also valid for nonnegative Radon measures μ on $\mathbf{R}^N$ (i.e., locally bounded nonnegative Borel measures; see Section 4.2). For a proof consult [229].

PROOF OF THEOREM 4.1.1 CONTINUED. Let A be a set of finite Lebesgue measure in $\mathbf{R}^N$ and, for each $\eta > 0$, let U be an open subset of $\mathbf{R}^N$ such that $A \subset U$ and $\mathcal{L}^N(U \setminus A) < \eta$. There exists a family $\mathcal{F}$ of closed balls with diameter less than δ, included in U, and covering finely U. Indeed, for all $x \in U$, there exists a closed ball $\overline{B}_{r(x)}(x)$ included in U and $\mathcal{F} = (\overline{B}_r(x))_{x \in U,\ r \le r(x) \wedge \delta/2}$ is such a suitable family. According to Lemma 4.1.2 one can extract a subfamily $(B_i)_{i=1,\dots,\infty}$ of pairwise disjoint elements with diameter less than δ, satisfying

$$\mathcal{L}^N\left(U \setminus \bigcup_{i=1}^{\infty} B_i\right) = 0, \ \bigcup_{i=1}^{\infty} B_i \subset U.$$

Set $A^* := \bigcup_{i=1}^{\infty}(B_i \cap A)$. We have $\mathcal{L}^N(A \setminus A^*) = 0$ and

$$\begin{aligned} \mathcal{H}^N_\delta(A^*) &\le \sum_{i=1}^{\infty} c_N (\operatorname{diam} B_i)^N \\ &= \sum_{i=1}^{\infty} \mathcal{L}^N(B_i) \\ &= \mathcal{L}^N(U) \le \mathcal{L}^N(A) + \eta. \end{aligned}$$

Letting $\delta \to 0$ and $\eta \to 0$, we obtain $\mathcal{H}^N(A^*) \le \mathcal{L}^N(A)$. It remains to establish $\mathcal{L}^N(A \setminus A^*) = 0 \Longrightarrow \mathcal{H}^N(A \setminus A^*) = 0$ and more generally that if E is a Borel set satisfying $\mathcal{L}^N(E) = 0$, then $\mathcal{H}^N(E) = 0$. This property is a corollary of the first assertion of Lemma 4.1.3 below, whose proof may be found in [229].

Lemma 4.1.3. *Let $\mathcal{F}$ be a family of closed balls with* $\sup\{\mathrm{diam}\, B : B \in \mathcal{F}\} < +\infty$. *Then there exists a countable subfamily $\mathcal{G}$ of pairwise disjoint elements of $\mathcal{F}$ such that*

$$\bigcup_{B\in\mathcal{F}} B \subset \bigcup_{B\in\mathcal{G}} B^*,$$

where B^* denotes the closed ball concentric with B with radius five times as big as that of B. Let E be a subset of $\mathbf{R}^N$, finely covered by $\mathcal{F}$; then, for all finite family $\mathcal{G}^* \subset \mathcal{G}$,

$$E \subset \Big(\bigcup_{B\in\mathcal{G}^*} B\Big)\bigcup\Big(\bigcup_{B\in\mathcal{G}\setminus\mathcal{G}^*} B^*\Big).$$

Indeed, for $\eta > 0$, consider an open subset U of $\mathbf{R}^N$ such that $E \subset U$ and $\mathcal{L}^N(U) \le \eta$. The open set U being a union of closed balls included in U with diameter less than δ, according to Lemma 4.1.3, there exists a family $\mathcal{G}$ of pairwise disjoint closed balls included in U such that $U \subset \bigcup_{B\in\mathcal{G}} B^*$ and

$$\begin{aligned}\mathcal{H}^N_\delta(E) &\le \sum_{B\in\mathcal{G}} H^N_\delta(B^*)\\ &\le \sum_{B\in\mathcal{G}} c_N 5^N (\mathrm{diam}(B))^N\\ &= 5^N \sum_{B\in\mathcal{G}} \mathcal{L}^N(B) = 5^N \mathcal{L}^N\left(\bigcup_{B\in\mathcal{G}} B\right) < 5^N\eta.\end{aligned}$$

The conclusion follows by letting $\delta \to 0$ and $\eta \to 0$.

We establish now that for all Borel subsets E of $\mathbf{R}^N$, $\mathcal{H}^s(E) = 0$ when $s > N$. One can write $E = \cup_{n\in\mathbf{N}} E_n$, where $\mathcal{H}^N(E_n) < +\infty$ (take $E_n = B_n \cap E$, where B_n denotes the ball of $\mathbf{R}^N$ with radius n, centered at 0). Let $\delta > 0$ and $(A_i)_{i\in\mathbf{N}}$ a δ-covering of E_n. We have

$$c_s \sum_{i\in\mathbf{N}} \mathrm{diam}(A_i)^s \le \frac{c_s}{c_N}\delta^{s-N} c_N \sum_{i\in\mathbf{N}} \mathrm{diam}(A_i)^N$$

which yields

$$\mathcal{H}^s_\delta(E_n) \le \frac{c_s}{c_N}\delta^{s-N}\mathcal{H}^N(E_n),$$

and finally, letting $\delta \to 0$, one obtains $\mathcal{H}^s(E_n) = 0$. Therefore, by subadditivity,

$$\mathcal{H}^s(E) \le \sum_{n=0}^{\infty} \mathcal{H}^s(E_n) = 0.$$

The fact that $\mathcal{H}^0$ is the counting measure is easy to establish and left to the reader. □

4.1.2 Hausdorff measures: Scaling properties and Lipschitz transformations

The scaling properties of length, area, or volume are well known. The two propositions below summarize and generalize these properties.

Proposition 4.1.4. *Let A be any subset of $\mathbf{R}^N$ and $\lambda > 0$. Then*

$$\mathcal{H}^s(\lambda A) = \lambda^s \mathcal{H}^s(A).$$

PROOF. If $(A_i)_i$ is a δ-covering of A, then $(\lambda A_i)_i$ is a $\lambda\delta$-covering of λA so that

$$\mathcal{H}^s_{\lambda\delta}(\lambda A) \leq \sum_{i=1}^{\infty} c_s(\lambda \mathrm{diam}(A_i))^s = \lambda^s \sum_{i=1}^{\infty} c_s \mathrm{diam}(A_i)^s.$$

Therefore $\mathcal{H}^s_{\lambda\delta}(\lambda A) \leq \lambda^s \mathcal{H}^s_\delta(A)$. Letting $\delta \to 0$, we obtain $\mathcal{H}^s(\lambda A) \leq \lambda^s \mathcal{H}^s(A)$. Replacing λ with $1/\lambda$ and A with λA gives the converse inequality. □

Proposition 4.1.5. *Let A be any subset of $\mathbf{R}^N$ and $f : A \longrightarrow \mathbf{R}^m$ satisfying for all x and y in A*

$$|f(x) - f(y)| \leq L|x - y|^\alpha,$$

where $L > 0$ and $\alpha > 0$ are two given constants. Then

$$\mathcal{H}^{s/\alpha}(f(A)) \leq \frac{c_{s/\alpha}}{c_s} L^{s/\alpha} \mathcal{H}^s(A).$$

Consequently, if f is a Lipschitz function with modulus L, then

$$\mathcal{H}^s(f(A)) \leq L^s\, \mathcal{H}^s(A),$$

and if f is an isometry, then

$$\mathcal{H}^s(f(A)) = \mathcal{H}^s(A).$$

PROOF. If $(A_i)_i$ is a δ-covering of A, then $(f(A \cap A_i))_i$ is an $L\delta^\alpha$-covering of $f(A)$, and

$$\mathcal{H}^{s/\alpha}_{L\delta^\alpha}(f(A)) \leq \frac{c_{s/\alpha}}{c_s} \sum_{i=1}^{\infty} c_s \mathrm{diam}(f(A \cap A_i))^{s/\alpha} \leq \frac{c_{s/\alpha}}{c_s} L^{s/\alpha} \sum_{i=1}^{\infty} c_s \mathrm{diam}(A_i)^s.$$

We deduce $\mathcal{H}^{s/\alpha}_{L\delta^\alpha}(f(A)) \leq \frac{c_{s/\alpha}}{c_s} L^{s/\alpha} \mathcal{H}^s_\delta(A)$ and the conclusion follows after letting $\delta \to 0$. □

The Hausdorff measure of an N-dimensional regular hypersurface of $\mathbf{R}^m$ is nothing but its classical area when the hypersurface is defined by means of a parametrization.

Proposition 4.1.6. *Let $m \in \mathbf{N}$ with $m \geq N$, $f : \mathbf{R}^N \longrightarrow \mathbf{R}^m$ be a one-to-one Lipschitz map and E a Borel subset of $\mathbf{R}^N$. Then*

$$\mathcal{H}^N(f(E)) = \int_E \left(\sum_{i=1}^{C_m^N} |J_i|^2 \right)^{1/2} d\mathcal{L}^N,$$

where J_i, $i = 1, \ldots, C_m^N$, are the $N \times N$-minors of the Jacobian matrix of f.

PROOF. To illustrate how the definition of the one-dimensional Hausdorff measure is well adapted to the local geometry of arcs, we only give the proof in the case $N = 1$, where $f : [0,1] \longrightarrow \mathbf{R}^m$ is denoted by $t \mapsto (x_i(t))_{i=1,\dots,m}$. For a complete proof, consult Ambrosio [16] or Evans and Gariepy [132]. We must establish

$$\begin{aligned}\mathcal{H}^1\big(f([0,1])\big) &= \int_{[0,1]} \left(\sum_{i=1}^{m} |x_i'(t)|^2\right)^{1/2} dt \\ &= \int_{[0,1]} |f'(t)|\, dt,\end{aligned}$$

which is the length of the parametrized arc $\Gamma = f([0,1])$. We use the well-known classical result straightforwardly obtained from Taylor's formula and Riemann integration theory: if f belongs to $\mathbf{C}^1([\alpha,\beta], \mathbf{R}^m)$, we have

$$\int_{[\alpha,\beta]} |f'(t)|\, dt = \lim_{\eta\to 0} \sum_{i=0}^{n-1} |f(t_{i+1}) - f(t_i)| \geq |f(\alpha) - f(\beta)|,$$

where $t_0 = \alpha < t_1 < \dots\ t_i < t_{i+1} < \dots t_n = \beta$ is some finite subdivision of $[\alpha,\beta]$ with $\eta = \max_{i=0,\dots,n-1}(t_{i+1} - t_i)$.

Step 1. The function f is assumed to belong to $\mathbf{C}^1([0,1], \mathbf{R}^m)$ and we prove $\int_{[0,1]} |f'(t)|\, dt \geq \mathcal{H}^1(\Gamma)$. For all $\delta > 0$, let $\eta > 0$ be such that

$$|t - t'| < \eta \Longrightarrow |f(t) - f(t')| < \delta.$$

Let us consider a finite subdivision $t_0 = 0 < t_1 < \dots\ t_i < t_{i+1} < \dots t_n = 1$ of $[0,1]$ satisfying $\eta > \max_{i=0,\dots,n-1}(t_{i+1} - t_i)$ and set $\Gamma_i = f([t_i, t_{i+1}))$ for $i = 0, \dots, n-1$. Consider now $a_i,\ b_i$ in $[t_i, t_{i+1}]$ such that $\operatorname{diam}(\Gamma_i) = |f(a_i) - f(b_i)|$. We have

$$\Gamma \setminus \{f(1)\} = \bigcup_{i=0}^{n-1} \Gamma_i$$

with $\operatorname{diam}(\Gamma_i) < \delta$ and

$$\begin{aligned}\int_{[0,1]} |f'(t)|\, dt &= \sum_{i=0}^{n-1} \int_{[t_i,t_{i+1}]} |f'(t)|\, dt \\ &\geq \sum_{i=0}^{n-1} \int_{[a_i,b_i]} |f'(t)|\, dt \\ &\geq \sum_{i=0}^{n-1} |f(a_i) - f(b_i| \\ &= \sum_{i=0}^{n-1} \operatorname{diam}(\Gamma_i) \geq \mathcal{H}^1_\delta(\Gamma).\end{aligned}$$

We end this step by letting $\delta \to 0$.

Step 2. The function f is again assumed to belong to $\mathbf{C}^1([0,1],\mathbf{R}^m)$ and we prove the converse inequality $\int_{[0,1]}|f'(t)|\,dt \le \mathcal{H}^1(\Gamma)$. Let $n \in \mathbf{N}$, $h = 1/n$, $t_i = hi$, $i = 0,\dots,n-1$, and consider the covering

$$\Gamma = \bigcup_{i=0}^{n-1} f([t_i,t_{i+1})) \cup \{f(1)\}$$

by pairwise disjoint Borel subsets $f([t_i,t_{i+1}))$ of Γ (recall that f is one-to-one). We then have

$$\mathcal{H}^1(\Gamma) = \sum_{i=0}^{n-1} \mathcal{H}^1(f([t_i,t_{i+1}))).$$

For each $i = 1,\dots,n-1$, consider the orthogonal projection π of $f([t_i,t_{i+1}))$ onto the line $(f(t_i),\,f(t_{i+1}))$. As π does not increases distances, from Proposition 4.1.5 we have

$$\mathcal{H}^1\big(f([t_i,t_{i+1}))\big) \ge \mathcal{H}^1\big(\pi(f([t_i,t_{i+1})))\big),$$

which is greater than $\mathcal{H}^1([f(t_i),\,f(t_{i+1})))$. Indeed, a convexity argument yields

$$[f(t_i),\,f(t_{i+1})) \subset \pi(f([t_i,t_{i+1}))).$$

By using the definition, it is easy to prove for the segments $[a,b]$ in $\mathbf{R}^m$ that $\mathcal{H}^1([a,b]) = |b-a|$. Therefore $\mathcal{H}^1([f(t_i),\,f(t_{i+1}]) = |f(t_i) - f(t_{i+1})|$ and

$$\mathcal{H}^1(\Gamma) \ge \sum_{i=0}^{n-1} |f(t_i) - f(t_{i+1})|.$$

The conclusion follows after letting $n \to +\infty$.

Step 3. The function f is assumed to be Lipschitz continuous. The conclusion is a straightforward consequence of the two previous steps and the following approximating result: every Lipschitz map $f : E \to \mathbf{R}^m$ can be approximated in the Lusin sense by functions of class $\mathbf{C}^1$. More precisely, there exists a nondecreasing family $(K_i)_{i\in\mathbf{N}}$ of compact sets included in E such that

$$\mathcal{H}^N\left(E \setminus \bigcup_{i=0}^{\infty} K_i\right) = 0$$

and such that $f\lfloor K_i$ is the restriction of a function of class $\mathbf{C}^1$. For a proof of this property, consult [132]. □

Remark 4.1.5. The measure $\big(\sum_{i=1}^{C_m^N} |J_i|^2\big)^{1/2} \cdot \mathcal{L}^N$ is sometimes called the element of area of the N-dimensional surface $f(E)$ and, for all Borel subset B of $f(E)$,

$$B \mapsto \mathcal{H}^N(B) = \int_{f^{-1}(B)} \left(\sum_{i=1}^{C_m^N} |J_i|^2\right)^{1/2} d\mathcal{L}^N$$

is its corresponding surface measure.

When f is not one-to-one, the generalization of the formula in Proposition 4.1.6 is the so-called area formula:

$$\int_{f(E)} \mathcal{H}^0(E \cap f^{-1}(x))\, \mathcal{H}^N(dx) = \int_E \left(\sum_{i=1}^{m-N+1} |J_i|^2 \right)^{1/2} d\mathcal{L}^N,$$

where $x \mapsto \mathcal{H}^0(E \cap f^{-1}(x))$ is nothing but the multiplicity function as illustrated in the example of the parametrization of the unit circle: $t \mapsto (\cos(2n\pi t), \sin(2n\pi t))$ from $E = [0, 1[$ into $\mathbf{R}^2$.

One can also generalize this formula as follows: let $h : E \to [0, +\infty]$ be a Borel function; then

$$\int_{f(E)} \sum_{x \in f^{-1}(y)} h(x)\, \mathcal{H}^N(dy) = \int_E h(x) \left(\sum_{i=1}^{C_m^N} |J_i|^2 \right)^{1/2} \mathcal{L}^N(dx).$$

For these generalizations, see [16], [132], or Federer [134].

4.1.3 Hausdorff dimension

Among the wide variety of definitions of dimension, the Hausdorff dimension introduced in Theorem 4.1.2 has the advantage of being defined for any set. For alternative definitions of dimensions, consult, for instance, Falconer [133].

Lemma 4.1.4. *For all fixed set A of $\mathbf{R}^N$, the map $s \mapsto \mathcal{H}^s(A)$ is nonincreasing. More precisely, for all $\delta > 0$ and for $t > s$, we have*

$$\mathcal{H}^t_\delta(A) \leq \delta^{t-s} \frac{c_t}{c_s} \mathcal{H}^s_\delta(A). \tag{4.2}$$

PROOF. Let $(A_i)_{i \in \mathbf{N}}$ be a δ-covering of A. One has

$$\begin{aligned} \mathcal{H}^t_\delta(A) &\leq \sum_{i \in \mathbf{N}} c_t \mathrm{diam}(A_i)^t \\ &\leq \frac{c_t}{c_s} \sum_{i \in \mathbf{N}} c_s \mathrm{diam}(A_i)^s \delta^{t-s}. \end{aligned}$$

The conclusion then follows by taking the infimum on all the δ-covering of A. □

Theorem 4.1.2 (definition of the Hausdorff dimension). *Let A be a set of $\mathbf{R}^N$ and set*

$$s_0 := \inf\{t \geq 0 : \mathcal{H}^t(A) = 0\}.$$

Then s_0 satisfies

$$\mathcal{H}^s(A) = \begin{cases} +\infty \ \text{if } s < s_0, \\ 0 \ \text{if } s > s_0. \end{cases}$$

The real number s_0 is called the Hausdorff dimension of the set A and is denoted by $dim_H(A)$. At the critical value s_0, $\mathcal{H}^{s_0}(A)$ may be zero or infinite or may satisfy $0 < \mathcal{H}^{s_0}(A) < +\infty$. In this last case, A is called an s_0-set.

PROOF. Letting $\delta \to 0$ in inequality (4.2), we obtain

$$\mathcal{H}^s(A) < +\infty \Longrightarrow \forall t > s,\ \mathcal{H}^t(A) = 0.$$

Set $s_0 := \inf\{t \geq 0 : \mathcal{H}^t(A) = 0\}$ and take $s < s_0$. Assume that $\mathcal{H}^s(A) < +\infty$. For t satisfying $s < t < s_0$ we have $\mathcal{H}^t(A) = 0$, which contradicts the definition of s_0. Consequently, for $s < s_0$, $\mathcal{H}^s(A) = +\infty$. Since the map $s \mapsto \mathcal{H}^s(A)$ is nonincreasing, for $s > s_0$, we have $\mathcal{H}^s(A) = 0$.

Obviously $\dim_H(\mathbf{R}) = 1$ and $\mathcal{H}^1(\mathbf{R}) = +\infty$. On the other hand, Proposition 4.1.9 provides a nontrivial set A in $\mathbf{R}$ with $\dim_H(A) = 1$ and satisfying $\mathcal{H}^1(A) = 0$. □

Remark 4.1.6. Taking, as δ-covering, the class of balls of $\mathbf{R}^N$, one may define

$$\tilde{\mathcal{H}}^s_\delta(E) := \inf\left\{ c_s \sum_{i=1}^{\infty} \operatorname{diam}(B_i)^s : E \subset \bigcup_{i=1}^{\infty} B_i,\ 0 < \operatorname{diam}(B_i) \leq \delta,\ B_i \text{ ball of } \mathbf{R}^N \right\},$$

and the set mapping $\tilde{\mathcal{H}}^s$, by

$$\begin{aligned} \tilde{\mathcal{H}}^s(E) &:= \sup_{\delta > 0} \tilde{\mathcal{H}}^s_\delta(E) \\ &= \lim_{\delta \to 0} \tilde{\mathcal{H}}^s_\delta(E). \end{aligned}$$

It is easy to establish the following bounds:

$$\mathcal{H}^s(E) \leq \tilde{\mathcal{H}}^s(E) \leq 2^s \mathcal{H}^s(E).$$

Thanks to this estimate, the Hausdorff dimensions defined from the two mappings $\mathcal{H}^s$ and $\tilde{\mathcal{H}}^s$ are equal.

The following are useful for the Hausdorff dimension.

Proposition 4.1.7. *Let A be any set in $\mathbf{R}^N$.*

(i) *The following implications hold true:*

$$\begin{aligned} &\mathcal{H}^s(A) < +\infty \Longrightarrow dim_H(A) \leq s, \\ &\mathcal{H}^s(A) > 0 \Longrightarrow dim_H(A) \geq s; \end{aligned}$$

(ii) *Let $f : A \longrightarrow \mathbf{R}^m$ satisfying for all x and y in A,*

$$|f(x) - f(y)| \leq L|x - y|^\alpha,$$

where $L > 0$ and $\alpha > 0$ are two given positive constants. Then $dim_H(f(A)) \leq \frac{1}{\alpha} dim_H(A)$.

PROOF. The proof of assertion (i) is a straightforward consequence of the definition. Let us prove (ii). Indeed $s > \dim_H(A) \Longrightarrow \mathcal{H}^s(A) = 0$ and inequality $\mathcal{H}^{s/\alpha}(f(A)) \leq \frac{c_{s/\alpha}}{c_s} L^{s/\alpha}\mathcal{H}^s(A)$ established in Proposition 4.1.5 yields $\mathcal{H}^{s/\alpha}(f(A)) = 0$ so that

$$\frac{s}{\alpha} \geq \dim_H(f(A)).$$

The conclusion follows by letting $s \to \dim_H(A)$. □

Example 4.1.1. When U is an open subset of $\mathbf{R}^N$, then $\dim_H(U) = N$. Indeed, U contains an open ball B and $\mathcal{H}^N(U) \geq \mathcal{H}^N(B) > 0$, and thus $\dim_H(U) \geq N$.

Example 4.1.2. Every countable subset A of $\mathbf{R}^N$ is a set of zero Hausdorff dimension. Indeed, by σ-additivity, one has

$$\mathcal{H}^s(A) = \sum_{a \in A} \mathcal{H}^s(\{a\}).$$

Since $\mathcal{H}^0(\{a\}) = 1$, for all $s > 0$ one has $\mathcal{H}^s(\{a\}) = 0$, and thus $\mathcal{H}^s(A) = 0$. This proves that $\dim_H(A) = 0$.

Example 4.1.3. Let us consider for $N \leq m$ a one-to-one map $f : \mathbf{R}^N \longrightarrow \mathbf{R}^m$ of class $\mathbf{C}^1$. Let E be a compact subset of $\mathbf{R}^N$. We have $\dim_H(f(E)) = N$. Indeed, from Proposition 4.1.6, we have $0 < \mathcal{H}^N(f(E)) \leq C\mathcal{L}^N(E) < +\infty$, where C is a positive constant.

Example 4.1.4. Let us consider now the middle third Cantor set C of the interval $[0, 1]$. We have $\dim_H(C) = \frac{\ln 2}{\ln 3}$. Indeed,

$$C = \left(C \cap \left[0, \frac{1}{3}\right]\right) \cup \left(C \cap \left[\frac{2}{3}, 1\right]\right).$$

Set $C_1 := C \cap [0, \frac{1}{3}]$ and $C_2 := C \cap [\frac{2}{3}, 1]$. These two sets are geometrically similar to C by a ratio $\frac{1}{3}$. According to the properties of the Hausdorff measure established in Proposition 4.1.5, we derive

$$\mathcal{H}^s(C) = \frac{2}{3^s}\mathcal{H}^s(C).$$

The conclusion follows if we assume that C is an s-set, that is, $0 < \mathcal{H}^s(C) < +\infty$. For a proof of this property, consult [133], where various interesting methods are given for the calculation of Hausdorff dimensions of fractal sets.

More generally, let $S_1, \ldots, S_m : \mathbf{R}^N \to \mathbf{R}^N$ be m similarities, i.e., satisfying

$$|S_i(x) - S_i(y)| = c_i|x - y|$$

for all $x,\ y$ in $\mathbf{R}^N$, where $0 < c_i < 1$ (the ratio of S_i). We assume that the S_i satisfy the *open set condition*, that is, there exists a nonempty bounded open set U such that

$$\bigcup_{i=1}^{m} S_i(U) \subset U.$$

Consider now a so-called self-similar set F satisfying $F = \bigcup_{i=1}^{m} S_i(F)$. Then $\dim_H F = s$, where s is the solution of

$$\sum_{i=1}^{m} c_i^s = 1$$

and $0 < \mathcal{H}^s(F) < +\infty$. For a proof, consult [133].

Note that the middle third Cantor set is such a self-similar set by taking $S_1(x) = \frac{1}{3}x$ and $S_2(x) = \frac{1}{3}x + \frac{2}{3}$ and the *open set condition* holds for $U =]0, 1[$.

The Hausdorff dimension of a set gives us some information about its topological structure, as stated in the proposition below.

Proposition 4.1.8. *A subset E of $\mathbf{R}^N$ with $dim_H(E) < 1$ is totally disconnected: no two of its points lie in the same connected component.*

Proof. Let x and y be distinct elements of E and consider the distance function d_x to x in $\mathbf{R}^N$, that is, $d_x(z) = |z - x|$. As d_x does not increase distances, from Proposition 4.1.7 we deduce that $\dim_H(d_x(E)) \leq \dim_H(E) < 1$. Thus $d_x(E)$ is a subset of $\mathbf{R}$ of $\mathcal{H}^1$ measure (or Lebesgue measure) zero. Consequently, there exists $r > 0$ such that $r < d_x(y)$ and $r \notin d_x(E)$ (otherwise $(0, d_x(y)) \subset d_x(E)$). The set E is then the union of the two disjoint open sets

$$E = \{z \in E : d_x(z) < r\} \cup \{z \in E : d_x(z) > r\}$$

where x is in one set and y is in the other, so that x and y lie in different connected components. □

We end this section by giving a nontrivial set in $\mathbf{R}$, with null Lebesgue measure but with Hausdorff dimension equal to one.

Proposition 4.1.9. *There exists a compact set E of $[0, 1]$ such that $\mathcal{H}^1(E) = 0$ and $dim_H(E) = 1$.*

Proof. We make use of a Cantor-type set construction but reduce the proportion of intervals removed at each stage. More precisely, we define a family $(K_n)_{n\in\mathbf{N}}$ of closed sets as follows: $K_0 = [0, 1]$, $K_1 = K_0 \setminus I_1^1$, where I_1^1 is an open interval centered at $1/2$ with length l_0, $0 < l_0 < 1$; for all $n > 1$, K_n is the union of 2^n closed disjoint intervals I_i^n with the same length l_n and K_{n+1} is obtained by removing from each I_i^n an open interval of length $\frac{l_n}{n+1}$ and centered at the center of I_i^n. Let us show that the compact set $E = \cap_{n\in\mathbf{N}} K_n$ answers the question.

A straightforward calculation gives $l_n = 2^{-n}(1 - l_0)/n$ so that $\mathcal{H}^1(E) = \lim_{n\to+\infty} \mathcal{H}^1(K_n) = 0$. Now fix $0 < r < 1$ and define the integer (depending on r),

$$n_0 = \sup\{n \in \mathbf{N} : \exists i \in \{1, \ldots, 2^n\},\ B_r(x) \cap E \subset I_i^n\}.$$

It is easily seen that

$$r \geq \frac{l_{n_0}}{n_0 + 1} = \frac{2^{-n_0}(1 - l_0)}{n_0(n_0 + 1)}. \tag{4.3}$$

On the other hand, for each $0 < \alpha < 1$, there exists a positive constant (i.e., independent of n_0) $C(\alpha, l_0)$ such that

$$2^{-n_0} \leq C(\alpha, l_0) \frac{2^{-\alpha n_0}(1-l_0)^\alpha}{n_0^\alpha (n_0+1)^\alpha}; \tag{4.4}$$

indeed, it is enough to take

$$C(\alpha, l_0) = (1-l_0)^{-\alpha} \max\left\{x^\alpha (1+x)^\alpha 2^{-(1-\alpha)x} \ : \ x \geq 1\right\}.$$

Consider now the outer measure μ in $\mathbf{R}$ satisfying $\mu(I_i^n) = 2^{-n}$ for all n and all $i = 1, \ldots, 2^n$, defined following the general construction of Remark 4.1.2. Collecting (4.3) and (4.4), we obtain for all $x \in E$ and all $0 < r < 1$,

$$\mu(B_r(x) \cap E) \leq C(\alpha, l_0) r^\alpha.$$

This estimate yields $\dim_H(E) \geq \alpha$. Indeed, if $(B_i)_{i \in \mathbf{N}}$ is a δ-covering of E made up of balls B_i centered in E with $\delta < 1$,

$$\sum_{i \in \mathbf{N}} c_\alpha \mathrm{diam}(B_i)^\alpha \geq \frac{\omega_\alpha}{C(\alpha, l_0)} \sum_{i \in \mathbf{N}} \mu(B_i \cap E) \geq \frac{\omega_\alpha}{C(\alpha, l_0)} \mu(E) = \frac{\omega_\alpha}{C(\alpha, l_0)},$$

so that $\tilde{\mathcal{H}}^\alpha(E) > 0$, then $\mathcal{H}^\alpha(E) > 0$ and $\dim_H(E) \geq \alpha$.

Since $\alpha < 1$ is arbitrary, we have $\dim_H(E) \geq 1$. The opposite inequality is trivial; then $\dim_H(E) = 1$. □

4.2 Set functions and duality approach to Borel measures

4.2.1 Borel measures as set functions

Let Ω be a separable locally compact topological space (for example, $\mathbf{R}^N$ or an open subset of $\mathbf{R}^N$) and $\mathcal{B}(\Omega)$ its Borel field. We denote the set of all $\mathbf{R}^m$-valued Borel measures by $\mathbf{M}(\Omega, \mathbf{R}^m)$. Let us recall that $\mathbf{M}(\Omega, \mathbf{R}^m)$ is the vectorial space of all the set functions $\mu : \mathcal{B}(\Omega) \to \mathbf{R}^m$ satisfying $\mu(\emptyset) = 0$ and the σ-additivity condition:

$$\mu\left(\bigcup_{n \in \mathbf{N}} B_n\right) = \sum_{n \in \mathbf{N}} \mu(B_n) \text{ for all pairwise disjoint family } (B_n)_{n \in \mathbf{N}} \text{ in } \mathcal{B}(\Omega).$$

In the case when $m = 1$, we will use the notation $\mathbf{M}(\Omega)$ (or $\mathbf{M}_b(\Omega)$) and the elements of $\mathbf{M}(\Omega)$ are called signed Borel measures. The subset of its nonnegative elements is denoted by $\mathbf{M}^+(\Omega)$.

If A is a fixed Borel subset of Ω, the restriction to A of a Borel measure $\mu \in \mathbf{M}(\Omega, \mathbf{R}^m)$ is the Borel measure $\mu \lfloor A$ of $\mathbf{M}(\Omega, \mathbf{R}^m)$ defined for all Borel set E of Ω, by $\mu \lfloor A(E) = \mu(E \cap A)$. It is worth noticing that Section 4.1 provides many concrete Borel measures on $\mathbf{R}^N$. Indeed, if A is a Borel subset in $\mathbf{R}^N$ satisfying $\mathcal{H}^s(A) < +\infty$, then $\mu = \mathcal{H}^s \lfloor A$ belongs to $\mathbf{M}^+(\mathbf{R}^N)$.

The support of a measure $\mu \in \mathbf{M}(\Omega, \mathbf{R}^m)$ is the smallest closed set E of Ω, denoted by $\mathrm{spt}(\mu)$, such that $|\mu|(\Omega \setminus E) = 0$. As a straightforward consequence of the definition, we also have

$$\mathrm{spt}(\mu) = \{x \in \Omega \ : \ \forall r > 0, \ |\mu|(B_r(x)) > 0\}.$$

Let us recall that if $\mu \in \mathbf{M}^+(\Omega)$, the measure $\mu(B)$ of all Borel subset B of Ω can be approximated by the measures of the open or compact subsets of Ω. More precisely, the following holds.

Proposition 4.2.1. *The Borel measures μ in $\mathbf{M}^+(\Omega)$ are regular, i.e., for all Borel set B of Ω, one has*

$$\begin{aligned}\mu(B) &= \sup\{\mu(K) : K \subset B,\ K \text{ compact set of } \Omega\},\\ &= \inf\{\mu(U) : B \subset U,\ U \text{ open set of } \Omega\}.\end{aligned}$$

The total variation of a measure $\mu \in \mathbf{M}(\Omega, \mathbf{R}^m)$ is the real-valued set function $|\mu|$, defined for all Borel set B of Ω by

$$|\mu|(B) = \sup\left\{\sum_{i=0}^{\infty} |\mu(B_i)| \ :\ \bigcup_{i=0}^{\infty} B_i = B\right\},$$

where the supremum is taken over all the partitions of B in $\mathcal{B}(\Omega)$. We point out that for all μ in $\mathbf{M}(\Omega, \mathbf{R}^m)$ we automatically have $|\mu|(\Omega) < +\infty$ (μ is bounded) and $|\mu|$ is σ-additive, so that $|\mu|$ is a Borel measure in $\mathbf{M}^+(\Omega)$. Actually, it is easily seen that $|\mu|$ is the smallest nonnegative scalar Borel measure ν such that $|\mu(B)| \leq \nu(B)$ for all Borel sets B and that the mapping $\mu \mapsto |\mu|(\Omega)$ is a norm for which $\mathbf{M}(\Omega, \mathbf{R}^m)$ is a Banach space.

In the scalar case, for all $\mu \in \mathbf{M}(\Omega)$ we define in $\mathbf{M}^+(\Omega)$ the positive part μ^+ and the negative parts μ^- of μ by

$$\mu^+ = \frac{|\mu| + \mu}{2}, \qquad \mu^- = \frac{|\mu| - \mu}{2},$$

so that $\mu = \mu^+ - \mu^-$ and $|\mu| = \mu^+ + \mu^-$. We define now the nonnegative Radon measures as the locally finite nonnegative Borel measures.

Definition 4.2.1. *A set function $\mu : \mathcal{B}(\Omega) \to [0, +\infty]$ such that for all $\Omega' \subset\subset \Omega$ its restrictions to $\mathcal{B}(\Omega')$ is a Borel measure on Ω', is called a nonnegative Radon measure.*

Remark 4.2.1. It is easily seen that the nonnegative Radon measures are regular.

Given a measure λ in $\mathbf{M}^+(\Omega)$, we denote the set of all Borel functions $f : \Omega \to \mathbf{R}^m$ such that

$$\int_{\Omega} |f|\, d\lambda < +\infty$$

by $L^1_\lambda(\Omega, \mathbf{R}^m)$ or by $L^1_\lambda(\Omega)$ when $m = 1$. Given now two measures $\mu \in \mathbf{M}(\Omega, \mathbf{R}^m)$ and $\lambda \in \mathbf{M}^+(\Omega)$, we say that the measure μ is absolutely continuous with respect to the measure λ and we write $\mu << \lambda$ iff

$$\forall B \in \mathcal{B}(\Omega),\ \lambda(B) = 0 \Longrightarrow \mu(B) = 0.$$

We say that the measure μ is singular with respect to the measure λ and we write $\mu \perp \lambda$ iff there exists B in $\mathcal{B}(\Omega)$ such that $\lambda(B) = 0$ and μ is concentrated on B, i.e., $\mu(C) = 0$ for all Borel set C such that $B \cap C = \emptyset$.

One may establish that

$$\mu << \lambda \iff \exists f \in L^1_\lambda(\Omega, \mathbf{R}^m) \text{ s.t. } \mu = f\lambda.$$

The following theorem extends this result.

Theorem 4.2.1 (Radon–Nikodym). *Let μ and λ be two Borel measures, respectively, in $\mathbf{M}(\Omega, \mathbf{R}^m)$ and $\mathbf{M}^+(\Omega)$. Then, there exists a function f in $L^1_\lambda(\Omega, \mathbf{R}^m)$ and a measure μ^s in $\mathbf{M}(\Omega, \mathbf{R}^m)$ such that*

$$\mu = f\lambda + \mu^s, \quad \mu^s \perp \lambda.$$

Moreover, f is given by

$$f(x) = \lim_{\rho \to 0} \frac{\mu(B_\rho(x))}{\lambda(B_\rho(x))} \quad \text{for } \mu \text{ a.e. } x \in \Omega,$$

where $B_\rho(x)$ denotes the open ball in Ω centered at $x \in \Omega$ with radius ρ.

For details about these elementary notions and various proofs of these properties, we refer the reader to the books Buchwalter [94], Marle [178], and Rudin [206].

We now give three useful lemmas concerning Borel measures in $\mathbf{M}^+(\Omega)$. The first states that one can reduce families of pairwise disjoint Borel sets to a countable subfamily as long as we consider their measures. The second is a localization lemma in $\mathbf{M}^+(\Omega)$. The last lemma is a basic result concerning the s-density of a nonnegative Radon measure and is an essential tool in the study of the structure of sets with finite perimeter (see chapter 10).

Lemma 4.2.1. *Let μ be a Borel measure in $\mathbf{M}^+(\Omega)$ and $(B_i)_{i \in I}$ a family of pairwise disjoint Borel subsets of Ω. Then the subset of indices $i \in I$ such that $\mu(B_i) \neq 0$ is at most countable.*

Proof. Since $\{i \in I : \mu(B_i) \neq 0\} = \bigcup_{n \in \mathbf{N}^*} \{i \in I : \mu(B_i) > \frac{1}{n}\}$, it is enough to prove that each set $I_n = \{i \in I : \mu(B_i) > \frac{1}{n}\}$ is finite. Assume that I_n contains an infinite sequence of indices in I, that is, $\{i_0, \ldots, i_k, \ldots\} \subset I$. We then have

$$+\infty > \mu\left(\bigcup_{k \in \mathbf{N}} B_{i_k}\right) = \sum_{k \in \mathbf{N}} \mu(B_{i_k}) = +\infty,$$

a contradiction. □

Example 4.2.1. Let $\mu \in \mathbf{M}(\Omega, \mathbf{R}^m)$ and $B_r(x_0)$ be the open ball in Ω centered at x_0 with radius r. Then for all but countably many r in $\mathbf{R}^+$, one has

$$\int_{\partial B_r(x_0)} |\mu| = 0.$$

Lemma 4.2.2. *Let $\mu \in \mathbf{M}^+(\Omega)$, $(f_i)_{i \in \mathbf{N}}$ be a family of nonnegative functions in $L^1_\mu(\Omega)$ and set $f = \sup_i f_i$. Then*

$$\int_\Omega f \, d\mu = \sup \left\{ \sum_{i \in I} \int_{A_i} f_i \, d\mu \right\},$$

where the supremum is taken over all finite families $(A_i)_{i\in I}$ of pairwise disjont open subsets of Ω.

PROOF. Let n be a fixed element of $\mathbf{N}$ and consider the μ-measurable sets

$$E_i := \left\{ x \in \Omega \ : \ \sup_{0\le k\le n} f_k(x) = f_i(x) \right\}.$$

We now construct the following family $(\Omega_i)_{i=0,\dots,n}$ of pairwise disjoint μ-measurable sets:

$$\Omega_0 = E_0, \qquad \Omega_{i+1} = \left(\Omega \setminus \bigcup_{k=1}^{i} E_k \right) \cap E_{i+1}, \quad i = 0, \dots, n-1.$$

It is easy to check that $\Omega = \bigcup_{i=0}^n E_i = \bigcup_{i=0}^n \Omega_i$. Moreover,

$$\begin{aligned} \int_\Omega \sup_{0\le k\le n} f_k \, d\mu &= \int_\Omega 1_{\bigcup_{i=0}^n \Omega_i} \sup_{0\le k\le n} f_k \, d\mu \\ &= \sum_{i=0}^n \int_{\Omega_i} \sup_{0\le k\le n} f_k \, d\mu \\ &= \sum_{i=0}^n \int_{\Omega_i} f_i \, d\mu. \end{aligned}$$

Let $\mu_i = f_i \cdot \mu$ be the Borel measure in $\mathbf{M}^+(\Omega)$ whose density with respect to μ is f_i. Since each measure μ_i is regular, one has $\mu_i(\Omega_i) = \sup\{\mu_i(K) : K \text{ compact subset of } \Omega_i\}$, hence

$$\int_\Omega \sup_{0\le i\le n} f_i \, d\mu = \sup \left\{ \sum_{i=0}^n \int_{K_i} f_i \, d\mu : (K_i)_{i=1,\dots,n} \text{ pairwise disjoint compact sets} \right\}.$$

From the regularity property of μ_i again, and by compactness, for all $\varepsilon > 0$, there exists a family $(\mathcal{O}_i)_{i=0,\dots,n}$ of open sets, that one may assume pairwise disjoint, each $\mathcal{O}_i$ containing K_i, such that $\mu_i(\mathcal{O}_i \setminus K_i) < \varepsilon/n$. Therefore

$$\int_\Omega \sup_{0\le i\le n} f_i \, d\mu = \sup \left\{ \sum_{i=0}^n \int_{A_i} f_i \, d\mu : (A_i)_{i=1,\dots,n} \text{ pairwise disjoint open sets} \right\}.$$

Taking the supremum on n and by the monotone convergence theorem, we obtain

$$\int_\Omega f \, d\mu \le \sup \left\{ \sum_{i\in I} \int_{A_i} f_i \, d\mu \right\}.$$

Since the converse inequality is obvious, the proof is complete. □

Example 4.2.2. Let Ω be a bounded open set of $\mathbf{R}^N$ and let us consider the following measure associated with a given function u in the space $SBV(\Omega)$ defined in Section 10.5:

S_u is a hypersurface in Ω, which is the union of a negligible subset of Ω for the $N-1$-Hausdorff measure, and of countable many C^1-hypersurfaces with Hausdorff dimension $N-1$ (S_u is the jump set of u). For H^{N-1} almost every $x \in S_u$, $\nu_u(x)$ is a normal unit vector at x and $\mu := \mathcal{L}^N \lfloor \Omega + \nu_u \mathcal{H}^{N-1} \lfloor S_u$, where $\mathcal{L}^N$ denotes the Lebesgue measure in $\mathbf{R}^N$. On the other hand, let us consider the functions $f_\nu := |\nabla u \cdot \nu|^p + |\nu_u \cdot \nu| 1_{S_u}$ indexed by a countable dense subset D of ν in S^{N-1}. We have

$$\begin{aligned}\int_\Omega \sup_{\nu \in D} f_\nu \, d\mu &= \int_{\Omega \setminus S_u} \sup_{\nu \in D} f_\nu \, d\mu + \int_{S_u} \sup_{\nu \in D} f_\nu \, d\mu \\ &= \int_\Omega |\nabla u|^p \, dx + \mathcal{H}^{N-1}(S_u),\end{aligned}$$

which is the Mumford–Shah energy functional introduced in image segmentation and studied in Sections 12.4 and 14.3.

By applying Lemma 4.2.2 now, we then obtain the following formula which is a central point to extend to arbitrary dimension, the approximation of the Mumford–Shah energy functional established in one dimension:

$$\int_\Omega |\nabla u|^p \, dx + \mathcal{H}^{N-1}(S_u) = \sup \left\{ \sum_{i \in I} \left(\int_{A_i} |\nabla u \cdot \nu_i|^p \, dx + \int_{A_i} |\nu_v \cdot \nu_i| \, d\mathcal{H}^{N-1} \lfloor S_u \right) \right\}.$$

The supremum is taken over all finite families $(A_i)_{i \in I}$ of pairwise disjoint open subsets of Ω. Note that for $p = 1$, we obtain the expression of the total variation of the measure $Du := \nabla u \, \mathcal{L}^N \lfloor \Omega + [u] \nu_u \, \mathcal{H}^{N-1} \lfloor S_u$. For more details on the spaces $BV(\Omega)$ and $SBV(\Omega)$, see chapter 10, and for the definition and properties concerning the Mumford–Shah energy functional, consult Sections 12.4 and 14.3. For another application of the localization Lemma 4.2.2, we refer the reader to Section 13.3.

Lemma 4.2.3. *Let Ω be an open bounded subset of $\mathbf{R}^N$, μ a nonnegative Radon measure on Ω, and $0 < s \leq N$, $t > 0$. For each Borel subset E of Ω, the following implication holds:*

$$\forall x \in E, \ \limsup_{\rho \to 0} \frac{\mu(B_\rho(x))}{\rho^s} > t \Longrightarrow \mu \geq Ct \mathcal{H}^s \lfloor E,$$

where C is a positive constant depending only on s. When s is an integer, one may take $C = \omega_s$, the volume of the unit ball of $\mathbf{R}^s$.

Proof. One may assume $\mu(E) < +\infty$. Since $\mu(E) = \inf\{\mu(U) : E \subset U, \ U \text{ open set of } \Omega\}$, one may choose an arbitrary open subset U of Ω such that $E \subset U$ and $\mu(U) < +\infty$. Let now $\delta > 0$ and consider the family of closed balls

$$\mathcal{F} := \left\{ \overline{B}_\rho(x) \subset U : x \in E, \ \rho \leq \frac{\delta}{2}, \ \frac{\mu(B_\rho(x))}{\rho^s} \geq t \right\}.$$

From the hypothesis, it is easily seen that this family finely covers E (Definition 4.1.3). Therefore, according to Lemma 4.1.3, there exists a countable subfamily $\mathcal{G}$ of pairwise

disjoint elements of $\mathcal{F}$ such that

$$\bigcup_{B\in\mathcal{F}} B \subset \bigcup_{B\in\mathcal{G}} B^*,$$

where B^* denotes the closed ball concentric with B, with radius five times as big as that of B. Moreover, for each finite family $\mathcal{G}^* \subset \mathcal{G}$,

$$E \subset \left(\bigcup_{B\in\mathcal{G}^*} B\right) \cup \left(\bigcup_{B\in\mathcal{G}\setminus\mathcal{G}^*} B^*\right).$$

Thus,

$$\begin{aligned}\mathcal{H}^s_{5\delta}(E) &\le c_s \sum_{B\in\mathcal{G}^*} (\operatorname{diam}(B))^s + c_s 5^s \sum_{B\in\mathcal{G}\setminus\mathcal{G}^*} (\operatorname{diam}(B))^s \\ &\le 2^s c_s t^{-1} \left(\sum_{B\in\mathcal{G}^*} \mu(B)\right) + 2^s c_s 5^s t^{-1} \left(\sum_{B\in\mathcal{G}\setminus\mathcal{G}^*} \mu(B)\right).\end{aligned}$$

Since the second sum of the last inequality can be taken less than δ for an appropriate choice of $\mathcal{G}^*$, we obtain

$$\begin{aligned}\mathcal{H}^s_{5\delta}(E) &\le \omega_s t^{-1} \left(\sum_{B\in\mathcal{G}^*} \mu(B)\right) + \delta \\ &\le \omega_s t^{-1} \mu(U) + \delta.\end{aligned}$$

We end the proof by letting $\delta \to 0$ and taking the infimum on U. □

4.2.2 Duality approach

We recall now some definitions and important results concerning the Riesz functional analysis approach. We set $\mathbf{C}_0(\Omega, \mathbf{R}^m)$ to denote the space of all continuous functions which tend to zero at infinity, i.e.,

$$\forall \varepsilon > 0, \text{ there exists a compact set } K_\varepsilon \subset \Omega \text{ such that } \sup_{x\in\Omega\setminus K_\varepsilon} |\varphi(x)| \le \varepsilon.$$

Recall that equipped with the norm

$$\|\varphi\|_\infty = \sup_{x\in\Omega} |\varphi(x)|,$$

$\mathbf{C}_0(\Omega, \mathbf{R}^m)$ is a Banach space. We denote its subspace made up of all continuous functions with compact support in Ω by $\mathbf{C}_c(\Omega, \mathbf{R}^m)$. When $m = 1$, the two above spaces will be denoted, respectively, by $\mathbf{C}_0(\Omega)$ and $\mathbf{C}_c(\Omega)$.

According to the vectorial version of the Riesz–Alexandroff representation Theorem 2.4.6, the dual of $\mathbf{C}_0(\Omega, \mathbf{R}^m)$ (and then of $\mathbf{C}_c(\Omega, \mathbf{R}^m)$) can be isometrically identified with $\mathbf{M}(\Omega, \mathbf{R}^m)$. Then, any Borel measure is a continuous linear form on $\mathbf{C}_0(\Omega, \mathbf{R}^m)$ or

$\mathbf{C}_c(\Omega, \mathbf{R}^m)$ and the two dual norms $\|\cdot\|_{\mathbf{C}_c'(\Omega,\mathbf{R}^m)}$ and $\|\cdot\|_{\mathbf{C}_0'(\Omega,\mathbf{R}^m)}$ are equal to the total mass $|\cdot|(\Omega)$:

$$\begin{aligned} |\mu|(\Omega) &= \sup\{\langle \mu, \varphi\rangle : \varphi \in \mathbf{C}_0(\Omega, \mathbf{R}^m), ||\varphi||_\infty \le 1\} \\ &= \sup\{\langle \mu, \varphi\rangle : \varphi \in \mathbf{C}_c(\Omega, \mathbf{R}^m), ||\varphi||_\infty \le 1\}, \end{aligned}$$

where $\langle \mu, \varphi\rangle = \mu(\varphi)$. In what follows, $\langle \mu, \varphi\rangle$ will also be denoted by $\int_\Omega \varphi \, d\mu$. Note that $\mathbf{M}(\Omega, \mathbf{R}^m)$ is isomorphic to the product space $\mathbf{M}(\Omega)^m$ and that according to this isomorphism

$$\mu \in M(\Omega, \mathbf{R}^m) \iff \mu = (\mu_1, \ldots, \mu_m) \text{ and } \mu_i \in \mathbf{C}_0(\Omega)', \ i = 1, \ldots, m.$$

The following proposition is an easy generalization, for vectorial measures, of Proposition 2.4.13 and Corollary 2.4.1. To shorten notations, $\sigma(\mathbf{C}_0', \mathbf{C}_0)$ and $\sigma(\mathbf{C}_c', \mathbf{C}_c)$ denote, respectively, the two weak topologies $\sigma(\mathbf{C}_0'(\Omega, \mathbf{R}^m), \mathbf{C}_0(\Omega, \mathbf{R}^m))$ and $\sigma(\mathbf{C}_c'(\Omega, \mathbf{R}^m), \mathbf{C}_c(\Omega, \mathbf{R}^m))$.

Proposition 4.2.2. *The weak topologies $\sigma(\mathbf{C}_0', \mathbf{C}_0)$ and $\sigma(\mathbf{C}_c', \mathbf{C}_c)$ induce the same toplogy on bounded subsets of $\mathbf{M}(\Omega, \mathbf{R}^m)$. Moreover, from any bounded sequence of Borel measures $(\mu_n)_{n\in\mathbf{N}}$ in $\mathbf{M}(\Omega, \mathbf{R}^m)$, one can extract a subsequence $\sigma(\mathbf{C}_0', \mathbf{C}_0)$-converging (thus $\sigma(\mathbf{C}_c', \mathbf{C}_c)$-converging) to some Borel measure μ in $\mathbf{M}(\Omega, \mathbf{R}^m)$.*

Note that for unbounded sequences of $\mathbf{M}(\Omega, \mathbf{R}^m)$, the two weak topologies do not agree, as illustrated in the following example: take $\Omega = (0, +\infty)$ and $\mu_n = n\delta_n$. Then

$$n\delta_n \rightharpoonup 0 \text{ in the topology } \sigma(\mathbf{C}_c'(\Omega), \mathbf{C}_c(\Omega))$$

but $\langle n\delta_n, \varphi\rangle \to 1$ as $n \to +\infty$ for any function of $\mathbf{C}_0(\Omega)$ satisfying $\varphi \sim \frac{1}{x}$ in the neighborhood of $+\infty$.

According to Proposition 4.2.2, from now on, for bounded sequences in $\mathbf{M}(\Omega, \mathbf{R}^m)$, we do not distinguish the convergences associated with these two weak topologies and we will refer to these convergences as the weak convergence in $\mathbf{M}(\Omega, \mathbf{R}^m)$. In terms of probabilistic approach, we have the following properties.

Proposition 4.2.3 (Alexandrov). *Let $\mu, \ \mu_n$ in $\mathbf{M}^+(\Omega)$ such that μ_n weakly converges to μ; then*

$$\begin{aligned} &\text{for all open subset } U \text{ of } \Omega, \ \mu(U) \le \liminf_{n\to+\infty} \mu_n(U), \\ &\text{for all compact subset } K \text{ of } \Omega, \ \mu(K) \ge \limsup_{n\to+\infty} \mu_n(K). \end{aligned}$$

Consequently, for all relatively compact Borel subset B of Ω such that $\mu(\partial B) = 0$, we have

$$\mu(B) = \lim_{n\to+\infty} \mu_n(B).$$

PROOF. Let U be an open subset of Ω and 1_U its characteristic function. Classically, there exists a nondecreasing sequence $(\varphi_p)_{p\in\mathbf{N}}$ in $\mathbf{C}_c(\Omega)$ such that $1_U = \sup_p \varphi_p$. Therefore

$$\begin{aligned}\mu(U) &= \lim_{p\to+\infty} \int_\Omega \varphi_p \, d\mu \\ &= \lim_{p\to+\infty} \lim_{n\to+\infty} \int_\Omega \varphi_p \, d\mu_n \\ &\le \liminf_{n\to+\infty} \int_U d\mu_n = \liminf_{n\to+\infty} \mu_n(U).\end{aligned}$$

For the other assertion, it suffices to notice that if K is a compact subset of Ω, there exists a nonincreasing sequence $(\varphi_p)_{p\in\mathbf{N}}$ in $\mathbf{C}_c(\Omega)$ such that $1_K = \inf_p \varphi_p$ and to argue similarly.

Let us prove the last assertion. Since μ_n is a nondecreasing set function, according to the first assertions, we have

$$\begin{aligned}\mu(\overset{\circ}{B}) &\le \liminf_{n\to+\infty} \mu_n(\overset{\circ}{B}) \le \liminf_{n\to+\infty} \mu_n(B) \\ &\le \limsup_{n\to+\infty} \mu_n(B) \le \limsup_{n\to+\infty} \mu_n(\overline{B}) \le \mu(\overline{B}).\end{aligned}$$

The conclusion follows from $\mu(\overline{B}) = \mu(\overset{\circ}{B})$. □

The following corollary clarifies the relation between the weak convergence of sequences in $\mathbf{M}(\Omega, \mathbf{R}^m)$ and the convergence of the corresponding measures of suitable Borel sets.

Corollary 4.2.1. *Let μ, μ_n in $\mathbf{M}(\Omega, \mathbf{R}^m)$ be such that μ_n weakly converges to μ and $|\mu_n|$ weakly converges to some σ in $\mathbf{M}^+(\Omega)$. Then $|\mu| \le \sigma$ and, for all relatively compact Borel subset B of Ω such that $\sigma(\partial B) = 0$, we have $\mu(B) = \lim_{n\to+\infty} \mu_n(B)$.*

Proof. Let U be an open subset of Ω and φ any function in $\mathbf{C}_c(U, \mathbf{R}^m)$ with $\|\varphi\|_\infty \le 1$. Letting $n \to +\infty$ in

$$\Big| \int_U \varphi \, d\mu_n \Big| \le \int_U |\varphi| \, d|\mu_n|,$$

we obtain

$$\Big| \int_U \varphi \, d\mu \Big| \le \int_U |\varphi| \, d\sigma.$$

Taking the supremum on φ gives $|\mu|(U) \le \sigma(U)$. The conclusion $|\mu| \le \sigma$ follows from Proposition 4.2.1.

For proving the last assertion, let us denote the m components of μ_n and μ by μ_n^i and μ^i, $i = 1, \dots, m$, respectively, and the weak limits (for a subsequence not relabeled) of the positive and negative parts of μ_n^i by $\nu^{i,+}$ and $\nu^{i,-}$, respectively. Going to the limit when $n \to +\infty$ in $\mu_n^i = \mu_n^{i,+} - \mu_n^{i,-}$ and $\mu_n^{i,\pm} \le |\mu_n|$, we obtain $\mu^i = \nu^{i,+} - \nu^{i,-}$ and $\nu^{i,\pm} \le \sigma$. The conclusion then follows by applying Proposition 4.2.3 to the m components $\mu_n^{i\pm}$. □

We now restrict ourselves to the case when $\Omega = \mathbf{R}^N$ and we study the approximation of a Borel measure in $\mathbf{M}(\mathbf{R}^N, \mathbf{R}^m)$ by a regular function in the sense of the weak convergence of measures. To this end, we define the regularization (or the mollification) of a measure by means of a regularizing kernel. Let us recall that a regularizer ρ_ε is a function in $\mathbf{C}_c^\infty(\mathbf{R}^N)$

defined by $\rho_\varepsilon(x) = \varepsilon^{-N}\rho(x/\varepsilon)$, where ρ is some nonnegative real-valued function in $\mathbf{C}_c^\infty(\mathbf{R}^N)$ satisfying

$$\int_{\mathbf{R}^N} \rho(x)\,dx = 1, \qquad \mathrm{spt}\rho \subset \overline{B}_1(0).$$

Note that the support $\mathrm{spt}\,\rho_\varepsilon$ of ρ_ε is included in $\overline{B}_\varepsilon(0)$. For any measure μ in $\mathbf{M}(\mathbf{R}^N, \mathbf{R}^m)$, we define the function $\rho_\varepsilon * \mu$ defined on $\mathbf{R}^N$ by

$$\rho_\varepsilon * \mu(x) = \int_{\mathbf{R}^N} \rho_\varepsilon(x-y)\,\mu(dy)$$

and we aim to show that $\rho_\varepsilon * \mu$ is a suitable approximation of μ.

Theorem 4.2.2. *The functions $\rho_\varepsilon * \mu$ belong to $\mathbf{C}^\infty(\mathbf{R}^N, \mathbf{R}^m)$ and for any $\alpha \in \mathbf{N}^N$, $D^\alpha(\rho_\varepsilon * \mu) = D^\alpha\rho_\varepsilon * \mu$. Moreover, when ε goes to zero,*

(i) $\rho_\varepsilon * \mu \rightharpoonup \mu$, *weakly in* $\mathbf{M}(\mathbf{R}^N, \mathbf{R}^m)$;

(ii) $\displaystyle\int_{\mathbf{R}^N} |\rho_\varepsilon * \mu| \le \int_{\mathbf{R}^N} |\mu|$;

(iii) $\displaystyle\int_{\mathbf{R}^N} |\rho_\varepsilon * \mu| \to \int_{\mathbf{R}^N} |\mu|$.

Proof. The classical derivation theorem under the integral sign yields $D^\alpha(\rho_\varepsilon * \mu) = D^\alpha\rho_\varepsilon * \mu$. Let us establish (i). Let $\varphi \in \mathbf{C}_c(\mathbf{R}^N, \mathbf{R}^m)$. According to Fubini's theorem,

$$\begin{aligned}
&\left|\int_{\mathbf{R}^N} \varphi(x)\,\rho_\varepsilon * \mu(x)\,dx - \int_{\mathbf{R}^N} \varphi(y)\,\mu(dy)\right| \\
&= \left|\int_{\mathbf{R}^N}\int_{\mathbf{R}^N} \varphi(x)\rho_\varepsilon(x-y)\,\mu(dy)\,dx - \int_{\mathbf{R}^N} \varphi(y)\,\mu(dy)\right| \\
&= \left|\int_{\mathbf{R}^N}\int_{\mathbf{R}^N} \varphi(x)\,\rho_\varepsilon(x-y)\,\mu(dy)\,dx - \int_{\mathbf{R}^N}\left(\int_{\mathbf{R}^N} \rho_\varepsilon(x-y)\,dx\right)\varphi(y)\,\mu(dy)\right| \\
&\le \int_{\mathbf{R}^N}\int_{\mathbf{R}^N} |\varphi(x)-\varphi(y)|\rho_\varepsilon(x-y)\,dx\,|\mu|(dy) \\
&\le \sup_{\{(x,y)\in\mathbf{R}^{2N}:|x-y|\le\varepsilon\}} |\varphi(x)-\varphi(y)| \int_{\mathbf{R}^N} |\mu|,
\end{aligned}$$

which, thanks to the uniform continuity of φ, tends to zero when $\varepsilon \to 0$.

We establish now (ii). By Fubini's theorem and a change of scale, we have

$$\begin{aligned}
\int_{\mathbf{R}^N} |\rho_\varepsilon * \mu|(x)\,dx &= \varepsilon^{-N}\int_{\mathbf{R}^N}\left|\int_{\mathbf{R}^N} \rho\left(\frac{x-y}{\varepsilon}\right)\mu(dy)\right|dx \\
&\le \varepsilon^{-N}\int_{\mathbf{R}^N}\int_{\mathbf{R}^N} \rho\left(\frac{x-y}{\varepsilon}\right)|\mu|(dy)\,dx \\
&\le \int_{\mathbf{R}^N}\left(\int_{\mathbf{R}^N} \varepsilon^{-N}\rho\left(\frac{x-y}{\varepsilon}\right)dx\right)|\mu|(dy) = \int_{\mathbf{R}^N}|\mu|.
\end{aligned}$$

Assertion (iii) is a straightforward consequence of (i), the weak lower semicontinuity of the map $\mu \mapsto \int_{\mathbf{R}^N} |\mu|$ and (ii). □

Let $\mathbf{C}_b(\Omega, \mathbf{R}^m)$ be the set of all bounded continuous functions from Ω into $\mathbf{R}^m$. We introduce now a stronger notion of convergence induced by the weak topology $\sigma(\mathbf{C}_b'(\Omega, \mathbf{R}^m), \mathbf{C}_b(\Omega, \mathbf{R}^m))$.

Definition 4.2.2. *A sequence $(\mu_n)_n$ in $M(\Omega, \mathbf{R}^m)$ narrowly converges to μ in $M(\Omega, \mathbf{R}^m)$ iff*

$$\int \varphi \, d\mu_n \to \int \varphi \, d\mu$$

for all φ in $\mathbf{C}_b(\Omega, \mathbf{R}^m)$.

This convergence is strictly stronger than the weak convergence of measures. Indeed, let $\Omega = (0, 1)$ and $\mu_n = \delta_{1/n}$. Then μ_n weakly converges to 0 but $\int_\Omega \mu_n = 1$. Moreover, taking $\varphi \sim \sin(1/x)$ at 0^+, $\int \varphi \mu_n$ does not converge for any subsequence. This example shows that the unit ball of $\mathbf{M}(\Omega)$ is not weakly sequentially compact for this topology. Nevertheless, the Prokhorov theorem below asserts that the bounded sets of $\mathbf{M}^+(\Omega)$ are sequentially compact for the narrow topology as long as a uniform control is assumed outside a compact set whose measure is close to that of Ω.

Theorem 4.2.3 (Prokhorov). *Let $\mathcal{H}$ be a bounded subset of $\mathbf{M}^+(\Omega)$ satisfying*

$$\forall \varepsilon, \exists K_\varepsilon \text{ compact subset of } \Omega \text{ such that } \sup\{\mu(\Omega \setminus K_\varepsilon) : \mu \in \mathcal{H}\} \leq \varepsilon.$$

Then $\mathcal{H}$ is sequentially compact for the narrow topology.

For a proof, consult, for instance, Delacherie and Meyer [125]. Any subset of $\mathbf{M}^+(\Omega)$ satisfying the previous uniform bound is said to be tight. Note that in the previous example, $\{\delta_{1/n} : n \in \mathbf{N}^*\}$ is not tight. In the bounded subsets of $\mathbf{M}^+(\Omega)$, the weak and the narrow topology agree when there is no loss of mass, i.e., when $\mu(\Omega) = \lim_{n \to +\infty} \mu_n(\Omega)$. More precisely, we have the following.

Proposition 4.2.4. *Let μ_n, μ in $\mathbf{M}^+(\Omega)$. Then the following assertions are equivalent:*

(i) $\mu_n \rightharpoonup \mu$ *narrowly;*

(ii) $\mu_n \rightharpoonup \mu$ *weakly and* $\mu_n(\Omega) \to \mu(\Omega)$.

Proof. We prove (ii) $\Longrightarrow$ (i). The converse is obvious. Let $f \in \mathbf{C}_b(\Omega)$, $\varepsilon > 0$ and K a compact subset of Ω such that $\mu(\Omega \setminus K) \leq \varepsilon$. Let moreover $\varphi \in \mathbf{C}_c(\Omega)$ satisfying $0 \leq \varphi \leq 1$, $\varphi = 1$ in K. We have

$$\begin{aligned}\left|\int f\,d\mu_n - \int f\,d\mu\right| &\le \left|\int f\,d\mu_n - \int f\varphi\,d\mu_n\right| + \left|\int f\varphi\,d\mu_n - \int f\varphi\,d\mu\right| \\ &\quad + \left|\int f\varphi\,d\mu - \int f\,d\mu\right| \\ &\le \|f\|_\infty \int (1-\varphi)\,d\mu_n + \left|\int f\varphi\,d\mu_n - \int f\varphi\,d\mu\right| \\ &\quad + \|f\|_\infty \int (1-\varphi)\,d\mu\end{aligned}$$

so that

$$\limsup_{n\to+\infty}\left|\int f\,d\mu_n - \int f\,d\mu\right| \le 2\varepsilon\|f\|_\infty$$

and the conclusion follows after letting $\varepsilon \to 0$. □

In the probabilistic approach we have a similar statement.

Proposition 4.2.5. *Let $\mu_n,\ \mu$ in $\mathbf{M}^+(\Omega)$. Then the following assertions are equivalent:*

(i) *μ_n narrowly converges to μ;*

(ii) *$\mu_n(\Omega) \to \mu(\Omega)$ and for all open subset $U,\ \mu(U) \le \liminf_{n\to+\infty} \mu_n(U)$;*

(iii) *$\mu_n(\Omega) \to \mu(\Omega)$ and for all closed subset $F,\ \mu(F) \ge \limsup_{n\to+\infty} \mu_n(F)$;*

(iv) *for all Borel subset B such that $\mu(\partial B) = 0,\ \mu_n(B) \to \mu(B)$.*

Proof. The only implication we have to establish is (iv) $\Longrightarrow$ (i). Indeed, each of the others is an easy consequence of Propositions 4.2.4 and 4.2.3. According to Proposition 4.2.4, it is enough to establish the weak convergence of μ_n to μ. For this, let $\varphi \in \mathbf{C}_c(\Omega)$, $\|\varphi\|_{L^\infty(\Omega)} = M$, with compact support K, and consider the subdivision

$$\begin{cases} -M = a_0 < a_1 \cdots < a_i < a_{i+1} < \cdots < a_m = M, \\ a_{i+1} - a_i \le \varepsilon, \\ \mu([\varphi = a_i]) = 0. \end{cases}$$

Such a subdivision exists. Indeed, the last property is a consequence of Lemma 4.2.1. Consider now the Borel subsets $U_i = [\varphi < a_i] \cap K$ and the function

$$\varphi_\varepsilon = \sum_{i=1}^{m} a_i \chi_{U_i \setminus U_{i-1}}.$$

Since $\mu(\partial U_i) = 0$, from assertion (iv) we have $\mu_n(U_i) \to \mu(U_i)$ so that for all $\varepsilon > 0$

$$\int \varphi_\varepsilon d\mu_n \to \int \varphi_\varepsilon d\mu.$$

From the equiboundedness of the measures μ_n and the estimate $\|\varphi_\varepsilon - \varphi\|_\infty \le \varepsilon$, we easily deduce

$$\int \varphi d\mu_n \to \int \varphi d\mu,$$

which ends the proof. □

The following proposition is an extension of property (iv). For a proof, consult Marle [178, Proposition 9.9.4].

Proposition 4.2.6. *Let μ_n, μ be Borel measures in $\mathbf{M}^+(\Omega)$ such that μ_n narrowly converges to μ and let f be a μ_n-measurable (for every n) and bounded function from Ω into $\mathbf{R}$ such that the set of its discontinuity points has a null μ-measure. Then f is μ-measurable and*

$$\lim_{n\to+\infty}\int f\,d\mu_n = \int f\,d\mu.$$

We end this subsection by stating two theorems extending in some sense the classical Fubini's theorem. We prove only the first one. For the second, see, for instance, [132].

Let μ in $\mathbf{M}^+(\Omega\times\mathbf{R}^m)$. We denote the projection of μ on Ω by σ. Let us recall that σ is the measure of $\mathbf{M}^+(\Omega)$ defined for all Borel set E of Ω by $\sigma(E)=\mu(E\times\mathbf{R}^m)$. The following slicing decomposition holds.

Theorem 4.2.4. *There exists a family $(\mu_x)_{x\in\Omega}$ of probability measures on $\mathbf{R}^m$, unique up to equality σ-a.e., such that for all f in $\mathbf{C}_0(\Omega\times\mathbf{R}^m)$*

(i) $x\mapsto\displaystyle\int_{\mathbf{R}^m} f(x,y)\,d\mu_x(y)$ *is σ-measurable;*

(ii) $\displaystyle\int_{\Omega\times\mathbf{R}^m} f(x,y)\,d\mu(x,y)=\int_\Omega\Big(\int_{\mathbf{R}^m} f(x,y)\,d\mu_x(y)\Big)\,d\sigma(x).$

We will write $\mu=(\mu_x)_{x\in\Omega}\otimes\sigma$.

Proof. *First step.* We establish the result for all f of the form $f=g\otimes h$, where $g(x)=1_B(x)$, B is any Borel subset of Ω, and h belongs to $\mathbf{C}_0(\mathbf{R}^m)$.

Let us first assume that h belongs to a dense countable subset D of $\mathbf{C}_0(\mathbf{R}^m)$ and define $\gamma_h\in\mathbf{M}^+(\Omega)$ by

$$\gamma_h(B)=\int_{B\times\mathbf{R}^m} h(y)\,d\mu(x,y)\ \forall\ B\in\mathcal{B}(\Omega).$$

Since $\mu(B\times\mathbf{R}^m)=\sigma(B)=0\Longrightarrow\gamma_h(B)=0$, according to the Radon–Nikodym theorem, Theorem 4.2.1, the measure γ_h has a density $a_h\in L^1_\sigma(\Omega)$ with respect to σ, i.e.,

$$\gamma_h=a_h\cdot\sigma.$$

We then obtain

$$\int_{B\times\mathbf{R}^m} h(y)\,d\mu(x,y)=\int_B a_h(x)\,d\sigma(x). \tag{4.5}$$

Moreover, there exists a sequence $(N_h)_{h\in D}$ of σ-null sets such that a_h is given for all $x_0\in\Omega':=\Omega\setminus(\cup_{h\in D}N_h)$ by

$$\begin{aligned} a_h(x_0)&=\lim_{\rho\to0}\frac{\gamma_h(B_\rho(x_0))}{\sigma(B_\rho(x_0))}\\ &=\lim_{\rho\to0}\frac{\int_{(B_\rho(x_0))\times\mathbf{R}^m}h(y)\,d\mu(x,y)}{\mu(B_\rho(x_0))\times\mathbf{R}^m)}. \end{aligned}\tag{4.6}$$

For all fixed x_0 in Ω', let us now consider the linear map $\Gamma_{x_0} : D \to \mathbf{R}$ defined by

$$\Gamma_{x_0}(h) = a_h(x_0).$$

From (4.6), we easily check that $|\Gamma_{x_0}(h)| \leq \|h\|_\infty$. Therefore Γ_{x_0} may be extended by a measure μ_{x_0} in $\mathbf{M}^+(\mathbf{R}^m)$. Since the map $x \mapsto 1_B(x)\mu_x(h) = 1_B(x)a_h(x)$ is σ-measurable for all $h \in D$, it is also σ-measurable for all h in $\mathbf{C}_0(\mathbf{R}^m)$. Now, from (4.5), one can write

$$\int_{B\times\mathbf{R}^m} h(y)\, d\mu(x,y) = \int_B \left(\int_{\mathbf{R}^m} h(y)\, d\mu_x(y) \right) d\sigma(x)$$

with $\|\mu_x\| \leq 1$.

Second step. We establish the result for all f of the form $f = g \otimes h$, where $g \in L^1_\sigma(\Omega)$ and $h \in \mathbf{C}_0(\mathbf{R}^m)$.

For all $\varepsilon > 0$, let us consider the step function $g_\varepsilon = \sum_{i\in I} \alpha_i 1_{B_i}$, $B_i \in \mathcal{B}(\Omega)$, I finite, such that

$$\int_\Omega |g - g_\varepsilon|\, d\sigma < \varepsilon. \tag{4.7}$$

We have

$$\begin{aligned}
&\left| \int_{\Omega\times\mathbf{R}^m} g\otimes h\, d\mu - \int_\Omega g(x) \left(\int_{\mathbf{R}^m} h(y)\, d\mu_x(y) \right) d\sigma(x) \right| \\
\leq &\left| \int_{\Omega\times\mathbf{R}^m} g\otimes h\, d\mu - \int_{\Omega\times\mathbf{R}^m} g_\varepsilon \otimes h\, d\mu \right| \\
&+ \left| \int_{\Omega\times\mathbf{R}^m} g_\varepsilon \otimes h\, d\mu - \int_\Omega g_\varepsilon(x) \left(\int_{\mathbf{R}^m} h(y)\, d\mu_x(y) \right) d\sigma(x) \right| \\
&+ \left| \int_\Omega g_\varepsilon(x) \left(\int_{\mathbf{R}^m} h(y)\, d\mu_x(y) \right) d\sigma(x) - \int_\Omega g(x) \left(\int_{\mathbf{R}^m} h(y)\, d\mu_x(y) \right) d\sigma(x) \right|.
\end{aligned}$$

According to the first step, the second term of the right-hand side is equal to zero. According to (4.7), each of the two other terms is less than $\varepsilon\|h\|_\infty$. Since ε is arbitrary, we obtain

$$\int_{\Omega\times\mathbf{R}^m} g\otimes h\, d\mu = \int_\Omega g(x) \left(\int_{\mathbf{R}^m} h(y)\, d\mu_x(y) \right) d\sigma(x). \tag{4.8}$$

We are going to prove that μ_x is a probability measure. Let $(h_n)_{n\in\mathbf{N}}$ be a nondecreasing sequence of functions in $\mathbf{C}_0(\mathbf{R}^m)$ pointwise converging to $1_{\mathbf{R}^m}$. From (4.8) we deduce, for all Borel set B in Ω,

$$\int_{B\times\mathbf{R}^m} h_n(y)\, d\mu(x,y) = \int_B \left(\int_{\mathbf{R}^m} h_n(y)\, d\mu_x(y) \right) d\sigma(x),$$

and, by letting $n \to +\infty$,

$$\sigma(B) = \int_B \mu_x(\mathbf{R}^m)\, d\sigma(x).$$

Since for σ a.e. x in Ω, $\mu_x(\mathbf{R}^m) = \|\mu_x\| \leq 1$ we infer that $\mu_x(\mathbf{R}^m) = 1$ for σ a.e. x in Ω.

Third step. We assume that f belongs to $\mathbf{C}_0(\Omega \times \mathbf{R}^m)$. The result is now an easy consequence of the density of

$$\left\{ \sum_{i \in I} g_i \otimes h_i \ : \ g_i \in \mathbf{C}_c(\Omega),\ h_i \in \mathbf{C}_c(\mathbf{R}^m),\ I \in PF(\mathbf{N}) \right\}$$

in $\mathbf{C}_0(\Omega \times \mathbf{R}^m)$ for the uniform norm. ($PF(\mathbf{N})$ denotes the family of all finite subsets of $\mathbf{N}$.)

Last step. It remains to establish the uniqueness of the family $(\mu_x)_x$, up to equality σ-a.e. Take $f = 1_{B_\rho(x_0)} \times h$, where h is any function in D and $x_0 \in \Omega$ is such that the limit

$$\lim_{\rho \to 0} \frac{\int_{B_\rho(x_0)} \left(\int_{\mathbf{R}^m} h(y)\, d\mu_x(y) \right) d\sigma(x)}{\sigma(B_\rho(x_0))}$$

exists. According to the theory of differentiation of measures (Radon theorem), we know that there exists a Borel set N_h with $\sigma(N_h) = 0$ such that the above limit exists for $x_0 \in \Omega \setminus N_h$. Now, this limit exists for $x_0 \in \Omega' = \Omega \setminus \cup_{h \in D} N_h$ and for all $h \in D$.

From

$$\int_{\Omega \times \mathbf{R}^m} 1_{B_\rho(x_0)} h(y)\, d\mu(x, y) = \int_{B_\rho(x_0)} \left(\int_{\mathbf{R}^m} h(y)\, d\mu_x(y) \right) d\sigma(x)$$

we deduce that for $x_0 \in \Omega'$

$$\begin{aligned} \int_{\mathbf{R}^m} h(y)\, d\mu_{x_0}(y) &= \lim_{\rho \to 0} \frac{\int_{(B_\rho(x_0)) \times \mathbf{R}^m} h(y)\, d\mu(x, y)}{\sigma(B_\rho(x_0))} \\ &= \Gamma_{x_0}(h). \end{aligned}$$

Therefore $\mu_x = \Gamma_x$ for all $x \in \Omega'$ and all $h \in D$. This gives the required uniqueness. □

Theorem 4.2.5 (classical coarea formula). *For all Lipschitz functions $f : \mathbf{R}^N \longrightarrow \mathbf{R}$ and for all functions $g : \mathbf{R}^N \longrightarrow \mathbf{R}$ in $L^1(\mathbf{R}^N)$ we have*

$$\int_{\mathbf{R}^N} g(x) |Df|\, dx = \int_{-\infty}^{+\infty} \left(\int_{[f=t]} g(x)\, d\mathcal{H}^{N-1}(x) \right) dt.$$

As a corollary of Theorem 4.2.5 we obtain the so-called curvilinear Fubini theorem.

Corollary 4.2.2. *Let Ω be a bounded open subset of $\mathbf{R}^N$ and Γ_t the set $\{x \in \Omega : d(x, \mathbf{R}^N \setminus \overline{\Omega}) = t\}$. Then*

$$\int_\Omega g(x)\, dx = \int_{-\infty}^{+\infty} \left(\int_{\Gamma_t} g(x)\, d\mathcal{H}^{N-1}(x) \right) dt.$$

Proof. Take $f = d(\cdot, \mathbf{R}^N \setminus \overline{\Omega})$. □

Taking now $g = 1$ and f the truncation of $d(\cdot, S)$ between s and s', with $s < s'$, we obtain the next corollary.

Corollary 4.2.3. *Let S be a subset of $\mathbf{R}^N$. Then*

$$\mathcal{L}^N\big([s < d(\cdot, S) < s']\big) = \int_s^{s'} \mathcal{H}^{N-1}\big([d(\cdot, S) = t]\big)\, dt$$

and the distributional derivative of $\mathcal{L}^N\big([d(\cdot, S) < t]\big)$ is given by

$$\frac{d}{dt}\mathcal{L}^N\big([d(\cdot, S) < t]\big) = \mathcal{H}^{N-1}\big([d(\cdot, S) = t]\big).$$

4.3 Introduction to Young measures

We deal now with the notion of Young measure, a measure theoretical tool, well suited to the analysis of oscillations of minimizing sequences (see, for instance, [108]). We will give in chapter 11 an important application in the scope of relaxation in nonlinear elasticity. For an application to phase transitions for crystals, consult [110]. For a general exposition of the theory, see [53], [55], [104], [215], [223], [224], and the references therein.

4.3.1 Definition

In all this section, Ω is an open bounded subset of $\mathbf{R}^N$ and $E = \mathbf{R}^d$. In chapter 11, Section 11.4, we will consider the case when $d = mN$ so that E will be isomorphic to the space $\mathbf{M}^{m\times N}$ of $m \times N$ matrices. To shorten notations, we denote the N-dimensional Lebesgue measure restricted to Ω by $\mathcal{L}$.

Definition 4.3.1. *We call a Young measure on $\Omega \times E$ any positive measure $\mu \in \mathbf{M}^+(\Omega \times E)$ whose image $\pi_\Omega\#\mu$ by the projection π_Ω on Ω is the Lebesgue measure $\mathcal{L}$ on Ω, i.e., for all Borel subset B of Ω,*

$$\pi_\Omega\#\mu(B) := \mu(B \times E) = \mathcal{L}(B).$$

We denote the set of all Young measures on $\Omega \times E$ by $\mathcal{Y}(\Omega; E)$.

We now consider the space $\mathbf{C}_b(\Omega; E)$ of Carathéodory integrands, namely, the space of all functions $\psi : \Omega \times E \to \mathbf{R}$, $\mathcal{B}(\Omega) \otimes \mathcal{B}(E)$ measurable, and satisfying

(i) $\psi(x, \cdot)$ is bounded continuous on E for all $x \in \Omega$;

(ii) $x \mapsto \|\psi(x, \cdot)\|$ is Lebesgue integrable.

We equip $\mathcal{Y}(\Omega; E)$ with the narrow topology, i.e., the weakest topology which makes the maps

$$\mu \mapsto \int_{\Omega\times E} \psi\, d\mu$$

continuous, when ψ runs through $\mathbf{C}_b(\Omega; E)$. This topology induces the narrow convergence of Young measures defined as follows: let $(\mu_n)_{n\in\mathbf{N}}$ be a sequence of measures in $\mathcal{Y}(\Omega; E)$ and $\mu \in \mathcal{Y}(\Omega; E)$; then

$$\mu_n \overset{nar}{\rightharpoonup} \mu \Longleftrightarrow \lim_{n\to+\infty}\int_{\Omega\times E} \psi(x, \lambda) d\mu_n(x, \lambda) = \int_{\Omega\times E} \psi(x, \lambda) d\mu(x, \lambda)\ \forall \psi \in \mathbf{C}_b(\Omega; E).$$

Remark 4.3.1. Let $(\mu_n)_{n\in\mathbf{N}}$ be a sequence in $\mathcal{Y}(\Omega; E)$ and $\mu \in \mathcal{Y}(\Omega; E)$. It is easily seen (see Valadier [223], [224]) that

$$\mu_n \overset{nar}{\rightharpoonup} \mu \Longleftrightarrow \lim_{n\to+\infty}\int_{\Omega\times E} 1_B(x)\varphi(\lambda)d\mu_n = \int_{\Omega\times E} 1_B(x)\varphi(\lambda)d\mu \;\forall (B,\varphi) \in \mathcal{B}(\Omega)\times \mathbf{C}_b(E).$$

The space $\mathcal{Y}(\Omega; E)$ is closed in $\mathbf{M}(\Omega \times E)$ equipped with the narrow convergence; more precisely, we have the next proposition.

Proposition 4.3.1. *Let $(\mu_n)_{n\in\mathbf{N}}$ be a sequence in $\mathcal{Y}(\Omega; E)$ narrowly converging to some μ in $\mathbf{M}(\Omega \times E)$. Then μ belongs to $\mathcal{Y}(\Omega; E)$.*

Proof. From Remark 4.3.1, taking the test function $(x, \lambda) \mapsto \varphi(x, \lambda) = 1_B(x)1_E(\lambda)$, we obtain

$$\mathcal{L}(B) = \lim_{n\to+\infty} \mu_n(B \times E) = \mu(B \times E)$$

for all Borel subsets B of Ω. □

4.3.2 Slicing Young measures

According to Theorem 4.2.4, for each Young measure μ corresponds a unique family $(\mu_x)_{x\in\Omega}$ (up to equality a.e.) of probability measures on E such that $\mu = (\mu_x)_{x\in\Omega} \otimes \mathcal{L}$. Moreover, the map $x \mapsto \mu_x$ is measurable in the following sense:

$$\forall h \in \mathbf{C}_0(E),\; x \mapsto \int_E h\, d\mu_x \text{ is measurable.}$$

A Young measure μ is then also called a *parametrized measure* and Ω is the set of the parameters.

Let $L_w(\Omega, \mathbf{M}(E))$ be the space of all families $(\mu_x)_{x\in\Omega}$ of measures $\mu_x \in \mathbf{M}(E)$ (not necessarily probability measures) such that $x \mapsto \mu_x$ is measurable in the previous sense. By identifying $\mu = (\mu_x)_{x\in\Omega} \otimes \mathcal{L}$ with $(\mu_x)_{x\in\Omega}$, we have $\mathcal{Y}(\Omega; E) \subset L_w(\Omega, \mathbf{M}(E))$. We equip $L_w(\Omega, \mathbf{M}(E))$ with the following weak convergence:

$$(\mu_x^n)_{x\in\Omega} \overset{L_w}{\rightharpoonup} (\mu_x)_{x\in\Omega} \iff \forall\, h \in \mathbf{C}_0(E),\; \int_E h\, d\mu_x^n \rightharpoonup \int_E h\, d\mu_x \text{ in } L^\infty(\Omega) \text{ weak star,}$$

i.e.,

$$\int_\Omega g(x)\left(\int_E h\, d\mu_x^n\right) dx \to \int_\Omega g(x)\left(\int_E h\, d\mu_x\right) dx \;\forall\, h \in \mathbf{C}_0(E) \text{ and } \forall g \in L^1(\Omega).$$

Remark 4.3.2. The set $\mathcal{Y}(\Omega; E)$ is not closed in $L_w(\Omega, \mathbf{M}(E))$ equipped with this convergence. Indeed, take $\Omega = (0, 1)$, $E = \mathbf{R}$, and $\mu^n = \delta_n \otimes \mathcal{L}$. Then $(\mu_x^n)_{x\in\Omega}$ is the constant family δ_n and $(\mu_x^n)_{x\in\Omega} \overset{L_w}{\rightharpoonup} 0$ which is not a Young measure.

Let us define the tightness notion for Young measures.

Definition 4.3.2. *A subset $\mathcal{H}$ of $\mathcal{Y}(\Omega; E)$ is said to be tight if*

$$\forall \varepsilon > 0,\ \exists \mathcal{K}_\varepsilon \text{ compact subset of } E \text{ such that } \sup_{\mu \in \mathcal{H}} \mu(\Omega \times (E \setminus \mathcal{K}_\varepsilon)) < \varepsilon.$$

Tight subsets of $\mathcal{Y}(\Omega; E)$ are closed in $L_w(\Omega, \mathbf{M}(E))$. More precisely, we have the next proposition.

Proposition 4.3.2. *Let $(\mu^n)_{n\in\mathbf{N}}$ be a tight sequence in $\mathcal{Y}(\Omega; E)$ with $\mu^n = (\mu^n_x)_{x\in\Omega} \otimes \mathcal{L}$ and assume that $(\mu^n_x)_{x\in\Omega} \overset{L_w}{\rightharpoonup} (\mu_x)_{x\in\Omega}$ in $L_w(\Omega, \mathbf{M}(E))$. Then for a.e. x in Ω, μ_x is a probability measure on E so that $(\mu_x)_{x\in\Omega} \otimes \mathcal{L}$ belongs to $\mathcal{Y}(\Omega; E)$.*

Proof. Since $\sup_{n\in\mathbf{N}} \mu^n(\Omega \times E) = \mathcal{L}(\Omega) < +\infty$, there exists a subsequence (not relabeled) and some $\mu \in \mathbf{M}^+(\Omega \times E)$ such that

$$\mu^n \rightharpoonup \mu \text{ weakly in the sense of measures in } \mathbf{M}(\Omega \times E).$$

We claim that it suffices to prove that $\mu \in \mathcal{Y}(\Omega; E)$. Indeed, assuming $\mu \in \mathcal{Y}(\Omega; E)$ and denoting by $(\nu_x)_{x\in\Omega}$ the family of probability measures associated with μ, by using a density argument, we easily obtain

$$(\mu^n_x)_{x\in\Omega} \overset{L_w}{\rightharpoonup} (\nu_x)_{x\in\Omega} \text{ in } L_w(\Omega, \mathbf{M}(E)).$$

Therefore, by unicity of the weak limit in $L_w(\Omega, \mathbf{M}(E))$, up to a Lebesgue negligible subset of Ω, we will obtain $(\nu_x)_{x\in\Omega} = (\mu_x)_{x\in\Omega}$ so that $\mu = (\mu_x)_{x\in\Omega} \otimes \mathcal{L} \in \mathcal{Y}(\Omega; E)$.

We are going to establish that $\mu \in \mathcal{Y}(\Omega; E)$. According to Alexandrov's theorem (Proposition 4.2.3), for all open subsets U of Ω, one has

$$\pi_\Omega \# \mu(U) = \mu(U \times E) \leq \liminf_{n\to+\infty} \mu^n(U \times E) = \mathcal{L}(U).$$

Let now K be any compact subset of Ω and for all $\varepsilon > 0$ let $\mathcal{K}_\varepsilon$ be a compact subset of E given by the tightness hypothesis. According to Alexandrov's theorem again, one has

$$\begin{aligned} \pi_\Omega \# \mu(K) = \mu(K \times E) &\geq \mu(K \times \mathcal{K}_\varepsilon) \\ &\geq \limsup_{n\to+\infty} \mu^n(K \times \mathcal{K}_\varepsilon) \\ &\geq \limsup_{n\to+\infty} \mu^n(K \times E) - \varepsilon \\ &= \mathcal{L}(K) - \varepsilon \end{aligned}$$

so that, since ε is arbitrary, $\pi_\Omega \# \mu(K) \geq \mathcal{L}(K)$. Since the measure $\pi_\Omega \# \mu$ is regular, we deduce that $\pi_\Omega \# \mu(B) = \mathcal{L}(B)$ for all Borel subsets B of Ω. □

On $\mathcal{Y}(\Omega; E)$ the narrow convergence and the weak convergence of families of corresponding probability measures, are equivalent. More precisely, we have the following theorem.

Theorem 4.3.1. *Let $(\mu^n)_{n\in\mathbf{N}}$ be a sequence in $\mathcal{Y}(\Omega;E)$ and $\mu\in\mathcal{Y}(\Omega;E)$ with $\mu^n=(\mu^n_x)_{x\in\Omega}\otimes\mathcal{L}$ and $\mu=(\mu_x)_{x\in\Omega}\otimes\mathcal{L}$. Then*

$$\mu^n \overset{nar}{\rightharpoonup} \mu \iff (\mu^n_x)_{x\in\Omega} \overset{L_w}{\rightharpoonup} (\mu_x)_{x\in\Omega} \text{ in } L_w(\Omega,\mathbf{M}(E)).$$

Proof. Implication $\Longrightarrow$ is straightforward. We now prove the converse implication.

First step. We establish the tightness of $(\mu^n)_{n\in\mathbf{N}}$. Averaging each family, we define the two probability measures on E,

$$\nu_n:=\frac{1}{\mathcal{L}(\Omega)}\int_\Omega \mu^n_x\,dx,\quad \nu:=\frac{1}{\mathcal{L}(\Omega)}\int_\Omega \mu_x\,dx,$$

which act on all $\varphi\in\mathbf{C}_0(E)$ as follows:

$$\langle\nu_n,\varphi\rangle:=\frac{1}{\mathcal{L}(\Omega)}\int_\Omega\left(\int_E\varphi(\lambda)\,d\mu^n_x(\lambda)\right)dx,\quad \langle\nu,\varphi\rangle:=\frac{1}{\mathcal{L}(\Omega)}\int_\Omega\left(\int_E\varphi(\lambda)\,d\mu_x(\lambda)\right)dx.$$

Thus, the weak convergence of $(\mu^n_x)_{x\in\Omega}$ toward $(\mu_x)_{x\in\Omega}$ in $L_w(\Omega,\mathbf{M}(E))$ yields the weak convergence of ν_n toward ν in $\mathbf{M}(E)$. According to the regularity property satisfied by ν, for arbitrary $\varepsilon>0$, there exists a compact subset $\mathcal{K}_\varepsilon$ of E such that $\nu(E\setminus\mathcal{K}_\varepsilon)<\varepsilon$. From Lemma 4.2.1, one may assume $\nu(\partial\mathcal{K}_\varepsilon)=0$ so that, according to Alexandrov's theorem, Proposition 4.2.3,

$$\nu_n(\mathcal{K}_\varepsilon)\to\nu(\mathcal{K}_\varepsilon),$$

and, since ν_n and ν are probability measures,

$$\nu_n(E\setminus\mathcal{K}_\varepsilon)\to\nu(E\setminus\mathcal{K}_\varepsilon).$$

We then deduce $\sup_{n\geq N_\varepsilon}\nu_n(E\setminus\mathcal{K}_\varepsilon)<2\varepsilon$ for a certain N_ε in $\mathbf{N}$. Our claim then follows from $\mu^n(\Omega\times(E\setminus\mathcal{K}_\varepsilon))=\mathcal{L}(\Omega)\nu_n(E\setminus\mathcal{K}_\varepsilon)$.

Second step. We establish $\mu^n\overset{nar}{\rightharpoonup}\mu$. According to Remark 4.3.1, it suffices to prove

$$\lim_{n\to+\infty}\int_{\Omega\times E}1_B(x)\varphi(\lambda)\,d\mu^n(x,\lambda)=\int_{\Omega\times E}1_B(x)\varphi(\lambda)\,d\mu(x,\lambda)\;\forall B\in\mathcal{B}(\Omega),\ \forall\varphi\in\mathbf{C}_b(E).$$

For $\varepsilon>0$ let $\mathcal{K}_\varepsilon$ be the compact subset of E given by the tightness of $(\mu^n,\mu)_{n\in\mathbf{N}}$ and consider $\phi_\varepsilon\in\mathbf{C}_c(E)$ satisfying $0\leq\phi_\varepsilon\leq1$ and $\phi_\varepsilon=1$ on $\mathcal{K}_\varepsilon$. We now write

$$\begin{aligned}
&\left|\int_{\Omega\times E}1_B(x)\varphi(\lambda)\,d\mu^n(x,\lambda)-\int_{\Omega\times E}1_B(x)\varphi(\lambda)\,d\mu(x,\lambda)\right|\\
&\leq\left|\int_{\Omega\times E}1_B(x)\varphi(\lambda)\,d\mu^n(x,\lambda)-\int_{\Omega\times E}1_B(x)\phi_\varepsilon(\lambda)\varphi(\lambda)\,d\mu^n(x,\lambda)\right|\\
&+\left|\int_{\Omega\times E}1_B(x)\phi_\varepsilon(\lambda)\varphi(\lambda)\,d\mu^n(x,\lambda)-\int_{\Omega\times E}1_B(x)\phi_\varepsilon(\lambda)\varphi(\lambda)\,d\mu(x,\lambda)\right|\\
&+\left|\int_{\Omega\times E}1_B(x)\phi_\varepsilon(\lambda)\varphi(\lambda)\,d\mu(x,\lambda)-\int_{\Omega\times E}1_B(x)\varphi(\lambda)\,d\mu(x,\lambda)\right|.
\end{aligned}\tag{4.9}$$

According to the tightness of $(\mu^n,\mu)_{n\in\mathbf{N}}$, the first and the last term in the right-hand side of (4.9) are less than $\varepsilon\|\varphi\|_\infty$. On the other hand, since $\phi_\varepsilon\varphi\in\mathbf{C}_0(E)$, by hypothesis, the second term tends to zero when n goes to $+\infty$. Therefore, since ε is arbitrary, we end the proof by letting $n\to+\infty$ in (4.9). □

4.3.3 Prokhorov's compactness theorem

The theorem below may be considered as a parametrized version of the classical Prokhorov compactness Theorem 4.2.3.

Theorem 4.3.2 (Prokhorov's compactness theorem for Young measures). *Let $(\mu^n)_{n\in\mathbf{N}}$ be a tight sequence in $\mathcal{Y}(\Omega; E)$. Then, there exists a subsequence $(\mu^{n_k})_{k\in\mathbf{N}}$ of $(\mu^n)_{n\in\mathbf{N}}$ and μ in $\mathcal{Y}(\Omega; E)$ such that*

$$\mu^{n_k} \overset{nar}{\rightharpoonup} \mu \text{ in } \mathcal{Y}(\Omega; E).$$

PROOF. Since $\sup_{n\in\mathbf{N}} \mu^n(\Omega\times E) = \mathcal{L}(\Omega) < +\infty$, there exists a subsequence (not relabeled) and $\mu \in \mathbf{M}^+(\Omega\times E)$ such that $\mu^n \rightharpoonup \mu$ weakly in the sense of measures in $\mathbf{M}(\Omega\times E)$. Since $(\mu^n)_{n\in\mathbf{N}}$ is tight, arguing as in the proof of Proposition 4.3.2, one may assert that μ belongs to $\mathcal{Y}(\Omega; E)$. It remains to establish the narrow convergence of μ^n toward μ, or, equivalently, according to Theorem 4.3.1, the weak convergence of $(\mu^n_x)_{x\in\Omega}$ toward $(\mu_x)_{x\in\Omega}$ in $L_w(\Omega, \mathbf{M}(E))$. Let $\Phi \in L^1(\Omega)$, $\varphi \in \mathbf{C}_0(E)$, and $\Phi_\varepsilon \in \mathbf{C}_c(\Omega)$ satisfying

$$\int_\Omega |\Phi - \Phi_\varepsilon|\, dx < \varepsilon. \tag{4.10}$$

Since μ^n weakly converges to μ in $\mathbf{M}(\Omega\times E)$ and $(x,\lambda) \mapsto \Phi_\varepsilon(x)\varphi(\lambda)$ belongs to $\mathbf{C}_0(\Omega\times E)$, according to the slicing Theorem 4.2.4, one has

$$\lim_{n\to+\infty} \left| \int_\Omega \Phi_\varepsilon(x) \left(\int_E \varphi(\lambda)\, d\mu^n_x \right) dx - \int_\Omega \Phi_\varepsilon(x) \left(\int_E \varphi(\lambda)\, d\mu_x \right) dx \right| = 0. \tag{4.11}$$

Let us write

$$\begin{aligned}
&\left| \int_\Omega \Phi(x) \left(\int_E \varphi(\lambda)\, d\mu^n_x \right) dx - \int_\Omega \Phi(x) \left(\int_E \varphi(\lambda)\, d\mu_x \right) dx \right| \\
&\le \left| \int_\Omega \Phi(x) \left(\int_E \varphi(\lambda)\, d\mu^n_x \right) dx - \int_\Omega \Phi_\varepsilon(x) \left(\int_E \varphi(\lambda)\, d\mu^n_x \right) dx \right| \\
&+ \left| \int_\Omega \Phi_\varepsilon(x) \left(\int_E \varphi(\lambda)\, d\mu^n_x \right) dx - \int_\Omega \Phi_\varepsilon(x) \left(\int_E \varphi(\lambda)\, d\mu_x \right) dx \right| \\
&+ \left| \int_\Omega \Phi_\varepsilon(x) \left(\int_E \varphi(\lambda)\, d\mu_x \right) dx - \int_\Omega \Phi(x) \left(\int_E \varphi(\lambda)\, d\mu_x \right) dx \right|.
\end{aligned} \tag{4.12}$$

From (4.10), the first and the last term of the right-hand side of (4.12) are less than $\varepsilon\|\varphi\|_\infty$. Since ε is arbitrary, the claim follows from (4.11) by letting $n \to +\infty$ in (4.12). □

4.3.4 Young measures associated with functions and generated by functions

Let $u : \Omega \to E$ be a given Borel function and consider the image $\mu = G\#\mathcal{L}$ of the measure $\mathcal{L}$ by the graph function $G : \Omega \to \Omega\times E,\ x \mapsto (x, u(x))$.

Since the image of the measure μ by the projection π_Ω on Ω is the Lebesgue measure $\mathcal{L}$, μ belongs to $\mathcal{Y}(\Omega; E)$. This measure, concentrated on the graph of u, is called the *Young*

measure associated with the function u. By definition of the image of a measure, μ "acts" on $\mathbf{C}_b(\Omega; E)$ as follows:

$$\forall \varphi \in \mathbf{C}_b(\Omega; E), \quad \int_{\Omega\times E} \varphi(x,\lambda)\, d\mu(x,\lambda) = \int_\Omega \varphi(x, u(x))\, dx.$$

This shows that the probability family $(\mu_x)_{x\in\Omega}$ associated with μ is $(\delta_{u(x)})_{x\in\Omega}$.

Let $(u_n)_{n\in\mathbf{N}}$ be a sequence of Borel functions $u_n : \Omega \to E$ and consider the sequence of their associated Young measures $(\mu_n)_{n\in\mathbf{N}}$, $\mu_n = (\delta_{u_n(x)})_{x\in\Omega} \otimes \mathcal{L}$. If $\mu_n \overset{nar}{\rightharpoonup} \mu$ in $\mathcal{Y}(\Omega; E)$, the Young measure μ is said to be *generated by* the sequence $(u_n)_{n\in\mathbf{N}}$. In general (see Examples 4.3.1 and 4.3.2), μ is not associated with a function.

Let us rephrase the tightness of a sequence $(\mu_n)_{n\in\mathbf{N}}$ in terms of the associated sequence $(u_n)_{n\in\mathbf{N}}$. We easily obtain the following equivalence: the sequence $(\mu_n)_{n\in\mathbf{N}}$ is tight iff

$$\forall \varepsilon > 0,\ \exists \mathcal{K}_\varepsilon, \text{ compact subset of } E, \text{ such that } \sup_{n\in\mathbf{N}} \mathcal{L}\big\{x \in \Omega \ :\ u_n(x) \in E \setminus \mathcal{K}_\varepsilon\big\} < \varepsilon.$$

Remark 4.3.3. It is worth noticing that a sequence $(\mu_n)_{n\in\mathbf{N}}$ of Young measures associated with a bounded sequence $(u_n)_{n\in\mathbf{N}}$ in $L^1(\Omega, E)$ is tight. Indeed according to the Markov inequality, one has

$$\begin{aligned}\mathcal{L}\big\{x \in \Omega \ :\ |u_n(x)| > M\big\} &\le \frac{1}{M}\int_\Omega |u_n|\, dx \\ &\le \frac{1}{M}\sup_{n\in\mathbf{N}}\int_\Omega |u_n|\, dx,\end{aligned}$$

which tends to zero when $M \to +\infty$. Therefore, according to Prokhorov's theorem, Theorem 4.3.2, for each bounded sequence $(u_n)_{n\in\mathbf{N}}$ in $L^1(\Omega, E)$, one can extract a subsequence generating a Young measure μ, i.e.,

$$(\delta_{u_n(x)})_{x\in\Omega} \otimes \mathcal{L} \overset{nar}{\rightharpoonup} \mu.$$

4.3.5 Semicontinuity and continuity properties

Here is a first semicontinuity result related to extended real-valued nonnegative functions.

Proposition 4.3.3. *Let $\varphi : \Omega \times E \to [0, +\infty]$ be a $\mathcal{B}(\Omega) \otimes \mathcal{B}(E)$ measurable function such that $\lambda \mapsto \varphi(x,\lambda)$ is lsc for a.e. x in Ω. Moreover, let $(\mu_n)_{n\in\mathbf{N}}$ be a sequence of Young measures in $\mathcal{Y}(\Omega; E)$, narrowly converging to some Young measure μ in $\mathcal{Y}(\Omega; E)$. Then*

$$\int_{\Omega\times E} \varphi(x,\lambda)\, d\mu(x,\lambda) \le \liminf_{n\to+\infty} \int_{\Omega\times E} \varphi(x,\lambda)\, d\mu_n(x,\lambda).$$

Proof. Let us consider the Lipschitz regularization of φ

$$\varphi_p(x,\lambda) := \inf_{\xi\in E} \big\{\varphi(x,\xi) + p|\lambda - \xi|\big\}$$

for $p \in \mathbf{N}$ intended to go to $+\infty$, and set $\psi_p = \varphi_p \wedge p$. It is easily seen that ψ belongs to $\mathbf{C}_b(\Omega; E)$ (see Theorem 9.2.1) and that $(\psi_p)_{p\in\mathbf{N}}$ is a nondecreasing sequence which pointwise converges to φ. Consequently,

$$\begin{aligned}\int_{\Omega\times E} \psi_p(x,\lambda)\, d\mu(x,\lambda) &= \lim_{n\to+\infty} \int_{\Omega\times E} \psi_p(x,\lambda)\, d\mu_n(x,\lambda)\\ &\le \liminf_{n\to+\infty} \int_{\Omega\times E} \varphi(x,\lambda)\, d\mu_n(x,\lambda),\end{aligned}$$

and we complete the proof thanks to the monotone convergence theorem, by letting $p \to +\infty$ in the left-hand side. □

We would like to improve Proposition 4.3.3 for functions φ taking negative values or more generally for functions φ which are not necessarily bounded from below. We restrict ourselves to sequences of Young measures associated with functions. Let us first recall the notion of uniform integrability: a sequence $(f_n)_{n\in\mathbf{N}}$ of functions $f_n : \Omega \to \mathbf{R}$ in $L^1(\Omega)$ is said to be *uniformly integrable* if

$$\lim_{R\to+\infty} \sup_{n\in\mathbf{N}} \int_{[|f_n|>R]} |f_n| = 0.$$

It is classical that this definition is equivalent to Definition 2.4.4 (see Delacherie and Meyer [125]).

Proposition 4.3.4. *Let $(\mu_n)_{n\in\mathbf{N}}$ be a sequence of Young measures associated with a sequence of functions $(u_n)_{n\in\mathbf{N}}$, narrowly converging to some Young measure μ. On the other hand, let $\varphi : \Omega \times E \to \mathbf{R}$ be a $\mathcal{B}(\Omega)\otimes\mathcal{B}(E)$ measurable function such that $\lambda \mapsto \varphi(x,\lambda)$ is lsc for a.e. x in Ω. Assume moreover that the negative part $x \mapsto \varphi(x,u_n(x))^-$ is uniformly integrable. Then*

$$\int_{\Omega\times E} \varphi(x,\lambda)\, d\mu(x,\lambda) \le \liminf_{n\to+\infty} \int_\Omega \varphi(x,u_n(x))\, dx.$$

Proof. Let $R > 0$ intended to tend to $+\infty$ and set $\varphi_R = \sup(-R,\varphi)+R$. Since $\varphi_R \ge 0$ and $\lambda \mapsto \varphi_R(x,\lambda)$ is lsc, one may apply Proposition 4.3.3 so that, removing the term $R\mathcal{L}(\Omega)$, one obtains

$$\begin{aligned}\int_{\Omega\times E} \varphi(x,\lambda)\, d\mu(x,\lambda) &\le \int_{\Omega\times E} \sup(-R,\varphi(x,\lambda))\, d\mu(x,\lambda)\\ &\le \liminf_{n\to+\infty} \int_{\Omega\times E} \sup(-R,\varphi(x,\lambda))\, d\mu_n(x,\lambda)\\ &= \liminf_{n\to+\infty} \int_\Omega \sup(-R,\varphi(x,u_n(x)))\, dx. \qquad (4.13)\end{aligned}$$

On the other hand,

$$\begin{aligned}\int_\Omega \sup(-R,\varphi(x,u_n(x)))\, dx &= \int_{[\varphi(.,u_n(.))\ge -R]} \varphi(x,u_n(x))\, dx + \int_{[\varphi(.,u_n(.))<-R]} (-R)\, dx\\ &\le \int_\Omega \varphi(x,u_n(x))dx - \int_{[\varphi(.,u_n(.))<-R]} \varphi(x,u_n(x))dx. \qquad (4.14)\end{aligned}$$

But

$$
\begin{aligned}
-\int_{[\varphi(.,u_n(.))<-R]} \varphi(x,u_n(x))\,dx &= \int_{[\varphi(.,u_n(.))<-R]} \varphi(x,u_n(x))^-\,dx \\
&= \int_{[\varphi(.,u_n(.))^->R]} \varphi(x,u_n(x))^-\,dx \\
&\le \sup_{n\in\mathbf{N}} \int_{[\varphi(.,u_n(.))^->R]} \varphi(x,u_n(x))^-\,dx, \qquad (4.15)
\end{aligned}
$$

which, by hypothesis, tends to 0 when $R \to +\infty$. We end the proof by collecting (4.13), (4.14), and (4.15) and letting $R \to +\infty$. □

As a straightforward consequence of Proposition 4.3.4, we obtain the following useful theorem, a key tool in relaxation theory (see Section 11.4.2).

Theorem 4.3.3. *Let $(\mu_n)_{n\in\mathbf{N}}$ be a sequence of Young measures associated with a sequence of functions $(u_n)_{n\in\mathbf{N}}$, narrowly converging to some Young measure μ. On the other hand, let $\varphi : \Omega \times E \to \mathbf{R}$ be a $\mathcal{B}(\Omega)\otimes\mathcal{B}(E)$ measurable function such that $\lambda \mapsto \varphi(x,\lambda)$ is continuous for a.e. x in Ω. Assume moreover that $x \mapsto \varphi(x,u_n(x))$ is uniformly integrable. Then*

$$
\int_{\Omega\times E} \varphi(x,\lambda)\,d\mu(x,\lambda) = \lim_{n\to+\infty} \int_\Omega \varphi(x,u_n(x))\,dx.
$$

Proof. Since $x \mapsto \varphi(x,u_n(x))$ is uniformly integrable, $x \mapsto \varphi(x,u_n(x))^-$ is uniformly integrable so that, according to Proposition 4.3.4,

$$
\int_{\Omega\times E} \varphi(x,\lambda)\,d\mu(x,\lambda) \le \liminf_{n\to+\infty} \int_\Omega \varphi(x,u_n(x))\,dx.
$$

Let us set $\tilde\varphi := -\varphi$. Since $x \mapsto \tilde\varphi(x,u_n(x))^-$ is equal to $x \mapsto \varphi(x,u_n(x))^+$, which is also uniformly integrable, we infer from Proposition 4.3.4

$$
\int_{\Omega\times E} \tilde\varphi(x,\lambda)\,d\mu(x,\lambda) \le \liminf_{n\to+\infty} \int_\Omega \tilde\varphi(x,u_n(x))\,dx,
$$

that is

$$
\int_{\Omega\times E} \varphi(x,\lambda)\,d\mu(x,\lambda) \ge \limsup_{n\to+\infty} \int_\Omega \varphi(x,u_n(x))\,dx,
$$

which completes the proof. □

Remark 4.3.4. Theorem 4.3.3 is often applied when $(\psi(.,u_n(.))_{n\in\mathbf{N}}$ is a sequence of real-valued functions weakly converging in $L^1(\Omega)$. Indeed, according to the Dunford–Pettis theorem, Theorem 2.4.5, there is equivalence between the weak relative compactness of the sequence $(\psi(.,u_n(.))_{n\in\mathbf{N}}$ in $L^1(\Omega)$ and its uniform integrability.

4.3.6 Young measures capture oscillations

Let $(\mu_n)_{n\in\mathbf{N}}$ be a sequence of Young measures associated with a sequence $(u_n)_{n\in\mathbf{N}}$ of Borel functions $u_n : \Omega \to E$ and assume that $\mu_n \overset{nar}{\rightharpoonup} \mu$ in $\mathcal{Y}(\Omega; E)$. Let us show that, roughly speaking, for a.e. x_0 in Ω, each μ_{x_0} gives the limit probability distribution of the values of u_n when x are taken randomly, with the uniform probability law around x_0. Indeed, it is easily seen that the slicing Theorem 4.2.4 holds for functions $(x, \lambda) \mapsto f(x, \lambda)$ of the form $(x, \lambda) \mapsto 1_{B_\rho(x_0)\times A}(x, \lambda)$, where $B_\rho(x_0)$ is the open ball with radius $\rho > 0$ centered at $x_0 \in \Omega$, and A is any open set of E. Then we have

$$\mu(B_\rho(x_0) \times A) = \int_{\Omega\times E} 1_{B_\rho(x_0)\times A}(x, \lambda) = \int_{B_\rho(x_0)} \mu_x(A)\, dx.$$

Therefore, according to Lebesgue's differentiation theorem, for a.e. $x_0 \in \Omega$,

$$\mu_{x_0}(A) = \lim_{\rho\to 0} \frac{1}{\mathcal{L}(B_\rho(x_0))} \mu(B_\rho(x_0) \times A). \tag{4.16}$$

Note that the negligible set of all the points x_0 for which (4.16) does not hold depends on the choice of A in E. But reasoning with a given countable family of sets A, there exists a negligable set $\mathcal{N}$ in Ω such that for all $x_0 \in \Omega \setminus \mathcal{N}$, (4.16) holds for all set A of this family. Let then $x_0 \in \Omega \setminus \mathcal{N}$ and A be chosen so that (4.16) holds. Choose moreover $\rho > 0$ such that $\mu(\partial(B_\rho(x_0) \times A)) = 0$. Such a choice is possible thanks to Lemma 4.2.1. Since $\mu_n \overset{nar}{\rightharpoonup} \mu$, in particular μ_n weakly converges in the sense of measures in $\mathbf{M}(\Omega \times E)$ so that, according to Theorem 4.2.3,

$$\begin{aligned}\mu(B_\rho(x_0) \times A) &= \lim_{n\to+\infty} \mu_n(B_\rho(x_0) \times A) \\ &= \lim_{n\to+\infty} \mathcal{L}\big(\{x \in B_\rho(x_0) : u_n(x) \in A\}\big).\end{aligned} \tag{4.17}$$

Collecting (4.16) and (4.17), we finally obtain

$$\mu_{x_0}(A) = \lim_{\rho\to 0} \lim_{n\to+\infty} \frac{\mathcal{L}\big(\{x \in B_\rho(x_0) : u_n(x) \in A\}\big)}{\mathcal{L}(B_\rho(x_0))}, \tag{4.18}$$

which proves the thesis.

For a given sequence $(u_n)_{n\in\mathbf{N}}$ the first mode of behavior which can cause a defect of strong convergence is the presence of rapid oscillations in the functions u_n. Estimate (4.18) shows that Young measures capture some information on such oscillations. We will illustrate this property with a few examples.

Let $u : Y = (0, 1)^N \to E$ be a given function in $L^p(Y, E)$, $p \geq 1$ extended by Y-periodicity to $\mathbf{R}^N$ and define the sequence $(u_n)_{n\in\mathbf{N}}$ by setting $u_n(x) = u(nx)$ for all $x \in \mathbf{R}^N$. Classically one has

$$u_n \rightharpoonup \overline{u} \text{ in } L^p(\Omega, E),$$

where $\overline{u}$ is the mean value of u defined by $\overline{u} = \int_Y u(y)\, dy$. Obviously, if u is not a constant function, we have not strong convergence in $L^p(\Omega, \mathbf{R}^N)$ nor a.e. pointwise convergence on Ω of the sequence $(u_n)_{n\in\mathbf{N}}$ toward $\overline{u}$. Let $(\mu_n)_{n\in\mathbf{N}}$ denote the sequence of Young measures

associated with the sequence $(u_n)_{n\in\mathbf{N}}$, i.e. $\mu_n = (\delta_{u_n(x)})_{x\in\Omega} \otimes \mathcal{L}$. Then we have the next proposition.

Proposition 4.3.5. *The sequence $(\mu_n)_{n\in\mathbf{N}}$ narrowly converges to $\mu = (\mu_x)_{x\in\Omega} \otimes \mathcal{L}$ in $\mathcal{Y}(\Omega; E)$, where, for a.e. x in Ω, the probability measure μ_x is the image $u\#\mathcal{L}\lfloor Y$ of the Lebesgue measure $\mathcal{L}\lfloor Y$ by the function u. In other words, μ acts on all $\varphi \in \mathbf{C}_b(\Omega; E)$ as follows:*

$$\int_{\Omega\times E} \varphi(x,\lambda)\, d\mu(x,\lambda) = \int_\Omega \left(\int_Y \varphi(x, u(y))\, dy \right) dx.$$

PROOF. It is enough to establish

$$\lim_{n\to+\infty} \int_{\Omega\times E} \varphi(x,\lambda)\, d\mu^n(x,\lambda) = \int_\Omega \left(\int_Y \varphi(x, u(y))\, dy \right) dx$$

when φ is of the form $\varphi(x,\lambda) = 1_B(x)\phi(\lambda)$, where B belongs to $\mathcal{B}(\Omega)$ and ϕ is a bounded continuous function on E (see Remark 4.3.1).

Since, classically, $x \mapsto \phi(u(nx))$ weakly converges to $\int_Y \phi(u(y))\, dy$ in $L^\infty(\Omega)$ weak star, we have

$$\begin{aligned}
\lim_{n\to+\infty} \int_{\Omega\times E} \varphi(x,\lambda)\, d\mu^n(x,\lambda) &= \lim_{n\to+\infty} \int_\Omega 1_B(x)\phi(u(nx))\, dx \\
&= \int_\Omega 1_B(x) \left(\int_Y \phi(u(y))\, dy \right) dx \\
&= \int_\Omega \left(\int_Y \varphi(x, u(y))\, dy \right) dx
\end{aligned}$$

and the thesis is proved. □

Example 4.3.1. Take $u : (0,1) \to \mathbf{R}$ the function defined by

$$u(x) = \begin{cases} -1 & \text{if } x \in (0, \frac{1}{2}), \\ +1 & \text{if } x \in (\frac{1}{2}, 1), \end{cases}$$

and consider the sequence $(u_n)_{n\in\mathbf{N}}$ defined as previously with, for instance, $\Omega = (0,1)$. Using Proposition 4.3.5 (or (4.18)), it is easily seen that $\mu_x = \frac{1}{2}\delta_{-1} + \frac{1}{2}\delta_{+1}$ so that μ is the measure on $(0,1)\times\mathbf{R}$, concentrated on the union of the two segments $(0,1)\times\{-1\} \cup (0,1)\times\{1\}$, with the mass 1/2 on each two segments. Therefore, the measure μ, encoding the oscillations, is not associated with a function.

Example 4.3.2. Take $u : (0,1) \to \mathbf{R}$ the function defined by $u(x) = \sin(2\pi x)$ and consider the sequence $(u_n)_{n\in\mathbf{N}}$ defined as previously with, for instance, $\Omega = (0,1)$. An elementary calculation gives $\mu_x = \frac{1}{\pi\sqrt{1-y^2}}\mathcal{L}\lfloor(-1,1)$ so that μ is not associated with a function but concentrated on all the rectangle $(0,1)\times(-1,1)$.

The next proposition shows that Young measures encode oscillations on the weak limits.

Proposition 4.3.6. *Let $(u_n)_{n\in\mathbf{N}}$ be a given sequence of functions in $L^p(\Omega, E)$, weakly converging to some u in $L^p(\Omega, E)$, $p \geq 1$, and assume that the sequence $(\mu_n)_{n\in\mathbf{N}}$ of their associated Young measures narrowly converges to some Young measure μ. Then, for a.e. x in Ω, $u(x)$ is the barycenter (or the expectation) of the probability measure μ_x:*

$$u(x) = \int_E \lambda d\mu_x(\lambda).$$

Proof. Reasoning with each component of u_n, one may assume, without restrictions, that u_n is a real-valued function and that $E = \mathbf{R}$. Let us first apply Theorem 4.3.3 with φ defined by $\varphi(x, \lambda) = \phi(x)\lambda$, where $\phi \in \mathbf{C}_c(\Omega)$. Since $(u_n)_{n\in\mathbf{N}}$ weakly converges in $L^p(\Omega)$, the sequence $\varphi(., u_n(.))_{n\in\mathbf{N}}$ is uniformly integrable (see Remark 4.3.4) so that, according to Theorem 4.3.3,

$$\begin{aligned}\int_{\Omega\times E} \phi(x)\lambda \, d\mu(x, \lambda) &= \lim_{n\to+\infty} \int_\Omega \phi(x)u_n(x) \, dx \\ &= \int_\Omega \phi(x)u(x) \, dx.\end{aligned}$$

According now to Theorem 4.2.4, we infer

$$\int_\Omega \phi(x) \left(\int_{\mathbf{R}} \lambda \, d\mu_x(\lambda) \right) dx = \int_\Omega \phi(x)u(x) \, dx.$$

Since ϕ is arbitrary, one obtains $u(x) = \int_{\mathbf{R}} \lambda \, d\mu_x(\lambda)$ for a.e. x in Ω. □

One can now establish the Dunford–Pettis theorem, Theorem 2.4.5.

Proposition 4.3.7. *Let $(u_n)_{n\in\mathbf{N}}$ be a given sequence of uniformly integrable functions in $L^1(\Omega, E)$. Then, there exists a subsequence $(u_{n_k})_{k\in\mathbf{N}}$ and u in $L^1(\Omega, E)$ such that*

$$u_{n_k} \rightharpoonup u \qquad \sigma(L^1, L^\infty).$$

Proof. Since $(u_n)_{n\in\mathbf{N}}$ is uniformly integrable, one can easily establish that

$$\sup_{n\in\mathbf{N}} \int_\Omega |u_n| \, dx < +\infty$$

so that (see Remark 4.3.3) the sequence of Young measures $\mu_n = (\delta_{u_n(x)})_{x\in\Omega} \otimes \mathcal{L}$ is tight. According to Prokhorov compactness Theorem 4.3.2, there exist a subsequence of $(\mu_n)_{n\in\mathbf{N}}$ (not relabeled) and μ in $\mathcal{Y}(\Omega; E)$ satisfying $\mu_n \overset{nar}{\rightharpoonup} \mu$.

Consider $g \in L^\infty(\Omega, E)$ and set $\varphi(x, \lambda) := g(x).\lambda$. The sequence $(\varphi(x, u_n(x))_{n\in\mathbf{N}}$ obviously satisfies hypotheses of Theorem 4.3.3 so that

$$\lim_{n\to+\infty} \int_\Omega g(x).u_n(x) \, dx = \int_\Omega g(x). \left(\int_E \lambda \, d\mu_x \right) dx.$$

The barycenter

$$u : x \mapsto \int_E \lambda \, d\mu_x$$

then satisfies $u_n \rightharpoonup u \;\; \sigma(L^1, L^\infty)$. □

4.3.7 Young measures do not capture concentrations

Another mode of behavior which causes a defect of strong convergence for a sequence $(u_n)_{n\in\mathbf{N}}$ weakly converging to some u in $L^p(\Omega, E)$ is the *concentration effect*. Such concentration effects appear when $u_n - u$ converges to zero in measure and when the total mass $\int_\Omega |u_n - u|^p\, dx$ is concentrated at the limit to a set of zero Lebesgue measure. Note that in examples of Section 4.3.6 no concentration effects occured because the sequences $(u_n)_{n\in\mathbf{N}}$ did not converge in measure. Let us illustrate the concentration phenomenon with the following elementary example.

Let $\Omega = (-1, 1)$, $p = 2$ and consider the real-valued function u_n, $n \in \mathbf{N}^*$, defined by

$$u_n(x) = \begin{cases} \sqrt{n} & \text{if } x \in (-\frac{1}{n}, \frac{1}{n}), \\ 0 & \text{otherwise.} \end{cases}$$

It is easily seen that u_n converges to 0 a.e. in $(-1, 1)$, in measure, and weakly in $L^2(-1, 1)$. On the other hand, the total mass is $\int_{(-1,1)} |u_n|^2\, dx = 2$, while the measure $|u_n|^2\mathcal{L}$ weakly converges to $2\delta_0$ in the sense of measures in $\mathbf{M}(-1, 1)$. Therefore the total mass is concentrated at the point $x = 0$.

Proposition 4.3.8 shows that Young measures generated by sequences converging in measure are trivial and therefore do not capture concentration effects. For the analysis of both oscillations and concentration effects, see [217] and [137].

Proposition 4.3.8. *A sequence $(u_n)_{n\in\mathbf{N}}$ of Borel functions converges to u in measure iff the associated sequence of Young measures $(\mu_n)_{n\in\mathbf{N}}$ narrowly converges to the Young measure associated with u, i.e., $\mu = (\delta_{u(x)})_{x\in\Omega} \otimes \mathcal{L}$.*

On the other hand, let $(\mu_n)_{n\in\mathbf{N}}$ be a sequence of Young measures associated with a sequence of Borel functions $(u_n)_{n\in\mathbf{N}}$, narrowly converging to some Young measure μ. Moreover let $(v_n)_{n\in\mathbf{N}}$ be another sequence of Borel functions $v_n : \Omega \to E$ such that $v_n - u_n$ converges to 0 in measure. Then the sequence $(v_n)_{n\in\mathbf{N}}$ generates the same Young measure μ. In other words, $v_n = (v_n - u_n) + u_n$ generates the Young measure μ generated by u_n so that the perturbation by $v_n - u_n$, for which a concentration phenomenon may occur, has no effect on μ.

Proof. *First step.* We first claim

$$u_n \text{ converges in measure toward } u \Longrightarrow (\delta_{u_n(x)})_{x\in\Omega} \overset{L_w}{\rightharpoonup} (\delta_{u(x)})_{x\in\Omega},$$

which, according to Theorem 4.3.1, is equivalent to the narrow convergence of corresponding Young measures. Now, after using an easy density argument, it is enough to test the convergence with $\varphi(x, \lambda) = 1_B(x)\phi(\lambda)$, where $B \in \mathcal{B}(\Omega)$ and $\phi \in \mathbf{C}_0(E)$. Let $\varepsilon > 0$ given arbitrary. From uniform continuity of ϕ, there exists $\eta > 0$ such that

$$|\lambda - \lambda'| < \eta \Longrightarrow |\phi(\lambda) - \phi(\lambda')| < \varepsilon.$$

Let us write

$$\begin{aligned}
&\left| \int_\Omega \varphi(x, u_n(x))\, dx - \int_\Omega \varphi(x, u(x))\, dx \right| \le \int_\Omega |\phi(u_n(x)) - \phi(u(x))|\, dx \\
&= \int_{[|u_n - u| > \eta]} |\phi(u_n(x)) - \phi(u(x))|\, dx + \int_{[|u_n - u| \le \eta]} |\phi(u_n(x)) - \phi(u(x))|\, dx \\
&\le 2\|\phi\|_\infty \mathcal{L}([|u_n - u| > \eta]) + \varepsilon \mathcal{L}(\Omega). \qquad (4.19)
\end{aligned}$$

Now, by hypothesis, $\lim_{n\to+\infty} \mathcal{L}([|u_n - u| > \eta]) = 0$ and, since ε is arbitrary, the claim follows after letting $n \to +\infty$ in (4.19).

Second step. We establish the converse implication. Let us consider $\varphi \in \mathbf{C}_b(\Omega; E)$ defined by $\varphi(x, \lambda) = |\lambda - u(x)| \wedge C$, where C is any positive constant. Since $\mu_n \overset{nar}{\rightharpoonup} \mu$, one has

$$\lim_{n\to+\infty} \int_{\Omega\times E} \varphi(x, \lambda)\, d\mu_n(x, \lambda) = \int_{\Omega\times E} \varphi(x, \lambda)\, d\mu(x, \lambda),$$

that is,

$$\lim_{n\to+\infty} \int_\Omega |u_n(x) - u(x)| \wedge C\, dx = \int_\Omega |u(x) - u(x)| \wedge C\, dx = 0. \qquad (4.20)$$

On the other hand, for any $\eta > 0$,

$$\mathcal{L}([|u_n - u| > \eta]) \le \frac{1}{\min(\eta, C)} \int_\Omega |u_n(x) - u(x)| \wedge C\, dx.$$

Consequently, (4.20) yields $\lim_{n\to+\infty} \mathcal{L}([|u_n - u| > \eta]) = 0$.

Last step. We establish the second assertion. As previously, according to Theorem 4.3.1, it suffices to establish

$$\lim_{n\to+\infty} \int_B \phi(v_n)\, dx = \int_{\Omega\times E} 1_B(x)\phi(\lambda)\, d\mu(x, \lambda)$$

for all Borel subset B of Ω and all $\phi \in \mathbf{C}_c(E)$.

Let $\varepsilon > 0$. Since ϕ is uniformly continuous on E, there exists $\eta > 0$ such that $|\phi(v_n(x)) - \phi(u_n(x))| < \varepsilon$ for all x in the set $[|v_n - u_n| < \eta]$. On the other hand, since $v_n - u_n$ tends to 0 in measure, $\lim_{n\to+\infty} \mathcal{L}([|v_n - u_n| \ge \eta]) = 0$. Therefore

$$\begin{aligned}
\left| \int_B \phi(v_n)\, dx - \int_B \phi(u_n)\, dx \right| &\le \int_\Omega |\phi(v_n) - \phi(u_n)|\, dx \\
&= \int_{[|v_n - u_n| \ge \eta]} |\phi(v_n) - \phi(u_n)|\, dx \\
&\quad + \int_{[|v_n - u_n| < \eta]} |\phi(v_n) - \phi(u_n)|\, dx \\
&\le 2\|\phi\|_\infty \mathcal{L}([|v_n - u_n| \ge \eta]) + \varepsilon \mathcal{L}(\Omega)
\end{aligned}$$

and, since ε is arbitrary, we conclude by letting $n \to +\infty$. □

Chapter 5

Sobolev spaces

In this chapter we introduce the key notion of Sobolev spaces, which can be considered as one of the main tools that made possible the wide development of the theory of PDEs in the last several decades.

Motivations. In chapter 2, Section 2.3.1, it was shown that the Dirichlet problem

$$\begin{cases} -\Delta u = f & \text{on } \Omega, \\ u = 0 & \text{on } \partial\Omega \end{cases}$$

can be formulated in a weak sense as

$$\begin{cases} \text{find } u \in V \text{ such that} \\ \displaystyle\int_\Omega Du \cdot Dv\, dx = \int_\Omega f v\, dx \quad \forall v \in V, \end{cases} \tag{5.1}$$

where V is a functional space, containing $\mathcal{D}(\Omega)$, which has to be adequately chosen. Indeed, one can observe that problem (5.1) can be naturally attacked thanks to the Lax–Milgram theorem. One has to find a Hilbert space V such that the bilinear form

$$a(u, v) = \int_\Omega Du \cdot Dv\, dx$$

is continuous and coercive and the linear form

$$L(v) = \int_\Omega f v dx$$

is continuous. Noticing that

$$\begin{aligned} a(v, v) &= \int_\Omega |Dv|^2\, dx, \\ |a(u, v)| &\le \left(\int_\Omega |Du|^2\, dx\right)^{1/2} \left(\int_\Omega |Dv|^2\, dx\right)^{1/2}, \\ |L(v)| &\le \left(\int_\Omega f^2\, dx\right)^{1/2} \left(\int_\Omega v^2\, dx\right)^{1/2}, \end{aligned}$$

it is natural to take V equal to the completion of the space $D(\Omega)$ with respect to the norm

$$\|v\| = \left(\int_\Omega v^2 + |Dv|^2 \, dx \right)^{1/2}.$$

That is precisely the Sobolev space $V = H_0^1(\Omega)$. For pedagogical reasons, we prefer to introduce these spaces in a direct analytical way, by using the concept of distributional derivative. Then, we prove that regular functions are dense in the Sobolev spaces with respect to the corresponding Sobolev norms, which involve the L^p-norms of the functions and of their derivatives. This, combined with the completeness of the Sobolev spaces, establishes that Sobolev spaces are precisely the spaces provided by the above completion approach.

The key for a simple direct definition of Sobolev spaces is the notion of derivative in the sense of distributions developed in Section 2.2

5.1 Sobolev spaces: Definition, density results

Unless specified, in the following Ω is a general open subset of $\mathbf{R}^N$.

Definition 5.1.1. *The Sobolev space $H^1(\Omega)$ is defined by*

$$H^1(\Omega) = \left\{ v \in L^2(\Omega) \ : \ \frac{\partial v}{\partial x_i} \in L^2(\Omega) \quad i = 1, \ldots, N \right\},$$

where $\frac{\partial v}{\partial x_i}$ is taken in the distributional sense. The space $H^1(\Omega)$ is equipped with the scalar product

$$\langle u, v \rangle = \int_\Omega \left(uv + \sum_{i=1}^N \frac{\partial u}{\partial x_i} \frac{\partial v}{\partial x_i} \right) dx$$

and the corresponding norm

$$\|v\|_{H^1(\Omega)} = \left[\int_\Omega \left(v^2 + \sum_{i=1}^N \left(\frac{\partial v}{\partial x_i} \right)^2 \right) dx \right]^{1/2}.$$

Remark 5.1.1. By definition of distributional derivative, the following conditions are equivalent:

(a) $v \in H^1(\Omega)$;

(b) $v \in L^2(\Omega)$ and there exist $g_1, g_2, \ldots, g_N \in L^2(\Omega)$ such that

$$\forall \varphi \in \mathcal{D}(\Omega) \quad \int_\Omega v \, \frac{\partial \varphi}{\partial x_i} \, dx = - \int_\Omega g_i \varphi \, dx.$$

Then, by definition, $\frac{\partial v}{\partial x_i} = g_i$ in distribution sense.

The above definition can be naturally extended when replacing the $L^2(\Omega)$ space by a general $L^p(\Omega)$ space, $1 \leq p \leq +\infty$.

Definition 5.1.2. *For any $1 \le p \le +\infty$, the Sobolev space $W^{1,p}(\Omega)$ is defined by*

$$W^{1,p}(\Omega) = \left\{ v \in L^p(\Omega) \ : \ \frac{\partial v}{\partial x_i} \in L^p(\Omega) \quad i = 1, 2, \dots, N \right\},$$

where $\frac{\partial v}{\partial x_i}$ is taken in the distribution sense. The space $W^{1,p}(\Omega)$ is equipped with the norm

$$\|v\|_{W^{1,p}(\Omega)} = \left[\int_\Omega \left(|v|^p + \sum_{i=1}^N \left| \frac{\partial v}{\partial x_i} \right|^p \right) dx \right]^{1/p} \quad \text{for } 1 \le p < +\infty,$$

$$\|v\|_{W^{1,\infty}(\Omega)} = \max \left\{ \|v\|_\infty ; \left\| \frac{\partial v}{\partial x_1} \right\|_\infty ; \dots ; \left\| \frac{\partial v}{\partial x_N} \right\|_\infty \right\} \quad \text{for } p = +\infty.$$

When $p = 2$, the space $W^{1,2}(\Omega)$ is often denoted by $H^1(\Omega)$. Both notations are commonly used, the notation H^1 recalling the Hilbert structure which is so obtained when $p = 2$ (H is the initial of Hilbert). The next extension of the above notions is obtained when considering higher-order derivatives.

Definition 5.1.3. *Take m a nonnegative integer and $1 \le p \le +\infty$. The Sobolev space $W^{m,p}(\Omega)$ is defined by*

$$W^{m,p}(\Omega) = \{ v \in L^p(\Omega) \ : \ D^\alpha v \in L^p(\Omega) \ \forall \, \alpha \text{ with } |\alpha| \le m \},$$

where $D^\alpha v$ is the distribution derivative of v of symbol α. We recall that for $\alpha = (\alpha_1, \dots, \alpha_N)$,

$$D^\alpha v = \frac{\partial^{|\alpha|} v}{\partial x_1^{\alpha_1} \partial x_2^{\alpha_2} \dots \partial x_N^{\alpha_N}}$$

with $|\alpha| = \alpha_1 + \alpha_2 + \cdots + \alpha_N$. The space $W^{m,p}(\Omega)$ is equipped with the norm

$$\|v|_{W^{m,p}(\Omega)} = \left[\sum_{0 \le |\alpha| \le m} \int_\Omega |D^\alpha v|^p \, dx \right]^{1/p} \quad \text{for } 1 \le p < +\infty,$$

$$\|v\|_{W^{m,\infty}(\Omega)} = \max_{0 \le |\alpha| \le m} \|D^\alpha v\|_\infty \quad \text{for } p = +\infty.$$

When $p = 2$, we also use the notation $H^m(\Omega)$ for $W^{m,2}(\Omega)$ to enhance the Hilbertian structure of the space $W^{m,2}(\Omega)$.

The space $W^{1,p}(\Omega)$ may be also equipped with the equivalent norm

$$\|v\|_{L^p} + \sum_{i=1}^N \left\| \frac{\partial v}{\partial x_i} \right\|_{L^p},$$

but we prefer the choice of the norm made in Definition 5.1.2, because when $p = 2$ it yields a Hilbertian norm (which is not the case of the above one).

Clearly, $\mathcal{D}(\Omega)$ is always a subspace of $W^{m,p}(\Omega)$ for any $m \in \mathbf{N}$, $1 \le p \le +\infty$, so one can consider its closure in $W^{m,p}(\Omega)$.

Definition 5.1.4. *By definition,*

$H_0^1(\Omega)$ = *closure of* $\mathcal{D}(\Omega)$ *in* $H^1(\Omega)$,

$W_0^{1,p}(\Omega)$ = *closure of* $\mathcal{D}(\Omega)$ *in* $W^{1,p}(\Omega)$,

$W_0^{m,p}(\Omega)$ = *closure of* $\mathcal{D}(\Omega)$ *in* $W^{m,p}(\Omega)$.

As we will see, the elements of $H_0^1(\Omega)$ are precisely the elements of $H^1(\Omega)$ whose trace on $\partial\Omega$ is equal to zero. This, obviously, requires introduction of a notion of trace for functions of $H^1(\Omega)$ which extends the classical restriction operation for regular functions.

Example 5.1.1. Let us first examine the case $N = 1$, Ω being an open interval of $\mathbf{R}$.

(a) Take $\Omega =]-1, 1[$ an open interval in $\mathbf{R}$ and $v(x) = |x|$. Clearly, v is not differentiable in the classical sense at $x = 0$. Let us compute its distribution derivative. Given $\varphi \in \mathcal{D}(]-1, 1[)$ we have

$$\begin{aligned}\int_{-1}^{1} v(x)\varphi'(x)dx &= \int_{-1}^{0}(-x)\varphi'(x)dx + \int_{0}^{1} x\varphi'(x)dx \\ &= \int_{-1}^{0}\varphi(x)dx - \int_{0}^{1}\varphi(x)dx \\ &= -\int_{-1}^{1}\operatorname{sign}(x)\varphi(x)dx,\end{aligned}$$

where

$$\operatorname{sign}(x) = \begin{cases} -1 & \text{if} \quad -1 \le x < 0, \\ 1 & \text{if} \quad 0 < x \le 1. \end{cases}$$

So, $v'(x) = \operatorname{sign}(x)$ belongs to $L^\infty(\Omega)$ and v belongs to $W^{1,\infty}(-1, 1)$ and thus to any $W^{1,p}(-1, 1)$, $1 \le p \le +\infty$. Clearly, the above argument can be easily iterated and any continuous piecewise affine (and in fact piecewise $\mathbf{C}^1$) belongs to $W^{1,\infty}(a, b)$ for any bounded interval (a, b).

(b) Let us now exhibit a continuous function on a bounded interval of $\mathbf{R}$ which does not belong to H^1. Take $\Omega =]-1, 1[$ and consider for $0 < \alpha \le 1$ $v_\alpha(x) = |x|^\alpha$. A similar computation as above yields $v'_\alpha(x) = \alpha \operatorname{sign} x|x|^{\alpha-1}$.

Hence

$$\int_{-1}^{1} v'_\alpha(x)^2\,dx = \alpha^2 \int_{-1}^{1} x^{2\alpha-2}\,dx,$$

which is finite iff $2 - 2\alpha < 1$, i.e., $\alpha > \frac{1}{2}$. Thus, $|x|^\alpha$ belongs to $H^1(-1, 1)$ iff $\alpha > \frac{1}{2}$. As a consequence, the function $u(x)=\sqrt{|x|}$ does not belong to $H^1(-1, 1)$!

(c) We are going to prove that any element $v \in H^1(a, b)$ has a continuous representative. Indeed, this is a very special property of the one-dimensional case. It is no more true as soon as $N \ge 2$. We will further carefully study the regularity properties of elements of Sobolev spaces. Before proving this result, let us observe that a function v which has

a jump at one point in $\mathbf{R}$ does not belong to a Sobolev space $W^{1,p}$. For simplicity, take $\Omega =]-1, 1[$ and

$$v(x) = \begin{cases} 1 & \text{if} \quad x \geq 0, \\ 0 & \text{if} \quad x < 0. \end{cases}$$

An elementary computation yields that the distributional derivative of v is equal to the Dirac mass at the origin, indeed,

$$v' = \delta_0,$$

which, as we already observed, is a distribution which is not attached to a function $f \in L^1(\Omega)$.

Theorem 5.1.1. *Take $1 \leq p \leq +\infty$. Let $\Omega = (a, b)$ be an open interval of $\mathbf{R}$.*

(i) *Let $v \in W^{1,p}(a, b)$ and denote by $v' \in L^p(a, b)$ its first distributional derivative. Then there exists a continuous function $\tilde{v} \in \mathbf{C}([a, b])$ such that*

$$\begin{cases} v(x) = \tilde{v}(x) \text{ for almost every } x \in (a, b), \\ \tilde{v}(x) - \tilde{v}(y) = \displaystyle\int_y^x v'(t)dt \ \forall\, x, y \in [a, b]. \end{cases}$$

We say in this case that v admits a continuous representative $\tilde{v}$. Indeed $\tilde{v}$ is unique, and, when $p > 1$, it belongs to $\mathbf{C}^{0,\alpha}([a, b])$ with $\alpha = \frac{1}{p'}$, $\frac{1}{p} + \frac{1}{p'} = 1$, i.e.,

$$|\tilde{v}(x) - \tilde{v}(y)| \leq C\,|x - y|^\alpha \quad \forall x, y \in [a, b].$$

As a consequence,

$$v(x) - v(y) = \int_y^x v'(t)dt \quad \text{for almost every } x, y \in (a, b).$$

(ii) *Conversely, let us assume that $v \in L^p(\Omega)$ and there exists a function $g \in L^p(\Omega)$ such that*

$$v(x) - v(y) = \int_y^x g(t)dt \quad \text{for almost every } x, y \in (a, b).$$

Then, $v \in W^{1,p}(a, b)$ and $v' = g$ in the distributional sense.

Before proving Theorem 5.1.1, let us recall that the fundamental theorem of calculus states that for any $\mathbf{C}^1$ function v we have

$$v(x) = v(a) + \int_a^x v'(t)\,dt.$$

The above theorem allows us to extend such formula to the class of functions $v \in W^{1,p}$, the derivative v' being taken in the distributional sense and the integral in the Lebesgue sense.

Proof of Theorem 5.1.1. Let $v \in W^{1,p}(a, b)$ and $v' \in L^p(a, b)$ its distributional derivative. Let us denote by $w(\cdot)$ the function

$$w(x) = \int_a^x v'(t)\,dt.$$

When $1 < p \le +\infty$, $w(\cdot)$ is Hölder continuous: indeed, for all $x, y \in (a, b)$,

$$\begin{aligned} |w(y) - w(x)| &= \left| \int_x^y v'(t)\, dt \right| \\ &\le \int_x^y |v'(t)|\, dt \\ &\le |y - x|^{1/p'} \left(\int_x^y |v'(t)|^p\, dt \right)^{1/p} \\ &\le \|v'\|_{L^p(a,b)} |y - x|^{1/p'}. \end{aligned}$$

Let us prove that there exists a constant $C \in \mathbf{R}$ such that

$$v(x) - w(x) = C \quad \text{for almost every } x \in (a, b).$$

To that end, let us compute w' in a distributional sense and prove that $w' = v'$. Taking $\varphi \in \mathcal{D}(a, b)$ we have

$$\begin{aligned} \langle w', \varphi \rangle_{(D',D)} &= -\int_a^b w(x)\varphi'(x)\, dx \\ &= -\int_a^b \left(\int_a^x v'(t)\, dt \right) \varphi'(x)\, dx \\ &= -\int_a^b \left(\int_a^b 1_{[a,x]}(t) v'(t)\, dt \right) \varphi'(x)\, dx. \end{aligned}$$

The function $(t, x) \longmapsto 1_{[a,x]}(t)v'(t)\varphi'(x)$ belongs to $L^1((a, b) \times (a, b))$, which allows us to apply the Fubini theorem to obtain

$$\begin{aligned} \langle w', \varphi \rangle_{(D',D)} &= -\int_a^b v'(t) \left(\int_a^b 1_{[a,x]}(t)\varphi'(x)\, dx \right) dt \\ &= -\int_a^b v'(t) \left(\int_t^b \varphi'(x)\, dx \right) dt \\ &= \int_a^b v'(t)\varphi(t)\, dt \\ &= \langle v', \varphi \rangle_{(\mathcal{D}',\mathcal{D})}. \end{aligned}$$

Hence, $w' = v'$ in $\mathcal{D}'(a, b)$, and as a consequence $(w - v)' = 0$. We conclude the proof of assertion (i) by the help of the lemma below; the proof of assertion (ii) is easy and thus left to the reader. □

Lemma 5.1.1. *Let $f \in L^1_{loc}(a, b)$ such that $f' = 0$ in distributional sense. Then, there exists a constant $C \in \mathbf{R}$ such that $f(x) = C$ for almost every $x \in (a, b)$.*

Proof. By assumption $f' = 0$ in $\mathcal{D}'(a, b)$ which is equivalent to saying that

$$\int_a^b f(x)\varphi'(x)dx = 0 \quad \forall\, \varphi \in \mathcal{D}(a, b).$$

To exploit this information we need to understand what is the space

$$W = \{\varphi' \,:\, \varphi \in \mathcal{D}(a,b)\}.$$

Clearly, for any $\varphi \in \mathcal{D}(a,b)$ we have that φ' still belongs to $\mathcal{D}(a,b)$ and

$$\int_a^b \varphi'(x)dx = \varphi(b) - \varphi(a) = 0.$$

Conversely, take $\psi \in \mathcal{D}(a,b)$ such that $\int_a^b \psi(x)dx = 0$ and prove that $\psi = \varphi'$ for some $\varphi \in \mathcal{D}(a,b)$. Indeed

$$\varphi(x) = \int_a^x \psi(t)dt$$

belongs to $\mathbf{C}^\infty(a,b)$ and satisfies $\varphi' = \psi$. Assuming that $\psi \equiv 0$ outside of $[c,d]$ with $a < c < d < b$, we have $\varphi(x) = 0$ for all $x \le c$ and, for all $x > d$,

$$\varphi(x) = \int_a^x \psi(t)dt = \int_a^b \psi(t)dt = 0.$$

Hence $\varphi \in \mathcal{D}(a,b)$. We have obtained that

$$\begin{aligned} W &= \{\varphi' \,:\, \varphi \in \mathcal{D}(a,b)\} \\ &= \{\psi \in \mathcal{D}(a,b) : \int_a^b \psi(x)dx = 0\}. \end{aligned}$$

To construct functions $\psi \in W$, we first take some function $\theta \in D(a,b)$ such that $\int_a^b \theta(x)\,dx = 1$. Then, observe that for any $\chi \in D(a,b)$ the function $h(\cdot)$ which is defined by

$$h(x) := \chi(x) - \left(\int_a^b \chi(t)\,dt\right)\theta(x)$$

satisfies $h \in \mathcal{D}(a,b)$ and $\int_a^b h(x)dx = 0$. Hence $h \in W$ and

$$\int_a^b f(x)h(x)\,dx = 0.$$

Equivalently,

$$\int_a^b f(x)\chi(x)dx = \left(\int_a^b \chi(t)dt\right)\int_a^b f(x)\theta(x)\,dx.$$

Denoting $C := \displaystyle\int_a^b f(x)\theta(x)dx$, we obtain

$$\forall \chi \in \mathcal{D}(a,b) \quad \int_a^b (f(x) - C)\chi(x)\,dx = 0.$$

We conclude thanks to Theorem 2.2.1. □

As soon as the dimension N is greater than or equal to 2, the situation is more complex, and, in general, elements of $H^1(\Omega)$ have no continuous representative. Let us illustrate that fact with the help of the following example.

Example 5.1.2. Take $N = 2$ and $\Omega = B(0, R)$ with $R < 1$, that is, $\Omega = \{x = (x_1, x_2) \in \mathbf{R}^2 : |x| = \sqrt{x_1^2 + x_2^2} < R\}$. On Ω, we consider the function $v(x) = \big|\ln|x|\big|^k$, where k is a real parameter.

Let us examine, depending on the value of the parameter k, whether the function v belongs to $H^1(\Omega)$. To do so, it is natural to use radial coordinates and take $r = |x|$. As a classical result, we have $|\nabla v|^2 = (\frac{\partial \tilde{v}}{\partial r})^2$, where $v(x) = \tilde{v}(r)$. Here $\tilde{v}(r) = |\ln r|^k$. It follows

$$\int_{B(0,R)} v^2(x)dx = 2\pi \int_0^R (\ln r)^{2k} r\, dr,$$

$$\int_{B(0,R)} |\nabla v(x)|^2 dx = 2\pi \int_0^R \frac{k^2}{r^2} (\ln r)^{2k-2} r\, dr.$$

Let us make the change of variable $t = -\ln r$ to obtain

$$\int_\Omega (v^2 + |\nabla v|^2)dx = 2\pi \int_{-\ln R}^{+\infty} t^{2k} e^{-2t}\, dt + 2\pi k^2 \int_{-\ln R}^{+\infty} \frac{dt}{t^{2-2k}}.$$

It follows that v belongs to $H^1(\Omega)$ iff $2 - 2k > 1$, i.e., $k < \frac{1}{2}$. By taking $0 < k < \frac{1}{2}$, we obtain a function belonging to $H^1(\Omega)$ which blows up to $+\infty$ at the origin and which does not have a continuous representative.

Remark 5.1.2. It is a central question to know the best regularity results on the elements of the Sobolev spaces. As a general rule, for a function $v \in L^p(\Omega)$, which is a priori defined only almost everywhere on Ω, to know that some of its distribution derivatives belong to $L^p(\Omega)$ allows us, even if the function v has no continuous representative, to treat it more precisely than almost everywhere with respect to x. We will describe further three distinct approaches to this question: the Sobolev embedding theorem, the trace theory, and the capacity theory.

Let us now examine the general properties of the Sobolev spaces.

Theorem 5.1.2. *Let Ω be an open subset of $\mathbf{R}^N$. For any nonnegative integer m and any real number p with $1 \le p \le +\infty$, $W^{m,p}(\Omega)$ is a Banach space. When $p = 2$, $W^{m,p}(\Omega) = H^m(\Omega)$ is a Hilbert space.*

PROOF. Let $(v_n)_{n\in\mathbf{N}}$ be a Cauchy sequence in $W^{m,p}(\Omega)$. For any multi-index $\alpha = (\alpha_1, \alpha_2, \ldots, \alpha_N)$ with $|\alpha| \le m$, we have

$$\|D^\alpha v_n - D^\alpha v_n\|_{L^p} \le \|v_n - v_m\|_{W^{m,p}}.$$

Hence, the sequence $(D^\alpha v_n)_{n\in\mathbf{N}}$ is a Cauchy sequence in $L^p(\Omega)$, which is a Banach space. Let

$$\begin{array}{lll} v_n \to v & \text{in} \quad L^p(\Omega) & \text{as } n \to +\infty, \\ D^\alpha v_n \to g_\alpha & \text{in} \quad L^p(\Omega) & \text{as } n \to +\infty. \end{array}$$

Convergence in $L^p(\Omega)$ clearly implies convergence in the distribution sense.

By continuity of the derivation operation with respect to the convergence in the distribution sense (see Proposition 2.2.7), we obtain

$$D^\alpha v_n \to g_\alpha = D^\alpha v \quad \text{in } L^p(\Omega) \text{ as } n \to +\infty.$$

Hence $v \in W^{m,p}(\Omega)$ and the sequence $(v_n)_{n\in\mathbf{N}}$ converges to v in $W^{m,p}(\Omega)$. □

Remark 5.1.3. Notice that the above result relies essentially on the fact that the Sobolev spaces are built by means of the Lebesgue spaces $L^p(\Omega)$ (which are Banach spaces) and of the generalized notion of derivation, namely, the distributional derivative.

The following important result makes the link between the definition of Sobolev spaces relying on the notion of distributional derivative and the completion approach.

Theorem 5.1.3. *Let $\Omega = \mathbf{R}^N$. For any $1 \le p < +\infty$ and $m \in \mathbf{N}$, $\mathcal{D}(\mathbf{R}^N)$ is a dense subspace of $W^{m,p}(\mathbf{R}^N)$. Equivalently,*

$W_0^{m,p}(\mathbf{R}^N) = W^{m,p}(\mathbf{R}^N)$,
$W^{m,p}(\mathbf{R}^N)$ is the completion of $\mathcal{D}(\mathbf{R}^N)$ with respect to the $\|\cdot\|_{W^{m,p}}$ norm.

Proof. For simplicity of notation, let us consider the case $m = 1$ and prove that $\mathcal{D}(\mathbf{R}^N)$ is dense in $W^{1,p}(\mathbf{R}^N)$, $(1 \le p < +\infty)$. This is a two-step approximation procedure:

(a) Truncation (of the domain). Let $M \in \mathcal{D}(\mathbf{R}^N)$ such that $M(0) = 1$. For any $n \in \mathbf{N}^*$, let us define $M_n(\xi) = M(\frac{\xi}{n})$. Clearly, $M_n \in \mathcal{D}(\mathbf{R}^N)$ and, for any $\xi \in \mathbf{R}^N$, $\lim_{n\to+\infty} M_n(\xi) = M(0) = 1$.

The most commonly used truncation function is a function $M : \mathbf{R}^N \to \mathbf{R}^+$ such that $0 \le M \le 1$, $M(x) = 1$ on $B(0, \frac{1}{2})$, and $M(x) = 0$ for $\|x\| \ge 1$. Such a function exists by using the Tietze–Urysohn lemma, or it can be explicitly described if needed. But notice that for our purpose, we will only exploit the fact that $M \in \mathcal{D}(\mathbf{R}^N)$ and $M(0) = 1$.

For any $v \in W^{1,p}(\mathbf{R}^N)$, let us define $v_n(x) = M_n(x)v(x)$. Clearly, v_n has a compact support, since

$$\operatorname{spt} v_n \subset \operatorname{spt} M_n \subset n \operatorname{spt} M.$$

Let us verify that v_n still belongs to $W^{1,p}(\mathbf{R}^N)$.

$$\int_{\mathbf{R}^N} |v_n|^p dx \le \|M\|_\infty^p \int_{\mathbf{R}^N} |v|^p dx < +\infty.$$

On the other hand, the classical differentiation rule is still valid in this context. It is worthwhile to state it as a lemma.

Lemma 5.1.2. *Let $v \in W^{1,p}(\Omega)$ and $M \in \mathcal{D}(\Omega)$. Then $Mv \in W^{1,p}(\Omega)$ and*

$$\frac{\partial}{\partial x_i}(Mv) = M\frac{\partial v}{\partial x_i} + \frac{\partial M}{\partial x_i}v.$$

PROOF. Take $\varphi \in \mathcal{D}(\Omega)$ as a test function. By definition,

$$\left\langle \frac{\partial}{\partial x_i}(Mv), \varphi \right\rangle_{(\mathcal{D}'(\Omega),\mathcal{D}(\Omega))} = -\left\langle Mv, \frac{\partial \varphi}{\partial x_i} \right\rangle_{(\mathcal{D}',\mathcal{D})} = -\int_\Omega M(x)v(x)\frac{\partial \varphi}{\partial x_i}\,dx.$$

Since $M \in \mathcal{D}(\Omega)$, we can use the classical differentiation rule

$$\frac{\partial}{\partial x_i}(M\varphi) = M\frac{\partial \varphi}{\partial x_i} + \frac{\partial M}{\partial x_i}\varphi.$$

Hence

$$\left\langle \frac{\partial}{\partial x_i}(Mv), \varphi \right\rangle_{(D',D)} = -\int_\Omega v.\frac{\partial}{\partial x_i}(M\varphi)dx + \int_\Omega v\,\frac{\partial M}{\partial x_i}\,\varphi\,dx.$$

Noticing that $M\varphi \in \mathcal{D}(\Omega)$ we obtain

$$\left\langle \frac{\partial}{\partial x_i}(Mv), \varphi \right\rangle_{(D',D)} = \left\langle \frac{\partial v}{\partial x_i}, M\varphi \right\rangle_{(D',D)} + \int v\frac{\partial M}{\partial x_i}\,\varphi\,dx.$$

Since $v \in W^{1,p}(\Omega)$,

$$\left\langle \frac{\partial}{\partial x_i}(Mv), \varphi \right\rangle_{(D',D)} = \int_\Omega \left(\frac{\partial v}{\partial x_i}M + v\frac{\partial M}{\partial x_i}\right)\varphi\,dx = \left\langle \frac{\partial v}{\partial x_i}M + v\frac{\partial M}{\partial x_i},\ \varphi \right\rangle_{(D',D)}.$$

We conclude by noticing that

$$\frac{\partial v}{\partial x_i}M + v\frac{\partial M}{\partial x_i} \in L^p(\Omega). \quad \square$$

PROOF OF THEOREM 5.1.3 CONTINUED. Applying Lemma 5.1.2 we have

$$\frac{\partial v_n}{\partial x_i} = M_n\frac{\partial v}{\partial x_i} + \frac{\partial M_n}{\partial x_i}v.$$

Since M_n, $\frac{\partial M_n}{\partial x_i}$ belong to $\mathcal{D}(\mathbf{R}^N)$, we have $\frac{\partial v_n}{\partial x_i} \in L^p(\mathbf{R}^N)$.

Let us prove that the sequence $(v_n)_{n\in\mathbf{N}}$ norm-converges to v in $W^{1,p}(\mathbf{R}^N)$ as $n \to +\infty$:

$$\int_{\mathbf{R}^N} |v_n - v|^p dx = \int_{\mathbf{R}^N} |1 - M_n|^p\ |v|^p\,dx.$$

Then notice that

$$|1 - M_n|^p|v|^p \to 0 \text{ pointwise as } n \to +\infty,$$

$$|1 - M_n|^p|v|^p \le (1 + \|M\|_\infty)^p|v|^p \text{ for all } n \in \mathbf{N},$$

and apply the Lebesgue dominated convergence theorem to obtain that $v_n \to v$ in $L^p(\mathbf{R}^N)$. Similarly,

$$M_n\frac{\partial v}{\partial x_i} \to \frac{\partial v}{\partial x_i} \text{ in } L^p(\mathbf{R}^N) \text{ as } n \to +\infty.$$

To prove that $\frac{\partial v_n}{\partial x_i} \to \frac{\partial v}{\partial x_i}$ in $L^p(\mathbf{R}^N)$, we just need to prove that

$$\frac{\partial M_n}{\partial x_i} v \to 0 \text{ in } L^p(\mathbf{R}^N) \text{ as } n \to +\infty.$$

This follows immediately from

$$\begin{aligned}\left|\frac{\partial M_n}{\partial x_i}(x)v(x)\right| &= \frac{1}{n}\left|\frac{\partial M}{\partial x_i}\left(\frac{x}{n}\right)\right| |v(x)| \\ &\le \frac{1}{n}\left\|\frac{\partial M}{\partial x_i}\right\|_\infty |v(x)|,\end{aligned}$$

and hence

$$\left\|\frac{\partial M_n}{\partial x_i} v\right\|_{L^p} \le \frac{1}{n}\left\|\frac{\partial M}{\partial x_i}\right\|_\infty \|v\|_{L^p}.$$

(b) Regularization. The second step consists in proving that any element $u \in W^{1,p}(\mathbf{R}^N)$ with compact support can be approximated in $W^{1,p}(\mathbf{R}^N)$ by a sequence $(u_n)_{n\in\mathbf{N}}$ of elements in $\mathcal{D}(\mathbf{R}^N)$. To that end, we use the regularization by convolution method. With the same notation as in Section 2.2.2, we introduce $\rho \in \mathcal{D}(\mathbf{R}^N)$, $\rho \ge 0$ with $\operatorname{spt}\rho \subset B(0,1)$, and $\int_{\mathbf{R}^N} \rho(x)\,dx = 1$ and define for each $n \in \mathbf{N}^*$

$$\rho_n(x) := n^N \rho(nx),$$

which satisfies

$$\begin{cases} \rho_n \in \mathcal{D}(\mathbf{R}^N),\ \rho_n \ge 0, \\ \operatorname{spt}\rho_n \subset \mathbf{B}(0,1/n), \\ \displaystyle\int_{\mathbf{R}^N} \rho_n(x)\,dx = 1. \end{cases}$$

Given $u \in W^{1,p}(\mathbf{R}^N)$, let us define for all $n \in \mathbf{N}^*$, $u_n := u * \rho_n$, that is,

$$u_n(x) = \int_{\mathbf{R}^N} u(x-y)\rho_n(y)\,dy \tag{5.2}$$

$$= \int_{\mathbf{R}^N} u(y)\rho_n(x-y)\,dy. \tag{5.3}$$

To pass from one equality to the other in the formula above, one just has to make a change of variable and use that the Lebesgue measure on $\mathbf{R}^N$ is invariant by translation. Depending on the situation, we will use one formula or the other: note that the x variable appears in (5.2) in u, while in (5.3) it appears in ρ_n.

Let us verify that for each $n \in \mathbf{N}$, u_n belongs to $\mathcal{D}(\mathbf{R}^N)$. This means that u_n has a compact support and that u_n belongs to $\mathbf{C}^\infty(\mathbf{R}^N)$. Take $R > 0$ such that $u(x) = 0$ for $|x| > R$. By construction, for each $n \in \mathbf{N}^*$, $\rho_n(x) = 0$ for $|x| > \frac{1}{n}$. Let us verify that $u_n(x) = u * \rho_n(x)$ is equal to zero when $|x| > R + \frac{1}{n}$. This follows from the fact that the function $y \longmapsto u(x-y)\rho_n(y)$ is identically zero when $|x| > R + \frac{1}{n}$. Indeed, either $|y| > \frac{1}{n}$, in which case $\rho_n(y) = 0$, or $|y| \le \frac{1}{n}$, in which case

$$|x-y| \ge |x| - |y| > R + \frac{1}{n} - \frac{1}{n} = R$$

and $u(x-y) = 0$.

(One should notice that the above argument can be easily generalized to obtain that for any two functions f and g, $\mathrm{spt}(f \star g) \subset \mathrm{spt}\, f + \mathrm{spt}\, g$).

Let us now verify that $u_n = u * \rho_n$ belongs to $\mathbf{C}^\infty(\mathbf{R}^N)$. At this point, one has to use precisely formula (5.3), where the x variable, with respect to which one wants to derivate, appears in ρ_n. It is the differentiability property of ρ_n, which makes $u_n = u * \rho_n$ differentiable too!

One has to derivate under the sum sign, which is a direct consequence of the differentiability properties of ρ_n and of the Lebesgue dominated convergence theorem. As an illustration, let us prove that for each $i = 1, 2, \ldots, N$, for each $x \in \mathbf{R}^N$, $\frac{\partial u_n}{\partial x_i}(x)$ exists and

$$\frac{\partial u_n}{\partial x_i}(x) = \left(\frac{\partial u}{\partial x_i} * \rho_n\right)(x).$$

To avoid any confusion, we prefer to write $\frac{\partial u_n}{\partial e_i}$ instead of $\frac{\partial u_n}{\partial x_i}$, where e_i is the ith vector of the canonical basis of $\mathbf{R}^N$. For any $t \neq 0$ let us consider the differential quotient

$$\frac{1}{t}[u_n(x + te_i) - u_n(x)] = \int_{\mathbf{R}^N} u(y) \frac{\rho_n(x - y + te_i) - \rho_n(x - y)}{t}\, dy.$$

Clearly

$$u(y) \frac{\rho_n(x - y + te_i) - \rho_n(x - y)}{t} \to u(y) \left(\frac{\partial \rho_n}{\partial e_i}\right)(x - y) \text{ as } t \to 0,$$

$$\left| u(y) \frac{\rho_n(x - y + te_i) - \rho_n(x - y)}{t} \right| \leq |u(y)| \sup_{z \in \mathbf{R}^N} \left| \frac{\partial \rho_n}{\partial e_i}(z) \right| \leq M|u(y)|,$$

which belongs to $L^p(\mathbf{R}^N)$. Note that in the above inequality, we have used that for each $n \in \mathbf{N}$, since ρ_n belongs to $\mathcal{D}(\mathbf{R}^N)$, ρ_n and its partial derivatives are bounded functions on $\mathbf{R}^N$. (M here is a constant which depends on n.) By using the Lebesgue dominated convergence theorem, we obtain

$$\frac{\partial u_n}{\partial e_i}(x) = \int_{\mathbf{R}^N} u(y) \left(\frac{\partial \rho_n}{\partial e_i}\right)(x - y)\, dy.$$

Let us now notice that for x fixed, the function $y \longmapsto (\frac{\partial \rho_n}{\partial e_i})(x - y)$ can be written as a partial derivative of a function of $\mathcal{D}(\mathbf{R}^N)$. Take $\xi_n(y) = -\rho_n(x - y)$. Then

$$\begin{aligned} \frac{\partial \xi_n}{\partial e_i}(y) &= \lim_{t \to 0} \frac{1}{t}[\xi_n(y + te_i) - \xi_n(y)] \\ &= \lim_{t \to 0} \frac{\rho_n(x - y - te_i) - \rho_n(x - y)}{-t} \\ &= \left(\frac{\partial \rho_n}{\partial e_i}\right)(x - y). \end{aligned}$$

Thus

$$\frac{\partial u_n}{\partial e_i}(x) = \int_{\mathbf{R}^N} u(y) \frac{\partial \xi_n}{\partial e_i}(y)\, dy.$$

Since $u \in W^{1,p}(\mathbf{R}^N)$, and by definition of the distribution derivative

$$\begin{aligned}\frac{\partial u_n}{\partial e_i}(x) &= -\int_{\mathbf{R}^N} \frac{\partial u}{\partial e_i}(y)\xi_n(y)dy \\ &= \int_{\mathbf{R}^N} \frac{\partial u}{\partial e_i}(y)\ \rho_n(x-y)dy,\end{aligned}$$

that is,

$$\frac{\partial u_n}{\partial e_i} = \frac{\partial u}{\partial e_i} * \rho_n.$$

Returning to the usual notation with $\frac{\partial}{\partial x_i}$ instead of $\frac{\partial}{\partial e_i}$ we have

$$\frac{\partial u_n}{\partial x_i} = \frac{\partial u}{\partial x_i} * \rho_n.$$

Let us apply Proposition 2.2.4(iii) to obtain

$$\frac{\partial u_n}{\partial x_i} \to \frac{\partial u}{\partial x_i} \text{ in } L^p(\mathbf{R}^N) \text{ as } n \to +\infty,$$

that is, $u_n \to u$ in $W^{1,p}(\mathbf{R}^N)$. □

In general, when $\Omega \neq \mathbf{R}^N$, the two spaces $W_0^{1,p}(\Omega)$ and $W^{1,p}(\Omega)$ do not coincide. The space $W_0^{1,p}(\Omega)$ is strictly included in $W^{1,p}(\Omega)$. We will justify this fact a little further by proving that $W_0^{1,p}(\Omega)$ consists of functions whose trace on $\partial\Omega$ is equal to zero.

The strict inclusion $W_0^{1,p}(\Omega) \subsetneq W^{1,p}(\Omega)$ for $\Omega \neq \mathbf{R}^N$ (at least for Ω with a smooth boundary) can be seen also as a consequence of the following result of independent interest.

Proposition 5.1.1. *Let Ω be an open set in $\mathbf{R}^N$ and let $u \in W_0^{m,p}(\Omega)$. Then, the function $\tilde{u}$ which is equal to u on Ω and zero on $\mathbf{R}^N \setminus \Omega$ belongs to $W^{m,p}(\mathbf{R}^N)$. The linear mapping p defined by $p(u) = \tilde{u}$ is an isometry from $W_0^{m,p}(\Omega)$ into $W^{m,p}(\mathbf{R}^N)$.*

Proof. Let us consider the linear mapping

$$\begin{aligned} p\ :\ \mathcal{D}(\Omega) &\longrightarrow \mathcal{D}(\mathbf{R}^N) \\ u &\longmapsto p(u) = \tilde{u}.\end{aligned}$$

It is important to notice that since $u \in \mathcal{D}(\Omega)$, for each $x \in \partial\Omega$ there exists a neighborhood of x on which u is equal to zero. This clearly implies that $\tilde{u} \in \mathcal{D}(\mathbf{R}^N)$. Moreover, p is an isometry for the $W^{m,p}$ norms:

$$\forall u \in \mathcal{D}(\Omega) \quad \|p(u)\|_{W^{m,p}(\mathbf{R}^N)} = \|u\|_{W^{m,p}(\Omega)}.$$

Thus, p is a uniformly continuous mapping

$$p : \mathcal{D}(\Omega) \to W^{m,p}(\mathbf{R}^N),$$

and moreover $W^{m,p}(\mathbf{R}^N)$ is a Banach space. It can be continuously extended into a mapping (that we still denote by p)

$$p : \overline{\mathcal{D}(\Omega)}^{W^{m,p}(\Omega)} = W_0^{m,p}(\Omega) \to W^{m,p}(\mathbf{R}^N).$$

For $u \in W_0^{m,p}(\Omega)$, there exists a sequence $u_n \in \mathcal{D}(\Omega)$ such that $u_n \to u$ in $W_0^{m,p}(\Omega)$, and so

$$p(u_n) = \tilde{u}_n \to p(u) \quad \text{in } W^{m,p}(\mathbf{R}^N).$$

The convergence in $W^{m,p}$ implies the convergence in L^p, which implies the convergence almost everywhere of a subsequence. It follows that

$$p(u) = \tilde{u}, \text{ that is } \tilde{u} \in W^{m,p}(\mathbf{R}^N).$$

Moreover, for any $u \in W_0^{m,p}(\Omega)$, $\|p(u)\|_{W^{m,p}(\mathbf{R}^N)} = \|u\|_{W^{m,p}(\Omega)}$. □

Example 5.1.3. Take Ω a smooth, bounded domain in $\mathbf{R}^N$, for example, $\Omega = \mathbf{B}(0,1)$ the open ball centered at the origin with radius one. The constant function $u \equiv 1$ on Ω belongs to $W^{1,p}(\Omega)$ but does not belong to $W_0^{1,p}(\Omega)$. Otherwise, its extension $\tilde{u}$ by zero outside of Ω would belong to $W^{1,p}(\mathbf{R}^N)$, which is not true since its first partial derivatives $\frac{\partial \tilde{u}}{\partial x_i}$ are measures supported by the sphere $\partial\Omega$. When $N = 1$, $\frac{du}{dx} = \delta_{\{-1\}} - \delta_{\{1\}}$, a distribution which cannot be represented by an integrable function!

When $\Omega \subset \mathbf{R}^N$, one can adapt the previous argument, that is, truncation on the domain and regularization by convolution.

Theorem 5.1.4 (Meyers–Serrin). *If $1 \le p < \infty$, then $\mathbf{C}^\infty(\Omega) \cap W^{m,p}(\Omega)$ is dense in $W^{m,p}(\Omega)$.*

Proof. For each integer $k \ge 1$ set

$$\Omega_k = \left\{ x \in \Omega \ : \ \|x\| \le k \text{ and } \operatorname{dist}(x, \partial\Omega) > \frac{1}{k} \right\}$$

and $\Omega_0 = \emptyset$ the empy set. Define for each $k = 1, 2, \ldots$

$$G_k := \Omega_{k+1} \cap (\overline{\Omega}_{k-1})^c.$$

Then $(G_k)_{k\ge1}$ is an open covering of Ω. Let $(\alpha_k)_{k\ge1}$ be a $\mathbf{C}^\infty$ partition of unity for Ω subordinate to the open cover $(G_k)_{k\ge1}$, that is,

$$\begin{cases} \operatorname{spt}\alpha_k \subset G_k, \alpha_k \in \mathcal{D}(G_k), \alpha_k \ge 0, \\ \displaystyle\sum_{k=1}^{+\infty} \alpha_k = 1. \end{cases}$$

Let $u \in W^{m,p}(\Omega)$. Given $\varepsilon > 0$, we are going to construct an element $\varphi \in \mathbf{C}^\infty(\Omega)$ such that $\|u - \varphi\|_{W^{m,p}(\Omega)} < \varepsilon$.

Consider the truncated function $\alpha_k u$. The same argument as in Theorem 5.1.3 applies and $\alpha_k u \in W_0^{m,p}(G_k)$. It can be extended by zero outside of G_k into a function belonging to $W^{m,p}(\mathbf{R}^N)$. We now use a regularization by convolution method with a kernel $(\rho_n)_{n\in\mathbf{N}}$. By taking $n = n(k)$ sufficiently large, we have that

$$\|\rho_{n(k)} * (\alpha_k u) - \alpha_k u\|_{W^{m,p}(\Omega)} \leq \frac{\varepsilon}{2^k}$$

and

$$\mathrm{spt}[\rho_{n(k)} * (\alpha_k u)] \subset G_k.$$

Define $\varphi := \sum_{k=1}^{+\infty} \rho_{n(k)} * (\alpha_k u)$ and note that for $x \in G_k$ we have

$$\varphi(x) = \sum_{j=-1}^{+1} [\rho_{n(k+j)} * (\alpha_{k+j} u)](x),$$

i.e., there are at most three terms which are nonzero in this sum. Hence $\varphi \in \mathbf{C}^\infty(\Omega)$ and

$$\begin{aligned} \|u - \varphi\|_{W^{m,p}(\Omega)} &= \left\| \sum_{k=1}^{\infty} \big(\alpha_k u - (\alpha_k u) * \rho_{n(k)}\big) \right\|_{W^{m,p}} \\ &\leq \sum_{k=1}^{\infty} \|\alpha_k u - (\alpha_k u) * \rho_{n(k)}\|_{W^{m,p}} \\ &\leq \varepsilon \end{aligned}$$

and the proof is complete. □

Remark 5.1.4. The above result shows that, equivalently, the space $W^{m,p}(\Omega)$ can be defined as the completion with respect to the $\|\cdot\|_{W^{m,p}(\Omega)}$ norm of the space $\mathcal{V} = \{v \in \mathbf{C}^\infty(\Omega) : \|v\|_{W^{m,p}(\Omega)} < +\infty\}$. This result was obtained in 1964 by Meyers and Serrin.

5.2 The topological dual of $H_0^1(\Omega)$. The space $H^{-1}(\Omega)$.

When studying a functional space, a fundamental question which arises is the description of its topological dual space. Let us first consider the Sobolev space $H_0^1(\Omega)$, in which case this question can be easily solved thanks to the Hilbert structure of $H_0^1(\Omega)$ and the density of $\mathcal{D}(\Omega)$.

Let us take $T \in H_0^1(\Omega)^*$ an element of the topological dual of $H_0^1(\Omega)$. By using the Riesz represensation theorem of the elements of the topological dual of a Hilbert space, we obtain the existence of a unique element $g \in H_0^1(\Omega)$ such that

$$\forall v \in H_0^1(\Omega) \quad T(v) = \langle v, g\rangle = \int_\Omega \left(vg + \sum_{i=1}^{N} \frac{\partial v}{\partial x_i} \frac{\partial g}{\partial x_i} \right) dx. \tag{5.4}$$

Moreover,

$$\|T\|_{H_0^1(\Omega)^*} = \|g\|_{H_0^1(\Omega)} = \left[\int_\Omega \left(g^2 + \sum_{i=1}^{N} \left(\frac{\partial g}{\partial x_i} \right)^2 \right) dx \right]^{1/2}. \tag{5.5}$$

This is a first description of $H_0^1(\Omega)^*$. Indeed, one can give another description of the elements of $H_0^1(\Omega)^*$ as distributions. To that end, let us introduce the identity mapping

$$i : \mathcal{D}(\Omega) \hookrightarrow H_0^1(\Omega).$$

To each element $T \in H_0^1(\Omega)^*$ one can associate its restriction $T \circ i$ to $\mathcal{D}(\Omega)$

$$T \circ i : \mathcal{D}(\Omega) \stackrel{i}{\to} H_0^1(\Omega) \stackrel{T}{\to} \mathbf{R}.$$

Noticing that $\mathcal{D}(\Omega)$ is dense in $H_0^1(\Omega)$ for the $\|\cdot\|_{H_0^1}$ norm, it is equivalent to know T or its restriction $T \circ i$ to $\mathcal{D}(\Omega)$. Moreover, the convergence for the topology of $\mathcal{D}(\Omega)$ (Definition 2.2.1) is stronger than the norm convergence in $H_0^1(\Omega)$. Therefore, $T \circ i$ is a continuous linear form on $\mathcal{D}(\Omega)$ and $T \circ i$ is a distribution. One can summarize the above results by saying that the mapping

$$\begin{aligned} H_0^1(\Omega)^* &\hookrightarrow \mathcal{D}'(\Omega), \\ T &\longmapsto T \circ i = T|_{\mathcal{D}(\Omega)} \end{aligned}$$

is a continuous embedding of $H_0^1(\Omega)^*$ into the space of distributions $\mathcal{D}'(\Omega)$.

Let us now give a precise description of the distributions which are so obtained: for any $\varphi \in \mathcal{D}(\Omega)$

$$\langle T \circ i, \varphi \rangle = T(\varphi) = \int_\Omega \left(g\varphi + \sum_{i=1}^N \frac{\partial \varphi}{\partial x_i} \frac{\partial g}{\partial x_i} \right) dx.$$

Let us write $g_0 = g$ and $g_i = -\dfrac{\partial g}{\partial x_i}$ for $i = 1, \ldots, N$. Then,

$$\begin{aligned} \langle T \circ i, \varphi \rangle &= \int_\Omega g_0 \varphi - \sum_{i=1}^N g_i \frac{\partial \varphi}{\partial x_i} \, dx \\ &= \left\langle g_0 + \sum_{i=1}^N \frac{\partial g_i}{\partial x_i}, \varphi \right\rangle_{(\mathcal{D}'(\Omega), \mathcal{D}(\Omega))}. \end{aligned}$$

Therefore, we can identify T with a distribution of the form $g_0 + \sum_{i=1}^N \frac{\partial g_i}{\partial x_i}$ with $g_0, g_1, \ldots, g_N \in L^2(\Omega)$. Moreover,

$$\|T\|_{H_0^1(\Omega)^*} = \|g\|_{H_0^1(\Omega)} = \left(\sum_{i=0}^N \int_\Omega g_i(x)^2 \, dx \right)^{1/2}. \tag{5.6}$$

Conversely, any distribution $T \in \mathcal{D}'(\Omega)$ which can be written as $T = g_0 + \sum_{i=1}^N \frac{\partial g_i}{\partial x_i}$ with $g_0, g_1 \ldots g_N \in L^2(\Omega)$, clearly satisfies

$$\begin{aligned} \forall \varphi \in \mathcal{D}(\Omega) \quad \langle T, \varphi \rangle_{(D',D)} &= \int_\Omega \left(g_0 \varphi - \sum_{i=1}^N g_i \frac{\partial \varphi}{\partial x_i} \right) dx \\ &\leq \left(\sum_{i=0}^N \int_\Omega g_i^2(x) \, dx \right)^{1/2} \|\varphi\|_{H_0^1(\Omega)}. \end{aligned}$$

As a consequence, T can be uniquely extended by density to a continuous linear form on $\overline{\mathcal{D}(\Omega)}^{H_0^1(\Omega)} = H_0^1(\Omega)$ which we still denote by T. Moreover,

$$\|T\|_{H_0^1(\Omega)^*} \leq \left(\sum_{i=0}^{N} \int_\Omega g_i(x)^2 \, dx \right)^{1/2}. \tag{5.7}$$

It is important to notice that there is not a unique way to write T as a sum $T = g_0 + \sum \frac{\partial g_i}{\partial x_i}$. Indeed, take any g such that $\Delta g = 0$, then $\sum \frac{\partial}{\partial x_i}(\frac{\partial g}{\partial x_i}) = 0$, and T can, as well, be written as $T = g_0 + \sum \frac{\partial}{\partial x_i}(g_i + \frac{\partial g}{\partial x_i})$. By using (5.6) and (5.7), we obtain that

$$\|T\|_{H_0^1(\Omega)^*} = \min \left\{ \left(\sum_{i=0}^{N} \int_\Omega g_i(x)^2 \, dx \right)^{1/2} : T = g_0 + \sum_{i=1}^{N} \frac{\partial g_i}{\partial x_i} \right\}. \tag{5.8}$$

Note that the minimum is achieved precisely by taking the $(g_0, \ldots, g_N)$ provided by the Riesz representation (5.4), (5.5).

Let us summarize the above results in the following statement.

Theorem 5.2.1. *Let us define $H^{-1}(\Omega)$ as the topological dual space of $H_0^1(\Omega)$, i.e., $H^{-1}(\Omega) = H_0^1(\Omega)^*$. Then, $H^{-1}(\Omega)$ is isometrically isomorphic to the Hilbert space of distributions $T \in \mathcal{D}'(\Omega)$ satisfying*

$$T = g_0 + \sum_{i=1}^{N} \frac{\partial g_i}{\partial x_i} \quad \text{for some } g_0, g_1, \ldots, g_N \in L^2(\Omega)$$

with

$$\|T\|_{H^{-1}(\Omega)} = \inf \left\{ \left(\sum_{i=0}^{N} \|g_i\|_{L^2(\Omega)}^2 \right)^{1/2} : T = g_0 + \sum_{i=1}^{N} \frac{\partial g_i}{\partial x_i} \right\}.$$

The above results can be easily extended to higher-order Sobolev spaces and to Sobolev spaces built on $L^p(\Omega)$ spaces instead of $L^2(\Omega)$. In that case, one uses the duality between $L^p(\Omega)$ and $L^{p'}(\Omega)$ with $\frac{1}{p} + \frac{1}{p'} = 1$, to obtain the following result.

Theorem 5.2.2. *Let $1 \leq p < +\infty$. The topological dual space of $W_0^{m,p}(\Omega)$ is denoted by $W^{-m,p'}(\Omega)$, where $\frac{1}{p} + \frac{1}{p'} = 1$. It is isometrically isomorphic to the Banach space consisting of those distributions $T \in \mathcal{D}'(\Omega)$ satisfying*

$$T = \sum_{0 \leq |\alpha| \leq m} D^\alpha g_\alpha \quad \text{with each } g_\alpha \in L^{p'}(\Omega)$$

with

$$\|T\|_{W^{-m,p'}(\Omega)} = \inf \left\{ \left(\sum_{0 \leq |\alpha| \leq m} \int_\Omega |g_\alpha|^{p'}(x) \, dx \right)^{1/p'} : T = \sum_{0 \leq |\alpha| \leq m} D^\alpha g_\alpha \right\}.$$

5.3 Poincaré inequality and Rellich–Kondrakov theorem in $W_0^{1,p}(\Omega)$

The Poincaré inequality is a basic ingredient of the variational approach to the Dirichlet problem. It provides the coercivity of the Dirichlet integral $\int_\Omega |Dv|^2\,dx$ on the space $H_0^1(\Omega)$.

Theorem 5.3.1 (Poincaré inequality). *Let Ω be an open subset of $\mathbf{R}^N$ which is bounded in one direction. Then, for each $1 \le p < +\infty$, there exists a constant $C_{p,N}(\Omega)$ which depends only on Ω, p, and N such that*

$$\left(\int_\Omega |v(x)|^p\,dx\right)^{1/p} \le C_{p,N}(\Omega)\left(\int_\Omega \sum_{i=1}^N \left|\frac{\partial v}{\partial x_i}\right|^p dx\right)^{1/p} \quad \forall\, v \in W_0^{1,p}(\Omega). \tag{5.9}$$

Proof. Since Ω is bounded in a direction, we can find a system of coordinates, which for simplicity we still denote $(x_1, x_2, \dots, x_N)$, such that Ω is contained in the strip $a \le x_N \le b$. Let us write $x = (x', x_N)$ with $x' = (x_1, x_2, \dots, x_{N-1})$.

Since $\mathcal{D}(\Omega)$ is dense in $W_0^{1,p}(\Omega)$ (by definition of $W_0^{1,p}(\Omega)$) let us first take $v \in \mathcal{D}(\Omega)$. We will then extend the result by a density and continuity argument. Let us define

$$\tilde v = \begin{cases} v & \text{on } \Omega, \\ 0 & \text{on } \mathbf{R}^N \setminus \Omega. \end{cases}$$

We have, for any $x \in \mathbf{R}^N$, with $a \le x_N \le b$, $x' \in \mathbf{R}^{N-1}$

$$\tilde v(x) = \tilde v(x', x_N) = \tilde v(x', a) + \int_a^{x_N} \frac{\partial \tilde v}{\partial x_N}(x', t)\,dt.$$

Since $\tilde v(x', a) = 0$

$$\tilde v(x', x_N) = \int_a^{x_N} \frac{\partial \tilde v}{\partial x_N}(x', t)\,dt.$$

Applying the Hölder inequality we obtain ($\frac{1}{p} + \frac{1}{p'} = 1$)

$$\begin{aligned} |\tilde v(x', x_N)|^p &\le (x_N - a)^{p/p'} \int_a^{x_N} \left|\frac{\partial \tilde v}{\partial x_N}(x', t)\right|^p dt \\ &\le (x_N - a)^{p/p'} \int_{\mathbf{R}} \left|\frac{\partial \tilde v}{\partial x_N}(x', t)\right|^p dt. \end{aligned}$$

Let us first integrate with respect to $x' \in \mathbf{R}^{N-1}$. We obtain

$$\int_{\mathbf{R}^{N-1}} |\tilde v(x', x_N)|^p\,dx' \le (x_N - a)^{p/p'} \int_{\mathbf{R}^N} \left|\frac{\partial \tilde v}{\partial x_N}(x)\right|^p dx.$$

Let us now integrate with respect to x_N ($a \le x_N \le b$) to obtain

$$\int_{\mathbf{R}^N} |\tilde v(x)|^p\,dx \le \frac{(b-a)^{1+p/p'}}{1 + p/p'} \int_{\mathbf{R}^N} \left|\frac{\partial \tilde v}{\partial x_N}(x)\right|^p dx.$$

Since $\frac{\partial \tilde{v}}{\partial x_N} = Dv \cdot n$, where n is a unit normal vector to the strip containing Ω, we have

$$\left|\frac{\partial \tilde{v}}{\partial x_N}\right|^p = \left|\sum_{i=1}^N \frac{\partial \tilde{v}}{\partial x_i} n_i\right|^p \le \left(\sum_{i=1}^N \left|\frac{\partial \tilde{v}}{\partial x_i}\right|^p\right)\left(\sum_{i=1}^N n_i^{p'}\right)^{p/p'}$$
$$\le N^{p/p'} \sum_{i=1}^N \left|\frac{\partial \tilde{v}}{\partial x_i}\right|^p.$$

Hence, noticing that $1 + \frac{p}{p'} = p$

$$\int_{\mathbf{R}^N} |\tilde{v}(x)|^p \, dx \le \frac{(b-a)^p}{p} N^{p/p'} \int_{\mathbf{R}^N} \sum_{i=1}^N \left|\frac{\partial \tilde{v}}{\partial x_i}\right|^p dx.$$

Since $\tilde{v}$ and all the $\frac{\partial \tilde{v}}{\partial x_i}$ $(i = 1, \dots, N)$ are equal to zero outside of Ω and are equal, respectively, to v and $\frac{\partial v}{\partial x_i}$ on Ω, we obtain

$$\left(\int_\Omega |v(x)|^p \, dx\right)^{1/p} \le C_{p,N}(\Omega) \left(\int_\Omega \sum_{i=1}^N \left|\frac{\partial v}{\partial x_i}\right|^p dx\right)^{1/p} \qquad \forall\, v \in \mathcal{D}(\Omega)$$

with $C_{p,N}(\Omega) = (b-a)\frac{N^{1/p'}}{p^{1/p}}$.

By the density of $\mathcal{D}(\Omega)$ in $W_0^{1,p}(\Omega)$, this inequality can be directly extended to $v \in W_0^{1,p}(\Omega)$. □

Definition 5.3.1. *The Poincaré constant is the smallest constant C for which the inequality (5.9) holds for all $v \in W_0^{1,p}(\Omega)$. We denote it by $\bar{C}_{p,N}(\Omega)$:*

$$\bar{C}_{p,N}(\Omega) = \inf\left\{ C \,:\, \left(\int_\Omega |v|^p \, dx\right)^{1/p} \le C \left(\int_\Omega \sum_{i=1}^N \left|\frac{\partial v}{\partial x_i}\right|^p dx\right)^{1/p} \quad \forall v \in W_0^{1,p}(\Omega) \right\}.$$

Equivalently,

$$\frac{1}{\bar{C}_{p,N}(\Omega)} = \inf\left\{ \left(\int_\Omega \sum_{i=1}^N \left|\frac{\partial v}{\partial x_i}\right|^p dx\right)^{1/p} \,:\, \int_\Omega |v|^p \, dx = 1,\ v \in W_0^{1,p}(\Omega) \right\}.$$

In some instances (see the "cloud of ice" in Attouch [28], for example), it is useful to know precisely how this constant depends on the size of Ω.

Proposition 5.3.1. *For any $R > 0$, for any Ω in $\mathbf{R}^N$,*

$$\bar{C}_{p,N}(R\Omega) = R\bar{C}_{p,N}(\Omega).$$

Proof. Let $v \in W_0^{1,p}(R\Omega)$. Let us define $v_R(x) = v(Rx)$. Clearly $v_R \in W_0^{1,p}(\Omega)$ and

$$\int_\Omega |v_R(x)|^p \, dx = \int_\Omega |v(Rx)|^p \, dx = R^{-N} \int_{R\Omega} |v(y)|^p \, dy.$$

$$\int_\Omega |Dv_R(x)|^p\,dx = \int_\Omega \sum_{i=1}^N \left|\frac{\partial v_R}{\partial x_i}(x)\right|^p dx = \int_\Omega \sum_{i=1}^N \left|\frac{\partial}{\partial x_i} v(Rx)\right|^p dx$$
$$= R^p \int_\Omega \sum_{i=1}^N \left|\frac{\partial v}{\partial x_i}(Rx)\right|^p dx$$
$$= R^{p-N} \int_{R\Omega} \sum_{i=1}^N \left|\frac{\partial v}{\partial x_i}(\xi)\right|^p d\xi.$$

Hence

$$\int_{R\Omega} |v(y)|^p\,dy = R^N \int_\Omega |v_R(x)|^p\,dx$$
$$\leq R^N \bar{C}_p(\Omega)^p \int_\Omega |Dv_R|^p\,dx$$
$$\leq R^N\, R^{p-N} \bar{C}_p(\Omega)^p \int_{R\Omega} |Dv|^p\,dx$$
$$= (R\,\bar{C}_p(\Omega))^p \int_{R\Omega} |Dv|^p\,dx.$$

This being true for any $v \in W_0^{1,p}(R\Omega)$ it follows

$$\bar{C}_{p,N}(R\Omega) \leq R\bar{C}_{p,N}(\Omega).$$

Conversely, since $\Omega = \frac{1}{R}(R\Omega)$, we have $\bar{C}_{p,N}(\Omega) \leq \frac{1}{R}\bar{C}_{p,N}(R\Omega)$ and thus the equality $\bar{C}_{p,N}(R\Omega) = R\bar{C}_{p,N}(\Omega)$ holds. □

We will see later when $p = 2$ in chapter 8, Theorem 8.4.1 how the Poincaré constant can be related to the first eigenvalue of the Laplacian operator with Dirichlet boundary condition.

Another basic result is the compact embedding theorem of Rellich–Kondrakov. For that purpose, we need to recall the Kolmogorov compactness criteria in $L^p(\mathbf{R}^N)$.

Theorem 5.3.2 (Kolmogorov). *Let $p \in [1, +\infty[$ and let $\mathcal{F}$ be a subset of $L^p(\mathbf{R}^N)$. Then, $\mathcal{F}$ is relatively compact in $L^p(\mathbf{R}^N)$ iff the three following conditions are satisfied:*

(i) *$\mathcal{F}$ is bounded in $L^p(\mathbf{R}^N)$;*

(ii) $\lim_{R\to+\infty} \int_{\{|x|>R\}} |v(x)|^p\,dx = 0$ *uniformly with respect to $v \in \mathcal{F}$;*

(iii) $\lim_{h\to 0} \|\tau_h v - v\|_{L^p(\mathbf{R}^N)} = 0$ *uniformly with respect to $v \in \mathcal{F}$, where $\tau_h v$ is the translated function $(\tau_h v)(x) := v(x-h)$.*

Proof. Let us prove the implication which is useful for applications, that is, (i), (ii), (iii) $\Longrightarrow$ $\mathcal{F}$ relatively compact in $L^p(\mathbf{R}^N)$.

Equivalently, we have to prove that $\mathcal{F}$ is precompact, which means that for any $\varepsilon > 0$, there exists a finite number of balls $B(v_1, \varepsilon), \ldots, B(v_k, \varepsilon)$ which cover $\mathcal{F}$. So, let us give

$\varepsilon > 0$. By (ii) there exists some $R > 0$ such that

$$\forall v \in \mathcal{F} \quad \int_{|x|>R} |v(x)|^p \, dx < \varepsilon.$$

Let $(\rho_n)_{n\in\mathbf{N}}$ be a mollifier. It follows from Proposition 2.2.4 that

$$\forall n \geq 1, \ \forall v \in L^p(\mathbf{R}^N) \quad \|v - v * \rho_n\|_{L^p}^p \leq \int_{\mathbf{R}^N} \rho_n(y) \|v - \tau_y v\|_{L^p}^p dy.$$

Hence

$$\|v - v * \rho_n\|_{L^p} \leq \sup_{|y|\leq 1/n} \|v - \tau_y v\|_{L^p}.$$

By (iii), there exists some integer $N(\varepsilon) \in \mathbf{N}$ such that

$$\forall v \in \mathcal{F} \quad \|v - v * \rho_{N(\varepsilon)}\|_{L^p} < \varepsilon.$$

On the other hand, for any $x, x' \in \mathbf{R}^N$, for any $v \in L^p(\mathbf{R}^N)$ and $n \in \mathbf{N}$,

$$\begin{aligned} |(v * \rho_n)(x) - (v * \rho_n)(x')| &\leq \int |v(x - y) - v(x' - y)| \rho_n(y) dy \\ &\leq \|\tau_x \check{v} - \tau_{x'} \check{v}\|_{L^p} \, \|\rho_n\|_{L^{p'}} \\ &\leq \|\tau_{x-x'} v - v\|_{L^p} \, \|\rho_n\|_{L^{p'}}. \end{aligned}$$

(Note that this last property follows from the invariance property of the Lebesgue measure.) Moreover,

$$|(v * \rho_n)(x)| \leq \|v\|_{L^p} \, \|\rho_n\|_{L^{p'}}.$$

Let us consider the family $\mathcal{H} = \{v * \rho_{N(\varepsilon)} \ : \ B(0, R) \to \mathbf{R}, \ v \in \mathcal{F}\}$. By using (i) and (iii), we have that it satisfies the conditions of the Ascoli theorem. Hence, it is precompact for the topology of the uniform convergence on $B(0, R)$, and we have the existence of a finite set $\{v_1, \ldots, v_k\}$ of elements of $\mathcal{F}$ such that

$$\bigcup_{i=1}^{k} \mathbf{B}(v_i * \rho_{N(\varepsilon)}, \varepsilon R^{-N/p}) \supset \mathcal{H}.$$

So, for all $v \in \mathcal{F}$, there exists some $j \in \{1, 2, \ldots, k\}$ such that

$$\forall x \in B(0, R) \quad |v * \rho_{N(\varepsilon)}(x) - v_j * \rho_{N(\varepsilon)}(x)| \leq \varepsilon |B(0, R)|^{-1/p}.$$

Hence,

$$\begin{aligned} \|v - v_j\|_{L^p(\mathbf{R}^N)} \leq &\left(\int_{|x|>R} |v|^p \, dx \right)^{1/p} + \left(\int_{|x|>R} |v_j|^p \, dx \right)^{1/p} \\ &+ \|v - v * \rho_{N(\varepsilon)}\|_{L^p} + \|v_j - v_j * \rho_{N(\varepsilon)}\|_{L^p} \\ &+ \|v * \rho_{N(\varepsilon)} - v_j * \rho_{N(\varepsilon)}\|_{L^p(B(0,R)}. \end{aligned}$$

The last term can be majorized as follows:

$$\|v * \rho_{N(\varepsilon)} - v_j * \rho_{N(\varepsilon)}\|_{L^p(B(0,R))} = \left(\int_{B(0,R)} |v * \rho_{N(\varepsilon)}(x) - v_j * \rho_{N(\varepsilon)}(x)|^p \, dx \right)^{1/p}$$
$$\leq \varepsilon |B(0,R)|^{-1/p} |B(0,R)|^{1/p} = \varepsilon.$$

Finally,

$$\|v - v_j\|_{L^p(\mathbf{R}^N)} \leq 5\varepsilon,$$

which proves that $\mathcal{F}$ is precompact in $L^p(\mathbf{R}^N)$. □

The rate of convergence in $L^p(\mathbf{R}^N)$ of $\tau_h v$ to v as $|h| \to 0$ can be made precise for functions v belonging to $W^{1,p}(\mathbf{R}^N)$.

Proposition 5.3.2. *For all $1 \leq p \leq +\infty$ and all $v \in W^{1,p}(\mathbf{R}^N)$ the following inequality holds:*

$$\forall h \in \mathbf{R}^N \quad \|\tau_h v - v\|_{L^p(\mathbf{R}^N)} \leq \|Dv\|_{L^p(\mathbf{R}^N)} |h|,$$

where $\|Dv\|_{L^p(\mathbf{R}^N)} = (\int_{\mathbf{R}^N} |Dv(x)|^p \, dx)^{1/p}$ and $|Dv(x)|$ is the Euclidean norm of $Dv(x)$.

Proof. Since $\mathcal{D}(\mathbf{R}^N)$ is dense in $W^{1,p}(\mathbf{R}^N)$, we just need to prove this inequality for $v \in \mathcal{D}(\mathbf{R}^N)$. We have

$$(\tau_h v)(x) - v(x) = v(x-h) - v(x)$$
$$= -\int_0^1 Dv(x - th)\, h \, dt.$$

Hence, by the Cauchy–Schwarz inequality

$$|(\tau_h v)(x) - v(x)| \leq \int_0^1 |Dv(x - th)|\, |h| \, dt,$$

and then, by Hölder's inequality,

$$|(\tau_h v)(x) - v(x)|^p \leq |h|^p \int_0^1 |Dv(x - th)|^p \, dt.$$

Integrating over $\mathbf{R}^N$ we have

$$\int_{\mathbf{R}^N} |(\tau_h v)(x) - v(x)|^p \, dx \leq |h|^p \int_{\mathbf{R}^N} \left(\int_0^1 |Dv(x - th)|^p \, dt \right) dx.$$

Let us use the Fubini–Tonelli theorem and the invariance by translation of the Lebesgue measure in $\mathbf{R}^N$ to obtain

$$\int_{\mathbf{R}^N} |\tau_h v - v|^p \, dx \leq |h|^p \int_{\mathbf{R}^N} |Dv(x)|^p \, dx,$$

which ends the proof. □

We can now state the main compactness theorem in Sobolev spaces.

Theorem 5.3.3 (Rellich–Kondrakov). *Let Ω be a bounded open subset of $\mathbf{R}^N$. Then the canonical embedding $W_0^{1,p}(\Omega) \hookrightarrow L^p(\Omega)$ is compact. In other words, every bounded subset of $W_0^{1,p}(\Omega)$ is relatively compact in $L^p(\Omega)$.*

PROOF. Let us denote by p the natural extension operator by zero outside of Ω, which is a continuous operator from $W_0^{1,p}(\Omega)$ into $W^{1,p}(\mathbf{R}^N)$. Indeed, it follows from Proposition 5.1.1 that p is a linear isometry.

The restriction operator $r : L^p(\mathbf{R}^N) \to L^p(\Omega)$ defined by $r(v) = v|_\Omega$ is clearly linear continuous (with norm less than or equal to one). So, the embedding

$$i : W_0^{1,p}(\Omega) \hookrightarrow L^p(\Omega)$$

can be written as the composition $i = r \circ j \circ p$,

$$W_0^{1,p}(\Omega) \xrightarrow{p} W^{1,p}(\mathbf{R}^N) \xrightarrow{j} L^p(\mathbf{R}^N) \xrightarrow{r} L^p(\Omega),$$

where j is the canonical embedding from $W^{1,p}(\mathbf{R}^N)$ into $L^p(\mathbf{R}^N)$.

Let $\mathbf{B}$ be the unit ball in $W_0^{1,p}(\Omega)$. Since r is continuous and the image of a compact set by a continuous mapping is still compact, we need to show that $(j \circ p)(\mathbf{B})$ is relatively compact in $L^p(\mathbf{R}^N)$. To that end, we use the Kolmogorov compactness criteria in $L^p(\mathbf{R}^N)$; cf. Theorem 5.3.2.

(i) Since j and p are linear continuous, $(j \circ p)(\mathbf{B})$ is bounded in $L^p(\mathbf{R}^N)$.

(ii) Since Ω is bounded, there exists $R > 0$ such that $\Omega \subset B(0, R)$. Hence for all $v \in \mathbf{B}$, $p(v) = 0$ on $\mathbf{R}^N \setminus B(0, R)$ and $\int_{\{|x|>R\}} |p(v)|^p \, dx = 0$.

(iii) Since $p(\mathbf{B})$ is contained in the unit ball of $W^{1,p}(\mathbf{R}^N)$, it follows from Proposition 5.3.2 that there exists a constant $C > 0$ such that

$$\forall v \in \mathbf{B} \quad \|\tau_h p(v) - p(v)\|_{L^p(\mathbf{R}^N)} \leq C|h|,$$

which proves that $\tau_h p(v)$ tends to $p(v)$ in $L^p(\mathbf{R}^N)$ as $h \to 0$, uniformly with respect to $v \in \mathbf{B}$.

This completes the proof. □

By a proof similar to the one above we obtain the following useful result.

Corollary 5.3.1. *Let $\mathcal{F}$ be a subset of $W^{1,p}(\mathbf{R}^N)$, $1 \leq p < +\infty$ which satisfies the two following conditions:*

(i) *$\mathcal{F}$ is bounded in $W^{1,p}(\mathbf{R}^N)$, i.e., $\sup_{v\in\mathcal{F}} \|v\|_{W^{1,p}(\mathbf{R}^N)} < +\infty$.*

(ii) *$\mathcal{F}$ is L^p-equi-integrable at infinity, i.e., $\lim_{R\to+\infty} \int_{\{|x|>R\}} |v(x)|^p \, dx = 0$ uniformly with respect to $v \in \mathcal{F}$.*

Then, $\mathcal{F}$ is relatively compact in $L^p(\mathbf{R}^N)$.

5.4 Extension operators from $W^{1,p}(\Omega)$ into $W^{1,p}(\mathbf{R}^N)$. Poincaré inequalities and the Rellich–Kondrakov theorem in $W^{1,p}(\Omega)$

We have been able to prove some important properties of the space $W_0^{1,p}(\Omega)$, without any regularity assumptions on the boundary of the open set Ω, because there always exists a continuous extension operator from $W_0^{1,p}(\Omega)$ into $W^{1,p}(\mathbf{R}^N)$, namely, the extension by zero (Proposition 5.1.1). When working with the space $W^{1,p}(\Omega)$, the situation is more delicate and the regularity of the boundary of Ω will play a crucial role in the proofs and the statements of the results. Once more, a basic ingredient will be the obtainment of an extension operator. So doing, we will be able to use the previous results in $W^{1,p}(\mathbf{R}^N)$, where techniques like convolution and translation naturally apply.

Notation. given $x \in \mathbf{R}^N$ we write $x = (x', x_N)$ with $x' \in \mathbf{R}^{N-1}$, $x' = (x_1, x_2, \ldots, x_{N-1})$. We write

$\mathbf{R}^N_+ = \{x = (x', x_N) \ : \ x_N > 0\}$ the open upper half-space;

$B = B(0,1) = \{x \in \mathbf{R}^N \ : \ |x| = (\sum_{i=1}^N x_i^2)^{1/2} < 1\}$ the open unit ball in $\mathbf{R}^N$;

$B_+ = B(0,1) \cap \mathbf{R}^N_+$;

$B_0 = B \cap \mathbf{R}^{N-1} = \{x = (x', x_N) \in \mathbf{R}^N \ : \ |x'| \leq 1 \text{ and } x_N = 0\}$.

A $\mathbf{C}^1$-diffeomorphism from an open set $U \subset X$ into an open set $V \subset Y$, where X and Y are normed linear spaces, is a one-to-one mapping φ from U into V which is continuously differentiable and such that its inverse φ^{-1} is continuously differentiable from V onto U.

Definition 5.4.1. *Let Ω be an open subset of $\mathbf{R}^N$. We say that Ω is of class $\mathbf{C}^1$ if for all $x \in \Gamma = \partial\Omega$ (the topological boundary of Ω), there exists an open neighborhood G of x and a $\mathbf{C}^1$-diffeomorphism φ from $B(0,1)$ onto G such that*

$$\varphi(B_+) = G \cap \Omega,$$
$$\varphi(B_0) = G \cap \Gamma.$$

Theorem 5.4.1. *Let Ω be an open subset of class $\mathbf{C}^1$ which is bounded (or $\Omega = \mathbf{R}^N_+$). Then, there exists an extension operator $\mathbf{P} : W^{1,p}(\Omega) \to W^{1,p}(\mathbf{R}^N)$ which is linear and continuous. More precisely, for all $v \in W^{1,p}(\Omega)$,*

(i) $\mathbf{P}v|_\Omega = v$;

(ii) $\|\mathbf{P}v\|_{L^p(\mathbf{R}^N)} \leq C \|v\|_{L^p(\Omega)}$;

(iii) $\|\mathbf{P}v\|_{W^{1,p}(\mathbf{R}^N)} \leq C \|v\|_{W^{1,p}(\Omega)}$.

The constant C above depends only on Ω and p.

The proof of Theorem 5.4.1 first proves the result in the case where $\Omega = \mathbf{R}^N_+$ is a half-space then passes to the general situation by using partition of unity and local coordinates.

Lemma 5.4.1. *There exists an extension operator* $\mathbf{P} : W^{1,p}(\mathbf{R}^N_+) \to W^{1,p}(\mathbf{R}^N)$ *which is obtained by reflection:*

$$(\mathbf{P}u)(x', x_N) = \begin{cases} u(x', x_N) & \text{if } x_N > 0, \\ u(x', -x_N) & \text{if } x_N < 0. \end{cases}$$

Moreover, $\mathbf{P}$ *is linear and continuous, with*

$$\|\mathbf{P}u\|_{L^p(\mathbf{R}^N)} \le 2\|u\|_{L^p(\mathbf{R}^N_+)};$$
$$\|\mathbf{P}u\|_{W^{1,p}(\mathbf{R}^N)} \le 2\|u\|_{W^{1,p}(\mathbf{R}^N_+)}.$$

Proof. Clearly $\mathbf{P}u \in L^p(\mathbf{R}^N)$ and

$$\left(\int_{\mathbf{R}^N} |\mathbf{P}u(x)|^p\, dx\right)^{1/p} = \left(2\int_{\mathbf{R}^N_+} |u(x)|^p\, dx\right)^{1/p} = 2^{1/p}\|u\|_{L^p} \le 2\|u\|_{L^p}.$$

Let us prove that

$$\frac{\partial}{\partial x_i}(\mathbf{P}u) = \mathbf{P}\left(\frac{\partial u}{\partial x_i}\right) \quad \text{for } i = 1, 2, \ldots, N-1,$$
$$\frac{\partial}{\partial x_N}(\mathbf{P}u) = \mathbf{S}\left(\frac{\partial u}{\partial x_N}\right),$$

where we have set

$$(\mathbf{S}v)(x', x_N) = \begin{cases} v(x', x_N) & \text{if } x_N > 0, \\ -v(x', -x_N) & \text{if } x_N \le 0. \end{cases}$$

We need to introduce a truncation (on the domain) function. Let $\eta \in \mathbf{C}^\infty(\mathbf{R})$ be such that

$$\eta(t) = \begin{cases} 0 & \text{if } t < \frac{1}{2}, \\ 1 & \text{if } t > 1 \end{cases}$$

and define $\eta_k(t) = \eta(kt)$ for $k \in \mathbf{N}^*$. Given $\varphi \in \mathcal{D}(\mathbf{R}^N)$ let us compute $\langle \frac{\partial}{\partial x_i}\mathbf{P}u, \varphi\rangle_{D',D}$:

(a) Take first $1 \le i \le N-1$. By definition

$$\left\langle \frac{\partial}{\partial x_i}(\mathbf{P}u), \varphi \right\rangle_{(\mathcal{D}'(\mathbf{R}^N), \mathcal{D}(\mathbf{R}^N))} = -\int_{\mathbf{R}^N} \mathbf{P}u . \frac{\partial \varphi}{\partial x_i}\, dx. \tag{5.10}$$

By definition of $\mathbf{P}$,

$$\begin{aligned} \int_{\mathbf{R}^N} \mathbf{P}u \frac{\partial \varphi}{\partial x_i}\, dx &= \int_{\mathbf{R}^{N-1}} dx' \int_{\mathbf{R}^+} u(x', x_N) \frac{\partial \varphi}{\partial x_i}(x', x_N)\, dx_N \\ &\quad + \int_{\mathbf{R}^{N-1}} dx' \int_{\mathbf{R}^-} u(x', -x_N) \frac{\partial \varphi}{\partial x_i}(x', x_N)\, dx_N \\ &= \int_{\mathbf{R}^N_+} u(x) \frac{\partial \psi}{\partial x_i}\, dx, \end{aligned} \tag{5.11}$$

where $\psi(x', x_N) = \varphi(x', x_N) + \varphi(x', -x_N)$.

But ψ does not belong to $\mathcal{D}(\mathbf{R}^N_+)$; this is where we need to use a truncation method.

Take as a test function $(\eta_k\psi)(x', x_N) = \eta_k(x_N)\psi(x', x_N)$. Note that $\eta_k\psi$ belongs to $\mathcal{D}(\mathbf{R}^N_+)$ and since $u \in W^{1,p}(\mathbf{R}^N_+)$

$$\int_{\mathbf{R}^N_+} u \frac{\partial}{\partial x_i}(\eta_k\psi)\, dx = -\int_{\mathbf{R}^N_+} \frac{\partial u}{\partial x_i}\eta_k\psi\, dx.$$

Noticing that $\frac{\partial}{\partial x_i}(\eta_k\psi) = \eta_k \frac{\partial \psi}{\partial x_i}$ (since $1 \le i \le N-1$), we obtain

$$\int_{\mathbf{R}^N_+} u\, \eta_k \frac{\partial \psi}{\partial x_i}\, dx = -\int_{\mathbf{R}^N_+} \frac{\partial u}{\partial x_i}\eta_k\psi\, dx.$$

Then, pass to the limit as $k \to +\infty$. By using the Lebesgue dominated convergence theorem,

$$\int_{\mathbf{R}^N_+} u \frac{\partial \psi}{\partial x_i}\, dx = -\int_{\mathbf{R}^N_+} \frac{\partial u}{\partial x_i}\psi\, dx. \tag{5.12}$$

Then, combine (5.10), (5.11), (5.12) to obtain

$$\begin{aligned}\left\langle \frac{\partial}{\partial x_i}(\mathbf{P}u, \varphi \right\rangle &= \int_{\mathbf{R}^N_+} \frac{\partial u}{\partial x_i}\psi\, dx \\ &= \int_{\mathbf{R}^N_+} \frac{\partial u}{\partial x_i}\varphi(x', x_N)\, dx + \int_{\mathbf{R}^N_+} \frac{\partial u}{\partial x_i}\varphi(x', -x_N)\, dx \\ &= \int_{\mathbf{R}^N} \mathbf{P}\left(\frac{\partial u}{\partial x_i}\right)\varphi\, dx.\end{aligned}$$

Hence $\frac{\partial}{\partial x_i}(\mathbf{P}u) = \mathbf{P}(\frac{\partial u}{\partial x_I})$ for $i = 1, 2, \ldots, N-1$.

(b) Take now $i = N$ and compute

$$\begin{aligned}\int_{\mathbf{R}^N} \mathbf{P}u.\frac{\partial \varphi}{\partial x_N}\, dx &= \int_{\mathbf{R}^{N-1}} dx' \int_{\mathbf{R}^+} u(x', x_N)\frac{\partial \varphi}{\partial x_N}(x', x_N)\, dx_N \\ &\quad + \int_{\mathbf{R}^{N-1}} dx' \int_{\mathbf{R}^-} u(x', -x_N)\frac{\partial \varphi}{\partial x_N}(x', x_N)\, dx_N.\end{aligned}$$

Then note that

$$\int_{\mathbf{R}^-} u(x', -x_N)\frac{\partial \varphi}{\partial x_N}(x', x_N)\, dx_N = -\int_{\mathbf{R}^+} u(x', x_N)\frac{\partial}{\partial x_N}(\varphi(x', -x_N))\, dx_N.$$

Let us introduce $\chi(x', x_N) := \varphi(x', x_N) - \varphi(x', -x_N)$. We have

$$\int_{\mathbf{R}^N} \mathbf{P}u.\frac{\partial \varphi}{\partial x_N}\, dx = \int_{\mathbf{R}^N_+} u\frac{\partial \chi}{\partial x_N}\, dx. \tag{5.13}$$

Since $\chi(x', 0) = 0$, there exists some constant $M > 0$ such that

$$|\chi(x', x_N)| \le M|x_N| \quad \text{for } |x_N| \le R, \quad \text{where } \operatorname{spt}\varphi \subset B(0, R).$$

The same argument as before yields, with η_k as a truncation function (note that $\eta_k \chi \in \mathcal{D}(\mathbf{R}^N_+)$),

$$\int_{\mathbf{R}^N_+} u \frac{\partial}{\partial x_N}(\eta_k \chi)\, dx = -\int_{\mathbf{R}^N_+} \frac{\partial u}{\partial x_N} \eta_k \chi\, dx.$$

We have

$$\frac{\partial}{\partial x_N}(\eta_k \chi) = \eta_k \frac{\partial \chi}{\partial x_N} + \eta_k' \chi$$

and

$$\eta_k'(x_N) = k\eta'(k\, x_N).$$

Hence

$$\int_{\mathbf{R}^N_+} \frac{\partial u}{\partial x_N} \eta_k \chi\, dx = -\int_{\mathbf{R}^N_+} u \eta_k \frac{\partial \chi}{\partial x_N}\, dx - \int_{\mathbf{R}^N_+} u\, k\eta'(k\, x_N) \chi(x', x_N)\, dx. \tag{5.14}$$

Let now pass to the limit as $k \to +\infty$. By the Lebesgue dominated convergence theorem,

$$\int_{\mathbf{R}^N_+} \frac{\partial u}{\partial x_N} \eta_k \chi\, dx \to \int_{\mathbf{R}^N_+} \frac{\partial u}{\partial x_N} \chi\, dx,$$
$$\int_{\mathbf{R}^N_+} u \eta_k \frac{\partial \chi}{\partial x_N}\, dx \to \int_{\mathbf{R}^N_+} u \frac{\partial \chi}{\partial x_N}\, dx,$$

and the last integral in (5.14) can be majorized as

$$\left| \int_{\mathbf{R}^N_+} ku\eta'(k\, x_N) \chi(x', x_N)\, dx \right| \le k\, MC \int_{0<x_N<1/k} |u| x_N\, dx$$
$$\le MC \int_{0<x_N<1/k} |u(x)|\, dx,$$

where $C = \sup_{t\in[0,1]} \eta'(t)$. Hence

$$\int_{\mathbf{R}^N_+} \frac{\partial u}{\partial x_N} \chi\, dx = -\int_{\mathbf{R}^N_+} u \frac{\partial \chi}{\partial x_N}\, dx,$$

and, returning to (5.13),

$$\begin{aligned}
\int_{\mathbf{R}^N} \mathbf{P}u . \frac{\partial \varphi}{\partial x_N}\, dx &= -\int_{\mathbf{R}^N_+} \frac{\partial u}{\partial x_N} \chi\, dx \\
&= -\int_{\mathbf{R}^N_+} \frac{\partial u}{\partial x_N} (\varphi(x', x_N) - \varphi(x', -x_N))\, dx \\
&= -\int_{\mathbf{R}^N_+} \frac{\partial u}{\partial x_N} \varphi + \int_{\mathbf{R}^N_-} \left(\frac{\partial u}{\partial x_N} \right)(x', -x_N) \varphi\, dx \\
&= -\int_{\mathbf{R}^N} \mathbf{S}\left(\frac{\partial u}{\partial x_N} \right) \varphi\, dx,
\end{aligned}$$

i.e.,

$$\frac{\partial}{\partial x_N}(\mathbf{P}u) = \mathbf{S}\left(\frac{\partial u}{\partial x_N} \right). \quad \square$$

Let us now consider a general open set. We will use a partition of unity. Since Ω is an open set of class $\mathbf{C}^1$ which is bounded, there exists a finite number of open sets $G_0, G_1, \ldots, G_k$ such that $\bar{\Omega} \subset \bigcup_{i=0}^k G_i$ with $\bar{G}_0 \subset \Omega$ and for each $i = 1, \ldots, k$ a system of local coordinates $\varphi_i : B(0,1) \to G_i$.

Let us introduce a partition of the unity relatively to the open covering $\{G_0, G_1, \ldots, G_k\}$ of the compact set $\bar{\Omega}$: there exists $\{\alpha_0, \alpha_1, \ldots, \alpha_k\}$ with $\alpha_i \in \mathcal{D}(G_i)$, $0 \le i \le k$, and $\sum_{i=0}^k \alpha_i = 1$ on $\bar{\Omega}$. We have now all the elements to prove Theorem 5.4.1.

PROOF OF THEOREM 5.4.1. Given $v \in W^{1,p}(\Omega)$, since $1 = \sum_{i=0}^k \alpha_i$ on Ω we have

$$v = \left(\sum_{i=0}^k \alpha_i \right) v = \sum_{i=0}^k \alpha_i v = \sum_{i=0}^k v_i, \quad \text{where } v_i = \alpha_i v.$$

We now extend each of the functions v_i to the whole of $\mathbf{R}^N$. We have to distinguish the case $i = 0$ from the case $i \ge 1$.

(a) Extension of v_0. The natural extension consists in taking

$$\tilde{v}_0(x) = \begin{cases} v_0(x) & \text{if } x \in \Omega, \\ 0 & \text{if } x \in \mathbf{R}^N \setminus \Omega. \end{cases}$$

One can easily verify that

$$\frac{\partial}{\partial x_i} \tilde{v}_0 = \alpha_0 \frac{\widetilde{\partial v}}{\partial x_i} + \frac{\partial \alpha_0}{\partial x_i} \tilde{v}_0,$$

which belongs to $L^p(\mathbf{R}^N)$. Hence $\tilde{v}_0 \in W^{1,p}(\mathbf{R}^N)$.

(b) Extension of v_i, $1 \le i \le k$. We use the local coordinates φ_i on G_i to define, for $1 \le i \le k$,

$$w_i = \begin{cases} (\alpha_i v) \circ \varphi_i & \text{on } B_+, \\ 0 & \text{on } \mathbf{R}^N_+ \setminus B_+. \end{cases}$$

This function belongs to $W^{1,p}(\mathbf{R}^N_+)$. We can apply the extension by reflection operator $\mathbf{P}$ (cf. Lemma 5.4.1) to obtain

$$\mathbf{P}w_i \in W^{1,p}(\mathbf{R}^N) \quad \text{and} \quad \operatorname{spt} \mathbf{P}w_i \subset B(0,1).$$

We then return to G_i, by considering the function $\mathbf{P}w_i \circ \varphi_i^{-1}$ and define

$$\hat{v}_i = \begin{cases} \mathbf{P}w_i \circ \varphi_i^{-1} & \text{on } G_i, \\ 0 & \text{on } \mathbf{R}^N \setminus G_i. \end{cases}$$

Since $v = \Sigma v_i$, and since we look for a linear extension operator, we set

$$\mathbf{P}v = \sum_{i=0}^k \hat{v}_i.$$

It is easy to verify that Pv coincides with v on Ω and since all the operations involved in this construction are continuous on $W^{1,p}$, so is $\mathbf{P}$. □

As a direct consequence of the extension Theorem 5.4.1, we have the following result, where we denote $\mathcal{D}(\bar{\Omega}) = \{v|_\Omega \ : \ v \in \mathcal{D}(\mathbf{R}^N)\}$.

Proposition 5.4.1. *Let Ω be a bounded open subset of class $\mathbf{C}^1$ in $\mathbf{R}^N$, or $\Omega = \mathbf{R}^N_+$. Then $\mathcal{D}(\bar{\Omega})$ is dense in $W^{1,p}(\Omega)$ $(1 \le p < +\infty)$.*

PROOF. Let $v \in W^{1,p}(\Omega)$. By Theorem 5.4.1, $\mathbf{P}v \in W^{1,p}(\mathbf{R}^N)$. By the density of $\mathcal{D}(\mathbf{R}^N)$ in $W^{1,p}(\mathbf{R}^N)$ (see Theorem 5.1.3), there exists a sequence $(v_n)_{n\in\mathbf{N}}$ in $\mathcal{D}(\mathbf{R}^N)$ with $v_n \underset{n\to+\infty}{\longrightarrow} \mathbf{P}v$ in $W^{1,p}(\mathbf{R}^N)$. Then $v_n|_\Omega \in \mathcal{D}(\bar{\Omega})$ and $v_n|_\Omega \to \mathbf{P}v|_\Omega = v$ in $W^{1,p}(\Omega)$.

When $\Omega = \mathbf{R}^N_+$ the conclusion can be achieved by a truncation argument (see [90, Corollary IX.8]). □

Another important consequence of Theorem 5.4.1 is the Rellich–Kondrakov compact embedding theorem in $W^{1,p}(\Omega)$.

Theorem 5.4.2. *Let Ω be a bounded open set in $\mathbf{R}^N$ of class $\mathbf{C}^1$. For $1 \le p \le +\infty$, the canonical embedding $W^{1,p}(\Omega) \to L^p(\Omega)$ is compact.*

PROOF. Let $(v_n)_{n\in\mathbf{N}}$ be a sequence which is bounded in $W^{1,p}(\Omega)$. By Theorem 5.4.1, the sequence $(\mathbf{P}(v_n))$ is bounded in $W^{1,p}(\mathbf{R}^N)$. Since Ω is bounded, there exists some $R > 0$ such that $\Omega \subset B(0, R)$. Take $\alpha \in \mathcal{D}(\mathbf{R}^N)$ such that $\alpha = 1$ on $B(0, R)$ and $\alpha = 0$ on $\mathbf{R}^N \setminus B(0, 2R)$. The sequence $(\alpha\, \mathbf{P}(v_n))_{n\in\mathbf{N}}$ is bounded in $W^{1,p}(\mathbf{R}^N)$ and is identically equal to zero on $\mathbf{R}^N \setminus B(0, 2R)$. It follows from the Rellich–Kondrakov theorem in $W^{1,p}(\mathbf{R}^N)$ (cf. Corollary 5.3.1) that the sequence $(\alpha\mathbf{P}(v_n))_{n\in\mathbf{N}}$ is relatively compact in $L^p(\mathbf{R}^N)$. Let $\alpha\mathbf{P}(v_{n_k}) \to v$ in $L^p(\mathbf{R}^N)$. Then

$$\alpha\mathbf{P}(v_{n_k})|_\Omega \to v|_\Omega \text{ in } L^p(\Omega).$$

Since $\alpha\mathbf{P}(v_{n_k})|_\Omega = v_{n_k}$, we have $v_{n_k} \to v|_\Omega$ in $L^p(\Omega)$. □

As an application of the Rellich–Kondrakov theorem in $W^{1,p}(\Omega)$ we have the following result.

Theorem 5.4.3. *Let $\Omega \subset \mathbf{R}^N$ be an open connected set which is bounded and of class $\mathbf{C}^1$. Let $V \subset W^{1,p}(\Omega)$ be a linear subspace of $W^{1,p}(\Omega)$ which is closed and such that the only constant function belonging to V is the function which is identically zero. Then, there exists a constant $C > 0$ such that*

$$\|v\|_{L^p(\Omega)} \le C \left(\int_\Omega \sum_{i=1}^N \left| \frac{\partial v}{\partial x_i} \right|^p dx \right)^{1/p} \qquad \forall\, v \in V.$$

PROOF. Let us argue by contradiction. There exists a sequence $(v_n)_{n\in\mathbf{N}}$ in V, $v_n \not\equiv 0$, such that

$$\|v_n\|_{L^p} > n\, \|Dv_n\|_{L^p},$$

where we briefly denote $\|Dv\|_{L^p} = (\Sigma |\frac{\partial v}{\partial x_i}|^p)^{1/p}$.

Take $u_n := v_n/\|v_n\|_{L^p}$. We have $u_n \in V$ (since V is a linear subspace), $\|u_n\|_{L^p} = 1$, and $\|Du_n\|_{L^p} < 1/n$, that is, $\lim_{n\to+\infty} \|Du_n\|_{L^p} = 0$. The sequence $(u_n)_{n\in\mathbf{N}}$ is bounded in $W^{1,p}(\Omega)$. By the Rellich–Kondrakov theorem, it is relatively compact in $L^p(\Omega)$. So, we can extract a subsequence $(u_{n_k})_{k\in\mathbf{N}}$ such that

$$u_{n_k} \to u \text{ in } L^p(\Omega).$$

Since

$$Du_{n_k} \to 0 \text{ in } L^p(\Omega)$$

we conclude that $Du = 0$. Since Ω is connected we have $u \equiv C$ and

$$u_{n_k} \to C \text{ in } W^{1,p}(\Omega).$$

Since V is closed in $W^{1,p}(\Omega)$, C belongs to V, which by assumption forces C to be equal to zero. On the other hand, since $\|u_{n_k}\|_{L^p} = 1$ and $u_{n_k} \to C$ in L^p we have $|C||\Omega|^{1/p} = 1$, a contradiction with the fact that $C = 0$. □

Corollary 5.4.1 (Poincaré–Wirtinger inequality). *Let Ω satisfy the assumptions of Theorem* 5.4.3. *Then, there exists a constant $C_p > 0$ such that*

$$\forall v \in W^{1,p}(\Omega) \quad \left\| v - \frac{1}{|\Omega|}\int_\Omega v(x)\,dx \right\|_{L^p(\Omega)} \le C_p \|Dv\|_{L^p(\Omega)}.$$

Proof. Take $V = \{v \in W^{1,p}(\Omega) \ : \ \int_\Omega v(x)\,dx = 0\}$. Clearly, V is a closed linear subspace of $W^{1,p}(\Omega)$ and the only constant function which is in V is the function which vanishes identically. Then the proof is concluded by noticing that $v - \frac{1}{|\Omega|}\int_\Omega v\,dx$ belongs to V for every $v \in W^{1,p}(\Omega)$. □

5.5 The Fourier approach to Sobolev spaces. The space $H^s(\Omega)$, $s \in \mathbf{R}$

We give here a description of the space $H^1(\mathbf{R}^N)$ by using the Fourier transform. We recall that for any $v \in L^1(\mathbf{R}^N)$, its Fourier transform $\hat{v}$ is defined by

$$\hat{v}(\xi) = \frac{1}{(2\pi)^{N/2}} \int_{\mathbf{R}^N} e^{-i\,\xi\cdot x} v(x)\,dx. \tag{5.15}$$

It can be shown that $\hat{v}$ belongs to $\mathbf{C}_0(\mathbf{R}^N)$. When $\hat{v} \in L^1(\mathbf{R}^N)$, one can invert the Fourier transform and obtain v from $\hat{v}$ by using the following formula:

$$v(x) = \frac{1}{(2\pi)^{N/2}} \int_{\mathbf{R}^N} e^{i\xi\cdot x} \hat{v}(\xi)\,d\xi. \tag{5.16}$$

Indeed, the condition $\hat{v} \in L^1(\mathbf{R}^N)$ is not always easy to handle. One often prefers to work with the Fourier–Plancherel transform, which is defined as follows.

The basic property is that if $v \in L^1(\mathbf{R}^N) \cap L^2(\mathbf{R}^N)$, then $\hat{v} \in L^2(\mathbf{R}^N)$ and, in fact,

$$\|\hat{v}\|_{L^2(\mathbf{R}^N)} = \|v\|_{L^2(\mathbf{R}^N)}. \tag{5.17}$$

Note that, since the Lebesgue measure of $\mathbf{R}^N$ is infinite, $L^2(\mathbf{R}^N)$ is not a subspace of $L^1(\mathbf{R}^N)$ and one cannot define directly $\hat{v}$ for $v \in L^2(\mathbf{R}^N)$ by using formula (5.15). However, if $v \in L^1 \cap L^2$, formula (5.15) makes sense and (5.17) tells us that $v \to \hat{v}$ is an isometry for the L^2 norm.

By using the density of $L^1 \cap L^2$ into L^2 (note, for example, that $L^1 \cap L^2$ contains the continuous functions with compact support), one can extend the Fourier transform to L^2. The so-obtained mapping is called the Fourier–Plancherel transform and we denote it by $\mathcal{F}$. The basic properties of $\mathcal{F}$ are summarized in the following proposition.

Proposition 5.5.1. *One can associate to each function $v \in L^2(\mathbf{R}^N)$ a function $\mathcal{F}v \in L^2(\mathbf{R}^N)$, called the Fourier–Plancherel transform of v, so that the following properties hold:*

(i) $\forall v \in L^1(\mathbf{R}^N) \cap L^2(\mathbf{R}^N),\ \mathcal{F}v = \hat{v}$.

(ii) $\forall v \in L^2(\mathbf{R}^N),\ \|\mathcal{F}v\|_{L^2(\mathbf{R}^N)} = \|v\|_{L^2(\mathbf{R}^N)}$.

(iii) *The mapping $\mathcal{F}$ is an isometric isomorphism from $L^2(\mathbf{R}^N)$ onto $L^2(\mathbf{R}^N)$.*

(iv) *For $v \in L^2(\mathbf{R}^N)$, setting for each $n \in \mathbf{N}$*

$$w_n(\xi) = \frac{1}{(2\pi)^{N/2}} \int_{|x| \le n} e^{-i\xi \cdot x} v(x)\, dx,$$

we have $w_n \to \mathcal{F}v$ in $L^2(\mathbf{R}^N)$ as $n \to +\infty$.

(v) *Conversely, for $v \in L^2(\mathbf{R}^N)$, setting for each $n \in \mathbf{N}$*

$$v_n(x) = \frac{1}{(2\pi)^{N/2}} \int_{\{|\xi| \le n\}} e^{i\xi \cdot x} \mathcal{F}(v)(\xi)\, d\xi,$$

we have $v_n \to v$ in $L^2(\mathbf{R}^N)$ as $n \to +\infty$.

We can now give the following characterization of the Sobolev space $H^1(\mathbf{R}^N)$.

Theorem 5.5.1. *We have*

$$H^1(\mathbf{R}^N) = \{v \in L^2(\mathbf{R}^N) : (1 + |\xi|^2)^{1/2} \mathcal{F}(v) \in L^2(\mathbf{R}^N)\}$$

and, for any $v \in H^1(\mathbf{R}^N)$,

$$\|v\|_{H^1(\mathbf{R}^N)} = \|(1 + |\xi|^2)^{1/2} \mathcal{F}(v)\|_{L^2(\mathbf{R}^N)}.$$

Proof. Take first $v \in \mathcal{D}(\mathbf{R}^N)$. By definition,

$$\widehat{\frac{\partial v}{\partial x_k}}(\xi) = \frac{1}{(2\pi)^{N/2}} \int_{\mathbf{R}^N} e^{-i\,\xi \cdot x} \frac{\partial v}{\partial x_k}(x)\, dx.$$

Let us integrate by parts the above formula. Since v has compact support, we obtain

$$\widehat{\frac{\partial v}{\partial x_k}}(\xi) = (i\,\xi_k)\hat{v}(\xi). \tag{5.18}$$

We now use (see Theorem 5.1.3) the density of $\mathcal{D}(\mathbf{R}^N)$ in $H^1(\mathbf{R}^N)$. For any $v \in H^1(\mathbf{R}^N)$ there exists a sequence $(v_n)_{n\in\mathbf{N}}$ in $\mathcal{D}(\mathbf{R}^N)$ such that $v_n \to v$ with respect to the norm topology of $H^1(\mathbf{R}^N)$. So for each $n \in \mathbf{N}$,

$$\widehat{\frac{\partial v_n}{\partial x_k}}(\xi) = (i\,\xi_k)\hat{v}_n(\xi). \tag{5.19}$$

By definition of the Fourier–Plancherel transform (which coincides with the classical Fourier transform for functions in $\mathcal{D}(\mathbf{R}^N)$) we have

$$\mathcal{F}\left(\frac{\partial v_n}{\partial x_k}\right)(\xi) = (i\,\xi_k)\mathcal{F}(v_n)(\xi). \tag{5.20}$$

Since $v_n \to v$ for the $\|\cdot\|_{H^1(\mathbf{R}^N)}$ norm, we have that $v_n \to v$ in $L^2(\mathbf{R}^N)$ and $\frac{\partial v_n}{\partial x_k} \longrightarrow \frac{\partial v}{\partial x_k}$ in $L^2(\mathbf{R}^N)$. We now use the continuity property of $\mathcal{F}$ for the $L^2(\mathbf{R}^N)$ norm and pass to the limit in (5.20): one can extract a subsequence $n(p)$ such that

$$\begin{aligned}
&\mathcal{F}(v_{n(p)})(\xi) \longrightarrow \mathcal{F}(v)(\xi) && \text{for almost every } \xi \in \mathbf{R}^N,\\
&\mathcal{F}\left(\frac{\partial v_{n(p)}}{\partial x_k}\right)(\xi) \longrightarrow \mathcal{F}\left(\tfrac{\partial v}{\partial x_k}\right)(\xi) && \text{for almost every } \xi \in \mathbf{R}^N
\end{aligned}$$

and so obtain

$$\mathcal{F}\left(\frac{\partial v}{\partial x_k}\right)(\xi) = i\,\xi_k\,\mathcal{F}(v)(\xi) \quad \text{for almost every } \xi \in \mathbf{R}^N. \tag{5.21}$$

It follows from the above argument and the isometry property of $\mathcal{F}$ in $L^2(\mathbf{R}^N)$ that

$$\begin{aligned}
v \in H^1(\mathbf{R}^N) &\Longleftrightarrow \mathcal{F}(v) \in L^2(\mathbf{R}^N) \text{ and } i\,\xi_k\,\mathcal{F}(v) \in L^2(\mathbf{R}^N)\\
&\qquad \forall\, k = 1, 2, \ldots, N,\\
&\Longleftrightarrow (1+|\xi|^2)^{1/2}\mathcal{F}(v) \in L^2(\mathbf{R}^N).
\end{aligned}$$

Moreover,

$$\begin{aligned}
\|v\|^2_{H^1(\mathbf{R}^N)} &= \int_{\mathbf{R}^N} v^2(x) + \sum_{k=1}^N \left(\frac{\partial v}{\partial x_k}\right)^2 (x)\,dx\\
&= \int_{\mathbf{R}^N} \left(|\mathcal{F}(v)(\xi)|^2 + \sum_{k=1}^N \left|\mathcal{F}\left(\frac{\partial v}{\partial x_k}\right)(\xi)\right|^2\right) d\xi\\
&= \int_{\mathbf{R}^N} (1+|\xi|^2)\,|\mathcal{F}(v)(\xi)|^2\,d\xi\\
&= \|(1+|\xi|^2)^{1/2}\mathcal{F}(v)\|^2_{L^2(\mathbf{R}^N)}
\end{aligned}$$

and the proof is complete. □

The above characterization of the space $H^1(\mathbf{R}^N)$ via the Fourier–Plancherel transform is quite useful. It permits us to obtain in an elegant way a number of important properties of Sobolev spaces. Let us first show how to obtain by this way the Rellich–Kondrakov theorem.

Theorem 5.5.2 (Rellich–Kondrakov, second proof). *Let Ω be a bounded open subset of $\mathbf{R}^N$ with a boundary $\partial\Omega$ of class $\mathbf{C}^1$. Then, the embedding of $H^1(\Omega)$ into $L^2(\Omega)$ is compact.*

PROOF. Let $(v_n)_{n\in\mathbf{N}}$ be a sequence in $H^1(\Omega)$ which is bounded for the $\|\cdot\|_{H^1(\Omega)}$ norm, i.e., $\sup_{n\in\mathbf{N}} \|v_n\|_{H^1(\Omega)} < +\infty$. By using the extension operator $\mathbf{P}$ defined in Theorem 5.4.1, which is linear and continuous from $H^1(\Omega)$ into $H^1(\mathbf{R}^N)$, and by using a truncation argument (Ω is assumed to be bounded), we can assume that the sequence v_n is bounded in $H^1(\mathbf{R}^N)$ and $v_n \equiv 0$ outside a ball $B(0, R)$ (with R independent of $n \in \mathbf{N}$).

Since $(v_n)_{n\in\mathbf{N}}$ is bounded in $L^2(\mathbf{R}^N)$, we can extract a weakly convergent subsequence

$$v_{n(k)} \rightharpoonup v \text{ in } w - L^2(\mathbf{R}^N).$$

We are going to prove that, because of the uniform bound on the $H^1(\mathbf{R}^N)$ norm of the $v_{n(k)}$, the convergence is actually strong in $L^2(\mathbf{R}^N)$. Without loss of generality we can assume $v = 0$ (replacing $v_{n(k)}$ by $v_{n(k)} - v$!).

Let us simplify the notations and write v_k instead of $v_{n(k)}$. We have

$$v_k \to 0 \text{ in } w - L^2(\mathbf{R}^N), \tag{5.22}$$

$$v_k \equiv 0 \text{ outside of } B(0, R), \tag{5.23}$$

$$\sup_{k\in\mathbf{N}} \|v_k\|_{H^1(\mathbf{R}^N)} < +\infty, \tag{5.24}$$

and we want to prove that $v_k \to 0$ in $s - L^2(\mathbf{R}^N)$. To that end, we use the Fourier–Plancherel transformation and its isometrical property from $L^2(\mathbf{R}^N)$ onto $L^2(\mathbf{R}^N)$. We need to prove that

$$\|\mathcal{F}(v_k)\|_{L^2(\mathbf{R}^N)} \longrightarrow 0 \text{ as } k \to +\infty,$$

and we know (see Theorem 5.5.1) that

$$\sup_{k\in\mathbf{N}} \|(1 + |\xi|^2)^{1/2}\mathcal{F}(v_k)\|_{L^2(\mathbf{R}^N)} := C < +\infty.$$

Let us write

$$\begin{aligned}
\int_{\mathbf{R}^N} |\mathcal{F}(v_k)(\xi)|^2\, d\xi &= \int_{|\xi|\le M} |\mathcal{F}(v_k)(\xi)|^2\, d\xi + \int_{|\xi|\ge M} |\mathcal{F}(v_k)(\xi)|^2\, d\xi \\
&\le \int_{|\xi|\le M} |\mathcal{F}(v_k)(\xi)|^2\, d\xi \\
&\qquad + \frac{1}{1+M^2} \int_{\mathbf{R}^N} (1 + |\xi|^2)|\mathcal{F}(v_k)(\xi)|^2\, d\xi \\
&\le \int_{|\xi|\le M} |\mathcal{F}(v_k)(\xi)|^2\, d\xi + \frac{C^2}{1+M^2}.
\end{aligned} \tag{5.25}$$

On the other hand, since $v_k \equiv 0$ outside of $B(0, R)$, we have $v_k \in L^2(\mathbf{R}^N) \cap L^1(\mathbf{R}^N)$ and $\mathcal{F}(v_k) = \hat{v}_k$, i.e.,

$$\mathcal{F}(v_k)(\xi) = \frac{1}{(2\pi)^{N/2}} \int_{\mathbf{R}^N} e^{-i\,\xi\cdot x} v_k(x)\,dx$$
$$= \frac{1}{(2\pi)^{N/2}} \int_{\mathbf{R}^n} \mathbf{1}_{B(0,R)}(x) e^{-i\,\xi\cdot x} v_k(x)\,dx.$$

For any $\xi \in \mathbf{R}^N$, the function $x \mapsto \mathbf{1}_{B(0,R)}(x)e^{-i\,\xi\cdot x}$ belongs to $L^2(\mathbf{R}^N)$.

By using (5.22), we obtain

$$\forall \xi \in \mathbf{R}^N \quad \mathcal{F}(v_k)(\xi) \to 0 \text{ as } k \to +\infty. \tag{5.26}$$

To apply the Lebesgue dominated convergence theorem, we notice that by the Cauchy–Schwarz inequality,

$$|\mathcal{F}(v_k)(\xi)|^2 \le \frac{1}{(2\pi)^N} \|v_k\|^2_{L^2(\mathbf{R}^N)}\, |B(0, R)|, \tag{5.27}$$

where $|B(0, R)|$ is the Lebesgue measure of the ball $B(0, R)$.

By using (5.26) and (5.27), we obtain

$$\int_{|\xi|\le M} |\mathcal{F}(v_k)(\xi)|^2\,d\xi \to 0 \text{ as } k \to +\infty. \tag{5.28}$$

Returning to (5.25), by using (5.28), we obtain

$$\limsup_{k\to+\infty} \int_{\mathbf{R}^N} |\mathcal{F}(v_k)(\xi)|^2\,d\xi \le \frac{C^2}{1+M^2}.$$

This being true for arbitrarily large M, by letting $M \to +\infty$, we finally obtain

$$\lim_{k\to+\infty} \|\mathcal{F}(v_k)\|_{L^2(\mathbf{R}^N)} = 0,$$

which completes the proof. □

Clearly, the characterization of the space $H^1(\mathbf{R}^N)$ by means of the Fourier–Plancherel transform can be easily extended to higher-order Sobolev spaces $H^m(\Omega)$, $m \in \mathbf{N}$.

Theorem 5.5.3. *For any $m \in \mathbf{N}$*

$$H^m(\mathbf{R}^N) = \{v \in L^2(\mathbf{R}^N) : (1 + |\xi|^2)^{m/2}\mathcal{F}(v) \in L^2(\mathbf{R}^N)\}$$

and, for any $v \in H^m(\mathbf{R}^N)$,

$$\|v\|_{H^m(\mathbf{R}^n)} = \|(1 + |\xi|^2)^{m/2}\mathcal{F}(v)\|_{L^2(\mathbf{R}^N)}.$$

A major interest of the above approach is that it suggests a natural definition of the space $H^s(\mathbf{R}^N)$ for $s \in \mathbf{R}$, s being not necessarily an integer (s being possibly a fraction or

more generally any real number), and s being not necessarily positive. The central point is that the property for the integral

$$\int_{\mathbf{R}^N} (1+|\xi|^2)^m \, |\mathcal{F}(v)(\xi)|^2 \, d\xi$$

to be finite makes sense for any exponent $m \in \mathbf{R}$, since the quantity a^x makes sense for any $x \in \mathbf{R}$ as soon as $a > 0$. (Here $a = 1 + |\xi|^2$, which is clearly positive.)

We are led to the following definition.

Definition 5.5.1. *Let $s \geq 0$ be a nonnegative real number. Let us define*

$$H^s(\mathbf{R}^N) = \{v \in L^2(\mathbf{R}^N) : (1+|\xi|^2)^{s/2}\mathcal{F}(v) \in L^2(\mathbf{R}^N)\},$$

which is equipped with the scalar product, for any $u, v \in H^s(\mathbf{R}^N)$

$$\langle u, v\rangle_{H^s(\mathbf{R}^N)} := \int_{\mathbf{R}^N} (1+|\xi|^2)^s \mathcal{F}(u)(\xi)\overline{\mathcal{F}(v)}(\xi)\, d\xi,$$

and the corresponding norm,

$$\|v\|_{H^s(\mathbf{R}^N)} := \left(\int_{\mathbf{R}^N} (1+|\xi|^2)^s |\mathcal{F}(v)(\xi)|^2 \, d\xi\right)^{1/2}.$$

Let us summarize in the following statement some of the basic properties of the Sobolev spaces $H^s(\mathbf{R}^N)$.

Proposition 5.5.2. *For any $s \in \mathbf{R}^+$, $H^s(\mathbf{R}^N)$ is a Hilbert space. When $s = m \in \mathbf{N}$ we have that $H^s(\mathbf{R}^N) = H^m(\mathbf{R}^N) = W^{m,2}(\Omega)$ is the classical Sobolev space.*

Proof. Clearly $\mathcal{F}$ is an isomorphism from $H^s(\mathbf{R}^N)$ onto the Lebesgue space $L^2(a\,dm)$ where $a\,dm$ is the measure with density $a(\xi) = (1+|\xi|^2)^s$ with respect to the Lebesgue measure dm on $\mathbf{R}^N$. Moreover, $\mathcal{F}$ is an isometry and the Hilbert structure of the weighted Lebesgue space $L^2(a\,dm)$ is transfered to $H^s(\mathbf{R}^N)$ by the Fourier–Plancherel mapping $\mathcal{F}$. So

$$H^s(\mathbf{R}^N) \cong L^2((1+|\xi|^2)^s dm) \quad \text{(isomorphism)}$$

and the proof is complete. □

In accordance with the definition of $H^{-1}(\Omega)$ as the topological dual space of $H_0^1(\Omega)$, let us give the following definition of the Sobolev space $H^{-s}(\mathbf{R}^N)$ when $s \geq 0$.

Definition 5.5.2. *Let s be a nonnegative real number. By definition*

$$H^{-s}(\mathbf{R}^N) = H^s(\mathbf{R}^N)^*$$

is the topological dual of $H^s(\mathbf{R}^N)$.

5.6 Trace theory for $W^{1,p}(\Omega)$ spaces

To study the Dirichlet problem by variational techniques one needs to solve the following problem: "For an arbitrary $v \in H^1(\Omega)$, is it possible to give a meaning to the boundary condition $v = g$ on $\partial\Omega$?"

Clearly, considering v only as an element of $L^2(\Omega)$ is not sufficient information to talk about v on $\partial\Omega$, because the Lebesgue measure of $\partial\Omega$ is zero (for Ω smooth enough). Therefore, one has to rely on the additional information on v, namely, "$\frac{\partial v}{\partial x_i}$ belongs to $L^2(\Omega)$ for any $i = 1, \ldots, N$," to give meaning to v on $\partial\Omega$.

In this section and the next two, we give different answers to this question by using very different techniques. Note that except when $N = 1$, the space $H^1(\Omega)$ is not embedded in the space of continuous functions $\mathbf{C}(\bar{\Omega})$. In this section, we use the geometrical properties of $\partial\Omega$ (the fact that it is locally an $(N-1)$-dimensional manifold) to prove that for a general $v \in W^{1,p}(\Omega)$, $1 \leq p \leq +\infty$, it is possible to give meaning to v on $\partial\Omega$, that is, the trace of v on $\partial\Omega$. Note that for a regular function v, this notion has to reduce to the restriction of v on $\partial\Omega$. This naturally suggests defining the notion of trace by using a density and extension by continuity argument.

Theorem 5.6.1. *Let Ω be a bounded open set whose boundary $\partial\Omega$ is of class $\mathbf{C}^1$. Then, for any $1 \leq p < +\infty$, $\mathcal{D}(\bar{\Omega})$ is dense in $W^{1,p}(\Omega)$ and the restriction mapping*

$$\begin{aligned} \gamma_0 : \quad & \mathcal{D}(\bar{\Omega}) \longrightarrow L^p(\partial\Omega), \\ & v \longmapsto \gamma_0(v) = v|_{\partial\Omega}, \end{aligned}$$

which to each element $v \in \mathcal{D}(\bar{\Omega})$ associates its restriction to $\partial\Omega$, can be extended by continuity into a linear continuous mapping from $W^{1,p}(\Omega)$ into $L^p(\partial\Omega)$, which we still denote by γ_0. (Without ambiguity, one can still use the simpler notation $v|_{\partial\Omega}$.)

The so-defined mapping

$$\gamma_0 : W^{1,p}(\Omega) \longrightarrow L^p(\partial\Omega)$$

is called the trace operator of order zero.

Proof. By Proposition 5.4.1, and the regularity property of the boundary $\partial\Omega$, we know that $\mathcal{D}(\bar{\Omega})$ is dense in $W^{1,p}(\Omega)$. For any $v \in \mathcal{D}(\bar{\Omega})$, without any ambiguity, we can define the restriction of v to $\partial\Omega$, setting $\gamma_0(v) := v|_{\partial\Omega}$.

Assume for a moment that we have been able to prove that

$$\gamma_0 : (\mathcal{D}(\bar{\Omega}), \|\cdot\|_{W^{1,p}(\Omega)}) \longrightarrow (L^p(\partial\Omega), \|\cdot\|_{L^p(\partial\Omega)})$$

is continuous. Since γ_0 is linear, it is uniformly continuous. The space $L^p(\partial\Omega)$ is a Banach space (it is complete). Therefore, all the conditions of the extension by continuity theorem are fulfilled, which provides the existence of $\hat{\gamma}_0$, the unique linear and continuous extension of γ_0

$$\hat{\gamma}_0 : W^{1,p}(\Omega) \longrightarrow L^p(\partial\Omega).$$

For simplicity of notation, we still denote by γ_0 the so-defined operator, which is the trace operator.

Thus, we just need to prove that γ_0 is continuous. To do so, we first consider the case of a half-space $\Omega = \mathbf{R}^N_+$ and then use local coordinates.

Lemma 5.6.1. *Take $\Omega = \mathbf{R}^N_+ = \{x = (x', x_N) \in \mathbf{R}^{N-1} \times \mathbf{R} : x_N > 0\}$. Then, for any $1 \le p < +\infty$, the following inequality holds:*

$$\forall v \in \mathcal{D}(\bar{\mathbf{R}}^N_+) \quad \|\gamma_0(v)\|_{L^p(\mathbf{R}^{N-1})} \le p^{1/p} \|v\|_{W^{1,p}(\mathbf{R}^N_+)}.$$

Proof. Let $v \in \mathcal{D}(\bar{\mathbf{R}}^N_+)$. For any $x' \in \mathbf{R}^{N-1}$ we have

$$\begin{aligned} |v(x', 0)|^p &= -\int_0^{+\infty} \frac{\partial}{\partial x_N} |v(x', x_N)|^p dx_N \\ &\le p \int_0^{+\infty} |v(x', x_N)|^{p-1} \left| \frac{\partial v}{\partial x_N}(x', x_N) \right| dx_N. \end{aligned}$$

Let us apply the Young convexity inequality

$$ab \le \frac{1}{p} a^p + \frac{1}{p'} b^{p'}$$

with $\frac{1}{p} + \frac{1}{p'} = 1$ to the following situation:

$$a = \left| \frac{\partial v}{\partial x_N}(x', x_N) \right| \quad \text{and} \quad b = |v(x', x_N)|^{p-1}.$$

We obtain

$$|v(x', 0)|^p \le p \left[\int_0^{+\infty} \left(\frac{1}{p} \left| \frac{\partial v}{\partial x_N}(x', x_N) \right|^p + \frac{1}{p'} |v(x', x_N)|^{(p-1)p'} \right) dx_N \right].$$

By using the relation $(p-1)p' = p$ we obtain

$$|v(x', 0)|^p \le (p-1) \int_0^{+\infty} |v(x', x_N)|^p \, dx_N + \int_0^{+\infty} \left| \frac{\partial v}{\partial x_N}(x', x_N) \right|^p dx_N.$$

Integrating over $\mathbf{R}^{N-1}$ yields

$$\begin{aligned} \int_{\mathbf{R}^{N-1}} |v(x', 0)|^p \, dx' &\le (p-1) \int_{\mathbf{R}^N_+} |v(x)|^p \, dx + \int_{\mathbf{R}^N_+} \left| \frac{\partial v}{\partial x_N}(x) \right|^p dx \\ &\le (p-1) \int_{\mathbf{R}^N_+} |v(x)|^p dx + \int_{\mathbf{R}^N_+} \sum_{i=1}^N \left| \frac{\partial v}{\partial x_i} \right|^p dx \\ &\le p \int_{\mathbf{R}^N_+} \left(|v(x)|^p + \sum_{i=1}^N \left| \frac{\partial v}{\partial x_i} \right|^p \right) dx. \end{aligned}$$

Hence

$$\|v(\cdot, 0)\|_{L^p(\mathbf{R}^{N-1})} \le p^{1/p} \|v\|_{W^{1,p}(\mathbf{R}^N_+)}$$

and the proof is complete. □

PROOF OF THEOREM 5.6.1 (CONTINUATION). We use a system of local coordinates. With the same notation as in Section 5.4, we set

$$\bar{\Omega} \subset \bigcup_{i=0}^{k} G_i \qquad \text{with } \bar{G}_0 \subset \Omega,\ G_i \text{ open } \forall i = 0, \ldots, k,$$

while $\varphi_i : B(0,1) \to G_i$, $i = 1, 2, \ldots, k$, are the local coordinates $\{\alpha_0, \ldots, \alpha_i\}$, is an associated partition of unity, i.e., $\alpha_i \in \mathcal{D}(G_i)$, $\alpha_i \geq 0$, $\sum_{i=0}^{k} \alpha_i = 1$ on $\bar{\Omega}$.

Take $v \in \mathcal{D}(\bar{\Omega})$. For any $1 \leq i \leq k$, let us define

$$w_i = \begin{cases} (\alpha_i v) \circ \varphi_i & \text{on } B_+, \\ 0 & \text{on } \mathbf{R}^N_+ \setminus B_+. \end{cases}$$

Clearly w_i belongs to $\mathcal{D}(\mathbf{R}^N_+)$. By Lemma 5.6.1, we have

$$\|w_i(\cdot, 0)\|_{L^p(\mathbf{R}^{N-1})} \leq p^{1/p} \|w_i\|_{W^{1,p}(\mathbf{R}^N_+)}. \tag{5.29}$$

By using classical differential calculus rules (note that all the functions α_i, v, φ_i are continuously differentiable), one obtains the existence, for any $i = 1, \ldots, k$, of a constant C_i such that

$$\|w_i\|_{W^{1,p}(\mathbf{R}^N_+)} \leq C_i \|v\|_{W^{1,p}(\Omega)}. \tag{5.30}$$

Combining the two inequalities (5.29) and (5.30), we obtain

$$\|w_i(\cdot, 0)\|_{L^p(\mathbf{R}^{N-1})} \leq C_i\, p^{1/p} \|v\|_{W^{1,p}(\Omega)}. \tag{5.31}$$

We now use the definition of the $L^p(\partial\Omega)$ norm which is based on the use of local coordinates. One can show that an equivalent norm to the $L^p(\partial\Omega)$ norm can be obtained by using local coordinates: denoting by $\sim$ the extension by zero outside of $\mathbf{R}^{N-1} \setminus \{y \in \mathbf{R}^{N-1} : |y| < 1\}$, we have that

$$L^p(\partial\Omega) = \{v : \partial\Omega \to \mathbf{R} \ :\ \widetilde{(\alpha_i v)} \circ \varphi_i(\cdot, 0) \in L^p(\mathbf{R}^{N-1}),\ 1 \leq i \leq k\}$$

and

$$v \longmapsto \left(\sum_{i=1}^{k} \|\widetilde{(\alpha_i v)} \circ \varphi_i\|^p_{L^p(\mathbf{R}^{N-1})} \right)^{1/p} \tag{5.32}$$

is an equivalent norm to the $L^p(\partial\Omega)$ norm.

This definition of the $L^p(\partial\Omega)$ norm and the inequality (5.31) (note that $w_i = \widetilde{(\alpha_i v)} \circ \varphi_i$) yield

$$\|v\|_{L^p(\partial\Omega)} \leq C(p, N, \Omega) \|v\|_{W^{1,p}(\Omega)}$$

for some constant $C(p, N, \Omega)$. Thus, γ_0 is continuous, which ends the proof of Theorem 5.6.1. □

Let us now give some of the most important properties of the trace operator γ_0.

Proposition 5.6.1. *Let us assume that Ω is an open bounded subset of $\mathbf{R}^N$ whose boundary $\partial\Omega$ is $\mathbf{C}^1$. Then, for any $1 \leq p < \infty$, $W_0^{1,p}(\Omega)$ is equal to the kernel of γ_0, i.e.,*

$$W_0^{1,p}(\Omega) = \{v \in W^{1,p}(\Omega) : \gamma_0(v) = 0\}.$$

PROOF. We first show the inclusion $W_0^{1,p}(\Omega) \subset \ker \gamma_0$.

Take $v \in W_0^{1,p}(\Omega)$. By definition of $W_0^{1,p}(\Omega)$, there exists a sequence of functions $(v_n)_{n\in\mathbf{N}}$, $v_n \in \mathcal{D}(\Omega)$ such that $v_n \to v$ in $W^{1,p}(\Omega)$. Since $\gamma_0(v_n) = v_n|_{\partial\Omega} = 0$, by continuity of γ_0 we obtain that $\gamma_0(v) = 0$, i.e., $v \in \ker \gamma_0$.

The other inclusion is a bit more involved. We just sketch the main lines of its proof. Using local coordinates, we prove the following result.

Take $v \in W^{1,p}(\mathbf{R}^N_+)$ such that $\gamma_0(v) = 0$. Prove that $v \in W_0^{1,p}(\mathbf{R}^N_+)$. Let us first extend v by zero outside of $\mathbf{R}^N_+$. By using the information $\gamma_0(v) = 0$ one can verify that the so-obtained extension $\tilde{v}$ belongs to $W^{1,p}(\mathbf{R}^N)$. Then, let us translate $\tilde{v}$, and consider for any $h > 0$

$$\tau_h\tilde{v}(x', x_N) = \tilde{v}(x', x_N - h).$$

Finally, one regularizes by convolution the function $\tau_h\tilde{v}$. We have that for ε sufficiently small, $\rho_\varepsilon \star (\tau_h\tilde{v})$ belongs to $\mathcal{D}(\mathbf{R}^N_+)$ and $\rho_\varepsilon \star (\tau_h\tilde{v})$ tends to v in $W^{1,p}(\mathbf{R}^N_+)$ as $h \to 0$ and $\varepsilon \to 0$. Hence $v \in W_0^{1,p}(\mathbf{R}^N_+)$. □

Proposition 5.6.2 (Green's formula). *Let Ω be an open bounded set in $\mathbf{R}^N$ whose boundary $\partial\Omega$ is of class $\mathbf{C}^1$. Then, for any $u, v \in H^1(\Omega)$ and for any $1 \leq i \leq N$, we have*

$$\int_\Omega \frac{\partial u}{\partial x_i} v dx = -\int_\Omega u \frac{\partial v}{\partial x_i} dx + \int_{\partial\Omega} \gamma_0(u)\gamma_0(v)(\vec{n}.\vec{e}_i) d\sigma.$$

PROOF. Let us first establish the Green's formula for smooth functions $u, v \in \mathcal{D}(\bar{\Omega})$. Let us start from the divergence theorem, which states that for all $\mathbf{C}^1$ real-valued function u and vector-valued function $\vec{V}$,

$$\int_\Omega \operatorname{div}(u\vec{V}) dx = \int_{\partial\Omega} u(\vec{V} \cdot \vec{n})\, d\sigma.$$

Take now $\vec{V} = (0, \ldots, v, \ldots, 0) = v\vec{e}_i$, all components being equal to zero, except the component of rank i. We obtain

$$\int_\Omega \left(u \frac{\partial v}{\partial x_i} + v \frac{\partial u}{\partial x_i} \right) dx = \int_{\partial\Omega} uv(\vec{n} \cdot \vec{e}_i)\, d\sigma.$$

Let us now consider arbitrary elements $u, v \in H^1(\Omega)$ and use a density argument.

By Proposition 5.4.1, there exist approximating sequences $(u_k)_{k\in\mathbf{N}}$ (respectively, $(v_k)_{k\in\mathbf{N}}$) of elements of $\mathcal{D}(\bar{\Omega})$ such that u_k converges to u in $H^1(\Omega)$ (respectively, v_k converges to v in $H^1(\Omega)$). For each $k \in \mathbf{N}$, we have

$$\int_\Omega \left(u_k \frac{\partial v_k}{\partial x_i} + v_k \frac{\partial u_k}{\partial x_i} \right) dx = \int_{\partial\Omega} u_k v_k (\vec{n} \cdot \vec{e}_i)\, d\sigma. \tag{5.33}$$

Let us now apply Theorem 5.6.1. By definition of γ_0, and the continuity property of γ_0 from $H^1(\Omega)$ into $L^2(\partial\Omega)$, we have

$$u_k|_{\partial\Omega} \longrightarrow \gamma_0(u) \text{ in } L^2(\partial\Omega),$$
$$v_k|_{\partial\Omega} \longrightarrow \gamma_0(v) \text{ in } L^2(\partial\Omega).$$

Hence

$$\int_{\partial\Omega} u_k v_k (\vec{n}\cdot\vec{e}_i)\, d\sigma \longrightarrow \int_{\partial\Omega} \gamma_0(u)\gamma_0(v)(\vec{n}\cdot\vec{e}_i)\, d\sigma.$$

One can pass to the limit, without any difficulty, on the left-hand side of (5.33). We finally obtain

$$\int_\Omega \Big(u\frac{\partial v}{\partial x_i} + v\frac{\partial u}{\partial x_i}\Big)\, dx = \int_{\partial\Omega} \gamma_0(u)\gamma_0(v)(\vec{n}\cdot\vec{e}_i)\, d\sigma,$$

which ends the proof. □

Remark 5.6.1. (a) One should retain from the above argument the general method of the proof of Green's formulas: first establish it for smooth functions, then pass to the limit by using the continuity properties of the trace operators.

(b) The same formula holds for $u \in W^{1,p}(\Omega)$ and $v \in W^{1,q}(\Omega)$ with $\frac{1}{p} + \frac{1}{q} = 1$.

Let us now examine the space of traces $\{\gamma_0(v) : v \in W^{1,p}(\Omega)\}$. By Theorem 3.6.1 we know that the space trace is included in $L^p(\partial\Omega)$. Indeed, we are going to show that the range of γ_0 is a strict subspace of $L^p(\partial\Omega)$, namely, $\gamma_0(v) \in W^{1-1/p,p}(\partial\Omega)$. This result requires the use of fractional Sobolev spaces on a manifold, which is a quite involved subject. When $p = 2$, one can give a simpler proof of it thanks to the Fourier approach to Sobolev spaces.

Proposition 5.6.3 (range of γ_0). *Let Ω be an open bounded set in $\mathbf{R}^N$ whose boundary $\partial\Omega$ is of class $\mathbf{C}^1$. Then the trace operator γ_0 is linear continuous and onto from $H^1(\Omega) \to H^{1/2}(\partial\Omega)$.*

Proof. The definition of $H^{1/2}(\partial\Omega)$ is obtained by using local coordinates. Thus, we just need to prove Proposition 5.6.3 when $\Omega = \mathbf{R}^N_+$ and $\partial\Omega = \mathbf{R}^{N-1}$. By using the continuity of the reflection operator $\mathbf{P} : H^1(\mathbf{R}^N_+) \to H^1(\mathbf{R}^N)$ (see Lemma 5.4.1), we just need to prove that the trace operator

$$H^1(\mathbf{R}^N) \xrightarrow{\gamma_0} H^{1/2}(\mathbf{R}^{N-1})$$

is continuous. By using the density of $\mathcal{D}(\mathbf{R}^N)$ in $H^1(\mathbf{R}^N)$ and the definition of γ_0, we can finally reduce our study to the following situation.

Prove the existence of some constant $C \geq 0$ such that for any $v \in \mathcal{D}(\mathbf{R}^N)$

$$\|v(\cdot, 0)\|_{H^{1/2}(\mathbf{R}^{N-1})} \leq C\, \|v\|_{H^1(\mathbf{R}^N)}. \tag{5.34}$$

To prove (5.34), we use the Fourier transform as defined in Section 5.5.

Denoting $m_N = (1/(2\pi)^{N/2})dx$, we have

$$\mathcal{F}_N(v)(\xi) = \int_{\mathbf{R}^N} e^{-i\,\xi.x} v(x) dm_N,$$

where $\mathcal{F}_N$ refers to the Fourier transform in $\mathbf{R}^N$. Let us use the classical notations $x = (x', x_N) \in \mathbf{R}^{N-1} \times \mathbf{R}$. For any $v \in \mathcal{D}(\mathbf{R}^N)$, by using the Fubini theorem, we obtain

$$\begin{aligned}\mathcal{F}_N(v)(x', x_N) &= \int_{\mathbf{R}} e^{-i\,x_N\,t} \left(\int_{\mathbf{R}^{N-1}} e^{-i\,x'\cdot y} v(y,t) dm_{N-1}(y) \right) dm_1(t) \\ &= \mathcal{F}_1(\mathcal{F}_{N-1} v(x', \cdot))(x_N) \\ &= \mathcal{F}_{N-1}(\mathcal{F}_1 v(\cdot, x_N))(x'). \end{aligned} \tag{5.35}$$

Let us compute the $H^{1/2}(\mathbf{R}^{N-1})$ norm of $\gamma_0(v)$. We have

$$\gamma_0(v)(x') = v(x', 0) = \bar{\mathcal{F}}_1 \mathcal{F}_1(v(x', \cdot))(0),$$

where we used the Fourier inversion formula (Proposition 5.5.1). Hence

$$\gamma_0(v)(x') = \int_{-\infty}^{+\infty} \mathcal{F}_1(v(x', \cdot))(t)\, dt. \tag{5.36}$$

Let us apply $\mathcal{F}_{N-1}$ to the two sides of (5.36) and use (5.35) to obtain

$$\mathcal{F}_{N-1}(\gamma_0(v))(x') = \int_{-\infty}^{+\infty} (\mathcal{F}_N v)(x', t)\, dt. \tag{5.37}$$

From Theorem 5.6.1 we have

$$\begin{aligned}\|v\|_{H^1(\mathbf{R}^N)} &= \int_{\mathbf{R}^N} (1 + |x|^2) |(\mathcal{F}_N v)(x)|^2 dx \\ &= \int_{\mathbf{R}^N} (1 + |x'|^2 + t^2) |(\mathcal{F}_N v)(x', t)|^2 dx' dt. \end{aligned}$$

We may rewrite (5.37) as

$$\mathcal{F}_{N-1}(\gamma_0(v))(x') = \int_{-\infty}^{+\infty} (\mathcal{F}_N v)(x', t).(1 + |x'|^2 + t^2)^{1/2} \frac{dt}{(1 + |x'|^2 + t^2)^{1/2}}$$

and use the Cauchy–Schwarz inequality to obtain

$$|\mathcal{F}_{N-1}(\gamma_0(v))(x')|^2 \le \int_{-\infty}^{+\infty} |(\mathcal{F}_N v)(x', t)|^2 (1 + |x'|^2 + t^2)\, dt \int_{-\infty}^{+\infty} \frac{dt}{1 + |x'|^2 + t^2}. \tag{5.38}$$

An elementary computation yields (after the change of variable $t = (1 + |x'|^2)^{1/2} s$)

$$\int_{-\infty}^{+\infty} \frac{dt}{1 + |x'|^2 + t^2} = \frac{\pi}{(1 + |x'|^2)^{1/2}}. \tag{5.39}$$

Combining (5.38) and (5.39), we obtain

$$|(1 + |x'|^2)^{1/4} \mathcal{F}_{N-1}(\gamma_0(v))(x')|^2 \le \pi \int_{-\infty}^{+\infty} |(\mathcal{F}_N v)(x', t)|^2 (1 + |x'|^2 + t^2) dt. \tag{5.40}$$

Let us integrate (5.40) over $\mathbf{R}^{N-1}$ to obtain

$$\int_{\mathbf{R}^{N-1}} |(1+|x'|^2)^{1/4} \mathcal{F}_{N-1}(\gamma_0(v))(x')|^2 dx'$$
$$\leq \pi \int_{\mathbf{R}^N} |(\mathcal{F}_N v)(x',t)|^2 (1+|x'|^2+t^2) dx' dt. \tag{5.41}$$

By using again Theorem 5.5.1 and the Definition 5.5.1 of $H^{1/2}$, we thus get

$$\|\gamma_0(v)\|_{H^{1/2}(\mathbf{R}^{N-1})} \leq \sqrt{\pi} \|v\|_{H^1(\mathbf{R}^N)},$$

which expresses that γ_0 is continuous from $H^1(\mathbf{R}^N)$ into $H^{1/2}(\mathbf{R}^{N-1})$. □

Remark 5.6.2. When v belongs to $W^{2,p}(\Omega)$, by a similar argument one can give a meaning to $\frac{\partial v}{\partial n}$. Just notice that $\nabla v \in W^{1,p}(\Omega)^N$, and hence the trace of ∇v on $\partial\Omega$ belongs to $L^p(\partial\Omega)^N$. One defines

$$\frac{\partial v}{\partial n} := \gamma_0(\nabla v) \cdot \vec{n},$$

which belongs to $L^p(\partial\Omega)$. Indeed, one can show that

$$\frac{\partial v}{\partial n} \in W^{1-1/p,p}(\partial\Omega).$$

When $p = 2$, for $v \in H^2(\Omega)$ we have $\frac{\partial v}{\partial n} \in H^{1/2}(\partial\Omega)$.

One can also show that the operator $v \mapsto \{v|_{\partial\Omega}, \frac{\partial v}{\partial n}\}$ is linear continuous and onto from $W^{2,p}(\Omega)$ onto $W^{2-1/p,p}(\partial\Omega) \times W^{1-1/p,p}(\partial\Omega)$.

5.7 Sobolev embedding theorems

Let Ω be an open set in $\mathbf{R}^N$. We have seen in Section 5.1 (Theorem 5.1.1) that each element of the Sobolev space $W^{1,p}(a,b)$, $1 \leq p \leq +\infty$, has a continuous representative. This is no longer true for the elements of the space $W^{1,2}(\Omega)$ as soon as the dimension of the space $N \geq 2$. This raises a natural question: Is there a general relation between the numbers m, p, N which allows us to conclude that $W^{m,p}(\Omega) \hookrightarrow \mathbf{C}(\bar{\Omega})$?

Indeed, the answer is yes, and the Sobolev embedding Theorem 5.7.2 establishes that this is true as soon as $mp > N$. Another important aspect of the Sobolev embedding theorem is that, even if $v \in W^{m,p}(\Omega)$ and $mp < N$, one can say better than $v \in L^p(\Omega)$: indeed $v \in L^q(\Omega)$ with $\frac{1}{q} = \frac{1}{p} - \frac{m}{N}$.

We stress the fact that the Sobolev embedding theorem plays a crucial role in the variational approach to partial differential equations. It allows us to make the link between the two scales of spaces: $W^{m,p}(\Omega)$ and $\mathbf{C}^{k,\alpha}(\Omega)$.

When $p = 2$, an incisive approach to this question consists in using the Fourier–Plancherel transformation, as developed in Section 5.5.

Theorem 5.7.1. *Let $s > 0$ and assume that $2s > N$. Then $H^s(\mathbf{R}^N)$ is continuously embedded in $\mathbf{C}(\mathbf{R}^N)$.*

PROOF. We recall (see Theorem 5.5.1 and Definition 5.5.1) that

$$v \in H^s(\mathbf{R}^N) \Longleftrightarrow (1+|\xi|^2)^{s/2}\mathcal{F}(v) \in L^2(\mathbf{R}^N).$$

Let us set $g = (1+|\xi|^2)^{s/2}\mathcal{F}(v)$, which, by assumption, belongs to $L^2(\mathbf{R}^N)$. We have $\mathcal{F}(v) = g(1+|\xi|^2)^{-s/2}$ and $v = \bar{\mathcal{F}}(\mathcal{F}v)$. When $\mathcal{F}(v) = g(1+|\xi|^2)^{-s/2}$ belongs to $L^1(\mathbf{R}^N)$, then $\bar{\mathcal{F}}$ coincides with the classical inverse Fourier transform, and consequently $v \in \mathbf{C}(\mathbf{R}^N)$. Since g belongs to $L^2(\mathbf{R}^N)$, the function $g(1+|\xi|^2)^{-s/2}$ belongs to $L^1(\mathbf{R}^N)$ as soon as $(1+|\xi|^2)^{-s/2}$ belongs to $L^2(\mathbf{R}^N)$, i.e., $\int_{\mathbf{R}^N}(1+|\xi|^2)^{-s}\,d\xi < +\infty$. We have

$$\int_{\mathbf{R}^N} \frac{d\xi}{(1+|\xi|^2)^s} = c\int_0^{+\infty} \frac{r^{N-1}}{(1+r^2)^s}\,dr$$

and this last integral is finite iff $2s-(N-1) > 1$, i.e., $2s > N$. □

Let us now examine the general case $1 \le p \le +\infty$ and first consider the space $W^{1,p}(\Omega)$. By induction on m, we will then derive the result for $W^{m,p}(\Omega)$.

Theorem 5.7.2 (Sobolev). *Let Ω be an open bounded subset of $\mathbf{R}^N$ with a $\mathbf{C}^1$ boundary $\partial\Omega$. Let $1 \le p \le +\infty$ and consider the Sobolev space $W^{1,p}(\Omega)$. Then, the following continuous embedding results hold:*

(i) *If $1 \le p < N$, then $W^{1,p}(\Omega) \hookrightarrow L^{p^*}(\Omega)$ with $\frac{1}{p^*} = \frac{1}{p} - \frac{1}{N}$. More precisely, any element $v \in W^{1,p}(\Omega)$ belongs to $L^{p^*}(\Omega)$ and there exists a constant C, depending only on p, N, and Ω such that for all $v \in W^{1,p}(\Omega)$,*

$$\|v\|_{L^{p^*}(\Omega)} \le C\|v\|_{W^{1,p}(\Omega)}.$$

(ii) *If $p = N$, then $W^{1,p}(\Omega) \hookrightarrow L^q(\Omega)$ for all $1 \le q < +\infty$.*

(iii) *If $p > N$, then $W^{1,p}(\Omega) \hookrightarrow \mathbf{C}(\bar{\Omega})$. More precisely, we have $W^{1,p}(\Omega) \hookrightarrow \mathbf{C}^{0,\alpha}(\Omega)$ with $\alpha = 1 - \frac{N}{p}$, i.e., each element $v \in W^{1,p}(\Omega)$ is Hölder continuous with exponent α, and there exists a constant C depending only on p, N, and Ω such that for all $v \in W^{1,p}(\Omega)$,*

$$|v(x)-v(y)| \le C\|v\|_{W^{1,p}(\Omega)}|x-y|^\alpha \quad \text{for a.e. } x, y \in \Omega.$$

PROOF. The proof method is similar to the one used in Section 5.4. We first prove the Sobolev continuous embeddings (i), (ii), (iii) in the case $\Omega = \mathbf{R}^N$: indeed, one obtains in this case slightly more precise results as stated in Theorem 5.7.3 (Sobolev–Gagliardo–Nirenberg for the case $1 \le p < N$), Theorem 5.7.4 (Morrey for the case $p > N$), and Theorem 5.7.5 for the critical case $p = N$.

Let us introduce the linear continuous extension operator

$$\mathbf{P} : W^{1,p}(\Omega) \hookrightarrow W^{1,p}(\mathbf{R}^N),$$

whose properties are described in Theorem 5.4.1.

Let us consider, for example, the case $1 \leq p < N$. The composition of the continuous operators

$$W^{1,p}(\Omega) \overset{\mathbf{P}}{\hookrightarrow} W^{1,p}(\mathbf{R}^N) \overset{thm4.7.3}{\hookrightarrow} L^{p^*}(\mathbf{R}^N) \overset{r}{\hookrightarrow} L^{p^*}(\Omega)$$

(where r is the restriction operator to Ω) is still continuous, and it is the canonical embedding from $W^{1,p}(\Omega)$ into $L^{p^*}(\Omega)$, which is the identity. The same argument works for the cases $p > N$ and $p = N$.

Note that to use the extension operator technique, which is developed in Theorem 5.4.1, we need to make some regularity assumptions on Ω, namely, $\partial\Omega$ is assumed to be of class $\mathbf{C}^1$. □

From now on in this section, we work on the whole space $\mathbf{R}^N$.

5.7.1 Case $1 \leq p < N$

Theorem 5.7.3 (Sobolev–Gagliardo–Nirenberg). *Let $1 \leq p < N$. Then $W^{1,p}(\mathbf{R}^N) \hookrightarrow L^{p^*}(\mathbf{R}^N)$ with $\frac{1}{p^*} = \frac{1}{p} - \frac{1}{N}$. More precisely, there exists a constant $C = C(p, N)$ such that*

$$\forall v \in W^{1,p}(\mathbf{R}^N) \quad \|v\|_{L^{p^*}(\mathbf{R}^N)} \leq C\|\nabla v\|_{L^p(\mathbf{R}^N)}.$$

Remark 5.7.1. Before proving Theorem 5.7.3, it is worth pointing out that by an elementary homogeneity argument, one can obtain the precise value of $p^* = pN/(N-p)$. Let us assume that there exists a constant C and some $1 \leq q < +\infty$ such that for all $v \in W^{1,p}(\mathbf{R}^N)$,

$$\|v\|_{L^q(\mathbf{R}^N)} \leq C\|\nabla v\|_{L^p(\mathbf{R}^N)}.$$

Let us prove that, necessarily, $q = p^*$. In the above inequality, instead of v, let us take $v_\lambda(x) = v(\lambda x)$ with $\lambda > 0$. We have

$$\left(\int_{\mathbf{R}^N} |v(\lambda x)|^q dx\right)^{1/q} \leq C\left(\int_{\mathbf{R}^N} \lambda^p |\nabla v(\lambda x)|^p dx\right)^{1/p}.$$

Let us make the change of variable $y = \lambda x$. We obtain

$$\frac{1}{\lambda^{N/q}}\|v\|_{L^q(\mathbf{R}^N)} \leq C\lambda\frac{1}{\lambda^{N/p}}\|\nabla v\|_{L^p(\mathbf{R}^N)},$$

which, after simplification, yields

$$\|v\|_{L^q(\mathbf{R}^N)} \leq C\,\lambda^{(1-N/p+N/q)}\|\nabla v\|_{L^p(\mathbf{R}^N)}.$$

Clearly, this formula makes sense only when $1 - \frac{N}{p} + \frac{N}{q} = 0$, i.e., $q = p^*$. Otherwise, if $1 - \frac{N}{p} + \frac{N}{q} > 0$, let $\lambda \to 0$ in the above inequality, and if $1 - \frac{N}{p} + \frac{N}{q} < 0$, let $\lambda \to +\infty$. In both cases, one obtains that for any $v \in W^{1,p}(\mathbf{R}^N)$, $v \equiv 0$, which is an absurd statement!

To prove Theorem 5.7.3 we will use the following lemma, which is due to Gagliardo.

Lemma 5.7.1. *Let $N \geq 2$ and $g_1, g_2, \ldots, g_N \in L^{N-1}(\mathbf{R}^{N-1})$. Then the function g defined by $g(x) = g_1(\tilde{x}_1)g_2(\tilde{x}_2)\ldots g_N(\tilde{x}_N)$ belongs to $L^1(\mathbf{R}^N)$ and*

$$\|g\|_{L^1(\mathbf{R}^N)} \leq \prod_{i=1}^{N} \|g_i\|_{L^{N-1}(\mathbf{R}^{N-1})}.$$

PROOF. The inequality is obvious when $N = 2$. So let us argue by induction, assume that the result has been proved until N, and prove it for $N + 1$. Let us give $N + 1$ functions $g_1, g_2, \ldots, g_{N+1}$ belonging to $L^N(\mathbf{R}^N)$ and consider the function g defined for any $x = (x_1, x_2, \ldots, x_{N+1}) \in \mathbf{R}^{N+1}$ by

$$\begin{aligned} g(x) &= g_1(\tilde{x}_1)g_2(\tilde{x}_2)\ldots g_{N+1}(\tilde{x}_{N+1}) \\ &= [g_1(\tilde{x}_1)\ldots g_N(\tilde{x}_N)]g_{N+1}(\tilde{x}_{N+1}). \end{aligned}$$

Let us fix x_{N+1} and apply the Hölder inequality

$$\int_{\mathbf{R}^N} |g(x)|dx_1dx_2\ldots dx_N \leq \|g_{N+1}\|_{L^N(\mathbf{R}^N)} \left(\int_{\mathbf{R}^N} |g_1(\tilde{x}_1)\ldots g_N(\tilde{x}_N)|^{N'} dx_1 \ldots dx_N \right)^{1/N'} \tag{5.42}$$

with $N' = N/(N-1)$. Then, note that the functions $|g_1(\tilde{x}_1)|^{N'}, \ldots, |g_N(\tilde{x}_N)|^{N'}$ belong to $L^{N-1}(\mathbf{R}^{N-1})$. By the induction hypothesis,

$$\int_{\mathbf{R}^N} |g_1(\tilde{x}_1)\ldots g_N(\tilde{x}_N)|^{N'} dx_1 \ldots dx_N \leq \prod_{i=1}^{N} \|g_i(., x_{N+1})\|^{N'}_{L^N(\mathbf{R}^{N-1})}. \tag{5.43}$$

Combining (5.42) and (5.43) we obtain

$$\int_{\mathbf{R}^N} |g(x)|dx_1\ldots dx_N \leq \|g_{N+1}\|_{L^N(\mathbf{R}^N)} \prod_{i=1}^{N} \|g_i(\cdot, x_{N+1})\|_{L^N(\mathbf{R}^{N-1})}. \tag{5.44}$$

Let us now make x_{N+1} vary and integrate (5.44) over $\mathbf{R}$. We have

$$\int_{\mathbf{R}^{N+1}} |g(x)|dx_1\ldots dx_{N+1} \leq \|g_{N+1}\|_{L^N(\mathbf{R}^N)} \int_{\mathbf{R}} \prod_{i=1}^{N} \|g_i(\cdot, x_{N+1})\|_{L^N(\mathbf{R}^{N-1})}\, dx_{N+1}. \tag{5.45}$$

Let us notice that for all $i = 1, \ldots, N$ the function $\|g_i(\cdot, x_{N+1})\|_{L^N(\mathbf{R}^{N-1})}$ belongs to $L^N(\mathbf{R})$. Applying Hölder's inequality to (5.45) with $\frac{1}{N} + \cdots + \frac{1}{N} = 1$ (N times) we obtain

$$\int_{\mathbf{R}^{N+1}} |g(x)|dx_1\ldots dx_{N+1} \leq \|g_{N+1}\|_{L^N(\mathbf{R}^N)} \prod_{i=1}^{N} \|g_i\|_{L^N(\mathbf{R}^N)},$$

i.e.,

$$\|g\|_{L^1(\mathbf{R}^N)} \leq \prod_{i=1}^{N+1} \|g_i\|_{L^N(\mathbf{R}^N)},$$

which completes the induction and the proof. □

PROOF OF THEOREM 5.7.3. (a) Let us use a density argument and prove that it is equivalent to know that the inequality

$$\|v\|_{L^{p^*}(\mathbf{R}^N)} \leq C\|\nabla v\|_{L^p(\mathbf{R}^N)} \tag{5.46}$$

holds for any $v \in \mathcal{D}(\mathbf{R}^N)$ or for any $v \in W^{1,p}(\mathbf{R}^N)$. Given $v \in W^{1,p}(\mathbf{R}^N)$, by using Theorem 5.1.3, we can find a sequence $(v_n)_{n\in\mathbf{N}}$ of elements $v_n \in \mathcal{D}(\mathbf{R}^N)$ such that

$$v_n \to v \text{ in } W^{1,p}(\mathbf{R}^N) \text{ and } v_n(x) \to v(x) \text{ for almost every } x \in \mathbf{R}^N.$$

Let us write

$$\|v_n\|_{L^{p^*}(\mathbf{R}^N)} \leq C\|\nabla v_n\|_{L^p(\mathbf{R}^N)}.$$

Hence,

$$\begin{aligned}\|v\|_{L^{p^*}(\mathbf{R}^N)} &\leq \liminf_n \|v_n\|_{L^{p^*}(\mathbf{R}^N)} \\ &\leq C \lim_n \|\nabla v_n\|_{L^p(\mathbf{R}^N)} = C\|\nabla v\|_{L^p(\mathbf{R}^N)},\end{aligned}$$

where the first above inequality is obtained by using Fatou's lemma.

(b) Let us now verify that it is enough to prove (5.46) when $p = 1$. To do so, let us assume that it is true for $p = 1$, i.e.,

$$\forall v \in W^{1,1}(\mathbf{R}^N) \qquad \|v\|_{L^{1^*}(\mathbf{R}^N)} \leq C_1(N)\|\nabla v\|_{L^1(\mathbf{R}^N)}, \tag{5.47}$$

and prove that it is true for all $1 \leq p < N$.

Let us observe that if $v \in \mathcal{D}(\mathbf{R}^N)$, then $|v|^{p^*/1^*}$ belongs to $W^{1,1}(\mathbf{R}^N)$. Indeed, since $p > 1$ we have $p^* > 1^*$ and the function $|v|^{p^*/1^*}$ is continuously differentiable with compact support. Hence $|v|^{p^*/1^*} \in W^{1,1}(\mathbf{R}^N)$ and

$$\nabla\left(|v|^{p^*/1^*}\right) = \frac{p^*}{1^*} \operatorname{sign} v|v|^{(p^*/1^*)-1}\nabla v. \tag{5.48}$$

Let us replace v by $|v|^{p^*/1^*}$ in (5.47). Applying (5.48) we obtain

$$\left(\int_{\mathbf{R}^N} |v|^{p^*}dx\right)^{1/1^*} \leq C_1(N)\frac{p^*}{1^*}\int_{\mathbf{R}^N} |v|^{(p^*/1^*)-1}|\nabla v|dx. \tag{5.49}$$

Let us apply Hölder's inequality with $\frac{1}{p} + \frac{1}{p'} = 1$ to the right-hand side of (5.49):

$$\left(\int_{\mathbf{R}^N} |v|^{p^*}dx\right)^{1/1^*} \leq C_1(N)\frac{p^*}{1^*}\left(\int_{\mathbf{R}^N} |v|^{(\frac{p^*}{1^*}-1)p'}dx\right)^{1/p'}\left(\int_{\mathbf{R}^N} |\nabla v|^p dx\right)^{1/p}. \tag{5.50}$$

An elementary computation yields the equality $\frac{1}{1^*} - \frac{1}{p^*} = \frac{1}{p'}$, which is equivalent to $(\frac{p^*}{1^*} - 1)p' = p^*$, and allows us to simplify (5.50),

$$\|v\|_{L^{p^*}(\mathbf{R}^N)} \leq C_1(N)\frac{p^*}{1^*}\|\nabla v\|_{L^p(\mathbf{R}^N)}, \tag{5.51}$$

which is precisely (5.46) for an arbitrary $1 \leq p < N$. Note that we have obtained that one can take

$$C(p, N) = C_1(N)\frac{p^*}{1^*}. \tag{5.52}$$

(c) Thus, we just need to prove that for any $v \in \mathcal{D}(\mathbf{R}^N)$,

$$\|v\|_{L^{1^*}(\mathbf{R}^N)} \leq C_1(N)\|\nabla v\|_{L^1(\mathbf{R}^N)}. \tag{5.53}$$

For any $x = (x_1, x_2, \ldots, x_N) \in \mathbf{R}^N$ we have

$$\begin{aligned} |v(x)| &= \left| \int_{-\infty}^{x_i} \frac{\partial v}{\partial x_i}(x_1, x_2, \ldots, x_{i-1}, t, x_{i+1}, \ldots, x_N)\, dt \right| \\ &\leq \int_{-\infty}^{x_i} \left| \frac{\partial v}{\partial x_i}(x_1, x_2, \ldots, x_{i-1}, t, x_{i+1}, \ldots, x_N) \right| dt. \end{aligned}$$

Symmetrically,

$$|v(x)| \leq \int_{x_i}^{+\infty} \left| \frac{\partial v}{\partial x_i}(x_1, x_2, \ldots, x_{i-1}, t, x_{i+1}, \ldots, x_N) \right| dt.$$

Adding these two inequalities we obtain

$$|v(x)| \leq \frac{1}{2}\int_{-\infty}^{+\infty} \left| \frac{\partial v}{\partial x_i}(x_1, \ldots, x_{i-1}, t, x_{i+1}, \ldots, x_N) \right| dt. \tag{5.54}$$

Let us adopt the following notation. For any $x \in \mathbf{R}^N$ and $i = 1, 2, \ldots, N$

$$\tilde{x}_i := (x_1, x_2, \ldots, x_{i-1}, x_{i+1}, \ldots, x_N),$$

$$f_i(\tilde{x}_i) := \int_{-\infty}^{+\infty} \left| \frac{\partial v}{\partial x_i}(x_1, x_2, \ldots, x_{i-1}, t, x_{i+1}, \ldots, x_N) \right| dt.$$

We can rewrite (5.54) as $|v(x)| \leq \frac{1}{2} f_i(\tilde{x}_i)\ \forall i = 1, \ldots, N$, which implies, as a more symmetric expression,

$$|v(x)|^N \leq \frac{1}{2^N}\prod_{i=1}^{N} f_i(\tilde{x}_i). \tag{5.55}$$

Indeed, in (5.53) we need to majorize $\|v\|_{L^{1^*}}$. Noticing that $1^* = \frac{N}{N-1}$, let us write (5.55) as

$$|v(x)|^{1^*} \leq \frac{1}{2^{N/N-1}}\prod_{i=1}^{N} f_i(\tilde{x}_i)^{\frac{1}{N-1}}. \tag{5.56}$$

We note that for each $i = 1, \dots, N$ the function $g_i(\tilde{x}_i) := f_i(\tilde{x}_i)^{\frac{1}{N-1}}$ belongs to $L^{N-1}(\mathbf{R}^{N-1})$ and

$$\|g_i\|_{L^{N-1}(\mathbf{R}^{N-1})} = \left\| \frac{\partial v}{\partial x_i} \right\|_{L^1(\mathbf{R}^N)}^{\frac{1}{N-1}}. \tag{5.57}$$

Applying Lemma 5.7.1 to (5.56) and using (5.57) we obtain

$$\begin{aligned} \int_{\mathbf{R}^N} |v(x)|^{1^*} dx &\le \frac{1}{2^{N/N-1}} \left\| \prod_{i=1}^N g_i(\tilde{x}_i) \right\|_{L^1(\mathbf{R}^N)} \\ &\le \frac{1}{2^{N/N-1}} \prod_{i=1}^N \|g_i\|_{L^{N-1}(\mathbf{R}^{N-1})} \\ &\le \frac{1}{2^{N/N-1}} \prod_{i=1}^N \left\| \frac{\partial v}{\partial x_i} \right\|_{L^1(\mathbf{R}^N)}^{\frac{1}{N-1}}. \end{aligned}$$

Since $1^* = \frac{N}{N-1}$, it follows that

$$\|v\|_{L^{1^*}(\mathbf{R}^N)} \le \frac{1}{2} \prod_{i=1}^N \left\| \frac{\partial v}{\partial x_i} \right\|_{L^1(\mathbf{R}^N)}^{1/N}. \tag{5.58}$$

From the convexity inequality $(\prod_{i=1}^N a_i)^{1/N} \le \frac{1}{N} \sum_{i=1}^N a_i$, we finally obtain

$$\|v\|_{L^{1^*}(\mathbf{R}^N)} \le \frac{1}{2N} \sum_{i=1}^N \left\| \frac{\partial v}{\partial x_i} \right\|_{L^1(\mathbf{R}^N)} = \frac{1}{2N} \|\nabla v\|_{L^1(\mathbf{R}^N)}, \tag{5.59}$$

which ends the proof. □

Remark 5.7.2. Combining (5.47), (5.52), and (5.59), we obtain that one can take

$$C(p, N) = \frac{1}{2N} \frac{p^*}{1^*} = \frac{p(N-1)}{2N(N-p)}.$$

But this is not the best estimate. The best constant is strictly less than this one; it is known and quite involved (cf. Aubin [44], Talenti [216], and Lieb [171]).

As a direct consequence of the Sobolev–Gagliardo–Nirenberg theorem, one obtains the following Poincaré–Sobolev inequality.

Proposition 5.7.1. *There exists a constant $C(p, N)$ such that for any open subset Ω in $\mathbf{R}^N$ and for any $1 \le p < N$, the following inequality holds:*

$$\forall v \in W_0^{1,p}(\Omega) \quad \|v\|_{L^{p^*}(\Omega)} \le C(p, N) \|\nabla v\|_{L^p(\Omega)}.$$

Proof. Given $v \in W_0^{1,p}(\Omega)$, let $\tilde{v}$ be the extension of v by zero outside of Ω. By Proposition 5.1.1, $\tilde{v}$ belongs to $W^{1,p}(\mathbf{R}^N)$. Applying Theorem 5.7.3 to $\tilde{v}$ and noticing that $\|\tilde{v}\|_{L^{p^*}(\mathbf{R}^N)} = \|v\|_{L^{p^*}(\Omega)}$ and $\|\nabla \tilde{v}\|_{L^p(\mathbf{R}^N)} = \|\nabla v\|_{L^p(\Omega)}$, we obtain the desired conclusion. □

Remark 5.7.3. (a) A striking feature of the above Poincaré–Sobolev inequality is that it is valid for an arbitrary open set Ω (not necessarily bounded) and the constant $C(p, N)$ is independent of Ω. Indeed, one can take as a value of $C(p, N)$ the number $\frac{p(N-1)}{2N(N-p)}$.

These properties rely on the fact that one estimates $\|v\|_{L^{p^*}}$ from above by $\|\nabla v\|_{L^p}$. This makes a great contrast with the classical Poincaré inequality, where one estimates from above $\|v\|_{L^p}$ by $\|\nabla v\|_{L^p}$: in the classical Poincaré inequality (Theorem 5.3.1) one has to assume that Ω is bounded (at least in one direction) and the Poincaré constant does depend on Ω.

(b) When Ω is bounded, one can recover the classical Poincaré inequality from Proposition 5.7.1 just by using Hölder's inequality. Indeed, assuming that $1 \le p < N$,

$$\int_\Omega |v(x)|^p dx \le |\Omega|^{1-p/p^*} \left(\int_\Omega |v(x)|^{p^*} dx \right)^{p/p^*}.$$

Hence

$$\begin{aligned} \|v\|_{L^p(\Omega)} &\le |\Omega|^{1/p-1/p^*} \|v\|_{L^{p^*}(\Omega)} \\ &\le |\Omega|^{1/p-1/p^*} C(p, N) \|\nabla v\|_{L^p(\Omega)}. \end{aligned}$$

Using the equality $\frac{1}{p} - \frac{1}{p^*} = \frac{1}{N}$, we obtain

$$\|v\|_{L^p(\Omega)} \le |\Omega|^{1/N} C(p, N) \|\nabla v\|_{L^p(\Omega)}.$$

This is another way to see how the Poincaré constant $\bar{C}_p(\Omega)$ depends on Ω. Observing that $|R\Omega|^{1/N} = R|\Omega|$, we find again the conclusion of Proposition 5.3.1, which is $\bar{C}_p(R\Omega) = R\bar{C}_p(\Omega)$.

Let us summarize the above results in the following statement.

Corollary 5.7.1. *Let Ω be a bounded open subset of $\mathbf{R}^N$ and take $1 \le p < N$. Then, for any $v \in W_0^{1,p}(\Omega)$, the following inequality holds:*

$$\|v\|_{L^p(\Omega)} \le |\Omega|^{1/N} C(p, N) \|\nabla v\|_{L^p(\Omega)}.$$

For example, when $\Omega = B(0, r)$, we obtain

$$\int_{B_r} v(x)^2 dx \le Cr^2 \int_{B_r} |\nabla v(x)|^2 \, dx \qquad \forall v \in H_0^1(B_r).$$

We also have a Poincaré–Wirtinger–Sobolev inequality.

Proposition 5.7.2. *Let Ω be a bounded, connnected, open set in $\mathbf{R}^N$ whose boundary $\partial\Omega$ is of class $\mathbf{C}^1$. Then, there exists a constant $C(p, N, \Omega)$ such that for any $1 \le p < +\infty$, the following inequality holds:*

$$\forall v \in W^{1,p}(\Omega) \qquad \left\| v - \frac{1}{|\Omega|} \int_\Omega v(x) \, dx \right\|_{L^{p^*}(\Omega)} \le C(p, N, \Omega) \|\nabla v\|_{L^p(\Omega)}.$$

PROOF. Let us denote $M(v) = \frac{1}{|\Omega|}\int_\Omega v(x)\,dx$ and apply the Sobolev embedding Theorem 5.7.2(i) to the function $v - M(v)$. We obtain

$$\begin{aligned}\|v - M(v)\|_{L^{p^*}(\Omega)} &\le C_1(p, N, \Omega)\|v - M(v)\|_{W^{1,p}(\Omega)}\\ &\le C_1(p, N, \Omega)\big[\|v - M(v)\|_{L^p(\Omega)} + \|\nabla v\|_{L^p(\Omega)}\big].\end{aligned}$$

Let us now apply the classical Poincaré–Wirtinger inequality (Corollary 5.4.1) to v:

$$\|v - M(v)\|_{L^p(\Omega)} \le C_2(p, N, \Omega)\|\nabla v\|_{L^p(\Omega)}.$$

Combining the two last inequalities, we obtain

$$\|v - M(v)\|_{L^{p^*}(\Omega)} \le C(p, N, \Omega)\|\nabla v\|_{L^p(\Omega)},$$

which ends the proof. □

5.7.2 Case $p > N$

We now consider the space $W^{1,p}(\mathbf{R}^N)$ with $p > N$. The following theorem is due to Morrey [185].

Theorem 5.7.4 (Morrey). *Assume that $p > N$. Then there exists a continuous embedding $W^{1,p}(\mathbf{R}^N) \hookrightarrow \mathbf{C}^{0,\alpha}(\mathbf{R}^N)$ with $\alpha = 1 - \frac{N}{p}$. More precisely, there exists a constant $C(p, N)$ such that for all $v \in W^{1,p}(\mathbf{R}^N)$,*

$$|v(y) - v(x)| \le C(p, N)\|\nabla v\|_{L^p(\mathbf{R}^N)}|y - x|^\alpha \quad \text{for a.e. } x, y \in \mathbf{R}^N.$$

PROOF. Let us first take $v \in \mathcal{D}(\mathbf{R}^N)$. The proof is then completed by a density argument. Let Q be a cube containing the origin and whose edges are parallel to the coordinate axes in $\mathbf{R}^N$ and have a common length equal to $r > 0$. For each $x \in Q$ we have

$$v(x) - v(0) = \int_0^1 \frac{d}{dt}v(tx)dt.$$

From this, we infer

$$\begin{aligned}|v(x) - v(0)| &\le \int_0^1 \sum_{i=1}^N |x_i|\left|\frac{\partial v}{\partial x_i}(tx)\right| dt\\ &\le r\sum_{i=1}^N \int_0^1 \left|\frac{\partial v}{\partial x_i}(tx)\right| dt. \qquad (5.60)\end{aligned}$$

Let $\bar v := \frac{1}{|Q|}\int_Q v(x)dx$ denote the mean value of v on Q. Integrating (5.60) on Q we obtain

$$|\bar v - v(0)| \le \frac{r}{|Q|}\int_Q dx \sum_{i=1}^N \int_0^1 \left|\frac{\partial v}{\partial x_i}(tx)\right| dt.$$

Let us exchange the order of integration (Fubini's theorem):

$$|\bar{v} - v(0)| \le \frac{1}{r^{N-1}} \int_0^1 dt \int_Q \sum_{i=1}^N \left| \frac{\partial v}{\partial x_i}(tx) \right| dx.$$

Making the change of variable $y = tx$, we obtain

$$|\bar{v} - v(0)| \le \frac{1}{r^{N-1}} \int_0^1 dt \int_{tQ} \sum_{i=1}^N \left| \frac{\partial v}{\partial x_i}(y) \right| \frac{dy}{t^N}. \tag{5.61}$$

We now use Hölder's inequality and majorize this last integral as follows:

$$\int_{tQ} \left| \frac{\partial v}{\partial x_i}(y) \right| dy \le |tQ|^{1/p'} \left(\int_{tQ} \left| \frac{\partial v}{\partial x_i}(y) \right|^p dy \right)^{1/p}. \tag{5.62}$$

Since $0 \in Q$, we have $tQ \subset Q$ for all $0 \le t \le 1$. From the above inequalities (5.61) and (5.62) it follows that

$$\begin{aligned} |\bar{v} - v(0)| &\le \frac{r^{N/p'}}{r^{N-1}} \int_0^1 \frac{t^{N/p'}}{t^N} dt \sum_{i=1}^N \left(\int_Q \left| \frac{\partial v}{\partial x_i}(y) \right|^p dy \right)^{1/p} \\ &\le \frac{r^{1-N/p}}{1 - N/p} \|\nabla v\|_{L^p(Q)}. \end{aligned}$$

By translation, this inequality remains true for any cube Q whose edges are parallel to the coordinate axes and have common length equal to r. Hence, for any $x \in Q$,

$$|\bar{v} - v(x)| \le \frac{r^{1-N/p}}{1 - N/p} \|\nabla v\|_{L^p(Q)}.$$

We use the triangle inequality

$$|v(y) - v(x)| \le |v(y) - \bar{v}| + |\bar{v} - v(x)|$$

to obtain

$$|v(y) - v(x)| \le \frac{2r^{1-N/p}}{1 - N/p} \|\nabla v\|_{L^p(Q)} \qquad \forall x, y \in Q.$$

Then observe that for any two points $x, y \in \mathbf{R}^N$, there exists an open cube Q which is constructed as above and with $r = 2|y - x|$. It follows that for any $x, y \in \mathbf{R}^N$,

$$|v(y) - v(x)| \le C(p, N) \|\nabla v\|_{L^p(Q)} |y - x|^{1-N/p},$$

where $C(p, N)$ depends only on p and N. Here we have obtained $C(p, N) = \frac{2^{2-N/p}}{1-N/p}$. The proof is then completed by a standard density argument. □

5.7.3 Case $p = N$

Let us first show that for any $1 \leq q < +\infty$, $W^{1,N}(\mathbf{R}^N)$ is continuously embedded in $L^q_{loc}(\mathbf{R}^N)$. This result follows easily from the Sobolev–Gagliardo–Nirenberg theorem in the case $1 \leq p < N$ and the fact that $p^* = \frac{pN}{N-p}$ tends to $+\infty$ as p goes to N.

We recall that $v \in L^q_{loc}(\mathbf{R}^N)$ means that for any $R > 0$, $\int_{B(0,R)} |v(x)|^q \, dx < +\infty$. The topology on $L^q_{loc}(\mathbf{R}^N)$ is generated by the family of seminorms $\{\| \cdot \|_k, k = 1, 2, \ldots\}$

$$\|v\|_k = \left(\int_{B(0,R_k)} |v(x)|^q dx \right)^{1/q},$$

where R_k is an arbitrary sequence tending to $+\infty$ with k. We obtain in this way a Frechet topology (metrizable and complete), which does not depend on the choice of the sequence $(R_k)_{k \in \mathbf{N}}$, $R_k \to +\infty$. It is equivalent to say that $v_n \to v$ in $L^q_{loc}(\mathbf{R}^N)$ and

$$\forall R < +\infty \qquad \int_{B(0,R)} |v_n(x) - v(x)|^q dx \to 0.$$

We can now state the following result.

Theorem 5.7.5 (the limiting case $p = N$). *We have*

$$W^{1,p}(\mathbf{R}^N) \hookrightarrow L^q_{loc}(\mathbf{R}^N) \quad \forall \, 1 \leq q < +\infty$$

with a continuous injection.

Proof. Let $v \in W^{1,N}(\mathbf{R}^N)$. Let us use a truncation (on the domain) argument. Lemma 5.1.2 shows that for any $M \in \mathcal{D}(\mathbf{R}^N)$ the function Mv belongs to $W^{1,N}(\mathbf{R}^N)$ and has a compact support. As a consequence

$$Mv \in W^{1,N-\varepsilon}(\mathbf{R}^N) \quad \text{for all } \varepsilon > 0 \text{ arbitrarily small.}$$

(To obtain this result, we use that for any $R > 0$, for any $\varepsilon > 0$, $L^N(B(0,R)) \subset L^{N-\varepsilon}(B(0,R))$. Note that this is false on the whole of $\mathbf{R}^N$!) Let us apply Theorem 5.7.3 with $p = N - \varepsilon < N$. We obtain

$$Mv \in L^q(\mathbf{R}^N) \quad \text{with} \quad \frac{1}{q} = \frac{1}{N-\varepsilon} - \frac{1}{N} = \frac{\varepsilon}{N(N-\varepsilon)}.$$

Hence, for any $M \in \mathcal{D}(\mathbf{R}^N)$, for any $\varepsilon > 0$, $Mv \in L^{\frac{N(N-\varepsilon)}{\varepsilon}}(\mathbf{R}^N)$.

Noticing that $\frac{N(N-\varepsilon)}{\varepsilon} \to +\infty$ as $\varepsilon \to 0$, and taking $M \equiv 1$ on $B(0,R)$, we obtain that for any $1 \leq q < +\infty$, for any $R > 0$,

$$v \in L^q(B(0,R)).$$

One can easily verify that the above operations are continuous and so is the embedding $W^{1,N}(\mathbf{R}^N) \hookrightarrow L^q_{loc}(\mathbf{R}^N)$ for all $1 \leq q < +\infty$.

Remark 5.7.4. 1. Note that $W^{1,N}(\mathbf{R}^N) \hookrightarrow L^q_{loc}(\mathbf{R}^N)$ for all q finite. In general $v \in W^{1,N}(\mathbf{R}^N)$ does not imply that v is bounded on the bounded sets, as the example $v(x) = |\ln(x)|^k$ described in Section 5.1 shows.

2. One can show (see, for example, [90, Corollary IX.11] that $W^{1,N}(\mathbf{R}^N) \hookrightarrow L^q(\mathbf{R}^N)$ for all $1 \leq q < +\infty$. This result clearly implies the conclusion of Theorem 5.7.5.

3. Indeed, functions $v \in W^{1,N}(\mathbf{R}^N)$ are not in $L^\infty_{loc}(\mathbf{R}^N)$, but one can say something better than $v \in L^q_{loc}(\mathbf{R}^N)$ for each $1 \leq q < +\infty$. To do so, we need to use a sharper scaling of spaces than the classical Lebesgue $\{L^p; 1 \leq p \leq +\infty\}$ scaling. We use Orlicz spaces scaling as shown in the following proposition, which is again an easy consequence of Theorem 5.7.3.

Proposition 5.7.3. *There exist two constants $K > 0$ and $L > 0$ such that for any $R > 0$, for any function $v \in W^{1,N}(\mathbf{R}^N)$ which satisfies* spt $v \subset B(0, R)$ *and* $\|\nabla v\|_{L^N(\mathbf{R}^N)} \leq 1$, *we have*

$$\int_{B(0,R)} e^{Kv(x)} dx \leq L|B(0, R)|.$$

Proof. Let us return to Theorem 5.7.3. We have seen that for any $1 \leq p < N$, for any $v \in W^{1,p}(\mathbf{R}^N)$,

$$\|v\|_{L^{p^*}(\mathbf{R}^N)} \leq \frac{(N-1)p}{2N(N-p)} \|\nabla v\|_{L^p(\mathbf{R}^N)}.$$

We have

$$\frac{(N-1)p}{2N(N-p)} = \frac{(N-1)}{2N^2} \cdot \frac{Np}{N-p}.$$

Noticing that $p^* = \frac{Np}{N-p}$ and using the inequality $\frac{N-1}{2N^2} \leq 1$, we obtain that $\frac{(N-1)p}{2N(N-p)} \leq p^*$. Hence

$$\|v\|_{L^{p^*}(\mathbf{R}^N)} \leq p^* \|\nabla v\|_{L^p(\mathbf{R}^N)}.$$

Let us now suppose that spt $v \subset B(0, R)$. Using Hölder's inequality with $\frac{1}{p^*/p} + \frac{1}{N/p} = 1$, we obtain

$$\begin{aligned} \|\nabla v\|_{L^p(\mathbf{R}^N)} &= \left(\int_{B(0,R)} |\nabla v|^p dx \right)^{1/p} \\ &\leq |B_R|^{1/p^*} \left(\int_{B(0,R)} |\nabla v|^N dx \right)^{1/N}. \end{aligned}$$

Combining the two last inequalities, we obtain

$$\|v\|_{L^{p^*}(B_R)} \leq p^* |B_R|^{1/p^*} \|\nabla v\|_{L^N(B_R)}.$$

Then note that when $1 \leq p < N$, the corresponding $p^* = \frac{Np}{N-p}$ varies from $\frac{N}{N-1}$ to $+\infty$. Replacing p^* by q, we obtain the following intermediate result: for any q such that $\frac{N}{N-1} \leq q < +\infty$,

$$\|v\|_{L^q(B_R)} \leq q |B_R|^{1/q} \|\nabla v\|_{L^N(B_R)}. \tag{5.63}$$

We now use the asymptotic development

$$e^{K|v|} = 1 + K|v| + \frac{K^2}{2!}|v|^2 + \cdots + \frac{K^q}{q!}|v|^q + \cdots$$

to obtain

$$\int_{B_R} e^{K|v|}dx = |B_R| + K\int_{B_R} |v|\,dx + \sum_{q=2}^{+\infty} \frac{K^q}{q!}\int_{B_R} |v|^q\,dx. \tag{5.64}$$

Let us assume that $N \geq 2$. This implies $\frac{N}{N-1} \leq 2$, and (5.63) is valid for any $q \geq 2$. Since $\|\nabla v\|_{L^N(B_R)} \leq 1$, we have

$$\forall q \geq 2 \quad \frac{1}{|B_R|}\int_{B_R} |v|^q dx \leq q^q. \tag{5.65}$$

When $q = 1$, we use the Hölder inequality

$$\int_{B_R} |v(x)|dx \leq |B_R|^{1/2}\left(\int_{B_R} |v(x)|^2 dx\right)^{1/2}$$

and the inequality (5.65) with $q = 2$ to obtain

$$\frac{1}{|B_R|}\int_{B_R} |v(x)|dx \leq \left(\frac{1}{|B_R|}\int_{B_R} |v(x)|^2 dx\right)^{1/2} \leq 2. \tag{5.66}$$

Combining (5.64), (5.65), and (5.66), we obtain

$$\frac{1}{|B_R|}\int_{B_R} e^{K|v(x)|}dx \leq 1 + 2K + \sum_{q=2}^{+\infty} \frac{K^q q^q}{q!}. \tag{5.67}$$

Set $u_q := K^q q^q/q!$. We have

$$\lim_{q\to+\infty} \frac{u_{q+1}}{u_q} = K \lim_{q\to+\infty} \left(1 + \frac{1}{q}\right)^q = Ke.$$

Hence, the series $\sum K^q q^q/q!$ is convergent when $K < 1/e$. Choosing $K < 1/e$, we take $L = 1 + 2K + \sum_{q=2}^{+\infty} K^q q^q/q!$ and so we obtain for any $R > 0$

$$\frac{1}{|B_R|}\int_{B_R} e^{Kv(x)}dx \leq L,$$

which ends the proof. □

One can improve the conclusion of Proposition 5.7.3. Instead of (5.63), one can prove the sharper estimation

$$\|v\|_{L^q} \leq C_N q^{1-\frac{1}{N}} |B_R|^{1/q} \|\nabla v\|_{L^N}.$$

At this point, note the importance of getting the best constant $C(p, N)$ in the Sobolev embedding theorem. Then, by using the same argument as above, one obtains the following inequality, which is due to Trudinger [222] and Moser [189].

Proposition 5.7.4. *Under the assumptions of Proposition* 5.7.3*, there exist some constants* $\sigma > 0$ *and* $K(N) > 0$ *such that for any* v *in* $W^{1,N}(\mathbf{R}^N)$ *with* $\operatorname{spt} v \subset B_R$ *and* $\|\nabla v\|_{L^N(B_R)} \leq 1$*, we have*

$$\int_{B_R} e^{\sigma |v(x)|^{N/N-1}} dx \leq K(N)|B_R|.$$

Note that the power $\frac{N}{N-1}$ is the best possible power.

Let us now return to the general situation and complete the study of the Sobolev embedding theorem by the following results. Repeated applications of Theorem 5.7.2 enable us to obtain the statement below.

Theorem 5.7.6. *Let* Ω *be an open bounded subset of* $\mathbf{R}^N$ *with a* $\mathbf{C}^1$ *boundary* $\partial\Omega$*. Let* $1 \leq p < +\infty$ *and let* $m \geq 0$ *be an integer.*

(i) *If* $mp < N$*, then* $W^{m,p}(\Omega) \hookrightarrow L^q(\Omega)$ *with* $\frac{1}{q} = \frac{1}{p} - \frac{m}{N}$.

(ii) *If* $mp = N$*, then* $W^{m,p}(\Omega) \hookrightarrow L^q(\Omega)$ *for all* $1 \leq q < +\infty$.

(iii) *If* $mp > N$*, let us set* $k = [m - \frac{N}{p}]$ *and* $\alpha = m - \frac{N}{p} - k$ $(0 \leq \alpha < 1)$*. Then* $W^{m,p}(\Omega) \hookrightarrow \mathbf{C}^{k,\alpha}(\Omega)$*, where* $v \in C^{k,\alpha}(\Omega)$ *means that* $v \in \mathbf{C}^k(\Omega)$ *and* $D^l v \in \mathbf{C}^{0,\alpha}(\Omega)$ *for any* l *with* $|l| = k$.

Remark 5.7.5. As an illustration of the above theorem, we notice that $H^2(\Omega)$, where $\Omega \subset \mathbf{R}^N$, is continuously embedded in $\mathbf{C}(\bar{\Omega})$ as soon as $N < 4$, i.e., for $N = 1, 2, 3$.

Let us end this section by the following compactness embedding theorem. Using the Sobolev embedding theorem, one can improve the Rellich–Kondrakov theorem, Theorem 5.4.2.

Theorem 5.7.7. *Let* Ω *be an open bounded subset of* $\mathbf{R}^N$ *which has a* $\mathbf{C}^1$ *boundary* $\partial\Omega$*. Then, we have the following compact injections:*

(i) *If* $p < N$*,* $W^{1,p}(\Omega) \hookrightarrow L^q(\Omega)$ *for any* $q < p^*$*, with* $\frac{1}{p^*} = \frac{1}{p} - \frac{1}{N}$.

(ii) *If* $p = N$*,* $W^{1,p}(\Omega) \hookrightarrow L^q(\Omega)$ *for any* $1 \leq q < +\infty$.

(iii) *If* $p > N$*,* $W^{1,p}(\Omega) \hookrightarrow \mathbf{C}(\bar{\Omega})$.

Proof. The case (iii) follows from Theorem 5.7.2(iii) and Ascoli's theorem: any bounded subset of $W^{1,p}(\Omega)$ is bounded in $\mathbf{C}^{0,\alpha}(\Omega)$ with $\alpha = 1 - \frac{N}{p}$ and hence equicontinuous.

Let us consider the case (i). Take a bounded sequence $(v_n)_{n\in\mathbf{N}}$ in $W^{1,p}(\Omega)$. By the Sobolev embedding Theorem 5.7.2(i), the sequence $(v_n)_{n\in\mathbf{N}}$ is bounded in $L^{p^*}(\Omega)$. By the classical Rellich–Kondrakov theorem, Theorem 5.4.2, we can extract a convergent subsequence $v_{n_k} \to v$ in $L^p(\Omega)$. Let us prove that, indeed, the convergence of v_{n_k} to v holds in every $L^q(\Omega)$, with $1 \leq q < p^*$.

We use the following generalized version of Hölder's inequality: If $f_1, f_2, \dots, f_k$ satisfy $f_i \in L^{p_i}(\Omega)$ for $1 \le i \le k$ and

$$\frac{1}{p} = \frac{1}{p_1} + \frac{1}{p_2} + \cdots + \frac{1}{p_k} \le 1,$$

then the product $f = \prod_{i=1}^{k} f_i$ belongs to $L^p(\Omega)$ and

$$\|f\|_{L^p} \le \|f_1\|_{L^{p_1}} \dots \|f_k\|_{L^{p_k}}.$$

In particular, if $f \in L^p(\Omega) \cap L^q(\Omega)$ with $1 \le p \le q \le +\infty$, then $f \in L^r(\Omega)$ for all $p \le r \le q$ and the following interpolation formula holds:

$$\|f\|_{L^r} \le \|f\|_{L^p}^{\alpha} \|f\|_{L^q}^{1-\alpha} \quad \text{with} \quad \frac{1}{r} = \frac{\alpha}{p} + \frac{1-\alpha}{q} \quad (0 \le \alpha \le 1), \tag{5.68}$$

where we just applied Hölder's inequality to $|f|^{\alpha} \in L^{p/\alpha}$ and $|f|^{1-\alpha} \in L^{q/1-\alpha}$ noticing that

$$\frac{1}{r} = \frac{1}{p/\alpha} + \frac{1}{q/1-\alpha}\,.$$

So, let us take $p \le q < p^*$ and apply the above interpolation formula (5.68) to $|v_{n_k} - v|$ with $\frac{1}{q} = \frac{\alpha}{p} + \frac{1-\alpha}{p^*}$ and $\alpha > 0$ (because $q < p^*$). We have

$$\begin{aligned} \|v_{n_k} - v\|_{L^q(\Omega)} &\le \|v_{n_k} - v\|_{L^p}^{\alpha} \|v_{n_k} - v\|_{L^{p^*}}^{1-\alpha} \\ &\le C \|v_{n_k} - v\|_{L^p}^{\alpha} \end{aligned}$$

since $(v_{n_k})_{k \in \mathbf{N}}$ is bounded in $L^{p^*}(\Omega)$. Note that by Fatou's lemma, for example, one has also that v belongs to $L^{p^*}(\Omega)$. Since $\alpha > 0$ and $v_{n_k} \to v$ in $L^p(\Omega)$ we obtain that $v_{n_k} \to v$ in $L^q(\Omega)$ for any $p \le q < p^*$ and hence for any $1 \le q < p^*$ (Ω is bounded). □

Remark 5.7.6. Another proof of Theorem 5.7.7 relies on the use of Vitali's theorem: one can notice that for any $1 < q < p^*$ the sequence $\{|v_{n_k}|^q : k \in \mathbf{N}\}$ is equi-integrable (by a direct application of the classical Hölder's inequality) and converges in measure (after extraction of a further subsequence).

5.8 Capacity theory and elements of potential theory

In this section, we consider the Sobolev spaces $W^{1,p}(\Omega)$ and $W_0^{1,p}(\Omega)$, with $1 \le p < +\infty$, and pay particular attention to the case $p = 2$. We introduce the notion of capacity with respect to the energy functional $\Phi(v) = \int_\Omega |\nabla v(x)|^p \, dx$. As a key tool, we use that the contractions operate on the space $W^{1,p}(\Omega)$, as explained below.

5.8.1 Contractions operate on $W^{1,p}(\Omega)$

The central idea, which goes back to Deny and Beurling [65], is that the contractions operate on the spaces $W^{1,p}(\Omega)$ and $W_0^{1,p}(\Omega)$. Let us make this more precise. We say that $T : \mathbf{R} \to \mathbf{R}$

is a contraction if $T(0) = 0$ and $|T(r) - T(s)| \leq |r - s|$ for all $r, s \in \mathbf{R}$. To say that "the contractions operate on $W^{1,p}(\Omega)$" means that, for any $v \in W^{1,p}(\Omega)$, $T \circ v \in W^{1,p}(\Omega)$ and $\|T \circ v\|_{W^{1,p}(\Omega)} \leq \|v\|_{W^{1,p}(\Omega)}$.

We denote by $T \circ v$ the composition of the two functions $(T \circ v)(x) = T(v(x))$. When $p = 2$, this property is the basis of the so-called theory of Dirichlet spaces and Dirichlet forms.

The most commonly used contraction functions are $T(r) = r^+$, $T(r) = |r|$, and $T(r) = 1 \wedge r^+$ (fundamental contraction).

To state the results for a general open set Ω, we need to use the following density result, which is closely related to Theorem 5.1.3.

Theorem 5.8.1 (Friedrichs). *Let Ω be an arbitrary open set in $\mathbf{R}^N$. Then, for any $v \in W^{1,p}(\Omega)$, $1 \leq p < +\infty$, there exists a sequence $v_n \in \mathcal{D}(\mathbf{R}^N)$ such that*

$$v_n|_\Omega \to v \text{ in } L^p(\Omega) \quad \textit{and a.e.}$$

$$\frac{\partial v_n}{\partial x_i}|_\omega \to \frac{\partial v}{\partial x_i}|_\omega \text{ in } L^p(\omega) \quad \forall\, \omega \subset\subset \Omega,\ i = 1, \ldots, N.$$

PROOF. Let $\bar{v}$ be the extension by zero of v outside of Ω, i.e., $\bar{v}(x) = v(x)$ if $x \in \Omega$, $\bar{v}(x) = 0$ if $x \in \mathbf{R}^N \setminus \Omega$. Note that, except when $v \in W_0^{1,p}(\Omega)$, $\bar{v}$ does not belong to $W^{1,p}(\mathbf{R}^N)$!

We proceed analogously to the proof of Theorem 5.1.3. We regularize $\bar{v}$ by taking $v_n = M_n(\bar{v}(x) \star \rho_n)$, where M_n is a truncation function (on the domain) and ρ_n is a smoothing kernel. Clearly, v_n belongs to $\mathcal{D}(\mathbf{R}^N)$. Then note that for any $\omega \subset\subset \Omega$ one can find a function $\alpha \in \mathcal{D}(\Omega)$ such that $\alpha = 1$ on ω and $0 \leq \alpha \leq 1$. The point is that for n sufficiently large,

$$\bar{v} \star \rho_n = \alpha \bar{v} \star \rho_n \text{ on } \omega.$$

Since $\alpha \bar{v}$ belongs to $W^{1,p}(\mathbf{R}^N)$ one can apply the same argument as in the proof of Theorem 5.1.3 to obtain the result. After extraction of a subsequence one can also obtain the convergence almost everywhere. □

Let us first consider the case of smooth truncations.

Proposition 5.8.1. *Let $T \in \mathbf{C}^1(\mathbf{R})$ be a smooth truncation, i.e. $T : \mathbf{R} \to \mathbf{R}$ is a $\mathbf{C}^1$ function which satisfies $T(0) = 0$ and $|T'(r)| \leq 1$ for all $r \in \mathbf{R}$. Let Ω be an arbitrary open set in $\mathbf{R}^N$, and let $1 \leq p < +\infty$. Then, the following properties hold:*

(a) *for all $v \in W^{1,p}(\Omega)$, $T \circ v \in W^{1,p}(\Omega)$, $\frac{\partial}{\partial x_i}(T \circ v) = T'(v)\frac{\partial v}{\partial x_i}$ for all $1 \leq i \leq N$, and $\|T \circ v\|_{W^{1,p}(\Omega)} \leq \|v\|_{W^{1,p}(\Omega)}$;*

(b) *when $v \in W_0^{1,p}(\Omega)$ we have $T \circ v \in W_0^{1,p}(\Omega)$.*

PROOF. We have that for all $r, s \in \mathbf{R}$, $|T(r) - T(s)| \leq |r - s|$. This combined with $T(0) = 0$ yields $|T(r)| \leq |r|$ for all $r \in \mathbf{R}$. Hence

$$|T \circ v(x)| \leq |v(x)|$$

and $T \circ v \in L^p(\Omega)$. In a similar way,

$$\left| T'(v) \frac{\partial v}{\partial x_i} \right| \leq \left| \frac{\partial v}{\partial x_i} \right|$$

and $T'(v)\frac{\partial v}{\partial x_i}$ belongs to $L^p(\Omega)$. Thus, we just need to prove that $\frac{\partial}{\partial x_i}(T \circ v) = T'(v)\frac{\partial v}{\partial x_i}$ in the distribution sense, which means that

$$\forall \varphi \in \mathcal{D}(\Omega) \quad \int_\Omega (T \circ v) \frac{\partial \varphi}{\partial x_i} dx = -\int_\Omega T'(v) \frac{\partial v}{\partial x_i} \varphi \, dx. \tag{5.69}$$

To prove (5.69) we use Theorem 5.8.1, which provides us a sequence $(v_n)_{n\in\mathbf{N}}$ in $\mathcal{D}(\mathbf{R}^N)$ such that $v_n \to v$ in $L^p(\Omega)$ and a.e. and $\frac{\partial v_n}{\partial x_i} \to \frac{\partial v}{\partial x_i}$ in $L^p(\omega)$ for all $\omega \subset\subset \Omega$.

Since $T \circ v_n$ belongs to $\mathbf{C}^1(\Omega)$ and $\varphi \in \mathcal{D}(\Omega)$, we have by using classical differential calculus that for all $n \in \mathbf{N}$

$$\int_\Omega (T \circ v_n) \frac{\partial \varphi}{\partial x_i} dx = -\int_\Omega T'(v_n) \frac{\partial v_n}{\partial x_i} \varphi \, dx. \tag{5.70}$$

We have

$$|T \circ v_n - T \circ v| \leq |v_n - v|,$$

which implies that $T \circ v_n \to T \circ v$ in $L^p(\Omega)$.

On the other hand, let $\varphi \equiv 0$ outside of $\omega \subset\subset \Omega$. We have

$$\begin{aligned} \left| T'(v_n) \frac{\partial v_n}{\partial x_i} - T'(v) \frac{\partial v}{\partial x_i} \right| &\leq \left| T'(v_n) \left(\frac{\partial v_n}{\partial x_i} - \frac{\partial v}{\partial x_i} \right) \right| + \left| (T'(v_n) - T'(v)) \frac{\partial v}{\partial x_i} \right| \\ &\leq \left| \frac{\partial v_n}{\partial x_i} - \frac{\partial v}{\partial x_i} \right| + |T'(v_n) - T'(v)| \left| \frac{\partial v}{\partial x_i} \right|. \end{aligned}$$

Since $v_n \to v$ a.e. and T' is continuous, we have

$$T'(v_n) \to T'(v) \quad \text{a.e.}$$

and

$$|T'(v_n) - T'(v)|^p \left| \frac{\partial v}{\partial x_i} \right|^p \leq 2^p \left| \frac{\partial v}{\partial x_i} \right|^p.$$

We obtain, thanks to the Lebesgue dominated convergence theorem,

$$T'(v_n) \frac{\partial v_n}{\partial x_i} \to T'(v) \frac{\partial v}{\partial x_i} \text{ in } \quad L^p(\omega).$$

We can now pass to the limit in (5.70) to obtain (5.69).

Take now $v \in W_0^{1,p}(\Omega)$. The proof is even simpler, since one can take as an approximating sequence $v_n \in \mathcal{D}(\Omega)$ with $v_n \to v$ in $W^{1,p}(\Omega)$. We have $T \circ v_n \to T \circ v$ in $W^{1,p}(\Omega)$ and $T \circ v_n|_{\partial\Omega} = 0$. By continuity of the trace operator, we obtain $T \circ v \in W_0^{1,p}(\Omega)$. □

Let us now consider nonsmooth contractions. We start with the important case $T(r) = r^+ = r \vee 0$. The idea is to regularize the function $T(r)$ to reduce ourselves to the previous situation. This elementary construction is described below.

Lemma 5.8.1. *Let $T_n : \mathbf{R} \to \mathbf{R}$ be defined by*

$$T_n(r) = \begin{cases} r & \text{if} \quad r \geq 0, \\ r + n\frac{r^2}{2} & \text{if} \quad -\frac{1}{n} \leq r \leq 0, \\ -\frac{1}{2n} & \text{if} \quad r \leq -\frac{1}{n}. \end{cases}$$

Then, $T_n \in \mathbf{C}^1(\mathbf{R})$, $T_n(0) = 0$, $|T_n'(r)| \leq 1$ for all $r \in \mathbf{R}$ and, for all $r \in \mathbf{R}$, $T_n(r) \to r^+$ as $n \to +\infty$.

PROOF. The function $r \to r^+$ is not a $\mathbf{C}^1$ function. Its distribution derivative is not continuous. It is indeed the Heaviside function $\theta(r) = 1$ if $r \geq 0$, and $\theta(r) = 0$ elsewhere. Let us approximate θ by a sequence of continuous functions, taking for $n \geq 1$

$$\theta_n(r) = \begin{cases} 1 & \text{if} \quad r \geq 0, \\ nr + 1 & \text{if} \quad -\frac{1}{n} \leq r \leq 0, \\ 0 & \text{if} \quad r \leq -\frac{1}{n}. \end{cases}$$

Set

$$T_n(r) = \int_0^r \theta_n(s)ds.$$

Since $0 \leq \theta_n \leq 1$ we obtain

$$|T_n'(r)| = |\theta_n(r)| \leq 1,$$

and all the properties of T_n are easily verified. □

We can now state the following result.

Theorem 5.8.2. *Let Ω be an arbitrary open set in $\mathbf{R}^N$ and let $1 \leq p < +\infty$. Then, for any $v \in W^{1,p}(\Omega)$, $v^+ \in W^{1,p}(\Omega)$ and*

$$\frac{\partial}{\partial x_i} v^+ = \mathbf{1}_{\{v \geq 0\}} \frac{\partial v}{\partial x_i}, \quad i = 1, \ldots, N,$$

$$\|v^+\|_{W^{1,p}(\Omega)} \leq \|v\|_{W^{1,p}(\Omega)}.$$

Moreover, when $v \in W_0^{1,p}(\Omega)$, one still has $v^+ \in W_0^{1,p}(\Omega)$.

PROOF. Let T_n be the approximation of $T(r) = r^+$ which is defined in Lemma 5.8.1. Since T_n belongs to $\mathbf{C}^1(\mathbf{R})$, $T_n(0) = 0$ and $|T_n'(r)| \leq 1$ for all $r \in \mathbf{R}$, we can apply Proposition 5.8.1 to obtain

$$\frac{\partial}{\partial x_i} T_n \circ v = T_n'(v) \frac{\partial v}{\partial x_i}. \tag{5.71}$$

Let us pass to the limit on (5.71) as $n \to +\infty$. By Lemma 5.8.1, we have $T_n(r) \to r^+$ for all $r \in \mathbf{R}$. Hence

$$T_n \circ v \to v^+ \quad \text{a.e.}$$

and

$$|T_n \circ v| \leq |v|.$$

By the Lebesgue dominated convergence theorem, we have

$$T_n \circ v \to v^+ \text{ in } L^p(\Omega). \tag{5.72}$$

We now examine the right-hand side of (5.71). We have

$$T_n'(r) = 1 \quad \forall\, r \geq 0,$$

$$T_n'(r) \to 0 \text{ as } n \to +\infty \quad \forall\, r < 0.$$

Hence

$$T_n'(v) \to \mathbf{1}_{\{v \geq 0\}} \quad \text{a.e.}$$

and

$$T_n'(v)\frac{\partial v}{\partial x_i} \to \mathbf{1}_{\{v \geq 0\}}\frac{\partial v}{\partial x_i} \quad \text{a.e.}$$

Moreover,

$$\left| T_n'(v)\frac{\partial v}{\partial x_i} \right| \leq \left| \frac{\partial v}{\partial x_i} \right|.$$

We apply again the Lebesgue dominated convergence theorem and obtain

$$T_n'(v)\frac{\partial v}{\partial x_i} \to \mathbf{1}_{\{v \geq 0\}}\frac{\partial v}{\partial x_i} \quad \text{in } L^p(\Omega). \tag{5.73}$$

Combining (5.71), (5.72), and (5.73) we conclude that

$$\frac{\partial v^+}{\partial x_i} = \mathbf{1}_{\{v \geq 0\}}\frac{\partial v}{\partial x_i}. \tag{5.74}$$

In other words, the equality above follows from the continuity of the derivation with respect to the convergence in distribution and from the fact that the $L^p(\Omega)$-convergence implies the convergence in distribution.

From (5.74) we obtain

$$\left| \frac{\partial v^+}{\partial x_i} \right| \leq \left| \frac{\partial v}{\partial x_i} \right|.$$

This combined with $|v^+| \leq |v|$ yields

$$\|v^+\|_{W^{1,p}(\Omega)} \leq \|v\|_{W^{1,p}(\Omega)}.$$

Finally, note that (5.72) and (5.73) imply that $T_n \circ v \to v^+$ in $W^{1,p}(\Omega)$. If $v \in W_0^{1,p}(\Omega)$, we know by Proposition 5.8.1 that $T_n \circ v \in W_0^{1,p}(\Omega)$. Hence $v^+ \in W_0^{1,p}(\Omega)$. □

Let us derive from Theorem 5.8.2 some useful results.

Proposition 5.8.2. *Let $v \in W^{1,p}(\Omega)$, $1 \leq p < +\infty$. Then, for each $i = 1, \ldots, N$*

$$\mathbf{1}_{\{v=0\}}\frac{\partial v}{\partial x_i} = 0.$$

In other words,

$$\frac{\partial v}{\partial x_i} = 0 \quad a.e.\ in\ E = \{x \in \Omega : v(x) = 0\},\ i = 1, \ldots, N.$$

PROOF. Let us consider the truncation

$$T(r) = r^- = (-r) \vee 0.$$

An argument similar to the one of the proof of Theorem 5.8.2 shows that $v^- \in W^{1,p}(\Omega)$ and

$$\frac{\partial v^-}{\partial x_i} = -\mathbf{1}_{\{v \le 0\}} \frac{\partial v}{\partial x_i}.$$

From $v = v^+ - v^-$ we obtain

$$\begin{aligned}
\frac{\partial v}{\partial x_i} &= \frac{\partial v^+}{\partial x_i} - \frac{\partial v^-}{\partial x_i} \\
&= \left(\mathbf{1}_{\{v \ge 0\}} + \mathbf{1}_{\{v \le 0\}}\right) \frac{\partial v}{\partial x_i} \\
&= \left(\mathbf{1}_{\{v > 0\}} + \mathbf{1}_{\{v \le 0\}}\right) \frac{\partial v}{\partial x_i} + \mathbf{1}_{\{v = 0\}} \frac{\partial v}{\partial x_i} \\
&= \frac{\partial v}{\partial x_i} + \mathbf{1}_{\{v = 0\}} \frac{\partial v}{\partial x_i}.
\end{aligned}$$

Hence,

$$\mathbf{1}_{\{v = 0\}} \frac{\partial v}{\partial x_i} = 0,$$

which completes the proof. □

Corollary 5.8.1. *Let $v \in W^{1,p}(\Omega)$, $1 \le p < +\infty$. Then $|v| \in W^{1,p}(\Omega)$ and*

$$\frac{\partial}{\partial x_i}|v| = \mathbf{1}_{\{v \ge 0\}} \frac{\partial v}{\partial x_i} - \mathbf{1}_{\{v < 0\}} \frac{\partial v}{\partial x_i}.$$

Moreover, $\||v|\|_{W^{1,p}(\Omega)} = \|v\|_{W^{1,p}(\Omega)}$. Furthermore, if $v \in W_0^{1,p}(\Omega)$, then $|v| \in W_0^{1,p}(\Omega)$.

PROOF. We have $|v| = v^+ + v^-$. Hence

$$\begin{aligned}
\frac{\partial}{\partial x_i}|v| &= \frac{\partial v^+}{\partial x_i} + \frac{\partial v^-}{\partial x_i} \\
&= \mathbf{1}_{\{v \ge 0\}} \frac{\partial v}{\partial x_i} - \mathbf{1}_{\{v \le 0\}} \frac{\partial v}{\partial x_i}.
\end{aligned}$$

By using Proposition 5.8.2, we obtain

$$\frac{\partial}{\partial x_i}|v| = \mathbf{1}_{\{v \ge 0\}} \frac{\partial v}{\partial x_i} - \mathbf{1}_{\{v < 0\}} \frac{\partial v}{\partial x_i},$$

which ends the proof. □

Corollary 5.8.2. *The following facts hold.*

(a) *Let $u, v \in W^{1,p}(\Omega)$. Then, $u \wedge v$ and $u \vee v$ still belong to $W^{1,p}(\Omega)$, and*

$$\frac{\partial}{\partial x_i} u \wedge v = \mathbf{1}_{\{u>v\}} \frac{\partial u}{\partial x_i} + \mathbf{1}_{\{v \le u\}} \frac{\partial v}{\partial x_i}.$$

$$\frac{\partial}{\partial x_i} u \vee v = \mathbf{1}_{\{u,v\}} \frac{\partial u}{\partial x_i} + \mathbf{1}_{\{v \ge u\}} \frac{\partial v}{\partial x_i}.$$

(b) *The same holds true for $u, v \in W_0^{1,p}(\Omega)$.*

In particular, if $v \in W_0^{1,p}(\Omega)$ then $T_1 \circ v \in W_0^{1,p}(\Omega)$, where $T_1(r) = 1 \wedge r^+$, and

$$\int_\Omega |\nabla (T_1 \circ v)|^p dx \le \int_\Omega |\nabla v|^p dx.$$

PROOF. Just notice that

$$u \wedge v = u - (u - v)^+.$$

By Theorem 5.8.2, $u \wedge v \in W^{1,p}(\Omega)$ and

$$\begin{aligned} \frac{\partial}{\partial x_i} u \wedge v &= \frac{\partial u}{\partial x_i} - \frac{\partial}{\partial x_i} (u - v)^+ \\ &= \frac{\partial u}{\partial x_i} - \mathbf{1}_{\{u-v \ge 0\}} \left(\frac{\partial u}{\partial x_i} - \frac{\partial v}{\partial x_i} \right) \\ &= \mathbf{1}_{\{u-v<0\}} \frac{\partial u}{\partial x_i} + \mathbf{1}_{\{u-v \ge 0\}} \frac{\partial v}{\partial x_i} \\ &= \mathbf{1}_{\{u<v\}} \frac{\partial u}{\partial x_i} + \mathbf{1}_{\{u \ge v\}} \frac{\partial v}{\partial x_i}, \end{aligned}$$

which completes the proof for the function $u \wedge v$. The proof for $u \vee v$ is similar. □

5.8.2 Capacity

Let us introduce the capacity $\mathrm{Cap}_p(\cdot)$ which is associated to the energy functional defined by $\Phi(v) = \int_\Omega |\nabla v(x)|^p dx$. The capacity $\mathrm{Cap}_p(E)$ of a subset E of Ω is a nonnegative real number. The capacity theory allows us to study small sets in $\mathbf{R}^N$. Indeed, one can show that there are sets in $\mathbf{R}^N$ which are negligible with respect to the Lebesgue measure and whose capacity is strictly greater than zero. It makes sense to speak of the values of an arbitrary function $v \in W^{1,p}(\Omega)$ on sets $E \subset \Omega$ such that $\mathrm{Cap}_p(E) > 0$.

For example, a one-codimensional manifold in $\mathbf{R}^N$ (the boundary of a smooth open set, for example) has a strictly positive capacity, which highlights the fact that one can develop a trace theory on such sets (see Section 5.6). One can make these considerations precise and prove that each element of $W^{1,p}(\Omega)$ admits a quasi-continuous representative (see Evans and Gariepy [132], Ziemer [229]). This is a finer representation than the Lebesgue representation, where the functions are defined only a.e.

When $p > N$, the Morrey theorem, Theorem 5.7.4 says that each element $v \in W^{1,p}(\Omega)$ admits a continuous representative. In that case, the capacity theory is not useful

because only the empty set will be Cap_p-negligible. So we assume that $1 \leq p \leq N$. For simplicity of the exposition, we assume that Ω is bounded. We denote by $|\nabla v(x)|$ the Euclidean norm of $\nabla v(x)$, so that

$$\int_\Omega |\nabla v(x)|^p dx = \int_\Omega \left(\sum_{i=1}^N \left| \frac{\partial v}{\partial x_i}(x) \right|^2 \right)^{p/2} dx.$$

We may as well work with equivalent norms (like $|\nabla v|^p = \Sigma |\frac{\partial v}{\partial x_i}|^p$). This does not make any difference in the development of the capacity theory since the truly important notions for a set, such as "to have zero capacity" or "to have a capacity greater than zero," are invariant by using equivalent norms on $W_0^{1,p}(\Omega)$.

Definition 5.8.1. *Let Ω be an open bounded set in $\mathbf{R}^N$ and let $1 \leq p \leq N$.*

(a) *For any open subset G of Ω, the p-capacity of G with respect to Ω is defined by*

$$\mathrm{Cap}_p(G, \Omega) := \inf \left\{ \int_\Omega |\nabla v(x)|^p \, dx : v \in W_0^{1,p}(\Omega), v(x) \geq 1 \text{ a.e. on } G \right\}.$$

(b) *This definition is extended to any subset E of Ω in the following way:*

$$\mathrm{Cap}_p(E, \Omega) := \inf \left\{ \mathrm{Cap}_p(G, \Omega) : G \text{ open}, \ G \supset E \right\}.$$

Remark 5.8.1. (a) Let us notice that the above definition of the capacity Cap_p is a two-step procedure. For G open, one takes the infimum of the energy $\Phi(v)$ over all $v \in W_0^{1,p}(\Omega)$ which satisfy $v(x) \geq 1$ a.e. on G. Indeed, when G is open, the constraint $v \geq 1$ on G is easy to describe; it has just to be taken in the sense a.e. Then for an arbitrary set E, $\mathrm{Cap}_p(E)$ is the infimum of the capacity of the open sets G which contain E. We stress the fact that this definition is coherent since, clearly, for G open

$$\mathrm{Cap}_p(G, \Omega) = \inf\{\mathrm{Cap}_p(G', \Omega) : G' \text{ open}, \ G' \supset G\}.$$

(b) When $p = 2$, $\mathrm{Cap}_2(E, \Omega)$ is the harmonic capacity of E with respect to Ω. This notion comes from physics. In electrostatics, take a condensator whose internal part E has a potential equal to one and whose external part $\partial\Omega$ has a potential equal to zero. Then, $\mathrm{Cap}_2(E, \Omega)$ is, up to a constant factor, the total amount of charge (the energy) of the condensator.

Let us give equivalent formulations of $\mathrm{Cap}_p(E)$ for an arbitrary set E.

Proposition 5.8.3. *For any set $E \subset \Omega$,*

(a) $\mathrm{Cap}_p(E, \Omega) = \inf \left\{ \int_\Omega |\nabla v(x)|^p dx : v \in W_0^{1,p}(\Omega), v \geq 1 \text{ a.e. on a neighborhood of } E \right\}$ *where "$v \geq 1$ on a neighborhood of E" means that there exists an open set G which contains E and such that $v(x) \geq 1$ on G for almost every $x \in \Omega$.*

(b) $\mathrm{Cap}_p(E, \Omega) = \inf \left\{ \int_\Omega |\nabla v(x)|^p \, dx : v \in W_0^{1,p}(\Omega), v \geq 0, v = 1 \text{ a.e. on a neighborhood of } E \right\}$.

PROOF. (a) Take $v \in W_0^{1,p}(\Omega)$, $v \geq 1$ a.e. on a neighborhood of E. This means that there exists an open set $G \supset E$ such that $v \geq 1$ a.e. on G. By definition of $\mathrm{Cap}_p(G, \Omega)$, we have

$$\mathrm{Cap}_p(G, \Omega) \leq \int_\Omega |\nabla v(x)|^p dx.$$

Since $G \supset E$, G open, by definition of $\mathrm{Cap}_p(E, \Omega)$

$$\mathrm{Cap}_p(E, \Omega) \leq \mathrm{Cap}_p(G, \Omega),$$

which combined with the previous inequality yields

$$\mathrm{Cap}_p(E, \Omega) \leq \int_\Omega |\nabla v(x)|^p dx.$$

This being true for any such v, we infer that

$$\mathrm{Cap}_p(E, \Omega) \leq \inf \left\{ \int_\Omega |\nabla v(x)|^p dx : v \in W_0^{1,p}(\Omega),\ v \geq 1 \right.$$
$$\left. \text{a.e. on a neighborhood of } E \right\}.$$

Let us now prove the opposite inequality. If $\mathrm{Cap}_p(E, \Omega) = +\infty$, it is obvious. So, we assume $\mathrm{Cap}_p(E, \Omega) < +\infty$. For any $\varepsilon > 0$, there exists G_ε open, $G_\varepsilon \supset E$ such that

$$\mathrm{Cap}_p(G_\varepsilon, \Omega) \leq \mathrm{Cap}_p(E, \Omega) + \varepsilon.$$

By definition of $\mathrm{Cap}_p(G_\varepsilon, \Omega)$, there exists $v_\varepsilon \in W_0^{1,p}(\Omega)$ such that $v_\varepsilon \geq 1$ a.e. on G_ε and

$$\int_\Omega |\nabla v_\varepsilon(x)|^p dx \leq \mathrm{Cap}_p(G_\varepsilon, \Omega) + \varepsilon.$$

Adding the two last inequalities, we obtain

$$\int_\Omega |\nabla v_\varepsilon(x)|^p dx \leq \mathrm{Cap}_p(E, \Omega) + 2\varepsilon$$

with $v_\varepsilon \geq 1$ a.e. on a neighborhood of E. Hence

$$\inf \left\{ \int_\Omega |\nabla v|^p \, dx : v \in W_0^{1,p}(\Omega),\ v \geq 1 \quad \text{a.e. on a neighborhood of } E \right\}$$
$$\leq \mathrm{Cap}_p(E, \Omega) + 2\varepsilon.$$

Then, let $\varepsilon \to 0$ to obtain the result.

(b) Let $T(r) = 1 \wedge r^+$ be the fundamental contraction of the potential theory. Let us notice that if $v \in W_0^{1,p}(\Omega)$ satisfies "$v \geq 1$ a.e. on a neighborhood of E," then $T \circ v \geq 0$, $T \circ v = 1$ a.e. on a neighborhood of E.

Moreover, by Theorem 5.8.2 and Corollary 5.8.2 we have that $T \circ v$ belongs to $W_0^{1,p}(\Omega)$ and

$$\int_\Omega |\nabla(T \circ v)|^p dx \leq \int_\Omega |\nabla v|^p dx.$$

Hence,

$$\inf\left\{\int_\Omega |\nabla w|^p\,dx \;:\; w\in W_0^{1,p}(\Omega),\, w\ge 0,\, w=1 \text{ a.e. on a neighborhood of } E\right\}$$
$$\le \int_\Omega |\nabla (T\circ v)|^p\,dx$$
$$\le \int_\Omega |\nabla v|^p\,dx.$$

Then pass to the infimum on v to obtain

$$\inf\left\{\int_\Omega |\nabla w|^p\,dx \;:\; w\in W_0^{1,p}(\Omega),\, w\ge 0,\, w=1 \text{ a.e. on a neighborhood of } E\right\}$$
$$\le \inf\left\{\int_\Omega |\nabla v|^p\,dx \;:\; v\in W_0^{1,p}(\Omega),\, v\ge 1 \text{ a.e. on a neighborhood of } E\right\}.$$

The opposite inequality is clearly satisfied, since in the last infimum we can easily reduce ourselves to consider only nonnegative functions v. □

Let us now make the connection with the Choquet definition of capacities.

Proposition 5.8.4. (a) *For any compact set $K\subset\Omega$*

$$\mathrm{Cap}_p(K)=\inf\left\{\int_\Omega |\nabla v|^p dx : v\in\mathcal{D}(\Omega),\, v(x)\ge 1\ \forall\, x\in K\right\}.$$

(b) *For all G open set in Ω*

$$\mathrm{Cap}_p(G)=\sup\{\mathrm{Cap}_p(K) : K \textit{ compact, } K\subset G\}.$$

Proof. (a) The key point is that if G is an open set containing a compact K, then G contains an ε-enlargement K_ε of K with $\varepsilon>0$, where

$$K_\varepsilon=\{x\in\Omega : \mathrm{dist}\,(x,K)<\varepsilon\}.$$

To see this, just notice that $x\mapsto \mathrm{dist}(x,\Omega^c)$ is a continuous function on a compact set K which takes only positive values. Hence, its minimal value ε satisfies $\varepsilon>0$.

As a consequence, if $v\in W_0^{1,p}(\Omega)$ satisfies "$v\ge 1$ a.e. on a neighborhood of K," there exists some open set $G\supset K$ such that $v\ge 1$ a.e. on G and some $\varepsilon>0$ such that $G\supset K_\varepsilon$. Hence, $v\ge 1$ a.e. on K_ε.

Then use standard regularization techniques as developed in Theorem 5.1.3 to approximate v by a sequence $v_n\in\mathcal{D}(\Omega)$ which satisfies

"$v_n\ge 1$ a.e. on a neighborhood of K"

and

$$\int_\Omega |\nabla v_n|^p dx \longrightarrow \int_\Omega |\nabla v|^p dx.$$

To that end one should notice that, given $(\rho_n)_{n\in\mathbf{N}}$ a regularization kernel with $\rho_n \equiv 0$ outside of $\mathbf{B}(0, 1/n)$, we have

$$(v \star \rho_n)(x) = \int_{B(0,1/n)} v(x-y)\rho_n(y)\, dy.$$

When $1/n < \varepsilon$, $x \in K$, we have that for all $y \in \mathbf{B}(0, 1/n)$, $x - y \in K_\varepsilon$ and hence $v(x-y) \geq 1$. It follows that for $1/n < 1/\varepsilon$ and $x \in K$,

$$(v \star \rho_n)(x) \geq \int_{\mathbf{B}(0,1/n)} \rho_n(y) dy = 1.$$

Then combine this regularization by convolution method with a truncation on the domain (to preserve the Dirichlet boundary condition) to obtain v_n.

As a consequence,

$$\begin{aligned}\mathrm{Cap}_p(K) &= \inf\left\{ \int_\Omega |\nabla v|^p\, dx : v \in \mathcal{D}(\Omega),\, v(x) \geq 1 \quad \forall\, x \text{ in a} \right.\\ &\qquad\qquad \left. \text{neighborhood of } K \right\}\\ &\geq \inf\left\{ \int_\Omega |\nabla v|^p\, dx : v \in \mathcal{D}(\Omega),\, v(x) \geq 1\ \forall\, x \in K \right\},\end{aligned}$$

the last inequality being a consequence of the fact that one minimizes on a larger set.

(b) Since $\mathrm{Cap}_p(\cdot)$ is a monotone set function, we have

$$\mathrm{Cap}_p(G) \geq \sup\{\mathrm{Cap}_p(K) \ :\ K \text{ compact}, K \subset G\}.$$

To prove the opposite inequality, we consider only the nontrivial case when the right-hand side is finite. Take a sequence (K_n) of compact subsets of G such that $\cup_n K_n = G$ and for any integer n, by using statement (a) above, let $u_n \in \mathcal{D}(\Omega)$ such that

$$u_n \geq 1 \text{ on } K_n, \quad \int_\Omega |\nabla u_n|^p\, dx \leq \frac{1}{n} + \mathrm{Cap}_p(K_n).$$

The sequence (u_n) is bounded in $W_0^{1,p}(\Omega)$ and we may extract a subsequence (that we still denote by (u_n)) weakly converging to some $u \in W_0^{1,p}(\Omega)$. Then $u \geq 1$ a.e. on G and

$$\begin{aligned}\mathrm{Cap}_p(G) &\leq \int_\Omega |\nabla u|^p\, dx\\ &\leq \liminf_{n\to+\infty} \int_\Omega |\nabla u_n|^p\, dx\\ &\leq \liminf_{n\to+\infty} \mathrm{Cap}_p(K_n)\\ &\leq \sup\{\mathrm{Cap}_p(K) \ :\ K \text{ compact}, K \subset G\},\end{aligned}$$

which concludes the proof. □

Chapter 6

Variational problems: Some classical examples

In this chapter it is shown how the direct variational method can be used to solve some classical boundary value problems.

The following examples have been selected because of their importance in continuum mechanics, physics, biology, and so forth and also because of their relative simplicity. Indeed, in most of these examples, the unknown function is a scalar valued function and the variational problem can be expressed as a convex minimization problem in a reflexive Banach space V; typically V is a Sobolev space $W^{m,p}(\Omega)$ with $1 < p < \infty$. The existence of a (weak) solution can be obtained by application of the convex coercive minimization theorem, Theorem 3.3.4. This contrasts with more involved situations where the unknown function is a vector-valued function, and/or when the functional is no longer convex, and/or the functional space is no longer reflexive, and/or the functional is no longer lower semicontinuous and coercive. Most of these questions will be considered in the next chapters.

When the functional which has to be minimized is the sum of a convex quadratic (positive) form and a linear form, an equivalent approach consists in working with the Euler equation (Proposition 2.3.1) and the corresponding existence theorem, namely, the Lax–Milgram theorem, Theorem 3.1.2. In that case, one can treat the problem by either of the two above equivalent methods.

We stress the fact that it is part of the skill of the mathematician to find a variational formulation (if it exists) of the studied problem. It is not a priori given! In particular, one has to find a functional setting which is well adapted to the problem under consideration. In the examples which are considered in this chapter we use various Sobolev spaces like $H_0^1(\Omega)$, $H^1(\Omega)$, $H^2(\Omega)$, $H^2_{per}(\Omega)$, $W^{1,p}(\Omega)$.

As a general rule, the variational methods provide only weak solutions. It is an important (and often quite involved) question to study the regularity of the variational solution and hence to decide whether it is a classical solution. We will just give some indications on this question in the case of the Dirichlet problem.

Notation. Ω is a bounded open set in $\mathbf{R}^N$; $\partial\Omega$ is its topological boundary (also denoted

by Γ). Ω is said to be a regular open set if $\partial\Omega$ is piecewise of class $\mathbf{C}^1$, $n(x)$ is the outward unit normal vector to $\partial\Omega$ at x. Given $v : \bar{\Omega} \to \mathbf{R}$ we write

$$\frac{\partial v}{\partial n}(x) = \nabla v(x) \cdot n(x),$$

the outward normal derivative of v at $x \in \partial\Omega$.

6.1 The Dirichlet problem

Given $f : \Omega \to \mathbf{R}$, we are looking for a solution $u : \bar{\Omega} \to \mathbf{R}$ of the following boundary value problem:

$$\begin{cases} -\Delta u = f & \text{on } \Omega, \\ u = 0 & \text{on } \partial\Omega. \end{cases} \tag{6.1}$$

The condition "$u = 0$ on $\partial\Omega$" is called the homogeneous Dirichlet boundary condition. Problem (6.1) is called the homogeneous Dirichlet boundary value problem for the Laplace operator. We call it the Dirichlet problem.

The nonhomogeneous Dirichlet problem consists in finding the $u : \bar{\Omega} \to \mathbf{R}$ solution of

$$\begin{cases} -\Delta u = f & \text{on } \Omega, \\ u = g & \text{on } \partial\Omega, \end{cases}$$

where $f : \Omega \to \mathbf{R}$ and $g : \partial\Omega \to \mathbf{R}$ are given functions. The word *homogeneous* refers precisely to the case $g = 0$. We will see at the end of this section that, in general, the nonhomogeneous problem can be reduced to the homogeneous problem.

6.1.1 The homogeneous Dirichlet problem

Let us recall that when $N = 2$, problem (5.1) can be interpreted, for example, as describing the vertical motion of an elastic membrane under the action of a vertical force of density f. The Dirichlet boundary condition expresses that the membrane is fixed on its boundary.

Theorem 6.1.1. *The variational approach to the Dirichlet problem is described in the following statements.*

(a) *For every $f \in L^2(\Omega)$ there exists a unique $u \in H_0^1(\Omega)$ which satisfies*

$$\begin{cases} \displaystyle\int_\Omega \nabla u \cdot \nabla v \, dx = \int_\Omega f v \, dx \quad \forall v \in H_0^1(\Omega), \\ u \in H_0^1(\Omega). \end{cases} \tag{6.2}$$

(b) *The solution u of (6.2) satisfies*

$$\begin{cases} -\Delta u = f & \text{in } \mathcal{D}'(\Omega) \quad (\textit{equality as distributions}), \\ \gamma_0(u) = 0 & \text{on } \partial\Omega \quad (\gamma_0 \textit{ is the trace operator}). \end{cases} \tag{6.3}$$

Indeed, for $u \in H_0^1(\Omega)$ there is equivalence between (6.2) *and* (6.3)*. The solution u of* (6.2) *is called the weak solution of the Dirichlet problem* (6.1)*.*

(c) *The solution u of* (6.2) *is the unique solution of the minimization problem*

$$\min\left\{\frac{1}{2}\int_\Omega |\nabla v|^2\,dx - \int_\Omega f\,v\,dx \;:\; v \in H_0^1(\Omega)\right\}. \tag{6.4}$$

This is the Dirichlet variational principle. We also call u the variational solution of the Dirichlet problem.

PROOF. (a) Let us solve (6.2) by using the Lax–Milgram theorem. To that end, take $V = H_0^1(\Omega)$ equipped with the scalar product

$$\langle u, v\rangle = \int_\Omega (uv + \nabla u\cdot\nabla v)\,dx,$$

which makes V a Hilbert space. Then, set for any $u, v \in V$,

$$a(u,v) = \int_\Omega \nabla u\cdot\nabla v\,dx,$$

$$l(v) = \int_\Omega f\,v\,dx.$$

Let us first verify that the bilinear form $a : V\times V \to \mathbf{R}$ is continuous. For arbitrary $u, v \in V$, by using successively the Cauchy–Schwarz inequality in $\mathbf{R}^N$ and $L^2(\Omega)$, we obtain

$$\begin{aligned}|a(u,v)| &\le \int_\Omega |\nabla u|\,|\nabla v|\,dx\\ &\le \left(\int_\Omega |\nabla u|^2\,dx\right)^{1/2}\left(\int_\Omega |\nabla v|^2\,dx\right)^{1/2}\\ &\le \|u\|_V\,\|v\|_V.\end{aligned}$$

Let us now verify that the linear form $l : V \to \mathbf{R}$ is continuous. For arbitrary $v \in V$

$$\begin{aligned}|l(v)| &\le \int_\Omega |f||v|\,dx\\ &\le \left(\int_\Omega |f|^2\,dx\right)^{1/2}\left(\int_\Omega |v|^2\,dx\right)^{1/2}\\ &\le C\|v\| \quad \text{with } C = \|f\|_{L^2}.\end{aligned}$$

The only point which remains to verify is that the bilinear form a is coercive. To that end, we use the Poincaré inequality (Theorem 5.3.1). Since Ω has been assumed to be bounded, there exists some positive constant C such that

$$\forall v \in H_0^1(\Omega)\quad \int_\Omega v(x)^2\,dx \le C\int_\Omega |\nabla v(x)|^2\,dx.$$

By adding $\int_\Omega |\nabla v|^2\,dx$ to each side of the above inequality, we obtain

$$\int_\Omega (v(x)^2 + |\nabla v(x)|^2)\,dx \le (1+C)\int_\Omega |\nabla v(x)|^2\,dx.$$

Equivalently,

$$\forall v \in H_0^1(\Omega) \quad a(v,v) \ge \frac{1}{1+C}\|v\|_V^2,$$

and a is α-coercive (or α-elliptic) with $\alpha = \frac{1}{1+C} > 0$. Thus, all the assumptions of the Lax–Milgram theorem are satisfied. This implies existence and uniqueness of the solution u of problem (6.2).

(b) Let u be the solution of (6.2). Since $\mathcal{D}(\Omega) \subset H_0^1(\Omega)$, we have

$$\forall v \in \mathcal{D}(\Omega) \quad \int_\Omega \nabla u \cdot \nabla v\,dx = \int_\Omega f v\,dx, \tag{6.5}$$

which, by definition of the derivation in the distribution sense, is equivalent to

$$-\Delta u = f \quad \text{in } \mathcal{D}'(\Omega). \tag{6.6}$$

Moreover, we know by Proposition 5.6.1 that $H_0^1(\Omega) = \ker \gamma_0$, where γ_0 is the trace operator. Hence

$$\gamma_0(u) = 0 \quad \text{in trace sense}$$

and u satisfies (6.3). Conversely, if u satisfies $-\Delta u = f$ in the distribution sense, we have (6.5). Then use the density of $\mathcal{D}(\Omega)$ in $H_0^1(\Omega)$ and the fact that $u \in H_0^1(\Omega)$ and $f \in L^2(\Omega)$ to obtain (6.2).

(c) The equivalence between (6.5) and (6.3) is an immediate consequence of Proposition 2.3.1. To that end, just note that the bilinear form $a(\cdot,\cdot)$ is symmetric and positive.

The corresponding minimization problem is

$$\min\{J(v) \ : \ v \in H_0^1(\Omega)\},$$

where

$$\begin{aligned} J(v) &= \frac{1}{2}a(v,v) - l(v) \\ &= \frac{1}{2}\int_\Omega |\nabla v|^2\,dx - \int_\Omega f v\,dx. \end{aligned}$$

That's the Dirichlet variational principle. Note that the functional J is convex, continuous, and coercive on $H_0^1(\Omega)$. □

Let us now make the link between the notion of classical solution of the Dirichlet problem and the notion of weak solution which was introduced in Theorem 6.1.1. Let us first make precise the notion of classical solution.

Definition 6.1.1. *A function $u : \bar{\Omega} \to \mathbf{R}$ is said to be a classical solution of the Dirichlet problem if $u \in \mathbf{C}^2(\bar{\Omega})$ satisfies $-\Delta u = f$ in the sense of the classical differential calculus, while the restriction of u to $\partial\Omega$ is equal to zero:*

$$\begin{cases} -\Delta u(x) = f(x) & \forall x \in \Omega, \\ u(x) = 0 & \forall x \in \partial\Omega. \end{cases}$$

Proposition 6.1.1. (a) *If $u \in \mathbf{C}^2(\bar{\Omega})$ is a classical solution of the Dirichlet problem* (6.1)*, then it is equal to the weak solution of* (6.2)*. As a consequence, the classical solution, if it exists, is unique.*

(b) *If the weak solution u of* (6.2) *is regular, that is, $u \in \mathbf{C}^2(\bar{\Omega})$ and Ω is of class $\mathbf{C}^1$, then u is the classical solution of the Dirichlet problem* (6.1)*.*

Proof. (a) Let $u \in \mathbf{C}^2(\bar{\Omega})$ be a classical solution of the Dirichlet problem. Then u and $\frac{\partial u}{\partial x_i}$ for any $1 \le i \le N$ are continuous functions on the compact set $\bar{\Omega}$ (recall that Ω is bounded) and hence bounded on $\bar{\Omega}$. Since $L^\infty(\Omega) \subset L^2(\Omega)$ (we use again that Ω is bounded) we obtain that $u \in H^1(\Omega)$.

Since $u \in H^1(\Omega) \cap \mathbf{C}(\bar{\Omega})$, by using Proposition 5.6.1, we have $u = 0$ on $\partial\Omega$ so that $u \in H_0^1(\Omega)$. Note that this implication holds true without assuming any regularity assumption on $\partial\Omega$.

Take an arbitrary $v \in \mathcal{D}(\Omega)$. We have

$$-\int_\Omega v \Delta u \, dx = \int_\Omega f v \, dx.$$

Let us integrate by parts on Ω. We obtain

$$\forall v \in \mathcal{D}(\Omega) \quad \int_\Omega \nabla u \cdot \nabla v \, dx = \int_\Omega f v \, dx.$$

By density of $\mathcal{D}(\Omega)$ in $H_0^1(\Omega)$, we obtain

$$\forall v \in H_0^1(\Omega) \quad \int_\Omega \nabla u \cdot \nabla v \, dx = \int_\Omega f v \, dx.$$

Hence, u is the weak solution of (6.2).

(b) Let us assume that the weak solution u of (6.2) is regular, that is, $u \in \mathbf{C}^2(\bar{\Omega})$. Using again Proposition 5.6.1 and the fact that Ω is now assumed to be of class $\mathbf{C}^1$, we obtain

$$u \in \mathbf{C}(\bar{\Omega}) \cap H_0^1(\Omega) \Longrightarrow u = 0 \quad \text{on} \quad \partial\Omega \quad \text{(as a restriction to } \partial\Omega).$$

On the other hand, we have

$$-\Delta u = f \text{ in } \mathcal{D}'(\Omega).$$

Since $u \in \mathbf{C}^2(\Omega)$, the distributional derivatives of u (up to the second order) coincide with the classical derivatives and

$$-\Delta u(x) = f(x) \quad \forall x \in \Omega$$

in the classical sense. Hence, u is the classical solution of the Dirichlet problem. □

Remark 6.1.1. By Proposition 6.1.1, the question of the existence of a classical solution has been converted into the problem of the regularity of the weak solution. For this quite involved and important question, see [6], [7], [145], and [193].

6.1.2 The nonhomogeneous Dirichlet problem

Given $g : \partial\Omega \to \mathbf{R}$ and $f : \Omega \to \mathbf{R}$, we look for a solution $u : \bar{\Omega} \to \mathbf{R}$ of the following boundary value problem:

$$\begin{cases} -\Delta u = f & \text{on } \Omega, \\ u = g & \text{on } \partial\Omega. \end{cases} \tag{6.7}$$

Let us first give a variational formulation of this problem, as a constrained minimization problem on the set $C = \{v \in H^1(\Omega) \ : \ v = g \ \text{ on } \partial\Omega\}$. Then, we will see how this problem can be reduced to the homogeneous Dirichlet problem.

Theorem 6.1.2. *Let Ω be a bounded regular connected open set in $\mathbf{R}^N$. Let $g : \partial\Omega \to \mathbf{R}$ be a given function such that $g = \gamma_0(\tilde{g})$ for some $\tilde{g} \in H^1(\Omega)$, i.e., g belongs to $H^{1/2}(\partial\Omega)$, the trace space of $H^1(\Omega)$ on $\partial\Omega$. Let us denote*

$$C = \{v \in H^1(\Omega) \ : \ \gamma_0(v) = g \text{ on } \partial\Omega\}. \tag{6.8}$$

(i) *The set C is a closed convex nonempty subset of $H^1(\Omega)$. Indeed, $C = \tilde{g} + H_0^1(\Omega)$ is an affine subspace in $H^1(\Omega)$ which is parallel to $H_0^1(\Omega)$.*

(ii) *For any $f \in L^2(\Omega)$, there exists a unique solution u of the minimization problem*

$$\min\left\{\frac{1}{2}\int_\Omega |\nabla v(x)|^2\,dx - \int_\Omega f v\,dx \ : \ v \in C\right\}. \tag{6.9}$$

(iii) *The solution u of* (6.9) *is characterized by*

$$\begin{cases} \displaystyle\int_\Omega \nabla u \cdot \nabla v\,dx = \int_\Omega f v\,dx \qquad \forall v \in H_0^1(\Omega), \\ u \in C, \end{cases} \tag{6.10}$$

and it is a weak solution of the nonhomogeneous Dirichlet problem

$$\begin{cases} -\Delta u = f & \text{on } \Omega, \\ u = g & \text{on } \partial\Omega, \end{cases} \tag{6.11}$$

where $-\Delta u = f$ is interpreted in the distribution sense and $u = g$ in the trace sense.

PROOF. (i) The structure of C and its topological properties follow immediately from the fact that γ_0 is a linear continuous map from $H^1(\Omega)$ into $L^2(\partial\Omega)$.

(ii) The only point which is not immediate is the fact that the functional

$$J(v) := \frac{1}{2}\int_\Omega |\nabla v|^2\,dx - \int_\Omega f v\,dx + \delta_C(v) \tag{6.12}$$

is coercive on $H^1(\Omega)$. Given $\lambda \in \mathbf{R}$, let us prove that the set

$$\{J \le \lambda\} = \left\{ v \in C \ : \ \frac{1}{2}\int_\Omega |\nabla v|^2\,dx - \int_\Omega f v\,dx \le \lambda \right\}$$

is bounded in $H^1(\Omega)$. Set $v = w + \tilde{g}$ with $w \in H_0^1(\Omega)$. Equivalently, we have to prove that

$$K := \left\{ w \in H_0^1(\Omega) \ : \ \frac{1}{2}\int_\Omega |\nabla w + \nabla \tilde{g}|^2\,dx - \int_\Omega f(w+\tilde{g})\,dx \le \lambda \right\}$$

is bounded in $H_0^1(\Omega)$. Let us observe that

$$K \subset \left\{ w \in H_0^1(\Omega) \ : \ \int_\Omega |\nabla w|^2\,dx \le 2\int_\Omega |\nabla w||\nabla \tilde{g}|\,dx + 2\int_\Omega |f||w|\,dx + \gamma \right\}$$

with $\gamma = 2\lambda + 2\|f\|_{L^2}\|\tilde{g}\|_{L^2}$. By using the Poincaré inequality in $H_0^1(\Omega)$ and an elementary computation, one obtains that K is bounded, and hence J is coercive on $H^1(\Omega)$.

The functional J being convex lower semicontinuous (we use that C is closed convex) and coercive, the minimization problem of J on $H^1(\Omega)$, that is (6.9), admits a solution u. The uniqueness of the solution u follows from the strict convexity property of the Dirichlet integral: for any u_1, u_2 in C, we have

$$\begin{aligned}\int_\Omega \left|\nabla\left(\frac{u_1+u_2}{2}\right)\right|^2 dx = \frac{1}{2}\left[\int_\Omega |\nabla u_1|^2\,dx + \int_\Omega |\nabla u_2|^2\,dx\right] \\ -\frac{1}{4}\int_\Omega |\nabla(u_1-u_2)|^2\,dx.\end{aligned} \tag{6.13}$$

If u_1 and u_2 are two distinct solutions of (6.9), then the last term in (6.13) is strictly less than zero, which leads to a contradiction. We have used that $u_1 = u_2 = g$ on $\partial\Omega$ implies $u_1 - u_2 = 0$ on $\partial\Omega$ and that $\int_\Omega |\nabla(u_1-u_2)|^2\,dx = 0$ implies $u_1 - u_2$ is constant on Ω and hence, $u_1 = u_2$.

(iii) The general optimality condition for a minimization problem of the form

$$\min\left\{ \frac{1}{2}a(v,v) - l(v) \ : \ v \in C \right\}$$

is, according to Theorem 3.3.5,

$$\begin{cases} a(u, v-u) - l(v-u) \ge 0 \quad \forall v \in C, \\ u \in C. \end{cases} \tag{6.14}$$

Because of the particular structure of the set $C = \tilde{g} + H_0^1(\Omega)$, we have $v \in C$ iff $v - u \in H_0^1(\Omega)$ which is a subspace of $H^1(\Omega)$. As a consequence, (6.14) is equivalent to

$$\begin{cases} a(u,v) - l(v) = 0 \quad \forall v \in H_0^1(\Omega), \\ u \in C, \end{cases}$$

that is, (6.10).

Then use that $\mathcal{D}(\Omega)$ is dense in $H_0^1(\Omega)$ to obtain the equivalent formulation (6.11) in terms of distributions. □

Let us now make the link with the homogeneous Dirichlet problem. Indeed, we are going to show that just by taking as a new unknown function $w := u - \tilde{g}$, one can reduce the nonhomogeneous Dirichlet problem to some homogeneous Dirichlet problem. Formally, w satisfies

$$\begin{cases} -\Delta w = f + \Delta \tilde{g} & \text{on } \Omega, \\ w = 0 & \text{on } \partial\Omega, \end{cases}$$

which is a homogeneous Dirichlet problem, with a different right-hand side in the partial differential equation on Ω, namely, $f + \Delta\tilde{g}$, which now belongs to the space $H^{-1}(\Omega)$! Let us make this precise in the following statement.

Theorem 6.1.3. *Let us make the same assumptions and use the same notations as in Theorem* 6.1.2, *and fix some $\tilde{g} \in H^1(\Omega)$ such that $\gamma_0(\tilde{g}) = g$ on $\partial\Omega$.*

(i) *There exists a unique solution $w \in H_0^1(\Omega)$ of the problem*

$$\begin{cases} \displaystyle\int_\Omega \nabla w \cdot \nabla v \, dx = \int_\Omega f v \, dx - \int_\Omega \nabla \tilde{g} \cdot \nabla v \, dx \quad \forall v \in H_0^1(\Omega), \\ w \in H_0^1(\Omega), \end{cases} \tag{6.15}$$

and $u = w + \tilde{g}$ is the variational solution of the nonhomogeneous Dirichlet problem (6.9).

(ii) *The variational solution w of* (6.15) *is a weak solution of the boundary value problem*

$$\begin{cases} -\Delta w = f + \Delta \tilde{g} & \text{on } \Omega, \\ w = 0 & \text{on } \partial\Omega. \end{cases} \tag{6.16}$$

PROOF. Replacing u by $w + \tilde{g}$ in (6.10) gives exactly (6.15). Hence Theorem 6.1.3 is just an equivalent formulation of Theorem 6.1.2.

We could, as well, treat the nonhomogeneous Dirichlet problem by solving, in an independant way, the variational problem (6.15). Note that to apply the Lax–Milgram theorem, one needs to verify that

$$l(v) = \int_\Omega f v \, dx - \int_\Omega \nabla \tilde{g} \cdot \nabla v \, dx$$

is a linear continuous form on $H_0^1(\Omega)$, which is clear. This corresponds to the fact that in (6.16) w is a solution of the Laplace equation with a right-hand side $f + \Delta\tilde{g}$, which in general is no longer in $L^2(\Omega)$ but belongs to $H^{-1}(\Omega)$! □

Remark 6.1.2. From a practical and especially numerical point of view, it is much easier to work with $w = u - \tilde{g}$ and to solve the corresponding homogeneous Dirichlet problem by means of variational methods in the Sobolev space $H_0^1(\Omega)$. This justifies our special interest in studying the relations between the two approaches.

6.2 The Neumann problem

We are going to study successively the coercive Neumann problem (both in the homogeneous and the nonhomogeneous case) and then the semicoercive Neumann problem. In this section, Ω is a bounded open connected set in $\mathbf{R}^N$ which is assumed to be regular (piecewise of class $\mathbf{C}^1$). The outward unit normal vector to $\partial\Omega$ at $x \in \partial\Omega$ is denoted by $n(x)$. Generally speaking, the Neumann boundary condition expresses that the normal derivative $\frac{\partial u}{\partial n}$ of the unknown function is prescribed on $\partial\Omega$.

6.2.1 The coercive homogeneous Neumann problem

Let us give a function $a_0 \in L^\infty(\Omega)$ which satisfies the following assumption: there exists some positive real number $\alpha_0 > 0$ such that

$$a_0(x) \geq \alpha_0 \quad \text{for a.e. } x \in \Omega. \tag{6.17}$$

Given $f \in L^2(\Omega)$, we are looking for a solution u of the following boundary value problem:

$$\begin{cases} -\Delta u + a_0 u = f & \text{on } \Omega, \\ \frac{\partial u}{\partial n} = 0 & \text{on } \partial\Omega. \end{cases} \tag{6.18}$$

The word *homogeneous* refers to the fact that $\frac{\partial u}{\partial n}$ is prescribed being equal to zero on the boundary. The word *coercive* is related to the fact that problem (5.18) is a well-posed problem (existence and uniqueness of a solution) whose variational resolution involves a coercive bilinear form (equivalently, a coercive convex functional) on the space $H^1(\Omega)$. Let us make this precise in the statement below.

Theorem 6.2.1. *The following facts hold.*

(i) *There exists a unique solution $u \in H^1(\Omega)$ of the problem*

$$\begin{cases} \displaystyle\int_\Omega (\nabla u \cdot \nabla v + a_0 u v)\, dx = \int_\Omega f v\, dx \quad \forall v \in H^1(\Omega), \\ u \in H^1(\Omega). \end{cases} \tag{6.19}$$

(ii) *Equivalently, u is the unique solution of minimization problem*

$$\min \left\{ \frac{1}{2} \int_\Omega (|\nabla v|^2 + a_0 v^2)\, dx - \int_\Omega f v\, dx \ : \ v \in H^1(\Omega) \right\}. \tag{6.20}$$

(iii) *Let us assume that the solution u of* (6.19) *(or* (6.20)*) is regular, i.e., $u \in \mathbf{C}^2(\bar{\Omega})$. Then u is a classical solution of the Neumann problem* (6.18)*.*

(iv) *Conversely, if there exists a classical solution of the Neumann problem* (6.18)*, it is equal to the solution of the variational problem* (6.19) *(or* (6.20)*).*

Proof. (i) The functional space which is well adapted to the Neumann problem is $V = H^1(\Omega)$. Since Ω has been assumed to be regular, we know (see Proposition 5.4.1) that

$\mathcal{D}(\bar{\Omega})$ is a dense subspace of $H^1(\Omega)$. Recall that $\mathcal{D}(\bar{\Omega}) = \{v_{|\Omega} \ : \ v \in \mathcal{D}(\mathbf{R}^N)\}$. We will use elements $v \in \mathcal{D}(\bar{\Omega})$ as smooth test functions. Existence and uniqueness of a solution u of (6.19) is an immediate consequence of the Lax–Milgram theorem. Let us briefly give the proof. The bilinear form $a : H^1(\Omega) \times H^1(\Omega) \to \mathbf{R}$ which is defined by

$$a(u, v) = \int_\Omega (\nabla u \cdot \nabla v + a_0 u v)\, dx$$

verifies

$$\begin{aligned} |a(u, v)| &\leq \max\{1, \|a_0\|_{L^\infty}\} \int_\Omega (|\nabla u||\nabla v| + |u||v|) dx \\ &\leq \max\{1, \|a_0\|_{L^\infty}\} \|u\|_{H^1} \|v\|_{H^1} \end{aligned}$$

and a is continuous. On the other hand, for any $v \in H^1(\Omega)$

$$a(v, v) \geq \min\{1, \alpha_0\} \int_\Omega (|\nabla v|^2 + v^2)\, dx = \min\{1, \alpha_0\} \|v\|^2_{H^1}$$

and a is α-coercive on $V = H^1(\Omega)$ with $\alpha = \min\{1, \alpha_0\}$.

The linear form $l(v) = \int_\Omega f v\, dx$ verifies

$$|l(v)| \leq \|f\|_{L^2} \|v\|_{L^2} \leq \|f\|_{L^2} \|v\|_{H^1}$$

and l is continuous on $H^1(\Omega)$.

Thus, all the conditions of the Lax–Milgram theorem are satisfied and, as a consequence, there exists a solution u, which is unique, of problem (6.19).

(ii) The equivalence between (6.19) and the resolution of the minimization problem (6.20) follows immediately from Proposition 2.3.1 and the fact that a is positive and symmetric.

(iii) We now come to the most interesting question, which is the study of the relationship between the Neumann boundary value problem (6.18) and the variational problem (6.19) (or (6.20)). Indeed, the striking feature of the variational approach is that the Neumann boundary condition does not appear explicitly in the variational formulation. Indeed, it is implicitly contained in it; this is what we are going to verify now. Suppose that the solution u of (6.19) is regular, i.e., $u \in \mathbf{C}^2(\bar{\Omega})$. Take as a test function $v \in \mathcal{D}(\bar{\Omega}) \subset H^1(\Omega)$ and make an integration by parts on (6.19). (This is possible since u and v are regular up to the boundary.) We obtain

$$\forall v \in \mathcal{D}(\bar{\Omega}) \quad \int_\Omega (-\Delta u + a_0 u - f) v\, dx + \int_{\partial\Omega} v \frac{\partial u}{\partial n}\, d\sigma = 0. \tag{6.21}$$

Thanks to (6.21), we are going to test u successively on Ω and on $\partial\Omega$.

Let us first test u on Ω by taking $v \in \mathcal{D}(\Omega)$ in (6.21). Since $v = 0$ on $\partial\Omega$, the integral term on $\partial\Omega$ in (6.21) is equal to zero and we obtain

$$\forall v \in \mathcal{D}(\Omega) \quad \int_\Omega (-\Delta u + a_0 u - f) v\, dx = 0.$$

This implies that (see Theorem 9.3.1)

$$-\Delta u + a_0 u = f \quad \text{a.e. on } \Omega. \tag{6.22}$$

Let us now test u on $\partial\Omega$. To do so, we return to (6.21) with a general $v \in \mathcal{D}(\bar{\Omega})$ and use the previous information (6.22).

Formula (6.22) just expresses that $-\Delta u + a_0 u - f$, which is an $L^2(\Omega)$ function, is equal to zero almost everywhere on Ω. As a consequence, the first integral term on Ω in (6.21) is equal to zero and we obtain

$$\forall v \in \mathcal{D}(\bar{\Omega}) \quad \int_{\partial\Omega} v \frac{\partial u}{\partial n}\, d\sigma = 0. \tag{6.23}$$

We conclude thanks to the following lemma of independent interest.

Lemma 6.2.1. *Let Ω be a regular bounded open set in $\mathbf{R}^N$. Then, for any function $h \in L^2(\partial\Omega)$, we have*

$$\int_{\partial\Omega} hv\, d\sigma = 0 \quad \forall v \in \mathcal{D}(\bar{\Omega}) \Longrightarrow h = 0 \text{ on } \partial\Omega.$$

PROOF. Take $h \in L^2(\partial\Omega)$ which satisfies

$$\int_{\partial\Omega} hv\, d\sigma = 0 \quad \forall v \in \mathcal{D}(\bar{\Omega}). \tag{6.24}$$

By the density of $\mathcal{D}(\bar{\Omega})$ in $H^1(\Omega)$ and the continuity of the trace operator from $H^1(\Omega)$ into $L^2(\partial\Omega)$, property (6.24) can be extended by continuity to $H^1(\Omega)$, i.e.,

$$\int_{\partial\Omega} h\gamma_0(v)\, d\sigma = 0 \quad \forall v \in H^1(\Omega).$$

We know, by Proposition 5.6.3, that the image of $H^1(\Omega)$ by γ_0 is equal to $H^{1/2}(\partial\Omega)$. Hence

$$\int_{\partial\Omega} hv\, d\sigma = 0 \quad \forall v \in H^{1/2}(\partial\Omega).$$

The conclusion follows now from the density property of $H^s(\partial\Omega)$ in $L^2(\partial\Omega)$ for any $s > 0$. This last property can be easily deduced (by using local coordinates) from the fact that for any $s > 0$, $\mathcal{D}(\mathbf{R}^N)$ is included in $H^s(\mathbf{R}^N)$, and from the density property of $\mathcal{D}(\mathbf{R}^N)$ in $L^2(\mathbf{R}^N)$. □

PROOF OF THEOREM 6.2.1 (CONTINUATION). Let us complete the proof of Theorem 6.2.1 and prove the last point (iv). Let us assume that $u \in \mathbf{C}^2(\bar{\Omega})$ is a classical solution of the Neumann boundary value problem (6.18). Take $v \in \mathcal{D}(\bar{\Omega})$ an arbitrary test function, multiply (6.18) by v, and integrate on Ω. We obtain

$$\int_\Omega (-\Delta u + a_0 u) v\, dx = \int_\Omega f v\, dx \quad \forall v \in \mathcal{D}(\bar{\Omega}).$$

Integrating by parts, the above expression gives

$$\int_\Omega (\nabla u \cdot \nabla v + a_0 uv) dx - \int_{\partial\Omega} v \frac{\partial u}{\partial n}\, d\sigma = \int_\Omega f v\, dx \quad \forall v \in \mathcal{D}(\bar{\Omega}).$$

Since $\frac{\partial u}{\partial n} = 0$ on $\partial\Omega$, we get

$$\int_\Omega (\nabla u \cdot \nabla v + a_0 uv)\, dx = \int_\Omega f v\, dx \quad \forall v \in \mathcal{D}(\bar{\Omega}).$$

We now use the density of $\mathcal{D}(\bar{\Omega})$ in $H^1(\Omega)$ to extend this equality to any $v \in H^1(\Omega)$. Note also that $u \in \mathbf{C}^2(\bar{\Omega})$ clearly implies that $u \in H^1(\Omega)$. □

Remark 6.2.1. The following facts should be pointed out.

1. The solution u of (6.19) (equivalently, of the minimization problem (6.20)) is called the variational solution, or the weak solution of the Neumann boundary value problem (6.18). As in the case of the Dirichlet problem, the term *weak solution* is justified by the fact that the weak solution always exists, but on the counterpart, the equation on Ω and the Neumann boundary condition are satisfied only in a weak sense (respectively, distribution sense and trace sense). The existence of a classical solution is equivalent to the study of the regularity of the weak solution.

2. According to the variational approach to the Dirichlet problem, the most natural choice for the function space in which to solve the Neumann problem is given by the closure V in $H^1(\Omega)$ of

$$\mathcal{V} = \left\{ v \in \mathcal{D}(\bar{\Omega}) \ : \ \frac{\partial v}{\partial n} = 0 \quad \text{on} \quad \partial\Omega \right\}.$$

It is a good exercise to verify that $V = H^1(\Omega)$!

One may be convinced of this fact just by observing the following elementary situation: take $\Omega = (0, 1)$ and $v(x) = x$. For each $n \in \mathbf{N}$, take

$$v_n(x) = \begin{cases} \frac{1}{n} & \text{if} \quad 0 \le x \le \frac{1}{n}, \\ x & \text{if} \quad \frac{1}{n} \le x \le 1 - \frac{1}{n}, \\ 1 - \frac{1}{n} & \text{if} \quad 1 - \frac{1}{n} \le x \le 1. \end{cases}$$

Clearly $v_n \in H^1(0, 1)$, $\frac{dv_n}{dx}(0) = \frac{dv_n}{dx}(1) = 0$, and $v_n \to v$ in $H^1(0, 1)$ as $n \to +\infty$:

$$\|v_n - v\|^2_{H^1(0,1)} = 2 \int_0^{1/n} \left[\left(x - \frac{1}{n} \right)^2 + 1 \right] dx \le \frac{2}{n} + \frac{2}{n^3}\,.$$

Hence the property "$\frac{\partial v}{\partial x} = 0$ on $\partial\Omega$" is not stable for the convergence in the $H^1(\Omega)$ topology! As a conclusion, $H^1(\Omega)$ is the right functional space to solve the Neumann problem by variational methods.

3. As a general rule, homogeneous Neumann boundary conditions naturally occur in problems where the boundary is free, i.e., no constraints, no external action is exerted on $\partial\Omega$. This explains the importance of Neumann type boundary value problems in mechanics, physics, and so forth.

6.2.2 The coercive nonhomogeneous Neumann problem

We keep the same notation and assumptions as in the previous section. In addition, we suppose that there is some function $g \in L^2(\partial\Omega)$ which is given. We consider the (nonhomogeneous) Neumann problem

$$\begin{cases} -\Delta u + a_0 u = f & \text{on} \quad \Omega, \\ \frac{\partial u}{\partial n} = g & \text{on} \quad \partial\Omega. \end{cases} \tag{6.25}$$

The variational approach to (6.25) is similar to that in the proof of the homogeneous case (Theorem 6.2.1). One just needs to change the linear form $l : H^1(\Omega) \to \mathbf{R}$ by introducing the integral term $\int_{\partial\Omega} g v \, d\sigma$. Let us make this precise.

Theorem 6.2.2. (i) *Given $f \in L^2(\Omega)$ and $g \in L^2(\partial\Omega)$, there exists a unique solution $u \in H^1(\Omega)$ of the problem*

$$\begin{cases} \displaystyle\int_\Omega (\nabla u \cdot \nabla v + a_0 u v)\, dx = \int_\Omega f v \, dx + \int_{\partial\Omega} g v \, d\sigma \quad \forall v \in H^1(\Omega), \\ u \in H^1(\Omega). \end{cases} \tag{6.26}$$

(ii) *Equivalently, u is the unique solution of the minimization problem*

$$\min \left\{ \frac{1}{2}(|\nabla v|^2 + a_0 v^2)\, dx - \int_\Omega f v \, dx - \int_{\partial\Omega} g v \, d\sigma \ : \ v \in H^1(\Omega) \right\}. \tag{6.27}$$

(iii) *Let us assume that the solution u of* (6.26) *(equivalently,* (6.27)*) is regular, i.e., $u \in \mathbf{C}^2(\bar{\Omega})$. Then u is a classical solution of the nonhomogeneous Neumann problem* (6.25)*.*

(iv) *Conversely, if u is a classical solution of* (6.25)*, then it is equal to the solution of the variational problem* (6.26) *(equivalently,* (6.27)*).*

Proof. (i) We follow the lines of the proof of Theorem 6.2.1. We know, by Theorem 5.6.1, that the trace operator $\gamma_0 : H^1(\Omega) \to L^2(\partial\Omega)$ is continuous. From this, and by using the fact that $g \in L^2(\partial\Omega)$, we deduce that the linear form $v \in H^1(\Omega) \mapsto \int_{\partial\Omega} g v \, d\sigma$ is continuous on $H^1(\Omega)$. This clearly implies that the linear mapping $v \mapsto \int_\Omega f v \, dx + \int_{\partial\Omega} g v \, d\sigma$ is continuous on $H^1(\Omega)$. The assumptions of the Lax–Milgram theorem are satisfied. (The bilinear form $a : H^1(\Omega) \times H^1(\Omega) \to \mathbf{R}$ is the same as in the homogeneous case.) As a consequence, problem (6.26) admits a unique solution u.

(iii) Let us assume that u is a regular solution of (6.26). By taking first $v \in \mathcal{D}(\bar{\Omega})$ and by using the above equality, we infer that

$$\forall v \in \mathcal{D}(\bar{\Omega}) \quad \int_{\partial\Omega} v \left(\frac{\partial u}{\partial n} - g \right) d\sigma = 0.$$

By using Lemma 6.2.1, we conclude that $\frac{\partial u}{\partial n} = g$ on $\partial\Omega$. □

6.2.3 The semicoercive homogeneous Neumann problem

Let us now consider the following boundary value problem. Given $f \in L^2(\Omega)$, find $u : \bar{\Omega} \to \mathbf{R}$ which satisfies

$$\begin{cases} -\Delta u = f & \text{on} \quad \Omega, \\ \frac{\partial u}{\partial n} = 0 & \text{on} \quad \partial\Omega. \end{cases} \tag{6.28}$$

Note that in (6.28), the term a_0 which was supposed to be positive in the previous section is now equal to zero. This makes a big difference; problem (6.28) is no longer a well-posed problem. To see this, let us assume for a moment that u is a smooth solution of (6.28). Let us integrate (6.28) on Ω and use the divergence theorem. We obtain

$$\begin{aligned} \int_\Omega f(x)\,dx &= -\int_\Omega \Delta u\,dx \\ &= -\int_{\partial\Omega} \frac{\partial u}{\partial n}\,d\sigma \\ &= 0. \end{aligned}$$

Hence, a necessary condition for the existence of a (classical) solution of problem (6.28) is that $\int_\Omega f(x)\,dx = 0$. Moreover, we can observe also that if u is a solution of (6.28), then for any constant C, $u + C$ is also a solution.

Indeed, we are going to prove that the condition $\int_\Omega f(x)\,dx = 0$ is a necessary and sufficient condition for the existence of a solution of problem (6.28) and that any two solutions of this problem differ by a constant.

We are going to present two different variational approaches of independent interest. The first one consists in working with the space

$$V = \left\{ v \in H^1(\Omega) \ : \ \int_\Omega v(x)\,dx = 0 \right\},$$

and we will see that, in this framework, the problem becomes coercive. The second approach consists in regularizing problem (6.28) by considering for each $\varepsilon > 0$, the now coercive problem

$$\begin{cases} \varepsilon u_\varepsilon - \Delta u_\varepsilon = f & \text{on} \quad \Omega, \\ \frac{\partial u_\varepsilon}{\partial n} = 0 & \text{on} \quad \partial\Omega. \end{cases}$$

Then, by passing to the limit as $\varepsilon \to 0$, one obtains that the sequence $(u_\varepsilon)_{\varepsilon\to 0}$ converges to a particular solution of the initial problem (6.28).

Theorem 6.2.3. *Let us give $f \in L^2(\Omega)$ such that $\int_\Omega f(x)\,dx = 0$. We introduce the space $V = \{v \in H^1(\Omega) \ : \ \int_\Omega v(x)\,dx = 0\}$ which is equipped with the scalar product and norm of $H^1(\Omega)$, i.e.,*

$$\langle u, v\rangle = \int_\Omega (uv + \nabla u \cdot \nabla v)\,dx,$$

which makes V a closed subspace of $H^1(\Omega)$ and hence a Hilbert space.

(i) *There exists a unique $u \in V$ which satisfies*

$$\begin{cases} \displaystyle\int_\Omega \nabla u \cdot \nabla v \, dx = \int_\Omega f v \, dx \quad \forall v \in V, \\ u \in V. \end{cases} \tag{6.29}$$

Equivalently, u is the unique solution of the minimization problem

$$\min \left\{ \frac{1}{2} \int_\Omega |\nabla v|^2 \, dx - \int_\Omega f v \, dx \ : \ v \in V \right\}. \tag{6.30}$$

(ii) *Let us assume that the solution u of* (6.29) *is regular. Then u is a classical solution of the problem*

$$\begin{cases} -\Delta u = f & \text{on} \quad \Omega, \\ \frac{\partial u}{\partial n} = 0 & \text{on} \quad \partial\Omega, \\ \displaystyle\int_\Omega u(x) \, dx = 0, \end{cases} \tag{6.31}$$

and all the other classical solutions of (6.28) *are obtained by adding a constant to u.*

(iii) *Conversely if u is a classical solution of* (6.28)*, then $u - \frac{1}{|\Omega|} \int_\Omega u(x) \, dx$ is equal to the solution of problem* (6.29)*.*

PROOF. (i) The central point is to prove that the bilinear form $a : V \times V \to \mathbf{R}$, which is defined by

$$a(u, v) = \int_\Omega \nabla u \cdot \nabla v \, dx,$$

is coercive on V. Indeed,

$$\forall v \in V \quad a(v, v) = \int_\Omega |\nabla v|^2 \, dx; \tag{6.32}$$

so, we need some Poincaré inequality to conclude that a is coercive. In our setting, it is the Poincaré–Wirtinger inequality (cf. Corollary 5.4.1) which is well adapted. Let us recall that there exists some constant C such that

$$\forall v \in H^1(\Omega) \quad \left\| v - \frac{1}{|\Omega|} \int_\Omega v(x) \, dx \right\|_{L^2(\Omega)} \leq C \|\nabla v\|_{L^2(\Omega)^N}.$$

When $v \in V$, we have $\int_\Omega v(x) \, dx = 0$, and hence

$$\forall v \in V \quad \|v\|_{L^2(\Omega)} \leq C \|\nabla v\|_{L^2(\Omega)^N}. \tag{6.33}$$

From (6.32) and (6.33) we obtain

$$\forall v \in V \quad a(v, v) \geq \frac{1}{1 + C^2} \|v\|^2_{H^1(\Omega)}, \tag{6.34}$$

and a is coercive.

(ii) Let us now interpret the variational problem (6.29) in a classical sense when its solution u is regular. To that end, let us notice that

$$\forall v \in \mathcal{D}(\bar{\Omega}) \quad v - \frac{1}{|\Omega|}\int_\Omega v(x)\,dx \in V.$$

By using such test functions in (6.29) and by integration by parts, we obtain

$$\forall v \in \mathcal{D}(\bar{\Omega}) \quad \int_\Omega \nabla u \cdot \nabla v\,dx = \int_\Omega f\Big(v - \frac{1}{|\Omega|}\int_\Omega v(x)\,dx\Big)\,dx. \tag{6.35}$$

It is only now that we use the information $\int_\Omega f(x)\,dx = 0$ to obtain

$$\forall v \in \mathcal{D}(\bar{\Omega}) \quad \int_\Omega \nabla u \cdot \nabla v\,dx = \int_\Omega f v\,dx. \tag{6.36}$$

The end of the proof runs as before and we obtain that u satisfies (6.31).

If u^* is another classical solution of (6.29), then

$$\begin{cases} -\Delta(u - u^*) = 0 & \text{on } \Omega, \\ \frac{\partial}{\partial n}(u - u^*) = 0. \end{cases}$$

Let us multiply the first equation by $u - u^*$ and integrate by parts. We obtain

$$\int_\Omega |\nabla(u - u^*)|^2\,dx = 0$$

and, hence, $u - u^* \equiv C$ for some constant C.

(iii) Conversely, let u be a classical solution of (6.28). Then, $u - \frac{1}{|\Omega|}\int_\Omega u(x)\,dx := u^*$ belongs to V and satisfies

$$\begin{cases} -\Delta u^* = f & \text{on} \quad \Omega, \\ \frac{\partial u^*}{\partial n} = 0 & \text{on} \quad \partial\Omega. \end{cases}$$

Let us multiply the first equation by $v \in \mathcal{D}(\bar{\Omega})$ and integrate by parts

$$\forall v \in \mathcal{D}(\bar{\Omega}) \quad \int_\Omega \nabla u^* \cdot \nabla v\,dx = \int_\Omega f v\,dx.$$

By density of $\mathcal{D}(\bar{\Omega})$ in $H^1(\Omega)$ we obtain

$$\forall v \in H^1(\Omega) \quad \int_\Omega \nabla u^* \cdot \nabla v\,dx = \int_\Omega f v\,dx,$$

and since $V \subset H^1(\Omega)$ we obtain (6.29), i.e., u^* is the unique solution of problem (6.29). □

Let us now describe the other approach, which is an illustration of the so-called Tikhonov regularization method.

Theorem 6.2.4. *Let $f \in L^2(\Omega)$ be given such that $\int_\Omega f(x)\,dx = 0$. For any $\varepsilon > 0$, let u_ε be the unique solution of the variational problem*

$$\begin{cases} \displaystyle\int_\Omega (\varepsilon u_\varepsilon v + \nabla u_\varepsilon \cdot \nabla v)\,dx = \int_\Omega f v\,dx \quad \forall v \in H^1(\Omega), \\ u_\varepsilon \in H^1(\Omega). \end{cases} \tag{6.37}$$

Equivalently, u_ε is the weak solution of the homogeneous Neumann problem

$$\begin{cases} \varepsilon u_\varepsilon - \Delta u_\varepsilon = f & \text{on} \quad \Omega, \\ \frac{\partial u_\varepsilon}{\partial n} = 0 & \text{on} \quad \partial\Omega. \end{cases}$$

Then, the family $(u_\varepsilon)_{\varepsilon\to 0}$ norm converges in $H^1(\Omega)$, as $\varepsilon \to 0$, to a function $u \in H^1(\Omega)$, which is the solution of problem (6.29)*:*

$$\begin{cases} \displaystyle\int_\Omega \nabla u \cdot \nabla v\,dx = \int_\Omega f v\,dx \quad \forall v \in V, \\ u \in V, \end{cases}$$

where $V = \{v \in H^1(\Omega) \,:\, \int_\Omega v(x)\,dx = 0\}$. Equivalently, u is the weak solution of

$$\begin{cases} -\Delta u = f & \text{on} \quad \Omega, \\ \frac{\partial u}{\partial n} = 0 & \text{on} \quad \partial\Omega, \\ \int_\Omega u(x)\,dx = 0. \end{cases}$$

Proof. For $\varepsilon > 0$, take $a_0 \equiv \varepsilon$. We are in the coercive situation described by Theorem 6.2.1. Therefore, problem (6.37) admits a unique solution u_ε. Let us first prove that the family $(u_\varepsilon)_{\varepsilon\to 0}$ remains bounded in $H^1(\Omega)$. By taking $v = u_\varepsilon$ in (6.37), we obtain

$$\varepsilon \int_\Omega u_\varepsilon^2\,dx + \int_\Omega |\nabla u_\varepsilon|^2\,dx = \int_\Omega f u_\varepsilon\,dx. \tag{6.38}$$

By taking $v \equiv 1$ in (6.37), and using that $\int_\Omega f(x)\,dx = 0$, we obtain

$$\int_\Omega u_\varepsilon(x)\,dx = 0. \tag{6.39}$$

Let us invoke again the Poincaré–Wirtinger inequality (Corollary 5.4.1), which, in particular, implies the existence of a constant $C > 0$ such that

$$\forall v \in V \quad \|v\|_{L^2(\Omega)} \le C\left(\int_\Omega |\nabla v|^2\,dx\right)^{1/2}. \tag{6.40}$$

Combining (6.38), (6.39), and (6.40) we obtain

$$\begin{aligned} \int_\Omega u_\varepsilon^2\,dx &\le C^2 \int_\Omega |\nabla u_\varepsilon|^2\,dx \\ &\le C^2\left(\int_\Omega f^2\,dx\right)^{1/2}\left(\int_\Omega u_\varepsilon^2\,dx\right)^{1/2} \end{aligned}$$

and hence,

$$\left(\int_\Omega u_\varepsilon^2\,dx\right)^{1/2} \le C^2\left(\int_\Omega f^2\,dx\right)^{1/2}.$$

Returning to (6.38), we get

$$\int_\Omega |\nabla u_\varepsilon|^2\,dx \le C^2\int_\Omega f^2\,dx.$$

Finally,

$$\|u_\varepsilon\|_{H^1(\Omega)} \le C(1+C)\|f\|_{L^2(\Omega)}. \tag{6.41}$$

Let us extract a subsequence (which we still denote by (u_ε)) which weakly converges in $H^1(\Omega)$ to some $u^* \in H^1(\Omega)$. When passing to the limit on (6.37) we immediately obtain

$$\begin{cases} \displaystyle\int_\Omega \nabla u^*\cdot\nabla v\,dx = \int_\Omega f v\,dx \quad \forall v \in H^1(\Omega), \\ u^* \in V. \end{cases} \tag{6.42}$$

The problem (6.42) clearly admits at most a solution: take two solutions u_1^* and u_2^*, make the difference, and take $v = u_1^* - u_2^*$; one obtains $u_1^* - u_2^* =$ constant and $u_1^* - u_2^* \in V$, which implies $u_1^* = u_2^*$.

Hence, the whole sequence $(u_\varepsilon)_{\varepsilon\to 0}$ weakly converges to the unique solution u^* of (6.42), and $u^* = u$ which is the solution that we obtained in Theorem 6.2.3 by a different argument. Let us complete the argument by noticing that by passing to the limit on (6.38)

$$\lim_{\varepsilon\to 0}\int_\Omega |\nabla u_\varepsilon|^2\,dx = \int_\Omega fu.$$

On the other hand, by taking $v = u$ in (6.29) we have

$$\int_\Omega fu = \int_\Omega |\nabla u|^2\,dx.$$

Hence $u_\varepsilon \to u$ in $w - H^1(\Omega)$ and $\|u_\varepsilon\|_{H^1} \to \|u\|_{H^1}$. This implies that $u_\varepsilon \to u$ strongly in $H^1(\Omega)$. □

6.2.4 The semicoercive nonhomogeneous Neumann problem

Given $f : \Omega \to \mathbf{R}$ and $g : \partial\Omega \to \mathbf{R}$, we are looking for a solution $u : \bar\Omega \to \mathbf{R}$ of the following boundary value problem:

$$\begin{cases} -\Delta u = f & \text{on} \quad \Omega, \\ \frac{\partial u}{\partial n} = g & \text{on} \quad \partial\Omega. \end{cases}$$

Theorem 6.2.5. *Let $f \in L^2(\Omega)$ and $g \in L^2(\partial\Omega)$ be given functions which satisfy the so-called compatibility condition:*

$$\int_\Omega f(x)\,dx + \int_{\partial\Omega} g(x)\,d\sigma(x) = 0. \tag{6.43}$$

Set

$$V = \left\{ v \in H^1(\Omega) \ : \ \int_\Omega v(x)\,dx = 0 \right\}.$$

Then, there exists a unique solution u of the problem

$$\begin{cases} \displaystyle\int_\Omega \nabla u \cdot \nabla v\,dx = \int_\Omega f v\,dx + \int_{\partial\Omega} g v\,d\sigma \quad \forall v \in V, \\ u \in V. \end{cases} \tag{6.44}$$

Equivalently, u is the unique solution of the minimization problem

$$\min \left\{ \frac{1}{2}\int_\Omega |\nabla v|^2\,dx - \int_\Omega f v\,dx - \int_{\partial\Omega} g v\,d\sigma \ : \ v \in V \right\}. \tag{6.45}$$

The function u is a weak solution of the following nonhomogeneous Neumann boundary value problem

$$\begin{cases} -\Delta u = f & \text{on} \quad \Omega, \\ \frac{\partial u}{\partial n} = g & \text{on} \quad \partial\Omega. \end{cases} \tag{6.46}$$

Proof. Most of the ingredients of the proof were introduced in the previous sections. Therefore, we just briefly sketch the main lines of the proof. The only difference with the semicoercive homogeneous case comes from the linear form $l : V \to \mathbf{R}$

$$l(v) = \int_\Omega f v\,dx + \int_{\partial\Omega} g v\,d\sigma.$$

The continuity of l follows from the continuity of the trace operator γ_0 from $H^1(\Omega)$ into $L^2(\partial\Omega)$. The only point which deserves particular attention is the interpretation of (6.44) as a boundary value problem.

Let us assume that the solution u of (6.44) is regular. Given an arbitrary $v \in \mathcal{D}(\bar{\Omega})$, let us notice that $v - \frac{1}{|\Omega|}\int_\Omega v(x)\,dx$ belongs to V. Taking such a test function in (6.44), we infer that

$$\int_\Omega \nabla u \cdot \nabla v\,dx = \int_\Omega f v\,dx + \int_{\partial\Omega} g v\,d\sigma - \mathcal{M}(v)\left[\int_\Omega f(x)\,dx + \int_{\partial\Omega} g(x)\,d\sigma(x)\right],$$

where we set $\mathcal{M}(v) := \frac{1}{|\Omega|}\int_\Omega v(x)\,dx$. By using (6.43) we obtain

$$\forall v \in \mathcal{D}(\bar{\Omega}) \quad \int_\Omega \nabla u \cdot \nabla v\,dx = \int_\Omega f v\,dx + \int_{\partial\Omega} g v\,d\sigma,$$

and we conclude by using arguments similar to the ones in the proof of Theorem 6.2.2. □

Remark 6.2.2. (1) We have described two different methods to reduce the semicoercive Neumann problem to the coercive one. Each time, the idea is to reduce the problem to a situation where one can apply some Poincaré inequality. As another possibility let us mention (it is a good exercice to develop it) the parallel approach which consists in working with the space W (instead of V) defined by

$$W = \left\{ v \in H^1(\Omega) \ : \ \int_{\partial\Omega} v(x)\,d\sigma(x) = 0 \right\}.$$

(2) The term *semicoercive* comes from the fact that the lack of coercivity concerns only the lower-order terms (here only the zero-order term, namely, u!).

6.3 Mixed Dirichlet–Neumann problems

We are going to consider boundary value problems whose boundary conditions contain both u and $\frac{\partial u}{\partial n}$.

6.3.1 The Dirichlet-Neumann problem

Let Ω be an open bounded set in $\mathbf{R}^N$ which is assumed to be connected and regular. (Its boundary $\Gamma = \partial\Omega$ is piecewise of class $\mathbf{C}^1$.) Let us suppose that $\Gamma = \Gamma_0 \cup \Gamma_1$, $\Gamma_0 \cap \Gamma_1 = \emptyset$ with $H^{N-1}(\Gamma_0) > 0$, where Γ is the union of the two disjoint sets Γ_0 and Γ_1.

Given $f \in L^2(\Omega)$ and $g \in L^2(\Gamma_1)$ we are looking for a solution u of the following boundary value problem:

$$\begin{cases} -\Delta u = f & \text{on} \quad \Omega, \\ u = 0 & \text{on} \quad \Gamma_0, \\ \frac{\partial u}{\partial n} = g & \text{on} \quad \Gamma_1. \end{cases}$$

Note that two different types of boundary conditions are imposed to u: on Γ_0 it is a Dirichlet condition, while on the complementary $\Gamma_1 = \Gamma \setminus \Gamma_0$ it is a Neumann condition. This is a simplified model for an important situation in mechanics, where an elastic material is fixed on a part Γ_0 of its boundary, while on the complementary a surface density of force g is present.

Let us introduce the functional space

$$V := \{v \in H^1(\Omega) \ : \ \gamma_0(v) = 0 \text{ on } \Gamma_0\}. \tag{6.47}$$

The space V, as a subspace of $H^1(\Omega)$, is equipped with the scalar product of $H^1(\Omega)$ and the corresponding norm

$$\|v\|_V = \left(\int_\Omega (v^2 + |\nabla v|^2)\, dx \right)^{1/2}.$$

It follows immediately from the continuity of the trace operator $\gamma_0 : H^1(\Omega) \to L^2(\Gamma)$ that V is a closed subspace of $H^1(\Omega)$. Note that $\gamma_0(v) = 0$ in Γ_0 has to be understood in the following sense: $\gamma_0(v)(x) = 0$ for a.e. x with respect to the measure $H^{N-1}|\Gamma_0$. Hence V is a Hilbert space. We can now state the variational formulation of the Dirichlet–Neumann problem.

Theorem 6.3.1. (i) *Let $f \in L^2(\Omega)$ and $g \in L^2(\Gamma_1)$ be given functions. Let us assume moreover that $H^{N-1}(\Gamma_0) > 0$. Then, there exists a unique solution u of the problem*

$$\begin{cases} \displaystyle\int_\Omega \nabla u \cdot \nabla v \, dx = \int_\Omega f v \, dx + \int_{\Gamma_1} g v \, d\sigma \quad \forall v \in V, \\ u \in V. \end{cases} \tag{6.48}$$

Equivalently, u is the unique solution of the minimization problem

$$\min\left\{\frac{1}{2}\int_\Omega |\nabla v|^2\,dx - \int_\Omega fv\,dx - \int_{\Gamma_1} gv\,d\sigma \;:\; v\in V\right\}. \tag{6.49}$$

(ii) *Assume that Γ_0 is sufficiently regular to have the property*

$$\left\{v|_{\Gamma_1} \;:\; v\in\mathcal{D}(\bar\Omega),\ v=0 \text{ on } \Gamma_0\right\} \quad \text{is dense in } L^2(\Gamma_1). \tag{6.50}$$

Then, the solution u of the variational problem (6.48) *is a weak solution of the following Dirichlet–Neumann boundary value problem:*

$$\begin{cases} -\Delta u = f & \text{on} \quad \Omega,\\ u = 0 & \text{on} \quad \Gamma_0,\\ \frac{\partial u}{\partial n} = g & \text{on} \quad \Gamma_1. \end{cases} \tag{6.51}$$

PROOF. (i) Clearly, the bilinear form

$$a(u,v) = \int_\Omega \nabla u\cdot\nabla v\,dx$$

is continuous on $V\times V$. The coercivity of a is a consequence of the generalized Poincaré inequality (Theorem 5.4.3): we have seen that V is a closed subspace of $H^1(\Omega)$. On the other hand, the assumption $H^{N-1}(\Gamma_0)>0$ implies that the only constant function belonging to V is the zero function. Hence, there exists some $C>0$ such that

$$\forall v\in V \quad \|v\|_{L^2(\Omega)} \le C\left(\int_\Omega |\nabla v|^2\,dx\right)^{1/2}.$$

This immediately implies that

$$\forall v\in V \quad a(v,v) \ge \frac{1}{1+C^2}\|v\|^2_{H^1(\Omega)}.$$

Let us now consider the linear form $l: V\to\mathbf{R}$ defined by

$$l(v) = \int_\Omega fv\,dx + \int_{\Gamma_1} gv\,d\sigma.$$

The continuity of the trace operator $\gamma_0: H^1(\Omega)\to L^2(\Gamma)$ (Theorem 5.6.1) clearly implies that l is continuous.

(ii) Let us now suppose that the solution u of the variational problem (6.48) is smooth, and let us prove that u is a classical solution of (6.51).

Let us first take $v\in\mathcal{D}(\Omega)$ (note that $\mathcal{D}(\Omega)\subset V$). We obtain

$$\int_\Omega \nabla u\cdot\nabla v\,dx = \int_\Omega fv\,dx \quad \forall v\in\mathcal{D}(\Omega),$$

which classically implies

$$-\Delta u = f \text{ on } \Omega. \tag{6.52}$$

Let us now take $v \in \mathcal{D}(\bar{\Omega})$ such that $v = 0$ on Γ_0. Since $v \in V$, we have, by (6.48),

$$\int_\Omega \nabla u \cdot \nabla v \, dx = \int_\Omega f v \, dx + \int_{\Gamma_1} g v \, d\sigma \quad \forall v \in \mathcal{D}(\bar{\Omega}),\ v = 0 \text{ on } \Gamma_0.$$

After integration by parts, and by using (6.52), we obtain

$$\int_{\Gamma_1} \left(\frac{\partial u}{\partial n} - g \right) v \, d\sigma = 0 \qquad \forall v \in \mathcal{D}(\bar{\Omega}), v = 0 \text{ on } \Gamma_0. \tag{6.53}$$

From the regularity assumption (6.50) on Γ_0 and (6.53) we infer that

$$\frac{\partial u}{\partial n} = g \text{ on } \Gamma_1,$$

which completes the proof. □

Remark 6.3.1. (1) Conversely, when proving that a classical solution of (6.51) is a variational solution of (6.48), one needs to make the following assumption: $\mathcal{V}$ is dense in V for the $H^1(\Omega)$-norm, where

$$\mathcal{V} = \{v \in \mathcal{D}(\bar{\Omega}) \ : \ v = 0 \text{ on } \Gamma_0\}. \tag{6.54}$$

(2) The relationship between the classical and the variational solution for the Dirichlet–Neumann problem is quite delicate. Beside the classical regularity assumptions on Ω and its boundary $\Gamma = \partial\Omega$, it requires extra regularity assumptions on the portion Γ_0 of Γ, where the Dirichlet condition is imposed. Otherwise we could take some Γ_0, with $H^{N-1}(\Gamma_0) > 0$ which is dense in Γ; then, u regular and $u = 0$ on Γ_0 would force u to be equal to zero everywhere on Γ, which makes the above argumentation no more valid.

6.3.2 Mixed Dirichlet–Neumann boundary conditions

We still consider an open bounded regular set Ω in $\mathbf{R}^N$. Let us give some positive measurable function $a_0 : \partial\Omega \to \mathbf{R}^+$ such that there exists a positive real number $\alpha > 0$ with

$$a_0(x) \geq \alpha \text{ a.e. } x \in \partial\Omega \quad \text{with respect to } H^{N-1}|\partial\Omega. \tag{6.55}$$

Theorem 6.3.2. *Let $f \in L^2(\Omega)$ and $g \in L^2(\partial\Omega)$ be given functions and assume* (6.55).

(i) *Then, there exists a unique solution u of the following system:*

$$\begin{cases} \displaystyle\int_\Omega \nabla u \cdot \nabla v \, dx + \int_{\partial\Omega} a_0 u v \, d\sigma = \int_\Omega f v \, dx + \int_{\partial\Omega} g v \, d\sigma \quad \forall v \in H^1(\Omega). \\ u \in H^1(\Omega). \end{cases} \tag{6.56}$$

Equivalently, u is the unique solution of the minimization problem

$$\min \left\{ \frac{1}{2} \int_\Omega |\nabla v|^2 \, dx + \frac{1}{2} \int_{\partial\Omega} a_0 v^2 \, d\sigma - \int_\Omega f v \, dx - \int_{\partial\Omega} g v \, d\sigma \ : \ v \in H^1(\Omega) \right\}. \tag{6.57}$$

(ii) *The solution u of* (6.56) *is a weak solution of the following boundary value problem:*

$$\begin{cases} -\Delta u = f & \text{on} \quad \Omega, \\ a_0 u + \dfrac{\partial u}{\partial n} = g & \text{on} \quad \partial\Omega. \end{cases} \tag{6.58}$$

PROOF. (i) The only point which is not standard is to verify that the bilinear form

$$a(u, v) = \int_\Omega \nabla u \cdot \nabla v \, dx + \int_{\partial\Omega} a_0 u v \, d\sigma$$

is coercive on $H^1(\Omega)$. By assumption (6.55) we have

$$\forall v \in H^1(\Omega) \quad a(v, v) \geq \int_\Omega |\nabla v|^2 \, dx + \alpha \int_{\partial\Omega} v^2 \, d\sigma. \tag{6.59}$$

Let us apply the generalized Poincaré inequality (Theorem 5.4.3) to the space

$$V = \left\{ v \in H^1(\Omega) \ : \ \int_{\partial\Omega} v(x) \, d\sigma(x) = 0 \right\}, \tag{6.60}$$

where, as usual, for simplicity of the notation, we write v instead of $\gamma_0(v)$. By using again the continuity of the trace operator $\gamma_0 : H^1(\Omega) \to L^2(\partial\Omega)$, we have that V is a closed subspace of $H^1(\Omega)$. Moreover, if v is a constant function which belongs to V, say, $v \equiv C$, we necessarily have $C H^{N-1}(\partial\Omega) = 0$ which forces C to be equal to zero.

Let us finally observe that

$$\forall v \in H^1(\Omega) \quad v - \mathcal{M}_{\partial\Omega}(v) \in V,$$

where $\mathcal{M}_{\partial\Omega}(v) := \frac{1}{|\partial\Omega|} \int_{\partial\Omega} v(x) \, d\sigma(x)$. Thus, by applying the generalized Poincaré inequality to the space V, we obtain the existence of a positive constant C such that

$$\forall v \in H^1(\Omega) \quad \|v - \mathcal{M}_{\partial\Omega}(v)\|_{L^2(\Omega)} \leq C \left(\int_\Omega |\nabla v|^2 \, dx \right)^{1/2}. \tag{6.61}$$

From (6.61) we easily obtain the inequality

$$\forall v \in H^1(\Omega) \quad \int_\Omega v(x)^2 \, dx \leq 2C^2 \int_\Omega |\nabla v|^2 \, dx + 2|\Omega| \mathcal{M}_{\partial\Omega}(v)^2. \tag{6.62}$$

On the other hand, by using the Cauchy–Schwarz inequality in $L^2(\partial\Omega)$, we obtain

$$\mathcal{M}_{\partial\Omega}(v) = \frac{1}{|\partial\Omega|} \int_{\partial\Omega} v(x) \, d\sigma(x) \leq \frac{1}{|\partial\Omega|^{1/2}} \left(\int_{\partial\Omega} v(x)^2 \, d\sigma(x) \right)^{1/2}. \tag{6.63}$$

Combining (6.62) and (6.63), we obtain

$$\forall v \in H^1(\Omega) \quad \int_\Omega v(x)^2 \, dx \leq 2C^2 \int_\Omega |\nabla v|^2 \, dx + 2 \frac{|\Omega|}{|\partial\Omega|} \int_{\partial\Omega} v(x)^2 \, d\sigma(x). \tag{6.64}$$

It is an elementary computation to obtain, by using (6.59) and (6.64), the inequality

$$\forall v \in H^1(\Omega) \quad a(v,v) \geq \gamma \|v\|^2_{H^1(\Omega)} \tag{6.65}$$

with

$$\gamma = \frac{\min\{\alpha, 1\}}{1 + 2\max\{C^2, \frac{|\Omega|_N}{|\partial\Omega|_{N-1}}\}}.$$

(ii) Let us assume that the solution u of (6.56) is regular. Then, by taking $v \in \mathcal{D}(\bar{\Omega})$ in (6.56) and integrating by parts, we obtain

$$\int_\Omega (-\Delta u - f)dx + \int_{\partial\Omega} \left(a_0 u + \frac{\partial u}{\partial n} - g\right) v \, d\sigma = 0 \quad \forall v \in \mathcal{D}(\bar{\Omega}).$$

An analysis similar to the one used in the case of the Neumann problem gives that u satisfies (6.58). □

Remark 6.3.2. It is interesting to notice that the boundary condition $a_0 u + \frac{\partial u}{\partial n} = g$ contains, at least formally, all the previous boundary conditions that we have examined.

(i) Take $a_0 = 0$; then one obtains the Neumann boundary condition $\frac{\partial u}{\partial n} = g$. Note that in this case, the coercivity property is lost.

(ii) Take $a_0 \equiv +\infty$; then formally one obtains the Dirichlet boundary condition, $u = 0$ on $\partial\Omega$.

(iii) Take $a_0 = 0$ on Γ_1 and $a_0 = +\infty$ on $\Gamma_0 = \Gamma \setminus \Gamma_0$; then one obtains the Dirichlet–Neumann boundary condition.

Indeed, it is a good exercice to justify these formal results. For example, take $a_0(x) \equiv n$ in (6.56) and prove that the corresponding sequence (u_n) norm-converges to the solution u of the Dirichlet problem.

6.4 Heterogenous media: Transmission conditions

Let us consider the following model situation coming from electrostatics. In the open set Ω of $\mathbf{R}^N$ we have two distinct materials with respective conductivity coefficients α and β; take $0 < \alpha < \beta < +\infty$, for example. The subsets of Ω occupied by the two materials are denoted, respectively, by Ω_α and Ω_β. We assume that Ω_α and Ω_β are open sets with a common boundary denoted by Σ which is a $\mathbf{C}^1$ manifold and

$$\Omega = \Omega_\alpha \cup \Omega_\beta \cup \Sigma.$$

In this model, the conductivity coefficient $a : \Omega \to \mathbf{R}$ takes only two values, $a \equiv \alpha$ on Ω_α and $a \equiv \beta$ on Ω_β:

$$a(x) = \begin{cases} \alpha & \text{if} \quad x \in \Omega_\alpha, \\ \beta & \text{if} \quad x \in \Omega_\beta. \end{cases}$$

Note that a is not continuous (it is discontinuous through Σ). It belongs to $L^\infty(\Omega)$. Indeed, in the following developments, $a(\cdot)$ plays a role only as a function defined a.e. with respect to the Lebesgue measure. Since Σ has zero Lebesgue measure, we don't need to define a on Σ.

For each $x \in \Sigma$, $n(x)$ is the unit normal vector to Σ at x which is directed outward with respect to Ω_α (and inward with respect to Ω_β).

For any (electrostatic) potential function $v : \Omega \to \mathbf{R}$ the corresponding stored internal energy is

$$\Phi(v) = \int_\Omega a(x)|\nabla v(x)|^2\,dx.$$

Suppose that the conductor is connected to the earth on its boundary. For a given density of charge $f : \Omega \to \mathbf{R}$ the equilibrium potential function $u : \bar{\Omega} \to \mathbf{R}$ solves the minimization problem

$$\min\left\{\frac{1}{2}\int_\Omega a(x)|\nabla v(x)|^2\,dx - \int_\Omega f(x)v(x)\,dx \;:\; v = 0 \text{ on } \partial\Omega\right\}.$$

The variational approach to the above problem, and the corresponding transmission conditions through the interface Σ satisfied by the solution u, are described in the following statement.

Theorem 6.4.1. (a) *For every $f \in L^2(\Omega)$ there exists a unique $u \in H_0^1(\Omega)$ solution of the following minimization problem:*

$$\min\left\{\frac{1}{2}\int_\Omega a(x)|\nabla v(x)|^2\,dx - \int_\Omega f(x)v(x)\,dx \;:\; v \in H_0^1(\Omega)\right\}. \tag{6.66}$$

(b) *Equivalently, the solution u of* (6.66) *verifies*

$$\begin{cases} \displaystyle\int_\Omega a(x)\nabla u(x)\cdot\nabla v(x)\,dx = \int_\Omega f(x)v(x)\,dx \quad \forall v \in H_0^1(\Omega), \\ u \in H_0^1(\Omega). \end{cases} \tag{6.67}$$

(c) *Equivalently, the solution u of* (6.66) *is a weak solution of the boundary value problem*

$$\begin{cases} -\operatorname{div}(a(x)\nabla u(x)) = f \text{ on } \Omega, \\ u = 0 \text{ on } \partial\Omega, \end{cases} \tag{6.68}$$

in the following sense: $u \in H_0^1(\Omega)$, the first equation is satisfied in the distribution sense, while $u = 0$ is satisfied in the trace sense.

(d) *Let us denote $u_\alpha = u|_{\Omega_\alpha}$ and $u_\beta = u|_{\Omega_\beta}$, and let us assume that the variational (weak) solution of* (6.66) *is regular in the following sense: $u_\alpha \in \mathbf{C}^2(\bar{\Omega}_\alpha)$ and $u_\beta \in \mathbf{C}^2(\bar{\Omega}_\beta)$. Then, u is a classical solution of the following transmission problem:*

$$\begin{cases} -\alpha\Delta u_\alpha = f & \text{on } \Omega_\alpha, \\ -\beta\Delta u_\beta = f & \text{on } \Omega_\beta, \\ u_\alpha = u_\beta & \text{on } \Sigma, \\ \alpha\frac{\partial u_\alpha}{\partial n} = \beta\frac{\partial u_\beta}{\partial n} & \text{on } \Sigma, \\ u = 0 & \text{on } \partial\Omega. \end{cases} \tag{6.69}$$

In particular, u is continuous through Σ, but $\frac{\partial u}{\partial n}$ is discontinuous through Σ.

(e) *Conversely, if u is a classical solution, i.e., $u_\alpha \in \mathbf{C}^2(\bar{\Omega}_\alpha)$, $u_\beta \in \mathbf{C}^2(\bar{\Omega}_\beta)$, and u satisfies* (6.69)*, then it is equal to the variational (weak) solution of* (6.66)*.*

PROOF. (a) The argument is quite similar to the one developed in the variational approach to the Dirichlet problem. The functional $J : H_0^1(\Omega) \to \mathbf{R}$ defined by

$$J(v) := \frac{1}{2}\int_\Omega a(x)|\nabla v(x)|^2\,dx - \int_\Omega f v\,dx$$

is convex and continuous on $H_0^1(\Omega)$. The coercivity of J follows from the inequality

$$J(v) \geq \frac{\min(\alpha,\beta)}{2}\int_\Omega |\nabla v|^2\,dx - \|f\|_{L^2}\|v\|_{L^2}$$

and the Poincaré inequality on $H_0^1(\Omega)$.

The conditions of the convex minimization theorem, Theorem 3.3.4, are fulfilled. Hence, there exists a solution u to problem (6.66). It is unique because J is strictly convex. Indeed, this follows from the strict convexity of $\Phi(v) = \int a(x)|\nabla v|^2\,dx$ which is a positive definite quadratic form on $H_0^1(\Omega)$ (see Proposition 2.3.4).

(b) The equivalence between (6.66) and (6.67) is a direct consequence of Proposition 2.3.1 and the fact that the bilinear form

$$b(u,v) := \int_\Omega a(x)\nabla u(x)\cdot\nabla v(x)\,dx$$

is symmetric and positive. Indeed, we could as well solve the variational problem by applying the Lax–Milgram theorem, Theorem 3.1.2, to the formulation (6.67).

(c) It follows from the density of $\mathcal{D}(\Omega)$ in $H_0^1(\Omega)$ that it is equivalent in the variational formulation (6.67) to take only v belonging to $\mathcal{D}(\Omega)$. One then obtains the equivalent formulation (6.68) just by using the notion of derivation in the distribution sense.

(d) Let us now come to the point which deserves some particular attention, namely, the interpretation of the distribution formula

$$-\operatorname{div}(a(x)\nabla u(x)) = f \text{ on } \Omega.$$

As we just pointed out, all the information we have on u is contained in this distribution formula and the fact that $u \in H_0^1(\Omega)$. Equivalently, we have

$$\begin{cases} \displaystyle\int_\Omega a(x)\nabla u(x)\cdot\nabla v(x)\,dx - \int_\Omega f(x)v(x)\,dx = 0 \quad \forall v \in \mathcal{D}(\Omega), \\ u \in H_0^1(\Omega). \end{cases} \tag{6.70}$$

Let us particularize the test function v to test successively u on Ω_α, Ω_β and then on Σ. We assume that $u_\alpha \in \mathbf{C}^2(\bar{\Omega}_\alpha)$ and $u_\beta \in \mathbf{C}^2(\bar{\Omega}_\beta)$.

(1) Let us first take $v \in \mathcal{D}(\Omega_\alpha)$. From (6.70) and the fact that $a \equiv \alpha$ on Ω_α, we obtain that $u_\alpha = u|_{\Omega_\alpha}$ satisfies

$$\alpha\int_{\Omega_\alpha} \nabla u_\alpha\cdot\nabla v\,dx - \int_{\Omega_\alpha} f v\,dx = 0 \quad \forall v \in \mathcal{D}(\Omega_\alpha).$$

This yields

$$-\alpha \Delta u_\alpha = f \text{ on } \Omega_\alpha. \tag{6.71}$$

(2) Similarly, by taking test functions $v \in \mathcal{D}(\Omega_\beta)$ one obtains

$$-\beta \Delta u_\beta = f \text{ on } \Omega_\beta. \tag{6.72}$$

(3) Let us now analyze the transmission conditions through Σ which are satisfied by u.

Let us first observe that $u_\alpha = u_\beta$ on Σ. Indeed, since $u \in H_0^1(\Omega)$ there exists some approximating sequence $(u_n)_{u\in\mathbf{N}}$, $u_n \in \mathcal{D}(\Omega)$ for each $n \in \mathbf{N}$ such that $u_n \to u$ in $H^1(\Omega)$ as $n \to +\infty$. Let us identify u_n and its extension by zero outside of Ω. By definition of $\mathcal{D}(\bar{\Omega}_\alpha)$ and $\mathcal{D}(\bar{\Omega}_\beta)$ we have that for every $n \in \mathbf{N}$,

$$u_n|_{\Omega_\alpha} \in \mathcal{D}(\bar{\Omega}_\alpha) \quad \text{and} \quad u_n|_{\Omega_\beta} \in \mathcal{D}(\bar{\Omega}_\beta).$$

Moreover,

$$u_n|_{\Omega_\alpha} \to u_\alpha \text{ in } H^1(\Omega_\alpha),$$

$$u_n|_{\Omega_\beta} \to u_\beta \text{ in } H^1(\Omega_\beta).$$

By using the continuity of the trace operator $\gamma_{0,\alpha} : H^1(\Omega_\alpha) \to L^2(\Sigma)$ and the fact that for functions in $\mathcal{D}(\bar{\Omega}_\alpha)$ the trace coincides with the restriction, we obtain that

$$u_n|_\Sigma \to \gamma_{0,\alpha}(u_\alpha) \text{ in } L^2(\Sigma).$$

Similarly, we have

$$u_n|_\Sigma \to \gamma_{0,\beta}(u_\beta) \text{ in } L^2(\Sigma).$$

Hence $\gamma_{0,\alpha}(u_\alpha) = \gamma_{0,\beta}(u_\beta)$, that is, the traces of u from both sides of Σ are the same. Since u has been assumed to be $\mathbf{C}^2(\bar{\Omega}_\alpha)$ and $\mathbf{C}^2(\bar{\Omega}_\beta)$, these traces coincide with the respective values of u_α and u_β on Σ, that is,

$$u_\alpha = u_\beta \text{ on } \Sigma, \tag{6.73}$$

which makes the function u continuous on Ω.

Let us now take a general test function $v \in \mathcal{D}(\Omega)$ and rewrite (6.70) as

$$\alpha \int_{\Omega_\alpha} \nabla u_\alpha \cdot \nabla v \, dx + \beta \int_{\Omega_\beta} \nabla u_\beta \cdot \nabla v \, dx - \int_{\Omega_\alpha} f v \, dx - \int_{\Omega_\beta} f v \, dx = 0. \tag{6.74}$$

Integrating by parts gives

$$\alpha \int_{\Omega_\alpha} \nabla u_\alpha \cdot \nabla v \, dx = -\int_{\Omega_\alpha} (\alpha \Delta u_\alpha) v \, dx + \alpha \int_\Sigma v \frac{\partial u_\alpha}{\partial n} \, d\sigma. \tag{6.75}$$

Similarly (note that the outward normal to Ω_β on Σ is now the opposite vector $-\vec{n}$), we have

$$\beta \int_{\Omega_\beta} \nabla u_\beta \cdot \nabla v \, dx = -\int_{\Omega_\beta} (\beta \Delta u_\beta) v \, dx - \beta \int_\Sigma v \frac{\partial u_\beta}{\partial n} \, d\sigma. \tag{6.76}$$

Combining (6.74), (6.75), and (6.76) we obtain

$$-\int_{\Omega_\alpha}(\alpha\Delta u_\alpha+f)v\,dx-\int_{\Omega_\beta}(\beta\Delta u_\beta+f)v\,dx+\int_\Sigma\left(\alpha\frac{\partial u_\alpha}{\partial n}-\beta\frac{\partial u_\beta}{\partial n}\right)v\,d\sigma=0.$$

By using (6.71) and (6.72), we finally obtain

$$\forall v\in\mathcal{D}(\Omega)\quad\int_\Sigma\left(\alpha\frac{\partial u_\alpha}{\partial n}-\beta\frac{\partial u_\beta}{\partial n}\right)v(x)\,d\sigma(x)=0,\tag{6.77}$$

which implies

$$\alpha\frac{\partial u_\alpha}{\partial n}=\beta\frac{\partial u_\beta}{\partial n}\text{ on }\Sigma.\tag{6.78}$$

Let us finally notice that the condition $u=0$ on $\partial\Omega$ follows from the fact that $u\in H_0^1(\Omega)$ and the regularity assumptions on $\partial\Omega$ and u.

(e) To pass from the boundary value problem (6.69) to (6.67) we proceed in a similar way, just making the integration by parts in the reverse way. The only point which requires some attention is the proof of the property $u\in H^1(\Omega)$. Indeed, this is a consequence of the following lemma of independent interest. □

Lemma 6.4.1. *Let $u_\alpha\in H^1(\Omega_\alpha)$ and $u_\beta\in H^1(\Omega_\beta)$, Ω_α and Ω_β being two disjoint open sets with a common interface Σ. We define*

$$u=\begin{cases}u_\alpha\text{ on }\Omega_\alpha,\\ u_\beta\text{ on }\Omega_\beta.\end{cases}$$

Then the following derivation rule holds:

$$\nabla u=\chi_{\Omega_\alpha}\widetilde{\nabla u_\alpha}+\chi_{\Omega_\beta}\widetilde{\nabla u_\beta}+[u]_\Sigma\vec{n}\,d\sigma,$$

where $[u]_\Sigma:=\gamma_{0,\beta}(u_\beta)-\gamma_{0,\alpha}(u_\alpha)$ is the jump of u through Σ, $d\sigma=\mathcal{H}^{N-1}\lfloor\Sigma$, and $\widetilde{\nabla u_\alpha}$, $\widetilde{\nabla u_\beta}$ denote the extensions by zero of ∇u_α and ∇u_β, respectively, on $\Omega\setminus\Omega_\alpha$ and $\Omega\setminus\Omega_\beta$.

Proof. Take $1\le i\le N$ and compute $\frac{\partial u}{\partial x_i}$ in $\mathcal{D}'(\Omega)$:

$$\left\langle\frac{\partial u}{\partial x_i},\varphi\right\rangle_{(\mathcal{D}'(\Omega),\mathcal{D}(\Omega))}:=-\int_{\Omega_\alpha}u_\alpha\frac{\partial\varphi}{\partial x_i}\,dx-\int_{\Omega_\beta}u_\beta\frac{\partial\varphi}{\partial x_i}\,dx.$$

Let us integrate by parts by using the Green's formula (Proposition 5.6.2):

$$\begin{aligned}\left\langle\frac{\partial u}{\partial x_i},\varphi\right\rangle_{(\mathcal{D}',\mathcal{D})}&=\int_{\Omega_\alpha}\frac{\partial u_\alpha}{\partial x_i}\varphi\,dx-\int_\Sigma\gamma_{0,\alpha}(u_\alpha)\varphi n_i\,d\sigma+\int_{\Omega_\beta}\frac{\partial u_\beta}{\partial x_i}\varphi\,dx\\&\quad+\int_\Sigma\gamma_{0,\beta}(u_\beta)\varphi n_i\,d\sigma\\&=\int_\Omega\left(\widetilde{\frac{\partial u_\alpha}{\partial x_i}}\chi_{\Omega_\alpha}+\widetilde{\frac{\partial u_\beta}{\partial x_i}}\chi_{\Omega_\beta}\right)\varphi\,dx\\&\quad+\int_\Sigma\big(\gamma_{0,\beta}(u_\beta)-\gamma_{0,\alpha}(u_\alpha)\big)n_i\varphi\,d\sigma.\end{aligned}$$

Together with the equality

$$\frac{\partial u}{\partial x_i} = \widetilde{\frac{\partial u_\alpha}{\partial x_i}} \chi_{\Omega_\alpha} + \widetilde{\frac{\partial u_\beta}{\partial x_i}} \chi_{\Omega_\beta} + [u]_\Sigma n_i H^{N-1} | \Sigma,$$

this ends the proof. □

Remark 6.4.1. (1) We stress the fact that the transmission law through Σ is contained in the formula

$$-\operatorname{div}(a(x)\nabla u(x)) = f \quad \text{on } \Omega,$$

which holds in the distribution sense. Let us first notice that this formula makes sense. We have $u \in H_0^1(\Omega)$, $a \in L^\infty(\Omega)$. Hence, for each $i \in \mathbf{N}$, $a(x)\frac{\partial u}{\partial x_i} \in L^2(\Omega)$, which defines a distribution on Ω and $\frac{\partial}{\partial x_i}(a(x)\frac{\partial u}{\partial x_i})$ is well defined as a distribution. The point is that we have to treat $a(x)\frac{\partial u}{\partial x_i}$ as a block; we cannot derivate this product by using the classical derivation rule, because $a(\cdot) \in L^\infty(\Omega)$ and $\frac{\partial a}{\partial x_i}$ is a measure which is singular with respect to the Lebesgue measure. This fact reveals some of the strengths and weaknesses of the distribution theory. It is a powerful tool which allows us to formulate in a quite simple and unifying way a lot of phenomena. On the counterpart, the physical laws are often implicit and hidden in the distribution formulation.

(2) Taking $\Omega = (0,1)$, $\Omega_\alpha =]0, c[$, $\Omega_\beta =]c, 1[$, and $\Sigma = \{c\}$, the profile of the solution of the corresponding transmission problem corresponds to the classical Descartes law in geometrical optics, with the equality

$$\beta \frac{d^+ u}{dx}(c) = \alpha \frac{d^- u}{dx}(c).$$

6.5 Linear elliptic operators

Let Ω be an open subset of $\mathbf{R}^N$. Let us give a family $\{a_{ij}(\cdot) \,:\, 1 \le i, j \le n\}$ of functions belonging to $L^\infty(\Omega)$ and which satisfies the following condition: there exists some positive real number $\alpha > 0$ such that

$$\forall \xi \in \mathbf{R}^N \quad \sum_{i,j=1}^N a_{ij}(x)\xi_i\xi_j \ge \alpha |\xi|^2 \quad \text{a.e. on } \Omega. \tag{6.79}$$

Let us introduce the corresponding bilinear form $a : H^1(\Omega) \times H^1(\Omega) \to \mathbf{R}$,

$$\forall u, v \in H^1(\Omega) \quad a(u,v) := \int_\Omega \sum_{i,j=1}^N a_{ij}(x) \frac{\partial u}{\partial x_i} \frac{\partial v}{\partial x_j}\, dx. \tag{6.80}$$

Property (6.79) implies that the bilinear form enjoys a so-called ellipticity condition:

$$\begin{aligned} \forall v \in H^1(\Omega) \qquad a(v,v) &= \int_\Omega \sum_{i,j=1}^N a_{ij}(x) \frac{\partial v}{\partial x_i} \frac{\partial v}{\partial x_j}\, dx \\ &\ge \alpha \int_\Omega |\nabla v(x)|^2\, dx. \end{aligned} \tag{6.81}$$

By using (6.81) one can prove existence and uniqueness results for variational problems involving the bilinear form a. The proof is similar to the ones of the previous sections. Note that the case of the Dirichlet integral and of the corresponding Laplace equation corresponds to the particular situation $a_{ij}(x) = \delta_{ij}$.

Since we are now familiar with a number of boundary value problems (Dirichlet, Neumann, mixed Dirichlet–Neumann), let us give in the present situation a unified approach to these problems by taking as a functional space a closed subspace V of $H^1(\Omega)$ such that $H_0^1(\Omega) \subset V \subset H^1(\Omega)$.

Theorem 6.5.1. *Let us assume that Ω is a bounded open set in $\mathbf{R}^N$ which is regular (piecewise of class $\mathbf{C}^1$) and connected. Let V be a closed subspace of $H^1(\Omega)$ which satisfies conditions*

(a) $H_0^1(\Omega) \subset V \subset H^1(\Omega)$;

(b) $v \in V$, $v \equiv$ *constant* $\Longrightarrow v \equiv 0$.

On the other hand, let us give a family $(a_{ij})_{i,j=1,\dots,N}$ of $L^\infty(\Omega)$ functions which satisfies the ellipticity condition (6.79).

(i) *Then, for any $f \in L^2(\Omega)$ and any $g \in L^2(\partial\Omega)$, there exists a unique solution $u \in V$ of the problem*

$$\begin{cases} \displaystyle\int_\Omega \sum_{i,j=1}^N a_{ij}(x) \frac{\partial u}{\partial x_i} \frac{\partial v}{\partial x_j}\, dx = \int_\Omega f v\, dx + \int_{\partial\Omega} g v\, d\sigma \quad \forall v \in V, \\ u \in V. \end{cases} \tag{6.82}$$

(ii) *The solution u of* (6.82) *satisfies*

$$-\sum_{i,j=1}^N \frac{\partial}{\partial x_j}\left(a_{ij}(x) \frac{\partial u}{\partial x_i}\right) = f \text{ on } \Omega$$

in the distribution sense.

(iii) *When the matrix $(a_{ij}(x))_{1 \le i,j \le N}$ is symmetric,* (6.82) *is equivalent to saying that u is the unique solution of the minimization problem*

$$\min\left\{ \frac{1}{2} \int_\Omega \sum_{i,j=1}^N a_{ij}(x) \frac{\partial v}{\partial x_i} \frac{\partial v}{\partial x_j}\, dx - \int_\Omega f v\, dx - \int_{\partial\Omega} g v\, d\sigma \ : \ v \in V \right\}.$$

Proof. (i) To apply the Lax–Milgram theorem to problem (6.82), we are going to work in the space V which is considered as a subspace of $H^1(\Omega)$ and is equipped with the scalar product of $H^1(\Omega)$:

$$\langle u, v \rangle_V = \int_\Omega (uv + \nabla u \cdot \nabla v)\, dx.$$

Since V is a closed subspace of $H^1(\Omega)$, it is a Hilbert space.

Problem (6.82) can be written

$$\begin{cases} \text{find } u \in V \text{ such that} \\ a(u, v) = l(v) \quad \forall v \in V, \end{cases}$$

where a is given by (6.80) and $l(v) = \int_\Omega f v\,dx + \int_{\partial\Omega} g v\,d\sigma$.

Let us first verify that the bilinear form $a : V \times V \to \mathbf{R}$ is continuous on V. Set $M := \sup_{1\le i,j\le N} \|a_{ij}\|_{L^\infty(\Omega)}$. Then, for any $u, v \in V$

$$\begin{aligned} |a(u,v)| &\le M \int_\Omega \sum_{i,j=1}^N \Big|\frac{\partial u}{\partial x_i}\Big|\Big|\frac{\partial v}{\partial x_j}\Big|\,dx \\ &\le M \int_\Omega N|\nabla u|\,|\nabla v|\,dx \\ &\le MN\Big(\int_\Omega |\nabla u|^2\,dx\Big)^{1/2}\Big(\int_\Omega |\nabla v|^2\,dx\Big)^{1/2} \\ &\le MN\|u\|_V\,\|v\|_V, \end{aligned} \tag{6.83}$$

which implies that a is continuous. The continuity of l follows from the inequality

$$|l(v)| \le \|f\|_{L^2(\Omega)}\|v\|_{L^2(\Omega)} + \|g\|_{L^2(\partial\Omega)}\|\gamma_0(v)\|_{L^2(\partial\Omega)}$$

and the continuity of the trace operator from $H^1(\Omega)$ into $L^2(\Sigma)$. Let us now verify the crucial point, that is, the coercivity of $a : V \times V \to \mathbf{R}$. By (6.81), since $V \subset H^1(\Omega)$ we have

$$\forall v \in V \quad a(v,v) \ge \alpha\|\nabla v\|^2_{L^2(\Omega)^N}. \tag{6.84}$$

We now use the generalized Poincaré inequality (Theorem 5.4.3): the assumptions (a) and (b) on V imply the existence of a positive constant C such that

$$\forall v \in V \quad \|v\|_{L^2(\Omega)} \le C\|\nabla v\|_{L^2(\Omega)^N}. \tag{6.85}$$

As a consequence of (6.84) and (6.85) we obtain that

$$\forall v \in V \quad a(v,v) \ge \frac{\alpha}{1+C^2}\|v\|^2_V$$

and a is coercive on V.

(ii) is a direct consequence of (6.82) and of the fact that V contains $\mathcal{D}(\Omega)$.

(iii) is an equivalent formulation to (6.82) when the matrix $(a_{ij})_{i,j}$ is symmetric. Indeed, in that case the bilinear form $a : V \times V \to \mathbf{R}$ is symmetric, and the equivalence follows from Proposition 2.3.1. □

As a particular case of Theorem 6.5.1, let us consider the following situation: take

$$V = \{v \in H^1(\Omega) \ :\ \gamma_0(v) = 0 \text{ on } \Gamma_0\},$$

where $\Gamma_0 \subset \Gamma = \partial\Omega$ is a measurable subset of the boundary Γ with a strictly positive surface measure: $H^{N-1}(\Gamma_0) > 0$. We set $\Gamma_1 = \Gamma \setminus \Gamma_0$. We know (see Section 6.3) that V satisfies all the assumptions of Theorem 6.5.1. The question we are going to examine is the interpretation of the boundary conditions satisfied by the solution u of (6.82).

Proposition 6.5.1. *When $V = \{v \in H^1(\Omega) \ : \ \gamma_0(v) = 0 \text{ on } \Gamma_0\}$, where $H^{N-1}(\Gamma_0) > 0$, the unique solution u of the problem*

$$\begin{cases} \displaystyle\int_\Omega \sum_{i,j=1}^N a_{ij}(x) \frac{\partial u}{\partial x_i} \frac{\partial v}{\partial x_j}\, dx = \int_\Omega f v dx + \int_{\partial\Omega} g v \, d\sigma \quad \forall v \in V, \\ u \in V \end{cases} \tag{6.86}$$

is a weak solution of the following boundary value problem:

$$\begin{cases} \displaystyle -\sum_{i,j=1}^N \frac{\partial}{\partial x_j}(a_{ij} \frac{\partial u}{\partial x_i}) = f \text{ on } \Omega, \\ u = 0 \text{ on } \Gamma_0, \\ \frac{\partial u}{\partial \nu_A} = g \text{ on } \Gamma_1, \end{cases} \tag{6.87}$$

where $\frac{\partial u}{\partial \nu_A} := \sum_{i,j=1}^N a_{ij} \frac{\partial u}{\partial x_i} n_j$ is called the conormal derivative of u associated with the operator $A : v \mapsto -\Sigma \frac{\partial}{\partial x_j}(a_{ij} \frac{\partial v}{\partial x_i})$.

Proof. To find the boundary condition satisfied by u on Γ_1 let us make the following regularity assumptions: $u \in H^2(\Omega)$ and $a_{ij} \in \mathbf{C}^1(\bar{\Omega})$ for any $1 \le i, j \le N$.

Then, for any $i = 1, 2, \ldots, N$ we have

$$\xi_j := \sum_{i=1}^N a_{ij} \frac{\partial u}{\partial x_i} \in H^1(\Omega).$$

Let us write the Green's formula (Proposition 5.6.2)

$$\int_\Omega \sum_{j=1}^N \xi_j \frac{\partial v}{\partial x_j}\, dx = -\int_\Omega v \sum_{j=1}^N \frac{\partial \xi_j}{\partial x_j}\, dx + \int_{\Gamma_1} v \sum_{j=1}^N \xi_j n_j \, d\sigma. \tag{6.88}$$

From (6.86), (6.87), and (6.88) we infer that

$$\forall v \in V \quad \int_{\Gamma_1} \Big(\sum_{j=1}^N \xi_j n_j \Big) v \, d\sigma = \int_{\Gamma_1} g v \, d\sigma,$$

which implies

$$\sum_{j=1}^N \xi_j n_j = g,$$

that is, $\frac{\partial u}{\partial \nu_A} = g$ on Γ_1. □

Remark 6.5.1. When the above regularity properties on a_{ij} and u are not satisfied, the formula $\frac{\partial u}{\partial \nu_A} = g$ on Γ_1 is just a formal way to express the boundary conditions implicitly contained in (6.86).

6.6 The nonlinear Laplacian Δ_p

Most of the results of the previous sections have a natural extension when replacing the Dirichlet integral $\int_\Omega |\nabla v|^2\,dx$ by $\int_\Omega |\nabla v|^p\,dx$ with $1 < p < +\infty$ and the space $H^1(\Omega)$ by the space $W^{1,p}(\Omega)$. So doing, the Laplace operator Δ is replaced by the nonlinear Laplacian Δ_p, which is defined by

$$\begin{aligned} \Delta_p v &= \sum_{i=1}^{N} \frac{\partial}{\partial x_i}\left(|\nabla v|^{p-2}\frac{\partial v}{\partial x_i}\right) \\ &= \operatorname{div}(|\nabla v|^{p-2}\nabla v). \end{aligned} \tag{6.89}$$

Note that when $p = 2$ one obtains $\Delta_2 = \Delta$.

The assumption $1 < p < +\infty$ is crucial. When $1 < p < +\infty$, the space $W^{1,p}(\Omega)$ is a reflexive Banach space and most of the variational techniques which have been developed in the space $H^1(\Omega)$ can be generalized to this setting. As a variational tool, we use the convex minimization Theorem 3.3.4 which holds in general reflexive Banach spaces. The cases $p = 1$ and $p = +\infty$, which are important too for applications, are much more involved; a major reason is the lack of reflexivity of these spaces. As an illustration of a boundary value problem for the Δ_p operator, let us first consider the Dirichlet problem.

Theorem 6.6.1. *Let p be a positive real number $1 < p < +\infty$. Let Ω be a bounded open set in $\mathbf{R}^N$ and let $f \in L^\infty(\Omega)$ be a given function.*

(i) *There exists a unique solution $u \in W_0^{1,p}(\Omega)$ of the following minimization problem:*

$$\min\left\{\frac{1}{p}\int_\Omega |\nabla v(x)|^p\,dx - \int_\Omega f v\,dx \;:\; v \in W_0^{1,p}(\Omega)\right\}. \tag{6.90}$$

(ii) *Equivalently, u is solution of the problem*

$$\begin{cases} \displaystyle\int_\Omega |\nabla u|^{p-2}\nabla u\cdot\nabla v\,dx = \int_\Omega f v\,dx \quad \forall v \in W_0^{1,p}(\Omega), \\ u \in W_0^{1,p}(\Omega). \end{cases} \tag{6.91}$$

(iii) *The solution u of* (6.91) *is a weak solution of the boundary value problem*

$$\begin{cases} -\operatorname{div}(|\nabla u|^{p-2}\nabla u) = f \quad \text{on} \quad \Omega, \\ u = 0 \quad \text{on} \quad \partial\Omega, \end{cases} \tag{6.92}$$

where the first equation is satisfied in the sense of distributions and $u = 0$ in the sense of the trace operator $W^{1,p}(\Omega) \hookrightarrow L^p(\partial\Omega)$.

Proof. Let us recall that the space $W^{1,p}(\Omega)$ is equipped with the norm

$$\|v\|_{W^{1,p}(\Omega)} = \left(\int_\Omega (|v(x)|^p + |\nabla v(x)|^p)dx\right)^{1/p},$$

where $|\nabla v(x)|$ is the Euclidean norm of the vector $\nabla v(x)$ in $\mathbf{R}^N$.

We know (cf. Theorem 5.1.2) that $W^{1,p}(\Omega)$ is a Banach space, and since $1 < p < +\infty$, it is a reflexive Banach space.

By definition, $W_0^{1,p}(\Omega)$ is the closure of $\mathcal{D}(\Omega)$ in $W^{1,p}(\Omega)$. Let us equip $W_0^{1,p}(\Omega)$ with the norm of $W^{1,p}(\Omega)$. Hence, $W_0^{1,p}(\Omega)$, which is a closed subspace of a reflexive Banach space, is still a reflexive Banach space. Let us now consider the functional $J : W_0^{1,p}(\Omega) \to \mathbf{R}$ defined by

$$J(v) := \frac{1}{p}\int_\Omega |\nabla v(x)|^p\,dx - \int_\Omega f v\,dx \tag{6.93}$$

and verify that the assumptions of the convex minimization theorem, Theorem 3.3.4, are satisfied. Let us first verify that J is convex: For any $v_1, v_2 \in W_0^{1,p}(\Omega)$, for any $0 \le \lambda \le 1$

$$\begin{aligned} J(\lambda v_1 + (1-\lambda)v_2) &= \frac{1}{p}\int_\Omega |\lambda\nabla v_1(x) + (1-\lambda)\nabla v_2(x)|^p\,dx \\ &\quad -\lambda\int_\Omega f v_1\,dx - (1-\lambda)\int_\Omega f v_2\,dx \\ &\le \frac{1}{p}\int_\Omega (\lambda|\nabla v_1(x)| + (1-\lambda)|\nabla v_2(x)|)^p\,dx \\ &\quad -\lambda\int_\Omega f v_1\,dx - (1-\lambda)\int_\Omega f v_2\,dx. \end{aligned}$$

Let us now use the convexity of the function $\varphi(r) = r^p$ from $\mathbf{R}^+$ into $\mathbf{R}^+$ (note that $\varphi''(r) = p(p-1)r^{p-2}$ is nonnegative) to obtain

$$\begin{aligned} J(\lambda v_1 + (1-\lambda)v_2) &\le \lambda\left[\frac{1}{p}\int_\Omega |\nabla v_1|^p\,dx - \int_\Omega f v_1\,dx\right] \\ &\quad + (1-\lambda)\left[\frac{1}{p}\int_\Omega |\nabla v_2|^p\,dx - \int_\Omega f v_2\,dx\right] \\ &\le \lambda J(v_1) + (1-\lambda)J(v_2). \end{aligned}$$

Clearly J is continuous on $W_0^{1,p}(\Omega)$: note that

$$\int_\Omega \big||\nabla v(x)| - |\nabla w(x))|\big|^p\,dx \le \int_\Omega |\nabla(v-w)(x)|^p\,dx \le \|v-w\|^p_{W^{1,p}(\Omega)}.$$

The above inequality implies the continuity of the mapping $v \mapsto |\nabla v|$ from $W^{1,p}$ into L^p and hence of the functional $v \mapsto \int_\Omega |\nabla v|^p\,dx$ on $W^{1,p}$. Note that since $f \in L^\infty(\Omega)$,

$$\begin{aligned} \left|\int_\Omega f v\,dx\right| &\le \|f\|_\infty \|v\|_1 \\ &\le \|f\|_\infty |\Omega|^{1/p'}\|v\|_p \quad (\text{where } \frac{1}{p} + \frac{1}{p'} = 1) \\ &\le \|f\|_\infty |\Omega|^{1/p'}\|v\|_{W^{1,p}} \end{aligned} \tag{6.94}$$

and $v \mapsto \int_\Omega f v$ is a linear continuous form on $W^{1,p}(\Omega)$.

Thus, the only point which remains to verify is the coercivity of J on $W_0^{1,p}(\Omega)$. To that end, we use the classical Poincaré inequality on $W_0^{1,p}(\Omega)$ which does not require any regularity assumptions on Ω, just Ω to be bounded. By Theorem 5.3.1, there exists a positive constant C such that

$$\forall v \in W_0^{1,p}(\Omega) \quad \int_\Omega |v(x)|^p \, dx \le C \int_\Omega |\nabla v(x)|^p \, dx. \tag{6.95}$$

To verify that J is coercive, let us prove that its sublevel sets are bounded. (Indeed these two properties are equivalent; see Proposition 3.2.8.) Let us fix some $\lambda \in \mathbf{R}$ and consider

$$lev_\lambda J = \{v \in W_0^{1,p}(\Omega) \ : \ J(v) \le \lambda\}.$$

For $v \in lev_\lambda J$, by using (6.93) we have

$$\begin{aligned} \int_\Omega |\nabla v|^p \, dx &\le p \int_\Omega |fv| \, dx + p\lambda \\ &\le p\|f\|_\infty |\Omega|^{1/p'} \|v\|_p + p\lambda. \end{aligned} \tag{6.96}$$

By using the Poincaré inequality (6.95), we obtain from (6.96)

$$\|v\|_p^p \le pC\|f\|_\infty |\Omega|^{1/p'} \|v\|_p + pC\lambda.$$

Hence

$$\|v\|_p^{p-1} \le pC\|f\|_\infty |\Omega|^{1/p'} + \frac{pC\lambda}{\|v\|_p},$$

which implies

$$\|v\|_p \le \max\left\{1, (pC\|f\|_\infty |\Omega|^{1/p'} + pC\lambda)^{\frac{1}{p-1}}\right\}. \tag{6.97}$$

Returning to (6.96) we obtain that v remains in a bounded subset of $W_0^{1,p}(\Omega)$.

Let us summarize the previous results: $W_0^{1,p}(\Omega)$ is a reflexive Banach space and $J : W_0^{1,p}(\Omega) \to \mathbf{R}$ is a convex, continuous, coercive functional. All the assumptions of Theorem 3.3.4 are satisfied and the minimization problem (6.90) admits a solution. The uniqueness of the solution is a consequence of the strict convexity on $W_0^{1,p}(\Omega)$ of the functional $v \mapsto \int_\Omega |\nabla v|^p \, dx$, which is a consequence of the strict convexity of the function $r \to r^p$ from $\mathbf{R}^+$ into $\mathbf{R}^+$.

(ii) Let us now establish the corresponding Euler equation. To that end, let us write, for any $v \in W_0^{1,p}(\Omega)$, for any $t > 0$,

$$\frac{1}{t}[J(u + tv) - J(u)] \ge 0,$$

and pass to the limit on this inequality as $t \to 0^+$. We have

$$\forall v \in W_0^{1,p}(\Omega) \quad \frac{1}{p} \int_\Omega \frac{|\nabla u + t\nabla v|^p - |\nabla u|^p}{t} \, dx - \int_\Omega fv \, dx \ge 0. \tag{6.98}$$

To pass to the limit in (6.98) we use the Lebesgue dominated convergence theorem: Set $h(t) = |\nabla u + t\nabla v|^p$. We have $h'(t) = p|\nabla u + t\nabla v|^{p-2}(\nabla u + t\nabla v) \cdot \nabla v$.

Hence

$$\frac{1}{t}\left[|\nabla u + t\nabla v|^p - |\nabla u|^p\right] = \frac{1}{t}(h(t) - h(0))$$
$$= \frac{1}{t}\int_0^1 p|\nabla u + s\nabla v|^{p-2}(\nabla u + s\nabla v)\cdot\nabla v\, ds.$$

From this, by taking $0 < t \leq 1$, we obtain

$$\begin{aligned}\frac{1}{t}\left||\nabla u + t\nabla v|^p - |\nabla u|^p\right| &\leq \frac{p}{t}\int_0^t |\nabla u + s\nabla v|^{p-1}|\nabla v|\, ds \\ &\leq p(|\nabla u| + |\nabla v|)^{p-1}|\nabla v|.\end{aligned} \tag{6.99}$$

Let us notice that $|\nabla v| \in L^p(\Omega)$ and that $(|\nabla u| + |\nabla v|)^{p-1}$ belongs to $L^{p'}(\Omega)$ (because of the equality $(p-1)p' = p$). Hence the right-hand side of (6.99) is a function which belongs to $L^1(\Omega)$ and which is independent of $t \in]0, 1]$. We can now pass to the limit on (6.98) and obtain

$$\forall v \in W_0^{1,p}(\Omega) \quad \int_\Omega |\nabla u|^{p-2}\nabla u\cdot\nabla v\, dx - \int_\Omega f v\, dx \geq 0.$$

Let us now replace v by $-v$. We obtain

$$\forall v \in W_0^{1,p}(\Omega) \quad \int_\Omega |\nabla u|^{p-2}\nabla u\cdot\nabla v\, dx = \int_\Omega f v\, dx.$$

(iii) By taking $v \in \mathcal{D}(\Omega)$, we obtain, by definition of the derivation in distributions, the following equality:

$$-\operatorname{div}(|\nabla u|^{p-2}\nabla u) = f \text{ in } \mathcal{D}'(\Omega).$$

On the other hand, when Ω is regular, from $u \in W_0^{1,p}(\Omega)$ we infer that $u = 0$ on $\partial\Omega$ in the trace sense (Proposition 5.6.1). □

Remark 6.6.1. We could as well consider Neumann-type boundary value problems for the Δ_p operator. Let us just notice that, assuming $u \in \mathbf{C}^2(\bar{\Omega})$, we have the following integration by parts formula:

$$\forall v \in \mathcal{D}(\bar{\Omega}) \quad \int_\Omega |\nabla u|^{p-2}\nabla u\cdot\nabla v\, dx = -\int_\Omega v\Delta_p u\, dx + \int_{\partial\Omega} |\nabla u|^{p-2}\frac{\partial u}{\partial n} v\, d\sigma.$$

It follows that variational problems in $W^{1,p}(\Omega)$ lead to boundary conditions of the following type:

$$|\nabla u|^{p-2}\frac{\partial u}{\partial n} = g \quad \text{on } \partial\Omega.$$

(Note that when $p = 2$, one recovers the classical Neumann boundary condition.)

6.7 The Stokes system

In this section, we are going to make precise the variational approach to the Stokes system, which was introduced in Section 2.3.1. Let us recall that the Stokes system for an incompressible viscous fluid in a domain Ω of $\mathbf{R}^N$ consists in finding functions $u_1, u_2, \dots u_N : \Omega \longrightarrow \mathbf{R}$ and $p : \Omega \longrightarrow \mathbf{R}$ which satisfy

$$\begin{cases} -\mu \Delta u_i + \dfrac{\partial p}{\partial x_i} = f_i \text{ on } \Omega, \ i = 1, \dots, N, \\ \displaystyle\sum_{i=1}^{N} \frac{\partial u_i}{\partial x_i} = 0 \quad \text{on } \Omega, \\ u_i = 0 \quad \text{on } \partial\Omega, \ i = 1, \dots, N. \end{cases}$$

The given vector $f = (f_1, f_2, \dots, f_N) \in L^2(\Omega)^N$ represents a volumic density of forces, $\mu > 0$ is the viscosity coefficient. (It is a positive scalar which is inversely proportional to the Reynolds number.) The vector function $u = (u_1, \dots, u_N) : \Omega \longrightarrow \mathbf{R}^N$ is the velocity vector field of the fluid; it assigns to each point $x \in \Omega$ the velocity vector $u(x) = (u_i(x))_{i=1,\dots,N}$ of the fluid at x. The scalar function $p : \Omega \longrightarrow \mathbf{R}$ is the pressure; for each $x \in \Omega$, $p(x)$ is the pressure of the fluid at x.

The Stokes system can be written in the following form:

$$\begin{cases} -\mu \Delta u + \nabla p = f \quad \text{on } \Omega, \\ \operatorname{div}(u) = 0 \quad \text{on } \Omega, \\ u = 0 \quad \text{on } \partial\Omega. \end{cases}$$

The condition $\operatorname{div}(u) = 0$ expresses that the fluid is incompressible. The Stokes system is a linear system of $(N + 1)$ partial differential equations on Ω involving $(N + 1)$ unknown functions $(u_1, \dots u_N, p)$.

The variational formulation of the Stokes system was introduced by Leray around 1934. The idea is to work in the functional space

$$V = \left\{ v \in H_0^1(\Omega)^N : \operatorname{div}(v) = 0 \right\} \tag{6.100}$$

and make the pressure appear as a Lagrange multiplier of the constraint "$\operatorname{div}(v) = 0$." Let us assume that Ω is a bounded connected open subset of $\mathbf{R}^N$ whose boundary is piecewise $\mathbf{C}^1$. The space V is equipped with the scalar product of $H_0^1(\Omega)^N$

$$\langle u, v \rangle_{H_0^1(\Omega)^N} = \sum_{i=1}^{N} \langle u_i, v_i \rangle_{H_0^1(\Omega)},$$

where

$$\langle u_i, v_i \rangle_{H_0^1(\Omega)} = \int_\Omega (u_i v_i + \nabla u_i \cdot \nabla v_i) \, dx,$$

and the corresponding norm

$$\|v\|_{H_0^1(\Omega)^N} = \left(\sum_{i=1}^{N} \int_\Omega (u_i^2 + |\nabla u_i|^2) \, dx \right)^{1/2}.$$

The space V is equal to the kernel of the divergence operator div,

$$\operatorname{div} : v \in H_0^1(\Omega)^N \longrightarrow \operatorname{div}(v) \in L^2(\Omega),$$

which is a linear continuous operator from $H_0^1(\Omega)^N$ into $L^2(\Omega)$. The continuity of the div operator follows from the following inequality:

$$\begin{aligned} \forall v \in H_0^1(\Omega)^N \quad \|\operatorname{div}(v)\|^2_{L^2(\Omega)} &= \sum_{i=1}^N \int_\Omega \left| \frac{\partial v_i}{\partial x_i} \right|^2 dx \\ &\leq \sum_{i=1}^N \int_\Omega |\nabla v_i|^2 \, dx \\ &\leq \|v\|^2_{H_0^1(\Omega)^N}. \end{aligned}$$

Hence, V is a closed subspace of $H_0^1(\Omega)^N$ and V is an Hilbert space.

We can now state the variational formulation of the Stokes system.

Theorem 6.7.1. (a) *For every $f \in L^2(\Omega)^N$ there exists a unique $u \in V$ which satisfies*

$$\begin{cases} \mu \displaystyle\sum_{i=1}^N \int_\Omega \nabla u_i \cdot \nabla v_i \, dx = \sum_{i=1}^N \int_\Omega f_i v_i \, dx \qquad \forall v \in V, \\ u \in V. \end{cases} \tag{6.101}$$

(b) *Let u be the solution of* (6.101)*. Then the relation* (6.101) *determines a unique $p \in L^2(\Omega)$ (up to an additive constant) such that the couple $(u, p) \in V \times L^2(\Omega)$ satisfies*

$$\mu \sum_{i=1}^N \int_\Omega \nabla u_i \cdot \nabla v_i \, dx - \sum_{i=1}^N \int_\Omega f_i v_i \, dx = \int_\Omega p \operatorname{div}(v) \, dx \quad \forall v \in H_0^1(\Omega)^N. \tag{6.102}$$

(c) *The couple (u, p) is a weak solution of the Stokes system:*

$$\begin{cases} -\mu \Delta u + \nabla p = f \text{ in } \mathcal{D}'(\Omega)^N, \\ \operatorname{div}(u) = 0 \quad \text{in } \mathcal{D}'\Omega), \\ u = 0 \quad \text{on } \partial\Omega \text{ in the trace sense.} \end{cases}$$

The couple (u, p) is called the variational solution of the Stokes system.

Proof. (a) Let us consider the bilinear form $a : V \times V \longrightarrow \mathbf{R}$

$$a(u, v) = \mu \sum_{i=1}^N \int_\Omega \nabla u_i \cdot \nabla v_i \, dx$$

and the linear form $l : V \longrightarrow \mathbf{R}$

$$l(v) = \sum_{i=1}^N \int_\Omega f_i v_i \, dx.$$

The bilinear form a is clearly continuous and its coercivity follows, by standard argument, from the Poincaré inequality in $H_0^1(\Omega)$. The continuity of l is also immediate. Thus, all the assumptions of the Lax–Milgram theorem, Theorem 3.1.2, are satisfied. This implies the existence and uniqueness of the solution u of (6.101).

(b) The difficulty comes from the fact that $\mathcal{D}(\Omega)^N$ is not contained in the space V, and one cannot interpret directly (6.101) in terms of distributions. Moreover, up to now, the pressure p still has not appeared in the above variational formulation. Let us reformulate (6.101) as an orthogonality relation. Let us consider the linear form $L : H_0^1(\Omega)^N \longrightarrow \mathbf{R}$ which is defined by

$$L(v) = \mu \sum_{i=1}^{N} \int_\Omega \nabla u_i \cdot \nabla v_i \, dx - \sum_{i=1}^{N} \int_\Omega f_i v_i \, dx. \tag{6.103}$$

The linear form L is clearly continuous on $H_0^1(\Omega)^N$ and (6.101) precisely tells us that $L(v) = 0$ for all $v \in V$. In other words $L \in \left(H_0^1(\Omega)^N\right)^*$ and $L(v) = 0 \ \forall v \in V$, i.e., $L \in V^\perp$ the orthogonal subspace of V (for the pairing between $H_0^1(\Omega)^N$ and its topological dual). The precise description of such elements is provided by the following theorem, obtained by De Rham in 1955.

Theorem 6.7.2. *Let Ω be a bounded connected set in $\mathbf{R}^N$ whose boundary is piecewise $\mathbf{C}^1$. Let $L \in \left(H_0^1(\Omega)^N\right)^*$, a linear continuous form on $H_0^1(\Omega)^N$. Set $V = \left\{v \in H_0^1(\Omega)^N : \operatorname{div}(v) = 0\right\}$. Then*

$$L(v) = 0 \quad \forall v \in V \iff \exists p \in L^2(\Omega) \text{ such that } L(v) = \int_\Omega p \operatorname{div}(v) \, dx \quad \forall v \in H_0^1(\Omega)^N.$$

Proof of Theorem 6.7.1 continued. Let us admit the De Rham theorem (the implication $L \in V^\perp \Longrightarrow \exists p \ldots$, which is the interesting part of the theorem, is a nontrivial result), and apply it to the bilinear form L which is defined in (6.103). We thus have the existence of $p \in L^2(\Omega)$ such that

$$\mu \sum_{i=1}^{N} \int_\Omega \nabla u_i \cdot \nabla v_i \, dx - \sum_{i=1}^{N} \int_\Omega f_i v_i \, dx = \int_\Omega p \operatorname{div}(v) \, dx \quad \forall v \in H_0^1(\Omega)^N.$$

This is precisely (6.102).

(c) Since $\mathcal{D}(\Omega)^N$ is dense in $H_0^1(\Omega)^N$, the solution of (6.102) is characterized by $(u, p) \in V \times L^2(\Omega)$ and

$$\mu \sum_{i=1}^{N} \int_\Omega \nabla u_i \cdot \nabla v_i \, dx - \int_\Omega p \operatorname{div}(v) \, dx = \sum_{i=1}^{N} \int_\Omega f_i v_i \, dx \quad \forall v \in \mathcal{D}(\Omega)^N.$$

Taking $v = (0, \ldots, v_i, \ldots 0), \ i = 1, \ldots, N$ yields

$$\mu \int_\Omega \nabla u_i \cdot \nabla v_i \, dx - \int_\Omega p \frac{\partial v_i}{\partial x_i} \, dx = \int_\Omega f_i v_i \, dx \quad \forall v_i \in \mathcal{D}(\Omega),$$

that is,

$$-\mu \Delta u_i + \frac{\partial p}{\partial x_i} = f_i \text{ in } \mathcal{D}'(\Omega).$$

Moreover, $u \in V$ contains the information

$$\begin{aligned} &\operatorname{div}(u) = 0 \quad \text{in } \mathcal{D}'(\Omega), \\ &u = 0 \quad \text{on } \partial\Omega \text{ in the trace sense.} \end{aligned}$$

Hence, (u, p) is a weak solution of the Stokes system. □

Chapter 7

The finite element method

One of the major interests of variational methods is to provide both a theory for existence of solutions and numerical methods for computing accurate approximations of these solutions. Certainly, the most celebrated of these variational approximation methods is the finite element method. It is a Galerkin approximation scheme where the elements of the finite dimensional approximating subspaces V_n are piecewise polynomial functions. This method has been proved to be very successful, the main reason being that, because of the local character of several problems, by choosing a basis of the space V_n whose functions have small supports, one obtains approximated problems with sparse matrices, i.e., with most entries equal to zero. This is a decisive property in order to be able to solve numerically the corresponding linear system: one should notice that engineering problems involving systems of PDEs from continuum mechanics usually give rise to large linear systems (100×100 or 1000×1000 are quite frequent!).

Our scope is to introduce the main ideas of the finite element method and then to describe a typical example.

For simplicity of the exposition, we restrict ourselves to linear problems whose variational formulation enters into the abstract setting of the Lax–Milgram theorem: find $u \in V$ such that

$$a(u, v) = l(v) \text{ for all } v \in V. \tag{7.1}$$

Let us first recall and make precise some aspects of the Galerkin method (which was introduced in Section 3.1.2).

7.1 The Galerkin method: Further results

Let us briefly recall the assumptions on the abstract variational problem (7.1): V is a Hilbert space, $a : V \times V \to \mathbf{R}$ is a continuous coercive bilinear form, i.e., there exists some constants $M \in \mathbf{R}^+$ and $\alpha > 0$ such that

$$\forall u, v \in V, \quad |a(u, v)| \leq M\|u\|\|v\|, \tag{7.2}$$

$$\forall v \in V, \quad a(v, v) \geq \alpha\|v\|^2. \tag{7.3}$$

The linear form $l : V \to \mathbf{R}$ is supposed to be continuous. Then, Lax–Milgram's theorem asserts the existence and uniqueness of a solution $u \in V$ of problem (7.1).

Typically, as in the boundary value problems which were studied in the previous sections, V is a Sobolev space, like $H^1(\Omega)$ or $H_0^1(\Omega)$. It is an infinite dimensional space; this is a common feature of all methods from functional analysis, and the numerical computation of the solution u requires a further step which is the reduction to a finite dimensional problem.

The Galerkin method is based on the approximation of the infinite dimensional space V by a sequence of finite dimensional subspaces $(V_n)_{n\in\mathbf{N}}$. More precisely, for each $n \in \mathbf{N}$, V_n is a finite dimensional subspace of V and one supposes that the following approximation property holds:

$$\forall v \in V,\ \exists (v_n)_{n\in\mathbf{N}},\ v_n \in V_n\ \forall\, n \in \mathbf{N},\ \text{and } v_n \to v \text{ in } V. \tag{7.4}$$

For each $n \in \mathbf{N}$ the approximated variational problem is

$$\begin{cases} \text{find } u_n \in V_n \text{ such that} \\ a(u_n, v) = l(v)\ \forall\ v \in V_n. \end{cases} \tag{7.5}$$

One should notice that the approximated problem (7.5) is still a variational problem: it has the same structure as the initial variational problem (7.1), except now it is posed on a finite dimensional space V_n. Indeed, existence and uniqueness of the solution u_n of (7.5) follows, in a similar way, from the Lax–Milgram theorem. When a is symmetric, problem (7.5) reduces to a minimization problem, namely, u_n is the solution of

$$\min\left\{\frac{1}{2}a(v, v) - l(v) : v \in V_n\right\}. \tag{7.6}$$

This alternate description of the finite dimensional approximation method is often called the Ritz method.

The term *variational approximation* is justified by the fact that the sequence of problems (7.5) does approximate the initial problem (7.1), in the sense that the sequence $(u_n)_{n\in\mathbf{N}}$ norm converges in V to u. More precisely, in Proposition 3.1.2 it was proved that

$$\|u - u_n\| \le \frac{M}{\alpha}\,\mathrm{dist}(u, V_n), \tag{7.7}$$

and one can observe that, clearly, the approximation property (7.4) implies that for any $u \in V$, $\mathrm{dist}(u, V_n) \to 0$ as $n \to +\infty$.

Let us now make precise the structure of the approximated problem (7.5). Let us introduce a basis $(\varphi_1, \varphi_2, \ldots, \varphi_{I(n)})$ of the vector space V_n with $I(n) = \dim V_n$. Let us write $u_n = \sum_{i=1}^{I(n)} \lambda_i \varphi_i$. Then (7.5) is equivalent to

$$\begin{cases} a(u_n, \varphi_j) = l(\varphi_j) \quad \forall\ j = 1, 2, \ldots, I(n), \\ u_n = \displaystyle\sum_{i=1}^{I(n)} \lambda_i \varphi_i. \end{cases}$$

This, in turn, is equivalent to finding $\lambda = (\lambda_i)_{i=1,2,\ldots,I(n)}$ in $\mathbf{R}^{I(n)}$ which is a solution of the linear system

$$\sum_{i=1}^{I(n)} \lambda_i a(\varphi_i, \varphi_j) = l(\varphi_j) \qquad \forall\ j = 1, 2, \ldots, I(n). \tag{7.8}$$

Let us set $A_n = (a(\varphi_j, \varphi_i))_{1\le i,j\le I(n)}$. It is an $I(n)\times I(n)$ square matrix which, by reference to the elasticity problem, is often called the stiffness matrix. Similarly, the vector $b_n = (l(\varphi_j))$ in $\mathbf{R}^{I(n)}$ is often called the load vector.

With these notations, one can write (7.8) in the following form:

$$A_n\lambda = b_n. \tag{7.9}$$

Let us now examine the properties of the matrix A_n: for any vector $\lambda \in \mathbf{R}^{I(n)}$ we have ($\langle\cdot,\cdot\rangle$ is the Euclidean scalar product in $\mathbf{R}^{I(n)}$ and $|\cdot|$ the Euclidean norm)

$$\begin{aligned}\langle A_n\lambda, \lambda\rangle &= \sum_{i=1}^{I(n)} (A_n\lambda)_i\lambda_i \\ &= \sum_{i=1}^{I(n)} \left(\sum_{j=1}^{I(n)} a(\varphi_j,\varphi_i)\lambda_j\right)\lambda_i \\ &= a\left(\sum_{j=1}^{I(n)}\lambda_j\varphi_j, \sum_{i=1}^{I(n)}\lambda_i\varphi_i\right) \\ &\ge \alpha\|\sum_{i=1}^{I(n)}\lambda_i\varphi_i\|^2.\end{aligned}$$

Since $(\varphi)_{i=1}^{I(n)}$ is a basis of V_n, one can easily verify that $\lambda \mapsto \|\sum_{i=1}^{I(n)}\lambda_i\varphi_i\|$ is a norm on $\mathbf{R}^{I(n)}$. All norms being equivalent on the finite dimensional space $\mathbf{R}^{I(n)}$, there exists some constant $c > 0$ such that

$$\forall\lambda \in \mathbf{R}^{I(n)} \qquad c|\lambda| \le \left\|\sum_{i=1}^{I(n)}\lambda_i\varphi_i\right\|.$$

Hence

$$\forall\lambda \in \mathbf{R}^{I(n)} \qquad \langle A_n\lambda,\lambda\rangle \ge \alpha c^2|\lambda|^2. \tag{7.10}$$

From (7.10) it follows that A_n is one to one (that is, $\ker(A_n) = \{0\}$), which, in this finite dimensional setting, implies that for any load vector b_n, problem (7.9) has a unique solution. This is another elementary way (without using Lax–Milgram's theorem) to prove existence and uniqueness of the solution λ of the approximated problem (7.9). Let us also notice that when a is symmetric, so is the matrix A_n.

We now come to the central point of this theory which is the effective construction of the finite dimensional approximating subspaces V_n and the resolution of the linear system (7.9).

As we have already stressed, it is crucial, from a numerical point of view, that the matrix A_n possesses as many zeroes as possible. At this point, there are different strategies: in the next section, we shall describe an approach which uses spectral analysis in infinite dimensional spaces and a special basis whose elements are eigenfunctions.

The finite element method relies on a different strategy that we now describe.

7.2 Description of finite element methods

Let us now assume that V is a closed subspace of $H^1(\Omega)$ and the bilinear form $a : V \times V \to \mathbf{R}$ is of the type

$$a(u, v) = \int_\Omega (\nabla u \cdot \nabla v + a_0 uv)\, dx, \tag{7.11}$$

with $a_0 \in L^\infty(\Omega)$, $a_0 \geq 0$. This allows us to cover various situations like Dirichlet, Neumann, and mixed problems which were studied in the previous sections. A key property of a is that it is a *local* bilinear form, that is,

$$a(\varphi, \psi) = 0,$$

as soon as φ and ψ are two elements of V whose supports do not intersect, or more generally such that the Lebesgue measure of the intersection of their supports is zero. Let us recall that the stiffness matrix A_n is equal to $(a(\varphi_i, \varphi_j))$, where (φ_i), $i = 1, \ldots, I(n)$, is a basis of V_n. The strategy is now clear: we have to choose V_n such that one can find a canonical basis in the space V_n whose corresponding functions have supports which are as small as possible.

This is made possible thanks to a *triangulation* of the set $\overline{\Omega}$. For simplicity of the exposition, we restrict ourselves to problems which are posed over sets $\overline{\Omega} \subset \mathbf{R}^2$ which are polyhedra; we also say that such set Ω is polygonal.

Definition 7.2.1. *A triangulation $\mathcal{T}$ of a polygonal set Ω of $\mathbf{R}^2$ is a finite decomposition of the set $\overline{\Omega}$ of the form*

$$\overline{\Omega} = \bigcup_{K \in \mathcal{T}} K$$

such that

(i) *each set $K \in \mathcal{T}$ is a triangle,*

(ii) *whenever K_1 and K_2 belong to $\mathcal{T}$, $K_1 \cap K_2$ is either empty, or reduced to a common vertex, or to a common face (edge).*

In particular, for each distinct $K_1, K_2 \in \mathcal{T}$, one has $int(K_1) \cap int(K_2) = \emptyset$. Two triangles K_1 and K_2 which have a common face are said to be adjacent. The triangles $K \in \mathcal{T}$ are called finite elements.

We set

$$h(\mathcal{T}) = \max_{K \in \mathcal{T}} \operatorname{diam} K, \tag{7.12}$$

where $\operatorname{diam} K = \sup\{|x - y| : x, y \in K\}$ is the diameter of K. By convention, we denote by $\mathcal{T}_h$ a triangulation $\mathcal{T}$ such that $h(\mathcal{T}) = h$. As we shall see, to each triangulation $\mathcal{T}_h$ will be associated a finite approximating dimensional subspace V_h. It is convenient to consider the family of approximating subspaces indexed by the positive parameter h, say, V_h with $h \to 0$. Of course, to reduce to an abstract Galerkin scheme as described in Section 7.1, one may take $V_h = V_{h_n}$ for some $h_n \to 0$. An example of triangulation is given in Figure 7.1.

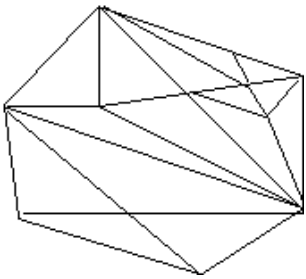


Figure 7.1. *Example of triangulation.*

One may already observe that one can use a fine triangulation in a subregion where a particular behavior of the solution is expected (for example, in special parts of airplanes). Figure 7.2 shows a forbidden situation where the intersection of two triangles K_1 and K_2 is not an edge of K_2:

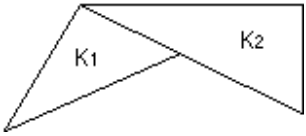


Figure 7.2. *A forbidden situation.*

Let us now describe the finite dimensional space V_h which is associated to a triangulation $\mathcal{T}_h$. At this point, we need to make precise the boundary condition; take, for example, the Dirichlet boundary condition $u = 0$ on $\partial\Omega$ and $V = H_0^1(\Omega)$. Then

$$V_h = \{v \in \mathbf{C}(\overline{\Omega}) : v \text{ is affine on each } K \in \mathcal{T}_h,\ v = 0 \text{ on } \partial\Omega\}.$$

In other words, V_h is the linear space of continuous functions on $\overline{\Omega}$ which are piecewise linear with respect to the triangulation $\mathcal{T}_h$ and which vanish on the boundary.

One can easily verify that an affine function on a triangle K is uniquely determined by its values at the vertices of K. Hence, any function $v \in V_h$ is uniquely determined by its values at the vertices (also called nodes) of the triangulation which are in the interior of $\overline{\Omega}$, i.e., in Ω (on the vertices which are on the boundary $\partial\Omega$, v is prescribed to be equal to zero).

For any vertex a_i of the triangulation, $i = 1, 2, \ldots, I(h)$, which is in the interior set Ω, let us denote by φ_i the element of V_h which satisfies

$$\varphi_i(a_j) = \delta_{ij}, \qquad 1 \le i, j \le I(h).$$

Equivalently, φ_i is the function of V_h which is equal to one at the vertex a_i and is equal to zero at all other vertices a_j with $j \neq i$. It is usually called a hat function. Clearly, $(\varphi_1, \varphi_2, \ldots, \varphi_{I(h)})$ is a basis of V_h and each element v of V_h can be uniquely written in the

form

$$v = \sum_{i=1}^{I(h)} v(a_i)\varphi_i .$$

One should notice that each element φ_i of this basis has small support: more precisely, the support of φ_i is the union of all triangles K of $\mathcal{T}_h$ such that a_i is a vertex of K. The stiffness matrix $A_h = (a(\varphi_i, \varphi_j))_{1\le i,j\le I(h)}$ is a sparse matrix, since $a(\varphi_i, \varphi_j) = 0$ except when a_i and a_j are two vertices of a same triangle K of the triangulation $\mathcal{T}_h$.

Let us now stress a technical but important point: the structure of A_h, i.e., the distribution of zeroes, for a given triangulation $\mathcal{T}_h$ highly depends on the enumeration of the vertices. Clearly, one has to use an enumeration to obtain A_h with a simple structure like, for example, tridiagonal matrices. Let us illustrate this in a concrete situation.

7.3 An example

Take $\Omega = (0,1) \times (0,1)$ the unit square in $\mathbf{R}^2$ and, given $f \in L^2(\Omega)$, let us consider the Dirichlet boundary value problem

$$\begin{cases} -\Delta u = f & \text{on } \Omega, \\ u = 0 & \text{on } \partial\Omega. \end{cases}$$

Its variational formulation is as follows: find $u \in H_0^1(\Omega)$ such that

$$\int_\Omega \nabla u \cdot \nabla v \, dx = \int_\Omega f v \, dx \qquad \forall\, v \in H_0^1(\Omega).$$

Let us consider the triangulation of Ω in Figure 7.3.

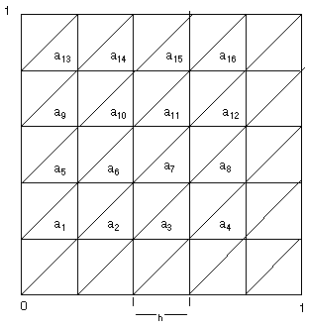


Figure 7.3. *Triangulation of Ω with indexed nodes.*

We denote by $a_{l,m} = (lh, mh)$, $0 \le l, m \le N+1$, the nodes of the triangulation. There are N^2 nodes which are in Ω, and the dimension of V_h is equal to N^2 with $h = 1/(N+1)$. Let us index the nodes of the triangulation as indicated in Figure 7.3, where N has been taken equal to 4. We draw, for example, the perspective of the hat function φ_6 which is an element of the finite element basis (Figure 7.4).

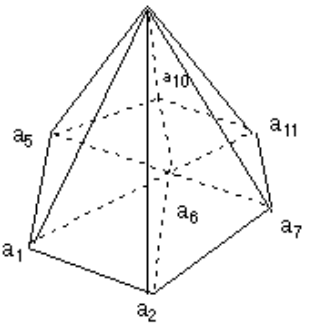


Figure 7.4. *The hat function* φ_6.

Recall that $a(\varphi_i, \varphi_j) \neq 0$ iff a_i and a_j are two vertices of the same triangle. Then, notice that a vertex a_i is connected to at most six other vertices a_j with $j \neq i$, which, a priori, yields a seven-point numerical scheme. The matrix A_h then has the following structure where each cross represents an element $a(\varphi_i, \varphi_j)$ which is, a priori, not equal to zero.

It is a matrix with a band structure, that is,

$$a(\varphi_i, \varphi_j) = 0 \text{ for } |i - j| > d_{max},$$

where d_{max}, the width of the band, is small with respect to the size of the matrix.

Indeed, for symmetry reasons

$$a(\varphi_{l,m}, \varphi_{l+1,m+1}) = 0 \text{ and } a(\varphi_{l,m}, \varphi_{l-1,m-1}) = 0$$

and we have a five-point scheme! Let $\lambda_{l,m} = u_h(\varphi_{l,m})$ be the component of u_h with respect to $\varphi_{l,m}$ so that $u_h = \sum \lambda_{l,m}\varphi_{l,m}$. An elementary computation yields

$$\begin{cases} -\lambda_{l-1,m} - \lambda_{l,m-1} + 4\lambda_{l,m} - \lambda_{l,m+1} - \lambda_{l+1,m} = h^2 f_{l,m}, & 1 \leq l, m \leq N, \\ \lambda_{l,0} = \lambda_{l,N+1} = 0, & 0 \leq l \leq N+1, \\ \lambda_{0,m} = \lambda_{N+1,m} = 0, & 0 \leq m \leq N+1. \end{cases}$$

This is the classical five-point scheme for the Laplacian. (Here $f_{l,m} = \int_\Omega f\varphi_{l,m}\,dx$ or, equivalently, an approximation of this integral.)

Remark 7.3.1. It is worth noticing that the tridiagonal block structure of A_h is intimately related to the enumeration of the elements of the basis. This structure may be lost by choosing a different enumeration!

7.4 Convergence of the finite element method

The convergence of the finite element method, which is a Galerkin approximation method, relies on Proposition 3.1.2:

$$\|u - u_h\|_{H_0^1(\Omega)} \leq \frac{M}{\alpha} \operatorname{dist}(u, V_h).$$

Therefore, the estimate of the error $\|u-u_h\|$ (and showing that the error goes to zero as $h \to 0$) can be reduced to a problem in approximation theory: one has to evaluate (majorize) the distance for the $H_0^1(\Omega)$ norm between a function $u \in H_0^1(\Omega)$ and the subspace V_h of continuous functions which are piecewise affine relative to a given triangulation $\mathcal{T}_h$.

To that end, we need to make a geometrical assumption on the family of triangulations $(\mathcal{T}_h)_{h\to 0}$.

Definition 7.4.1. *A family of triangulations $(\mathcal{T}_h)_{h>0}$ is said to be regular if there exists a constant σ ($\sigma \geq 0$) such that for any $h > 0$ and any $K \in \mathcal{T}_h$*

$$\frac{h_K}{\rho_K} \leq \sigma, \tag{7.13}$$

where h_K is the diameter of K and ρ_K is the supremum of the diameters of the balls contained in K.

It can be easily shown that this condition is equivalent to the following. There exists a constant $\theta_0 > 0$ such that for any $h > 0$ and for any $K \in \mathcal{T}_h$,

$$\theta_K \geq \theta_0,$$

where θ_K denotes the smallest angle of the triangle K.

Thus the regularity of a family of triangulations $(\mathcal{T}_h)_{h>0}$ in the sense of Definition 7.4.1 prevents the triangles from becoming "flat" in the limit when $h \to 0$.

As we shall see, this is a key assumption to obtain the convergence of the method. For example, the situation with $\varepsilon_h \to 0$ in Figure 7.5 is not allowed in the context of a regular family of triangulation (as $\varepsilon_h \to 0$, K_h becomes flat).

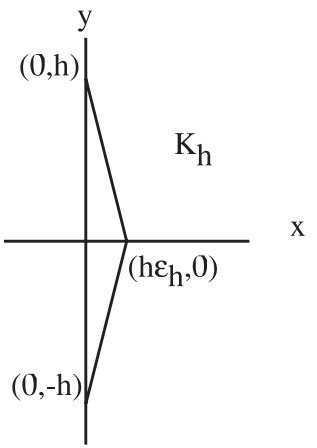


Figure 7.5. *A triangle K_h becoming flat.*

We shall return later to this example and show that in such a situation some of the following mathematical developments fail to be true.

The main result of this section, which is the convergence of the finite element method under the regularity assumption (7.13), is given below.

Theorem 7.4.1. *Let Ω be a polygon and let $(\mathcal{T}_h)_{h\to 0}$ be a regular family of triangulations of Ω. Then, the finite element method converges, i.e.,*

$$\lim_{h\to 0} \|u - u_h\|_{H_0^1(\Omega)} = 0.$$

Moreover, if u belongs to $H^2(\Omega)$, the following estimate holds: there exists some constant $C > 0$ such that for all $h > 0$

$$\|u - u_h\|_{H_0^1(\Omega)} \leq Ch\|u\|_{H^2(\Omega)}.$$

PROOF. (a) Let us first assume that $u \in H^2(\Omega)$. Recalling that $N = 2$, by Sobolev embedding theorems (see Section 5.7) we have $u \in \mathbf{C}(\overline{\Omega})$. Indeed this is true under the assumption $N \leq 3$. This allows us to talk about the value of u at any point of $\overline{\Omega}$, and especially at the nodes $(a_i)_{i=1,\dots,I(h)}$ of the triangulation $\mathcal{T}_h$. Let us introduce the function $\Pi_h(u)$ which is the continuous affine interpolant of u at the nodes of $\mathcal{T}_h$:

$$\Pi_h(u) = \sum_{i=1}^{I(h)} u(a_i)\varphi_i.$$

Recall that φ_i is the hat function related to the node a_i and that the $(\varphi_i)_{i=1,\dots,I(h)}$ form a basis of V_h. The above formula is just the linear decomposition of $\Pi_h(u)$ in the basis $(\varphi_i)_{i=1,\dots,I(h)}$. Since $\Pi_h(u) \in V_h$, by definition of $\mathrm{dist}(u, V_h)$, we have

$$\mathrm{dist}(u, V_h) \leq \|u - \Pi_h(u)\|_{H_0^1(\Omega)}.$$

This inequality, when combined with Proposition 3.1.2 (Cea's lemma) yields

$$\|u - u_h\|_{H_0^1(\Omega)} \leq \frac{M}{\alpha}\|u - \Pi_h(u)\|_{H_0^1(\Omega)}. \tag{7.14}$$

Let us now use the following approximation result that we admit for the moment (we shall return to this crucial result further): there exists a constant C independent of h such that for all $u \in H^2(\Omega)$

$$\|u - \Pi_h(u)\|_{H^1(\Omega)} \leq Ch\|u\|_{H^2(\Omega)}. \tag{7.15}$$

Let us notice that this estimate makes use in an essential way of the regularity assumption (7.13) on the family of triangulations $(\mathcal{T}_h)_{h\to 0}$ and of the fact that $u \in H^2(\Omega)$.

Let us now combine (7.14) and (7.15) to obtain

$$\|u - u_h\|_{H_0^1(\Omega)} \leq \frac{M}{\alpha}Ch\|u\|_{H^2(\Omega)}. \tag{7.16}$$

Thus, in the case $u \in H^2(\Omega)$, we have convergence of the finite element method, that is, norm convergence in $H_0^1(\Omega)$ of the sequence $(u_h)_{h\to 0}$ to u as $h \to 0$. More precisely, the estimate (7.16) provides information about the rate of convergence of the method.

(b) In the general case, that is, $u \in H_0^1(\Omega)$, one completes the proof by a density argument: for any $\varepsilon > 0$ let us introduce some $v_\varepsilon \in \mathcal{D}(\Omega) = \mathbf{C}_c^\infty(\Omega)$ such that

$$\|u - v_\varepsilon\|_{H_0^1(\Omega)} < \varepsilon. \tag{7.17}$$

Since $v_\varepsilon \in \mathcal{D}(\Omega) \subset H^2(\Omega)$, for each $\varepsilon > 0$ we can use the previous argument and, by (7.15), we have

$$\|v_\varepsilon - \Pi_h(v_\varepsilon)\|_{H_0^1(\Omega)} \leq Ch\|v_\varepsilon\|_{H^2(\Omega)}. \tag{7.18}$$

Let us now write the triangle inequality

$$\|u - \Pi_h(v_\varepsilon)\|_{H_0^1(\Omega)} \leq \|u - v_\varepsilon\|_{H_0^1(\Omega)} + \|v_\varepsilon - \Pi_h(v_\varepsilon)\|_{H_0^1(\Omega)}$$

and use inequalities (7.17) and (7.18) to obtain

$$\|u - \Pi_h(v_\varepsilon)\|_{H_0^1(\Omega)} \leq \varepsilon + Ch\|v_\varepsilon\|_{H^2(\Omega)}.$$

Since $\Pi_h(v_\varepsilon) \in V_h$, this implies

$$\operatorname{dist}(u, V_h) \leq \varepsilon + Ch\|v_\varepsilon\|_{H^2(\Omega)}.$$

Hence

$$\limsup_{h\to 0} \operatorname{dist}(u, V_h) \leq \varepsilon.$$

This being true for any $\varepsilon > 0$, we finally obtain

$$\lim_{h\to 0} \operatorname{dist}(u, V_h) = 0,$$

which, by Proposition 3.1.2 (Cea's lemma), implies the norm convergence in $H_0^1(\Omega)$ of u_h to u as $h \to 0$. □

Let us now give the proof of the piecewise affine interpolation inequality (7.15) which, together with the abstract Cea's lemma, is the key ingredient of the proof of Theorem 7.4.1. Because of its importance and its own interest let us state it independently.

Theorem 7.4.2. *Let Ω be a polygon and $(\mathcal{T}_h)_{h\to 0}$ a regular family of triangulations of Ω (i.e., (7.13) is supposed to be satisfied). Then, there exists a constant C, which is independent of h, such that for any $u \in H^2(\Omega)$*

$$\|u - \Pi_h(u)\|_{H^1(\Omega)} \leq Ch\|u\|_{H^2(\Omega)}.$$

We recall that $\Pi_h(u)$ is the piecewise affine interpolant of u relative to $\mathcal{T}_h$.

For pedagogical reasons, it is worthwhile to first prove Theorem 7.4.2 in one dimension, i.e., $\Omega = (a, b)$ is an interval of $\mathbf{R}$. Indeed, the role of the norms $H^1(\Omega)$ and $H^2(\Omega)$ and of the assumption $u \in H^2(\Omega)$ already appear quite naturally in this situation, and the proof just requires elementary tools. We shall then consider the two-dimensional case and show how, in that case, one has to do some geometrical assumptions on $\mathcal{T}_h$. (This is where the regularity assumption (7.13) on $\mathcal{T}_h$ plays a central role.)

PROOF OF THEOREM 7.4.2 IN THE ONE-DIMENSIONAL CASE. Take $\Omega = (a, b)$ be an interval of $\mathbf{R}$ with $-\infty < a < b < +\infty$. Let

$$a = a_0 < a_1 < a_2 < \cdots < a_n = b$$

be a discretization of Ω, and set $h = \max_i |a_{i+1} - a_i|$.

(a) Let us first assume that u is smooth, say, $u \in \mathbf{C}^\infty([a, b])$, and let $x \in (a_j, a_{j+1})$. The Taylor–Lagrange formula at order one yields

$$\begin{aligned}\Big(\Pi_h(u)\Big)'(x) &= \frac{u(a_{j+1}) - u(a_j)}{a_{j+1} - a_j} \\ &= u'(a_j + \theta_j)\end{aligned}$$

for some $0 < \theta_j < h$. Hence, for any $x \in (a_j, a_{j+1})$

$$\begin{aligned}\big|u'(x) - \big[\Pi_h(u)\big]'(x)\big| &= |u'(x) - u'(a_j + \theta_j)| \\ &\le \int_{a_j}^{a_{j+1}} |u''(s)|\, ds.\end{aligned}$$

Applying the Cauchy–Schwarz inequality, we obtain

$$\big|u'(x) - \big[\Pi_h(u)\big]'(x)\big|^2 \le h \int_{a_j}^{a_{j+1}} |u''(s)|^2\, ds.$$

The above inequality holds for any $x \in (a_j, a_{j+1})$. After integration on (a_j, a_{j+1}) one obtains

$$\int_{a_j}^{a_{j+1}} \big|u'(x) - \big[\Pi_h(u)\big]'(x)\big|^2\, dx \le h^2 \int_{a_j}^{a_{j+1}} |u''(s)|^2\, ds.$$

Summing the above inequality with respect to $j = 0, 1, \ldots, N-1$, finally yields

$$\int_a^b \big|u'(x) - \big[\Pi_h(u)\big]'(x)\big|^2\, dx \le h^2 \int_a^b |u''(s)|^2\, ds,$$

that is,

$$\|u' - \Pi_h(u)'\|_{L^2(\Omega)} \le h \|u''\|_{L^2(\Omega)}. \tag{7.19}$$

Let us prove that there exists some constant $C > 0$ such that

$$\|u - \Pi_h(u)\|_{L^2(\Omega)} \le Ch \|u''\|_{L^2(\Omega)}. \tag{7.20}$$

Using the same argument and notation as above, we can write for $x \in (a_j, a_{j+1})$

$$\begin{aligned}u(x) - \Pi_h(u)(x) &= u(x) - \left[u(a_j) + \frac{u(a_{j+1}) - u(a_j)}{a_{j+1} - a_j}(x - a_j)\right] \\ &= u(x) - u(a_j) - (x - a_j)u'(a_j + \theta_j) \\ &= (x - a_j)u'(a_j + \theta_{x,j}) - (x - a_j)u'(a_j + \theta_j),\end{aligned}$$

where $0 < \theta_{x,j} < x - a_j$. It follows that

$$|u(x) - \Pi_h(u)(x)| \leq (x - a_j) \int_{a_j}^{a_{j+1}} |u''(t)|\, dt,$$

which by similar arguments as above yields

$$\|u - \Pi_h(u)\|_{L^2(\Omega)} \leq \frac{h^2}{\sqrt{3}} \|u''\|_{L^2(\Omega)}.$$

Finally, by combining (7.19) and (7.20), one obtains

$$\|u - \Pi_h(u)\|_{H^1(\Omega)} \leq Ch\|u\|_{H^2(\Omega)}. \tag{7.21}$$

(b) Let us now extend the above inequality to an arbitrary $u \in H^2(\Omega)$. To that end one uses a density argument: noticing that $\mathbf{C}^\infty([a,b])$ is dense in $H^1(a,b)$ and $H^2(a,b)$, we just need to prove that the operator $\Pi_h : H^1(\Omega) \to H^1(\Omega)$ is continuous. More precisely, one can state the following result of independent interest, which concludes the proof in the one-dimensional case. □

Lemma 7.4.1. *Suppose $\Omega = (a,b)$ and $(\mathcal{T}_h)_{h\to 0}$ is a discretization of Ω. Then, there exists a constant $C > 0$ such that for any $h > 0$, for any $v \in H^1(\Omega)$,*

$$\|\Pi_h(v)\|_{H^1(\Omega)} \leq C\|v\|_{H^1(\Omega)}.$$

Proof. For any $x \in (a_j, a_{j+1})$

$$\begin{aligned}\Pi_h(v)'(x) &= \frac{v(a_j+1) - v(a_j)}{a_{j+1} - a_j} \\ &= \frac{1}{a_{j+1} - a_j} \int_{a_j}^{a_{j+1}} v'(t)\, dt\end{aligned}$$

(see Theorem 5.1.1). Hence

$$\big[\Pi_h(v)\big]'^2(x) \leq \frac{1}{(a_{j+1} - a_j)^2} \left(\int_{a_j}^{a_{j+1}} |v'(t)|\, dt \right)^2,$$

which by the Cauchy–Schwarz inequality yields

$$\big[\Pi_h(v)\big]'^2(x) \leq \frac{1}{a_{j+1} - a_j} \int_{a_j}^{a_{j+1}} |v'(t)|^2\, dt.$$

After integration on (a_j, a_{j+1}), and summation with respect to j, one obtains

$$\|\Pi_h(v)'\|_{L^2(\Omega)} \leq \|v'\|_{L^2(\Omega)}. \tag{7.22}$$

On the other hand,

$$\begin{aligned}\|\Pi_h(v)\|_{L^2(\Omega)} &\leq \sqrt{b-a}\, \|\Pi_h(v)\|_{L^\infty(\Omega)} \\ &\leq \sqrt{b-a}\, \|v\|_{L^\infty(\Omega)}.\end{aligned} \tag{7.23}$$

In the one-dimensional case ($N = 1$) it was proved in Theorem 5.1.1 that each element of $H^1(a, b)$ has a unique continuous representative. Let us verify by some elementary computation that this canonical embedding $H^1(a, b) \subset \mathbf{C}([a, b])$ is continuous, i.e., there exists some constant $C > 0$ such that

$$\forall\, v \in H^1(a, b) \qquad \|v\|_{L^\infty(a,b)} \le C\|v\|_{H^1(a,b)}. \tag{7.24}$$

Recall that we still denote $\tilde{v} = v$ the continuous representative of v and that for any $x_0, x \in [a, b]$

$$v(x_0) = v(x) + \int_{x_0}^{x} v'(t)\,dt.$$

Let us apply the Cauchy–Schwarz inequality to the above formula:

$$\begin{aligned} |v(x_0)| &\le |v(x)| + \int_{x_0}^{x} |v'(t)|\,dt \\ &\le |v(x)| + \sqrt{b-a}\left(\int_a^b |v'(t)|^2\,dt\right)^{1/2}. \end{aligned}$$

Let us now integrate this inequality with respect to $x \in [a, b]$:

$$\begin{aligned} (b-a)|v(x_0)| &\le \int_a^b |v(x)|\,dx + (b-a)^{3/2}\left(\int_a^b |v'(t)|^2\,dt\right)^{1/2} \\ &\le (b-a)^{1/2}\left(\int_a^b |v(x)|^2\,dx\right)^{1/2} + (b-a)^{3/2}\left(\int_a^b |v'(t)|^2\,dt\right)^{1/2}. \end{aligned}$$

The elementary inequality $(\alpha + \beta)^2 \le 2(\alpha^2 + \beta^2)$ now yields

$$\begin{aligned} \|v\|_{L^\infty(a,b)} &\le \sqrt{2}\left[\frac{1}{b-a}\int_a^b v(x)^2\,dx + (b-a)\int_a^b v'(x)^2\,dx\right]^{1/2} \\ &\le \sqrt{2}\max\left\{\frac{1}{b-a}, b-a\right\}^{1/2}\|v\|_{H^1(a,b)}. \end{aligned}$$

Combining (7.23) and (7.24) finally yields

$$\|\Pi_h(v)\|_{L^2(\Omega)} \le C\sqrt{b-a}\,\|v\|_{H^1(\Omega)},$$

which, together with (7.22) gives

$$\|\Pi_h(v)\|_{H^1(\Omega)} \le C\|v\|_{H^1(\Omega)},$$

and the proof of Lemma 7.4.1 is complete. □

Remark 7.4.1. Indeed one can prove the following result: for all $v \in H^1(a, b)$

$$\lim_{h\to 0} \|v - \Pi_h(v)\|_{H^1(a,b)} = 0. \tag{7.25}$$

This is slightly more precise than proving that for every $u \in H^1(a, b)$

$$\lim_{h\to 0} \operatorname{dist}(u, V_h) = 0,$$

which we have used in the proof of the convergence of the finite element method. The proof follows the lines of the previous arguments: given $v \in H^1(a,b)$, for any $\varepsilon > 0$, let $v_\varepsilon \in H^2(a,b)$ with $\|v - v_\varepsilon\|_{H^1(a,b)} < \varepsilon$. We have

$$\begin{aligned}\|v - \Pi_h(v)\|_{H^1(a,b)} &\le \|v - v_\varepsilon\|_{H^1(a,b)} + \|v_\varepsilon - \Pi_h(v_\varepsilon)\|_{H^1(a,b)} + \|\Pi_h(v - v_\varepsilon)\|_{H^1(a,b)} \\ &\le C\|v - v_\varepsilon\|_{H^1(a,b)} + Ch\|v_\varepsilon\|_{H^2(a,b)}.\end{aligned}$$

It follows that

$$\limsup_{h\to 0} \|v - \Pi_h(v)\|_{H^1(a,b)} \le C\varepsilon.$$

This being true for any $\varepsilon > 0$, the conclusion follows.

PROOF OF THEOREM 7.4.2 IN THE TWO-DIMENSIONAL CASE. Suppose now that Ω is a polygon. The proof of the basic estimate for $u \in H^2(\Omega)$,

$$\|u - \Pi_h(u)\|_{H^1(\Omega)} \le Ch\|u\|_{H^2(\Omega)}, \tag{7.26}$$

is much more involved than in the one-dimensional case.

To establish this result one needs to make some geometrical assumptions on the triangulation $\mathcal{T}_h$; this is where the regularity assumption (7.13) plays a central role. Indeed, to establish (7.26) we first are going to argue with a single triangle (the key step) and show the following result.

Theorem 7.4.3. *There exists a constant $C > 0$ such that for any triangle K, for any $u \in H^2(K)$,*

$$\|u - \Pi(u)\|_{H^2(K)} \le Ch\left(h + \frac{h}{\rho}\right)\|u\|_{H^2(K)}, \tag{7.27}$$

where $\Pi(u)$ is the affine interpolant of u at the vertices of K, h is the diameter of K, and ρ is the diameter of the largest ball contained in K.

PROOF OF THEOREM 7.4.2 (CONTINUATION). The basic estimate (7.26) and so Theorem 7.4.2 can be easily deduced from (7.27), as shown in the following. By taking the square of each member of (7.27), writing the corresponding inequalities for all the triangles of the triangulation $\mathcal{T}_h$, and then summing these inequalities, one obtains

$$\|u - \Pi_h(u)\|^2_{H^1(\Omega)} \le Ch^2\left(h + \frac{h}{\rho}\right)^2 \|u\|^2_{H^2(\Omega)}.$$

Let us now use the regularity assumption (7.13) on the triangulation ($h/\rho \le \sigma$) and take $h \le 1$ to obtain

$$\|u - \Pi_h(u)\|_{H^1(\Omega)} \le Ch(1+\sigma)\|u\|_{H^2(\Omega)},$$

which proves (7.26) □.

Thus our concern to complete the proof of the finite element method in the case $N = 2$ is to prove Theorem 7.4.3. A key idea in the process of getting the estimate (7.27) consists first in establishing such a formula for a fixed triangle $\widehat{K}$, which is used as a reference; take, for example, $\widehat{K}$ equal to the unit simplex.

Lemma 7.4.2. *Let $\widehat{K}$ be a given triangle. Then there exists a constant $C > 0$ such that for any $v \in H^2(\widehat{K})$*

$$\|v - \Pi_{\widehat{K}}(v)\|_{H^1(\widehat{K})} \leq C|D^2 v|_{L^2(\widehat{K})}, \tag{7.28}$$

where $|D^2 v|_{L^2(\widehat{K})} := \sum_{i_1+i_2=2} \int_{\widehat{K}} \left| \frac{\partial^2 v}{\partial x_1^{i_1} \partial x_2^{i_2}}(x) \right|^2 dx$.

This is the two-dimensional version of inequalities (7.19) and (7.20). At this stage, we don't need to know the precise value of the constant C, the point being just to know if such a constant exists.

Then, we shall pass from the reference triangle $\widehat{K}$ to a triangle K of $\mathcal{T}_h$ by using an affine transformation,

$$x = B\hat{x} + b = F(\hat{x}),$$

where B is an invertible matrix and $b \in \mathbf{R}^2$, which satisfies

$$K = F(\widehat{K}).$$

The geometrical properties of the triangulation will appear through this transformation.

Proof of Lemma 7.4.2. To prove (7.28), and since we don't need to know the precise value of C, we follow an analysis similar to the one in the proof of general Poincaré inequalities (Theorem 5.4.3, with the Poincaré–Wirtinger inequality as an example). The idea is to argue by contradiction and use the Rellich–Kondrakov compact embedding theorem, Theorem 5.4.2. Without ambiguity, for simplicity of the notation let us write K instead of $\widehat{K}$.

Suppose the assertion (7.28) is false. Then we could find a sequence $(v_n)_{n\in\mathbf{N}}$ such that

$$\begin{cases} v_n \in H^2(K) \quad \forall\, n \in \mathbf{N}, \\ \dfrac{1}{|D^2 v_n|_{L^2(K)}} \|v_n - \Pi_K(v_n)\|_{H^1(K)} \geq n. \end{cases}$$

Noticing that $D^2(\Pi_K(v_n)) = 0$, one can rewrite the above inequality in the following form:

$$\left| D^2 \left(\frac{v_n - \Pi_K(v_n)}{\|v_n - \Pi_K(v_n)\|_{H^1(K)}} \right) \right|_{L^2(K)} \leq \frac{1}{n}.$$

Let us introduce the function

$$u_n := \frac{v_n - \Pi_K(v_n)}{\|v_n - \Pi_K(v_n)\|_{H^1(K)}}.$$

We have

$$\begin{cases} u_n \in H^2(K), \\ \|u_n\|_{H^1(K)} = 1, \\ |D^2 u_n|_{L^2(K)} \leq 1/n, \\ u_n(a_j) = 0, \qquad j = 1, 2, 3, \text{ where } a_j \text{ are the vertices of } K. \end{cases}$$

Let us show how to obtain a contradiction from this set of properties. From $\|u_n\|_{H^1(K)} = 1$ and $|D^2 u_n|_{L^2(K)} \leq 1/n$ we obtain that the sequence $(u_n)_{n\in\mathbf{N}}$ is bounded in $H^2(K)$. Since K is bounded and piecewise $\mathbf{C}^1$ one can apply the Rellich–Kondrakov theorem, which

gives that the sequence $(u_n)_{n\in\mathbf{N}}$ is relatively compact in $H^1(K)$. We can then extract a subsequence, which we still denote u_n, such that

$$u_n \to u \quad \text{in } H^1(K).$$

Hence $\|u\|_{H^1(K)} = 1$. On the other hand, from $|D^2 u_n|_{L^2(K)} \le 1/n$, we obtain that

$$D^2 u_n \to 0 = D^2 u \quad \text{in } L^2(K).$$

Hence u is an affine function on K. The linear map $u \mapsto u(a_j)$ from $H^2(K)$ into $\mathbf{R}$ is continuous. Since $u_n \to u$ in $H^2(K)$ we have $u(a_j) = 0,\ j = 1, 2, 3$. The only affine function on K which is zero at the vertices is the function $u = 0$. This is a contradiction with $\|u\|_{H^1(K)} = 1$. This establishes (7.28) and concludes the proof of Lemma 7.4.2. □

Let us now consider an affine invertible transformation

$$F(\hat{x}) = B\hat{x} + b \tag{7.29}$$

with $K = F(\widehat{K})$ and examine how it affects the formula (7.28).

The following notation and definitions will be helpful. We use the mappings

$$\hat{x} \in \widehat{K} \overset{F}{\longmapsto} F(\hat{x}) = x \in K$$

and $F^{-1} : K \to \widehat{K}$, which is the inverse of F. To each function v defined on K one can associate the function $\hat{v} : \widehat{K} \to \mathbf{R}$, which is defined by

$$\hat{v}(\hat{x}) = v(F(\hat{x})),$$

which, with the above notation, gives

$$\hat{v}(\hat{x}) = v(x).$$

Recall that $F(\hat{x}) = B\hat{x} + b$ is an affine invertible map. The spectral norms of B and B^{-1} will play a central role in the following. Recall that these norms are defined by

$$\|B\| = \sup\big\{|B\xi| \ :\ |\xi| = 1\big\},$$
$$\|B^{-1}\| = \sup\big\{|B^{-1}\xi| \ :\ |\xi| = 1\big\},$$

where $|\xi|$ is the Euclidean norm.

The geometrical characteristic properties of the triangulation (h and ρ) do appear naturally in the evaluation of these norms. Let us denote by h_K and ρ_K (respectively, $h_{\widehat{K}}$ and $\rho_{\widehat{K}}$) the geometrical characteristic numbers of K (respectively, $\widehat{K}$) as defined in (7.13).

Lemma 7.4.3. *The following estimates hold:*

$$\|B\| \le \frac{h_K}{\rho_{\widehat{K}}}, \qquad \|B^{-1}\| \le \frac{h_{\widehat{K}}}{\rho_K}.$$

PROOF. We have

$$\begin{aligned}\|B\| &= \sup\big\{|B\xi| \ :\ |\xi| = 1\big\} \\ &= \frac{1}{\rho_{\widehat{K}}} \sup\big\{|B\xi| \ :\ |\xi| = \rho_{\widehat{K}}\big\}.\end{aligned}$$

By definition of $\rho_{\widehat{K}}$, for any ξ with $|\xi| = \rho_{\widehat{K}}$ one can find two points $\hat{x}_1$ and $\hat{x}_2$ in $\widehat{K}$ such that

$$\xi = \hat{x}_2 - \hat{x}_1.$$

Hence

$$\begin{aligned} B\xi &= B\hat{x}_2 - B\hat{x}_1 \\ &= F\hat{x}_2 - F\hat{x}_1 \\ &= x_2 - x_1, \end{aligned}$$

where x_2 and x_1 belong to $K = F(\widehat{K})$.

By definition of h_K, which is the diameter of K, we have

$$|B\xi| = |x_2 - x_1| \leq h_K.$$

This inequality being true for any ξ with $|\xi| = \rho_{\widehat{K}}$, we deduce

$$\|B\| \leq \frac{h_K}{\rho_{\widehat{K}}}.$$

The other inequality is obtained in a similar way, by reversing the role of K and $\widehat{K}$. □

The other basic ingredient of the proof is the change of variables in the integrals which are equal to Sobolev norms. Let us write, for a given integer $m \geq 0$

$$|v|_{m,K} = \left(\sum_{|\alpha|=m} \int_K |\partial^\alpha v(x)|^2 \, dx \right)^{1/2}. \tag{7.30}$$

Lemma 7.4.4. *Let K and $\widehat{K}$ be two finite elements which are affine equivalent, that is, $K = F(\widehat{K})$ with $F(\hat{x}) = B\hat{x} + b$ and B affine invertible. If a function v belongs to the space $H^m(K)$ for some integer $m \geq 0$, then the function $\hat{v} = v \circ F$ belongs to $H^m(\widehat{K})$ and there is a constant $C(m) > 0$ such that*

$$\forall v \in H^m(K) \qquad |\hat{v}|_{m,\widehat{K}} \leq C(m)\|B\|^m |\det B|^{-1/2} |v|_{m,K}.$$

Analogously, one has

$$\forall \hat{v} \in H^m(\widehat{K}) \qquad |v|_{m,K} \leq C(m)\|B^{-1}\|^m |\det B|^{1/2} |\hat{v}|_{m,K}.$$

Proof. By standard density arguments, one just needs to argue with $v \in \mathbf{C}^\infty(\overline{K})$. Hence $\hat{v} \in \mathbf{C}^\infty(\overline{\widehat{K}})$. It is convenient to introduce the first and second derivatives of v: then $Dv(x)$ is a linear form and

$$\frac{\partial v}{\partial x_i}(x) = Dv(x) \cdot e_i,$$

where (e_i) are the vectors of the canonical basis in $\mathbf{R}^N$ (here $N = 2$). Similarly, $D^2v(x)$ is the bilinear symmetric form associated to the Hessian matrix and

$$\frac{\partial^2 v}{\partial x_i \partial x_j}(x) = D^2v(x) \cdot (e_i, e_j).$$

One can unify these two situations (and much more) by writing for any multi-index $\alpha = (\alpha_1, \alpha_2)$ with length $|\alpha| = m$

$$\partial^\alpha v(x) = D^m v(x)(e_1, \dots, e_1, e_2, \dots, e_2),$$

where e_1 is repeated α_1 times, and e_2 is repeated α_2 times. Recall that $\partial^\alpha v(x) = (\partial^{|\alpha|} v / \partial x_1^{\alpha_1} \partial x_2^{\alpha_2})(x)$. Set

$$\|D^m v(x)\| = \sup\big\{|D^m v(x)(\xi_1, \dots, \xi_m)| \ : \ |\xi_i| \le 1, \ 1 \le i \le m\big\}.$$

Then

$$|\partial^\alpha v(x)| \le \|D^m v(x)\|$$

and

$$\begin{aligned} |v|_{m,K} &= \bigg(\int_K \sum_{|\alpha|=m} |\partial^\alpha v(x)|^2 \, dx \bigg)^{1/2} \\ &\le C_1 \bigg(\int_K \|D^m v(x)\|^2 \, dx \bigg)^{1/2}, \end{aligned} \tag{7.31}$$

where C_1^2 is the cardinal of the set of indices α such that $|\alpha| = m$, i.e., $C_1 = C_1(m)$. We can now perform the differentiation rule for composition of functions. Recalling that $\hat{v}(\hat{x}) = v(F(\hat{x})) = v(B\hat{x} + b)$ we have

$$D^m \hat{v}(\hat{x})(\xi_1, \dots, \xi_m) = D^m v(x)(B\xi_1, \dots, B\xi_m)$$

so that

$$\|D^m \hat{v}(\hat{x})\| \le \|D^m v(x)\| \|B\|^m.$$

Taking the square and after integration on $\widehat{K}$, one obtains

$$\int_{\widehat{K}} \|D^m \hat{v}(\hat{x})\|^2 \, d\hat{x} \le \|B\|^{2m} \int_{\widehat{K}} \|D^m v(F(\hat{x}))\|^2 \, d\hat{x}.$$

Using the formula of change of variables in multiple integrals we get

$$\int_{\widehat{K}} \|D^m \hat{v}(\hat{x})\|^2 \, d\hat{x} \le \|B\|^{2m} |\det(B^{-1})| \int_K \|D^m v(x)\|^2 \, dx. \tag{7.32}$$

Combining (7.31) and (7.32) we obtain

$$|\hat{v}|_{m,\widehat{K}} \le C_1(m) \|B\|^m |\det B|^{-1/2} \bigg(\int_K \|D^m v(x)\|^2 \, dx \bigg)^{1/2}.$$

Since conversely there exists a constant $C_2(m)$ such that

$$\bigg(\int_K \|D^m v(x)\|^2 \, dx \bigg)^{1/2} \le C_2(m) |v|_{m,K}, \tag{7.33}$$

we finally obtain

$$|\hat{v}|_{m,\widehat{K}} \le C(m) \|B\|^m |\det B|^{-1/2} |v|_{m,K}$$

with $C(m) = C_1(m)C_2(m)$.
Reversing the role of K and $\widehat{K}$ yields the other inequality. □

END OF THE PROOF OF THEOREM 7.4.3. We now have all the ingredients of the proof of the basic estimate (7.27) in Theorem 7.4.3:

$$\|u - \Pi(u)\|_{H^1(K)} \leq Ch\left(h + \frac{h}{\rho}\right)\|u\|_{H^2(K)}.$$

Let $u \in H^2(K)$. By Lemma 7.4.4 we have for $m = 0, 1$

$$|u - \Pi(u)|_{m,K} \leq C\|B^{-1}\|^m |\det B|^{1/2} |\hat{u} - \widehat{\Pi(u)}|_{m,\widehat{K}}. \tag{7.34}$$

The estimate (7.28) on $\widehat{K}$ yields for $m = 0, 1$

$$|\hat{u} - \widehat{\Pi(u)}|_{m,\widehat{K}} \leq C|\hat{u}|_{2,\widehat{K}}. \tag{7.35}$$

Applying again Lemma 7.4.4 we have

$$|\hat{u}|_{2,\widehat{K}} \leq C\|B\|^2 |\det B|^{-1/2} |u|_{2,K}. \tag{7.36}$$

Combining (7.34), (7.35), and (7.36) we obtain

$$|u - \Pi(u)|_{m,K} \leq C\|B^{-1}\|^m \|B\|^2 |u|_{2,K}. \tag{7.37}$$

Take the square of inequality (7.37) and sum over $m = 0, 1$ to obtain

$$\|u - \Pi(u)\|_{H^1(K)} \leq C\|B\|^2\big(1 + \|B^{-1}\|\big)|u|_{2,K}.$$

Using Lemma 7.4.3 we finally get

$$\begin{aligned}\|u - \Pi(u)\|_{H^1(K)} &\leq C\frac{h_K^2}{\rho_{\widehat{K}}^2}\left(1 + \frac{h_{\widehat{K}}}{\rho_K}\right)|u|_{2,K} \\ &\leq C\left(h_K^2 + \frac{h_K^2}{\rho_K}\right)|u|_{2,K}\end{aligned}$$

where $1/\widehat{\rho_K}$ and $\widehat{h_K}$ have been included in the constant C. Recall that $\widehat{K}$ is a fixed reference triangle. Noticing that $|u_{2,K}| \leq \|u\|_{H^2(K)}$, the proof is complete. □

Remark 7.4.2. Note that we have obtained a slightly more precise result than (7.27); indeed, we proved that for every $u \in H^2(K)$

$$\|u - \Pi_h(u)\|_{H^1(K)} \leq Ch\left(h + \frac{h}{\rho}\right)|u|_{2,K},$$

where $|u|_{2,K} = |D^2u|_{L^2(K)}$ just involves the L^2-norm of the second-order partial derivatives of u. Consequently, in Theorem 7.4.2, we have that for any $u \in H^2(\Omega)$

$$\|u - \Pi_h(u)\|_{H^1(\Omega)} \leq Ch|u|_{2,K}.$$

Similarly, in Theorem 7.4.1, we have that if $u \in H^2(\Omega)$,

$$\|u - u_h\|_{H^1(\Omega)} \leq Ch|u|_{2,K}.$$

7.5 Complements

7.5.1 Flat triangles

Let us return to the situation described in Figure 7.5, which illustrates a family of triangulations $(\mathcal{T}_h)_{h\to 0}$ involving triangles $K_h \in \mathcal{T}_h$ becoming flat as $h \to 0$. Let us show that on such triangles the affine interpolate can lead to significant errors.

Take a simple function which is not affine, for example, a quadratic function

$$u(x, y) = y^2.$$

Let us compute the affine interpolate $\Pi_h(u)$ of u on the triangle K_h whose vertices are $(0, \frac{h}{2})$, $(0, -\frac{h}{2})$, and $(h\varepsilon_h, 0)$. We have that $\Pi_h(u)$ vanishes at $(h\varepsilon_h, 0)$ and is equal to $h^2/4$ at the two other vertices. Hence

$$\frac{\partial}{\partial x}\Pi_h(u) = -\frac{h^2/4}{h\varepsilon_h} = -\frac{h}{4\varepsilon_h}.$$

Since $\frac{\partial u}{\partial x} = 0$ we obtain $|\frac{\partial}{\partial x}(u - \Pi_h(u))| = \frac{h}{4\varepsilon_h}$ and

$$\left(\int_{K_h} \left|\frac{\partial}{\partial x}(u - \Pi_h(u)\right|^2 (x)\, dx\right)^{1/2} = \frac{h^2}{4\sqrt{2}\sqrt{\varepsilon_h}}.$$

On the other hand, we have $\frac{\partial^2 u}{\partial x^2} = \frac{\partial^2 u}{\partial x \partial y} = 0$ and $\frac{\partial^2 u}{\partial y^2} = 2$, which give

$$|u|_{2,K_h} = \left(\int_{K_h} 4\, dx\, dy\right)^{1/2} = \sqrt{2}h\sqrt{\varepsilon_h}.$$

An inequality of the type

$$\|u - \Pi_h(u)\|_{H^1(K_h)} \le Ch|u|_{2,K_h}$$

would then imply

$$\left\|\frac{\partial}{\partial x}(u - \Pi_h(u)\right\|_{L^2(K_h)} \le Ch|u|_{2,K},$$

that is,

$$\frac{h^2}{4\sqrt{2}\sqrt{\varepsilon_h}} \le Ch\sqrt{2}h\sqrt{\varepsilon_h},$$

which is equivalent to

$$\inf_{h>0} \varepsilon_h > 0.$$

Thus, the convergence analysis developed in this chapter fails to be true without any geometrical assumption on the family of triangulations preventing the triangles from becoming flat.

7.5.2 $H^2(\Omega)$ regularity of the solution of the Dirichlet problem on a convex polygon

In the model situation studied in this section, we chose to take as Ω a polygon in $\mathbf{R}^2$, to make as simple as possible the description of the triangulation in the finite element method. (Otherwise, for general Ω one has to approximate it by such polygonal sets Ω_h.)

Conversely, we have a difficulty, which is to know if the solution u of the Dirichlet problem with $f \in L^2(\Omega)$

$$\begin{cases} -\Delta u = f \text{ on } \Omega, \\ u = 0 \text{ on } \partial\Omega, \end{cases}$$

satisfies the property $u \in H^2(\Omega)$. Indeed the estimate

$$\|u - u_h\|_{H_0^1(\Omega)} \leq h$$

has been established under the assumption $u \in H^2(\Omega)$.

We are in a situation where Ω is a polygon, its boundary is not smooth (it is only piecewise $\mathbf{C}^1$ or Lipschitz continuous), and the classical Agmon–Douglis–Nirenberg theorem which asserts that $u \in H^2(\Omega)$ under the assumption that Ω is of class $\mathbf{C}^2$ does not apply.

The answer to this question is quite involved. It was studied by Grisvard in [150], [151], who proved in particular that if Ω is a polygon which is supposed to be convex, then $u \in H^2(\Omega)$.

7.5.3 Finite element methods of type P_2

The method which has been developed in $\mathbf{R}^2$ with Ω a polygon and finite elements which are triangles can be naturally extended to $\mathbf{R}^3$ when replacing triangles by tetrahedrons. Functions of the approximating subspaces V_h are continuous and piecewise affine. This is what we call a finite element method of type P_1 (by reference to the degree of the polynomial functions which are used). To improve the quality of the approximation, one may naturally think to enrich the approximating subspaces and make them contain more functions. This can be done, for example, by considering functions which are piecewise polynomial of degree less than or equal to two. Let us briefly describe an example of such a finite element method of type P_2.

Take Ω a polygon in $\mathbf{R}^2$ and a given triangulation $\mathcal{T}_h$ of Ω. We introduce the space

$$V_h = \left\{ v \in \mathbf{C}(\overline{\Omega}) \ : \ v \lfloor K \in P_2 \text{ for every } K \in \mathcal{T}_h \right\},$$

where P_2 is the family of polynomial functions on $\mathbf{R}^2$ of degree less than or equal to two. The general form of an element $p \in P_2$ is then

$$p(x, y) = a + bx + cy + dx^2 + exy + fy^2,$$

and one can verify that P_2 is a vector space of dimension equal to 6. Then, to fix an element $p \in P_2$, it is not sufficient to give its values at the vertices of a triangle (as it was the case for $p \in P_1$): we need to give its values at six points carefully choosen. Take, for example,

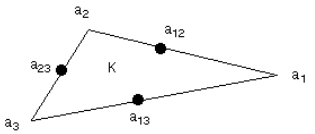


Figure 7.6. *Six points on the triangle K.*

the case of Figure 7.6, where a_1, a_2, a_3 are the vertices of K and $a_{ij} = \frac{1}{2}(a_i + a_j)$ are the midpoints of the edges of K.

This choice leads to triangulations whose nodes are the vertices of the triangles and the midpoints of all the edges. As an illustration consider the case of Figure 7.7.

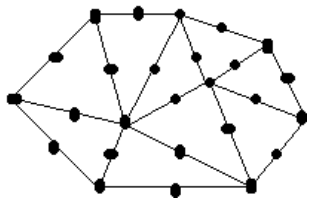


Figure 7.7. *A triangulation for finite element method of type P_2.*

Let us denote by (N_j) the nodes of $\mathcal{T}_h$, $j = 1, \ldots, I(h)$ (vertices + midpoints of edges). It is quite elementary to verify that

(a) V_h is a subspace of dimension $I(h)$ of $H^1(\Omega)$ and any function v of V_h is uniquely determined by its values at the nodes of the triangulation,

(b) a basis of V_h is given by the family of functions $(p_j)_{j=1,\ldots,I(h)}$ which is defined by

$$p_j \in V_h,$$
$$p_j(N_i) = \delta_{ij} \quad \text{for } i, j = 1, \ldots, I(h).$$

For any element $v \in V_h$ one has

$$v(x) = \sum_{j=1}^{I(h)} v(N_j)\, p_j(x).$$

The finite element method can now be developed in a way parallel to what we did before. Indeed, as expected, one can get a better order in the approximation by piecewise P_2 functions. Let us denote by

$$\Pi_h(v) = \sum_{j=1}^{I(h)} v(N_j)\, p_j$$

the element of V_h obtained by interpolation of v on the nodes of the triangulation ($\Pi_h(v) = v$ on the nodes). Then, one can show the following result (which we do not prove): if the family of triangulations $(\mathcal{T}_h)_{h\to 0}$ is regular, then there exists a constant $C > 0$ such that for any $u \in H^3(\Omega)$,

$$\|u - \Pi_h(u)\|_{H^1(\Omega)} \leq Ch^2 |u|_{3,\Omega}.$$

Chapter 8

Spectral Analysis of the Laplacian

8.1 Introduction

From the very beginning of the 19th century, the study of the eigenvalue problem for the Laplace equation emerged as a fundamental topic in the theory of partial differential equations. In 1922, J.B.J. Fourier was faced with this question to develop the so-called separation of variables method. Let us illustrate this method in the case of the wave equation with Dirichlet boundary data, which, for example, gives a model for the vibrations of an elastic membrane which is clamped on its boundary.

Given initial data $u_0, u_1 : \Omega \to \mathbf{R}$, one looks for a solution $u : Q = \Omega \times (0, +\infty) \to \mathbf{R}$ of the boundary value problem

$$\begin{cases} \dfrac{\partial^2 u}{\partial t^2} - \Delta u = 0 & \text{on } Q, \\ u = 0 & \text{on } \Sigma = \partial\Omega \times (0, +\infty), \\ u(x, 0) = u_0(x) & \text{on } \Omega, \\ \dfrac{\partial u}{\partial t}(x, 0) = u_1(x) & \text{on } \Omega. \end{cases}$$

The idea is to look for a solution u of the form

$$u(x, t) = w(x)\varphi(t), \tag{8.1}$$

where the dependence of u with respect to (x, t) has been separated. The wave equation then becomes

$$w(x)\varphi''(t) - \varphi(t)\Delta w(x) = 0$$

or, equivalently,

$$\frac{\varphi''(t)}{\varphi(t)} = \frac{\Delta w(x)}{w(x)}. \tag{8.2}$$

Since the left-hand side of (8.2) is a function only of t and the right-hand side only of x, this forces these two expressions to be constant, that is,

$$\frac{\varphi''(t)}{\varphi(t)} = \frac{\Delta w(x)}{w(x)} = -\lambda$$

for some constant λ.

This method leads to the study of the spectral problem for the so-called Laplace–Dirichlet operator,

$$\begin{cases} -\Delta w = \lambda w \quad \text{on } \Omega, \\ w = 0 \quad \text{on } \partial\Omega, \end{cases} \tag{8.3}$$

and the resolution of the ordinary differential equation

$$\varphi''(t) + \lambda\varphi(t) = 0. \tag{8.4}$$

One can easily verify that the eigenvalues of the spectral problem (8.3) are positive (it is enough to multiply by w and integrate by parts on Ω). Therefore, the solutions of (8.4) are of the following form:

$$\varphi(t) = A\cos\sqrt{\lambda}t + B\sin\sqrt{\lambda}t.$$

The question is now, can one, by linear combinations of such separate solutions, obtain a solution

$$u(x,t) = \sum_i w_i(x)\Big(A\cos\sqrt{\lambda_i}t + B\sin\sqrt{\lambda_i}t\Big) \tag{8.5}$$

which satisfies the initial data $u(x,0) = u_0(x)$ and $(\partial u/\partial t)(x,t) = u_1(x)$?

Indeed, this question is intimately related to the possibility of generating any given function by linear combinations (indeed, series!) of eigenvectors of the Laplace–Dirichlet operator.

We shall give a positive answer to this question and prove the following theorem (which is the main result of this chapter). Assume Ω is a bounded open set in $\mathbf{R}^N$. Then, there exists a complete orthonormal system of eigenvectors of the Lapalce–Dirichlet operator in the space $L^2(\Omega)$. A complete orthonormal system is also called a Hilbertian basis. This is a deep result of Rellich which is precisely based on the Rellich–Kondrakov theorem (compact embedding of $H_0^1(\Omega)$ into $L^2(\Omega)$; see Theorem 5.3.3) and on the abstract spectral decomposition theorem for compact self-adjoint operators. Indeed, variational methods play a central role in the theory of eigenvalues of elliptic partial differential equations. Another striking result in this direction is the variational characterization of eigenvalues of the Laplace–Dirichlet operator. One of the formulas provided by the Courant–Fisher min-max principle is the following: the first eigenvalue of the Laplace–Dirichlet operator is given by the variational formula

$$\lambda_1(-\Delta) = \min\left\{ \int_\Omega |\nabla v(x)|^2\,dx \ : \ v \in H_0^1(\Omega), \int_\Omega v(x)^2\,dx = 1 \right\}.$$

This formula and its companions provide powerful tools for studying the properties of the eigenvalues and eigenvectors of the Laplace–Dirichlet operator. In 1911, Weyl used this principle to solve the problem on the asymptotic distribution of the eigenvalues of the Laplace–Dirichlet operator. We shall briefly describe such a result in Section 8.5.

In the last two decades, spectral methods, just like finite element methods, have proved to be very efficient in the numerical treatment of some partial differential operators. They give raise to approximation methods where the finite dimensional approximating spaces V_n are based on (orthogonal) polynomials of high degree (by contrast with finite element methods, where the degree is fixed, one for P_1 method, two for P_2 method, for example).

Here, the degree of polynomials of V_n increases with n. These methods provide accurate approximations of the solution, which are limited only by the regularity of the solution. But on the counterpart, there are some restrictions on the geometry of Ω.

For pedagogical reasons and simplicity of exposition, we restrict our attention to the spectral analysis of the Laplace equation with Dirichlet boundary condition. In Section 8.6, we shall briefly survey some straight extensions of these results.

8.2 The Laplace–Dirichlet operator: Functional setting

Our objective is to study the eigenvalue problem for the Laplace equation on Ω with Dirichlet boundary conditions on $\partial\Omega$. We seek for $\lambda \in \mathbf{R}$ and $u \neq 0$ such that

$$\begin{cases} -\Delta u = \lambda u & \text{on } \Omega, \\ u = 0 & \text{on } \partial\Omega. \end{cases} \tag{8.6}$$

In all this chapter, Ω is assumed to be a bounded open set in $\mathbf{R}^N$.

Let us give a precise meaning to this definition and write its variational formulation.

Definition 8.2.1. *We say that $\lambda \in \mathbf{R}$ is an eigenvalue of the Laplace–Dirichlet operator if there exists some $u \in H_0^1(\Omega)$, $u \neq 0$ such that*

$$\begin{cases} \displaystyle\int_\Omega \nabla u(x) \cdot \nabla v(x)\, dx = \lambda \int_\Omega u(x)v(x)\, dx \quad \forall v \in H_0^1(\Omega), \\ u \in H_0^1(\Omega). \end{cases} \tag{8.7}$$

When such u exists it is called an eigenvector related to the eigenvalue λ.

Remark 8.2.1. Let us make some comments to the definition above:

(a) if (8.7) is satisfied, then

$$\begin{cases} -\Delta u = \lambda u & \text{in the distribution sense,} \\ u = 0 & \text{in the trace sense,} \end{cases}$$

i.e., (8.6) is satisfied in a weak sense. The next step consists in proving that u is regular and hence it is a classical solution of (8.6).

(b) One may wonder whether "$\lambda \in \mathbf{R}$" is not too restrictive, and take instead $\lambda \in \mathbf{C}$. Indeed, by taking $v = u$ in (8.7) one obtains

$$\lambda = \frac{\displaystyle\int_\Omega |\nabla u|^2\, dx}{\displaystyle\int_\Omega u^2(x)\, dx},$$

which implies that all the eigenvalues of the Laplace-Dirichlet operator are positive real numbers.

Let us now come to the central idea which will permit us to formulate the above problem in terms of classical operator theory.

The classical theory for spectral analysis of operators in infinite dimensional spaces works with operators

$$T : H \longrightarrow H,$$

where H is a Hilbert space and $T \in L(H)$ is a linear, continuous, and compact operator from H into H.

One cannot write $-\Delta$ in such a setting because, as for any differential operator, there is a loss of regularity when passing from u to $-\Delta u$. Indeed, $-\Delta$ is a linear continuous operator

$$-\Delta : H_0^1(\Omega) \longrightarrow H^{-1}(\Omega).$$

Another way to treat $-\Delta$ is to consider it as an operator from $L^2(\Omega)$ into $L^2(\Omega)$, but with a domain, i.e.,

$$\mathrm{dom}(-\Delta) = H^2(\Omega) \cap H_0^1(\Omega).$$

By contrast, and that is the central idea, the inverse operator $T = (-\Delta)^{-1}$ is a nice operator which fits well with the classical theory. The straight relation which connects the spectrum of an operator and the spectrum of its inverse operator will permit us to conclude our analysis. Let us now define the operator T as the inverse of the Laplace–Dirichlet operator.

Definition 8.2.2. *The inverse of the Laplace–Dirichlet operator is the operator* $T : L^2(\Omega) \longrightarrow L^2(\Omega)$ *which is defined for every* $h \in L^2(\Omega)$ *by the following:* $Th \in H_0^1(\Omega) \subset L^2(\Omega)$ *is the unique solution of the variational problem*

$$\begin{cases} \displaystyle\int_\Omega \nabla(Th)(x) \cdot \nabla v(x)\, dx = \int_\Omega h(x) v(x)\, dx \quad \forall v \in H_0^1(\Omega), \\ Th \in H_0^1(\Omega). \end{cases}$$

Equivalently, Th is the variational solution of the Dirichlet problem (see Theorem 5.1.1),

$$\begin{cases} -\Delta(Th) = h \quad \text{on } \Omega, \\ Th = 0 \quad \text{on } \partial\Omega. \end{cases}$$

At this point let us notice that

$$T : L^2(\Omega) \longrightarrow H_0^1(\Omega).$$

To consider T as a linear continuous operator from a space H into itself we have two possibilities: either

$$T : L^2(\Omega) \longrightarrow L^2(\Omega)$$

or

$$T : H_0^1(\Omega) \longrightarrow H_0^1(\Omega).$$

These two approaches lead to similar parallel developments. We choose to consider T as acting from $L^2(\Omega)$ into $L^2(\Omega)$. We then have

$$(-\Delta) \circ T = id_H, \qquad H = L^2(\Omega),$$

i.e., T is the right inverse of $-\Delta$.

The introduction of $T = (-\Delta)^{-1}$ is justified in the context of the spectral analysis of the Laplace–Dirichlet operator by the following result.

Lemma 8.2.1. *The real number λ is an eigenvalue of the Laplace–Dirichlet operator iff $1/\lambda$ is an eigenvalue of $T = (-\Delta)^{-1}$.*

PROOF. Let us assume that λ is an eigenvalue of the Laplace–Dirichlet operator, i.e., there exists some $u \in H_0^1(\Omega)$, $u \neq 0$, such that

$$\int_\Omega \nabla u \cdot \nabla v \, dx = \lambda \int_\Omega uv \, dx \quad \forall v \in H_0^1(\Omega).$$

By definition of $T = (-\Delta)^{-1}$ this is equivalent to

$$u = T(\lambda u).$$

By linearity of T (this is proved in the next proposition), and using the fact that $\lambda \neq 0$, we deduce

$$T(u) = \frac{1}{\lambda} u,$$

i.e., $1/\lambda$ is an eigenvalue of T. □

The following properties of the operator T will play a central role in its spectral analysis.

Proposition 8.2.1. *The operator T satisfies the following properties:*

(i) *$T : L^2(\Omega) \longrightarrow L^2(\Omega)$ is a linear continuous operator,*

(ii) *T is self-adjoint in $L^2(\Omega)$,*

(iii) *T is compact from $L^2(\Omega)$ into $L^2(\Omega)$,*

(iv) *T is positive definite.*

PROOF. (i)$_1$ Take $h_1,\ h_2 \in L^2(\Omega)$ and $\alpha_1, \alpha_2 \in \mathbf{R}$. By definition of T

$$\int_\Omega \nabla(Th_1) \cdot \nabla v \, dx = \int_\Omega h_1 v \, dx \quad \forall v \in H_0^1(\Omega),$$

$$\int_\Omega \nabla(Th_2) \cdot \nabla v \, dx = \int_\Omega h_2 v \, dx \quad \forall v \in H_0^1(\Omega).$$

By taking a linear combination of these two equalities we obtain

$$\begin{cases} \displaystyle\int_\Omega \nabla(\alpha_1 Th_1 + \alpha_2 Th_2) \cdot \nabla v \, dx = \int_\Omega (\alpha_1 h_1 + \alpha_2 h_2) v \, dx \quad \forall v \in H_0^1(\Omega), \\ \alpha_1 Th_1 + \alpha_2 Th_2 \in H_0^1(\Omega). \end{cases}$$

By uniqueness of the solution of the Dirichlet problem, we get

$$T(\alpha_1 h_1 + \alpha_2 h_2) = \alpha_1 T h_1 + \alpha_2 T h_2,$$

which is the linearity of T.

(i)$_2$ Let us now prove that $T : L^2(\Omega) \longrightarrow L^2(\Omega)$ is continuous. By definition of T, the equality

$$\int_\Omega \nabla(Th) \cdot \nabla v \, dx = \int_\Omega hv \, dx$$

holds true for any $v \in H_0^1(\Omega)$. In particular, it is satisfied by $v = Th \in H_0^1(\Omega)$, which gives

$$\int_\Omega |\nabla(Th)|^2 \, dx = \int_\Omega hT(h) \, dx. \tag{8.8}$$

Using the Cauchy–Schwarz inequality, we obtain

$$\int_\Omega |\nabla(Th)|^2 \, dx \le \left(\int_\Omega h^2 \, dx \right)^{1/2} \left(\int_\Omega (Th)^2 \, dx \right)^{1/2}. \tag{8.9}$$

Let us now use the Poincaré inequality and the fact that Ω is bounded (Theorem 5.3.1). There exists a constant $C > 0$ which depends only on Ω such that

$$\forall v \in H_0^1(\Omega) \quad \left(\int_\Omega v(x)^2 \, dx \right)^{1/2} \le C \left(\int_\Omega |\nabla v(x)|^2 \, dx \right)^{1/2}.$$

In particular, since for every $h \in L^2(\Omega)$ we have $Th \in H_0^1(\Omega)$, we can write the Poincaré inequality with $v = Th$, which gives

$$\left(\int_\Omega |Th(x)|^2 \, dx \right)^{1/2} \le C \left(\int_\Omega |\nabla(Th(x))|^2 \, dx \right)^{1/2}. \tag{8.10}$$

Let us combine inequalities (8.9) and (8.10) to obtain

$$\int_\Omega |\nabla(Th)|^2 \, dx \le C \left(\int_\Omega h^2 \, dx \right)^{1/2} \left(\int_\Omega |\nabla(Th)|^2 \, dx \right)^{1/2}.$$

Equivalently

$$\left(\int_\Omega |\nabla(Th)|^2 \, dx \right)^{1/2} \le C \left(\int_\Omega h^2 \, dx \right)^{1/2}, \tag{8.11}$$

which, together with (8.10), gives

$$\left(\int_\Omega |Th(x)|^2 \, dx \right)^{1/2} \le C^2 \left(\int_\Omega h^2 \, dx \right)^{1/2}. \tag{8.12}$$

Thus, we have obtained

$$\forall h \in L^2(\Omega) \quad \|Th\|_{L^2(\Omega)} \le C^2 \|h\|_{L^2(\Omega)}, \tag{8.13}$$

i.e., T is a linear continuous operator from $L^2(\Omega)$ into $L^2(\Omega)$. Indeed, we have obtained a sharper result: from (8.11) and (8.12) we deduce that

$$\forall h \in L^2(\Omega) \qquad \|Th\|_{H_0^1(\Omega)} \le C\sqrt{1 + C^2} \|h\|_{L^2(\Omega)}, \tag{8.14}$$

i.e., $T : L^2(\Omega) \longrightarrow H_0^1(\Omega)$ is a linear continuous operator. This proves that one can also treat T as a linear continuous operator from $H_0^1(\Omega)$ into $H_0^1(\Omega)$.

(ii) Let us now prove that T is a self-adjoint operator in $L^2(\Omega)$, i.e.,

$$\forall g, h \in L^2(\Omega) \quad \langle Th, g\rangle_{L^2(\Omega)} = \langle h, Tg\rangle_{L^2(\Omega)},$$

which means

$$\int_\Omega (Th)(x)g(x)\,dx = \int_\Omega h(x)(Tg)(x)\,dx.$$

By definition of Th and Tg we have

$$\int_\Omega \nabla(Th)\cdot\nabla v\,dx = \int_\Omega hv\,dx \quad \forall v \in H_0^1(\Omega),$$

$$\int_\Omega \nabla(Tg)\cdot\nabla v\,dx = \int_\Omega gv\,dx \quad \forall v \in H_0^1(\Omega).$$

Take $v = Tg \in H_0^1(\Omega)$ in the first equality and $v = Th \in H_0^1(\Omega)$ in the second equality. We obtain

$$\int_\Omega \nabla(Th)\cdot\nabla(Tg)\,dx = \int_\Omega hT(g)\,dx = \langle h, Tg\rangle_{L^2(\Omega)},$$

$$\int_\Omega \nabla(Tg)\cdot\nabla(Th)\,dx = \int_\Omega gT(h)\,dx = \langle g, Th\rangle_{L^2(\Omega)}.$$

Hence

$$\langle h, Tg\rangle_{L^2(\Omega)} = \langle g, Th\rangle_{L^2(\Omega)} = \int_\Omega \nabla(Th)\cdot\nabla(Tg)\,dx,$$

which shows that T is self-adjoint in $L^2(\Omega)$.

(iii) T is compact from $L^2(\Omega)$ into $L^2(\Omega)$.

Take B a bounded set in $L^2(\Omega)$. By (8.14), since T is linear and continuous from $L^2(\Omega)$ into $H_0^1(\Omega)$, the set $T(B)$ is bounded in $H_0^1(\Omega)$. We now use the fact that Ω is bounded and the Rellich–Kondrakov theorem, Theorem 5.3.3, to conclude that $T(B)$ is relatively compact in $L^2(\Omega)$.

(iv) T is positive definite. By (8.8) we have

$$\forall v \in L^2(\Omega) \quad \langle Th, h\rangle = \int_\Omega |\nabla(Th)|^2\,dx \geq 0,$$

that is, T is positive. Moreover, if $\langle Th, h\rangle = 0$, then $\nabla(Th) = 0$, that is, Th is locally constant. Since $Th \in H_0^1(\Omega)$, this forces Th to be equal to zero. Coming back to the definition of Th, $Th = 0$ means that $\int_\Omega hv\,dx = 0$ for all $v \in H_0^1(\Omega)$. By the density of $H_0^1(\Omega)$ into $L^2(\Omega)$, we conclude that $h = 0$. □

We can summarize the results of this section and say that the operator $T = (-\Delta)^{-1}$ is a linear continuous, self-adjoint, compact, positive operator from $L^2(\Omega)$ into $L^2(\Omega)$. This will allow us to obtain in the next section the spectral decomposition of the Laplace–Dirichlet operator.

8.3 Existence of a Hilbertian basis of eigenvectors of the Laplace–Dirichlet operator

Let us first recall the well-known (see, for instance, Brezis [90]) abstract "diagonalization" theorem for compact self-adjoint positive definite operators.

Theorem 8.3.1. *Let us assume that H is a separable Hilbert space with* $\dim H = +\infty$. *Let $T : H \longrightarrow H$ be a linear continuous self-adjoint compact and positive definite operator. Then we have the following:*

(i) *T is diagonalizable: there exists a Hilbertian basis of eigenvectors of T.*

(ii) *The set $\Lambda(T)$ of eigenvalues of T is countable. It can be written as a sequence $(\mu_n)_{n\in\mathbf{N}}$ of positive distinct real numbers that decreases to zero as $n \to +\infty$*

$$0 \leftarrow \mu_n < \cdots < \mu_3 < \mu_2 < \mu_1.$$

(iii) *For each $\mu_n \in \Lambda(T)$, $E_{\mu_n} = \ker(T - \mu_n I)$ is a finite dimensional subspace of H: it is the eigensubspace relative to the eigenvalue μ_n. Its dimension is called the multiplicity of μ_n.*

(iv) *For all $\mu_i \neq \mu_j,\ \mu_i, \mu_j \in \Lambda(T),\ E_{\mu_i} \perp E_{\mu_j}$ (orthogonal subspaces).*

(v) *$H = \oplus_{n\in\mathbf{N}} E_{\mu_n}$, i.e.,*

$$\forall x \in H \qquad x = \sum_{n\in\mathbf{N}} proj_{E_{\mu_n}}(x)$$

and

$$\forall x \in H \qquad Tx = \sum_{n\in\mathbf{N}} \mu_n proj_{E_{\mu_n}}(x).$$

Remark 8.3.1. The situation described in the above statement is simplified by the fact that here we have assumed T be positive definite, i.e., $\ker T = 0$, which allows us to avoid considering $\mu = 0$ in the spectral decomposition (the null space $\ker T$ may be infinite dimensional).

Sketch of the proof of Theorem 8.3.1. It is worthwhile to recall some of the basic ingredients of the proof of Theorem 8.3.1.

(a) First notice that $\Lambda(T)$ is a bounded subset of $(0, +\infty)$: if $\mu \in \Lambda(T)$, then there exists some $u \in H$, $u \neq 0$ such that

$$Tu = \mu u.$$

Hence

$$\langle Tu, u\rangle = \mu |u|^2.$$

Since $u \neq 0$ we have $\langle Tu, u\rangle > 0$ (T is positive definite), which forces μ to be positive. On the other hand, since T is linear continuous,

$$\mu |u|^2 \leq |Tu|_H |u|_H \leq \|T\|_{L(H,H)} |u|^2,$$

which gives

$$0 < \mu \leq \|T\|_{L(H,H)}.$$

(b) Now take $\mu \neq \nu$ with $\mu, \nu \in \Lambda(T)$. By definition of $\Lambda(T)$ and E_μ, E_ν we have

$$\begin{aligned} &\forall h \in E_\mu \quad Th = \mu h, \\ &\forall k \in E_\nu \quad Tk = \nu k. \end{aligned}$$

We deduce

$$\begin{aligned} \langle Th, k\rangle &= \mu\langle h, k\rangle, \\ \langle Tk, h\rangle &= \nu\langle k, h\rangle. \end{aligned}$$

Since T is self-adjoint, we have $\langle Th, k\rangle = \langle h, Tk\rangle$. Hence

$$(\mu - \nu)\langle h, k\rangle = 0.$$

We have assumed $\mu \neq \nu$. This forces h and k to satisfy $\langle h, k\rangle = 0$, that is, $E_\mu \perp E_\nu$.

(c) It is interesting to see where the compactness on T comes into play. Let us notice that for any $\mu \in \Lambda(T)$ the subspace E_μ is closed and invariant by T; if $h \in E_\mu$, i.e., $Th = \mu h$, then $T(Th) = \mu(Th)$, i.e., $Th \in E_\mu$. Hence $T : E_\mu \longrightarrow E_\mu$ and E_μ is a Hilbert space for the induced structure of H. Moreover, if B_{E_μ} denotes the unit ball in E_μ, we have

$$T(B_{E_\mu}) = \mu B_{E_\mu}.$$

Since $\mu \neq 0$ and T is compact, this forces B_{E_μ} to be relatively compact, that is, $\dim E_\mu < +\infty$. □

We have obtained the following decomposition of H as a Hilbertian sum of eigenspaces:

$$H = \bigoplus_{n\in\mathbf{N}} E_{\mu_n}. \tag{8.15}$$

To derive from this formula a Hilbertian basis of H we have to pick up in each E_{μ_n} an orthonormal basis whose cardinal is equal to the (finite) multiplicity of μ_n.

To keep a quite simple notation we adopt the following convention.

Definition 8.3.1. *We now decide to count the eigenvalues of T according to their multiplicity, that is,*

μ_1 is repeated k_1 times where k_1 is the multiplicity of μ_1

...

μ_n is repeated k_n times where k_n is the multiplicity of μ_n

...

and so on. Clearly, in this way, we obtain a sequence of positive real numbers, which we still denote by $(\mu_n)_{n\in\mathbf{N}}$, such that

$$0 \leftarrow \mu_n \leq \cdots \mu_3 \leq \mu_2 \leq \mu_1.$$

Note that now the μ_i are not necessarily distinct.

The convention above allows us to pick up an orthonormal basis in each finite dimensional eigensubspace to obtain a Hilbertian basis $(h_n)_{n\in\mathbf{N}}$ in H which satisfies

$$Th_n = \mu_n h_n$$

for every $n \in \mathbf{N}$.

Let us now come back to our model example. Clearly, by Proposition 8.2.1, the operator $T = (-\Delta)^{-1}$ which is considered as acting from $H = L^2(\Omega)$ into $L^2(\Omega)$ satisfies all the conditions of the abstract diagonalization theorem, Theorem 8.3.1. Thus, there exists a Hilbertian basis $(e_n)_{n\in\mathbf{N}}$ of $L^2(\Omega)$ such that for each $n \in \mathbf{N}$ e_n is an eigenvector of T. More precisely,

$$Te_n = \mu_n e_n,$$

and $(\mu_n)_{n\in\mathbf{N}}$ is a sequence of positive numbers which decreases to zero. By Lemma 8.2.1, we deduce that $1/\mu_n$ is an eigenvalue of the Laplace–Dirichlet operator and that e_n is a corresponding eigenvector. This means

$$\begin{cases} -\Delta e_n = \frac{1}{\mu_n} e_n & \text{on } \Omega, \\ e_n = 0 & \text{on } \partial\Omega, \end{cases}$$

the solution e_n being taken in the variational sense, i.e., $e_n \in H_0^1(\Omega)$ and

$$\int_\Omega \nabla e_n \cdot \nabla v \, dx = \frac{1}{\mu_n} \int_\Omega e_n v \, dx \quad \forall v \in H_0^1(\Omega).$$

Indeed, it is immediate to verify that the above equality is equivalent to $T(e_n/\mu_n) = e_n$, that is, $T(e_n) = \mu_n e_n$.

Let us now forget the operator $T = (-\Delta)^{-1}$ which was just a technical ingredient in our study and convert the previous results directly in terms of $-\Delta$, the Laplace–Dirichlet operator. Noticing that the sequence $(\lambda_n)_{n\in\mathbf{N}}$ with $\lambda_n = 1/\mu_n$ is now an increasing sequence of positive numbers which tends to $+\infty$ as $n \to +\infty$, we obtain the following theorem.

Theorem 8.3.2. *The Laplace–Dirichlet operator has a countable family of eigenvalues $(\lambda_n)_{n\in\mathbf{N}}$ which can be written as an increasing sequence of positive numbers which tends to $+\infty$ as $n \to +\infty$:*

$$0 < \lambda_1 \le \lambda_2 \le \cdots \le \lambda_n \le \cdots.$$

Each eigenvalue is repeated a number of times equal to its multiplicity (which is finite). There exists a Hilbertian basis $(e_n)_{n\in\mathbf{N}}$ of $L^2(\Omega)$ such that for each $n \in \mathbf{N}$, e_n is an eigenvector of the Laplace–Dirichlet operator relatively to the eigenvalue λ_n:

$$\begin{cases} -\Delta e_n = \lambda_n e_n & \text{on } \Omega, \\ e_n = 0 & \text{on } \partial\Omega. \end{cases}$$

We already mentioned that the spectral analysis of the Laplace–Dirichlet operator could have been, as well, developed in the space $H_0^1(\Omega)$. Indeed, there is a direct link between these two approaches, which is described below.

Proposition 8.3.1. *The family $\left(e_n/\sqrt{\lambda_n}\right)_{n\in\mathbf{N}}$ is a Hilbertian basis of the space $H_0^1(\Omega)$ equipped with the scalar product $\langle u, v\rangle = \int_\Omega \nabla u \cdot \nabla v \, dx$.*

PROOF. (a) Let us start from the definition of e_n:

$$\int_\Omega \nabla e_n \cdot \nabla v \, dx = \lambda_n \int_\Omega e_n v \, dx \quad \forall v \in H_0^1(\Omega).$$

By taking $v = e_n \in H_0^1(\Omega)$, we obtain

$$\int_\Omega |\nabla e_n|^2 \, dx = \lambda_n \int_\Omega e_n^2 \, dx = \lambda_n.$$

Hence

$$\|e_n\|^2_{H_0^1(\Omega)} = \lambda_n$$

and

$$\left\| \frac{e_n}{\sqrt{\lambda_n}} \right\|_{H_0^1(\Omega)} = 1.$$

(b) Let us verify the orthogonality property in $H_0^1(\Omega)$: take $n \neq m$

$$\begin{aligned} \left\langle \frac{e_n}{\sqrt{\lambda_n}}, \frac{e_m}{\sqrt{\lambda_m}} \right\rangle_{H_0^1(\Omega)} &= \frac{1}{\sqrt{\lambda_n \lambda_m}} \int_\Omega \nabla e_n \cdot \nabla e_m \, dx \\ &= \frac{1}{\sqrt{\lambda_n \lambda_m}} \lambda_n \int_\Omega e_n e_m \, dx, \end{aligned}$$

which is equal to zero because $(e_n)_{n \in \mathbf{N}}$ is an orthogonal system in $L^2(\Omega)$.

(c) Let us verify that $\left(e_n/\sqrt{\lambda_n}\right)_{n \in \mathbf{N}}$ generates a vector space which is dense in $H_0^1(\Omega)$. Equivalently, we have to verify that if $f \in H_0^1(\Omega)$ is such that, for all $n \in \mathbf{N}$,

$$\left\langle f, \frac{e_n}{\sqrt{\lambda_n}} \right\rangle_{H_0^1(\Omega)} = 0,$$

then $f = 0$. Let us notice that

$$\begin{aligned} \left\langle f, \frac{e_n}{\sqrt{\lambda_n}} \right\rangle_{H_0^1(\Omega)} &= \frac{1}{\sqrt{\lambda_n}} \int_\Omega \nabla f \cdot \nabla e_n \, dx \\ &= \sqrt{\lambda_n} \int_\Omega f e_n \, dx. \end{aligned}$$

Since $\lambda_n \neq 0$, our assumption becomes $\int_\Omega f e_n \, dx = 0$ for all $n \in \mathbf{N}$, which clearly implies $f = 0$ because $(e_n)_{n \in \mathbf{N}}$ is an orthonormal basis in $L^2(\Omega)$. □

8.4 The Courant–Fisher min-max and max-min formulas

Let us start with the variational characterization of the first eigenvalue $\lambda_1(-\Delta)$ of the Laplace–Dirichlet operator. Without ambiguity, we write λ_1. To introduce this result,

let us notice that if λ is an eigenvalue of $-\Delta$, then there exists some $u \in H_0^1(\Omega)$, $u \neq 0$ such that

$$\int_\Omega \nabla u \cdot \nabla v \, dx = \lambda \int_\Omega uv \, dx \quad \forall v \in H_0^1(\Omega).$$

By taking $v = u$ and noticing that $u \neq 0$ we obtain

$$\lambda = \frac{\displaystyle\int_\Omega |\nabla u(x)|^2 \, dx}{\displaystyle\int_\Omega u(x)^2 \, dx}. \tag{8.16}$$

The above expression plays a central role in the variational approach to eigenvalue problems for the Laplace equation.

Definition 8.4.1. *For any $v \in H_0^1(\Omega)$, $v \neq 0$, let us write*

$$\mathcal{R}(v) = \frac{\displaystyle\int_\Omega |\nabla v(x)|^2 \, dx}{\displaystyle\int_\Omega v(x)^2 \, dx}.$$

$\mathcal{R} : H_0^1(\Omega) \longrightarrow \mathbf{R}^+$ is called the Rayleigh quotient.

From (8.16) we immediately obtain that for any eigenvalue λ of the Laplace–Dirichlet operator,

$$\lambda \geq \inf \left\{ \mathcal{R}(v) \ : \ v \in H_0^1(\Omega), \ v \neq 0 \right\},$$

which is equivalent to saying

$$\lambda_1 \geq \inf \left\{ \mathcal{R}(v) \ : \ v \in H_0^1(\Omega), \ v \neq 0 \right\}. \tag{8.17}$$

Indeed, there is equality between these two expressions. That is the object of the following theorem.

Theorem 8.4.1 (Courant–Fisher formula). *Assume Ω is a bounded open subset of $\mathbf{R}^N$. The first eigenvalue λ_1 of the Laplace–Dirichlet operator on Ω is given by the following variational formula:*

$$\lambda_1 = \min \left\{ \frac{\displaystyle\int_\Omega |\nabla v(x)|^2 \, dx}{\displaystyle\int_\Omega v^2(x) \, dx} \ : \ v \in H_0^1(\Omega), \ v \neq 0 \right\}.$$

Moreover, the infimum above is achieved and the solutions of this variational problem are the eigenvectors relative to the first eigenvalue λ_1.

Proof. By (8.17) we need only to prove the inequality

$$\inf \left\{ \mathcal{R}(v) \ : \ v \in H_0^1(\Omega), \ v \neq 0 \right\} \geq \lambda_1,$$

that is,

$$\forall v \in H_0^1(\Omega),\ v \neq 0, \quad \mathcal{R}(v) \geq \lambda_1. \tag{8.18}$$

The idea is to express, for any $v \in H_0^1(\Omega)$, $\mathcal{R}(v)$ in a Hilbertian basis of eigenvectors of $(-\Delta)$. Indeed, we know by Theorem 8.3.2 and Proposition 8.3.1 that there exists a Hilbertian basis $(e_n)_{n\in\mathbf{N}}$ of $L^2(\Omega)$ such that e_n is an eigenvector of $-\Delta$ relatively to the eigenvalue λ_n and that $\big(e_n/\sqrt{\lambda_n}\big)$ is a corresponding Hilbertian basis in $H_0^1(\Omega)$.

By using the Bessel–Parceval inequality respectively in $H_0^1(\Omega)$ and $L^2(\Omega)$, we have

$$\|v\|^2_{H_0^1(\Omega)} = \int_\Omega |\nabla v(x)|^2\,dx = \sum_{n=1}^{+\infty} \langle v,\, \frac{e_n}{\sqrt{\lambda_n}}\rangle^2_{H_0^1(\Omega)}, \tag{8.19}$$

$$\|v\|^2_{L^2(\Omega)} = \int_\Omega v(x)^2\,dx = \sum_{n=1}^{+\infty} \langle v, e_n\rangle^2_{L^2(\Omega)}. \tag{8.20}$$

One can easily compare these two quantities by using that e_n is an eigenvalue of $-\Delta$ relatively to λ_n. Indeed, (8.19) gives

$$\begin{aligned}
\|v\|^2_{H_0^1(\Omega)} &= \sum_{n=1}^{+\infty} \frac{1}{\lambda_n}\Big(\int_\Omega \nabla v\cdot\nabla e_n\,dx\Big)^2 \\
&= \sum_{n=1}^{+\infty} \frac{\lambda_n^2}{\lambda_n}\Big(\int_\Omega e_n v\,dx\Big)^2 = \sum_{n=1}^{+\infty} \lambda_n \langle v, e_n\rangle^2_{L^2(\Omega)} \\
&\geq \lambda_1 \sum_{n=1}^{+\infty} \langle v, e_n\rangle^2_{L^2(\Omega)} = \lambda_1 \|v\|^2_{L^2(\Omega)},
\end{aligned}$$

where in the last equality we use (8.20). Hence, $\mathcal{R}(v) \geq \lambda_1$ for any $v \in H_0^1(\Omega),\ v \neq 0$, which concludes the first part of the theorem.

Let us now notice that by (8.16), any eigenvector v relative to the eigenvalue λ_1 is a solution of the Courant–Fisher minimization problem. Conversely, if v is such a solution, by the above Parceval equalities (8.19), (8.20) we must have

$$\mathcal{R}(v) = \frac{\|v\|^2_{H_0^1(\Omega)}}{\|v\|^2_{L^2(\Omega)}} = \frac{\sum_{n=1}^{+\infty} \lambda_n \langle v, e_n\rangle^2}{\sum_{n=1}^{+\infty} \langle v, e_n\rangle^2} = \lambda_1, \tag{8.21}$$

i.e.,

$$\sum_{n=1}^{+\infty} (\lambda_n - \lambda_1)\langle v, e_n\rangle^2 = 0.$$

This forces $\langle v, e_n\rangle$ to be equal to zero for all indices $n \in \mathbf{N}^*$ such that $\lambda_n \neq \lambda_1$. Equivalently, v must belong to the eigenspace relative to λ_1. □

Another elegant and direct approach to the Courant–Fisher formula relies on the theory of Lagrange multipliers. Let us start with the following equivalent formulation of the Courant–Fisher variational formula for λ_1:

$$\inf\left\{ \int_\Omega |\nabla v(x)|^2\,dx \ :\ v \in H_0^1(\Omega),\ \int_\Omega v(x)^2\,dx = 1 \right\}. \tag{$\mathcal{P}$}$$

Problem $(\mathcal{P})$ is now seen as a constrained minimization problem, namely, the minimization of the Dirichlet energy functional on the unit sphere of $L^2(\Omega)$. That is where Lagrange multipliers naturally come into play!

Let us first notice that by using the direct methods of the calculus of variations and the Rellich–Kondrakov theorem, one can easily prove that problem $(\mathcal{P})$ admits a solution u. To write some optimality condition satisfied by such solution u we take $V = H_0^1(\Omega)$ with $\langle u, v\rangle = \int_\Omega \nabla u \cdot \nabla v \, dx$ and $\|v\|^2 = \langle v, v\rangle$, and we consider the functionals

$$F(v) = \int_\Omega |\nabla v(x)|^2 \, dx = \|v\|^2,$$
$$G(v) = \int_\Omega v^2(x) \, dx.$$

Then $(\mathcal{P})$ is equivalent to

$$\min\big\{F(v) \ : \ G(v) = 1, \ v \in V\big\}.$$

The sphere $S = \big\{v \in V \ : \ G(v) = 1\big\}$ is a submanifold of class $\mathbf{C}^1$ and codimension 1 in V. Indeed, for any u, v we have, as $t \to 0$,

$$\frac{1}{t}\big[G(u + tv) - G(u)\big] \longrightarrow 2\int_\Omega uv \, dx.$$

Let us interpret this limit as a linear continuous form on $H_0^1(\Omega)$:

$$\int_\Omega uv \, dx = \langle h, v\rangle_{H_0^1(\Omega)} = \int_\Omega \nabla h \cdot \nabla v \, dx \qquad \forall v \in H_0^1(\Omega)$$

means

$$\begin{cases} -\Delta h = u & \text{on } \Omega, \\ h = 0 & \text{on } \partial\Omega, \end{cases}$$

i.e., $h = Tu$, where $T = (-\Delta)^{-1}$. Hence G is Frechet differentiable on V and $\nabla G(u) = 2Tu$.

The theory of Lagrange multipliers applies in our situation and we have that there exists $\mu \in \mathbf{R}$ such that

$$\nabla F(u) = \mu \nabla G(u),$$

that is,

$$u = \mu T u.$$

Equivalently,

$$\begin{cases} -\Delta u = \mu u & \text{on } \Omega, \\ u = 0 & \text{on } \partial\Omega, \end{cases}$$

which implies

$$\mu = \mathcal{R}(u) = \frac{\displaystyle\int_\Omega |\nabla u(x)|^2 \, dx}{\displaystyle\int_\Omega u(x)^2 \, dx} = \lambda_1.$$

Let us summarize the previous results in the following statement.

Proposition 8.4.1. *The first eigenvalue $\lambda_1(-\Delta)$ of the Laplace–Dirichlet operator is a Lagrange multiplier of the constrained minimization problem*

$$\min\left\{\int_\Omega |\nabla v(x)|^2\,dx \;:\; v\in H_0^1(\Omega),\ \int_\Omega v^2(x)\,dx=1\right\}.$$

The above approach provides a direct variational proof of the existence of an eigenvalue (indeed, the first $\lambda_1(-\Delta)$) of the Laplace–Dirichlet operator. The whole theory can then be developed in this way, by using a recursive argument: in the next step we can apply the same argument in the orthogonal subspace of V_1 (which is the eigenspace relative to λ_1) and so on. Let us make this precise in the following statement.

Proposition 8.4.2. *Let $0<\lambda_1\le\lambda_2\le\cdots\le\lambda_n\le\cdots$ be the sequence of the eigenvalues of the Laplace–Dirichlet operator (repeated in accordance with their multiplicities) and $(e_n)_{n\in\mathbf{N}}$ a corresponding Hilbertian basis of eigenvectors in $L^2(\Omega)$:*

$$\begin{cases} -\Delta e_n=\lambda_n e_n & \text{on } \Omega,\\ e_n=0 & \text{on } \partial\Omega. \end{cases}$$

Let us denote by V_n the subspace of $H_0^1(\Omega)$ generated by the first eigenvectors $e_1,\dots,e_n$,

$$V_n=\operatorname{span}\{e_1,e_2,\dots,e_n\},$$

and by $V_n^\perp$ the orthogonal of V_n in $H_0^1(\Omega)$ with respect to the scalar product of $H_0^1(\Omega)$:

$$\langle u,v\rangle=\int_\Omega \nabla u\cdot\nabla v\,dx.$$

Then the following variational formulas hold:

$$\begin{aligned}
&\lambda_1=\min\left\{\mathcal{R}(v)\;:\;v\in H_0^1(\Omega),\ v\neq 0\right\},\\
&\lambda_2=\min\left\{\mathcal{R}(v)\;:\;v\in V_1^\perp,\ v\neq 0\right\}\\
&\dots\\
&\lambda_n=\min\left\{\mathcal{R}(v)\;:\;v\in V_{n-1}^\perp,\ v\neq 0\right\}\\
&\dots.
\end{aligned}$$

Proof. By definition of λ_n and e_n

$$\int_\Omega \nabla e_n\cdot\nabla v\,dx=\lambda_n\int_\Omega e_n v\,dx\quad\forall v\in H_0^1(\Omega).$$

By taking $v=e_n$ we obtain

$$\mathcal{R}(e_n)=\lambda_n,\quad n=1,2,\dots.$$

Hence, by noticing that $e_n\in V_{n-1}^\perp$, we first obtain

$$\lambda_n\ge\inf\left\{\mathcal{R}(v)\;:\;v\in V_{n-1}^\perp,\ v\neq 0\right\}. \tag{8.22}$$

To obtain the reverse inequality, we use an argument similar to the one used in the proof of Theorem 8.4.1, which relies on the Bessel–Parceval equality. Recall that (see (8.21)) for any $v \in H_0^1(\Omega)$

$$\mathcal{R}(v) = \frac{\sum_{i=1}^{+\infty} \lambda_i \langle v, e_i \rangle^2_{L^2(\Omega)}}{\sum_{i=1}^{+\infty} \langle v, e_i \rangle^2_{L^2(\Omega)}}$$

and that $\langle v, e_i \rangle_{H_0^1(\Omega)} = \lambda_i \langle v, e_i \rangle^2_{L^2(\Omega)}$, which allows to pass from orthogonality in $H_0^1(\Omega)$ to orthogonality in $L^2(\Omega)$ and vice versa.

Hence, for any $v \in V_{n-1}^{\perp}$ we have

$$\mathcal{R}(v) = \frac{\sum_{i \geq n} \lambda_i \langle v, e_i \rangle^2_{L^2(\Omega)}}{\sum_{i \geq n} \langle v, e_i \rangle^2_{L^2(\Omega)}} \geq \lambda_n.$$

Therefore

$$\inf \left\{ \mathcal{R}(v) \ : \ v \in V_{n-1}^{\perp}, \ v \neq 0 \right\} \geq \lambda_n. \tag{8.23}$$

Combining inequalities (8.22) and (8.23) yields the result. The fact that the infimum is actually attained follows from the direct methods of the calculus of variations, being the functional $v \mapsto \mathcal{R}(v)$ coercive and weakly lower semicontinuous. □

We can now state the Courant–Fisher min-max and max-min formulas.

Theorem 8.4.2 (Courant–Fisher min-max and max-min formulas). *With the same notation as in Proposition* 8.4.2, *we have*

$$\begin{aligned} \lambda_n &= \min_{M \in \mathcal{L}_n} \max_{v \in M, \ v \neq 0} \mathcal{R}(v) \\ &= \max_{M \in \mathcal{L}_{n-1}} \min_{v \in M^{\perp}, \ v \neq 0} \mathcal{R}(v), \end{aligned}$$

where $\mathcal{L}_n$ is the class of all n-dimensional linear subspaces in $H_0^1(\Omega)$ and $M^{\perp}$ stands for the orthogonal subspace of M in $H_0^1(\Omega)$.

PROOF. (a) Take first $M = V_n = \text{span}\{e_1, \ldots, e_n\}$. For any $v \in V_n$, we have $\langle v, e_i \rangle_{H_0^1(\Omega)} = \langle v, e_i \rangle_{L^2(\Omega)}$ for all $i > n$. By expressing $\mathcal{R}(v)$ as (see (8.21))

$$\mathcal{R}(v) = \frac{\sum_{i=1}^{+\infty} \lambda_i \langle v, e_i \rangle^2_{L^2(\Omega)}}{\sum_{i=1}^{+\infty} \langle v, e_i \rangle^2_{L^2(\Omega)}},$$

we obtain

$$\mathcal{R}(v) = \frac{\sum_{i=1}^{n} \lambda_i \langle v, e_i \rangle^2_{L^2(\Omega)}}{\sum_{i=1}^{n} \langle v, e_i \rangle^2_{L^2(\Omega)}} \leq \lambda_n.$$

Hence

$$\max \left\{ \mathcal{R}(v) \ : \ v \in V_n, \ v \neq 0 \right\} \leq \lambda_n \tag{8.24}$$

(indeed equality holds taking $v = e_n$) and consequently

$$\min_{M\in\mathcal{L}_n}\max_{v\in M, v\neq 0}\mathcal{R}(v) \leq \lambda_n.$$

Let us now prove the reverse inequality. Equivalently, we have to prove that for any n-dimensional subspace M of $H_0^1(\Omega)$,

$$\lambda_n \leq \max\left\{\mathcal{R}(v) \ : \ v \in M, \ v \neq 0\right\}. \tag{8.25}$$

Take a subspace M of $H_0^1(\Omega)$ with $\dim M = n$. We claim that

$$M \cap V_{n-1}^{\perp} \neq \{0\}.$$

This is a consequence of the classical relation linking the dimension of the image and the kernel of a linear mapping: take $P : M \longrightarrow V_{n-1}$ to be the linear mapping which, to any $v \in M$, associates $P(v) = \mathrm{proj}_{V_{n-1}} v$, the projection of v on V_{n-1}. We have

$$\dim M = n = \dim(\ker P) + \dim(P(M)).$$

Since $P(M)$, the image of M per P, is contained in V_{n-1}, we have

$$\dim(P(M)) \leq n-1.$$

Hence

$$\dim(\ker P) \geq n - (n-1) = 1.$$

Equivalently, there exists some $v \in M$, $v \neq 0$ such that $\mathrm{proj}_{V_{n-1}} v = 0$, that is, $v \in M \cap V_{n-1}^{\perp}$, which proves the claim.

Take now any $\overline{v} \in M \cap V_{n-1}^{\perp}$, $\overline{v} \neq 0$. We know by Proposition 8.4.2 that

$$\lambda_n = \min\left\{\mathcal{R}(v) \ : \ v \in V_{n-1}^{\perp}\right\}.$$

Hence

$$\begin{aligned}\lambda_n &\leq \mathcal{R}(\overline{v})\\ &\leq \max\left\{\mathcal{R}(v) \ : \ v \in M, \ v \neq 0\right\},\end{aligned}$$

which proves (8.25) and completes the proof of the min-max formula.

(b) The proof of the max-min formula is very similar to the proof of the min-max formula.

First note that by Proposition 8.4.2,

$$\lambda_n = \min\left\{\mathcal{R}(v) \ : \ v \in V_{n-1}^{\perp}, \ v \neq 0\right\}. \tag{8.26}$$

Hence

$$\lambda_n \leq \sup_{M\in\mathcal{L}_{n-1}}\inf_{v\in M^{\perp}, \ v\neq 0}\mathcal{R}(v). \tag{8.27}$$

To prove the reverse inequality, we need to show that for any $M \in \mathcal{L}_{n-1}$ we have

$$\lambda_n \geq \inf_{v\in M^{\perp}, \ v\neq 0}\mathcal{R}(v). \tag{8.28}$$

We claim that there exists some $\overline{v} \in M^{\perp} \cap V_n$ with $\overline{v} \neq 0$. This can be obtained by considering the linear mapping $Q : V_n \longrightarrow M$ which, to any $v \in V_n$, associates $Q(v) = \mathrm{proj}_M v$. We have

$$n = \dim V_n = \dim(\ker Q) + \dim\big(Q(V_n)\big).$$

Since $\dim Q(V_n) \leq \dim M = n - 1$, we have $\dim(\ker Q) \geq 1$. Equivalently, there exists some $\overline{v} \in V_n$, $\overline{v} \neq 0$ such that $\mathrm{proj}_M \overline{v} = 0$, that is $\overline{v} \in M^{\perp} \cap V_n$.

We now use (8.24) to obtain

$$\begin{aligned} \lambda_n &\geq \mathcal{R}(\overline{v}) \\ &\geq \inf_{v \in M^{\perp},\ v \neq 0} \mathcal{R}(v), \end{aligned}$$

that is, (8.28). Hence $\lambda_n = \sup_{M \in \mathcal{L}_{n-1}} \inf_{v \in M^{\perp},\ v \neq 0} \mathcal{R}(v)$.

Moreover, by (8.26) we have that the sup is a max (it is precisely attained by taking $M = V_{n-1}$) and the inf is a min (take $v = e_n$). Finally,

$$\lambda_n = \max_{M \in \mathcal{L}_{n-1}} \min_{v \in M^{\perp}\ v \neq 0} \mathcal{R}(v),$$

which ends the proof. □

Remark 8.4.1. (1) It is worth pointing out that the Courant–Fisher min-max and max-min principles, which give a variational characterization of the eigenvalues of the Laplace–Dirichlet operator, hold for the sequence $(\lambda_n)_{n \in \mathbf{N}}$ of eigenvalues which is expressed according to the multiplicity condition. This is another justification of this convention which gives the information on the values of the eigenvalues and on their multiplicities.

For these reasons, we call the $(\lambda_n)_{n \in \mathbf{N}}$ with the multiplicity condition (i.e., λ_n is repeated a number of times equal to its multiplicity) *the sequence of eigenvalues of the Laplace–Dirichlet operator*.

(2) In Proposition 8.4.2, the $(\lambda_n)_{n \in \mathbf{N}}$ are obtained by a recursive formula: one has first to know V_{n-1} to obtain λ_n. By contrast, the Courant–Fisher min-max and max-min principles provide a direct variational formulation of the eigenvalues of the Laplace–Dirichlet operator.

(3) The Courant–Fisher min-max principle is the point of departure for the Ljusternik–Schnirelman theory of critical points. Indeed, in 1930, Ljusternik wrote, "The theory of eigenvalues of quadratic form developed by R. Courant enables one to discern their existence and reality without calculations. We shall generalize their theory to arbitrary functions having continuous second partial derivatives."

Typically, the Ljusternik theory deals with the variational approach to nonlinear eigenvalue problems of the type

$$F'(u) = \lambda u, \quad u \in X,\ \lambda \in \mathbf{R},\ \|u\| = 1,$$

where X is a separable Hilbert space, $\dim X = +\infty$, $F : X \longrightarrow \mathbf{R}$ is even, of class $\mathbf{C}^1$, with F' compact.

Let us conclude this section by making a direct connnection between the first eigenvalue $\lambda_1(-\Delta)$ of the Laplace–Dirichlet operator and the Poincaré constant (cf. Definition 5.3.1). Let us recall that the Poincaré constant is the smallest constant C such that for any $v \in H_0^1(\Omega)$

$$\left(\int_\Omega v(x)^2 \, dx \right)^{1/2} \leq C \left(\int_\Omega |\nabla v(x)|^2 \, dx \right)^{1/2}.$$

This is equivalent to saying that

$$\frac{1}{C^2} = \inf \left\{ \mathcal{R}(v) \ : \ v \in H_0^1(\Omega), \ v \neq 0 \right\},$$

i.e., $1/C^2 = \lambda_1$. In other words, we have obtained the following result.

Proposition 8.4.3. *The Poincaré constant C and the first eigenvalue λ_1 of the Laplace-Dirichlet operator are related by the following formula:*

$$\frac{1}{C^2} = \lambda_1.$$

8.5 Multiplicity and asymptotic properties of the eigenvalues of the Laplace–Dirichlet operator

The first eigenvalue $\lambda_1(-\Delta)$ plays a fundamental role, for example, in the analysis of the resonance phenomena for vibrating structures and in some related shape optimization problems. Indeed, the first eigenvalue $\lambda_1(-\Delta)$ enjoys remarkable properties as stated in the following result.

Theorem 8.5.1. *Let Ω be a bounded connected regular open set in $\mathbf{R}^N$. The first eigenvalue λ_1 of the Laplace–Dirichlet operator has multiplicity equal to one. Its eigenspace is generated by a vector $e_1 \in H_0^1(\Omega)$ such that $e_1 > 0$ on Ω.*

Proof. Let us denote by E_1 the eigenspace relative to the first eigenvalue λ_1. We recall that the Courant–Fisher theorem, Theorem 8.4.1, asserts that the elements $v \in E_1, \ v \neq 0$, are the solutions of the minimization problem

$$\min \left\{ \frac{\displaystyle\int_\Omega |\nabla v(x)|^2 \, dx}{\displaystyle\int_\Omega v(x)^2 \, dx} \ : \ v \in H_0^1(\Omega), \ v \neq 0 \right\}. \qquad (\mathcal{P})$$

An important consequence of this formula is that if $v \in E_1$, then automatically $|v| \in E_1$. This follows from the fact that the truncations operate on the space $H_0^1(\Omega)$. In particular, see Corollary 5.8.1, for any $v \in H_0^1(\Omega), \ |v| \in H_0^1(\Omega)$, and $\mathcal{R}(|v|) = \mathcal{R}(v)$.

The following argument proceeds by contradiction and makes use of the strong maximum principle. Suppose that one can find two elements v_1 and v_2 in the eigensubspace E_1 which are not proportional.

Because of the regularity assumption on Ω, v_1 and v_2 are smooth functions and it makes sense to consider their values at any point $x \in \Omega$. Thus we can find x_0 and $x_1 \in \Omega$ such that

$$\alpha_1 v_1(x_0) + \alpha_2 v_2(x_0) = 0,$$
$$\alpha_1 v_1(x_1) + \alpha_2 v_2(x_1) \neq 0$$

for some $\alpha_1,\ \alpha_2 \in \mathbf{R}$.

Now take $w = |\alpha_1 v_1 + \alpha_2 v_2|$. Since $v_1,\ v_2 \in E_1$, we have $\alpha_1 v_1 + \alpha_2 v_2 \in E_1$ and $w = |\alpha_1 v_1 + \alpha_2 v_2|$ still belongs to E_1 (as shown just above, as a consequence of the Courant–Rayleigh variational formula for λ_1). Let us summarize the properties of w:

$$\begin{cases} w \in E_1, \\ w \geq 0, \\ w(x_0) = 0, \\ w(x_1) \neq 0. \end{cases}$$

Since $-\Delta w = \lambda_1 w$, from $\lambda_1 > 0$ and $w \geq 0$, we deduce that

$$\begin{cases} -\Delta w \geq 0, \\ w = 0 \quad \text{on } \partial\Omega, \\ w(x_0) = 0, \quad x_0 \in \Omega. \end{cases}$$

The strong maximum principle property of Hopf now implies that $w = 0$ on Ω, a clear contradiction to the fact that $w(x_1) \neq 0$.

Thus E_1 has dimension one. By taking any vector $w \in E_1 \setminus \{0\}$ and $e_1 = |w|$ we obtain a vector in E_1 which satisfies, by using again the strong maximum principle, $e_1 > 0$ on Ω. □

Corollary 8.5.1. *Let Ω be as in Theorem* 8.5.1 *and take any eigenvector e_i of the Laplace–Dirichlet operator corresponding to an eigenvalue $\lambda_i > \lambda_1$. Then the sign of e_i is not constant on Ω.*

Proof. Since $\lambda_i \neq \lambda_1$ we have $E(\lambda_i) \perp E(\lambda_1)$ and $\int_\Omega e_i(x) e_1(x)\, dx = 0$. Since $e_1 > 0$ on Ω, this forces e_i to change sign on Ω. □

This means that the status of the first eigenvalue is very particular. It is the only eigenvalue which possesses an eigenvector with constant sign. Indeed, in the analysis of the second eigenvalue problem $\lambda_2(-\Delta)$, the nodal set of a second eigenvector (the set where it is equal to zero) plays a central role.

The explicit computation of the spectrum of the Laplace–Dirichlet operator is possible only in very particular situations. Nevertheless, even in situations where such a computation is not possible (or is too complicated), one can get rather precise information on the spectrum by using comparison arguments. To stress the dependance of eigenvalues with respect to Ω, let us denote by $(\lambda_n(\Omega))_{n\in\mathbf{N}}$ the sequence of eigenvalues of the Laplace–Dirichlet operator on Ω (with the multiplicity convention). Then, as a direct consequence of the Courant–Fisher min-max principle, we have the following comparison result.

Proposition 8.5.1. *Let Ω and $\tilde{\Omega}$ be two open bounded subsets of $\mathbf{R}^N$ with $\Omega \subset \tilde{\Omega}$. Then, for any $n \geq 1$,*

$$\lambda_n(\tilde{\Omega}) \leq \lambda_n(\Omega),$$

i.e., $\lambda_n(\Omega)$ is a decreasing function of Ω.

PROOF. For any $v \in H_0^1(\Omega)$, let us denote by $\tilde{v}$ the function which is equal to v on Ω and zero on $\tilde{\Omega} \setminus \Omega$. By Proposition 5.1.1, we have $\tilde{v} \in H_0^1(\tilde{\Omega})$. Moreover,

$$\int_\Omega |v(x)|^2\,dx = \int_{\tilde{\Omega}} |\tilde{v}(x)|^2\,dx,$$
$$\int_\Omega |\nabla v(x)|^2\,dx = \int_{\tilde{\Omega}} |\nabla \tilde{v}(x)|^2\,dx.$$

Hence, $H_0^1(\Omega)$ can be isometrically identified with a subspace of $H_0^1(\tilde{\Omega})$ by the mapping

$$v \in H_0^1(\Omega) \xrightarrow{\tilde{i}} \tilde{v} = \tilde{i}(v) \in H_0^1(\tilde{\Omega}).$$

If $M \in \mathcal{L}_n(\Omega)$ is an n-dimensional subspace of $H_0^1(\Omega)$, then $\tilde{i}(M) \in \mathcal{L}_n(\tilde{\Omega})$ is an n-dimensional subspace of $H_0^1(\tilde{\Omega})$. We can now apply the Courant–Fisher min-max formula (Theorem 8.4.2) to obtain

$$\begin{aligned}
\lambda_n(\Omega) &= \min_{M \in \mathcal{L}_n(\Omega)} \max_{v \in M,\ v \neq 0} \mathcal{R}(v, \Omega) \\
&= \min_{M \in \mathcal{L}_n(\Omega)} \max_{v \in M,\ v \neq 0} \mathcal{R}(\tilde{v}, \tilde{\Omega}) \\
&= \min_{W = \tilde{i}(M),\ M \in \mathcal{L}_n(\Omega)} \max_{w \in W \setminus \{0\}} \mathcal{R}(w, \tilde{\Omega}) \\
&\geq \min_{M \in \mathcal{L}_n(\tilde{\Omega})} \max_{w \in W \setminus \{0\}} \mathcal{R}(w, \tilde{\Omega}) = \lambda_n(\tilde{\Omega}),
\end{aligned}$$

which ends the proof. □

To go further and use this comparison result we need to know some particular situations where the spectrum of $-\Delta$ can be explicitely computed. Let us start with the simplest situation, that is, $N = 1$ and $\Omega = (0, 1)$.

Proposition 8.5.2. *Let $N = 1$ and $\Omega = (0, 1)$. Then, the eigenvalues $(\lambda_n)_{n \in \mathbf{N}}$ of the Laplace–Dirichlet operator are given by*

$$\lambda_n = n^2\pi^2, \quad n = 1, 2 \ldots,$$

and the corresponding orthonormal basis $(e_n)_{n \in \mathbf{N}}$ of eigenvectors in $L^2(\Omega)$ is given by

$$e_n(x) = \sqrt{2}\sin(n\pi x).$$

PROOF. The proof is elementary. When solving the ordinary differential equation

$$u'' + \lambda u = 0,$$

one obtains

$$u(x) = A\sin(\sqrt{\lambda}x) + B\cos(\sqrt{\lambda}x).$$

The boundary condition $u(0) = 0$ gives $B = 0$ and the boundary condition $u(1) = 0$ gives $\sin\sqrt{\lambda} = 0$, that is, $\lambda = n^2\pi^2$, for some $n \geq 1$. The corresponding solution is $u(x) = A\sin(n\pi x)$. After L^2-normalization one obtains $A = \sqrt{2}$. □

In this very simple situation, each eigenvalue has multiplicity one. Let us now study the Laplace equation with the Dirichlet boundary condition on the N-cube $\Omega = (0,1)^N$ and the corresponding eigenvalue problem.

Proposition 8.5.3. *Let $\Omega = (0,1)^N$. For each $p = (p_1, p_2, \ldots, p_N)$ with $p_i \in \mathbf{N}\setminus\{0\}$, $i = 1, 2, \ldots, N$ (i.e., $p \in (\mathbf{N}^*)^N$), the positive real number*

$$\lambda_p := \pi^2(p_1^2 + p_2^2 + \cdots + p_N^2)$$

is an eigenvalue of the Laplace–Dirichlet operator on $\Omega = (0,1)^N$ and the function

$$u_p(x) = 2^{N/2}\prod_{i=1}^{N}\sin(\pi p_i x_i)$$

is an eigenfunction corresponding to the eigenvalue λ_p. Indeed,

$$\Lambda(-\Delta, \Omega) = \left\{\lambda_p \ : \ p \in (\mathbf{N}^*)^N\right\},$$

i.e., all the eigenvalues of $-\Delta$ on $\Omega = (0,1)^N$ can be expressed in this way, and the family $\left\{u_p \ : \ p \in (\mathbf{N}^*)^N\right\}$ is an orthonormal basis of $L^2(\Omega)$.

Proof. Take $p = (p_1, p_2, \ldots, p_N)$, $v \in \mathcal{D}(\Omega)$ and compute

$$\int_\Omega \nabla u_p(x)\cdot\nabla v(x)\,dx = \int_\Omega \sum_{i=1}^{N}\frac{\partial u_p}{\partial x_i}(x)\frac{\partial v}{\partial x_i}(x)\,dx.$$

Let us notice that

$$u_p(x) = \prod_{i=1}^{N} e_{p_i}(x_i) \quad \text{with } e_{p_i}(x_i) = \sqrt{2}\sin(\pi p_i x_i).$$

Consequently,

$$\frac{\partial}{\partial x_i}u_p(x) = \left[\prod_{j\neq i} e_{p_j}(x_j)\right] e'_{p_i}(x_i).$$

(e'_{p_i} stands for the derivative of the function of one variable $e_{p_i}(\cdot)$.) Hence

$$\begin{aligned}\int_\Omega \nabla u_p(x)\cdot\nabla v(x)\,dx \;&=\; \sum_{i=1}^{N}\left(\int_0^1 e'_{p_i}(x_i)\frac{\partial v}{\partial x_i}(x)\,dx_i\right)\\ &\quad\cdot\int_{(0,1)^{N-1}}\prod_{j\neq i} e_{p_j}(x_j)\,dx_1\ldots dx_{i-1}dx_{i+1}\ldots dx_N.\end{aligned}$$

An integration by parts yields

$$\int_0^1 e'_{p_i}(x_i)\frac{\partial v}{\partial x_i}(x)\,dx_i = -\int_0^1 e''_{p_i}(x_i)v(x)\,dx_i$$
$$= \pi^2 p_i^2 \int_0^1 e_{p_i}(x_i)v(x)\,dx_i.$$

(The last equality follows from the fact that $e''_{p_i} + \pi^2 p_i^2 e_{p_i} = 0$, since e_{p_i} is an eigenvector relative to the eigenvalue $\pi^2 p_i^2$ of the one-dimensional Laplace–Dirichlet problem.) Thus

$$\int_\Omega \nabla u_p(x)\cdot\nabla v(x)\,dx = \pi^2\left(\sum_{i=1}^N p_i^2\right)\int_\Omega u_p(x)v(x)\,dx.$$

By a classical density and extension by continuity argument, this equality can be extended to an arbitrary $v \in H_0^1(\Omega)$, and

$$\begin{cases} \displaystyle\int_\Omega \nabla u_p(x)\cdot\nabla v(x)\,dx = \lambda_p \int_\Omega u_p v\,dx \quad \forall v \in H_0^1(\Omega), \\ u_p \in H_0^1(\Omega), \end{cases}$$

which precisely means that $\lambda_p = \pi^2\sum_{i=1}^N p_i^2$ is an eigenvalue of the Laplace–Dirichlet operator on $(0,1)^N$, and $u_p(x) = \prod_{i=1}^N e_{p_i}(x_i)$ is a corresponding eigenvector.

Let us now show that the family of eigenfunctions $\{u_p : p \in (\mathbf{N}^*)^N\}$ is an orthonormal basis of $L^2(\Omega)$.

First let us notice that if $p \neq q$, then there exists at least one $i \in \{1, 2, \ldots, N\}$ such that $p_i \neq q_i$. From the orthogonality in $L^2(0,1)$ of the two functions $\sin(p_i\pi x)$ and $\cos(q_i\pi x)$, we immediately obtain that the family $\{u_p : p \in (\mathbf{N}^*)^N\}$ is orthogonal in $L^2(\Omega)$.

The point that is more delicate is to prove that the family $\{u_p : p \in (\mathbf{N}^*)^N\}$ generates $L^2(\Omega)$ in the topological sense, that is, the vector space generated by this family of vectors is dense in $L^2(\Omega)$. Indeed, by a careful application of the Fubini theorem, one can prove the following result (which is quite classical in integration theory and we omit its proof).

Lemma 8.5.1. *Let $(v_p)_{p\in\mathbf{N}^*}$ and $(w_q)_{q\in\mathbf{N}^*}$ be two Hilbertian bases of $L^2(0,1)$. Then the family of functions*

$$(x, y) \mapsto v_p(x)w_q(y)$$

is a Hilbertian basis of $L^2((0,1)^2)$.

Thus, by iterating this result a finite number of times we obtain that the family $\{u_p : p \in (\mathbf{N}^*)^N\}$ is an orthonormal basis of $L^2(\Omega)$. This clearly implies that by taking $\Lambda = \{\lambda_p : p \in (\mathbf{N}^*)^N\}$ we have obtained all the eigenvalues; otherwise, there would exist some $v \in H_0^1(\Omega)$, $v \neq 0$, which is an eigenvector corresponding to some eigenvalue $\lambda \notin \Lambda$. By the orthogonality property this would imply that v is orthogonal in $L^2(\Omega)$ to all the $\{u_p : p \in (\mathbf{N}^*)^N\}$ which forms a basis, and hence $v = 0$, a clear contradiction. This completes the proof of the spectral analysis of the Laplace–Dirichlet operator in the case $\Omega = (0,1)^N$. □

Proposition 8.5.3 and Proposition 8.5.1 (comparison principle) permit us to obtain a sharp estimation of the asymptotic behavior of the sequence $(\lambda_n(\Omega))_{n\in\mathbf{N}^*}$ of the eigenvalues of the Laplace–Dirichlet operator in a bounded open set Ω in $\mathbf{R}^N$. Indeed, one can prove the following result.

Theorem 8.5.2. *Let $(\lambda_n(\Omega))_{n\in\mathbf{N}}$ be the sequence of the eigenvalues of the Laplace–Dirichlet operator in a bounded open set Ω in $\mathbf{R}^N$ (with the multiplicity convention). Then, there exist two positive constants c_Ω and d_Ω, which depend only on Ω, such that for all $n \geq 1$,*

$$c_\Omega n^{2/N} \leq \lambda_n(\Omega) \leq d_\Omega n^{2/N}.$$

SKETCH OF THE PROOF. The proof is quite technical but the idea is very simple. The idea consists in the comparison of Ω with two N-cubes Q_a and Q_b such that $Q_a \subset \Omega \subset Q_b$, $Q_a = (-a/2, a/2)^N$, $Q_b = (-b/2, b/2)^N$. Then Proposition 8.5.1 applies and one obtains

$$\lambda_n(Q_a) \leq \lambda_n(\Omega) \leq \lambda_n(Q_b).$$

Then the problem has been reduced to the evaluation of $\lambda_n(Q_a)$ and $\lambda_n(Q_b)$. Clearly $\lambda_n(Q_a) = \lambda_n(Q)/a^2$ and $\lambda_n(Q_b) = \lambda_n(Q)/b^2$, where $Q = (0, 1)^N$. By Proposition 8.5.3, the numbers $\lambda_n(Q)/\pi^2$ are precisely the positive integers of the form $\sum_{i=1}^N p_i^2$ with $p_i \in \mathbf{N} \setminus \{0\}$. Thus, one has to arrange the numbers $\left\{ \sum_{i=1}^N p_i^2 \ : \ p_i \in \mathbf{N}^* \right\}$ as an increasing sequence to obtain the sequence $\left\{\lambda_n(Q)/\pi^2 \ : \ n \in \mathbf{N}^*\right\}$: this is just a combinatorial problem! To that end, it is convenient to introduce for any $t > 0$ the quantity $\nu_N(t)$ which is the cardinal of all the elements $p \in (\mathbf{N}^*)^N$ such that $p = (p_1, \ldots, p_N)$ with $\sum_{i=1}^N p_i^2 \leq t$. Then, the key of the proof consists in showing the following estimate: $\nu_N(t) \sim C_N t^{N/2}$ for some constant $C_N > 0$. □

Remark 8.5.1. From Proposition 8.5.3, we can obtain, as indicated before, after some combinatorial argument, a complete description of the sequence $(\lambda_n)_{n\in\mathbf{N}}$ of the eigenvalues of the Laplace–Dirichlet operator in $\Omega = (0, 1)^N$. For example,

(a) $N = 2$. Then

$$\begin{aligned}
&\lambda_1(\Omega) = 2\pi^2, \quad \text{multiplicity } = 1 \text{ (no surprise!)},\\
&\lambda_2(\Omega) = 5\pi^2, \quad \text{multiplicity } = 2,\\
&\lambda_3(\Omega) = 8\pi^2, \quad \text{multiplicity } = 1,\\
&\lambda_4(\Omega) = 10\pi^2, \quad \text{multiplicity } = 2,\\
&\ldots.
\end{aligned}$$

(b) $N = 3$. Then

$$\begin{aligned}
&\lambda_1(\Omega) = 3\pi^2, \quad \text{multiplicity } = 1 \text{ (no surprise!)},\\
&\lambda_2(\Omega) = 6\pi^2, \quad \text{multiplicity } = 3,\\
&\lambda_3(\Omega) = 9\pi^2, \quad \text{multiplicity } = 3,\\
&\ldots.
\end{aligned}$$

Indeed, in the various examples we have encountered, the multiplicity of the second eigenvalue of the Laplace–Dirichlet operator does not obey a simple rule: it may be equal to one, two, three,….

This makes a sharp contrast to the first eigenvalue $\lambda_1(-\Delta)$, which always has multiplicity one, and makes the second eigenvalue more delicate to work with.

Remark 8.5.2. The previous analysis of the spectrum of the Laplace–Dirichlet operator on a general bounded open set Ω relies on the comparison of Ω with a reference set $\tilde{\Omega}$ such that either $\Omega \subset \tilde{\Omega}$ or $\tilde{\Omega} \subset \Omega$. There is another way to obtain comparison results, which consists in the use of rearrangement results (Steiner symmetrization). The idea is that this type of transformation preserves the measure and the L^2 norm of the functions and makes smaller the $(L^2)^N$ norm of the gradient of a function (Dirichlet integral). In this way it is possible to compare the corresponding Rayleigh quotients. This device is very useful in shape optimization; it permits us, for example, to solve an old problem from Rayleigh which consists in proving that the ball minimizes the first eigenvalue among all open sets of given volume.

8.6 A general abstract theory for spectral analysis of elliptic boundary value problems

So far, we have considered the spectral analysis of the Laplacian with Dirichlet boundary conditions. To be able to develop a similar analysis for more general linear elliptic operators and for different types of boundary conditions (like Neumann, mixed,…), let us introduce the following abstract setting.

(i) Let V and H be two real Hilbert spaces (infinite dimensional spaces) such that $V \xrightarrow{i} H$; we assume that

- V can be embedded in H by i which is linear continuous and one to one,
- V is dense in H (i.e., $\overline{i(V)}^H = H$),
- V is compactly embedded in H (i.e., i is compact)

(as a typical example, take $V = H_0^1(\Omega)$ and $H = L^2(\Omega)$ with their usual Hilbertian structures).

(ii) Let

$$\begin{aligned} a : V \times V &\longrightarrow \mathbf{R}, \\ (u, v) &\mapsto a(u, v), \end{aligned}$$

be a bilinear form on $V \times V$ which is symmetric continuous and coercive:

$$\exists \alpha > 0 \text{ such that } \forall\, v \in V \quad a(v, v) \geq \alpha \|v\|^2.$$

Here $\|\cdot\|$ stands for the norm in V.

The norm and the scalar product in H are respectively denoted by $|\cdot|_H$ ($|\cdot|$ without ambiguity) and $\langle \cdot, \cdot \rangle_H$ ($\langle \cdot, \cdot \rangle$ without ambiguity). Note that we are in the situation of the Lax–Milgram theorem, Theorem 3.1.2, and for any $L \in V^*$ there exists a unique $u \in V$ which satisfies

$$a(u, v) = L(v) \quad \forall v \in V.$$

Noticing that for any $h \in H$ the linear form

$$v \in V \mapsto \langle h, v \rangle_H$$

is continuous on V, we deduce from the Lax–Milgram theorem the existence of a unique solution $u = Th$ of the following problem:

$$\begin{cases} a(Th, v) = \langle h, v\rangle_H, \\ Th \in V. \end{cases}$$

By using the same device as in Proposition 8.2.1, we can easily prove that $T : H \longrightarrow H$ is a linear continuous, self-adjoint, compact, and positive definite operator. Thus, one can apply to T the abstract diagonalization theorem, Theorem 8.3.1, for compact, self-adjoint, positive definite operators and conclude that there exists a Hilbertian basis $(e_n)_{n\in\mathbf{N}}$ in H of eigenvectors,

$$Te_n = \mu_n e_n,$$

with $(\mu_n)_{n\in\mathbf{N}}$, the decreasing sequence of positive eigenvalues (with multiplicity condition), which tends to zero as $n \to +\infty$. Note that now the family $(e_n)_{n\in\mathbf{N}}$ is a Hilbertian basis of V, when V is equipped with the scalar product

$$\langle\langle u, v\rangle\rangle := a(u, v)$$

(which is equivalent to the initial one).

We can now give a precise description of the solutions of the abstract spectral problem: find $\lambda \in \mathbf{R}$ such that there exists $u \in V$, $u \neq 0$ which satisfies

$$a(u, v) = \lambda\langle u, v\rangle_H \quad \forall v \in V.$$

(When such $u \neq 0$ exists, it is called an eigenvector relative to the eigenvalue λ.)

Theorem 8.6.1. *Assume that the canonical injection of V into H is dense and compact and that the continuous bilinear form $a : V \times V \longrightarrow \mathbf{R}$ is symmetric and coercive on $V \times V$ (V-elliptic). Then the eigenvalues λ of the abstract variational problem*

$$\begin{cases} \text{find } \lambda \in \mathbf{R} \text{ such that there exists } u \in V,\ u \neq 0, \\ a(u, v) = \lambda\langle u, v\rangle_H \ \forall\, v \in V, \end{cases}$$

can be written as an increasing sequence of positive numbers $(\lambda_n)_{n\in\mathbf{N}}$ which tends to $+\infty$ as $n \to +\infty$

$$0 < \lambda_1 \leq \lambda_2 \leq \lambda_3 \leq \cdots \leq \lambda_n \leq \cdots.$$

(We again adopt the multiplicity convention: each eigenvalue is repeated a number of times equal to its multiplicity, which is finite.)

There exists an orthonormal basis (Hilbertian basis) $(e_n)_{n\in\mathbf{N}}$ of H such that for each $n \in \mathbf{N}$, e_n is an eigenvector relative to the eigenvalue λ_n:

$$\begin{cases} a(e_n, v) = \lambda_n\langle e_n, v\rangle_H \quad \forall v \in V, \\ e_n \in V. \end{cases}$$

Moreover, the sequence $\left(e_n/\sqrt{\lambda_n}\right)_{n\in\mathbf{N}}$ is a Hilbertian basis of V when this space is equipped with the (equivalent) scalar product $a(\cdot, \cdot)$.

We consider now some applications of the results above.

(1) NEUMANN PROBLEM. Take in this case

$$V = \left\{ v \in H^1(\Omega) \ : \ \int_\Omega v(x)\,dx = 0 \right\} \quad \text{and} \quad a(u,v) = \int_\Omega \nabla u(x) \cdot \nabla v(x)\,dx.$$

The coercivity of a on $V \times V$ follows from the Poincaré–Wirtinger inequality (Corollary 5.4.1) and the compact embedding $V \longrightarrow H = L^2(\Omega)$ from the Rellich–Kondrakov theorem, Theorem 5.4.2 (Ω is assumed smooth and bounded). From Theorem 8.6.1, we deduce the existence of a Hilbertian basis $(e_n)_{n\in\mathbf{N}}$ in $L^2(\Omega)$ such that

$$\begin{cases} -\Delta e_n = \lambda_n e_n \quad \text{on } \Omega, \\ \dfrac{\partial e_n}{\partial \nu} = 0 \quad \text{on } \partial\Omega, \end{cases}$$

where $\frac{\partial}{\partial \nu}$ stands for the normal outwards derivative on the boundary $\partial\Omega$.

(2) MIXED DIRICHLET–NEUMANN PROBLEM (see Section 6.3). By taking $V = \left\{ v \in H^1(\Omega) \ : \ \gamma_0(v) = 0 \text{ on } \Gamma_0 \right\}$ with $\mathcal{H}^{N-1}(\Gamma_0) > 0$, one obtains the existence of a Hilbertian basis in $L^2(\Omega)$, $(e_n)_{n\in\mathbf{N}}$ such that

$$\begin{cases} -\Delta e_n = \lambda_n e_n \quad \text{on } \Omega, \\ e_n = 0 \quad \text{on } \Gamma_0, \\ \dfrac{\partial e_n}{\partial \nu} = 0 \quad \text{on } \Gamma_1 = \Gamma \setminus \Gamma_0. \end{cases}$$

(3) One can obtain similar results by replacing $-\Delta$ by an elliptic linear operator A of the form

$$Av = -\sum_{i,j} \frac{\partial}{\partial x_i} \left(a_{i,j} \frac{\partial v}{\partial x_j} \right) + a_0 u$$

or more generally when considering elliptic systems (elasticity, Stokes,…).

Remark 8.6.1. The variational approach of Courant–Fisher works without any particular difficulty in such a general setting. One introduces the abstract Rayleigh quotient $\mathcal{R}(v) = a(v,v)/|v|_H^2$ and thus

$$\begin{aligned} \lambda_1 &= \min \left\{ \mathcal{R}(v) \ : \ v \in V, \ v \neq 0 \right\}, \\ \lambda_n &= \min_{M \in \mathcal{L}_n} \max_{v \in M, \ v \neq 0} \mathcal{R}(v) \\ &= \max_{M \in \mathcal{L}_{n-1}} \min_{v \in M^\perp, \ v \neq 0} \mathcal{R}(v). \end{aligned}$$

Let us end this chapter and return to the situation which was our first motivation for this study, the method of separation of variables of Fourier, applied to the wave equation

$$\begin{cases} \dfrac{\partial^2 u}{\partial t^2} - \Delta u = 0 \quad \text{on } Q = \Omega \times (0, +\infty), \\ u = 0 \quad \text{on } \Sigma = \partial\Omega \times (0, +\infty), \\ u(x,0) = u_0(x) \quad \text{on } \Omega, \\ \dfrac{\partial u}{\partial t}(x,0) = u_1(x) \quad \text{on } \Omega. \end{cases}$$

Denote by $0 < \lambda_1 < \lambda_2 \leq \lambda_3 \leq \cdots \leq \lambda_n \leq \cdots$ the eigenvalues of the Laplace–Dirichlet operator and by $(e_n)_{n\in\mathbf{N}}$ a corresponding Hilbertian basis in $L^2(\Omega)$ of eigenvectors. Set $\omega_n = \sqrt{\lambda_n}$. Then the unique (variational) solution of the above problem is given by the following formula (for $u_0 \in H_0^1(\Omega)$ and $u_1 \in L^2(\Omega)$ given):

$$u(t) = \sum_{n=1}^{+\infty} \left[\langle u_0, e_n \rangle_{L^2(\Omega)} \cos(\omega_n t) + \frac{1}{\omega_n} \langle u_1, e_n \rangle_{L^2(\Omega)} \sin(\omega_n t) \right] e_n.$$

Here the variational solution is taken in the following sense: for any $0 < T < +\infty$

$$u \in \mathbf{C}\big((0,T); H_0^1(\Omega)\big) \cap \mathbf{C}^1\big((0,T); L^2(\Omega)\big)$$

and

$$\begin{cases} \dfrac{d^2}{dt^2}\langle u(t), v\rangle_{L^2(\Omega)} + \displaystyle\int_\Omega \nabla u(t) \cdot \nabla v \, dx = 0 & \text{in the distributional sense on } (0,T) \\ & \forall v \in H_0^1(\Omega), \\ u(0) = u_0, \quad \dfrac{du}{dt}(0) = u_1 & \end{cases}$$

(see, for example, Raviart–Thomas [199, Section 8.2] for further details).

Chapter 9

Convex duality and optimization

In this chapter, unless specified, $(V, \|\cdot\|_V)$ is a general normed linear space with topological dual V^*. For any $v \in V$ and $v^* \in V^*$, we write $v^*(v) = \langle v^*, v\rangle_{(V^*,V)}$. Recall that V^* is a Banach space when equipped with the dual norm

$$\|v^*\|_{V^*} = \sup\{\langle v^*, v\rangle \ : \ \|v\|_V \leq 1\}.$$

Without ambiguity, for simplicity of notation, we write $\|\cdot\|$ instead of $\|\cdot\|_V$, $\|\cdot\|_*$ instead of $\|\cdot\|_{V^*}$ and $\langle v^*, v\rangle$ instead of $\langle v^*, v\rangle_{(V^*,V)}$.

9.1 Dual representation of convex sets

We know that several basic geometrical objects in a normed linear space V can be described by using continuous linear forms, i.e., elements of the topological dual space. For example, a closed hyperplane H can be written

$$H = \{v \in V \ : \ \langle v^*, v\rangle = \alpha\}$$

for some $v^* \in V^*$, $v^* \neq 0$, and $\alpha \in \mathbf{R}$. Similarly, a closed half-space $\mathcal{H}$ can be written

$$\mathcal{H} = \{v \in V \ : \ \langle v^*, v\rangle \leq \alpha\}.$$

Intersections of finite collections of closed half-spaces yield convex polyhedra. Indeed, we are going to show that arbitrary closed convex sets in V can be described by using only linear continuous forms. This is what we call a dual representation. This theory is based on the Hahn–Banach theorem (which we stated in Theorem 3.3.1), which is formulated below.

Theorem 9.1.1. *Let C be a nonempty closed convex subset of a normed linear space V. Then, each point $u \notin C$ can be strongly separated from C by a closed hyperplane, which means*

$$\exists u^* \in V^*, \ u^* \neq 0, \ \exists \alpha \in \mathbf{R} \text{ such that } \langle u^*, u\rangle > \alpha \text{ and } \langle u^*, v\rangle \leq \alpha \ \forall \ v \in C.$$

From a geometrical point of view, this means that C is contained in the closed half-space

$$\mathcal{H}_{\{u^* \leq \alpha\}} := \Big\{ v \in V \ : \ \langle u^*, v \rangle \leq \alpha \Big\},$$

whereas u is in the complement: $\mathcal{H}_{\{u^* > \alpha\}} := \{v \in V \ : \ \langle u^*, v \rangle > \alpha\}$.

PROOF. Let us give the proof of Theorem 9.1.1 when V is a Hilbert space. In that case, one can give a constructive proof relying on the projection theorem on a closed convex subset. (In the general case of a normed linear space V, one can use the analytic version of the Hahn–Banach theorem, which itself is a consequence of the Zorn lemma.)

Let us denote by $P_C(u)$ the projection of u on C. It is characterized by the angle condition (optimality condition)

$$\begin{cases} \langle u - P_C(u), v - P_C(u) \rangle \leq 0 & \forall\, v \in C, \\ P_C(u) \in C. \end{cases}$$

Set $z := u - P_C(u)$. Since $u \notin C$, we have $z \neq 0$ and we can rewrite the above inequality in the following form:

$$\sup_{v \in C} \langle z, v \rangle \leq \langle z, P_C(u) \rangle. \tag{9.1}$$

On the other hand, by definition of z and since $z \neq 0$

$$\begin{aligned} 0 < |z|^2 &= \langle z, z \rangle \\ &= \langle z, u \rangle - \langle z, P_C(u) \rangle, \end{aligned}$$

which implies

$$\langle z, P_C(u) \rangle < \langle z, u \rangle. \tag{9.2}$$

Take $\alpha := \langle z, P_C(u) \rangle$. Combining (9.1) and (9.2) we obtain

$$\sup_{v \in C} \langle z, v \rangle \leq \alpha < \langle z, u \rangle,$$

i.e., $C \subset \mathcal{H}_{\{\langle z, . \rangle \leq \alpha\}}$ and $u \in \mathcal{H}_{\{\langle z, . \rangle > \alpha\}}$. □

As a direct consequence of Theorem 9.1.1 we obtain the following corollary.

Corollary 9.1.1. *Let C be a nonempty closed convex subset of a normed linear space V. Then C is equal to the intersection of all closed half-spaces that contain it:*

$$C = \bigcap_{C \subset \mathcal{H}_{\{v^* \leq \alpha\}}} \mathcal{H}_{\{v^* \leq \alpha\}}.$$

PROOF. Let us denote by $\mathcal{F}$ the set

$$\mathcal{F} = \Big\{ (v^*, \alpha) \in V^* \times \mathbf{R} \ : \ C \subset \mathcal{H}_{\{v^* \leq \alpha\}} \Big\}.$$

Clearly $C \subset \bigcap_{(v^*,\alpha)\in\mathcal{F}} \mathcal{H}_{\{v^*\le\alpha\}}$. Let us prove the converse inclusion $\bigcap_{(v^*,\alpha)\in\mathcal{F}} \mathcal{H}_{\{v^*\le\alpha\}} \subset C$. By taking the complement, this is equivalent to proving

$$V \setminus C \subset \bigcup_{(v^*,\alpha)\in\mathcal{F}} \Big(V \setminus \mathcal{H}_{\{v^*\le\alpha\}}\Big),$$

which is precisely the conclusion of the Hahn–Banach separation theorem, Theorem 9.1.1. □

Among closed convex sets, an important subclass is obtained by taking the intersection of a finite number of closed half-spaces.

Definition 9.1.1. *A closed convex polyhedron P is an intersection of finitely many closed half-spaces: in other words, there exist $v_1^*, \dots, v_k^* \in V^*$ with $v_i^* \neq 0$ and $\alpha_1, \dots, \alpha_k \in \mathbf{R}$ such that*

$$P = \big\{v \in V \ : \ \langle v_i^*, v\rangle \le \alpha_i \text{ for } i = 1, \dots, k\big\}.$$

In the representation of closed convex sets as the intersection of closed half-spaces, it is natural to look for the simplest representation. To that end, let us observe the following elementary facts:

(a) $\alpha' \ge \alpha$ and $C \subset \mathcal{H}_{\{v^*\le\alpha\}} \Longrightarrow C \subset \mathcal{H}_{\{v^*\le\alpha'\}}$;

(b) fixing $v^* \neq 0$ and making α vary provides parallel hyperplanes.

From Corollary 9.1.1 and the above observations, we deduce

$$C = \bigcap_{v^*\in V^*,\ v^*\neq 0} \ \bigcap_{\{\exists\alpha\in\mathbf{R}\,:\,C\subset\mathcal{H}_{\{v^*\le\alpha\}}\}} \mathcal{H}_{\{v^*\le\alpha\}}. \tag{9.3}$$

The question we have to examine is to describe, for a given $v^* \in V^*,\ v^* \neq 0$, such that there exists some $\alpha \in \mathbf{R}$ with $C \subset \mathcal{H}_{\{v^*\le\alpha\}}$, what is the intersection of all the parallel half-spaces $\mathcal{H}_{\{v^*\le\alpha\}}$ which contain C. The answer to this question gives rise to the notion of support function.

Proposition 9.1.1. *For any $v^* \in V^*,\ v^* \neq 0$ such that $C \subset \mathcal{H}_{\{v^*\le\alpha\}}$ for some $\alpha \in \mathbf{R}$ we have*

$$\bigcap_{\{\alpha\,:\,C\subset\mathcal{H}_{\{v^*\le\alpha\}}\}} \mathcal{H}_{\{v^*\le\alpha\}} = \mathcal{H}_{\{v^*\le\sigma_C(v^*)\}},$$

where $\sigma_C(v^*) := \sup\big\{\langle v^*, v\rangle \ : \ v \in C\big\}$. In other words, for any given $v^* \in V^*,\ v^* \neq 0$, such that $C \subset \mathcal{H}_{\{v^*\le\alpha\}}$ for some $\alpha \in \mathbf{R}$, the intersection of all the "parallel" closed half-spaces $\mathcal{H}_{\{v^*\le\alpha\}}$ containing C is the closed half-space $\mathcal{H}_{\{v^*\le\sigma_C(v^*)\}}$, where $\sigma_C(v^*)$ is defined as above.

It is convenient to extend the definition of σ_C to an arbitrary $v^* \in V^*$ by allowing it to take the value $+\infty$.

Definition 9.1.2. *For any subset C of V, the function $\sigma_C : V^* \to \mathbf{R} \cup \{+\infty\}$ defined by*

$$\sigma_C(v^*) = \sup\big\{\langle v^*, v\rangle \ : \ v \in C\big\}$$

is called the support function of the set C.

PROOF OF PROPOSITION 9.1.1 (a) For any $v \in C$, by definition of σ_C, we have

$$\langle v^*, v\rangle \le \sigma_C(v^*).$$

Hence, $C \subset \mathcal{H}_{\{v^* \le \sigma_C(v^*)\}}$, which clearly implies

$$\bigcap_{\{\alpha\,:\,C\subset\mathcal{H}_{\{v^*\le\alpha\}}\}} \mathcal{H}_{\{v^*\le\alpha\}} \subset \mathcal{H}_{\{v^*\le\sigma_C(v^*)\}}.$$

(b) For any $\alpha \in \mathbf{R}$ such that $C \subset \mathcal{H}_{\{v^*\le\alpha\}}$, we have

$$\alpha \ge \sup\big\{\langle v^*, v\rangle \;:\; v \in C\big\} = \sigma_C(v^*).$$

Hence, $\mathcal{H}_{\{v^*\le\sigma_C(v^*)\}} \subset \mathcal{H}_{\{v^*\le\alpha\}}$ and

$$\mathcal{H}_{\{v^*\le\sigma_C(v^*)\}} \subset \bigcap_{\{\alpha\,:\,C\subset\mathcal{H}_{\{v^*\le\alpha\}}\}} \mathcal{H}_{\{v^*\le\alpha\}},$$

which completes the proof. □

As a direct consequence of formula (9.3) and Proposition 9.1.1 we obtain the following important result.

Theorem 9.1.2. *Let C be a nonempty closed convex subset of a normed linear space V. Then*

$$C = \bigcap_{v^*\in V^*,\ v^*\ne 0} \mathcal{H}_{\{v^*\le\sigma_C(v^*)\}},$$

where σ_C is the support function of C. Equivalently,

$$C = \big\{v \in V \;:\; \langle v^*, v\rangle \le \sigma_C(v^*)\ \forall\, v^* \in V^*\big\}.$$

Remark 9.1.1. The dual representation of a closed convex set C has been obtained with the help of the support function $\sigma_C : V^* \to \mathbf{R} \cup \{+\infty\}$. As we will see in this chapter, the mapping $C \mapsto \sigma_C$ can be viewed as a particular case of the general duality correspondence, namely, the Legendre–Fenchel transform $f \mapsto f^*$. More precisely, by taking $f = \delta_C$ the indicator of C, we have $f^* = \sigma_C$. We examine below the properties of σ_C which are direct consequences of its definition.

Proposition 9.1.2. *The support function $\sigma_C : V^* \to \mathbf{R}\cup\{+\infty\}$ of a closed convex nonempty subset C is a function which is closed, convex, proper, and positively homogeneous of degree 1.*

PROOF. For any $v \in C$, the mapping

$$v^* \in V^* \mapsto \langle v^*, v\rangle$$

is a linear continuous form on V^*, hence convex and continuous. The function σ_C as a supremum of convex functions is still convex and, as a supremum of continuous functions,

it is closed (lower semicontinuous); see Proposition 3.2.3. Moreover, $\sigma_C(0) = 0$ and σ_C is proper. Finally, for any $v^* \in V^*$ and $t > 0$ we have

$$\begin{aligned}\sigma_C(tv^*) &= \sup\left\{t\langle v^*, v\rangle \ : \ v \in C\right\}\\ &= t\sup\left\{\langle v^*, v\rangle \ : \ v \in C\right\}\\ &= t\sigma_C(v^*),\end{aligned}$$

which expresses that σ_c is positively homogeneous of degree 1. □

To have a sharper view of the dual generation of closed convex sets, it is interesting to introduce the notion of supporting hyperplane. This notion is closely related to the question: in the definition of $\sigma_C(v^*) = \sup\left\{\langle v^*, v\rangle \ : \ v \in C\right\}$, is the supremum attained?

Definition 9.1.3. *An element $v^* \in V^*$, $v^* \neq 0$, is said to support C at a point $u \in C$ if*

$$\begin{aligned}\sigma_C(v^*) &= \langle v^*, u\rangle\\ &= \sup\left\{\langle v^*, v\rangle \ : \ v \in C\right\}.\end{aligned}$$

An equivalent terminology consists in saying that v^* is a supporting functional of C at $u \in C$.

The geometric terminology above comes from the fact that when v^* supports C at $u \in C$ we have that the closed half-space

$$\mathcal{H}_{\{v^* \leq \sigma_C(v^*)\}}$$

contains C and that the corresponding hyperplane

$$H = \left\{v \in V \ : \ \langle v^*, v\rangle = \sigma_C(v^*)\right\}$$

intersects C at u. (Note that the intersection of H with C may contain some other points.)

An interesting question is to know whether it is possible to obtain a dual representation of closed convex sets by supporting functionals. As we will see, this is a quite involved question which is intimately connected with the properties of the subdifferential of a closed convex function and the Bishop–Phelps theorem (density properties of the domain of the subdifferential).

Let us end this section with some elementary examples illustrating the concept of support function.

Example 9.1.1. (1) Take $C = \mathbf{B}(0, 1)$ the unit ball of V. Then, for any $v^* \in V^*$

$$\begin{aligned}\sigma_C(v^*) &= \sup\left\{\langle v^*, v\rangle \ : \ \|v\|_V \leq 1\right\}\\ &= \|v^*\|_{V^*},\end{aligned}$$

i.e., σ_C is the dual norm $\|\cdot\|_{V^*}$ of $\|\cdot\|_V$.

(2) Take C as a cone, i.e., $\lambda v \in C$ for all $v \in C$ and $\lambda \geq 0$ (note that necessarily $0 \in C$). Let us assume moreover that C is closed and convex. Then

$$\begin{aligned}\sigma_C(v^*) &= \sup\left\{\langle v^*, v\rangle \ : \ v \in C\right\} \\ &= \begin{cases} 0 \text{ whenever } \langle v^*, v\rangle \leq 0 \ \forall v \in C, \\ +\infty \text{ otherwise.}\end{cases}\end{aligned}$$

Let us notice that the set

$$C^* = \left\{v^* \in V^* \ : \ \langle v^*, v\rangle \leq 0 \text{ for all } v \in C\right\}$$

is a closed convex cone; it is called the polar cone of C. We have that σ_C is equal to the indicator function of this polar cone

$$\sigma_C = \delta_{C^*}.$$

9.2 Passing from sets to functions: Elements of epigraphical calculus

Our next goal is to apply the dual representation Theorem 9.1.2 to the set $C = \operatorname{epi} f$, where $f : V \to \mathbf{R} \cup \{+\infty\}$ is a closed convex proper function. So doing, we will obtain by a pure geometrical approach the Legendre–Fenchel duality theory for closed convex functions.

To that end, it will be useful to develop some tools of epigraphical calculus, which consists of viewing functions as sets, via their epigraphs. As stressed in Section 3.2.2, the epigraph of an extended real valued function is a geometrical object that carries most of the properties of the corresponding variational problems. In our context, given $f : V \to \mathbf{R} \cup \{+\infty\}$, recall that

$$f \text{ is closed (lsc)} \iff \operatorname{epi} f \text{ is closed},$$

$$f \text{ is convex} \iff \operatorname{epi} f \text{ is convex},$$

and that the basic operation in convex analysis and duality which consists in taking the supremum of a family of convex (affine) functions has an immediate epigraphical interpretation

$$\operatorname{epi}\left(\sup_{k\in I} f_k\right) = \bigcap_{k\in I} \operatorname{epi} f_k.$$

Beyond the classical operations on extended real valued functions (sum and multiplication by a positive scalar) let us introduce the epi-addition, also called inf-convolution.

Definition 9.2.1. *Let V be a linear space and $f, g : V \to \mathbf{R} \cup \{+\infty\}$ two extended real valued functions. The epi-sum of f and g (also called inf-convolution) is the function*

$$f \#_e g : V \to \overline{\mathbf{R}}$$

defined by

$$\begin{aligned}(f\#_e g)(v) &= \inf\big\{f(v_1)+g(v_2)\ :\ v_1+v_2=v,\ v_1,v_2\in V\big\}\\ &= \inf\big\{f(v-w)+g(w)\ :\ w\in V\big\}\\ &= \inf\big\{f(w)+g(v-w)\ :\ w\in V\big\}.\end{aligned}$$

We often briefly write $f\#g$.

Note that $f\#_e g$ may take the value $-\infty$ (for example, take $g=0$ and f not minorized). The term *epi-sum* comes from the following geometrical interpretation of this operation.

Proposition 9.2.1. *For any* $f, g : V \to \mathbf{R}\cup\{+\infty\}$

$$epi_S(f\#_e g) = epi_S f + epi_S g,$$

where $epi_S f$ *stands for the strict epigraph of* f*, i.e.,*

$$epi_S f = \big\{(v,\lambda)\in V\times\mathbf{R}\ :\ \lambda > f(v)\big\},$$

and the sum $epi_S f + epi_S g$ *is the vectorial sum (also called Minkowski sum) of the two sets* $epi_S f$ *and* $epi_S g$.

Proof. We have

$$\lambda > (f\#_e g)(v)$$

iff there exists $v_1, v_2 \in V$, with $v = v_1 + v_2$ such that

$$\lambda > f(v_1) + g(v_2).$$

This is clearly equivalent to the existence of $v_1, v_2 \in V$ and $\lambda_1, \lambda_2 \in \mathbf{R}$ such that $\lambda_1 > f(v_1)$, $\lambda_2 > g(v_2)$ and $v = v_1+v_2$, $\lambda = \lambda_1+\lambda_2$. Equivalently, $(\lambda, v) = (\lambda_1, v_1)+(\lambda_2, v_2)$ with $(\lambda_1, v_1) \in \mathrm{epi}_S f$ and $(\lambda_2, v_2) \in \mathrm{epi}_S g$. □

Remark 9.2.1. The term *inf-convolution* refers to the (formal) similarities of this operation with the usual convolution of functions on $\mathbf{R}^N$

$$(f * g)(x) = \int_{\mathbf{R}^N} f(x-y)g(y)\,dy,$$

where one has to replace $\int_{\mathbf{R}^N}$ by inf and product by addition. As we will see, there are many striking similarities between these two operations.

Proposition 9.2.2. *Let* $f, g : V \to \mathbf{R}\cup\{+\infty\}$ *be two convex functions. Then, their epi-sum* $f\#_e g$ *is still a convex function.*

Proof. This property is a clear consequence of the geometrical interpretation of the epi-sum via epigraphs (Proposition 9.2.1)

$$\mathrm{epi}_S(f\#_e g) = \mathrm{epi}_S f + \mathrm{epi}_S g$$

and of the fact that the Minkowski (vectorial) sum of two convex sets is still convex: indeed, given C and D two convex subsets of a vector space E, consider two points of $C + D$, let $v_1 = c_1 + d_1$, $v_2 = c_2 + d_2$ with $c_i \in C$ and $d_i \in D$ $(i = 1, 2)$. For any $0 \leq \lambda \leq 1$, one has

$$\begin{aligned}\lambda v_1 + (1-\lambda)v_2 &= \lambda(c_1 + d_1) + (1-\lambda)(c_2 + d_2)\\ &= \big(\lambda c_1 + (1-\lambda)c_2\big) + \big(\lambda d_1 + (1-\lambda)d_2\big),\end{aligned}$$

which still belongs to $C + D$. One can then easily verify that the convexity of an extended real-valued function is equivalent to the convexity of its strict epigraph. □

The fact that the epi-sum preserves the convexity, as we have just observed, follows clearly from its geometrical interpretation. On the other hand, it is somewhat surprising from the analytical point of view, since $(f\#_e g)(v)$ is expressed as an infimum of convex functions, namely $v \mapsto f(v-w) + g(w)$ and the class of convex functions is not stable by infimal operations. This calls for some explanation.

Indeed, the convexity of $f\#_e g$ when f and g are convex functions is a consequence of the observation, "the function of two variables $(v, w) \mapsto h(v, w) := f(v - w) + g(w)$ is convex with respect to the pair (v, w)," and of the following proposition.

Proposition 9.2.3. *Let V and W be two linear spaces and $h : V \times W \to \mathbf{R} \cup \{+\infty\}$ be a convex function. Then, the function $p : V \to \overline{\mathbf{R}}$ defined by*

$$p(v) = \inf_{w \in W} h(v, w)$$

is still convex.

Proof. Let us prove that for any $u, v \in V$ and $\lambda \in]0, 1[$

$$p(\lambda u + (1-\lambda)v) \leq \lambda p(u) + (1-\lambda)p(v).$$

Without any restriction, we can assume $p(u) < +\infty$ and $p(v) < +\infty$; otherwise the inequality is trivially satisfied. Take arbitrary $s > p(u)$ and $t > p(v)$. By definition of p, one can find elements $w_{u,s}$ and $w_{v,t}$ in W such that

$$s > h(u, w_{u,s}) \text{ and } t > h(v, w_{v,t}).$$

By convexity of $h(\cdot, \cdot)$ with respect to the couple of variables (v, w)

$$\begin{aligned}h(\lambda u + (1-\lambda)v, \lambda w_{u,s} + (1-\lambda)w_{v,t}) &\leq \lambda h(u, w_{u,s}) + (1-\lambda)h(v, w_{v,t})\\ &\leq \lambda s + (1-\lambda)t.\end{aligned}$$

By definition of p

$$p(\lambda u + (1-\lambda)v) \leq h(\lambda u + (1-\lambda)v, \lambda w_{u,s} + (1-\lambda)w_{v,t}).$$

We combine the two above inequalities to obtain

$$p(\lambda u + (1-\lambda)v) \leq \lambda s + (1-\lambda)t.$$

This being true for any $s > p(u)$ and $t > p(v)$, by letting s tend to $p(u)$ and t tend to $p(v)$, we obtain the required convexity inequality. □

We will see that the epi-sum (inf-convolution) is the dual operation of the usual sum. The epi-sum plays also an important role in the regularization of lower semicontinuous extended real-valued functions. The following theorem has a long history (it goes back to Hausdorff, Pasch, Baire, and has been revisited by many authors).

Theorem 9.2.1 (Lipschitz regularization via epi-sum). *Let $(V, \|\cdot\|)$ be a normed space and let $f : V \to \mathbf{R} \cup \{+\infty\}$ be a proper and lower semicontinuous function. Suppose moreover that f is conically minorized, i.e., there exists some $k_0 \geq 0$ such that for all $v \in V$*

$$f(v) \geq -k_0(1 + \|v\|).$$

Let us define for all $k \in \mathbf{R}^+$ the function $f_k := f \#_e k\|\cdot\|$, i.e.,

$$f_k(v) = \inf_{w \in V} \big\{ f(w) + k\|v - w\| \big\}.$$

Then we have

(a) *for all $k \geq k_0$, f_k is Lipschitz continuous on V with constant k, that is, for all $u, v \in V$*

$$|f_k(u) - f_k(v)| \leq k\|u - v\|;$$

(b) *for all $v \in V$ one has*

$$f(v) = \lim_{k \to +\infty} f_k(v).$$

More precisely, the sequence $(f_k)_k$ monotonically increases to f as $k \uparrow \infty$.

(c) *When f is convex, so is f_k for all $k \geq k_0$.*

Proof. (a) For all $k \geq k_0$, we have

$$\begin{aligned} f_k(v) &\geq \inf_{w \in V} \big\{ -k_0 - k_0\|w\| + k\|v - w\| \big\} \\ &\geq \inf_{w \in V} \big\{ -k_0 - k_0\|w\| + k\|w\| - k\|v\| \big\} \\ &\geq -k_0 - k\|v\| > -\infty. \end{aligned}$$

On the other hand, taking some $w_0 \in \operatorname{dom} f \neq \emptyset$ (f is proper)

$$f_k(v) \leq f(w_0) + k\|v - w_0\| < +\infty.$$

Hence, for all $k \geq k_0$ and all $v \in V$, $f_k(v)$ is a real number. Take now $u, v \in V$. The triangle inequality yields for any $w \in V$

$$\|v - w\| \leq \|u - w\| + \|v - u\|.$$

Hence, for all $w \in V$, for all $k \in \mathbf{R}^+$

$$f(w) + k\|v - w\| \leq f(w) + k\|u - w\| + k\|v - u\|.$$

Taking the infimum with respect to $w \in V$ yields

$$f_k(v) \leq f_k(u) + k\|v - u\|.$$

Exchanging the role of v and u and noticing that for $k \geq k_0$ both $f_k(v)$ and $f_k(u)$ are finitely valued yields

$$|f_k(v) - f_k(u)| \leq k\|v - u\|.$$

(b) By taking $w = v$ in the definition of $f_k(v)$, one has

$$f_k(v) \leq f(v).$$

Clearly, the sequence $(f_k)_k$ is increasing with respect to k. Hence

$$\lim_{k \to +\infty} f_k(v) \leq f(v).$$

Let us prove the reverse inequality

$$f(v) \leq \lim_{k \to +\infty} f_k(v).$$

If $\lim_{k\to+\infty} f_k(v) = +\infty$, there is nothing to prove. So let us assume that $\lim_{k\to+\infty} f_k(v) < +\infty$. For each $k \geq k_0$, let us introduce some $w_k \in V$ such that

$$f_k(v) \geq f(w_k) + k\|v - w_k\| - \varepsilon_k$$

for some $\varepsilon_k > 0$, with $\varepsilon_k \to 0$ as $k \to +\infty$. Using the growth condition on f we obtain

$$+\infty > \sup_{k>0} f_k(v) \geq -k_0(1 + \|w_k\|) + k\|v - w_k\| - \varepsilon_k,$$

which clearly implies $w_k \to v$ in $(V, \|\cdot\|)$ as $k \to +\infty$. Let us now pass to the limit on the inequality

$$f_k(v) \geq f(w_k) - \varepsilon_k$$

and use the lower semicontinuity of f to obtain

$$\begin{aligned} \lim_{k \to +\infty} f_k(v) &\geq \liminf_k f(w_k) \\ &\geq f(v). \end{aligned}$$

(c) Noticing that f and $k\|\cdot\|$ are both convex functions, the convexity of $f_k = f \#_e k\|\cdot\|$ is a straightforward consequence of Proposition 9.2.2. □

Let us end this section with the following striking property of closed convex functions. We know that a linear operator from a normed space into another normed space is continuous iff it is bounded on bounded sets. This is an important property since it reduces the study of continuity of a linear operator $A : E \to F$ to the establishment of majorizations of the following type: there exists some $M \in \mathbf{R}^+$ such that

$$\|v\|_E \leq 1 \Longrightarrow \|Av\|_F \leq M.$$

We will prove in Theorem 9.3.1 that any closed convex function is the supremum of all its continuous affine minorants.

Thus, it is not surprising that also for closed convex functions local boundedness implies continuity. Let us make this precise in the following statement.

Theorem 9.2.2. *Let $(V, \|\cdot\|)$ be a normed space and let $f : V \to \mathbf{R} \cup \{+\infty\}$ be a convex function which is majorized on a neighbourhood of a point $v_0 \in \operatorname{dom} f$, i.e.,*

$$\exists r > 0 \text{ such that } \sup_{\|v - v_0\| < r} f(v) := M < +\infty.$$

Then f is continuous at the point v_0. More precisely, f is Lipschitz continuous on all balls $\mathbf{B}(v_0, r')$ with $r' < r$ and

$$|f(w) - f(v)| \leq \frac{2(M + |f(v_0)|)}{r - r'} \|w - v\| \quad \forall v, w \in \mathbf{B}(v_0, r').$$

Proof. (a) Let us first prove that f is continuous at v_0. By translation (consider the function $f(v + v_0) - f(v_0)$), one can reduce the problem to the case $v_0 = 0$ and $f(0) = 0$. Take an arbitrary ε such that $1 \geq \varepsilon > 0$ and observe that for any $v \in \mathbf{B}(0, r\varepsilon)$, the following convex inequalities hold:

writing $v = (1 - \varepsilon)0 + \varepsilon\left(\frac{1}{\varepsilon}v\right)$ we have

$$f(v) \leq (1 - \varepsilon) f(0) + \varepsilon f\left(\frac{1}{\varepsilon}v\right) \leq \varepsilon M \ ;$$

writing $0 = \frac{1}{1+\varepsilon}v + \frac{\varepsilon}{1+\varepsilon}\left(\frac{-1}{\varepsilon}v\right)$ we have

$$0 = f(0) \leq \frac{1}{1+\varepsilon} f(v) + \frac{\varepsilon}{1+\varepsilon} f\left(\frac{-1}{\varepsilon}v\right) \leq \frac{1}{1+\varepsilon} f(v) + \frac{\varepsilon M}{1+\varepsilon},$$

which yields

$$f(v) \geq -\varepsilon M.$$

Combining the two above inequalities, we obtain

$$|f(v)| \leq \varepsilon M \text{ for } v \in \mathbf{B}(0, r\varepsilon),$$

which yields the continuity of f at the origin.

(b) First observe that in the above argument, when taking $\varepsilon = 1$ we have the existence of some positive constant, we still denote by M, such that

$$|f(v + v_0) - f(v_0)| \leq M \qquad \forall\, v \in \mathbf{B}(0, r).$$

Take arbitrary $v, w \in \mathbf{B}(v_0, r')$ with $v \neq w$. Set $\varepsilon = r - r' > 0$ and

$$u = w + \frac{\varepsilon}{\|w - v\|}(w - v)\ , \qquad \lambda = \frac{\|w - v\|}{\varepsilon + \|w - v\|}\ .$$

We have $u \in \mathbf{B}(v_0, r)$ and $\|w - v\|u = (\varepsilon + \|w - v\|)w - \varepsilon v$. Equivalently,

$$w = \lambda u + \frac{\varepsilon}{\varepsilon + \|w - v\|} v,$$

$$w = \lambda u + (1 - \lambda)v \text{ with } \lambda \in]0, 1[.$$

By convexity of f

$$f(w) \leq \lambda f(u) + (1 - \lambda) f(v) = f(v) + \lambda(f(u) - f(v)),$$

which yields (observe that $|f(u) - f(v)| \leq |f(u) - f(v_0)| + |f(v) - f(v_0)| \leq 2M$)

$$f(w) - f(v) \leq \frac{\|w - v\|}{\varepsilon + \|w - v\|} 2M \leq \frac{2M}{\varepsilon} \|w - v\|.$$

Exchanging the role of w and v, we obtain

$$|f(w) - f(v)| \leq \frac{2M}{\varepsilon} \|w - v\|,$$

which completes the proof. □

9.3 Legendre–Fenchel transform

Given $f : V \to \mathbf{R} \cup \{+\infty\}$ a closed convex proper function, we are going to introduce f^*, the Legendre–Fenchel transform of f, by considering the set $C = \operatorname{epi} f \subset V \times \mathbf{R}$ and its dual representation, as given by Theorem 9.1.2. To that end, we need to exploit the particular structure of the set $C = \operatorname{epi} f$ in $V \times \mathbf{R}$ and describe the family of the closed half-spaces in $V \times \mathbf{R}$ containing it.

Let us start with the following elementary result, which describes the closed half-spaces in $V \times \mathbf{R}$.

Lemma 9.3.1. *Let $l \in (V \times \mathbf{R})^*$ be a linear continuous form on $V \times \mathbf{R}$, $l \neq 0$. Then, there exist $u^* \in V^*$ and $\gamma \in \mathbf{R}$, $(u^*, \gamma) \neq 0$ such that*

$$l(v, t) = \langle u^*, v \rangle + \gamma t \quad \forall\, (v, t) \in V \times \mathbf{R}.$$

A closed half-space $\mathcal{H}$ in $V \times \mathbf{R}$ is of the following form:

$$\mathcal{H} = \mathcal{H}_{\{(u^*, \gamma) \leq \alpha\}} := \big\{(v, t) \in V \times \mathbf{R} \,:\, \langle u^*, v \rangle + \gamma t \leq \alpha\big\}.$$

Depending on the value of γ ($\gamma = 0$ or $\gamma \neq 0$), we have the two distinct situations:

(a) $\gamma = 0$. *Then $\mathcal{H} = \mathcal{H}_{\{(u^*, \gamma) \leq \alpha\}} = \big\{(v, t) \in V \times \mathbf{R} \,:\, \langle u^*, v \rangle \leq \alpha\big\} = \big\{u^* \leq \alpha\big\} \times \mathbf{R}$ is invariant by all the translations parallel to $\big\{0\big\} \times \mathbf{R}$. In that case, we say that the half-space is "vertical".*

(b) $\gamma \neq 0$. *By normalization (divide by $-\gamma$) one can rewrite the closed half-space $\mathcal{H}$ in the form*

$$\mathcal{H} = \big\{(v, t) \in V \times \mathbf{R} \,:\, t \geq \langle u^*, v \rangle - \alpha\big\}$$

or

$$\mathcal{H} = \big\{(v,t) \in V \times \mathbf{R} \ : \ t \leq \langle u^*, v\rangle - \alpha\big\}.$$

It is the epigraph or the hypograph of the affine continuous function

$$v \mapsto \langle u^*, v\rangle - \alpha.$$

Because of its particular structure, the epigraph of a proper function cannot be contained in a half-space of the form $\big\{(v,t) \in V \times \mathbf{R} \ : \ t \leq \langle u^*, v\rangle - \alpha\big\}$. We can summarize the previous results in the following lemma.

Lemma 9.3.2. *Let $f : V \longrightarrow \mathbf{R} \cup \{+\infty\}$ be a proper function. Then, a closed half-space $\mathcal{H}$ in $V \times \mathbf{R}$ containing the set $C = \operatorname{epi} f$ is either vertical or equal to the epigraph of an affine continuous function, i.e., there exist some $u^* \in V$ and $\alpha \in \mathbf{R}$ such that*

$$\mathcal{H} = \big\{(v,t) \in V \times \mathbf{R} \ : \ t \geq \langle u^*, v\rangle - \alpha\big\}.$$

Let us keep in mind that we are looking for the simplest dual representation of convex functions f. In this perspective, it is a striking and important property that one can get rid of the vertical half-spaces in their dual representations. Indeed, this is not a surprising result since one can approach them, and get arbitrarily "close" to closed vertical half-spaces, by epigraphs of continuous affine functions.

Let us make this precise in the following statement.

Theorem 9.3.1. *Let $(V, \|\cdot\|)$ be a normed space and let $f : V \to \mathbf{R} \cup \{+\infty\}$ be a closed convex proper function. Then, f is equal to the supremum of all the continuous affine functions which minorize f.*

Proof. (a) Let us first notice that among all the closed half-spaces containing $C = \operatorname{epi} f$, at least one of them is the epigraph of an affine continuous function. Otherwise, there would be only vertical half-spaces in this family, and C, as an intersection of such sets, would be vertically invariant. This is impossible, because f is proper. Indeed, for any $v_0 \in \operatorname{dom} f$, $C \cap (\{v_0\} \times \mathbf{R}) = \{v_0\} \times [f(v_0), +\infty[$ with $[f(v_0), +\infty[$ strictly included in $\mathbf{R}$. Then, notice that

$$\operatorname{epi} f \subset \operatorname{epi}\big\{v \mapsto \langle u^*, v\rangle - \alpha\big\}$$

is equivalent to

$$f(v) \geq \langle u^*, v\rangle - \alpha \ \forall v \in V,$$

i.e., f admits at least an affine continuous minorant.

(b) Let us recall the general property: $f = \sup f_i \iff \operatorname{epi} f = \bigcap_{i\in I} \operatorname{epi} f_i$. Thus, to establish the assertion of the theorem, it suffices to show that each point $(v_0, t_0) \notin \operatorname{epi} f$ is outside the epigraph of an affine continuous function that is majorized by f.

We know that $C = \operatorname{epi} f$ is equal to the intersection of all the closed half-spaces that contain it (Corollary 9.1.1) and that any such half-space either is vertical or is the epigraph of an affine continuous function (Lemma 9.3.2). To eliminate the vertical half-spaces in their

dual representation, we use the Lipschitz regularization theorem, Theorem 9.2.1. Since f admits an affine continuous minorant, it is conically minorized and $f = \sup_k f_k$ with f_k convex and Lipschitz continuous.

Since $t_0 < f(v_0)$, for some k_0 sufficiently large $t_0 < f_{k_0}(v_0) \leq f(v_0)$ and $(v_0, t_0) \notin \operatorname{epi} f_{k_0}$. Let us now use the dual representation of the closed convex set $\operatorname{epi} f_{k_0}$ as the intersection of all the closed half-spaces that contain it. Since f_{k_0} is everywhere defined, there is no vertical half-space containing $\operatorname{epi} f_{k_0}$. Thus, f_{k_0} is the supremum of all its affine continuous minorants. As a consequence, one can find an affine continuous function $v \mapsto l(v) = \langle u^*, v\rangle - \alpha$ with

$$(v_0, t_0) \notin \operatorname{epi} l \text{ and } l \leq f_{k_0}.$$

Since $f_{k_0} \leq f$ we have $l \leq f$ and $(v_0, t_0) \notin \operatorname{epi} l$, that is, l satisfies all required properties. □

Just like for convex sets, we are going to look for the simplest dual description of closed convex functions, i.e., using the simplest continuous affine minorants. Theorem 9.3.1 tells us that

$$f(v) = \sup\big\{\langle v^*, v\rangle - \alpha \ : \ \langle v^*, v\rangle - \alpha \leq f(v)\ \forall v \in V\big\}.$$

Let us now observe that for $v^* \in V^*$ being fixed, making α vary provides parallel minorizing affine continuous functions. Clearly, the best α is obtained by taking

$$\alpha = \sup\big\{\langle v^*, v\rangle - f(v) \ : \ v \in V\big\},$$

which, for $v^* \in V^*$ being fixed, is a real number iff f admits a continuous affine minorant with slope v^*. This is precisely the quantity which is classically denoted by

$$f^*(v^*) = \sup\big\{\langle v^*, v\rangle - f(v) \ : \ v \in V\big\}$$

and which makes sense for an arbitrary $v^* \in V^*$, with possibly $+\infty$ values.

The above geometrical considerations allow us to reformulate Theorem 9.3.1 in the following form:

$$\forall v \in V \quad f(v) = \sup\big\{\langle v^*, v\rangle - f^*(v^*) \ : \ v^* \in V^*\big\}. \tag{9.4}$$

We are now ready to introduce classical notation, terminology, and basic facts concerning the Legendre–Fenchel transform which is defined below for arbitrary proper function f.

Definition 9.3.1. *Let V be a normed linear space and let $f : V \to \mathbf{R} \cup \{+\infty\}$ be a proper function. The Legendre–Fenchel conjugate of f (L-F in short) is the function*

$$f^* : V^* \to \mathbf{R} \cup \{+\infty\}$$

defined by

$$f^*(v^*) = \sup\big\{\langle v^*, v\rangle - f(v) \ : \ v \in V\big\}.$$

Let us notice that since f is proper, by taking some $v_0 \in \operatorname{dom} f$

$$f^*(v^*) \geq \langle v^*, v_0 \rangle - f(v_0),$$

i.e., f^* admits an affine continuous minorant and $f^* : V^* \to \mathbf{R} \cup \{+\infty\}$.

Moreover, $v^* \in \operatorname{dom} f^*$ iff there exists $\alpha \in \mathbf{R}$ such that for all $v \in V$ one has $\langle v^*, v \rangle - f(v) \leq \alpha$, i.e.,

$$f(v) \geq \langle v^*, v \rangle - \alpha.$$

Let us now return to the case when f is closed convex and proper. We know that f admits at least one such affine continuous minorant. This implies that f^* is proper. Since f^* is a supremum of continuous affine functions it is a closed convex proper function from V^* into $\mathbf{R} \cup \{+\infty\}$.

Let us examine the two formulas

$$f^*(v^*) = \sup \left\{ \langle v^*, v \rangle - f(v) \ : \ v \in V \right\} \text{ (definition of } f^*),$$

$$f(v) = \sup \left\{ \langle v^*, v \rangle - f^*(v^*) \ : \ v^* \in V^* \right\} \text{ (Theorem 9.3.1).}$$

They are essentially the same. Let us make this precise.

Since f^* is closed convex and proper we can compute its conjugate $f^{**} : V^{**} \to \mathbf{R} \cup \{+\infty\}$. By using the canonical embedding of V into V^{**}, we can restrict f^{**} to V to obtain

$$\forall v \in V \quad f^{**}(v) = \sup \left\{ \langle v^*, v \rangle - f^*(v^*) \ : \ v^* \in V^* \right\}$$

(recall that $i : V \to V^{**}$ is defined by $i(v)(v^*) = v^*(v)$).

The dual representation theorem, Theorem 9.3.1, for closed convex functions then can be reformulated in the following form: $f = f^{**}$. This is the Fenchel–Moreau–Rockafellar theorem that we now state.

Theorem 9.3.2. *Let V be a normed space and let $f : V \to \mathbf{R} \cup \{+\infty\}$ be a closed convex proper function. Then*

$$f = f^{**},$$

i.e., f is equal to its biconjugate. Equivalently,

$$\forall v \in V \quad f(v) = \sup \left\{ \langle v^*, v \rangle - f^*(v^*) \ : \ v^* \in V^* \right\}.$$

It is worth noticing that the dual representation of closed convex proper functions has been given, in the above theorem, a quite simple formulation. On the counterpart, it is a precise analytic formulation which may hide the geometrical features of this duality theory. It is good to keep in mind both aspects.

At this point, it is interesting to observe that the duality for functions has been derived from duality for sets (via the representation of $f : V \to \mathbf{R} \cup \{+\infty\}$ by its epigraph $C = \operatorname{epi} f$). Conversely, the duality for sets can be obtained as a particular case of the duality for functions. Let us associate to a set C its indicator function δ_C and first observe

that whenever C is a nonempty closed convex set, then δ_C is a closed convex proper function. Then notice that

$$\begin{aligned}(\delta_C)^*(v^*) &= \sup\big\{\langle v^*, v\rangle - \delta_C(v) \ : \ v \in V\big\}\\ &= \sup\big\{\langle v^*, v\rangle \ : \ v \in C\big\}\\ &= \sigma_C(v^*),\end{aligned}$$

i.e., $(\delta_C)^*$ is the support function of C.

The Fenchel–Moreau–Rockafellar theorem, Theorem 9.3.2, says that $(\delta_C)^{**} = \delta_C$, which is equivalent to

$$\delta_C(v) = \sup\big\{\langle v^*, v\rangle - \sigma_C(v^*) \ : \ v^* \in V^*\big\}.$$

Noticing that $v \in C$ iff $\delta_C(v) = 0$ one gets

$$C = \big\{v \in V \ : \ \langle v^*, v\rangle \leq \sigma_C(v^*) \ \forall\, v^* \in V^*\big\}.$$

Let us summarize the previous results.

Proposition 9.3.1. *Let C be a nonempty closed convex subset of a normed linear space V; then*

$$(\delta_C)^* = \sigma_C \text{ and } (\sigma_C)^* = \delta_C,$$

that is,

$$C = \big\{v \in V \ : \ \langle v^*, v\rangle \leq \sigma_C(v^*) \text{ for all } v^* \in V^*\big\}.$$

It is worth noticing that the biconjugate operation $f \mapsto f^{**}$ enjoys nice properties for convex functions which are not necessarily closed.

We recall (see Section 3.2.4) that given (X, τ) a general topological space and $f : X \to \mathbf{R} \cup \{+\infty\}$, $cl_\tau f$ is the largest τ-lower semicontinuous function that minorizes f. We have

$$\operatorname{epi}(cl_\tau f) = cl(\operatorname{epi} f);$$

$cl_\tau f$ is called the lower semicontinuous regularization of f. Moreover (see Proposition 3.2.5 (d)), f is τ-lsc at x iff $f(x) = (cl_\tau f)(x)$.

For convex functions $f : V :\to \mathbf{R} \cup \{+\infty\}$, with V a normed space, we have the following elegant characterization of clf (for the topology of the norm of the space V) in terms of the biconjugate f^{**}.

Proposition 9.3.2. *Let V be a normed linear space and let $f : V \to \mathbf{R} \cup \{+\infty\}$ be a convex proper function. Let us assume that f admits a continuous affine minorant. Then the following equality holds:*

$$f^{**} = clf.$$

As a consequence,

$$f \text{ is lower semicontinuous at } u \in V \iff f(u) = f^{**}(u).$$

PROOF. By definition, for any $v \in V$

$$f^{**}(v) = \sup\left\{\langle v^*, v\rangle - f^*(v^*) \ : \ v^* \in V^*\right\},$$

which implies that f^{**} is the upper envelope of the continuous affine minorants of f. It is a closed (convex) proper function, hence

$$f^{**} \leq clf \leq f.$$

Let us now observe that clf is still convex, because the epigraph of clf is the closure of epi f which is a convex set. Hence, clf is a closed convex proper function.

Since f^{**} and clf are both closed convex proper functions, we apply Theorem 9.3.2 to obtain

$$(f^{**})^{**} = f^{**},$$

$$(clf)^{**} = clf.$$

The inequality $f^{**} \leq clf \leq f$ implies, by taking the biconjugate of each term, that

$$(f^{**})^{**} \leq (clf)^{**} \leq f^{**},$$

i.e.,

$$f^{**} \leq clf \leq f^{**},$$

and the equality $f^{**} = clf$ follows.

By Proposition 3.2.5 (d), for general functions $f : V \to \mathbf{R} \cup \{+\infty\}$ we have the equivalence

$$f \text{ is lower semicontinuous at } u \in V \iff f(u) = clf(u).$$

As a consequence, when f is proper and convex, we have

$$f \text{ is lower semicontinuous at } u \in V \iff f(u) = f^{**}(u),$$

which completes the proof. □

Example 9.3.1. (1) Take $C = \mathbf{B}(0, 1)$. By definition of the dual norm, for any $v^* \in V^*$

$$\sigma_{\mathbf{B}(0,1)}(v^*) = \sup_{v \in \mathbf{B}(0,1)} \langle v^*, v\rangle = \|v^*\|_{V^*}.$$

Thus, $\sigma_{\mathbf{B}(0,1)} = \|\cdot\|_{V^*}$. Conversely the convex duality theorem yields

$$(\|\cdot\|_*)^* = \delta_{\mathbf{B}(0,1)}. \tag{9.5}$$

(2) As suggested by the result above we have

$$(\|\cdot\|)^* = \delta_{\mathbf{B}^*(0,1)}. \tag{9.6}$$

Let us prove (9.6). Indeed by contrast with (9.5), (9.6) is an elementary result which does not use the Hahn–Banach theorem. Set $f(v) = \|v\|_V$. Thus, for any $v^* \in V^*$

$$f^*(v^*) = \sup\left\{\langle v^*, v\rangle - \|v\| \ : \ v \in V\right\}.$$

If $\|v^*\|_* \leq 1$, then $\langle v^*, v\rangle - \|v\| \leq 0$ for all $v \in V$ and $\langle v^*, v\rangle - \|v\| = 0$ for $v = 0$. Thus $f^*(v^*) = 0$.

If $\|v^*\|_* > 1$, by definition of $\|\cdot\|_*$ there exists some $v_0 \in V$ such that $\langle v^*, v_0\rangle > \|v_0\|$. It follows that for all $t > 0$

$$\langle v^*, tv_0\rangle - \|tv_0\| = t(\langle v^*, v_0\rangle - \|v_0\|).$$

Hence $\lim_{t\to+\infty} \langle v^*, tv_0\rangle - \|tv_0\| = +\infty$, which implies

$$f^*(v^*) = \sup_{v\in V} \big\{\langle v^*, v\rangle - f(v)\big\} = +\infty.$$

By taking the conjugate in (9.6) and applying the duality Theorem 9.3.2 we obtain for any $v \in V$

$$\|v\| = \sup\big\{\langle v^*, v\rangle : \|v^*\|_* \leq 1\big\};$$

this is the isometrical embedding theorem from V into its bidual V^{**}.

We give in the following proposition an important example of a dual convex function which indeed is an extension of Example 9.3.1, case 2.

Proposition 9.3.3. *Let $(V, \|\cdot\|)$ be a normed space with topological dual space $(V^*, \|\cdot\|_*)$. Let $\varphi : \mathbf{R} \to \mathbf{R} \cup \{+\infty\}$ be a closed convex function which is even (i.e., $\varphi(-t) = \varphi(t)$). Then the function*

$$f : V \to \mathbf{R} \cup \{+\infty\}, \quad f(v) = \varphi(\|v\|),$$

is a closed convex proper function and

$$f^*(v^*) = \varphi^*(\|v^*\|_*).$$

Proof. The assumptions on φ imply that $\varphi : \mathbf{R}^+ \to \mathbf{R} \cup \{+\infty\}$ is increasing. Thus f is still convex and clearly closed. Moreover,

$$\begin{aligned} f^*(v^*) &= \sup_{v\in V} \big\{\langle v^*, v\rangle - \varphi(\|v\|)\big\} \\ &= \sup_{t\geq 0} \sup_{v\in V, \|v\|=t} \big\{\langle v^*, v\rangle - \varphi(\|v\|)\big\} \\ &= \sup_{t\geq 0} \big\{t\|v^*\|_* - \varphi(t)\big\} \\ &= \sup_{t\in\mathbf{R}} \big\{t\|v^*\|_* - \varphi(t)\big\} \text{ (because } \varphi \text{ is even)} \\ &= \varphi^*(\|v^*\|_*), \end{aligned}$$

which completes the proof. □

As a straightforward consequence we obtain the following useful result.

Corollary 9.3.1. *Set $f(v) = \frac{1}{p}\|v\|^p$ with $1 < p < +\infty$. Then $f^*(v^*) = \frac{1}{p'}\|v^*\|_*^{p'}$ where p' is the Hölder conjugate exponent of p, i.e., $1/p + 1/p' = 1$. In particular, taking*

$V = L^p(\Omega, \mathcal{A}, \mu)$, $1 < p < +\infty$, we have $V^* = L^{p'}(\Omega, \mathcal{A}, \mu)$ and the conjugate function of

$$f(v) = \frac{1}{p} \int_\Omega \|v(x)\|^p \, d\mu(x)$$

is equal to

$$f^*(v^*) = \frac{1}{p'} \int_\Omega \|v^*(x)\|^{p'} \, d\mu(x).$$

This makes the transition with the next important example, which is concerned with integral functionals and which will be studied in detail in chapter 13.

Theorem 9.3.3. *Let $V = L^p(\Omega, \mathcal{A}, \mu)$, $1 < p < +\infty$, and*

$$f(v) = \int_\Omega j(x, v(x)) \, d\mu(x)$$

a convex integral functional associated to a convex normal integrand j. Then

$$f^* : V^* = L^{p'}(\Omega, \mathcal{A}, \mu) \to \mathbf{R} \cup \{+\infty\}$$

is given by

$$f^*(v^*) = \int_\Omega j^*(x, v^*(x)) \, d\mu(x),$$

where $j^*(x, \cdot)$ is the convex conjugate of $j(x, \cdot)$.

Remark 9.3.1. When $V = H$ is a Hilbert space the L-F transform $f \mapsto f^*$ is an involution from $\Gamma_0(H)$ into itself, where $\Gamma_0(H)$ is the set of closed convex proper functions on H,

$$\begin{gathered} \Gamma_0(H) \xrightarrow{*} \Gamma_0(H), \\ f \mapsto f^*, \end{gathered}$$

i.e., $f^{**} = f$. This transform has some analogy with the Fourier–Plancherel transform (F-P in short),

$$\mathcal{F} : L^2(\mathbf{R}^N) \to L^2(\mathbf{R}^N), \ f \mapsto \mathcal{F}(f),$$

where $\overline{\mathcal{F}}\mathcal{F} f = f$, which is indeed an isometry. Let us notice that $(\|\cdot\|^2/2)^* = \|\cdot\|^2/2$ i.e., $\|\cdot\|^2/2$ is invariant for the L-F transform while $f(x) = \frac{1}{2} e^{-\|x\|^2}$ is invariant for the F-P transform. A basic property of the F-P transform is

$$\mathcal{F}(f * g) = \mathcal{F}(f)\mathcal{F}(g),$$

where $f * g$ is the convolution of functions. This property can be seen as the (formal) analogue of the property

$$(f \#_e g)^* = f^* + g^*,$$

which is studied in Section 9.4.

Let us complete this section by studying the natural setting in which the Legendre–Fenchel transform acts as an operator. We will pay particular attention to the description

of its range. Our final result (Theorem 9.3.5) shows that the Legendre–Fenchel transform is a one-to-one mapping from $\Gamma_0(V)$ onto $\Gamma_0(V^*)$ (cf. Definition 9.3.2). Given a general normed space $(V, \|.\|)$, let us recall that the Legendre–Fenchel transform is the mapping which associates to a closed convex proper function $f : V \longrightarrow \mathbf{R} \cup \{+\infty\}$ its conjugate $f^* : V^* \longrightarrow \mathbf{R} \cup \{+\infty\}$ which is defined by

$$\forall v^* \in V^* \qquad f^*(v^*) = \sup\big\{\langle v^*, v\rangle - f(v) : v \in V\big\}.$$

Let us start with some simple observations.

(a) The Legendre–Fenchel transform is one-to-one. This results from the implication

$$f^* = g^* \Longrightarrow f^{**} = g^{**}$$

and $f^{**} = f,\ g^{**} = g$, which hold true for arbitrary closed convex proper functions f and g (Theorem 9.3.2).

(b) The description of the range of the L-F transform requires some attention. Let us return to the (above) definition of f^*. In this supremum, one just needs to consider $v \in \operatorname{dom} f$. For such $v \in \operatorname{dom} f$, the mapping $v^* \mapsto \langle v^*, v\rangle - f(v)$ is continuous on V^* for the topology $\sigma(V^*, V)$, which is the weak star topology of the dual. This follows directly from the definition of this topology. Hence, f^*, as a supremum of such affine, $\sigma(V^*, V)$ continuous functions, is a convex, proper function which is $\sigma(V^*, V)$ lower semicontinuous. Indeed, when V is not reflexive, for a convex function $g : V^* \longrightarrow \mathbf{R}\cup\{+\infty\}$, to be $\sigma(V^*, V)$ lower semicontinuous is a strictly stronger property than just to be lower semicontinuous for the topology of the norm. One can exhibit a closed convex function $g : V^* \longrightarrow \mathbf{R}$ which is not $\sigma(V^*, V)$ lower semicontinuous. Take any $\xi \in V^{**} \setminus I(V)$, where I is the canonical embedding of V into its bidual V^{**} (recall that $\langle I(v), v^*\rangle_{(V^{**},V^*)} = \langle v^*, v\rangle_{(V^*,V)}$). Then define $g(v^*) := \langle \xi, v^*\rangle_{(V^{**},V^*)}$. Clearly g is continuous on V^*, but it is not $\sigma(V^*, V)$ lower semicontinuous. Otherwise, by linearity, it would be continuous for the topology $\sigma(V^*, V)$, which would imply $\xi \in I(V)$.

It turns out that this $\sigma(V^*, V)$ lower semicontinuity property allows us to characterize the range of the L-F transform. Let us make this precise in the following statement.

Theorem 9.3.4. *Let $(V, \|.\|)$ be a normed space.*

(a) *For any $f : V \longrightarrow \mathbf{R} \cup \{+\infty\}$ which is closed, convex, and proper, its Legendre–Fenchel conjugate $f^* : V^* \longrightarrow \mathbf{R} \cup \{+\infty\}$ is a convex proper function which is $\sigma(V^*, V)$ lower semicontinuous.*

(b) *Conversely, let $g : V^* \longrightarrow \mathbf{R} \cup \{+\infty\}$ be a convex, proper, and $\sigma(V^*, V)$ lower semicontinuous function. Then,*

$$g = g^{**}$$

and g belongs to the range of the Legendre–Fenchel transform. More precisely, $g = (g^*)^*$ is equal to the Legendre–Fenchel transform of the closed convex proper function $g^* : V \longrightarrow \mathbf{R} \cup \{+\infty\}$ which is defined by

$$\forall v \in V \qquad g^*(v) = \sup\big\{\langle v^*, v\rangle - g(v^*) : v^* \in V^*\big\}.$$

Proof. Part (a) has already been proved. Proof of part (b) requires some further topological tools. When equipped with the topology $\sigma(V^*, V)$, the space V^* is a locally convex

topological vector space, whose dual can be identified with V. The Hahn–Banach theorem still holds in locally convex topological vector spaces, and from that point, the proof is essentially the same as in Theorem 9.3.2. □

To give a unified formulation of Theorem 9.3.2 and Theorem 9.3.4 where V and V^*, f and f^* play symmetrical roles, it is convenient to introduce the following notions and notation.

Definition 9.3.2. *Let $(V, \|.\|)$ be a normed space with topological dual V^*. We set*

$$\Gamma_{0,V^*}(V) = \big\{ f : V \longrightarrow \mathbf{R} \cup \{+\infty\},\ f \text{ is a pointwise supremum of a nonvoid family of affine functions with slopes in } V^*,\ f \not\equiv +\infty \big\};$$

$$\Gamma_{0,V}(V^*) = \big\{ g : V^* \longrightarrow \mathbf{R} \cup \{+\infty\},\ g \text{ is a pointwise supremum of a nonvoid family of affine functions with slopes in } V,\ g \not\equiv +\infty \big\}.$$

These definitions make explicit references to the pairing between the two spaces V and V^*, that is, $(v, v^*) \in V \times V^* \mapsto \langle v^*, v\rangle_{(V^*,V)} = v^*(v)$. Without ambiguity, one often omits the subscript refering to the coupled space and writes briefly $\Gamma_0(V)$ and $\Gamma_0(V^*)$. To be more precise, one has

$$f \in \Gamma_0(V) \iff f = \sup_{i \in I} f_i$$

with $f_i(v) = \langle v_i^*, v\rangle - \alpha_i$ for some index set I, $v_i^* \in V^*$ (slope), and $\alpha_i \in \mathbf{R}$;

$$g \in \Gamma_0(V^*) \iff g = \sup_{j \in J} g_j$$

with $g_j(v^*) = \langle v^*, v_j\rangle - \beta_j$ for some index set J, $v_j \in V$ (slope), and $\beta_j \in \mathbf{R}$.

We can now reformulate Theorem 9.3.1 and its corresponding version when considering the locally convex topological vector space $(V^*, \sigma(V^*, V))$, together with Theorems 9.3.2 and 9.3.4 in the following final statement.

Theorem 9.3.5. *Let $(V, \|.\|)$ be a normed space with topological dual V^*. Then,*

(a) *one has*

$$\begin{aligned}\Gamma_0(V) &= \big\{ f : V \longrightarrow \mathbf{R} \cup \{+\infty\},\ f \text{ closed, convex, proper} \big\} \\ &= \big\{ f : V \longrightarrow \mathbf{R} \cup \{+\infty\},\ f\ \sigma(V, V^*) \text{ closed, convex, proper} \big\},\end{aligned}$$

while

$$\Gamma_0(V^*) = \big\{ g : V^* \longrightarrow \mathbf{R} \cup \{+\infty\},\ g\ \sigma(V^*, V) \text{ closed, convex, proper} \big\}.$$

(b) *The Legendre–Fenchel transform is a one-to-one mapping from $\Gamma_0(V)$ onto $\Gamma_0(V^*)$:*

$$\begin{aligned}\Gamma_0(V) &\xrightarrow{*} \Gamma_0(V^*) \\ f &\longrightarrow f^*.\end{aligned}$$

For any $f \in \Gamma_0(V)$ one has $f = f^{**}$ and for any $g \in \Gamma_0(V^*)$ one has $g = g^{**}$.

Remark 9.3.2. The preceding theory can be developed in the general setting of two vector spaces V and W in separate duality.

Let us denote by $\langle v, w\rangle_{(V,W)}$ a given pairing between elements $v \in V$ and $w \in W$. It is a bilinear form with separating properties, namely,

$$\begin{cases} \forall v \in V,\ v \neq 0,\ \exists w \in W \text{ with } \langle v, w\rangle \neq 0, \\ \forall w \in W,\ w \neq 0,\ \exists v \in V \text{ with } \langle v, w\rangle \neq 0. \end{cases}$$

Then W is the dual of $(V, \sigma(V, W))$ and conversely. The set $\Gamma_0(V)$ (respectively, $\Gamma_0(W)$) is defined by taking suprema of affine functions with slopes in W (respectively, V), and the Legendre–Fenchel transform is a one-to-one mapping from $\Gamma_0(V)$ onto $\Gamma_0(W)$; see Moreau [182] for further details.

9.4 Legendre–Fenchel calculus

As we have already stressed, most optimization problems can be written as

$$\inf\big\{f(v)\ :\ v \in V\big\},$$

where $f = f_0 + \delta_C$ is the sum of the objective function f_0 and the indicator function of the constraint C. This explains the importance of getting a formula for the Legendre–Fenchel conjugate of a sum of functions. At this point, the epi-sum plays a central role, because of the following general property.

Proposition 9.4.1. *Let $\varphi, \psi : V \to \mathbf{R} \cup \{+\infty\}$ be two proper functions. Then*

$$(\varphi \# \psi)^* = \varphi^* + \psi^*.$$

Proof. It is enough to take $v^* \in V^*$ and compute

$$\begin{aligned} (\varphi\#\psi)^*(v^*) &= \sup_{v\in V}\big\{\langle v^*, v\rangle - (\varphi\#\psi)(v)\big\} \\ &= \sup_{v\in V}\Big\{\langle v^*, v\rangle - \inf_{v_1+v_2=v}\big(\varphi(v_1)+\psi(v_2)\big)\Big\} \\ &= \sup_{v\in V}\Big\{\langle v^*, v\rangle + \sup_{v_1+v_2=v}\big(-\varphi(v_1)-\psi(v_2)\big)\Big\} \\ &= \sup_{v\in V,\ v_1+v_2=v}\big\{\big(\langle v^*, v_1\rangle - \varphi(v_1)\big) + \big(\langle v^*, v_2\rangle - \psi(v_2)\big)\big\} \\ &= \varphi^*(v^*) + \psi^*(v^*), \end{aligned}$$

which completes the proof. □

Corollary 9.4.1. *Let $f, g : V \to \mathbf{R} \cup \{+\infty\}$ be two closed convex proper functions. Then*

$$(f+g)^* = (f^*\#g^*)^{**}.$$

As a consequence, when the convex function $f^*\#g^*$ is a $\sigma(V^*, V)$ closed proper function, we have

$$(f+g)^* = f^*\#g^*.$$

PROOF. By Proposition 9.4.1 we have

$$(f^*\#g^*)^* = f^{**} + g^{**}.$$

When f and g are assumed to be closed convex and proper, one gets

$$(f^*\#g^*)^* = f + g.$$

Taking again the Legendre–Fenchel conjugate, we obtain

$$(f+g)^* = (f^*\#g^*)^{**}.$$

The function $f^*\#g^*$, as the epi-sum of two convex functions, is still convex (Proposition 9.2.2). When it is $\sigma(V^*, V)$ closed and proper, Theorem 9.3.4 yields

$$(f+g)^* = f^*\#g^*,$$

which completes the proof. □

We can now state the following theorem from Rockafellar [201] and Moreau [182], which, under a so-called qualification assumption on f and g, asserts that $f^*\#g^*$ is $\sigma(V^*, V)$ closed and hence $(f+g)^* = f^*\#g^*$.

Theorem 9.4.1. *Let V be a normed linear space and let $f, g : V \to \mathbf{R} \cup \{+\infty\}$ be two closed convex and proper functions which satisfy the following qualification assumption:*

$$\text{there is a point } u_0 \in \operatorname{dom} f \cap \operatorname{dom} g \text{ where } f \text{ is continuous.} \qquad (Q)$$

Then, $f^*\#g^*$ is a $\sigma(V^*, V)$ closed convex proper function and the following equality holds:

$$(f+g)^* = f^*\#g^*.$$

Moreover, for any $v^* \in V^*$, the infimum in the definition of $f^*\#g^*$ is achieved.

PROOF. Corollary 9.4.1 tells us that the only point we need to verify is that $f^*\#g^*$ is $\sigma(V^*, V)$ closed. Equivalently, we have to prove that for $\lambda \in \mathbf{R}$, the sublevel set of $f^*\#g^*$

$$C = \left\{ v^* \in V^* \ : \ (f^*\#g^*)(v^*) \leq \lambda \right\}$$

is $\sigma(V^*, V)$ closed. Indeed, we are going to establish that for each $\rho > 0$, $C \cap \rho \mathbf{B}_{V^*}$ is $\sigma(V^*, V)$ closed, i.e., the traces of C on all closed balls of V^* are $\sigma(V^*, V)$ closed. It will follow from the Banach–Dieudonné–Krein–Smulian theorem (see, e.g., [127, Theorem V 5.7]) that C is $\sigma(V^*, V)$ closed. Let $(v_n^*)_{n\in\mathbf{N}}$ be a bounded sequence of elements of C with $v_n^* \to v^*$, $\sigma(V^*, V)$. When V is separable, it is not restrictive to consider sequences. For

general V the argument can be readily extended by considering generalized sequences. By definition of $f^*\#g^*$, for each $n \in \mathbf{N}$, there exists some $w_n \in V^*$ such that

$$f^*(v_n^* - w_n^*) + g^*(w_n^*) \leq \lambda + \frac{1}{n}. \tag{9.7}$$

The key point of the proof is to prove that the sequence $(w_n^*)_{n\in\mathbf{N}}$ is bounded in V^*. To that end, we use as an essential fact the qualification assumption (Q): there exist some $r > 0$ and some $M \in \mathbf{R}$ such that

$$\sup_{\|v\|_V \leq 1} f(u_0 + rv) \leq M. \tag{9.8}$$

For any $v \in \mathbf{B}(0, 1)$ let us majorize $\langle w_n^*, v\rangle$. To that end, let us write

$$\begin{aligned} r\langle w_n^*, v\rangle_{(V^*,V)} &= \langle w_n^*, rv\rangle \\ &= \langle w_n^*, u_0\rangle + \langle w_n^*, rv - u_0\rangle \\ &= \langle w_n^*, u_0\rangle + \langle v_n^* - w_n^*, u_0 - rv\rangle - \langle v_n^*, u_0 - rv\rangle \\ &\leq g(u_0) + g^*(w_n^*) + f(u_0 - rv) + f^*(v_n^* - w_n^*) - \langle v_n^*, u_0 - rv\rangle. \end{aligned}$$

We rewrite the above inequality in the form

$$r\langle w_n^*, v\rangle_{(V^*,V)} \leq \Big(f^*(v_n^* - w_n^*) + g^*(w_n^*)\Big) + f(u_0 - rv) + g(u_0) + \|u_0 - rv\|\|v_n^*\|_*$$

and use (9.7), (9.8) to obtain

$$r\langle w_n^*, v\rangle_{(V,V^*)} \leq \lambda + \frac{1}{n} + M + g(u_0) + \|v_n^*\|_*(\|u_0\| + r).$$

Using that the sequence (v_n^*) is bounded, we immediately obtain from the above inequality (which is valid for any $v \in B(0, 1)$) that $\sup_n \|w_n^*\|_* < +\infty$.

We now use the Banach–Alaoglu–Bourbaki theorem, Theorem 1.4.7, and Corollary 1.4.2: when V is separable (for a general V one can use a device of Attouch and Brezis [31]), the unit ball of V^* is $\sigma(V^*, V)$ sequentially compact. As a consequence, one can find a subsequence $(w_{n_k})_{k\in\mathbf{N}}$ and some $w^* \in V^*$ such that $w_{n_k} \rightharpoonup w^*$ in $\sigma(V^*, V)$.

Let us now use the lower semicontinuity of f^* and g^* for the topology $\sigma(V^*, V)$ and pass to the limit in (9.7) to obtain

$$f^*(v^* - w^*) + g^*(w^*) \leq \lambda.$$

As a consequence,

$$(f^*\#g^*)(v^*) \leq f^*(v^* - w^*) + g^*(w^*) \leq \lambda$$

and $v^* \in C$.

The same argument with $\lambda = f^*\#g^*$ and $v_n^* = v^*$ gives that the infimum in the definition of $f^*\#g^*$ is achieved. □

The qualification assumption (Q), because of its importance, has been intensively studied and many weakened versions of it have been established. Let us quote the following result (see Aubin [45]).

Theorem 9.4.2. *Let V be a Banach space and let $f, g : V \to \mathbf{R} \cup \{+\infty\}$ be two closed convex proper functions such that*

$$\text{dom } f - \text{dom } g \text{ is a neighborhood of the origin.}$$

Then, the same conclusions as Theorem 9.4.1 *hold and*

$$(f + g)^* = f^* \# g^*.$$

In the same spirit, the same result was established by Attouch and Brezis in [31] under the even weaker assumption

$$\bigcup_{\lambda > 0} \lambda(\text{dom } f - \text{dom } g) \text{ is a closed subspace of } V.$$

Note that by contrast with the Rockafellar theorem, which holds in general normed spaces, the Aubin and Attouch–Brezis theorems require that the space V is a Banach space. Indeed, an essential ingredient in the proof of these theorems is the Banach–Steinhaus theorem. Otherwise, the proof is essentially the same as in Theorem 9.4.1.

9.5 Subdifferential calculus for convex functions

To obtain the simplest possible dual representation of a closed convex set C of a normed linear space $(V, \|\cdot\|)$, we introduce the notion of supporting hyperplane. When taking $C = \text{epi} f$, the epigraph of a closed convex proper function, the corresponding notion is the exact minoration: a continuous affine function $l : V \to \mathbf{R}$ is an exact minorant of f at u if $l \leq f$ and $l(u) = f(u)$.

Equivalently, when setting $l(v) = \langle u^*, v\rangle + \alpha$, this becomes

$$\begin{cases} f(v) \geq \langle u^*, v\rangle + \alpha \quad \forall\, v \in V, \\ f(u) = \langle u^*, u\rangle + \alpha, \end{cases}$$

i.e., $\alpha = f(u) - \langle u^*, u\rangle,\ l(v) = f(u) + \langle u^*, v - u\rangle$, which is equivalent to

$$\forall v \in V \quad f(v) \geq f(u) + \langle u^*, v - u\rangle.$$

This leads to the following definition.

Definition 9.5.1. *Let $(V, \|\cdot\|)$ be a normed space and $f : V \to \mathbf{R} \cup \{+\infty\}$ be a closed convex proper function. We say that an element $u^* \in V^*$ belongs to the subdifferential of f at $u \in V$ if*

$$\forall v \in V \quad f(v) \geq f(u) + \langle u^*, v - u\rangle_{(V^*, V)}.$$

We then write $u^* \in \partial f(u)$.

The terminology reflects the fact that when f is continuously differentiable and convex, the following inequality holds:

$$\forall v \in V \quad f(v) \geq f(u) + \langle \nabla f(u), v - u \rangle.$$

Moreover, this inequality characterizes $\nabla f(u)$. For this reason, when $u^* \in \partial f(u)$, we say either that u^* belongs to the subdifferential of f at u or that u^* is a subgradient of f at u.

Note that if $u^* \in \partial f(u)$, then necessarily $u \in \operatorname{dom} f$ (take $v_0 \in \operatorname{dom} f \neq \emptyset$, we have $f(v_0) - \langle u^*, v_0 - u \rangle \geq f(u)$ and $f(u) < +\infty$). It is also important to notice that given $u \in \operatorname{dom} f$, the set $\partial f(u)$ may be empty; see Phelps [197, Example 3.8].

Proposition 9.5.1. *Let $(V, \|\cdot\|)$ be a normed space and $f : V \to \mathbf{R} \cup \{+\infty\}$ a closed convex proper function. Then the two following conditions are equivalent:*

(i) $u^* \in \partial f(u)$,

(ii) $f(u) + f^*(u^*) - \langle u^*, u \rangle = 0$.

PROOF. (a) Let us first give a geometrical proof: to say that $u^* \in \partial f(u)$ means that $v \mapsto \langle u^*, v \rangle + f(u) - \langle u^*, u \rangle$ is an exact minorant of f at u. This implies that it is a maximal minorant with slope u^*, i.e., $f(u) - \langle u^*, u \rangle = -f^*(u^*)$.

(b) The analytic proof is also immediate: the inequality

$$f(u) + f^*(u^*) - \langle u^*, u \rangle \geq 0$$

is always true. Thus the equality $f(u) + f^*(u^*) - \langle u^*, u \rangle = 0$ is equivalent to the inequality

$$f(u) + f^*(u^*) - \langle u^*, u \rangle \leq 0.$$

By definition of f^* this is equivalent to saying

$$\langle u^*, u \rangle - f(u) \geq \langle u^*, v \rangle - f(v) \; \forall \; v \in V,$$

i.e., $u^* \in \partial f(u)$. □

Remark 9.5.1. As we have already stressed, for any $v \in V$ and $v^* \in V^*$ the inequality

$$f(v) + f^*(v^*) - \langle v^*, v \rangle \geq 0$$

is always true. Thus, when writing the characterization of $u^* \in \partial f(u)$,

$$f(u) + f^*(u^*) - \langle u^*, u \rangle = 0,$$

we express that for the pair $(u, u^*) \in V \times V^*$, the function $(v, v^*) \mapsto f(v) + f^*(v^*) - \langle v^*, v \rangle$ takes its minimal value. For this reason, relation (ii) in Proposition 9.5.1 is called the *Fenchel extremality relation.*

A major interest of the Fenchel extremality characterization of subdifferentials is that f and f^* play a symmetric role in its formulation. This together with the F-M-R duality theorem, Theorem 9.3.2 (which expresses that $f = f^{**}$), yields the following result.

Theorem 9.5.1. *Let $(V, \|\cdot\|)$ be a normed space and let $f : V \to \mathbf{R} \cup \{+\infty\}$ be a closed convex and proper function. Then, for $u \in V$ and $u^* \in V^*$ we have*

$$u^* \in \partial f(u) \iff u \in \partial f^*(u^*).$$

PROOF. In fact we obtain

$$\begin{aligned} u^* \in \partial f(u) &\iff f(u) + f^*(u^*) - \langle u^*, u\rangle = 0 \\ &\iff f^{**}(u) + f^*(u^*) - \langle u^*, u\rangle = 0 \\ &\iff u \in \partial f^*(u^*), \end{aligned}$$

where we use the Fenchel extremality characterization of the subdifferential and the F-M-R duality theorem, Theorem 9.3.2 ($f = f^{**}$). □

Remark 9.5.2. When using the notation of set-valued analysis we can write

$$(\partial f)^{-1} = \partial f^*.$$

This is indeed the formulation, in terms of subdifferentials, of the convex duality theory.

For theoretical reasons it is important to know if a closed convex proper function can be uniquely determined (up to a constant) by its subdifferential. From a geometrical point of view, this can be formulated in the following form: *is a closed convex proper function the upper envelope of its exact continuous affine minorants?* Indeed the answer is yes when V is a Banach space. The proof of this result relies on the Ekeland's ε-variational principle. (See Section 3.4. We state it without proof, referring, for instance, to Phelps [197, Corollary 3.1.9]).

Theorem 9.5.2. *Suppose $(V, \|\cdot\|)$ is a Banach space and $f : V \to \mathbf{R} \cup \{+\infty\}$ is a closed convex proper function. Then, for any $u \in \operatorname{dom} f$*

$$\begin{aligned} f(u) &= \sup\left\{ f(v) + \langle v^*, u - v\rangle \ : \ v \in V,\ v^* \in V^* \text{ with } v^* \in \partial f(v)\right\} \\ &= \sup\left\{ \langle v^*, u\rangle - f^*(v^*) \ : \ \exists v \in V \text{ such that } v^* \in \partial f(v)\right\}. \end{aligned}$$

Note that this theorem, when specialized to convex sets, says that any closed convex nonempty set in a Banach space is the intersection of the closed half-spaces defined by its supporting hyperplanes (Phelps [197, Proposition 3.2.1]).

When proving the above theorem via Ekeland's variational principle one obtains in the process the following density result.

Theorem 9.5.3. *Suppose $(V, \|\cdot\|)$ is a Banach space and $f : V \to \mathbf{R} \cup \{+\infty\}$ is a closed convex proper function. Then,* $\operatorname{dom} \partial f$ *is dense in* $\operatorname{dom} f$*. More precisely, for any $v \in \operatorname{dom} f$, there exists a sequence $(v_n)_{n\in\mathbf{N}}$ with $v_n \in \operatorname{dom} \partial f$ for all $n \in \mathbf{N}$ such that*

$$v_n \to v \text{ and } f(v_n) \to f(v).$$

PROOF. For the proof, see Azé [49, Theorem 3.2.4] and Aubin–Ekeland [46, Theorem 3]. □

To develop a calculus for subdifferentials it is convenient to consider ∂f as a multivalued operator,

$$\partial f : V \rightrightarrows V^*,$$

and to identify ∂f with its graph

$$\partial f = \big\{(v, v^*) \in V \times V^* \ : \ v^* \in \partial f(v)\big\}.$$

We recall the basic definitions for calculus of set-valued mappings: given $A, B : V \rightrightarrows V^*$ we have

$$\begin{aligned} \operatorname{dom} A &= \big\{v \in V \ : \ \exists v^* \in V^* \text{ with } (v, v^*) \in A\big\}, \\ A^{-1} &= \big\{(v^*, v) \in V^* \times V \ : \ (v, v^*) \in A\big\}, \end{aligned}$$

$$\begin{cases} \operatorname{dom}(A + B) = \operatorname{dom} A \cap \operatorname{dom} B, \\ (A + B)(v) = Av + Bv \quad \text{in the sense of vectorial sum.} \end{cases}$$

Moreover, we say that $A \subset B$ if graph $A \subset$ graph B.

As we have already stressed, the convex duality theory can be expressed as

$$\partial f^* = (\partial f)^{-1}.$$

Theorem 9.5.4. *Let $(V, \|\cdot\|)$ be a normed space and let $f, g : V \to \mathbf{R} \cup \{+\infty\}$ be two closed convex proper functions.*

(a) *The following inclusion is always true:*

$$\partial f + \partial g \subset \partial(f + g).$$

(b) *If moreover the qualification assumption (Q) holds,*

$$f \text{ is finite and continuous at a point of } \operatorname{dom} g, \tag{Q}$$

then we have

$$\partial f + \partial g = \partial(f + g).$$

PROOF. (a) Take $u \in \operatorname{dom} \partial f \cap \operatorname{dom} \partial g$, $u^* \in \partial f(u)$, and $w^* \in \partial g(u)$. By definition of ∂f and ∂g, for any $v \in V$,

$$f(v) \geq f(u) + \langle u^*, v - u\rangle,$$

$$g(v) \geq g(u) + \langle w^*, v - u\rangle.$$

By adding these two inequalities, we obtain for any $v \in V$

$$(f + g)(v) \geq (f + g)(u) + \langle u^* + w^*, v - u\rangle,$$

i.e., $u^* + w^* \in \partial(f + g)(u)$.

(b) Take $u^* \in \partial(f+g)(u)$. Equivalently, by using the Fenchel extremality relation, we obtain

$$(f+g)(u) + (f+g)^*(u^*) - \langle u^*, u\rangle = 0.$$

By Theorem 9.4.1, we have

$$(f+g)^*(u^*) = (f^* \# g^*)(u^*)$$

and the infimum in the definition of $(f^*\#g^*)(u^*)$ is achieved. Consequently, there exists some $w^* \in V^*$ such that

$$(f+g)(u) + f^*(u^* - w^*) + g^*(w^*) - \langle u^*, u\rangle = 0.$$

Equivalently,

$$(f(u) + f^*(u^* - w^*) - \langle u^* - w^*, u\rangle) + (g(u) + g^*(w^*) - \langle w^*, u\rangle) = 0.$$

By the Fenchel inequality,

$$f(u) + f^*(u^* - w^*) - \langle u^* - w^*, u\rangle \geq 0,$$

$$g(u) + g^*(w^*) - \langle w^*, u\rangle \geq 0.$$

Since the sum of these two quantities is equal to zero, we obtain

$$f(u) + f^*(u^* - w^*) - \langle u^* - w^*, u\rangle = 0,$$

$$g(u) + g^*(w^*) - \langle w^*, u\rangle = 0.$$

These are the Fenchel extremality relations (Proposition 9.5.1) and they are equivalent to

$$u^* - w^* \in \partial f(u) \quad \text{and} \quad w^* \in \partial g(u).$$

Finally, we obtain

$$u^* = (u^* - w^*) + w^* \in \partial f(u) + \partial g(u),$$

i.e., $u^* \in (\partial f + \partial g)(u)$. □

We already stressed the fact that for $u \in \operatorname{dom} f$, the set $\partial f(u)$ may be empty. The following result, which can be viewed as a corollary of Theorem 9.5.4, gives a sufficient condition for the set $\partial f(u)$ to be nonempty. This result, as we will see in Sections 9.6 and 9.8, is quite useful for applications.

Proposition 9.5.2. *Let $(V, \|\cdot\|)$ be a normed space and let $f : V \to \mathbf{R} \cup \{+\infty\}$ be a closed convex and proper function. Let us assume that f is continuous at $u \in \operatorname{dom} f$. Then $\partial f(u) \neq \emptyset$ and $\partial f(u)$ is a closed convex and bounded subset of V^*.*

PROOF. Let us apply Theorem 9.4.1 to the sum of the two closed convex and proper functions f and $g = \delta_{\{u\}}$ (g is the indicator function of the singleton $\{u\}$). By assumption, f is continuous at the point u, and the qualification assumption (Q) of Theorem 9.4.1 is satisfied.

Hence, for any $v^* \in V^*$, the equality

$$(f + \delta_{\{u\}})^*(v^*) = (f^* \# \delta^*_{\{u\}})(v^*)$$

holds, and the infimum in the formulation of $(f^* \# \delta^*_{\{u\}})(v^*)$ is achieved. An elementary computation yields

$$(f + \delta_{\{u\}})^*(v^*) = \langle v^*, u\rangle - f(u),$$
$$\delta^*_{\{u\}}(w^*) = \langle w^*, u\rangle.$$

Hence, for any $v^* \in V^*$, there exists some $w^* \in V^*$ such that

$$\langle v^*, u\rangle - f(u) = f^*(v^* - w^*) + \langle w^*, u\rangle.$$

Equivalently,

$$f(u) + f^*(v^* - w^*) - \langle v^* - w^*, u\rangle = 0.$$

This is the Fenchel extremality relation. This is equivalent to

$$v^* - w^* \in \partial f(u),$$

which expresses that $\partial f(u) \neq \emptyset$.

Note that, as well, we may have applied Theorem 9.5.4 instead of Theorem 9.4.1 to obtain the above result.

As a general rule, the set $\partial f(u)$ is closed and convex. This is an immediate consequence of the definition of $\partial f(u)$. Let us now verify that under the continuity assumption of f at u, this set is bounded. Since f is continuous at u, it is bounded on a neighborhood of u. Let $r > 0$ and $M \geq 0$ be such that

$$f(u + rv) \leq M \text{ for all } v \in \mathbf{B}(0, 1).$$

Take $v^* \in \partial f(u)$. By definition of ∂f, we have for all $v \in \mathbf{B}(0, 1)$

$$f(u + rv) \geq f(u) + r\langle v^*, v\rangle.$$

Hence

$$\langle v^*, v\rangle \leq \frac{1}{r}(M + |f(u)|).$$

This being true for any $v \in \mathbf{B}(0, 1)$, we obtain

$$\|v^*\|_* \leq \frac{1}{r}(M + |f(u)|),$$

and, as a consequence, the set $\partial f(u)$ is bounded. □

Let us now come to the central role played by the subdifferential calculus in convex optimization. The following result, despite its elementary proof (it is a straightforward consequence of the definition of ∂f), shows the role of the subdifferential optimality rule $\partial f(u) \ni 0$ as a substitute to the classical Fermat rule.

Proposition 9.5.3. *Let $(V, \|\cdot\|)$ be a normed space and let $f : V \to \mathbf{R} \cup \{+\infty\}$ be a closed convex and proper function. Then, for an element $u \in V$ the two following statements are equivalent:*

(i) $f(u) \leq f(v)$ *for all* $v \in V$;

(ii) $\partial f(u) \ni 0$.

Let us stress that the above proposition gives a necessary and sufficient condition for an element $u \in V$ to be a solution of the convex minimization problem

$$\min \big\{ f(v) \ : \ v \in V \big\}.$$

This necessary and sufficient condition

$$\partial f(u) \ni 0$$

is an extension to nonsmooth convex functions of the classical first-order necessary and sufficient condition of optimality for convex $\mathbf{C}^1$ functions, namely,

$$\nabla f(u) = 0.$$

Thus, for a given convex optimization problem, the problem which consists in finding the optimal solutions can be attacked by using the subdifferential calculus and solving the generalized equation $\partial f(u) \ni 0$.

As we have stressed, Legendre–Fenchel calculus and subdifferential calculus are intimately connected; playing with both of them when passing from one formulation to the other gives a lot of flexibility and makes a rich calculus. This calculus is made even richer when exploiting some of its geometrical aspects (duality via polar cones, etc.).

Let us develop these ideas in the following general approach to optimization (both finite and infinite dimensional) problems. Let $f_0 : V \to \mathbf{R} \cup \{+\infty\}$ be a closed convex proper function (objective function) on a normed linear space V, and let $C \subset V$ be a closed convex nonempty subset of V (set of constraints). Consider the following optimization problem:

$$\inf \big\{ f_0(v) \ : \ v \in C \big\}. \tag{$\mathcal{P}$}$$

It can be written in the equivalent form

$$\inf \big\{ f(v) \ : \ v \in V \big\},$$

where $f := f_0 + \delta_C$.

An element $u \in V$ is an optimal solution of $(\mathcal{P})$ iff $\partial f(u) \ni 0$. To compute ∂f we assume that the qualification assumption (Q) is satisfied:

$$f_0 \text{ is continuous at a point of } C \text{ or } \operatorname{int} C \cap \operatorname{dom} f_0 \neq \emptyset. \tag{Q}$$

Then, Theorem 9.5.4 tells us that it is equivalent to look for a solution of the equation

$$\partial f_0(u) + \partial \delta_C(u) \ni 0.$$

To describe the subdifferential of the indicator function of a closed convex set C, we need to introduce the notion of tangent and normal cone to C at a point $u \in C$.

Definition 9.5.2. *Let C be a closed convex nonempty subset of a normed space V and let $u \in C$.*

(a) *The tangent cone to C at u, denoted by $T_C(u)$ is defined by*

$$T_C(u) = \overline{\bigcup_{\lambda \geq 0} \lambda(C - u)}.$$

It is the closure of the cone spanned by $C - u$.

(b) *The normal cone (also called outward normal cone) $N_C(u)$ to C at $u \in C$ is the polar cone of the tangent cone:*

$$\begin{aligned} N_C(u) &= \big\{v^* \in V^* \ : \ \langle v^*, v\rangle \leq 0 \quad \forall\, v \in T_C(u)\big\} \\ &= \big\{v^* \in V^* \ : \ \langle v^*, v - u\rangle \leq 0 \quad \forall\, v \in C\big\}. \end{aligned}$$

Proposition 9.5.4. *Let C be a closed convex nonempty subset of a normed space V. Then, for every $u \in C$,*

$$\partial\delta_C(u) = N_C(u).$$

Proof. By definition of the subdifferential

$$u^* \in \partial\delta_C(u) \iff \delta_C(v) \geq \delta_C(u) + \langle u^*, v - u\rangle \ \forall v \in C$$

$$\iff \begin{cases} u \in C, \\ \langle u^*, v - u\rangle \leq 0 \quad \forall\, v \in C \end{cases}$$

$$\iff \begin{cases} u \in C, \\ \langle u^*, v\rangle \leq 0 \quad \forall\, v \in T_C(u), \end{cases}$$

that is, $u^* \in N_C(u)$. □

An equivalent and quite useful characterization of $N_C(u)$ is given by using the Fenchel extremality relation:

$$\begin{aligned} u^* \in N_C(u) &\iff u^* \in \partial\delta_C(u) \\ &\iff \delta_C(u) + \delta_C^*(u^*) = \langle u^*, u\rangle \\ &\iff \sigma_C(u^*) = \langle u^*, u\rangle, \end{aligned}$$

where we have used that $\delta_C^* = \sigma_C$ (see Proposition 9.3.1). Let us formulate this result precisely.

Proposition 9.5.5. *Let C be a closed convex nonempty subset of a normed linear space V. For every $u \in C$ we have*

$$N_C(u) = \big\{u^* \in V^* \ : \ \langle u^*, u\rangle = \max\{\langle u^*, v\rangle \ : \ v \in C\}\big\}.$$

Equivalently, an element u^* of $N_C(u)$ is characterized by the fact that the linear form $v \mapsto \langle u^*, v\rangle$ attains its maximum on C at the point u.

Let us come back to the convex constrained optimization problem $(\mathcal{P})$. We can summarize the previous results in the following statement.

Theorem 9.5.5. *Let $(V, \|\cdot\|)$ be a normed space, let $f_0 : V \to \mathbf{R} \cup \{+\infty\}$ be a closed convex and proper function, and let $C \subset V$ be a closed convex nonempty subset. We assume that one of the two following qualification assumptions (Q_1) or (Q_2) is satisfied:*

$$f_0 \text{ is continuous at some point of } C, \tag{Q_1}$$

$$\operatorname{dom} f_0 \cap \operatorname{int} C \neq \emptyset. \tag{Q_2}$$

Then the following statements are equivalent:

(i) *u is an optimal solution of the minimization problem $(\mathcal{P})$*

$$\min\big\{f_0(v) \ : \ v \in C\big\}; \tag{$\mathcal{P}$}$$

(ii) *u is a solution of the equation*

$$\partial f_0(u) + N_C(u) \ni 0;$$

(iii) *there exists some $u^* \in V^*$ such that*

$$\begin{cases} u \in C, \\ u^* \in \partial f_0(u), \\ \langle u^*, v - u\rangle \geq 0 \ \forall\, v \in C. \end{cases}$$

To go further we need to enrich the model and give more information on the structure of the set of constraints C. Because of its practical importance, in the next subsection we are going to pay particular attention to the mathematical convex programming theory (and in particular to linear programming) and the theory of multipliers. We will see how the notion of dual problem naturally occurs.

When f_0 is a smooth convex function, say, $f_0 \in \mathbf{C}^1(V, \mathbf{R})$, Theorem 9.5.5 takes the following simpler equivalent form: u is an optimal solution of the above minimization problem $(\mathcal{P})$ iff

(iii) $$u \in C \text{ and } \langle \nabla f_0(u), v - u\rangle \geq 0 \text{ for every } v \in C.$$

Problem (iii) is a particular case of the following general *variational inequality problem*: given an operator $A : V \to V^*$ and $z \in V^*$

$$\begin{cases} \text{find } u \in C \text{ such that} \\ \langle Au, v - u\rangle \geq \langle z, v - u\rangle \quad \forall\, v \in C. \end{cases}$$

Note that when $C = V$ (i.e., there are no constraints), the above problem reduces to the standard equation $Au = z$.

As an example, let us examine the important case where C is a closed convex cone such that $C \cap (-C) = \{0\}$. Then, C is equal to the positive cone for the partial ordering $v \geq u \iff v - u \in C$. Then problem (iii) takes the following equivalent form:

$$\begin{cases} \nabla f_0(u) \geq 0, \\ u \geq 0, \\ \langle \nabla f_0(u), u \rangle = 0. \end{cases}$$

(The last equality is obtained by taking successively $v = 0$ and $v = 2u$ in (iii).) This type of problem is called a *complementarity problem*.

Take now a closely related problem where $C = \{v \in V \ : \ v \geq g\}$, where $g \in V$ is given. One can easily obtain that (iii) becomes

$$\begin{cases} \nabla f_0(u) \geq 0, \\ u \geq g, \\ \langle \nabla f_0(u), u - g \rangle = 0. \end{cases}$$

When $V = H_0^1(\Omega)$ and $f_0(v) = \frac{1}{2} \int_\Omega |\nabla v(x)|^2 \, dx$ is the Dirichlet integral, we obtain

$$\begin{cases} -\Delta u \geq 0, \\ u \geq g, \\ \langle \Delta u, u - g \rangle = 0. \end{cases}$$

The first condition expresses that $-\Delta u = \mu \geq 0$ is a nonnegative Radon measure. The last condition (complementary condition) can be recognized as

$$\int_\Omega (\tilde{u} - \tilde{g}) \, d\mu = 0,$$

where $\tilde{u}$ and $\tilde{g}$ are the quasi-continuous representatives of u and g. It expresses that $\mu = -\Delta u$ does not charge the set where $\tilde{u} > \tilde{g}$.

In other words, μ is concentrated on the *contact set* $\omega = \{\tilde{u} = \tilde{g}\}$ and we have to solve the free boundary value problem:

$$\begin{cases} -\Delta u = 0 \text{ on } \Omega \setminus \omega, \\ u = g \text{ on } \omega, \\ u = 0 \text{ on } \partial\Omega. \end{cases}$$

9.6 Mathematical programming: Multipliers and duality

In this section, $(V, \|.\|)$ is a normed space. Mathematical programming is concerned with optimization problems of the form

$$\min \{f_0(v) \ : \ f_1(v) \leq 0, \ldots, f_n(v) \leq 0\}, \tag{$\mathcal{P}$}$$

where f_i $(i = 1, \ldots, n)$ are given functions from V into $\mathbf{R}$.

Thus, a mathematical programming problem is an optimization problem where the constraint C has the following specific form:

$$C = \{v \in V \ : \ f_i(v) \leq 0, \quad i = 1, \ldots, n\}.$$

This problem is of fundamental importance; a large number of problems in decision sciences, engineering, and so forth can be written as mathematical programming problems.

The mathematical analysis of this kind of problem depends heavily on the geometrical properties of the functions f_i $(i = 0, \ldots, n)$. When the functions f_i are affine, $(\mathcal{P})$ is called a linear programming problem. When f_i $(i = 1, \ldots, n)$ are affine and f_0 is quadratic, $(\mathcal{P})$ is called a quadratic programming problem.

In this section, we study the situation where $f_0, f_1, \ldots, f_n$ are supposed to be convex functions. Thus $(\mathcal{P})$ is a convex minimization problem (f_0 and C are convex); it is called a convex mathematical programming problem.

9.6.1 Karush–Kuhn–Tucker optimality conditions

The following theorem, which is the central result of this section, will be obtained by applying Theorem 9.5.5 to our situation.

Because of the specific form of the constraint C, the constraint qualification assumption (Q) takes a quite simple form (this is the Slater qualification assumption). The computation of the normal cone $N_C(u)$ provides, as fundamental mathematical objects, the Karush–Kuhn–Tucker optimality conditions and the corresponding Lagrange multipliers.

Theorem 9.6.1. *Suppose that V is a normed space, $f_0 : V \to \mathbf{R} \cup \{+\infty\}$ is closed convex proper, and $f_1, \ldots, f_n : V \to \mathbf{R}$ are convex and continuous. Suppose moreover that the following Slater qualification assumption is satisfied:*

there exists some $v_0 \in V$ such that $f_0(v_0) < +\infty$ and such that $f_i(v_0) < 0 \ \forall i = 1, \ldots, n$.

Then, the following statements are equivalent:

(i) *u is a solution of problem $(\mathcal{P})$ above;*

(ii) *there exist $\lambda_1, \lambda_2, \ldots, \lambda_n$ in $\mathbf{R}^+$ such that*

$$\begin{cases} \partial f_0(u) + \lambda_1 \partial f_1(u) + \ldots \lambda_n \partial f_n(u) \ni 0, \\ \lambda_i \geq 0 \quad \forall i = 1, \ldots, n, \\ f_i(u) \leq 0 \quad \forall i = 1, \ldots, n, \\ \lambda_i f_i(u) = 0 \quad \forall i = 1, \ldots, n. \end{cases}$$

The central point of the proof of Theorem 9.6.1 is the computation of the normal cone $N_C(u)$. We are going to do it first when C is a closed half-space (that is, when $C = \{v \in V \ : \ f(v) \leq 0\}$ with f affine continuous) and then in the general case.

Lemma 9.6.1. *Let $(V, \|.\|)$ be a normed space and $u^* \in V^*$ with $u^* \neq 0$. Let us consider the closed half-space*

$$\mathcal{H} = \{v \in V \ : \ \langle u^*, v - u \rangle \leq 0\}.$$

Then, $N_{\mathcal{H}}(u) = \mathbf{R}_+ u^*$.

In other words, $v^* \in V^*$ belongs to the normal cone to $\mathcal{H}$ at u iff there exists some $\lambda \geq 0$ such that $v^* = \lambda u^*$.

PROOF. The inclusion $\mathbf{R}_+ u^* \subset N_{\mathcal{H}}(u)$ is immediate: by definition of $\mathcal{H}$, we have $\langle u^*, v - u\rangle \leq 0$ for all $v \in \mathcal{H}$. Hence $u^* \in N_{\mathcal{H}}(u)$ and $\mathbf{R}_+ u^* \subset N_{\mathcal{H}}(u)$.

Conversely, take $v^* \in N_{\mathcal{H}}(u)$, $v^* \neq 0$ (the case $v^* = 0$ is trivial). By definition of $N_{\mathcal{H}}(u)$, we have

$$\langle v^*, v - u\rangle \leq 0 \ \forall\, v \in \mathcal{H}. \tag{9.9}$$

As a particular subset of $\mathcal{H}$, let us consider the affine subspace

$$W = \big\{v \in V \ : \ \langle u^*, v - u\rangle = 0\big\}.$$

We have $W = u + M$ where $M = \ker u^*$ is the hyperspace

$$M = \big\{v \in V \ : \ \langle u^*, v\rangle = 0\big\}.$$

By taking in (9.9) elements v belonging to $W = u + M$, we obtain

$$\langle v^*, v\rangle \leq 0 \ \forall\, v \in M.$$

Then, replace v by $-v$ (M is a subspace) to obtain

$$\langle v^*, v\rangle = 0 \ \forall\, v \in M.$$

We now follow a standard device in linear algebra. Take an arbitrary element $w \in V$, $w \notin M$; noticing that

$$\left\langle u^*, v - \frac{\langle u^*, v\rangle}{\langle u^*, w\rangle} w \right\rangle = 0,$$

we deduce that for every $v \in V$,

$$v - \frac{\langle u^*, v\rangle}{\langle u^*, w\rangle} w \in M = \ker u^*.$$

Since $v^* = 0$ on M, we have

$$\langle v^*, v\rangle = \left\langle v^*, \frac{\langle u^*, v\rangle}{\langle u^*, w\rangle} w \right\rangle,$$

that is,

$$\langle v^*, v\rangle = \left\langle \frac{\langle v^*, w\rangle}{\langle u^*, w\rangle} u^*, v \right\rangle.$$

This being true for all $v \in V$, we finally obtain

$$v^* = \frac{\langle v^*, w\rangle}{\langle u^*, w\rangle} u^*,$$

i.e., $v^* = t u^*$ for some $t \in \mathbf{R}$.

Until now, we have exploited only a part of the information given by (9.9). Returning to (9.9), t must satisfy

$$t\langle u^*, v-u\rangle \leq 0 \ \forall\, v \in \mathcal{H}.$$

Since for all $v \in \mathcal{H}$ $\langle u^*, v-u\rangle \leq 0$, we necessarily have $t \geq 0$. □

Let us now examine the situation where $C = \{v \in V \ : \ f(v) \leq 0\}$ and compute the normal cone $N_C(u)$ at an arbitrary point u of C.

Proposition 9.6.1. *Suppose that $f : V \to \mathbf{R}$ is a convex continuous function on a normed linear space V. Set*

$$C = \{v \in V \ : \ f(v) \leq 0\}$$

and assume that C satisfies the following Slater property:

$$\text{there exists some } v_0 \in C \text{ such that } f(v_0) < 0.$$

Then, for every $u \in C$

$$N_C(u) = \begin{cases} \{0\} & \text{if } f(u) < 0, \\ \mathbf{R}^+\partial f(u) & \text{if } f(u) = 0. \end{cases}$$

As a consequence,

$$u^* \in N_C(u) \iff \exists \lambda \geq 0 \text{ such that } u^* \in \lambda \partial f(u) \text{ and } \lambda f(u) = 0.$$

Proof. Take $u \in C$. If $f(u) < 0$, because of the continuity of f, we have $u \in \operatorname{int} C$, which yields $T_C(u) = V$ and hence $N_C(u) = \{0\}$.

If on the contrary $f(u) = 0$, let us prove that $N_C(u) = \mathbf{R}_+\partial f(u)$. The inclusion $\mathbf{R}_+\partial f(u) \subset N_C(u)$ is quite easy to verify: take $u^* \in \partial f(u)$; by definition of the subdifferential $\partial f(u)$ of f at u

$$\forall v \in V \quad f(v) \geq f(u) + \langle u^*, v-u\rangle.$$

Noticing that $f(u) = 0$ and $f(v) \leq 0$ for all $v \in C$, we obtain

$$\langle u^*, v-u\rangle \leq 0 \ \forall\, v \in C,$$

i.e., $u^* \in N_C(u)$. Since $N_C(u)$ is a cone, we obtain $\mathbf{R}_+\partial f(u) \subset N_C(u)$.

Let us now prove the opposite inclusion, which is the delicate part of the proof: $N_C(u) \subset \mathbf{R}_+\partial f(u)$. Equivalently, we have to prove that if $f(u) = 0$ and $u^* \in N_C(u)$, then there exists some $\lambda \geq 0$ such that $u^* \in \lambda\partial f(u)$. The case $u^* = 0$ is trivial, so we assume in the following that $u^* \neq 0$. We are going to prove the existence of such λ by using a variational argument. As a direct consequence of the definition of the normal cone, we have (see Proposition 9.5.5) the equivalence

$$u^* \in N_C(u)$$

$$\Updownarrow$$

the linear form $v \mapsto \langle u^*, v\rangle$ attains its maximal value on C at $u \in C$.

As a general property of a linear form, the maximum of the linear form $v \mapsto \langle u^*, v\rangle$ on C is attained on its boundary and

$$v \in \operatorname{int} C \Longrightarrow \langle u^*, v\rangle < \langle u^*, u\rangle.$$

Hence

$$f(v) < 0 \Longrightarrow \langle u^*, v\rangle < \langle u^*, u\rangle.$$

Therefore

$$\langle u^*, v-u\rangle \geq 0 \Longrightarrow f(v) \geq 0,$$

that is, on the closed half-space $\mathcal{H} = \big\{v \in V \ : \ \langle u^*, v-u\rangle \geq 0\big\}$ we have $f(v) \geq 0$. Noticing that $u \in \mathcal{H}$ and $f(u) = 0$, we have the following variational property:

"f achieves its minimal value on the half-space $\mathcal{H}$ at the point u."

Hence, $\partial(f + \delta_{\mathcal{H}})(u) \ni 0$. Since f is continuous, we can apply Theorem 9.5.5 to obtain

$$\partial f(u) + N_{\mathcal{H}}(u) \ni 0.$$

We are in the situation described in Lemma 9.6.1. Noticing that

$$\mathcal{H} = \big\{v \in V \ : \ \langle -u^*, v-u\rangle \leq 0\big\},$$

we thus have $N_{\mathcal{H}}(u) = \mathbf{R}_+(-u^*) = \mathbf{R}_-(u^*)$.

As a consequence, there exists some $t \leq 0$ such that

$$\partial f(u) + tu^* \ni 0.$$

Let us finally prove that $t < 0$. Otherwise, $t = 0$ and $\partial f(u) \ni 0$, which expresses that f attains its minimal value at u. This is impossible because $f(u_0) < 0$ (Slater condition) and $f(u) = 0$. Thus $t < 0$, and, dividing by t the above relation, we obtain $u^* \in -\frac{1}{t}\partial f(u)$, i.e., $u^* \in \mathbf{R}_+\partial f(u)$. □

We have now all the elements to prove Theorem 9.6.1.

Proof of Theorem 9.6.1. Let us first verify that all the assumptions of Theorem 9.5.5 are satisfied. Since the functions f_i are continuous, the Slater condition implies that $v_0 \in \operatorname{int} C$. Since $f_0(v_0) < +\infty$, we have $\operatorname{dom} f \cap \operatorname{int} C \neq \emptyset$ and the qualification assumption (Q_2) is satisfied.

Thus u is a solution of the convex programming problem $(\mathcal{P})$ iff

$$\partial f_0(u) + N_C(u) \ni 0.$$

Then notice that $C = \bigcap_{i=1}^n C_i$, where $C_i = \big\{v \in V \ : \ f_i(v) \leq 0\big\}$, which is equivalent to saying that $\delta_C = \delta_{C_1} + \cdots + \delta_{C_n}$. The Slater condition implies that each of the closed convex functions $f_i = \delta_{C_i}$ is continuous at the point v_0. Thus, the subdifferential rule for the sum of convex functions (see Theorem 9.5.4) gives

$$\partial\delta_C = \partial\delta_{C_1} + \cdots + \partial\delta_{C_n},$$

that is, for any $u \in C$,

$$N_C(u) = N_{C_1}(u) + \cdots + N_{C_n}(u).$$

We now combine these results with Proposition 9.6.1 to obtain the existence of real numbers $\lambda_1 \geq 0, \ldots, \lambda_n \geq 0$ such that

$$\partial f_0(u) + \lambda_1 \partial f_1(u) + \cdots + \lambda_n \partial f_n(u) \ni 0,$$
$$\lambda_i = 0 \text{ if } f_i(u) < 0.$$

Thus, in all cases $\lambda_i f_i(u) = 0$. □

9.6.2 The marginal approach to multipliers

Let us first restate Theorem 9.6.1 in a variational way.

Proposition 9.6.2. *Assume that the hypotheses of Theorem* 9.6.1 *are satisfied. Let u be an optimal solution of the minimization problem*

$$\min \big\{ f_0(v) \ : \ f_i(v) \leq 0, \qquad i = 1, \ldots, n \big\}. \tag{$\mathcal{P}$}$$

(a) *Then, there exists some vector $\lambda \in \mathbf{R}^n_+$ such that u is a solution of the unconstrained minimization problem:*

$$\min \left\{ f_0(v) + \sum_{i=1}^{n} \lambda_i f_i(v) \ : \ v \in V \right\}. \tag{$\mathcal{P}_\lambda$}$$

Moreover, the complementarity slackness condition holds:

$$\lambda_i f_i(u) = 0, \qquad i = 1, \ldots, n.$$

(b) *Conversely, if for some $\lambda \in \mathbf{R}^n_+$, u is a solution of the unconstrained minimization problem $(\mathcal{P}_\lambda)$ and*

$$\begin{cases} f_i(u) \leq 0, & i = 1, \ldots, n, \\ \lambda_i f_i(u) = 0, & i = 1, \ldots, n, \end{cases}$$

then u is an optimal solution of the minimization problem $(\mathcal{P})$.

Proof. Just notice that, since for all $i = 1, \ldots, n$ the functions f_i are supposed to be continuous, the additivity rule for subdifferentials holds,

$$\partial f_0 + \lambda_1 \partial f_1 + \cdots + \lambda_n \partial f_n = \partial (f_0 + \lambda_1 f_1 + \cdots + \lambda_n f_n),$$

and the Karush–Kuhn–Tucker condition can be written in the form

$$\partial \left(f_0 + \sum_{i=1}^{n} \lambda_i f_i \right)(u) \ni 0.$$

This expresses that u is a solution of the convex unconstrained minimization problem $(\mathcal{P}_\lambda)$. □

Definition 9.6.1. *Let u be an optimal solution of the minimization problem $(\mathcal{P})$ above. We call a vector $\lambda \in \mathbf{R}^n_+$ a Lagrange multiplier vector for u if*

$$\partial f_0(u) + \sum_{i=1}^{n} \lambda_i \partial f_i(u) \ni 0$$

and

$$\lambda_i f_i(u) = 0, \qquad i = 1, \ldots, n.$$

The determination of Lagrange multipliers is a central question since, if we are able to compute a Lagrange multiplier $\lambda(u)$ of an optimal solution u, then u can be obtained as a solution of the unconstrained minimization problem

$$\min \left\{ f_0(v) + \sum_{i=1}^{n} \lambda_i f_i(v) \ : \ v \in V \right\}. \tag{$\mathcal{P}_\lambda$}$$

Let us first notice that the set of Lagrange multipliers does not depend on the solution u, i.e., if u_1 and u_2 are two solutions of the minimization problem $(\mathcal{P})$, then $M(u_1) = M(u_2)$, where $M(u_i)$ is the set of Lagrange multipliers of the solution u_i. Indeed, this is a consequence of the following characterization of Lagrange multipliers.

Proposition 9.6.3. *Let u be an optimal solution of the minimization problem*

$$\min \big\{ f_0(v) \ : \ f_i(v) \le 0, \qquad i = 1, \ldots, n \big\}. \tag{$\mathcal{P}$}$$

Then, the set of Lagrange multipliers for u is equal to

$$M = \left\{ \lambda \in \mathbf{R}^n_+ \ : \ \inf_C f_0 = \inf_V \left(f_0 + \sum_{i=1}^{n} \lambda_i f_i \right) \right\},$$

where $C = \big\{ v \in V \ : \ f_i(v) \le 0$ for all $i = 1, \ldots, n \big\}$ is the set of constraints.

Proof. Take a Lagrange multiplier λ for u. Then u is a solution of the unconstrained minimization problem

$$f_0(u) + \sum_{i=1}^{n} \lambda_i f_i(u) = \inf_V \left\{ f_0(v) + \sum_{i=1}^{n} \lambda_i f_i(v) \ : \ v \in V \right\}.$$

Because of the complementarity slackness property we deduce

$$f_0(u) = \inf_V \left\{ f_0 + \sum_{i=1}^{n} \lambda_i f_i \right\}.$$

On the other hand, since u is an optimal solution of $(\mathcal{P})$, we have

$$f_0(u) = \inf \big\{ f_0 + \delta_C \big\},$$

which proves that $\lambda \in M$.

Conversely, let us suppose that $\lambda \in M$. Then,

$$f_0(u) \leq f_0(u) + \sum_{i=1}^{n} \lambda_i f_i(u)$$

and $\sum_{i=1}^{n} \lambda_i f_i(u) \geq 0$. Since $\lambda_i \geq 0$ and $f_i(u) \leq 0$, this implies $\lambda_i f_i(u) = 0$ for all $i = 1, \ldots, n$. Hence

$$f_0(u) + \sum_{i=1}^{n} \lambda_i f_i(u) = \inf_V \left\{ f_0(v) + \sum_{i=1}^{n} \lambda_i f_i(v) \ : \ v \in V \right\},$$

which expresses that u is a solution of the unconstrained minimization problem

$$\min \left\{ f_0(v) + \sum_{i=1}^{n} \lambda_i f_i(v) : v \in V \right\}.$$

As a consequence,

$$\partial f_0(u) + \sum_{i=1}^{n} \lambda_i f_i(u) \ni 0,$$

which, together with $\lambda_i \geq 0$, $f_i(u) \leq 0$, and $\lambda_i f_i(u) = 0$, tells us that λ is a Lagrange multiplier vector for u. □

Clearly, the set M is independent of u solution of $(\mathcal{P})$. Thus we can speak of the set of Lagrange multipliers of a convex program. Indeed, the definition of the set M makes sense, and the set M may be nonempty, even when there is no solution of the convex program $(\mathcal{P})$. This leads us to give the following definition.

Definition 9.6.2. *For a given convex program*

$$\inf \left\{ f_0(v) \ : \ f_i(v) \leq 0, \qquad i = 1, \ldots, n \right\}, \tag{$\mathcal{P}$}$$

the set M of generalized Lagrange multiplier vectors is defined by

$$M = \left\{ \lambda \in \mathbf{R}^n_+ \ : \ \inf_V \left(f_0 + \delta_C \right) = \inf_V \left(f_0 + \sum_{i=1}^{n} \lambda_i f_i \right) \right\},$$

where $C = \left\{ v \in V \ : \ f_i(v) \leq 0, \qquad i = 1, \ldots, n \right\}$. When the problem $(\mathcal{P})$ has a solution, then M is the set of Lagrange multiplier vectors for $(\mathcal{P})$.

Without ambiguity, in what follows we will omit the word "generalized." We are going to characterize the set M by using marginal analysis.

Definition 9.6.3. *The value function attached to a convex program*

$$\inf \left\{ f_0(v) \ : \ f_i(v) \leq 0 \ \forall \ i = 1, \ldots, n \right\} \tag{$\mathcal{P}$}$$

is the function $p : \mathbf{R}^n \to \overline{\mathbf{R}}$ *which is defined, for every* $y = (y_1, y_2 \ldots, y_n) \in \mathbf{R}^n$, *by*

$$p(y) := \inf \big\{ f_0(v) \ : \ f_i(v) \leq y_i \ \forall\, i = 1, \ldots, n \big\}.$$

The function p *is also called the marginal function.*

Let us observe that the value function is the optimal value of the perturbed convex program $(\mathcal{P}_y)$

$$\inf \big\{ f_0(v) \ : \ f_i(v) \leq y_i \ \forall\, i = 1, \ldots, n \big\}. \tag{$\mathcal{P}_y$}$$

The initial problem, or unperturbed problem corresponds to the case $y = 0$, i.e., $(\mathcal{P}) = (\mathcal{P}_0)$. We also notice that the value function may take the value $-\infty$, which may be a source of difficulties.

We are going to show that Lagrange multiplier vectors for problem $(\mathcal{P})$ correspond to subgradients of the value function p.

Theorem 9.6.2. *Consider the convex minimization problem*

$$\inf \big\{ f_0(v) \ : \ f_i(v) \leq 0 \ \forall\, i = 1, \ldots, n \big\} \tag{$\mathcal{P}$}$$

and its value function $p : \mathbf{R}^n \to \overline{\mathbf{R}}$

$$p(y) = \inf \big\{ f_0(v) \ : \ f_i(v) \leq y_i \ \forall\, i = 1, \ldots, n \big\}.$$

Then, the following properties hold:

(a) *the value function* p *is convex;*

(b) *if* $p(0) \in \mathbf{R}$, *then* $M = -\partial p(0)$, *i.e., the set of generalized Lagrange multiplier vectors for* $(\mathcal{P})$ *is equal to the opposite of the subdifferential of* p *at the origin;*

(c) *if* $p(0) \in \mathbf{R}$ *and the Slater qualification assumption is satisfied, then* p *is continuous at the origin and, as a consequence,* M *is a nonempty, closed, convex, bounded set in* $\mathbf{R}^n_+$.

Proof. (a) We notice that

$$p(y) = \inf_{v \in V} f(v, y),$$

where $f(v, y) = f_0(v) + \delta_{C(y)}(v)$ and $C(y) = \big\{ v \in V \ : \ f_i(v) \leq y_i \text{ for all } i = 1, \ldots, n \big\}$. Let us verify that the mapping $(v, y) \mapsto \delta_{C(y)}(v)$ is convex.

We just need to verify that for every $(u, z) \in V \times \mathbf{R}^n$ and $(v, y) \in V \times \mathbf{R}^n$ such that $u \in C(z)$ and $v \in C(y)$, we still have $\lambda u + (1 - \lambda)v \in C(\lambda z + (1 - \lambda)y)$ for all $\lambda \in [0, 1]$.

Indeed, this is an immediate consequence of the convexity of functions f_i: we have

$$\begin{aligned} f_i(\lambda u + (1 - \lambda)v) &\leq \lambda f_i(u) + (1 - \lambda) f_i(v) \\ &\leq \lambda z_i + (1 - \lambda) y_i = (\lambda z + (1 - \lambda)y)_i. \end{aligned}$$

Since f_0 is convex, we obtain that f is convex with respect to the pair (v, y). The convexity of the value function p is then a consequence of Proposition 9.2.3.

(b) We first prove that every generalized Lagrange multiplier vector $\lambda \in \mathbf{R}^n_+$ satisfies $-\lambda \in \partial p(0)$. Equivalently, we need to prove that

$$\forall y \in \mathbf{R}^n \quad p(y) \geq p(0) - \sum_{i=1}^{n} \lambda_i y_i.$$

By definition of p and by Definition 9.6.2 of generalized Lagrange multiplier vectors, we have

$$\begin{aligned} p(0) &= \inf\left\{f_0 + \delta_C\right\} \\ &= \inf\left\{f_0 + \sum_{i=1}^{n} \lambda_i f_i\right\}. \end{aligned}$$

Take an arbitrary $y \in \mathbf{R}^n$ and denote by $C(y)$ the set

$$C(y) = \left\{v \in V \ : \ f_i(v) \leq y_i, \qquad i = 1, \ldots, n\right\}.$$

For every $v \in C(y)$ we have $\sum_{i=1}^{n} \lambda_i f_i(v) \leq \sum_{i=1}^{n} \lambda_i y_i$ (recall that $\lambda_i \geq 0$ for all $i = 1, \ldots, n$). Hence, for all $v \in C(y)$,

$$p(0) \leq f_0(v) + \sum_{i=1}^{n} \lambda_i y_i.$$

As a consequence, by taking the infimum with respect to $v \in C(y)$, we obtain

$$p(0) \leq p(y) + \sum_{i=1}^{n} \lambda_i y_i.$$

Let us now prove that, conversely, if $-\lambda \in \partial p(0)$, then λ is a generalized Lagrange multiplier vector for $(\mathcal{P})$.

We first prove that $\lambda \in \mathbf{R}^n_+$. Indeed, for every $y \in \mathbf{R}^n_+$ we have $C \subset C(y)$, and as a consequence

$$p(y) \leq p(0).$$

Combining this inequality and the subdifferential inequality

$$p(y) \geq p(0) - \sum_{i=1}^{n} \lambda_i y_i,$$

we obtain

$$\sum_{i=1}^{n} \lambda_i y_i \geq 0.$$

This being true for all $y \in \mathbf{R}^n_+$, we obtain that $\lambda \in \mathbf{R}^n_+$.

Let us now prove that

$$\inf_V \left(f_0 + \delta_C\right) = \inf_V \left(f_0 + \sum_{i=1}^{n} \lambda_i f_i\right).$$

Equivalently, we need to prove that

$$p(0) = \inf_V \left(f_0 + \sum_{i=1}^{n} \lambda_i f_i\right).$$

The inequality $p(0) \geq \inf_V \left(f_0 + \sum_{i=1}^n \lambda_i f_i\right)$ is always true for arbitrary $\lambda \in \mathbf{R}^n_+$: indeed, for every $v \in C$, we have $f_i(v) \leq 0$ and hence $\lambda_i f_i(v) \leq 0$. This immediately yields

$$\begin{aligned}\inf_V \left(f_0 + \sum_{i=1}^n \lambda_i f_i\right) &\leq \inf_C \left(f_0 + \sum_{i=1}^n \lambda_i f_i\right) \\ &\leq \inf_C f_0 = p(0).\end{aligned}$$

The opposite inequality $p(0) \leq \inf \left(f_0 + \sum_{i=1}^n \lambda_i f_i\right)$ relies on the fact that $-\lambda \in \partial p(0)$. We thus have for each $y \in \mathbf{R}^n$

$$p(y) + \sum_{i=1}^n \lambda_i y_i \geq p(0).$$

Take an arbitrary $v \in V$ and choose correspondingly $y_i = f_i(v)$ for all $i = 1, \ldots, n$. Thus, we have $v \in C(y)$ and $p(y) \leq f_0(v)$. As a consequence,

$$f_0(v) + \sum_{i=1}^n \lambda_i f_i(v) \geq p(0).$$

This being true for all $v \in V$, by taking the infimum with respect to v, we obtain

$$\inf_V \left(f_0 + \sum_{i=1}^n \lambda_i f_i\right) \geq p(0).$$

Finally, we have proved that $\lambda \in \mathbf{R}^n_+$ and

$$\inf_V (f_0 + \delta_C) = \inf_V \left(f_0 + \sum_{i=1}^n \lambda_i f_i\right).$$

By Definition 9.6.2, λ is a generalized Lagrange multiplier vector.

(c) By the Slater qualification assumption, there exists some $v_0 \in \operatorname{dom} f_0$ such that $f_i(v_0) < 0$ for all $i = 1, \ldots, n$. Thus, we can find a neighborhood of the origin in $\mathbf{R}^n$, say, $B(0, r)$ with $r > 0$, such that

$$\forall y \in B(0, r),\ \forall\, i = 1, \ldots, n \quad f_i(v_0) < y_i.$$

(It is enough to take, for example, $r = \frac{1}{2} \inf \left\{|f_i(v_0)| \ :\ i = 1, \ldots, n\right\}$.)

By definition of the value function p, we have

$$\forall y \in B(0, r) \quad p(y) \leq f_0(v_0).$$

Since $f_0(v_0) < +\infty$, p is bounded from above on the ball $B(0, r)$. Let us prove that this property, together with $p(0) \in \mathbf{R}$, implies

$$\forall y \in \mathbf{R}^n \quad p(y) > -\infty.$$

We first formulate the properties above in terms of epigraphs. We have

$$B(0,r) \times [f_0(v_0), +\infty[\subset \text{epi } p.$$

If $p(y) = -\infty$ for some $y \in \mathbf{R}^n$, we would have

$$\{y\} \times \mathbf{R} \subset \text{epi } p.$$

Take $\xi = -\alpha y$ with $\alpha > 0$ to have $|\xi| < r$, for example, $\alpha = r/(2|y|)$. We can write $\alpha = (1-\lambda)/\lambda$ for some $0 < \lambda < 1$, which gives $\lambda\xi + (1-\lambda)y = 0$. Then we observe that

$$\begin{cases} (\xi, f_0(v_0)) \in B(0,r) \times [f_0(v_0), +\infty[, \\ (y,t) \in \{y\} \times \mathbf{R} \text{ for every } t \in \mathbf{R}. \end{cases}$$

By the convexity of epi p we obtain

$$\big(\lambda\xi + (1-\lambda)y, \lambda f_0(v_0) + (1-\lambda)t\big) \in \text{epi } p,$$

i.e.,

$$p(0) \le \lambda f_0(v_0) + (1-\lambda)t \text{ for every } t \in \mathbf{R}.$$

Since $0 < \lambda < 1$, this implies $p(0) = -\infty$, a contradiction.

We can now apply Theorem 9.2.2: the function $p : \mathbf{R}^n \to \mathbf{R} \cup \{+\infty\}$ is convex and majorized on a neighborhood of the origin. Hence, p is continuous at the origin. By Proposition 9.5.2, $\partial p(0) \neq \emptyset$, and the set $M = -\partial p(0)$ is a nonempty closed convex bounded set in $\mathbf{R}^n_+$. □

Let us now introduce a dual minimization problem $(\mathcal{P}^*)$ to the convex program $(\mathcal{P})$ and show that the generalized Lagrange multiplier vectors are the solutions of this dual problem $(\mathcal{P}^*)$.

Theorem 9.6.3 (dual convex program). *Let us consider a convex program*

$$\inf \big\{ f_0(v) \ : \ f_i(v) \le 0 \ \forall\, i = 1, \dots, n \big\} \tag{$\mathcal{P}$}$$

and let $p : \mathbf{R}^n \to \overline{\mathbf{R}}$ be its value function. We assume that $p(0) \in \mathbf{R}$ and that the Slater qualification assumption holds. Then

(a) *The generalized Lagrange multiplier vectors of $(\mathcal{P})$ are the solutions of the maximization problem*

$$\sup \big\{ -p^*(-\lambda) \ : \ \lambda \in \mathbf{R}^n_+ \big\}, \tag{$\mathcal{P}^*$}$$

which is called the dual problem of $(\mathcal{P})$. The set of solutions of $(\mathcal{P}^*)$ is a nonempty closed convex bounded subset of $\mathbf{R}^n_+$.

(b) *For every $\lambda \in \mathbf{R}^n_+$, the equality*

$$-p^*(-\lambda) = \inf_{v \in V} \left\{ f_0(v) + \sum_{i=1}^n \lambda_i f_i(v) \right\}$$

holds and, as a consequence, the dual problem $(\mathcal{P}^*)$ can be written in the form

$$\sup_{\lambda \in \mathbf{R}^n_+} \inf_{v \in V} \left\{ f_0(v) + \sum_{i=1}^n \lambda_i f_i(v) \right\}. \tag{$\mathcal{P}^*$}$$

PROOF. By Theorem 9.6.2, we have the following equivalence:

$$\lambda \in M \iff -\lambda \in \partial p(0).$$

We know that p is a convex function. Thus, by using the Fenchel extremality relation (Proposition 9.5.1), we obtain

$$\lambda \in M \iff p(0) + p^*(-\lambda) = 0.$$

Theorem 9.6.2 also tells us that p is continuous at the origin. By Proposition 9.3.2, we thus have $p(0) = p^{**}(0)$. Noticing that $p^{**}(0) = \sup\{-p^*(\mu) \ : \ \mu \in \mathbf{R}^n\}$, we obtain

$$\begin{aligned} \lambda \in M \iff -p^*(-\lambda) &= \sup_{\mu \in \mathbf{R}^n} -p^*(\mu) \\ &= \sup_{\mu \in \mathbf{R}^n} -p^*(-\mu). \end{aligned}$$

Thus, $\lambda \in M$ iff λ is a solution of the maximization problem

$$\sup_{\mu \in \mathbf{R}^n} -p^*(-\mu).$$

Let us now compute p^*:

$$\begin{aligned} p^*(\mu) &= \sup_y \big\{\langle \mu, y\rangle - p(y)\big\} \\ &= \sup_y \Big\{\langle \mu, y\rangle - \inf\big\{f_0(v) \ : \ f_i(v) \le y_i \ \forall\, i = 1, \dots, n\big\}\Big\} \\ &= \sup_y \big\{\langle \mu, y\rangle - f_0(v) \ : \ y \in \mathbf{R}^n, \ f_i(v) \le y_i \ \forall\, i = 1, \dots, n\big\}. \end{aligned}$$

If for some $i \in \{1, \dots, n\}$ we have $\mu_i > 0$ then $p^*(\mu) = +\infty$. Otherwise, when $-\mu \in \mathbf{R}^n_+$ we have

$$\begin{aligned} p^*(\mu) &= \sup_{v \in V} \left\{ -f_0(v) + \sup_{y_i \ge f_i(v)} \sum_{i=1}^n \mu_i y_i \right\} \\ &= \sup_{v \in V} \left\{ \sum_{i=1}^n \mu_i f_i(v) - f_0(v) \right\}. \end{aligned}$$

Hence

$$-p^*(-\mu) = \begin{cases} \inf\limits_{v \in V} \left\{ f_0(v) + \sum\limits_{i=1}^n \mu_i f_i(v) \right\} & \text{if } \mu \in \mathbf{R}^n_+, \\ -\infty \qquad \text{otherwise,} \end{cases}$$

and the dual problem can be written

$$\sup_{\mu \in \mathbf{R}^n_+} \inf_{v \in V} \left\{ f_0(v) + \sum_{i=1}^n \mu_i f_i(v) \right\}. \quad \square$$

9.6.3 The Lagrangian approach to duality

In the framework of convex problems, and thanks to the Legendre–Fenchel transform, we have seen that a large number of mathematical objects can be paired with a dual one. Indeed we are going to go further and see how to realize the duality of optimization problems themselves. In the previous section, we introduced a variational problem $(\mathcal{P}^*)$, called the dual problem of $(\mathcal{P})$. We are going to justify this terminology and explore how the primal problem $(\mathcal{P})$ and its dual $(\mathcal{P}^*)$ are related to each other in remarkable ways.

Let us introduce some basic notations and concepts. The convex program

$$\inf \left\{ f_0(v) \ : \ f_i(v) \leq 0 \ \forall \ i = 1, \ldots, n \right\} \tag{$\mathcal{P}$}$$

is called the *primal problem*.

The key notion in the duality theory for optimization problems is the *Lagrangian*.

Definition 9.6.4. *The Lagrangian function attached to the convex program* $(\mathcal{P})$ *is the function* $\mathbf{L} : V \times \mathbf{R}^n_+ \to \mathbf{R} \cup \{+\infty\}$ *defined by*

$$\mathbf{L}(v, \lambda) = f_0(v) + \sum_{i=1}^{n} \lambda_i f_i(v).$$

We already noticed that this expression plays a central role in the theory of Lagrange multipliers. The new aspect in the definition above is to consider this expression as a bivariate function, i.e., $\mathbf{L}$ is a function of the two variables v and λ. This is a big step, since we are no longer concerned only with the primal problem, its solutions, and the characterization of its solutions: we choose to give from the very beginning an equivalent status to the variables v and λ.

Let us first observe that the Lagrangian function $\mathbf{L}$ encapsulates all the information of the primal problem $(\mathcal{P})$. Clearly

$$\sup_{\lambda \in \mathbf{R}^n_+} \mathbf{L}(v, \lambda) = f_0(v) + \delta_C(v),$$

where $C = \left\{ v \in V \ : \ f_i(v) \leq 0 \text{ for all } i = 1, \ldots, n \right\}$ is the set of constraints. Thus, the primal problem can be equivalently written as an inf-sup problem, namely,

$$\inf_{v \in V} \sup_{\lambda \in \mathbf{R}^n_+} \mathbf{L}(v, \lambda). \tag{$\mathcal{P}$}$$

Let us denote by $\alpha \in \overline{\mathbf{R}}$ the optimal value of $(\mathcal{P})$

$$\alpha := \inf_{v \in V} \sup_{\lambda \in \mathbf{R}^n_+} \mathbf{L}(v, \lambda);$$

α is called the *primal value*.

This formulation of $(\mathcal{P})$ makes it rather natural to consider the associated variational problem

$$\sup_{\lambda \in \mathbf{R}^n_+} \inf_{v \in V} \mathbf{L}(v, \lambda), \tag{$\mathcal{P}^*$}$$

called the *dual problem*, which is obtained by interchanging the order of the sup and inf operators.

Indeed, this formulation fits perfectly with the conclusion of Theorem 9.6.3 where it is shown that, under some assumptions, Lagrange multipliers are solutions of the dual problem $(\mathcal{P}^*)$. Let us denote by $\beta \in \overline{\mathbf{R}}$ the optimal value of $(\mathcal{P}^*)$

$$\beta := \sup_{\lambda \in \mathbf{R}^n_+} \inf_{v \in V} \mathbf{L}(v, \lambda);$$

β is called the *dual value*.

The dual problem therefore consists in maximizing over vectors $\lambda \in \mathbf{R}^n_+$ the *dual function*

$$d(\lambda) := \inf_{v \in V} \mathbf{L}(v, \lambda).$$

Note that the dual problem $(\mathcal{P}^*)$ is well defined without any assumptions on the functions $f_i,\ i = 1, \ldots, n$.

One has always $\beta \leq \alpha$, because $\sup_X \inf_Y \leq \inf_Y \sup_X$ is always true. It can happen that the primal value α is strictly larger than the dual value β. In this case, we say that there is a *duality gap*.

A basic question is to find conditions ensuring that there is no duality gap. When this is the case, the primal and the dual problem are connected through a rich calculus involving value functions, Legendre–Fenchel transform and subdifferentials, minimax and saddle value problems.

The notion of saddle value and saddle point of the Lagrangian function is also fundamental: it permits us to treat in a unifying way the primal and the dual aspects of optimization problems and will allow us to develop all these ideas in a far more general setting in the next section.

Definition 9.6.5. *Let $L : X \times Y \to \overline{\mathbf{R}}$ be a bivariate function where X and Y are arbitrary spaces. A point $(\overline{x}, \overline{y}) \in X \times Y$ is called a saddle point of L if*

$$\max_{y \in Y} L(\overline{x}, y) = L(\overline{x}, \overline{y}) = \min_{x \in X} L(x, \overline{y}).$$

Equivalently, $(\overline{x}, \overline{y})$ is a saddle point of L if

$$L(\overline{x}, y) \leq L(\overline{x}, \overline{y}) \leq L(x, \overline{y}) \qquad \forall\, x \in X,\ y \in Y.$$

Another way to say this is

$$\begin{cases} \text{(a) } \overline{x} \text{ is a solution of the minimization problem } \inf_{x \in X} L(x, \overline{y}), \\ \text{(b) } \overline{y} \text{ is a solution of the maximization problem } \max_{y \in Y} L(\overline{x}, y). \end{cases}$$

Note that the existence of a saddle point $(\overline{x}, \overline{y})$ implies that there is no duality gap. This follows from the equalities

$$\sup_{y \in Y} L(\overline{x}, y) = L(\overline{x}, \overline{y}) = \inf_{x \in X} L(x, \overline{y}),$$

which imply

$$\alpha = \inf_x \sup_y L(x, y) \leq L(\overline{x}, \overline{y}) \leq \sup_y \inf_x L(x, y) = \beta.$$

Since $\alpha \geq \beta$ is always true we obtain

$$L(\overline{x}, \overline{y}) = \inf_x \sup_y L(x, y) = \sup_y \inf_x L(x, y).$$

The converse is not true in general: it is possible to have no duality gap without the existence of saddle points.

We can now reformulate the conclusions of Theorem 9.6.3 in the following form.

Theorem 9.6.4. *Consider a convex program $(\mathcal{P})$ and assume that the Slater condition holds. Then the following facts hold true.*

(a) *There is no duality gap, i.e., the primal and the dual values are equal; let us call it the optimal value.*

(b) *(dual attainment) Assuming moreover that the optimal value is finite, then the set of solutions of the dual problem $(\mathcal{P}^*)$ is nonempty: it is the set of generalized Lagrange multipliers of problem $(\mathcal{P})$, and it is convex and bounded.*

(c) *(saddle point formulation of primal solutions) The following assertions are equivalent:*

(i) *u is a solution of the primal problem $(\mathcal{P})$;*

(ii) *there exists a vector $\lambda \in \mathbf{R}^n_+$ such that (u, λ) is a saddle point of the Lagrangian function $\mathbf{L}$.*

If (ii) *is satisfied, then λ is a Lagrange multiplier of the optimal solution u, and it is a solution of the dual problem $(\mathcal{P}^*)$.*

Proof. (a) If $\alpha = -\infty$, there is nothing to prove, since we know that $\alpha \geq \beta$. Otherwise, if α is finite, we are in the situation which was studied in Theorem 9.6.2: the Slater condition implies $\partial p(0) \neq \emptyset$ and the set M of generalized Lagrange multiplier vectors is nonempty. For every $\overline{\lambda} \in M$ we have

$$\begin{aligned} \alpha &= \inf_v \sup_\lambda \mathbf{L}(v, \lambda) \\ &= \inf_v \mathbf{L}(v, \overline{\lambda}) \\ &\leq \sup_\lambda \inf_v \mathbf{L}(v, \lambda) = \beta. \end{aligned}$$

Since $\alpha \geq \beta$ is always true, we obtain $\alpha = \beta$.

(b) It is just a reformulation of Theorem 9.6.3: the set of solutions of the dual problem $(\mathcal{P}^*)$ has been characterized with the help of the value function:

$$\begin{aligned} \lambda \text{ solution of } (\mathcal{P}^*) &\iff -\lambda \in \partial p(0) \\ &\iff \lambda \text{ generalized Lagrange multiplier of } (\mathcal{P}). \end{aligned}$$

(c) Let us first prove the implication (i) $\Longrightarrow$ (ii).

Let u be a solution of problem $(\mathcal{P})$. Then, the Slater condition implies the existence of a Lagrange multiplier $\overline{\lambda}$ associated to u (Theorem 9.6.1 and Proposition 9.6.3). Therefore,

$$\mathbf{L}(u, \overline{\lambda}) = \min_{v \in V} \mathbf{L}(v, \overline{\lambda}).$$

On the other hand, because of the complementary slackness property ($\overline{\lambda}_i f_i(u) = 0$ for all $i = 1, \ldots, n$), we have for any $\lambda \in \mathbf{R}^n_+$

$$\begin{aligned} \mathbf{L}(u, \overline{\lambda}) &= f_0(u) + \sum_{i=1}^n \overline{\lambda}_i f_i(u) \\ &= f_0(u) \\ &\geq f_0(u) + \sum_{i=1}^n \lambda_i f_i(u) \end{aligned}$$

(use that $\lambda_i \geq 0$ and $f_i(u) \leq 0$). Hence

$$\mathbf{L}(u, \overline{\lambda}) \geq \sup_{\lambda \in \mathbf{R}^n_+} \mathbf{L}(u, \lambda).$$

Finally

$$\inf_{v \in V} \mathbf{L}(v, \overline{\lambda}) \geq \mathbf{L}(u, \overline{\lambda}) \geq \sup_{\lambda \in \mathbf{R}^n_+} \mathbf{L}(u, \lambda),$$

which expresses that $(u, \overline{\lambda})$ is a saddle point of $\mathbf{L}$.

Let us now prove the implication (ii) $\Longrightarrow$ (i). If $(u, \overline{\lambda})$ is a saddle point of $\mathbf{L}$ on $V \times \mathbf{R}^n_+$, we have

$$\begin{aligned} f_0(u) + \delta_C(u) = \sup_{\lambda \in \mathbf{R}^n_+} \mathbf{L}(u, \lambda) &\leq \mathbf{L}(u, \overline{\lambda}) \\ &\leq \inf_v \mathbf{L}(v, \overline{\lambda}) \leq \sup_{\lambda \in \mathbf{R}^n_+} \inf_v \mathbf{L}(v, \lambda) \\ &\leq \inf_v \sup_{\lambda \in \mathbf{R}^n_+} \mathbf{L}(v, \lambda) = \inf_v \big(f_0(v) + \delta_C(v)\big). \end{aligned}$$

Hence, u is a solution of the primal problem $(\mathcal{P})$ and

$$\inf_v \big(f_0 + \delta_C\big) = \inf_v \left(f_0 + \sum_{i=1}^n \overline{\lambda}_i f_i\right),$$

which expresses (see Proposition 9.6.3) that $\overline{\lambda}$ is a Lagrange multiplier and, hence, $\overline{\lambda}$ is a solution of the dual problem. □

9.6.4 Duality for linear programming

Take $V = \mathbf{R}^n$. Given vectors $a^1, a^2, \ldots, a^m$, c in $\mathbf{R}^n$ and a vector b in $\mathbf{R}^m$ consider the primal linear program

$$\inf \big\{ \langle c, x \rangle \ : \ \langle a^i, x \rangle - b_i \leq 0, \ i = 1, \ldots, m \big\}, \qquad (\mathcal{P})$$

where $\langle \cdot, \cdot \rangle$ is the usual Euclidean scalar product in $\mathbf{R}^n$. This is clearly a problem of linear programming, with

$$\begin{cases} f_0(x) = \langle c, x \rangle, \\ f_i(x) = \langle a^i, x \rangle - b_i \qquad i = 1, \dots, m. \end{cases}$$

Indeed, f_0 is linear and $C = \{x \in \mathbf{R}^n \ : \ \langle a^i, x \rangle - b_i \leq 0\}$ is a polyhedral set (finite intersection of closed half-spaces). The Lagrangian function $\mathbf{L} : \mathbf{R}^n \times \mathbf{R}^m_+ \to \mathbf{R}$ is given by

$$\mathbf{L}(x, \lambda) = \langle c, x \rangle + \sum_{i=1}^{m} \lambda_i \big(\langle a^i, x \rangle - b_i \big).$$

Therefore, the primal problem can be rewritten as

$$\inf_x \sup_{\lambda \in \mathbf{R}^m_+} \mathbf{L}(x, \lambda) \tag{$\mathcal{P}$}$$

and the dual problem $(\mathcal{P}^*)$ is given by

$$\sup_{\lambda \in \mathbf{R}^m_+} \inf_x \mathbf{L}(x, \lambda). \tag{$\mathcal{P}^*$}$$

Let us compute the dual function

$$\begin{aligned} d(\lambda) &= \inf_x \mathbf{L}(x, \lambda) \\ &= \inf_{x \in \mathbf{R}^n} \left(\left\langle x, c + \sum_{i=1}^{m} \lambda_i a^i \right\rangle - \sum_{i=1}^{m} \lambda_i b_i \right). \end{aligned}$$

We find

$$d(\lambda) = \begin{cases} -\sum_{i=1}^m \lambda_i b_i & \text{if } c + \sum_{i=1}^m \lambda_i a^i = 0, \\ -\infty & \text{otherwise.} \end{cases}$$

The dual problem $(\mathcal{P}^*)$ is then given by

$$\begin{cases} \sup -\langle b, \lambda \rangle \\ \text{subject to} \quad \begin{cases} \sum_{i=1}^m \lambda_i a^i = -c, \\ \lambda \in \mathbf{R}^m_+ , \end{cases} \end{cases} \tag{$\mathcal{P}^*$}$$

and the Kuhn–Tucker optimality conditions are the following:

$$\begin{cases} \sum_{i=1}^m \lambda_i a^i = -c, \\ \lambda_i \geq 0, \ x \in \mathbf{R}^n, \\ \langle a^i, x \rangle - b_i \leq 0, \\ \lambda_i \big(\langle a^i, x \rangle - b_i \big) = 0. \end{cases}$$

9.7 A general approach to duality in convex optimization

In Section 9.6, we developed a duality theory for convex programs. This is an important class of convex optimization problems, but it is far from covering the whole field of convex optimization. Thus a number of natural questions arise: Is it possible to develop a duality theory for general convex optimization, and if yes, is there a unique dual minimization problem? What are the relations between primal and dual problems, and what is the interpretation of the solutions of the dual problem?

At the center of all these questions is the notion of Lagrangian function which we now introduce. Let us consider the primal problem

$$\inf\big\{f(v) \ : \ v \in V\big\}, \tag{$\mathcal{P}$}$$

where $f : V \to \mathbf{R} \cup \{+\infty\}$ is a general convex, lower semicontinuous, and proper function whose definition usually includes the constraints.

The basic idea is to introduce a bivariate function $\mathbf{L} : V \times W \to \overline{\mathbf{R}}$ which satisfies

$$\forall v \in V \quad f(v) = \sup_{w \in W} \mathbf{L}(v, w).$$

We say that $\mathbf{L}$ is a Lagrangian function associated to problem $(\mathcal{P})$. In this way, the primal problem can be written as an inf $-$ sup problem:

$$\inf_{v \in V} \sup_{w \in W} \mathbf{L}(v, w). \tag{$\mathcal{P}$}$$

As we did in Section 9.6, one can associate to problem $(\mathcal{P})$ a dual problem $(\mathcal{P}^*)$ which is obtained by interchanging the order of inf and sup:

$$\sup_{w \in W} \inf_{v \in V} \mathbf{L}(v, w). \tag{$\mathcal{P}^*$}$$

Equivalently, $(\mathcal{P}^*)$ can be written as a maximization problem, in the form

$$\sup_{w \in W} d(w) \tag{$\mathcal{P}^*$}$$

with $d : W \to \overline{\mathbf{R}}$ being defined by

$$d(w) := \inf_{v \in V} \mathbf{L}(v, w).$$

We call $d(\cdot)$ the dual function (attached to the Lagrangian function $\mathbf{L}$). The interesting situation occurs when there is no duality gap, i.e.,

$$\inf(\mathcal{P}) = \sup(\mathcal{P}^*),$$

which is equivalent to saying

$$\inf_{v \in V} \sup_{w \in W} \mathbf{L}(v, w) = \sup_{w \in W} \inf_{v \in V} \mathbf{L}(v, w).$$

In this general abstract setting, we have the following result.

Proposition 9.7.1. *Let $\mathbf{L} : V \times W \to \overline{\mathbf{R}}$ be a general bivariate function. Then, the following facts are equivalent:*

(i) $(\overline{v}, \overline{w})$ *is a saddle point of* $\mathbf{L}$*;*

(ii) $\overline{v}$ *is a solution of the primal problem* $(\mathcal{P})$*,* $\overline{w}$ *is a solution of the dual problem* $(\mathcal{P}^*)$*, and there is no duality gap:* $\inf(\mathcal{P}) = \sup(\mathcal{P}^*)$.

PROOF. (i) $\Longrightarrow$ (ii). By definition of saddle point

$$\mathbf{L}(\overline{v}, w) \leq \mathbf{L}(\overline{v}, \overline{w}) \leq \mathbf{L}(v, \overline{w}) \qquad \forall v \in V,\ \forall w \in W.$$

Hence, for every $v \in V$

$$\sup_{w \in W} \mathbf{L}(\overline{v}, w) \leq \mathbf{L}(\overline{v}, \overline{w}) \leq \mathbf{L}(v, \overline{w}) \leq \sup_{w \in W} \mathbf{L}(v, w).$$

This being true for all $v \in V$, we obtain that $\overline{v}$ is a solution of the minimization problem

$$\inf_{v \in V} \left(\sup_{w \in W} \mathbf{L}(v, w) \right),$$

which is precisely the primal problem $(\mathcal{P})$.

In a similar way, for every $w \in W$

$$\inf_{v \in V} \mathbf{L}(v, w) \leq \mathbf{L}(\overline{v}, w) \leq \mathbf{L}(\overline{v}, \overline{w}) \leq \inf_{v \in V} \mathbf{L}(v, \overline{w}).$$

Hence, $\overline{w}$ is a solution of the maximization problem

$$\sup_{w \in W} \left(\inf_{v \in V} \mathbf{L}(v, w) \right),$$

which is the dual problem $(\mathcal{P}^*)$.

(ii) $\Longrightarrow$ (i). Since $\overline{v}$ is a solution of the primal problem $(\mathcal{P})$, we have

$$\sup_{w \in W} \mathbf{L}(\overline{v}, w) = \inf_{v \in V} \sup_{w \in W} \mathbf{L}(v, w).$$

Similarly, since $\overline{w}$ is a solution of the dual problem $(\mathcal{P}^*)$, we have

$$\inf_{v \in V} \mathbf{L}(v, \overline{w}) = \sup_{w \in W} \inf_{v \in V} \mathbf{L}(v, w).$$

Since there is no duality gap, $\inf \sup = \sup \inf$ and we obtain

$$\sup_{w \in W} \mathbf{L}(\overline{v}, w) = \inf_{v \in V} \mathbf{L}(v, \overline{w}),$$

that is,

$$\forall v \in V,\ \forall w \in W \quad \mathbf{L}(\overline{v}, w) \leq \mathbf{L}(v, \overline{w}).$$

This clearly implies $\mathbf{L}(\overline{v}, \overline{w}) \leq \mathbf{L}(v, \overline{w})$ for all $v \in V$, and $\mathbf{L}(\overline{v}, w) \leq \mathbf{L}(\overline{v}, \overline{w})$ for all $w \in W$, i.e.,

$$\mathbf{L}(\overline{v}, \overline{w}) = \min_{v} \mathbf{L}(v, \overline{w}) = \max_{w} \mathbf{L}(\overline{v}, w),$$

which expresses that $(\overline{v}, \overline{w})$ is a saddle point of $\mathbf{L}$. □

As we described above, duality theory for minimization problems follows very naturally from the Lagrangian formulation: it just consists in the permutation of the inf and the sup.

We stress the fact that the primal and the dual problems are intimately paired as soon as there is no duality gap and there exist saddle points of $\mathbf{L}$.

Thus the question is, for which class of bivariate functions $\mathbf{L}$ can one expect to have such properties? This is a central question in game theory, fixed point theory, economics, and so forth. Let us quote the celebrated Von Neumann's minimax theorem. Indeed, we give a slightly more general formulation to recover as a particular case the existence theorem for convex minimization problems (see Aubin [45], for example).

Theorem 9.7.1 (Von Neumann's minimax theorem). *Let V and W be two reflexive Banach spaces and let $M \subset V$ and $N \subset W$ be two closed convex nonempty sets. Let $\mathbf{L} : M \times N \to \mathbf{R}$ be a bivariate function which satisfies the following properties:*

$$\begin{cases} (\mathrm{i_a})\ \forall w \in N,\ v \mapsto \mathbf{L}(v, w) \text{ is convex and lower semicontinuous,} \\ (\mathrm{i_b})\ \forall v \in M\ w \mapsto \mathbf{L}(v, w) \text{ is concave and upper semicontinuous.} \end{cases}$$

$$\begin{cases} (\mathrm{ii_a})\ M \text{ is bounded or there exists some } w_0 \in N \text{ such that } v \mapsto \mathbf{L}(v, w_0) \text{ is coercive,} \\ (\mathrm{ii_b})\ N \text{ is bounded or there exists some } v_0 \in M \text{ such that } w \mapsto -\mathbf{L}(v_0, w) \text{ is coercive.} \end{cases}$$

Then $\mathbf{L}$ possesses a saddle point $(\overline{v}, \overline{w}) \in M \times N$, i.e.,

$$\min_{v \in M} \mathbf{L}(v, \overline{w}) = \mathbf{L}(\overline{v}, \overline{w}) = \max_{w \in N} \mathbf{L}(\overline{v}, w).$$

In particular $\inf_v \sup_w \mathbf{L}(v, w) = \sup_w \inf_v \mathbf{L}(v, w)$, i.e., there is no duality gap.

It follows from the previous results that the key property to developing a duality theory for optimization problems is the possibility to write the function f in the following form:

$$f(v) = \sup_{w \in W} \mathbf{L}(v, w)$$

with $\mathbf{L} : V \times W \to \overline{\mathbf{R}}$ a convex-concave bivariate function.

We are going to see how the Legendre–Fenchel transform permits us to produce such convex-concave Lagrangian functions in a systematic and elegant way.

The idea is first to introduce a perturbation function $F : V \times Y \to \mathbf{R} \cup \{+\infty\}$ such that $F(v, 0) = f(v)$ for all $v \in V$. The primal problem $(\mathcal{P})$ can now be written as

$$\inf \big\{ F(v, 0) \ :\ v \in V \big\}. \tag{$\mathcal{P}$}$$

The key property which allows us to produce a convex-concave Lagrangian function from F is that F is convex with respect to $(v, w) \in V \times W$. For example, in convex programming, the duality scheme that was studied in Section 9.6 is associated with the perturbation function:

$$F(v, y) = \begin{cases} f_0(v) & \text{if } f_i(v) \le y_i, \quad i = 1, \dots, n, \\ +\infty & \text{otherwise.} \end{cases}$$

One can easily verify that when f_0 and f_i $(i = 1, \dots, n)$ are convex functions, so is F.

Let us now describe how one can associate a Lagrangian function to F.

Proposition 9.7.2 (definition of Lagrangian). *Let $F : V \times Y \to \mathbf{R} \cup \{+\infty\}$ be a convex function. We associate to F a Lagrangian function $\mathbf{L} : V \times Y^* \to \overline{\mathbf{R}}$ by the following formula:*

$$\forall v \in V,\ \forall y^* \in Y^* \quad -\mathbf{L}(v, y^*) = \sup_{y \in Y} \big\{ \langle y^*, y \rangle - F(v, y) \big\},$$

i.e., $-\mathbf{L}(v, \cdot)$ is the Legendre–Fenchel conjugate of $F(v, \cdot)$. We have that $\mathbf{L}$ is a convex-concave function. More precisely,

$$\begin{cases} (1) \text{ for all } v \in V,\ y^* \mapsto \mathbf{L}(v, y^*) \text{ is concave, upper semicontinuous on } Y^*; \\ (2) \text{ for all } y^* \in Y^*,\ v \mapsto \mathbf{L}(v, y^*) \text{ is convex.} \end{cases}$$

Proof. The proof is immediate; just note that (2) follows from Proposition 9.2.3 and from the fact that the function $(v, y) \mapsto -\langle y^*, y \rangle + F(v, y)$ is convex. □

As expected, we can reformulate problem $(\mathcal{P})$ by using the Lagrangian function $\mathbf{L}$ attached to the (convex) perturbation function F.

Proposition 9.7.3. *Let $F : V \times Y \to \mathbf{R} \cup \{+\infty\}$ be a convex, lower semicontinuous, proper function and let $\mathbf{L} : V \times Y^* \to \overline{\mathbf{R}}$ be the corresponding Lagrangian function, given by*

$$-\mathbf{L}(v, y^*) = \sup_{y \in Y} \big\{ \langle y^*, y \rangle - F(v, y) \big\}.$$

(a) *Primal problem: we have*

$$\forall v \in V \quad F(v, 0) = \sup_{y^* \in Y^*} \mathbf{L}(v, y^*).$$

As a consequence, with the notation $f(v) = F(v, 0)$, the primal problem

$$\inf \big\{ f(v) \ :\ v \in V \big\} \tag{$\mathcal{P}$}$$

can be written as

$$\inf_{v \in V} \sup_{y^* \in Y^*} \mathbf{L}(v, y^*). \tag{$\mathcal{P}$}$$

(b) *Dual problem: the dual problem, which, by definition, is*

$$\sup_{y^* \in Y^*} \inf_{v \in V} \mathbf{L}(v, y^*) \tag{$\mathcal{P}^*$}$$

can be written as

$$\sup_{y^* \in Y^*} d(y^*) \tag{$\mathcal{P}^*$}$$

where the dual function $d : Y^* \to \mathbf{R} \cup \{+\infty\}$ is given by

$$\begin{aligned} d(y^*) &= \inf_{v \in V} \mathbf{L}(v, y^*) \\ &= -F^*(0, y^*), \end{aligned}$$

and F^* is the Legendre–Fenchel conjugate of F with respect to (v, y).

PROOF. (a) Since F is closed convex and proper on $V \times Y$, for every $v \in V$ the function $\varphi_v : y \mapsto F(v, y)$ is closed convex on Y. Hence, for all $v \in V$

$$\begin{aligned}\varphi_v(y) &= \varphi_v^{**}(y) \\ &= \sup_{y^* \in Y^*} \big\{\langle y^*, y\rangle - \varphi_v^*(y^*)\big\}.\end{aligned}$$

By definition of $\mathbf{L}$ we have

$$\begin{aligned}\varphi_v^*(y^*) &= \sup_{y \in Y} \big\{\langle y^*, y\rangle - \varphi_v(y)\big\} \\ &= \sup_{y \in Y} \big\{\langle y^*, y\rangle - F(v, y)\big\} \\ &= -\mathbf{L}(v, y^*).\end{aligned}$$

Hence,

$$\varphi_v(y) = \sup_{y^* \in Y^*} \big\{\langle y^*, y\rangle + \mathbf{L}(v, y^*)\big\}.$$

Take now $y = 0$ to obtain

$$f(v) = F(v, 0) = \varphi_v(0) = \sup_{y^* \in Y^*} \mathbf{L}(v, y^*).$$

(b) By definition, the dual function $d : Y^* \to \mathbf{R} \cup \{+\infty\}$ is equal to

$$d(y^*) = \inf_{v \in V} \mathbf{L}(v, y^*).$$

By definition of $\mathbf{L}$,

$$\begin{aligned}\mathbf{L}(v, y^*) &= -\sup_{y} \big\{\langle y^*, y\rangle - F(v, y)\big\} \\ &= \inf_{y} \big\{-\langle y^*, y\rangle + F(v, y)\big\}.\end{aligned}$$

Hence,

$$d(y^*) = \inf_{v \in V, y \in Y} \big\{-\langle y^*, y\rangle + F(v, y)\big\}.$$

Thus,

$$\begin{aligned}d(y^*) &= -\sup_{v \in V, y \in Y} \big\{\langle 0, v\rangle + \langle y^*, y\rangle - F(v, y)\big\} \\ &= -F^*(0, y^*),\end{aligned}$$

where F^* is the conjugate of F with respect to (v, y). □

The other fundamental mathematical object which is attached to the perturbation function F is the value function.

Proposition 9.7.4 (definition of the value function). *Let $F : V \times Y \to \mathbf{R} \cup \{+\infty\}$ be a convex function. The value function (also called marginal function) attached to F is the function $p : Y \to \mathbf{R} \cup \{+\infty\}$, which is defined by*

$$\forall y \in Y \quad p(y) := \inf_{v \in V} F(v, y).$$

It is a convex function. Moreover, for every $y^* \in Y^*$

$$p^*(y^*) = F^*(0, y^*).$$

Thus, the dual problem $(\mathcal{P}^*)$ can be formulated in terms of p as follows:

$$\sup_{y^* \in Y^*} (-p^*(y^*)) = - \inf_{y^* \in Y^*} p^*(y^*). \qquad (\mathcal{P}^*)$$

Proof. The convexity of p follows from the convexity of F and by applying Proposition 9.2.3. For $y^* \in Y^*$ we have

$$\begin{aligned} p^*(y^*) &= \sup_{y \in Y} \big\{ \langle y^*, y \rangle - p(y) \big\} \\ &= \sup_{y \in Y} \big\{ \langle y^*, y \rangle - \inf_{v \in V} F(v, y) \big\} \\ &= \sup_{(v,y) \in V \times Y} \big\{ \langle 0, v \rangle + \langle y^*, y \rangle - F(v, y) \big\} \\ &= F^*(0, y^*) \\ &= -d(y^*), \end{aligned}$$

where d is the dual function introduced in Proposition 9.7.3(b). □

We have now all the ingredients for developing a general convex duality theory. The following theorem may be proved in the same way as Theorems 9.6.2 and 9.6.3. Let us notice that the qualification assumption

"there exists some $v_0 \in V$ such that $F(v_0, .)$ is finite and continuous at the origin"

plays the role of the Slater qualification assumption in convex programming. For this reason, we call it (GS) (for generalized Slater).

Theorem 9.7.2. *Let $f : V \to \mathbf{R} \cup \{+\infty\}$ be a closed convex and proper function such that $\inf f > -\infty$. Let $F : V \times Y \to \mathbf{R} \cup \{+\infty\}$ be a perturbation function which satisfies the following conditions:*

(i) $F(v, 0) = f(v) \quad \forall v \in V$;

(ii) *F is a closed convex proper function;*

(iii) *(generalized Slater): there exists some $v_0 \in V$ such that $y \mapsto F(v_0, y)$ is finite and continuous at the origin.*

Then, the following properties hold:

(a) *The value function p is continuous at the origin. As a consequence, the set of solutions of the dual problem $(\mathcal{P}^*)$, which is equal to $\partial p(0)$, is nonempty. Indeed, it is a nonempty closed convex and bounded subset of Y^*.*

(b) *There is no duality gap, i.e.,* $\inf(\mathcal{P}) = \max(\mathcal{P}^*)$.

(c) *Let $\overline{u}$ be a solution of the primal problem $(\mathcal{P})$. Then, for every element $\overline{y}^*$ which is a solution of the dual problem (by property* (a) *there exist such elements), $(\overline{u}, \overline{y}^*)$ is a saddle point of the Lagrangian function* $\mathbf{L}$ *associated to F. Conversely, if $(\overline{u}, \overline{y}^*)$ is a saddle point of* $\mathbf{L}$*, then $\overline{u}$ is a solution of $(\mathcal{P})$ and $\overline{y}$ is a solution of $(\mathcal{P}^*)$.*

(d) *$(\overline{u}, \overline{y}^*)$ is a saddle point of* $\mathbf{L}$ *iff it satisfies the extremality relation:*

$$F(\overline{u}, 0) + F^*(0, \overline{y}^*) = 0.$$

PROOF. (a) The generalized Slater condition implies the existence of some neighborhood of 0 in Y in which the function $y \mapsto F(v_0, y)$ is bounded from above: let $r > 0$ and $M \in \mathbf{R}$ be such that

$$F(v_0, y) \leq M \quad \forall\, y \in B_Y(0, r).$$

As a consequence, the value function p satisfies

$$p(y) = \inf_{v \in V} F(v, y) \leq F(v_0, y) \leq M \quad \forall\, y \in B_Y(0, r).$$

The function p also satisfies $p(0) = \inf(\mathcal{P})$ and p is bounded fom above on a neighborhood of the origin. By Theorem 9.3.2 and Proposition 9.5.2 we obtain that p is subdifferentiable at 0, i.e., $\partial p(0) \neq \emptyset$. We know by Proposition 9.7.4 that the dual problem $(\mathcal{P}^*)$ can be expressed in terms of the value function p. Indeed, the fact that $\overline{y}^*$ is a solution of $(\mathcal{P}^*)$ is equivalent to saying that $\overline{y}^*$ is a solution of the minimization problem

$$\inf_{y^* \in Y^*} p^*(y^*).$$

Thus,

$$\begin{aligned} \overline{y}^* \text{ is a solution of } (\mathcal{P}^*) &\iff \partial p^*(y^*) \ni 0 \\ &\iff y^* \in \partial p(0) \quad \text{(Theorem 9.5.1).} \end{aligned}$$

By the argument above, $\partial p(0)$ is a closed convex bounded nonempty subset of Y^*. Thus, there exist solutions of the dual problem $(\mathcal{P}^*)$ and the set of these solutions is a closed convex bounded subset of Y^*.

(b) Let $\overline{y}^*$ be any solution of the dual problem $(\mathcal{P}^*)$. We know by (a) that there exist such elements and that they are characterized by the relation $\overline{y}^* \in \partial p(0)$ or, equivalently, by the Fenchel extremality relation

$$p(0) + p^*(\overline{y}^*) = \langle \overline{y}^*, 0 \rangle = 0.$$

Hence

$$\begin{aligned} \inf(\mathcal{P}) = p(0) &= -p^*(\overline{y}^*) \\ &= \sup_{y^* \in Y^*} -p^*(y^*) = \sup(\mathcal{P}^*). \end{aligned}$$

(c) We use the Lagrangian formulation of problem $(\mathcal{P})$ and $(\mathcal{P}^*)$ given by Proposition 9.7.3:

$$\begin{cases} (\mathcal{P}) & \inf_{v\in V} \sup_{y^*\in Y^*} \mathbf{L}(v, y^*), \\ (\mathcal{P}^*) & \sup_{y^*\in Y^*} \inf_{v\in V} \mathbf{L}(v, y^*). \end{cases}$$

The characterization of pairs of optimal solutions $(\overline{u}, \overline{y}^*)$ of problems $(\mathcal{P})$ and $(\mathcal{P}^*)$, respectively, as saddle points of the Lagrangian $\mathbf{L}$ is a direct consequence of Proposition 9.7.1.

(d) Let $(\overline{u}, \overline{y}^*)$ be a saddle point of $\mathbf{L}$. Let us reformulate the extremality relation

$$p(0) + p^*(\overline{y}^*) = 0 \quad \text{(see (b) above)}$$

in terms of the function F: we have

$$\begin{cases} p(0) = \inf(\mathcal{P}) = F(\overline{u}, 0), \\ p^*(\overline{y}^*) = \sup(\mathcal{P}^*) = F^*(0, \overline{y}^*) \quad \text{(Proposition 9.7.4).} \end{cases}$$

Hence

$$F(\overline{u}, 0) + F^*(0, \overline{y}^*) = 0,$$

that is, $(0, \overline{y}^*) \in \partial F(\overline{u}, 0)$. □

9.8 Duality in the calculus of variations: First examples

As a model example, let us consider the Dirichlet problem: given $h \in L^2(\Omega)$, find $u \in H_0^1(\Omega)$ such that

$$\begin{cases} -\Delta u = h & \text{in } \Omega, \\ u = 0 & \text{on } \partial\Omega. \end{cases}$$

The variational formulation of this problem has been extensively studied in chapter 5: the solution u of the Dirichlet problem is the minimizer, on the Sobolev space $H_0^1(\Omega)$, of the functional

$$f(v) = \frac{1}{2}\int_\Omega |\nabla v(x)|^2\, dx - \int_\Omega h(x)v(x)\, dx.$$

The primal problem $(\mathcal{P})$ can be expressed as

$$\min_{v\in H_0^1(\Omega)} \left\{ \frac{1}{2}\int_\Omega |\nabla v(x)|^2\, dx - \int_\Omega h(x)v(x)\, dx \right\}. \tag{$\mathcal{P}$}$$

We now introduce the perturbation function

$$F(v, y) = \frac{1}{2}\int_\Omega |\nabla v(x) + y(x)|^2\, dx - \int_\Omega h(x)v(x)\, dx,$$

which we consider as a function $F : H_0^1(\Omega) \times L^2(\Omega)^N \to \mathbf{R}$.

To compute the Lagrangian function $\mathbf{L}$ associated to F and describe the corresponding dual problem, we start to analyze the structure of this problem.

The primal function f can be written as

$$f(v) = \Phi(Av) + \Psi(v),$$

where $\Phi : L^2(\Omega)^N \to \mathbf{R}$ is the convex integral functional

$$\Phi(w) = \frac{1}{2}\int_\Omega |w(x)|^2\,dx$$

and A is the gradient operator, which can be viewed as a linear continuous operator from $H_0^1(\Omega)$ into $L^2(\Omega)^N$. The functional $\Psi : H_0^1(\Omega) \to \mathbf{R}$ is the linear and continuous mapping

$$\Psi(v) = -\int_\Omega h(x)v(x)\,dx.$$

The perturbation function $F : H_0^1(\Omega) \times L^2(\Omega)^N \to \mathbf{R}$ can then be written as

$$F(v, y) = \Phi(Av + y) + \Psi(v).$$

Theorem 9.8.1. *Let V and Y be two Banach spaces and suppose $\Phi \in \Gamma_0(Y)$, $\Psi \in \Gamma_0(V)$, and $A \in L(V, Y)$ (A is a linear continuous operator). Consider the primal problem*

$$\inf_{v\in V}\big\{\Phi(Av) + \Psi(v)\big\} \tag{$\mathcal{P}$}$$

and the perturbation function $F : V \times Y \to \mathbf{R} \cup \{+\infty\}$ defined by

$$F(v, y) = \Phi(Av + y) + \Psi(v).$$

Then the following facts hold.

(a) *The Lagrangian function $L : V \times Y^* \to \overline{\mathbf{R}}$ associated to F is given by*

$$\mathbf{L}(v, y^*) = \Psi(v) + \langle y^*, Av\rangle - \Phi^*(y^*)$$

and the dual problem ($\mathcal{P}^*$) is equal to

$$\sup_{y^*\in Y^*}\big\{-\Psi^*(-A^*y^*) - \Phi^*(y^*)\big\}, \tag{$\mathcal{P}^*$}$$

where A^* is the adjoint operator of A.

(b) *Let us assume that there exists some $v_0 \in V$ such that $\Psi(v_0) < +\infty$, $\Phi(A(v_0)) < +\infty$ and Φ is continuous at Av_0. Then ($\mathcal{P}^*$) has at least one solution $\overline{y}^*$. If $\overline{u}$ is a solution of ($\mathcal{P}$), one has the following extremality relations:*

$$\begin{cases} -A^*\overline{y}^* \in \partial\Psi(\overline{u}), \\ \overline{y}^* \in \partial\Phi(A\overline{u}). \end{cases}$$

Proof. (a) By definition of Lagrangian (Proposition 9.7.2)

$$\begin{aligned}
\mathbf{L}(v, y^*) &= \inf_{y\in Y}\big\{-\langle y^*, y\rangle + F(v, y)\big\} \\
&= \inf_{y\in Y}\big\{-\langle y^*, y\rangle + \Phi(Av + y) + \Psi(v)\big\} \\
&= \Psi(v) - \sup_{y\in Y}\big\{\langle y^*, y\rangle - \Phi(Av + y)\big\} \\
&= \Psi(v) - \sup_{y\in Y}\big\{\langle y^*, Av + y\rangle - \Phi(Av + y) - \langle y^*, Av\rangle\big\} \\
&= \Psi(v) + \langle y^*, Av\rangle - \Phi^*(y^*).
\end{aligned}$$

The perturbation function F is convex and lower semicontinuous: this is an immediate consequence of the facts that $\Phi \in \Gamma_0(Y)$, $\Psi \in \Gamma_0(V)$, and $A \in L(V, Y)$. Therefore, the primal and the dual problems can be expressed in terms of the Lagrangian function L and we have (Proposition 9.7.3)

$$\begin{cases} (\mathcal{P}) & \inf_{v\in V} \sup_{y^*\in Y^*} \mathbf{L}(v, y^*), \\ (\mathcal{P}^*) & \sup_{y^*\in Y^*} \inf_{v\in V} \mathbf{L}(v, y^*). \end{cases}$$

Let us compute the dual function d:

$$\begin{aligned} d(y^*) &= \inf_{v\in V} \mathbf{L}(v, y^*) \\ &= \inf_{v\in V} \left\{\Psi(v) + \langle y^*, Av\rangle - \Phi^*(y^*)\right\} \\ &= -\Phi^*(y^*) - \sup_{v\in V} \left\{\langle -y^*, Av\rangle - \Psi(v)\right\} \\ &= -\Phi^*(y^*) - \sup_{v\in V} \left\{\langle -A^*y^*, v\rangle - \Psi(v)\right\} \\ &= -\Phi^*(y^*) - \Psi^*(-A^*y^*). \end{aligned}$$

Therefore, from the analysis made in Section 9.7, part (a) is proved.

(b) The existence of $v_0 \in V$ such that $\Psi(v_0) < +\infty$ with Φ continuous at Av_0 clearly implies that the generalized Slater condition is satisfied. Hence $(\mathcal{P}^*)$ admits at least a solution (Theorem 9.7.2). Let $\overline{y}^*$ be such a solution. Let $\overline{u}$ be a solution of the primal problem $(\mathcal{P})$. We know that there is no duality gap. Hence $\inf(\mathcal{P}) = \sup(\mathcal{P}^*)$, i.e.,

$$\Phi(A\overline{u}) + \Psi(\overline{u}) = -\Phi(\overline{y}^*) - \Psi^*(-A^*\overline{y}^*).$$

Thus

$$\Phi(A\overline{u}) + \Phi^*(\overline{y}^*) + \Psi(\overline{u}) + \Psi^*(-A^*\overline{y}^*) = 0.$$

Equivalently,

$$\left(\Phi(A\overline{u}) + \Phi^*(\overline{y}^*) - \langle \overline{y}^*, A\overline{u}\rangle\right) + \left(\Psi(\overline{u}) + \Psi^*(-A^*\overline{y}^*) + \langle A^*\overline{y}^*, \overline{u}\rangle\right) = 0.$$

Since by the Fenchel inequality the quantities $\Phi(A\overline{u}) + \Phi^*(\overline{y}^*) - \langle \overline{y}^*, A\overline{u}\rangle$ and $\Psi(\overline{u}) + \Psi^*(-A^*\overline{y}^*) + \langle A^*\overline{y}^*, \overline{u}\rangle$ are nonnegative, we obtain

$$\begin{cases} \Phi(A\overline{u}) + \Phi^*(\overline{y}^*) - \langle \overline{y}^*, A\overline{u}\rangle = 0, \\ \Psi(\overline{u}) + \Psi^*(-A^*\overline{y}^*) + \langle A^*\overline{y}^*, \overline{u}\rangle = 0. \end{cases}$$

These are the Fenchel extremality relations, which are equivalent to

$$\begin{cases} \overline{y}^* \in \partial\Phi(A\overline{u}), \\ -A^*\overline{y}^* \in \partial\Psi(\overline{u}). \end{cases} \quad \square$$

Remark 9.8.1. It is often more convenient to write the dual problem as a minimization problem

$$\inf_{y^*\in Y^*} \left\{\Phi^*(y^*) + \Psi^*(-A^*y^*)\right\}.$$

Let us come back to the Dirichlet problem and apply the above results: we recall that $V = H_0^1(\Omega)$ and $Y = L^2(\Omega)^N$.

(a) $A : H_0^1(\Omega) \to L^2(\Omega)^N$ is the gradient operator. The adjoint operator

$$A^* : L^2(\Omega)^N \to H^{-1}(\Omega) = H_0^1(\Omega)^*$$

is defined by

$$\begin{aligned}\langle A^* y, v\rangle_{(H^{-1},H_0^1)} &= \langle y, Av\rangle_{L^2(\Omega)^N} \\ &= \sum_{i=1}^N \int_\Omega y_i \frac{\partial v}{\partial x_i}\, dx \\ &= \left\langle -\sum_{i=1}^N \frac{\partial y_i}{\partial x_i},\ v \right\rangle_{(\mathcal{D}'(\Omega),\mathcal{D}(\Omega))},\end{aligned}$$

where in the last equality v varies in $\mathcal{D}(\Omega)^N$, that is, $A^* y = -\sum_{i=1}^N D_i y_i = -\operatorname{div} y$ in the distribution sense, i.e., A^* is the opposite of the divergence operator.

(b) We know by Theorem 9.3.3 that the conjugate of the function

$$\Phi(y) = \frac{1}{2}\int_\Omega |y(x)|^2\, dx$$

is equal to

$$\Phi^*(y^*) = \frac{1}{2}\int_\Omega |y^*(x)|^2\, dx.$$

(c) Let $\Psi(v) = -\int_\Omega h(x)v(x)\, dx$. An easy computation gives

$$\begin{aligned}\Psi^*(v^*) &= \sup_{v\in H_0^1(\Omega)} \left\{\langle v^*, v\rangle_{(H^{-1},H_0^1)} - \langle -h, v\rangle_{L^2(\Omega)}\right\} \\ &= \sup_{v\in H_0^1(\Omega)} \langle v^* + h, v\rangle_{(H^{-1},H_0^1)} \\ &= \begin{cases} 0 & \text{if } v^* + h = 0, \\ +\infty & \text{otherwise.}\end{cases}\end{aligned}$$

We now collect all these results to obtain, thanks to Theorem 9.8.1, the description of the dual problem $(\mathcal{P}^*)$ of the Dirichlet problem

$$\sup\left\{-\frac{1}{2}\int_\Omega |y^*(x)|^2\, dx \ :\ y^* \in L^2(\Omega)^N,\ \operatorname{div} y^* = h\right\}. \qquad (\mathcal{P}^*)$$

Clearly, the generalized Slater condition is satisfied (F is everywhere continuous!). Thus there exists a solution $\overline{y}*$ of $(\mathcal{P}^*)$ and this solution is unique because of the strict convexity of the mapping $y^* \mapsto \int_\Omega |y^*(x)|^2\, dx$. On the other hand, we know that $(\mathcal{P})$ admits a unique solution u. The extremality relations yield

$$\begin{cases} \overline{y}^* = Au = \operatorname{grad} u, \\ \operatorname{div} \overline{y}^* = -h, \end{cases}$$

i.e., we have $-\operatorname{div}(\operatorname{grad} u) = h$ which is in accordance with the definition of u.

Part II

Advanced Variational Analysis

Chapter 10
Spaces BV and SBV

The modelization of a large number of problems in physics, mechanics, or image processing requires the introduction of new functional spaces permitting discontinuities of the solution. In phase transitions, image segmentation, plasticity theory, the study of cracks and fissures, the study of the wake in fluid dynamics, and so forth, the solution of the problem presents discontinuities along one-codimensional manifolds. Its first distributional derivatives are now measures which may charge zero Lebesgue measure sets, and the solution of these problems cannot be found in classical Sobolev spaces. Thus, the classical theory of Sobolev spaces must be completed by the new spaces $BV(\Omega)$ and $SBV(\Omega)$.

10.1 The space $BV(\Omega)$: Definition, convergences, and approximation

In this section Ω is an open subset of $\mathbf{R}^N$. Let us recall (see chapter 4) that $\mathbf{M}(\Omega, \mathbf{R}^N)$ denotes the space of all $\mathbf{R}^N$-valued Borel measures, which is also, according to the Riesz theory, the dual of the space $\mathbf{C}_0(\Omega, \mathbf{R}^N)$ of all continuous functions φ vanishing at infinity, equipped with the uniform norm $||\varphi||_\infty = \left(\sum_{i=1}^N \sup_{x\in\Omega} |\varphi_i(x)|^2\right)^{1/2}$. Note that $\mathbf{M}(\Omega, \mathbf{R}^N)$ is isomorphic to the product space $\mathbf{M}^N(\Omega)$ and that $\mu = (\mu_1, \ldots, \mu_N) \in \mathbf{M}(\Omega, \mathbf{R}^N) \iff \mu_i \in \mathbf{C}_0'(\Omega),\ i = 1, \ldots, N$.

Definition 10.1.1. *We say that a function $u : \Omega \to \mathbf{R}$ is a function of bounded variation iff it belongs to $L^1(\Omega)$ and its gradient Du in the distributional sense belongs to $\mathbf{M}(\Omega, \mathbf{R}^N)$. We denote the set of all functions of bounded variation by $BV(\Omega)$. The four following assertions are then equivalent:*

(i) $u \in BV(\Omega)$;

(ii) $u \in L^1(\Omega)$ *and* $\forall i = 1 \ldots N,\ \dfrac{\partial u}{\partial x_i} \in \mathbf{M}(\Omega)$;

(iii) $u \in L^1(\Omega)$ *and* $||Du|| := \sup\{\langle Du, \varphi\rangle\ : \varphi \in \mathbf{C}_c(\Omega, \mathbf{R}^N),\ ||\varphi||_\infty \le 1\} < +\infty$;

(iv) $u \in L^1(\Omega)$ *and* $||Du|| = \sup\{\int_\Omega u \text{ div } \varphi \, dx : \varphi \in \mathbf{C}_c^1(\Omega, \mathbf{R}^N),\ ||\varphi||_\infty \leq 1\} < +\infty$, *where the bracket* $<,>$ *in* (iii) *is defined by*

$$\langle Du, \varphi \rangle := \sum_{i=1}^{N} \int_\Omega \varphi_i \, \frac{\partial u}{\partial x_i}.$$

Equivalence between (ii) and (iii) is a direct consequence of the density of the space $\mathbf{C}_c(\Omega, \mathbf{R}^N)$ in $\mathbf{C}_0(\Omega, \mathbf{R}^N)$ equipped with the uniform norm. Equivalence between (iii) and (iv) can easily be established by the density of $\mathbf{C}_c^\infty(\Omega, \mathbf{R}^N)$ in $\mathbf{C}_c(\Omega, \mathbf{R}^N)$ and $\mathbf{C}_c^1(\Omega, \mathbf{R}^N)$.

Remark 10.1.1. According to the vectorial version of the Riesz–Alexandroff representation theorem, Theorem 2.4.6, the dual norm $||Du||$ is also the total mass $|Du|(\Omega) = \int_\Omega |Du|$ of the total variation $|Du|$ of the measure Du. Moreover, from classical integration theory, the integral $\int_\Omega f \, Du$ can be defined for all Du-integrable functions f from Ω into $\mathbf{R}^N$ as, for example, for functions in $\mathbf{C}_b(\Omega, \mathbf{R}^N)$. For the same reasons, $\int_\Omega f \, |Du|$ is well defined for all $|Du|$-integrable real-valued functions f as, for example, for functions in $\mathbf{C}_b(\Omega)$.

According to the Radon–Nikodym theorem, Theorem 4.2.1, there exist $\nabla u \in L^1(\Omega, \mathbf{R}^N)$ and a measure $D_s u$, singular with respect to the N-dimensional Lebesgue measure $\mathcal{L}^N \lfloor \Omega$ restricted to Ω, such that $Du = \nabla u \, \mathcal{L}^N \lfloor \Omega + D_s u$. Consequently, $W^{1,1}(\Omega)$ is a subspace of the vectorial space $BV(\Omega)$ and $u \in W^{1,1}(\Omega)$ iff $Du = \nabla u \, \mathcal{L}^N \lfloor \Omega$. For functions in $W^{1,1}(\Omega)$, we will sometimes write ∇u for Du. The space $BV(\Omega)$ is equipped with the following norm, which extends the classical norm in $W^{1,1}(\Omega)$:

$$||u||_{BV(\Omega)} := |u|_{L^1(\Omega)} + ||Du||.$$

We will define two weak convergence processes in $BV(\Omega)$. The first is too weak to ensure continuity of the trace operator defined in Section 10.2 but sufficient to provide compactness of bounded sequences. The second is an intermediate convergence between the weak and the strong convergence associated with the norm.

Definition 10.1.2. *A sequence* $(u_n)_{n \in \mathbf{N}}$ *in* $BV(\Omega)$ *weakly converges to some* u *in* $BV(\Omega)$, *and we write* $u_n \rightharpoonup u$ *iff the two following convergences hold:*

$$u_n \to u \text{ strongly in } L^1(\Omega);$$
$$Du_n \rightharpoonup Du \text{ weakly in } \mathbf{M}(\Omega, \mathbf{R}^N).$$

We will see later that when Ω is regular, the boundedness of a sequence in $BV(\Omega)$ is sufficient to ensure the existence of a weak cluster point (Theorem 10.1.4). In the proposition below we establish a compactness result related to this convergence, together with the lower semicontinuity of the total mass.

Proposition 10.1.1. *Let* $(u_n)_{n \in \mathbf{N}}$ *be a sequence in* $BV(\Omega)$ *strongly converging to some* u *in* $L^1(\Omega)$ *and satisfying* $\sup_{n \in \mathbf{N}} \int_\Omega |Du_n| < +\infty$. *Then*

(i) $u \in BV(\Omega)$ *and* $\int_\Omega |Du| \le \liminf_{n\to+\infty} \int_\Omega |Du_n|$;

(ii) u_n *weakly converges to* u *in* $BV(\Omega)$.

PROOF. For all φ in $\mathbf{C}^1_c(\Omega, \mathbf{R}^N)$ such that $||\varphi||_\infty \le 1$, we have

$$\int_\Omega u \ \operatorname{div}\varphi \, dx = \lim_{n\to+\infty} \int_\Omega u_n \ \operatorname{div}\varphi \, dx \le \liminf_{n\to+\infty} \int_\Omega |Du_n|$$

and assertion (i) is proved by taking the supremum in the first member, over all the elements φ in $\mathbf{C}^1_c(\Omega, \mathbf{R}^N)$ satisfying $||\varphi||_\infty \le 1$.

We now establish (ii). Since u_n strongly converges to u in $L^1(\Omega)$, for all $\varphi \in \mathbf{C}^\infty_c(\Omega, \mathbf{R}^N)$, one has

$$\langle Du_n, \varphi\rangle = -\int_\Omega u_n \ \operatorname{div} \ \varphi \, dx \to -\int_\Omega u \ \operatorname{div} \ \varphi \, dx = \langle Du, \varphi\rangle.$$

By using the density of $\mathbf{C}^\infty_c(\Omega, \mathbf{R}^N)$ in $\mathbf{C}_0(\Omega, \mathbf{R}^N)$ for the uniform norm and the boundedness of $(Du_n)_{n\in\mathbf{N}}$, we easily conclude that the sequence $(Du_n)_{n\in\mathbf{N}}$ weakly converges to Du. □

As a consequence of the semicontinuity property (i), $BV(\Omega)$ is a complete normed space.

Theorem 10.1.1. *Equipped with its norm,* $BV(\Omega)$ *is a Banach space.*

PROOF. Let $(u_n)_{n\in\mathbf{N}}$ be a Cauchy sequence in $BV(\Omega)$. Then $(u_n)_{n\in\mathbf{N}}$ is a Cauchy sequence in $L^1(\Omega)$ and for all $\varepsilon > 0$ there exists N_ε in $\mathbf{N}$ such that

$$\text{for all } p, q > N_\varepsilon, \ \int_\Omega |Du_p - Du_q| < \varepsilon. \tag{10.1}$$

Thus, there exists $u \in L^1(\Omega)$ such that $u_n \to u$ strongly in $L^1(\Omega)$. In particular $u_p - u_q \to u - u_q$ strongly in $L^1(\Omega)$ when p goes to $+\infty$. According to the lower semicontinuity property (i) of Proposition 10.1.1, (10.1) yields, for $q > N_\varepsilon$,

$$\int_\Omega |D(u - u_q)| \le \liminf_{p\to+\infty} \int_\Omega |D(u_p - u_q)| \le \varepsilon.$$

This estimate yields first $u \in BV(\Omega)$ then $\lim_{q\to+\infty} \int_\Omega |D(u - u_q)| = 0$, so that $u_n \to u$ in $BV(\Omega)$. □

To define the second weak convergence process, let us recall the notion of narrow convergence defined in Section 4.2.2. As said in Remark 10.1.1, the integral $\int_\Omega f \, |Du|$ is well defined for all f in the space $\mathbf{C}_b(\Omega)$ of bounded continuous functions on Ω. Thus

$|Du|$ may be considered as an element of $\mathbf{C}_b'(\Omega)$. We now say that a sequence $(|Du_n|)_{n\in\mathbf{N}}$ narrowly converges to μ in $\mathbf{M}(\Omega)$ iff $|Du_n| \rightharpoonup \mu$ for the $\sigma(\mathbf{C}_b'(\Omega), \mathbf{C}_b(\Omega))$ convergence (see Section 4.2.2).

Definition 10.1.3. *Let $(u_n)_{n\in\mathbf{N}}$ be a sequence in $BV(\Omega)$ and $u \in BV(\Omega)$. We say that u_n converges to u in the sense of the intermediate convergence iff*

$$u_n \to u \text{ strongly in } L^1(\Omega),$$
$$\int_\Omega |Du_n| \to \int_\Omega |Du|.$$

The term *intermediate convergence* is due to Temam [219] and is also called strict convergence. Let us notice that according to Proposition 10.1.1(i), when u_n strongly converges to u in $L^1(\Omega)$, there is in general loss of the total mass at the limit. The proposition below states that this convergence is stronger than the weak convergence, therefore justifying the terminology.

Proposition 10.1.2. *The three following assertions are equivalent:*

(i) $u_n \rightharpoonup u$ *in the sense of the intermediate convergence;*

(ii) $\begin{cases} u_n \rightharpoonup u \text{ weakly in } BV(\Omega), \\ \int_\Omega |Du_n| \to \int_\Omega |Du|; \end{cases}$

(iii) $\begin{cases} u_n \to u \text{ strongly in } L^1(\Omega), \\ |Du_n| \rightharpoonup |Du| \text{ narrowly in } \mathbf{M}(\Omega). \end{cases}$

Proof. (i) $\Longrightarrow$ (ii). This implication is a straightforward consequence of Proposition 10.1.1. We are going to prove (ii) $\Longrightarrow$ (iii). Let us recall (cf. Proposition 4.2.5) that for nonnegative Borel measures μ_n and μ in $\mathbf{M}(\Omega)$, there is equivalence between $\mu_n \rightharpoonup \mu$ narrowly and

$$\mu_n(\Omega) \to \mu(\Omega),$$

$$\mu(U) \le \liminf_{n\to+\infty} \mu_n(U) \ \forall \text{ open subset } U \text{ of } \Omega.$$

Set $\mu_n = |Du_n|$ and $\mu = |Du|$. At first we have $\mu_n(\Omega) = \int_\Omega |Du_n| \to \mu(\Omega) = \int_\Omega |Du|$. Let now U be any open subset of Ω. Obviously, u_n and u belong to $BV(U)$ and $u_n \to u$ strongly in $L^1(U)$. Applying Proposition 10.1.1 with $\Omega = U$ (note that $\sup_{n\in\mathbf{N}} \int_U |Du_n| \le \sup_{n\in\mathbf{N}} \int_\Omega |Du_n| < +\infty$), we obtain

$$\int_U |Du| \le \liminf_{n\to+\infty} \int_U |Du_n|.$$

Implication (iii) $\Longrightarrow$ (i) is straightforward. □

Remark 10.1.2. It results from Propositions 4.2.5 and 10.1.2 that if $u_n \rightharpoonup u$ in the sense of the intermediate convergence, for all Borel subset B of Ω such that $\int_{\partial B} |Du| = 0$, one has

$$\int_B |Du_n| \to \int_B |Du|.$$

More generally, according to Proposition 4.2.6, for all bounded Borel function $f : \Omega \to \mathbf{R}$ such that the set of its discontinuity points has a null $|Du|$-measure, one has

$$\int_\Omega f|Du_n| \to \int_\Omega f|Du|.$$

Remark 10.1.3. The intermediate convergence is strictly finer than the weak convergence in $BV(\Omega)$. Indeed, the sequence of functions in $BV(0,1)$ defined by

$$u_n(x) = \begin{cases} nx \ \text{ if } 0 < x \leq \frac{1}{n}, \\ 1 \ \text{ if } x > \frac{1}{n}, \end{cases}$$

weakly converges to the function 1 in $BV(0,1)$ and does not converge in the sense of the intermediate convergence. Indeed, the total mass $|Du_n|(0,1)$ is the constant 1.

The space $\mathbf{C}^\infty(\overline{\Omega})$ is not dense in $BV(\Omega)$ when $BV(\Omega)$ is equipped with its strong convergence. Indeed, its closure is the space $W^{1,1}(\Omega)$ (see Proposition 5.4.1). Nevertheless, one can approximate every element of $BV(\Omega)$ by a function of $\mathbf{C}^\infty(\overline{\Omega})$ in the sense of the intermediate convergence.

Theorem 10.1.2. *The space $C^\infty(\Omega) \cap BV(\Omega)$ is dense in $BV(\Omega)$ equipped with the intermediate convergence. Consequently $\mathbf{C}^\infty(\overline{\Omega})$ is also dense in $BV(\Omega)$ for the intermediate convergence.*

Proof. First notice that $\mathbf{C}^\infty(\Omega) \cap BV(\Omega) = \mathbf{C}^\infty(\Omega) \cap W^{1,1}(\Omega)$. The second assertion is then a straightforward consequence of the density of the space $\mathbf{C}^\infty(\overline{\Omega})$ in $W^{1,1}(\Omega)$ equipped with its strong convergence (Proposition 5.4.1) which is finer than the intermediate convergence.

Let $\varepsilon > 0$, intended to go to zero, and $u \in BV(\Omega)$. We are going to construct u_ε in $\mathbf{C}^\infty(\Omega) \cap W^{1,1}(\Omega)$ such that

$$\int_\Omega |u - u_\varepsilon| \, dx < \varepsilon \text{ and } \Big| \int_\Omega |Du_\varepsilon| - \int_\Omega |Du| \, \Big| < 4\varepsilon. \tag{10.2}$$

The following construction is similar to the proof of the Meyers–Serrin theorem, Theorem 5.1.4. Let us consider a family $(\Omega_i)_{i \in \mathbf{N}}$ of open subsets of Ω such that

$$\begin{aligned} &\int_{\Omega \setminus \Omega_0} |Du| < \varepsilon; \\ &\Omega_i \subset\subset \Omega_{i+1}; \\ &\Omega = \bigcup_{i=0}^{\infty} \Omega_i. \end{aligned} \tag{10.3}$$

We construct the open covering $(C_i)_{i\in\mathbf{N}^*}$ of Ω as follows: set $C_1 = \Omega_2$ and, for $i \geq 2$, $C_i = \Omega_{i+1} \setminus \overline{\Omega}_{i-1}$. Let now $(\varphi_i)_{i\in\mathbf{N}^*}$ be a partition of unity subordinate to the covering $(C_i)_{i\in\mathbf{N}^*}$. The functions φ_i satisfy : $\varphi_i \in \mathbf{C}_c^\infty(C_i)$, $0 \leq \varphi_i \leq 1$, $\sum_{i=1}^\infty \varphi_i = 1$. Note that $\varphi_1 = 1$ on Ω_1. For each i, choose $\varepsilon_i > 0$ such that

$$\operatorname{spt}(\rho_{\varepsilon_i} * \varphi_i u) \subset C_i, \tag{10.4}$$

$$\left| \int_\Omega |\rho_{\varepsilon_1} * (\varphi_1 Du)|\, dx - \int_\Omega |\varphi_1 Du| \right| < \varepsilon, \tag{10.5}$$

$$\int_\Omega |\rho_{\varepsilon_i} * (u\varphi_i) - u\varphi_i|\, dx < \varepsilon 2^{-i}, \tag{10.6}$$

$$\int_\Omega |\rho_{\varepsilon_i} * (uD\varphi_i) - uD\varphi_i|\, dx < \varepsilon 2^{-i}, \tag{10.7}$$

where ρ_{ε_i} are the regularizers defined in Theorem 4.2.2. Estimate (10.5) is obtained by applying Theorem 4.2.2(iii) to the measure $\varphi_1 Du$. Estimates (10.6) and (10.7) are straightforward consequences of the convergence of $\rho_\varepsilon * (u\varphi_i)$ and $\rho_\varepsilon * (uD\varphi_i)$, respectively, to $u\varphi_i$ and $uD\varphi_i$ in $L^1(\Omega)$ (see Proposition 2.2.4). We define u_ε by

$$u_\varepsilon = \sum_{i=1}^\infty \rho_{\varepsilon_i} * (u\varphi_i).$$

Note that each x in Ω belongs to at most two of the sets C_i and that, by (10.4), u_ε is well defined and clearly belongs to $\mathbf{C}^\infty(\Omega)$. From (10.6) we obtain

$$\int_\Omega |u - u_\varepsilon|\, dx \leq \sum_{i=1}^\infty \int_\Omega |\rho_{\varepsilon_i} * (u\varphi_i) - u\varphi_i|\, dx < \varepsilon.$$

We are going to establish the last estimate of (10.2). In the distributional sense, we have $D(u\varphi_i) = \varphi_i Du + uD\varphi_i \mathcal{L}\lfloor\Omega$ so that

$$\begin{aligned} Du_\varepsilon &= \sum_{i=1}^\infty D(\rho_{\varepsilon_i} * (u\varphi_i)) = \sum_{i=1}^\infty \rho_{\varepsilon_i} * (D(u\varphi_i)) \\ &= \sum_{i=1}^\infty \rho_{\varepsilon_i} * (\varphi_i Du) + \sum_{i=1}^\infty \rho_{\varepsilon_i} * (uD\varphi_i) \\ &= \sum_{i=1}^\infty \rho_{\varepsilon_i} * (\varphi_i Du) + \sum_{i=1}^\infty \Big(\rho_{\varepsilon_i} * (uD\varphi_i) - uD\varphi_i\Big). \end{aligned}$$

Therefore, according to (10.7), Theorem 4.2.2(ii), and (10.3),

$$\begin{aligned} \left| \int_\Omega |\rho_{\varepsilon_1} * (\varphi_1 Du)|\, dx - \int_\Omega |Du_\varepsilon| \right| &\leq \sum_{i=2}^\infty \int_\Omega |\rho_{\varepsilon_i} * (\varphi_i Du)|\, dx \\ &\quad + \sum_{i=1}^\infty \int_\Omega |\rho_{\varepsilon_i} * (uD\varphi_i) - uD\varphi_i|\, dx \end{aligned}$$

$$\leq \sum_{i=2}^{\infty} \int_{\Omega} |\rho_{\varepsilon_i} * (\varphi_i Du)|\, dx + \varepsilon$$
$$\leq \sum_{i=2}^{\infty} \int_{\Omega} |\varphi_i Du| + \varepsilon$$
$$\leq \int_{\Omega\setminus\Omega_0} |Du| + \varepsilon < 2\varepsilon. \tag{10.8}$$

On the other hand, from (10.5), (10.3) and because $\varphi_1 = 1$ on Ω_1,

$$\left| \int_{\Omega} |\rho_{\varepsilon_1} * (\varphi_1 Du)|\, dx - \int_{\Omega} |Du| \right| \leq \varepsilon + \int_{\Omega} (1-\varphi_1)|Du|$$
$$\leq \varepsilon + \int_{\Omega\setminus\Omega_0} |Du| \leq 2\varepsilon. \tag{10.9}$$

Collecting (10.8), (10.9) we obtain

$$\left| \int_{\Omega} |Du_\varepsilon| - \int_{\Omega} |Du| \right| < 4\varepsilon,$$

which completes the proof of (10.2). □

Theorem 10.1.2 allows us to extend Sobolev's inequalities and compactness embedding results on $W^{1,1}(\Omega)$ (see Section 5.7) to the space $BV(\Omega)$.

Theorem 10.1.3. *Let Ω be a 1-regular open bounded subset of $\mathbf{R}^N$. For all p, $1 \leq p \leq \frac{N}{N-1}$, the embedding*

$$BV(\Omega) \hookrightarrow L^p(\Omega)$$

is continuous. More precisely, there exists a constant C which depends only on Ω, p, and N, such that for all u in $BV(\Omega)$,

$$\left(\int_{\Omega} |u|^p\, dx \right)^{\frac{1}{p}} \leq C|u|_{BV(\Omega)}.$$

Proof. Let $(u_n)_{n\in\mathbf{N}}$ be a sequence of functions in $\mathbf{C}^\infty(\Omega)\cap BV(\Omega)$ which converges to some u in $BV(\Omega)$ for the intermediate convergence. Since the embedding $W^{1,1}(\Omega) \hookrightarrow L^p(\Omega)$ is continuous for $1 \leq p \leq \frac{N}{N-1}$, there exists a constant C, which depends only on Ω, p, and N such that

$$\left(\int_{\Omega} |u_n|^p\, dx \right)^{\frac{1}{p}} \leq C \left(|u_n|_{L^1(\Omega)} + \int_{\Omega} |Du_n|\, dx \right) < +\infty.$$

We deduce that $u_n \rightharpoonup u$ in $L^p(\Omega)$ and, according to the weak lower semicontinuity of the norm of $L^p(\Omega)$,

$$\left(\int_{\Omega} |u|^p\, dx \right)^{\frac{1}{p}} \leq \liminf_{n\to+\infty} \left(\int_{\Omega} |u_n|^p\, dx \right)^{\frac{1}{p}}$$
$$\leq \liminf_{n\to+\infty} C \left(|u_n|_{L^1(\Omega)} + \int_{\Omega} |Du_n|\, dx \right)$$
$$= C\|u\|_{BV(\Omega)},$$

where we have used the intermediate convergence in the last equality. □

Theorem 10.1.4. *Let Ω be a 1-regular open bounded subset of $\mathbf{R}^N$. Then for all p, $1 \leq p < \frac{N}{N-1}$ the embedding*

$$BV(\Omega) \hookrightarrow L^p(\Omega)$$

is compact.

Proof. According to Theorem 10.1.3, every element of $BV(\Omega)$ belongs to $L^p(\Omega)$ for $1 \leq p \leq \frac{N}{N-1}$, so that we can slightly improve the density Theorem 10.1.2 as follows: for all $u \in BV(\Omega)$, there exists $u_n \in \mathbf{C}^\infty(\Omega) \cap BV(\Omega)$ satisfying

$$\begin{cases} u_n \to u \text{ in } L^p(\Omega), \\ \displaystyle\int_\Omega |Du_n| \to \int_\Omega |Du|. \end{cases}$$

We conclude thanks to the compactness of the embedding of $W^{1,1}(\Omega) \hookrightarrow L^p(\Omega)$. Indeed, let u_n be such that $\|u_n\|_{BV(\Omega)} \leq 1$ and $v_n \in \mathbf{C}^\infty(\Omega) \cap BV(\Omega)$ be such that

$$\begin{cases} |v_n - u_n|_{L^p(\Omega)} \leq \frac{1}{n}, \\ \displaystyle\int_\Omega |Dv_n|\, dx \leq 2. \end{cases}$$

Since $\|v_n\|_{W^{1,1}(\Omega)} \leq 4$ for n large enough there exists a subsequence $(v_{n_k})_{k\in\mathbf{N}}$ and u in $L^p(\Omega)$ such that

$$v_{n_k} \to u \text{ strongly in } L^p(\Omega),$$

thus

$$u_{n_k} \to u \text{ strongly in } L^p(\Omega).$$

By the lower semicontinuity of the total mass, and since u_{n_k} strongly converges to u in $L^1(\Omega)$, we obtain

$$\begin{aligned} |u|_{L^1(\Omega)} + \int_\Omega |Du| &\leq \lim_{k\to+\infty} |u_{n_k}|_{L^1(\Omega)} + \liminf_{k\to+\infty} \int_\Omega |Du_{n_k}| \\ &\leq \liminf_{k\to+\infty} \|u_{n_k}\|_{BV(\Omega)} \leq 1 \end{aligned}$$

and the proof is complete. □

10.2 The trace operator, the Green's formula, and its consequences

In all this section, Ω is a domain of $\mathbf{R}^N$ with a Lipschitz boundary Γ (i.e., a Lipschitz domain). Under a weaker hypothesis on the regularity of the set Ω, we extend in the space $BV(\Omega)$ the notion of trace developed for Sobolev functions in Section 5.6. It is worth noticing that the method used for establishing the trace theorem below also applies to Sobolev functions.

Theorem 10.2.1. *There exists a linear continuous map γ_0 from $BV(\Omega)$ onto $L^1_{\mathcal{H}^{N-1}}(\Gamma)$ satisfying*

(i) *for all* u *in* $\mathbf{C}(\overline{\Omega}) \cap BV(\Omega)$, $\gamma_0(u) = u\lfloor\Gamma$;

(ii) *the generalized Green's formula holds:* $\forall \varphi \in \mathbf{C}^1(\overline{\Omega}, \mathbf{R}^N)$,

$$\int_\Omega \varphi Du = -\int_\Omega u \ \text{div}\ \varphi \, dx + \int_\Gamma \gamma_0(u)\, \varphi.\nu \, d\mathcal{H}^{N-1},$$

where $\nu(x)$ *is the outer unit normal at* $\mathcal{H}^{N-1}$ *almost all* x *in* Γ.

PROOF. Each generic element x in $\mathbf{R}^N$ will be denoted by $x = (\tilde{x}, x_N)$, where $\tilde{x} = (x_1, \ldots, x_{N-1}) \in \mathbf{R}^{N-1}$ and $x_N \in \mathbf{R}$. Let us consider a finite cover of Γ by the open cylinders

$$C_R(y) = S_R(\tilde{y}) \times (y_N - R, y_N + R), \ \ y = (\tilde{y}, y_N) \in \Gamma,$$

where $S_R(\tilde{y})$ is the open ball of $\mathbf{R}^{N-1}$ with radius R centered at $\tilde{y} = (y_1, \ldots, y_{N-1})$. Since Γ is Lipschitz regular, relabeling the coordinate axes if necessary, there exists $\varepsilon_0 > 0$ and a Lipschitz function f such that $\Omega \cap C_R(y)$ contains the open set

$$C_{R,\varepsilon_0}(y) := \{x \in \mathbf{R}^N : \tilde{x} \in S_R(\tilde{y}), \ f(\tilde{x}) - \varepsilon_0 < x_N < f(\tilde{x})\},$$

and

$$\Sigma(y) := \{(\tilde{x}, x_N) : \tilde{x} \in S_R(\tilde{y}), \ x_N = f(\tilde{x})\}$$

is a neighborhood of y in Γ. Let u be a fixed function in $BV(\Omega)$. According to Lemma 4.2.2, since $\int_{C_{R,\varepsilon_0}(y)} |Du| < +\infty$, R and ε_0 can be chosen so that the measure $|Du|$ does not charge $\partial C_{R,\varepsilon_0}(y) \setminus \Sigma(y)$, i.e., $\int_{\partial C_{R,\varepsilon_0}(y)\setminus\Sigma(y)} |Du| = 0$.

First step. We fix y in Γ and, to shorten notation, we denote the sets $\Sigma(y)$, $S_R(\tilde{y})$, and $C_{R,\varepsilon_0}(y)$ by Σ, S_R, and C_{R,ε_0}, respectively. Since Γ is Lipschitz, according to Rademacher's theorem (see [132]), the outer unit normal $\nu(x)$ exists at $\mathcal{H}^{N-1}$ a.e. x on Γ. This step is devoted to the existence of u^+ in $L^1_{\mathcal{H}^{N-1}}(\Sigma)$ satisfying

$$\begin{cases} \displaystyle\int_\Sigma |u^+| \, d\mathcal{H}^{N-1} \le C\Big(\int_{C_{R,\varepsilon_0}} |u| \, dx + \int_{C_{R,\varepsilon_0}} |Du|\Big) \\ \forall \varphi \in \mathbf{C}^1_c(C_{R,\varepsilon_0} \cup \Sigma, \mathbf{R}^N) \displaystyle\int_{C_{R,\varepsilon_0}} \varphi Du = -\int_{C_{R,\varepsilon_0}} u \ \text{div}\ \varphi \, dx + \int_\Sigma u^+ \varphi \nu \, d\mathcal{H}^{N-1}, \end{cases} \tag{10.10}$$

where C is a positive constant depending only on f.

For all regular function v defined on C_{R,ε_0} and all ε in $]0, \varepsilon_0[$, we adopt the notation $v^\varepsilon(\tilde{x}) := v(\tilde{x}, f(\tilde{x}) - \varepsilon)$. Consider now a sequence $(u_n)_{n\in\mathbf{N}}$ in $\mathbf{C}^\infty(C_{R,\varepsilon_0}) \cap BV(C_{R,\varepsilon_0})$ which converges to u in $BV(C_{R,\varepsilon_0})$ in the sense of intermediate convergence (see Theorem 10.1.2). We have

$$\begin{cases} \displaystyle\int_{C_{R,\varepsilon_0}} |u_n - u| \, dx \to 0, \\ \displaystyle\int_{C_{R,\varepsilon_0}} |Du_n| \, dx \to \int_{C_{R,\varepsilon_0}} |Du|. \end{cases} \tag{10.11}$$

Since the function u_n is smooth, for $\varepsilon > \varepsilon'$ in $]0, \varepsilon_0[$ one has

$$u_n^{\varepsilon}(\tilde{x}) - u_n^{\varepsilon'}(\tilde{x}) = \int_{-\varepsilon'}^{-\varepsilon} \frac{\partial u_n}{\partial x_N}(\tilde{x}, f(\tilde{x}) + s)\, ds,$$

so that

$$|u_n^{\varepsilon}(\tilde{x}) - u_n^{\varepsilon'}(\tilde{x})| \leq \int_{\varepsilon'}^{\varepsilon} |Du_n(\tilde{x}, f(\tilde{x}) - s)|\, ds.$$

Thus, according to Proposition 4.1.6 and Remark 4.1.5 applied to the map $S_R \subset \mathbf{R}^{N-1} \to \mathbf{R}^N$, $\tilde{x} \mapsto (\tilde{x}, f(\tilde{x}))$, we deduce

$$\begin{aligned}
\int_{\Sigma} |u_n^{\varepsilon}(\tilde{x}) - u_n^{\varepsilon'}(\tilde{x})|\, d\mathcal{H}^{N-1}(x) &\leq \int_{S_R} |u_n^{\varepsilon}(\tilde{x}) - u_n^{\varepsilon'}(\tilde{x})| \sqrt{1 + |Df(\tilde{x})|^2}\, d\tilde{x} \\
&\leq C \int_{S_R} \int_{\varepsilon'}^{\varepsilon} |Du_n(\tilde{x}, f(\tilde{x}) - s)|\, ds\, d\tilde{x} \\
&= C \int_{C_{R,\varepsilon,\varepsilon'}} |Du_n(x)|\, dx,
\end{aligned}$$

where C is a positive constant depending only on f and

$$C_{R,\varepsilon,\varepsilon'} = \{x \in \mathbf{R}^N : \tilde{x} \in S_R,\ f(\tilde{x}) - \varepsilon < x_N < f(\tilde{x}) - \varepsilon'\}.$$

Therefore, with the notation made precise above,

$$\int_{\Sigma} |u_n^{\varepsilon} - u_n^{\varepsilon'}|\, d\mathcal{H}^{N-1} \leq C \int_{C_{R,\varepsilon,\varepsilon'}} |Du_n|\, dx. \tag{10.12}$$

We intend to go to the limit on n in (10.12). From the coarea formula (Theorem 4.2.5), one has

$$\int_{-\varepsilon_0}^{0} \left(\int_{\Sigma} |u_n^{\varepsilon}(\tilde{x}) - u^{\varepsilon}(\tilde{x})|\, d\mathcal{H}^{N-1}(x) \right) d\varepsilon \leq C \int_{C_{R,\varepsilon_0}} |u_n(x) - u(x)|\, dx,$$

and the first limit in (10.11) yields the existence of a subsequence on n (not relabeled) such that for a.e. ε in $]0, \varepsilon_0[$, $u_n^{\varepsilon} \to u^{\varepsilon}$ in $L^1_{\mathcal{H}^{N-1}}(\Sigma)$. On the other hand, according to Proposition 10.1.2, (10.11) ensures the narrow convergence of the measure $|Du_n|$ to the measure $|Du|$ in $\mathbf{M}^+(C_{R,\varepsilon_0})$. Let us consider $\varepsilon,\ \varepsilon'$ such that $\int_{\partial C_{R,\varepsilon,\varepsilon'}} |Du| = 0$. This choice is indeed valid in the complementary I of a countable subset of $]0, \varepsilon_0[$, by using Lemma 4.2.2 and because $|Du|$ does not charge $\partial C_{R,\varepsilon_0} \setminus \Sigma$. According to the properties of the narrow convergence (cf. Proposition 4.2.5), we have, for ε and ε' in I,

$$\int_{C_{R,\varepsilon,\varepsilon'}} |Du_n| \to \int_{C_{R,\varepsilon,\varepsilon'}} |Du|.$$

Going to the limit on n in (10.12), we finally obtain, for ε in $]0, \varepsilon_0[\backslash\mathcal{N}$, where $\mathcal{N}$ is an $\mathcal{L}^1$-negligible set,

$$\int_\Sigma |u^\varepsilon - u^{\varepsilon'}|\, d\mathcal{H}^{N-1} \le C \int_{C_{R,\varepsilon,\varepsilon'}} |Du|. \tag{10.13}$$

From now on, ε denotes a sequence of numbers in $]0, \varepsilon_0[\backslash\mathcal{N}$, going to zero and (10.13) must be taken in the sense

$$\int_\Sigma |u^{\varepsilon_p} - u^{\varepsilon_q}|\, d\mathcal{H}^{N-1} \le C \int_{C_{R,\varepsilon_p,\varepsilon_q}} |Du|.$$

From (10.13), $(u^\varepsilon)_\varepsilon$ is a Cauchy sequence in $L^1_{\mathcal{H}^{N-1}}(\Sigma)$, then strongly converges to some function u^+ in $L^1(\Sigma)$. It remains to prove that u^+ satisfies (10.10). Letting $\varepsilon' \to 0$ in (10.13) and integrating over $]-\varepsilon_0, 0[$, we obtain

$$\int_\Sigma |u^+|\, d\mathcal{H}^{N-1} \le C \left(\int_{C_{R,\varepsilon_0}} |u|\, dx + \int_{C_{R,\varepsilon_0}} |Du| \right). \tag{10.14}$$

On the other hand, since $u_n \in \mathbf{C}^\infty(C_{R,\varepsilon_0})$ and $\varphi \in \mathbf{C}^1_c(C_{R,\varepsilon_0} \cup \Sigma, \mathbf{R}^N)$, going to the limit on n in the classical Green's formula

$$\int_{C_{R,\varepsilon_0,\varepsilon}} u_n \ \text{div}\ \varphi\, dx = -\int_{C_{R,\varepsilon_0,\varepsilon}} Du_n.\varphi\, dx + \int_\Sigma u_n^\varepsilon \varphi^\varepsilon.\nu\, d\mathcal{H}^{N-1},$$

where

$$C_{R,\varepsilon_0,\varepsilon} := \{x \in \mathbf{R}^N : \tilde{x} \in S_R(\tilde{y}),\ f(\tilde{x}) - \varepsilon_0 < x_N < f(\tilde{x}) - \varepsilon\},$$

we claim that

$$\int_{C_{R,\varepsilon_0,\varepsilon}} u \ \text{div}\ \varphi\, dx = -\int_{C_{R,\varepsilon_0,\varepsilon}} \varphi Du + \int_\Sigma u^\varepsilon \varphi^\varepsilon.\nu\, d\mathcal{H}^{N-1}. \tag{10.15}$$

We must justify the convergence

$$\int_{C_{R,\varepsilon_0,\varepsilon}} Du_n.\varphi\, dx \to \int_{C_{R,\varepsilon_0,\varepsilon}} \varphi Du. \tag{10.16}$$

The two others are straightforward. We reason by truncation: let $\varepsilon' < \varepsilon$ and consider $\tilde{\varphi} = \varphi\theta$ in $\mathbf{C}^1_c(C_{R,\varepsilon_0,\varepsilon'}, \mathbf{R}^N)$ where the scalar function θ belongs to $\mathbf{C}_c(C_{R,\varepsilon_0,\varepsilon'})$ and satisfies $\theta = 1$ in $C_{R,\varepsilon_0,\varepsilon}$, with $||\theta||_\infty \le 1$. We have

$$\int_{C_{R,\varepsilon_0,\varepsilon}} Du_n\varphi\, dx = \int_{C_{R,\varepsilon_0,\varepsilon'}} Du_n\tilde{\varphi}\, dx - \int_{C_{R,\varepsilon,\varepsilon'}} Du_n\tilde{\varphi}\, dx. \tag{10.17}$$

From

$$\left| \int_{C_{R,\varepsilon,\varepsilon'}} Du_n \tilde{\varphi}\, dx \right| \leq ||\tilde{\varphi}||_\infty \int_{C_{R,\varepsilon,\varepsilon'}} |Du_n|$$

and according to the narrow convergence of $|Du_n|$ to $|Du|$, and because $|Du|$ does not charge $\partial C_{R,\varepsilon,\varepsilon'}$, for ε and ε' in $]0, \varepsilon_0[\setminus\mathcal{N}$, we obtain

$$\lim_{\varepsilon' \to \varepsilon} \limsup_{n \to +\infty} \left| \int_{C_{R,\varepsilon,\varepsilon'}} Du_n \tilde{\varphi}\, dx \right| = 0.$$

The convergence in (10.16) is obtained by letting $n \to +\infty$ and $\varepsilon' \to \varepsilon$ in (10.17) and the claim is proved.

Finally going to the limit on ε in (10.15) we obtain

$$\int_{C_{R,\varepsilon_0}} u \ \text{div}\ \varphi\, dx = -\int_{C_{R,\varepsilon_0}} \varphi Du - \int_{S_R} u^+ \varphi . e_N \ \mathcal{H}^{N-1}.$$

Second step. According to the first step and from a straightforward argument using a partition of unity subordinate to a finite cover $(C_R(y_i))_{i=1,\dots,r}$ of Γ, there exists $\gamma_0(u)$ in $L^1_{\mathcal{H}^{N-1}}(\Gamma)$ satisfying

$$\int_\Gamma |\gamma_0(u)|\, d\mathcal{H}^{N-1} \leq C\Big(\int_\Omega |u|\, dx + \int_\Omega |Du| \Big) \tag{10.18}$$

and such that for all $\varphi \in \mathbf{C}(\overline{\Omega}, \mathbf{R}^N)$,

$$\int_\Omega \varphi Du = -\int_\Omega u \ \text{div}\ \varphi\, dx + \int_\Gamma \gamma_0(u)\, \varphi . \nu\, d\mathcal{H}^{N-1}, \tag{10.19}$$

where $\nu(x)$ is the outer normal unit at $\mathcal{H}^{N-1}$ almost all x in Γ and C a positive constant depending only on Ω. The operator γ_0 is well defined by $\gamma_0(u) = u_i^+$, in Σ_i, $i = 1, \dots, r$. Indeed, Green's formula (10.10) established in the first step yields

$$\int_{\Sigma_i \cap \Sigma_j} (u_i^+ - u_j^+) \varphi . \nu\, d\mathcal{H}^{N-1} = 0$$

for all functions φ in $\mathbf{C}^1_c((C_{R,\varepsilon_0}(y_i) \cap C_{R,\varepsilon_0}(y_j)) \cup (\Sigma_i \cap \Sigma_j), \mathbf{R}^N)$ so that $u_i^+ = u_j^+$ in $\Sigma_i \cap \Sigma_j$ up to sets of $\mathcal{H}^{N-1}$ measure zero.

The generalized Green's formula (10.19) yields the linearity of γ_0. The continuity is a consequence of (10.18). The identity $\gamma_0(u) = u \lfloor S_R$ for all u in $\mathbf{C}(\overline{\Omega}) \cap BV(\Omega)$ is a straightforward consequence of the definition of $\gamma_0(u)$. The operator γ_0 then agrees with the trace operator defined in $W^{1,1}(\Omega)$. Since $\gamma_0(W^{1,1}(\Omega)) = L^1(\Omega)$, we also have $\gamma_0(BV(\Omega)) = L^1(\Omega)$. □

Remark 10.2.1. Let Ω be a Lipschitz open bounded subset of $\mathbf{R}^N$. The density theorem, Theorem 10.1.2, may be slightly improved in the sense that one may further assert that the trace of each regular approximating function of $u \in BV(\Omega)$, belonging to $\mathbf{C}^\infty(\Omega) \cap BV(\Omega)$,

coincides with the trace of u on the boundary of Ω. Indeed, it is easily seen, with the notations of Theorem 10.1.2, that $u_\varepsilon - u$ is the strong limit in $BV(\Omega)$ of the functions

$$u_{\varepsilon,n} - u_n := \sum_{i=0}^{n} \left(\rho_{\varepsilon_i} * (u\varphi_i) - u\varphi_i\right)$$

whose traces on $\Gamma := \partial\Omega$ are the function null. The result then follows from the strong continuity of the trace operator.

Remark 10.2.2. One may define the space $BV(\Omega, \mathbf{R}^m)$ as the space of all functions $u : \Omega \to \mathbf{R}^m$ in $L^1(\Omega, \mathbf{R}^m)$ whose distributional derivative Du belongs to the space $\mathbf{M}(\Omega, M^{m\times N})$ of $m \times N$ matrix-valued measures. Arguing as in the proof of Theorem 10.2.1 with each component of u, one may prove the existence of the trace operator γ_0 from $BV(\Omega, \mathbf{R}^m)$ onto $L^1_{\mathcal{H}^{N-1}}(\Gamma, \mathbf{R}^m)$ satisfying

(i) $\forall u \in \mathbf{C}(\overline{\Omega}, \mathbf{R}^m) \cap BV(\Omega, \mathbf{R}^m)$, $\gamma_0(u) = u\lfloor\Gamma$;

(ii) the Green's formula holds : $\forall \varphi \in \mathbf{C}^1(\overline{\Omega}, M^{m\times N})$

$$\int_\Omega \varphi : Du = -\int_\Omega u.\,\mathrm{div}\ \varphi\, dx + \int_\Gamma \gamma_0(u) \otimes \nu : \varphi\, d\mathcal{H}^{N-1},$$

where $\nu(x)$ is the outer unit normal at $\mathcal{H}^{N-1}$ almost all x in Γ, $\gamma_0(u) \otimes \nu$ is the $M^{m\times N}$ valued function $(\gamma_0(u)_i \nu_j)_{i=1\ldots m,\, j=1\ldots N}$. The integral with respect to the measure Du in the first member is defined by

$$\int_\Omega \varphi : Du := \sum_{i=1}^{N}\sum_{j=1}^{m} \int_\Omega \varphi_{i,j} \frac{\partial u_j}{\partial x_i}$$

and

$$\gamma_0(u) \otimes \nu : \varphi := \sum_{i=1}^{m}\sum_{j=1}^{N} \gamma_0(u)_i \nu_j\, \varphi_{i,j}.$$

The divergence of φ is the vector valued distribution $(\mathrm{div}\ \varphi)_j := \sum_{i=1}^{N} \frac{\partial u_j}{\partial x_i}$, $j = 1\ldots m$.

The density Theorem 10.1.2 and Remark 10.2.1 also hold in $BV(\Omega, \mathbf{R}^m)$.

We now give some consequences of Theorem 10.2.1.

Example 10.2.1. Consider two disjoint Lipschitz domains Ω_1 and Ω_2, included in an open bounded subset Ω of $\mathbf{R}^N$, such that $\overline{\Omega} = \overline{\Omega_1} \cup \overline{\Omega_2}$, and set $\Gamma_{1,2} := \partial\Omega_1 \cap \partial\Omega_2$ which is assumed to satisfy $\mathcal{H}^{N-1}(\Gamma_{1,2}) > 0$ (see Figure 10.1). We respectively denote the trace operators from $BV(\Omega_1)$ onto $L^1(\partial\Omega_1)$ and $BV(\Omega_2)$ onto $L^1(\partial\Omega_2)$ by γ_1 and γ_2. Let u_1 and u_2 be, respectively, two functions in $BV(\Omega_1)$ and $BV(\Omega_2)$ and define

$$u = \begin{cases} u_1 \text{ in } \Omega_1, \\ u_2 \text{ in } \Omega_2. \end{cases}$$

Then u belongs to $BV(\Omega)$ and

$$Du = Du_1 \lfloor \Omega_1 + Du_2 \lfloor \Omega_2 + [u]\nu \mathcal{H}^{N-1} \lfloor \Gamma_{1,2},$$

where $[u] = \gamma_1(u_1) - \gamma_2(u_2)$ and $\nu(x)$ is the unit inner normal at x to $\Gamma_{1,2}$ considered as a part of the boundary of Ω_1 (see Figure 10.1). In particular, if $u \in W^{1,1}(\Omega \setminus \Gamma_{1,2})$

$$Du = \nabla u \ \mathcal{L}_N + [u]\nu \mathcal{H}^{N-1} \lfloor \Gamma_{1,2},$$

where ∇u is the gradient of u in $L^1(\Omega)$.

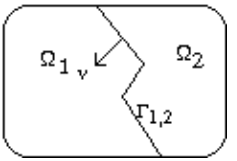


Figure 10.1. *The set* Ω

PROOF. For all $\varphi \in \mathbf{C}_c^1(\Omega, \mathbf{R}^N)$,

$$\langle Du, \varphi \rangle = -\int_\Omega u \ \text{div} \ \varphi \, dx = -\int_{\Omega_1} u_1 \ \text{div}\, \varphi \, dx - \int_{\Omega_2} u_2 \ \text{div}\, \varphi \, dx.$$

Since φ belongs to $\mathbf{C}^1(\overline{\Omega}_1, \mathbf{R}^N) \cap \mathbf{C}^1(\overline{\Omega}_2, \mathbf{R}^N)$, applying the generalized Green's formula in $BV(\Omega_1)$ and $BV(\Omega_2)$, we have

$$\begin{cases} \displaystyle\int_{\Omega_1} u_1 \,\text{div} \ \varphi \, dx = -\int_{\Omega_1} \varphi \ Du_1 + \int_{\Gamma_{1,2}} \gamma_1(u_1)\varphi.(-\nu) \ \mathcal{H}^{N-1}, \\ \displaystyle\int_{\Omega_2} u_2 \ \text{div} \ \varphi \, dx = -\int_{\Omega_2} \varphi \ Du_2 + \int_{\Gamma_{1,2}} \gamma_2(u_2)\varphi.\nu \ \mathcal{H}^{N-1}. \end{cases}$$

By summing these two equalities, we obtain

$$\langle Du, \varphi \rangle = \int_{\Omega_1} \varphi \ Du_1 + \int_{\Omega_2} \varphi \ Du_2 + \int_{\Gamma_{1,2}} \big(\gamma_1(u_1) - \gamma_2(u_2)\big)\varphi.\nu \mathcal{H}^{N-1}. \tag{10.20}$$

Assume now $||\varphi||_\infty \leq 1$. According to the continuity of γ_1 and γ_2, there exists a positive constant C depending only on Ω_1 and Ω_2 such that

$$\left| \int_{\Gamma_{1,2}} \big(\gamma_1(u_1) - \gamma_2(u_2)\big)\varphi.\nu \mathcal{H}^{N-1} \right| \leq C\big(\|u_1\|_{BV(\Omega_1)} + \|u_2\|_{BV(\Omega_2)}\big),$$

so that u belongs to $BV(\Omega)$ and

$$\|Du\| := \sup\{\langle Du, \varphi \rangle \ : \varphi \in \mathbf{C}_c(\Omega, \mathbf{R}^N), \ ||\varphi||_\infty \leq 1\} \leq C\big(\|u_1\|_{BV(\Omega_1)} + \|u_2\|_{BV(\Omega_2)}\big).$$

Finally, (10.20) can be written

$$\langle Du, \varphi \rangle = \int_\Omega \varphi \ \big(Du_1 \lfloor \Omega_1 + Du_2 \lfloor \Omega_2\big) + \int_\Omega \big(\gamma_1(u_1) - \gamma_2(u_2)\big)\varphi.\nu \mathcal{H}^{N-1} \lfloor \Gamma_{1,2}.$$

This shows that $Du = Du_1 \lfloor \Omega_1 + Du_2 \lfloor \Omega_2 + \big(\gamma_1(u_1) - \gamma_2(u_2)\big) \nu \mathcal{H}^{N-1} \lfloor \Gamma_{1,2}$ on $\mathbf{C}^1_c(\Omega, \mathbf{R}^N)$, and thus are equal by density. □

Example 10.2.2. Let us slightly modify the previous example by considering the function

$$v = \begin{cases} u \text{ in } \Omega, \\ 0 \text{ in } \mathbf{R}^N \setminus \overline{\Omega}, \end{cases}$$

where Ω is a Lipschitz domain of $\mathbf{R}^N$ and $u \in BV(\Omega)$. We see that v belongs to $BV(\mathbf{R}^N)$ and

$$Dv = Du \lfloor \Omega + u^+ \, \nu \, \mathcal{H}^{N-1} \lfloor \Gamma,$$

where Γ is the boundary of Ω, ν denotes the inner unit vector normal to Γ, and u^+ the trace of u on Γ. Thus, taking for instance u equal to the constant 1 in Ω, the characteristic function 1_Ω of Ω belongs to $BV(\mathbf{R}^N)$ (more precisely in a subspace $SBV(\mathbf{R}^N)$ introduced in Section 10.5) and

$$D1_\Omega = \nu \, \mathcal{H}^{N-1} \lfloor \Gamma.$$

This formula will be generalized in Section 10.3 when Ω possesses a *reduce boundary* $\partial_r \Omega$ (which in general does not coincides with the topological boundary) and a generalized unit normal ν to $\partial_r \Omega$. More precisely, for 1_Ω belonging to $BV(\mathbf{R}^N)$, we will obtain

$$D1_\Omega = \nu \, \mathcal{H}^{N-1} \lfloor \partial_r \Omega.$$

Example 10.2.3. *Jump through a family of hypersurfaces.* Let Ω be a Lipschitz domain of $\mathbf{R}^N$, u a function in $BV(\Omega)$, and $(\Sigma_t)_{t \in I}$ a family of Lipschitz hypersurfaces such that $\Sigma_t \subset \Omega$ is the boundary of a Lipschitz open bounded subset Ω_t of Ω with $t < t' \Longrightarrow \Omega_t \subset\subset \Omega_{t'}$. As a consequence of Example 10.2.1, one has for all but countably many t in I, the jumps $[u]_t$ of u through Σ_t are null.

Proof. Indeed, from Example 10.2.1,

$$\int_{\Sigma_t} |Du| = \int_{\Sigma_t} |[u]_t| \, d\mathcal{H}^{N-1},$$

and we conclude by Lemma 4.2.1. □

Example 10.2.4. Let Ω be a Lipschitz domain of $\mathbf{R}^N$ and $u \in BV(\Omega)$. For all $t > 0$, consider the Lipschitz domains $\Omega_t = \{x \in \Omega : d(x, \mathbf{R}^N \setminus \Omega) > t\}$ with boundary Γ_t and denote, respectively, by u_t^+ and u_t^- the traces of u when u is considered as an element of $BV(\Omega_t)$ or $BV(\Omega \setminus \overline{\Omega_t})$. We have for all but countably many t in $\mathbf{R}^+$: the traces $u_t^+(x)$ and $u_t^-(x)$ agree with $u(x)$ for $\mathcal{H}^{N-1}$ almost x in Γ_t.

Proof. By using arguments of the proof of Theorem 10.2, one may assume that Ω is a cylinder. With the notations of this theorem, for all t in a complementary of a countable subset of $\mathbf{R}^+$, estimate (10.13) becomes

$$\int_\Sigma |u^{t+\varepsilon} - u^t| \, d\mathcal{H}^{N-1} \leq C \int_{C_{R,t+\varepsilon,t}} |Du|.$$

We deduce, for these t, that $u^{t+\varepsilon}$ tends to u^t in $L^1(\Sigma)$ when ε goes to 0^+. On the other hand, according to the proof of the trace theorem, $u^{t+\varepsilon}$ tends to u_t^+ in $L^1(\Sigma_t)$ when ε goes to 0^+. Therefore, for $\mathcal{H}^{N-1}$ almost all x in Σ_t, $u_t^+ = u$. With the same arguments, but now reasoning on $C_{R,t,t+\varepsilon}$, we obtain that for $\mathcal{H}^{N-1}$ almost all x in Γ_t, $u_t^- = u$.

We are going to establish the continuity of the trace operator with respect to the intermediate convergence. Let us recall that the trace operator is continuous from $W^{1,p}(\Omega)$ into $L^p(\Gamma)$, $p \geq 1$ when these two spaces are equipped with their weak topology. Indeed, this is a consequence of the following property (cf. [90, Proposition III 9]): if T is a continuous linear operator from a Banach space E into a Banach space F, then T is continuous from E equipped with the $\sigma(E, E')$ topology into F equipped with the $\sigma(F, F')$ topology. We cannot apply this general principle to the trace operator defined in $BV(\Omega)$. Indeed, the weak toplogy in $BV(\Omega)$ is not the $\sigma(BV(\Omega), BV(\Omega)')$ topology and we must be careful with the terminology "weak" convergence in $BV(\Omega)$.

The example of Remark 10.1.3 shows that γ_0 is not continuous from $BV(\Omega)$ into $L^1(\Gamma)$ when $BV(\Omega)$ and $L^1(\Gamma)$ are equipped with their weak convergence: define $u_n \in BV(\Omega)$, $\Omega = (0,1)$ by $u_n(x) = nx$ if $x \in (0, \frac{1}{n}]$ and $u(x) = 1$ if $x \in [\frac{1}{n}, 1)$. It is easily seen that u_n weakly converges to the constant 1 in $BV(\Omega)$ whereas $u_n(x) = 0$ for $x \in \{0\}$. Note that in this example, u_n does not converge to $u = 1$ in the sense of intermediate convergence because there is " loss of total mass." Indeed $\int_\Omega |Du_n| = 1$ but $\int_\Omega |Du| = 0$. When the total mass is preserved at the limit, the trace operator is continuous.

Theorem 10.2.2. *Let Ω be a Lipschitz domain of $\mathbf{R}^N$. The trace operator γ_0 is continuous from $BV(\Omega)$ equipped with the intermediate convergence onto $L^1(\Gamma)$ equipped with the strong convergence.*

Proof. For each $t > 0$, consider the Lipschitz domain $\Omega_t = \{x \in \Omega : d(x, \mathbf{R}^N \setminus \Omega) > t\}$ with Lipschitz boundary Γ_t, and $(u_n)_{n\in\mathbf{N}}$ and u in $BV(\Omega)$ such that

$$\begin{cases} \displaystyle\int_\Omega |u_n - u \mid dx \to 0, \\ \displaystyle\int_\Omega |Du_n| \to \int_\Omega |Du|. \end{cases} \tag{10.21}$$

Possibly passing to a subsequence on n, and for almost all $t \in \mathbf{R}^+$, we have

$$\begin{cases} \displaystyle\int_{\Gamma_t} |Du| = 0, \\ u_n - u = (u_n - u)_t \text{ for } \mathcal{H}^{N-1} \text{ a.e. } x \text{ on } \Gamma_t, \\ \displaystyle\lim_{n\to+\infty} \int_{\Gamma_t} |u_n - u| \, d\mathcal{H}^{N-1} \to 0, \end{cases} \tag{10.22}$$

where $(u_n - u)_t$ denotes the trace on Γ_t of the function $u_n - u$ in $BV(\Omega_t)$. Indeed, the two first assertions are satisfied for all but countably many t in $\mathbf{R}^+$, thanks to Lemma 4.2.1 and to Example 10.2.4. The last assertion is a consequence of the curvilinear Fubini theorem

(cf. Corollary 4.2.2):

$$\int_0^{+\infty}\int_{\Gamma_t}|u_n-u|\,d\mathcal{H}^{N-1}\,dt=\int_0^{+\infty}\int_{[d(x,\mathbf{R}^N\setminus\overline{\Omega})=t]}|u_n-u|\,d\mathcal{H}^{N-1}\,dt$$
$$=\int_\Omega|u_n-u|\,dx.$$

For a fixed t, for which these assertions hold, let us define the function $u_{n,t}$ in $BV(\Omega)$ by

$$u_{n,t}=\begin{cases}u_n-u \text{ in } \Omega\setminus\overline{\Omega}_t,\\ 0 \text{ in } \Omega_t.\end{cases}$$

We have

$$Du_{n,t}=D(u_n-u)\lfloor(\Omega\setminus\overline{\Omega}_t)+(u_n-u)_t\,\nu_t\,\mathcal{H}^{N-1}\lfloor\Gamma_t,$$

where ν_t is the outer unit normal vector to Γ_t. According to the strong continuity of the trace operator γ_0 from $BV(\Omega)$ onto $L^1(\Gamma)$, we finally deduce

$$\int_\Gamma|\gamma_0(u_n-u)|\,d\mathcal{H}^{N-1}=\int_\Gamma|\gamma_0(u_{n,t})|\,d\mathcal{H}^{N-1}$$
$$\le C\Big(\int_{\Omega\setminus\Omega_t}|u_n-u|\,dx+\int_{\Omega\setminus\overline{\Omega}_t}|Du_n-Du|\,dx$$
$$+\int_{\Gamma_t}|u_n-u|\,d\mathcal{H}^{N-1}\Big).$$

Letting $n\to\infty$, (10.21) and (10.22) yield

$$\limsup_{n\to+\infty}\int_\Gamma|\gamma_0(u_n-u)|\,d\mathcal{H}^{N-1}\le 2C\int_{\Omega\setminus\overline{\Omega}_t}|Du|\,dx.$$

We have used the narrow convergence of $|Du_n|$ to $|Du|$ and $\int_{\Gamma_t}|Du|=0$ to assert that $\int_{\Omega\setminus\overline{\Omega}_t}|Du_n|$ tends to $\int_{\Omega\setminus\overline{\Omega}_t}|Du|$ (see Proposition 4.2.5). We complete the proof by letting t go to zero. □

10.3 The coarea formula and the structure of BV functions

It is well known that each real-valued function u of bounded variation on an interval I of $\mathbf{R}$ is the difference between two monotonous functions and, consequently, possesses at every point $x_0\in I$, two limits $u(x_0-0)$ and $u(x_0+0)$. One can then define its jump set $S_u:=\{x\in\Omega: u(x-0)\neq u(x+0)\}$. The main objective of this section is to establish that one can associate a set S_u to each u in $BV(\Omega)$, which generalize in any dimension the jump set of u in one dimension. In the particular case of a simple function $u=\chi_\Omega\in BV(\Omega)$, we will see that S_u is a part $\partial_M\Omega$ of the topological boundary $\partial\Omega$, which may differ from $\partial\Omega$ of a set of null $\mathcal{H}^{N-1}$ measure. The structure theorem, Theorem 10.3.4, will be a straightforward consequence of this property thanks to the generalized coarea formula in Section 10.3.3. For a first reading, the reader is advised to go directly to the definitions of the approximate limit sup and approximate limit inf (Definition 10.3.4) and to Theorem 10.3.4.

10.3.1 Notion of density and regular points

In this subsection, we generalize the notions of interior, exterior of subsets of $\mathbf{R}^N$ as well as the notions of limit, continuity, and jump for measurable functions. Indeed, we intend to extend the expression $D\chi_\Omega = \nu\, \mathcal{H}^{N-1}\lfloor\Gamma$ previously obtained for Lipschitz domains Ω thanks to the theory of traces (see Example 10.2.3) to sets Ω whose characteristic function χ_Ω belongs to $BV(\mathbf{R}^N)$. Actually, for these sets, we will obtain $D\chi_\Omega = \nu\, \mathcal{H}^{N-1}\lfloor\partial_M\Omega$, where the generalized boundary $\partial_M\Omega$ (the measure theoretical boundary) is, up to a set of $\mathcal{H}^{N-1}$ measure zero, the set of all points that are neither in the generalized interior nor in the generalized exterior of Ω.

In what follows, $B_\rho(x_0)$ denotes the open ball of $\mathbf{R}^N$ with radius $\rho > 0$ and centered at x_0.

Definition 10.3.1. *Let E be a Borel subset of $\mathbf{R}^N$. A point x_0 in $\mathbf{R}^N$ is a density point of E iff*

$$\lim_{\rho\to 0} \frac{\mathcal{L}^N(B_\rho(x_0)\cap E)}{\mathcal{L}^N(B_\rho(x_0))} = 1.$$

A point x_0 is a rarefaction point of E iff

$$\lim_{\rho\to 0} \frac{\mathcal{L}^N(B_\rho(x_0)\cap E)}{\mathcal{L}^N(B_\rho(x_0))} = 0.$$

The set of all density points and all rarefaction points of E are respectively called measure theoretical interior and measure theoretical exterior of E and denoted by E_* and E^*.

Example 10.3.1. When O is an open subset of $\mathbf{R}^N$, it is easily seen that all the points of O are density points and all the points of $\mathbf{R}^N \setminus \overline{O}$ are rarefaction points. Nevertheless, if x_0 belongs to the boundary Γ of O, various situations may occur as shown in Figure 10.2.

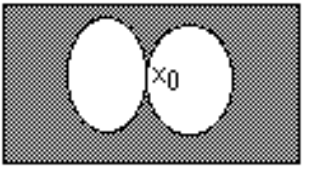


Figure 10.2. *The point x_0 is a density point of the union of the two discs but is a rarefaction point of the complementary of this union.*

We generalize these definitions relatively to a fixed Borel subset F of $\mathbf{R}^N$ as follows.

Definition 10.3.2. *Let F and E be two Borel subsets of $\mathbf{R}^N$ and assume that x_0 in $\mathbf{R}^N$ is such that $\mathcal{L}^N(B_\rho(x_0)\cap F) > 0$ for all $\rho > 0$ small enough. The point x_0 is an F-density point of E iff*

$$\lim_{\rho\to 0} \frac{\mathcal{L}^N(B_\rho(x_0)\cap F\cap E)}{\mathcal{L}^N(B_\rho(x_0)\cap F)} = 1.$$

The point x_0 is an F-rarefaction point of E iff

$$\lim_{\rho\to 0} \frac{\mathcal{L}^N(B_\rho(x_0)\cap F\cap E)}{\mathcal{L}^N(B_\rho(x_0)\cap F)} = 0.$$

These definitions allow us to adopt the following notion of boundary.

Definition 10.3.3. *Let E be a Borel subset of $\mathbf{R}^N$. The measure theoretical boundary of E is the subset of $\mathbf{R}^N$ denoted by $\partial_M E$, made up of all the elements of $\mathbf{R}^N$ which are neither density points nor rarefaction points of E.*

As a straightforward consequence of this definition, one can easily establish that the measure theoretical boundary is a part of the classical topological boundary as stated in the following proposition. The proof is left to the reader.

Proposition 10.3.1. *The measure theoretical boundary of E is the subset of the topological boundary ∂E defined by*

$$\partial_M E = \left\{ x\in\mathbf{R}^N : \limsup_{\rho\to 0} \frac{\mathcal{L}^N(B_\rho(x)\cap E)}{\mathcal{L}^N(B_\rho(x))} > 0 \text{ and } \limsup_{\rho\to 0} \frac{\mathcal{L}^N(B_\rho(x)\setminus E)}{\mathcal{L}^N(B_\rho(x))} > 0 \right\}.$$

Remark 10.3.1. The measure theoretical boundary may differ from the topological boundary of a set of nonnull $\mathcal{H}^{N-1}$ measure. Indeed, consider $E = \mathbf{B}\setminus[(0,0),(0,1)[$, where $\mathbf{B}$ is the unit open ball of $\mathbf{R}^2$. The measure theoretical boundary is the sphere but the (topological) boundary is the union of the sphere and the interval $[(0,0),(0,1)]$.

Let E be an open subset of $\mathbf{R}^N$ satisfying the following property: for all point x_0 of ∂E, there exists a normal vector $\nu(x_0)$ to ∂E such that E is included in the half-space $\pi_\nu(x_0) := \{x\in\mathbf{R}^N : \langle x-x_0, \nu(x_0)\rangle > 0\}$, where $\langle\ ,\ \rangle$ denotes the scalar product in $\mathbf{R}^N$. Then one has $\partial E = \partial_M E$. Indeed, it is easily seen that $\chi_{\frac{1}{\rho}(E-x_0)}$ strongly converges to the characteristic function $\chi_{\pi_\nu(x_0)}$ of $\pi_\nu(x_0)$ in $L^1_{loc}(\mathbf{R}^N)$ so that for x_0 in ∂E,

$$\begin{aligned}\lim_{\rho\to 0} \frac{\mathcal{L}^N(B_\rho(x_0)\cap E)}{\mathcal{L}^N(B_\rho(x_0))} &= \lim_{\rho\to 0} \frac{\mathcal{L}^N(B_1(x_0)\cap \frac{1}{\rho}(E-x_0)}{\mathcal{L}^N(B_1(x_0))}\\ &= 1/2 = \lim_{\rho\to 0} \frac{\mathcal{L}^N(B_\rho(x_0)\setminus(E-x_0))}{\mathcal{L}^N(B_\rho(x_0))}.\end{aligned}$$

We now look into the notion of approximate limit.

Definition 10.3.4. *Let $f : \mathbf{R}^N \longrightarrow \mathbf{R}$ be a measurable function and $x_0\in\mathbf{R}^N$. A real number α is the approximate limit of f at x_0 iff*

$$\forall\varepsilon>0,\ x_0 \text{ is a density point of the set } [\,|f-\alpha|<\varepsilon\,]$$

or equivalently

$$\forall \varepsilon > 0,\ x_0 \text{ is a rarefaction point of the set } [\ |f-\alpha| > \varepsilon\ ].$$

We then write $\alpha = ap\lim_{x\to x_0} f(x)$.

More generally, we define in $\overline{\mathbf{R}}$ *the approximate limit sup and approximate limit inf of* f *at* x_0 *by*

$$ap\limsup_{x\to x_0} f(x) = \inf\left\{ t\in\mathbf{R} : \lim_{\rho\to 0} \frac{\mathcal{L}^N(B_\rho(x_0)\cap[f>t])}{\mathcal{L}^N(B_\rho(x_0))} = 0 \right\},$$

$$ap\liminf_{x\to x_0} f(x) = \sup\left\{ t\in\mathbf{R} : \lim_{\rho\to 0} \frac{\mathcal{L}^N(B_\rho(x_0)\cap[f<t])}{\mathcal{L}^N(B_\rho(x_0))} = 0 \right\}.$$

Let F *be a fixed Borel subset of* $\mathbf{R}^N$. *A real number* α *is called the* F*-approximate limit of* f *at* x_0 *iff*

$$\forall \varepsilon > 0,\ x_0 \text{ is an } F\text{-density point of } [\ |f-\alpha| < \varepsilon\ ]$$

and we write $\alpha = ap\lim_{x\to x_0, x\in F} f(x)$.

Example 10.3.2. If $\alpha = \lim_{x\to x_0} f(x)$, it is straightforward to show that $\alpha = \mathrm{ap}\lim_{x\to x_0} f(x)$. When f is the characteristic function $\chi_{D_1\cup D_2}$ of the union of the two discs in Example 10.3.1, $[\ |f-1| < \varepsilon\ ] = D_1 \cup D_2$, so that $\mathrm{ap}\lim_{x\to x_0} f(x) = 1$ although the classical limit does not exist. When f is the characteristic function of the complementary of the union of these two discs, the approximate limit of f at x_0 is zero.

It is easy to establish uniqueness of the approximate limit when it exists. For further details related to these notions, see [156]. We only give, without proof, four elementary useful properties.

Proposition 10.3.2. *For all Borel subsets* A *and* B *of* $\mathbf{R}^N$, *the four following assertions hold:*

(i) *Let* $C = A\cup B$. *If the approximate limits*

$$ap\lim_{x\to x_0,\ x\in A} f(x) \text{ and } ap\lim_{x\to x_0,\ x\in B} f(x)$$

exist and coincide, then $ap\lim_{x\to x_0,\ x\in C} f(x)$ *exists.*

(ii) *If* x_0 *is not a rarefaction point of* A, $A\subset B$, *and if* x_0 *is a* B*-rarefaction point of* $B\setminus A$, *then the existence of*

$$ap\lim_{x\to x_0,\ x\in A} f(x)$$

implies the existence of

$$ap\lim_{x\to x_0,\ x\in B} f(x).$$

(iii) *If $A \subset B$ and if x_0 is not a rarefaction point of A, then the existence of*

$$ap \lim_{x \to x_0,\ x \in B} f(x)$$

implies the existence of

$$ap \lim_{x \to x_0,\ x \in A} f(x).$$

(iv) *If x_0 is not a rarefaction point of $A \cap B$, then the existence of*

$$ap \lim_{x \to x_0,\ x \in A} f(x) \text{ and } ap \lim_{x \to x_0,\ x \in B} f(x)$$

implies equality of these two approximate limits.

In the following proposition, we show that the approximate limit at x_0 is a classical limit for the restriction of f to a suitable Borel set. Moreover, when the approximate lim inf and approximate lim sup coincide, their common value is the approximate limit.

Proposition 10.3.3. *A measurable function $f : \mathbf{R}^N \longrightarrow \mathbf{R}$ possesses an approximate limit α at x_0 iff there exists a Borel subset B of $\mathbf{R}^N$ such that x_0 is a rarefaction point of $\mathbf{R}^N \setminus B$ and such that the restriction $f \lfloor B$ of f to B possesses the classical limit α at x_0. On the other hand, one always has*

$$ap \liminf_{x \to x_0} f \leq ap \limsup_{x \to x_0} f$$

in $\overline{\mathbf{R}}$ and $ap \liminf_{x \to x_0} f = ap \limsup_{x \to x_0} f := \alpha \in \mathbf{R}$ if and only if $ap \lim_{x \to x_0} f = \alpha$.

Proof. Let us prove the first assertion. It is easily seen that the given condition is sufficient. Conversely, assume that f possesses an approximate limit α at x_0. Without loss of generality one may assume $\alpha = 0$. For all integer i, x_0 is then a rarefaction point of

$$A_i := \mathbf{R}^N \setminus \{x \in \mathbf{R}^N : |f(x)| < 1/i\}.$$

Consider a nonincreasing sequence $\rho_1 > \rho_2 > \ldots > \rho_i > \ldots$ in $\mathbf{R}^+$ such that

$$\frac{\mathcal{L}^N(B_\rho(x_0) \cap A_i)}{\mathcal{L}^N(B_\rho(x_0))} \leq 2^{-i}$$

when $0 \leq \rho \leq \rho_i$, and denote the complementary set of $\cup_{i \in \mathbf{N}^*}(A_i \cap B_{\rho_i}(x_0))$ by B. It is straightforward to show that $f \lfloor B$ has the limit zero at x_0. We now show that x_0 is a rarefaction point of $\cup_{i \in \mathbf{N}^*}(A_i \cap B_{\rho_i}(x_0))$. Let $\rho_i > \rho > \rho_{i+1}$. Then

$$\begin{aligned} \mathcal{L}^N(B_\rho(x_0) \cap (\mathbf{R}^N \setminus B)) &\leq \mathcal{L}^N(B_\rho(x_0) \cap A_i) + \sum_{k=1}^{\infty} \mathcal{L}^N(B_{\rho_{i+k}}(x_0) \cap A_{i+k}) \\ &\leq C \rho^N \, 2^{-(i-1)}. \end{aligned}$$

We conclude the proof of the assertion by letting ρ go to zero and then i go to $+\infty$ in inequality

$$\frac{\mathcal{L}^N(B_\rho(x_0)\cap(\mathbf{R}^N\setminus B)}{\mathcal{L}^N(B_\rho(x_0))}\le C2^{-(i-1)}.$$

We are going to establish the second assertion. Assume that $\operatorname{ap\,lim\,inf}_{x\to x_0} f(x) = \operatorname{ap\,lim\,sup}_{x\to x_0} f(x)=\alpha$. Let $\varepsilon>0$, $t_\varepsilon,\ t'_\varepsilon$ be such that $\alpha\le t_\varepsilon<\alpha+\varepsilon,\ \alpha-\varepsilon<t'_\varepsilon\le\alpha$ and

$$\lim_{\rho\to0}\frac{\mathcal{L}^N(B_\rho(x_0)\cap[f>t_\varepsilon])}{\mathcal{L}^N(B_\rho(x_0))}=0,\ \lim_{\rho\to0}\frac{\mathcal{L}^N(B_\rho(x_0)\cap[f<t'_\varepsilon])}{\mathcal{L}^N(B_\rho(x_0))}=0.$$

Since

$$[|f-\alpha|>\varepsilon]=[f>\alpha+\varepsilon]\cup[f<\alpha-\varepsilon]$$

and

$$[f>\alpha+\varepsilon]\subset[f>t_\varepsilon],\ [f<\alpha-\varepsilon]\subset[f<t'_\varepsilon],$$

we infer

$$\lim_{\rho\to0}\frac{\mathcal{L}^N(B_\rho(x_0)\cap[|f-\alpha|>\varepsilon]}{\mathcal{L}^N(B_\rho(x_0))}=0,$$

and thus $\alpha=\operatorname{ap\,lim}_{x\to x_0} f(x)$. Conversely, assume that $\alpha=\operatorname{ap\,lim}_{x\to x_0} f(x)$, i.e., for all $\varepsilon>0$

$$\lim_{\rho\to0}\frac{\mathcal{L}^N(B_\rho(x_0)\cap[|f-\alpha|>\varepsilon]}{\mathcal{L}^N(B_\rho(x_0))}=0.$$

We deduce that $\forall\varepsilon>0$

$$\lim_{\rho\to0}\frac{\mathcal{L}^N(B_\rho(x_0)\cap[f>\alpha+\varepsilon])}{\mathcal{L}^N(B_\rho(x_0))}=0\quad\text{and}\quad\lim_{\rho\to0}\frac{\mathcal{L}^N(B_\rho(x_0)\cap[f<\alpha-\varepsilon])}{\mathcal{L}^N(B_\rho(x_0))}=0.$$

Therefore

$$\alpha+\varepsilon\le\operatorname{ap}\liminf_{x\to x_0} f(x)\text{ and }\alpha-\varepsilon\ge\operatorname{ap}\limsup_{x\to x_0} f(x),$$

and thus

$$\operatorname{ap}\liminf_{x\to x_0} f(x)=\operatorname{ap}\limsup_{x\to x_0} f(x),$$

provided that we have established $\operatorname{ap\,lim\,inf}_{x\to x_0} f(x)\le\operatorname{ap\,lim\,inf}_{x\to x_0} f(x)$.

Let us show $\operatorname{ap\,lim\,inf}_{x\to x_0} f(x)\le\operatorname{ap\,lim\,sup}_{x\to x_0} f(x)$. One may assume $\operatorname{ap\,lim\,sup}_{x\to x_0} f(x)<+\infty$. Let any $\tau\in\mathbf{R}$ be such that

$$\lim_{\rho\to0}\frac{\mathcal{L}^N(B_\rho(x_0)\cap[f<\tau])}{\mathcal{L}^N(B_\rho(x_0))}=0.$$

We claim that $\tau\le\operatorname{ap\,lim\,sup}_{x\to x_0} f(x)$. Assume, on the contrary, that $\tau>\operatorname{ap\,lim\,sup}_{x\to x_0} f(x)$ and set $\varepsilon=\tau-\operatorname{ap\,lim\,sup}_{x\to_0} f(x)$. We deduce that there exists t_ε such that $\tau>t_\varepsilon$ and

$$\lim_{\rho\to0}\frac{\mathcal{L}^N(B_\rho(x_0)\cap[f>t_\varepsilon])}{\mathcal{L}^N(B_\rho(x_0))}=0.$$

The inclusion $[f \geq \tau] \subset [f > t_\varepsilon]$ then yields

$$\lim_{\rho \to 0} \frac{\mathcal{L}^N(B_\rho(x_0) \cap [f \geq \tau])}{\mathcal{L}^N(B_\rho(x_0))} = 0$$

which is in contradiction with

$$\lim_{\rho \to 0} \frac{\mathcal{L}^N(B_\rho(x_0) \cap [f \geq \tau])}{\mathcal{L}^N(B_\rho(x_0)} = 1. \quad \square$$

Remark 10.3.2. Every measurable function $f : \mathbf{R}^N \longrightarrow \mathbf{R}$ possesses an approximate limit almost everywhere. For a proof, consult Morgan [184].

Let f be a function in $L^1_{loc}(\mathbf{R}^N)$. Then each Lebesgue point of f is a point of approximate limit. This property is indeed the straightforward consequence of

$$\frac{\mathcal{L}^N(B_\rho(x_0) \cap [|f - f(x_0)| > \varepsilon])}{\mathcal{L}^N(B_\rho(x_0)} \leq \frac{1}{\varepsilon\, \mathcal{L}^N(B_\rho(x_0))} \int_{B_\rho(x_0)} |f(x) - f(x_0)|\, dx.$$

Nevertheless, there exist points of approximate limit which are not Lebesgue points. Consult for instance, Morgan [184, Exercise 2.7].

In the next sections, for all functions f in $L^1(\Omega)$, Ω open subset of $\mathbf{R}^N$, we will adopt the following convention: we choose a representative of f, still denoted f, such that at every point x_0 of approximate limit, $f(x_0) = ap \lim_{x \to x_0} f(x)$. Such a representative is said to be approximately continuous at its points of approximate limit.

We now generalize the concept of left and right limits $u(x_0 - 0)$ and $u(x_0 + 0)$ for functions defined on $\mathbf{R}^N$. We denote the unit sphere of $\mathbf{R}^N$ by S^{N-1} and for all a in S^{N-1} and all x_0 in $\mathbf{R}^N$, $\pi_a(x_0)$ denotes the open half-space

$$\pi_a(x_0) := \{x \in \mathbf{R}^N : \langle x - x_0, a \rangle > 0\}.$$

We also denote the hemiball $\pi_a(x_0) \cap B_\rho(x_0)$ by $H_{\rho,a}(x_0)$.

Definition 10.3.5. *A point x_0 in $\mathbf{R}^N$ is called a regular point for the measurable function $f : \mathbf{R}^N \to \mathbf{R}$ iff there exists $a \in S^{N-1}$ such that the two following approximate limits exist:*

$$f_a(x) := ap \lim_{x \to x_0,\ x \in \pi_a(x_0)} f(x) \text{ and } f_{-a}(x) := ap \lim_{x \to x_0,\ x \in \pi_{-a}(x_0)} f(x).$$

Example 10.3.3. Let us consider Example 10.3.1 and take for f the characteristic function of the union of the two discs. The point x_0 is a regular point of f and satisfies $1 = f_a(x_0) = f_{-a}(x_0)$, where a is one of the two unit vectors orthogonal to the common tangent hyperplan at the two discs at x_0. One could say that x_0 is a point of approximate continuity for f. Let us point out that we also have $1 = ap \lim_{x \to x_0} f(x)$.

Consider now the characteristic function f of one of the two discs. The point x_0 is also a regular point of f and satisfies $f_a(x_0) \neq f_{-a}(x_0)$.

The following theorem asserts that regular points always satisfy the alternative of Example 10.3.3.

Theorem 10.3.1 (structure of the set of regular points, jump points, and jump sets). *Let x_0 be a regular point of a measurable function $f : \mathbf{R}^N \to \mathbf{R}$. We have the following alternative:*

(i) *If $f_a(x_0) = f_{-a}(x_0)$, then $ap\lim_{x\to x_0} f(x)$ exists and for all b in S^{N-1}, $f_b(x_0) = ap\lim_{x\to x_0} f(x)$. The point x_0 is called a point of approximate continuity of f.*

(ii) *If $f_a(x_0) \neq f_{-a}(x_0)$, then $\pm a$ is the unique element of S^{N-1} such that f_a and f_{-a} exist. The point x_0 is called a jump point of f. The real number $|f_a(x_0) - f_{-a}(x_0)|$ is the jump of f at x_0 and $\big(f_a(x_0) - f_{-a}(x_0)\big)a$ is the oriented jump. The set of all jump points of f is called the jump set of f and denoted by S_f.*

Proof. Assume that $f_a(x_0) = f_{-a}(x_0)$. The existence of $ap\lim_{x\to x_0} f(x)$ is a consequence of Proposition 10.3.2(i), with $A = \pi_a(x_0)$ and $B = \pi_{-a}(x_0)$. Thus, according to Proposition 10.3.2(iii), with $B = \mathbf{R}^N$ and $A = \pi_b(x_0)$, $f_b(x_0)$ exists for all $b \in S^{N-1}$. According to Proposition 10.3.2(iv), we obtain $f_b(x_0) = ap\lim_{x\to x_0} f(x)$.

We finally establish (ii). Let b be any element of S^{N-1}. We show that $f_b(x_0)$ does not exist if $b \neq \pm a$. Otherwise, as x_0 is not a rarefaction point of the sets $\pi_a(x_0) \cap \pi_b(x_0)$ and $\pi_{-a}(x_0) \cap \pi_b(x_0)$, from Proposition 10.3.2(iv), $f_a(x_0) = f_b(x_0)$ and $f_{-a}(x_0) = f_b(x_0)$, a contradiction. □

Example 10.3.4. Let us consider the function f defined by

$$f = \begin{cases} \alpha \text{ if } |x| \leq 1, \\ \beta \text{ if } |x| > 1, \end{cases}$$

where $\alpha \neq \beta$. The jump set is the sphere S^{N-1}. In Example 10.3.1, if $f = \chi_{D_1 \cup D_2}$, the jump set is the boundary (topological) except the point x_0.

The next proposition characterizes the jump set of simple functions.

Proposition 10.3.4 (inner measure theoretic normal). *Let E be a Borel subset of $\mathbf{R}^N$ and χ its characteristic function. The jump set S_χ of χ is the set of all points x_0 of $\mathbf{R}^N$ for which there exists a in S^{N-1} satisfying*

$$\begin{cases} \lim\limits_{\rho\to 0} \dfrac{\mathcal{L}^N\big(H_{\rho,a}(x_0) \cap E\big)}{\mathcal{L}^N\big(H_{\rho,a}(x_0)\big)} = 1, \\ \lim\limits_{\rho\to 0} \dfrac{\mathcal{L}^N\big(H_{\rho,-a}(x_0) \cap E\big)}{\mathcal{L}^N\big(H_{\rho,a}(x_0)\big)} = 0. \end{cases}$$

For such points, the unit vector a is unique and called the inner measure theoretic normal to E at x_0. Moreover, $S_\chi \subset \partial_M E$.

Proof. From the definition, every point satisfying the two above properties is a jump point of χ, and the uniqueness of a is a consequence of Theorem 10.3.1(ii).

Conversely, if x_0 is a jump point of χ, there exists a unique $\pm a$ in S^{N-1} such that $\chi_a(x_0) \neq \chi_{-a}(x_0)$. It is easily seen that $\chi_a(x_0)$ and $\chi_{-a}(x_0)$ belong to $\{0, 1\}$. Then,

exchanging, if necessary, a and $-a$, we have $\chi_a(x_0) = 1$ and $\chi_{-a}(x_0) = 0$, which gives the two required limits.

We must now prove $S_{\chi_E} \subset \partial_M E$. Let $x_0 \in S_{\chi_E}$. One has

$$\begin{aligned}\frac{\mathcal{L}^N(B_\rho(x_0) \cap E)}{\mathcal{L}^N(B_\rho(x_0))} &= \frac{\mathcal{L}^N(B_\rho(x_0) \cap E)}{\mathcal{L}^N(B_\rho(x_0) \cap E \cap \pi_a(x_0))} \\ &\times \frac{\mathcal{L}^N(B_\rho(x_0) \cap E \cap \pi_a(x_0))}{\mathcal{L}^N(B_\rho(x_0) \cap \pi_a(x_0))} \\ &\times \frac{\mathcal{L}^N(B_\rho(x_0) \cap \pi_a(x_0))}{\mathcal{L}^N(B_\rho(x_0))} \\ &\geq \frac{1}{2} \frac{\mathcal{L}^N(B_\rho(x_0) \cap E \cap \pi_a(x_0))}{\mathcal{L}^N(B_\rho(x_0) \cap \pi_a(x_0))}.\end{aligned}$$

Since the second factor tends to 1 when $\rho \to 0$, the inequality above yields

$$\limsup_{\rho \to 0} \frac{\mathcal{L}^N(B_\rho(x_0) \cap E)}{\mathcal{L}^N(B_\rho(x_0))} \geq \frac{1}{2} > 0.$$

Exchanging the roles of a and $-a$ and E and $\mathbf{R}^N \setminus E$, we also obtain

$$\limsup_{\rho \to 0} \frac{\mathcal{L}^N(B_\rho(x_0) \setminus E)}{\mathcal{L}^N(B_\rho(x_0))} \geq \frac{1}{2} > 0;$$

thus, according to Proposition 10.3.1, $x_0 \in \partial_M E$. □

10.3.2 Sets of finite perimeter, structure of simple BV functions

To clarify the structure of BV functions, we now establish that up to a set of $\mathcal{H}^{N-1}$ measure zero, the jump set of all simple function χ_E which belongs to $BV(\mathbf{R}^N)$ is essentially the measure theoretical boundary of E.

Definition 10.3.6. *A Borel subset E of $\mathbf{R}^N$ is called a set of finite perimeter in Ω iff its characteristic function χ_E belongs to $BV(\Omega)$. The total mass $|D_{\chi_E}|(\Omega)$ is called the perimeter of E in Ω and is denoted by $P(E, \Omega)$ or $P(E)$ when $\Omega = \mathbf{R}^N$. A Borel subset E of $\mathbf{R}^N$ is called a set of locally finite perimeter if it is a set of finite perimeter in U for all bounded open subset U of $\mathbf{R}^N$.*

Remark 10.3.3. When E is a Lipschitz open bounded subset of Ω, according to the trace Theorem 10.2.1 and to Example 10.2.2, we have $P(E, \Omega) = \mathcal{H}^{N-1}(\Omega \cap \partial E)$.

Theorem 10.3.2 (structure of simple functions of $BV(\Omega)$). *Let E be a set of finite perimeter in Ω. Then,*

(i) *up to a Borel subset of $\mathcal{H}^{N-1} \lfloor \Omega$-zero measure, $\partial_M E \cap \Omega$ is the jump set of χ_E;*

(ii) *the set $\partial_M E \cap \Omega$ is countably $N-1$-rectifiable, i.e., $\partial_M E \cap \Omega \subset \bigcup_{i \in \mathbf{N}} A_n$, where $\mathcal{H}^{N-1} \lfloor \Omega(A_0) = 0$ and for each $i = 1, \ldots, +\infty$ there exists a Lipschitz function $f_i : \mathbf{R}^{N-1} \to \mathbf{R}^N$ such that $A_i = f_i(\mathbf{R}^{N-1})$;*

(iii) *the following generalized Gauss–Green formula holds: for $\mathcal{H}^{N-1}$ almost all x in Ω, there exists $\nu(x) \in S^{N-1}$, called the generalized inner normal vector to E at x, such that for all φ in $\mathbf{C}^1_c(\Omega, \mathbf{R}^N)$,*

$$\int_\Omega \chi_E \ \text{div}\ \varphi \, dx = \int_{\partial_M E \cap \Omega} \varphi.(-\nu) d\mathcal{H}^{N-1},$$

that is, $D\chi_E = \nu \, \mathcal{H}^{N-1} \lfloor \partial_M E \cap \Omega$.

PROOF. The proof is divided into the five following steps:

Step 1. We define the generalized inner normal vector $\nu(x)$ to E at x and the reduced boundary $\partial_r E$ of E.

Step 2. This step consists of establishing $\partial_r E \subset S_{\chi_E} \cap \Omega \subset \partial_M E \cap \Omega$.

Step 3. We prove that $\partial_M E \cap \Omega$, $S_{\chi_E} \cap \Omega$, and $\partial_r E$ are essentially the same sets, more precisely, $\mathcal{H}^{N-1}(\partial_M E \cap \Omega \setminus \partial_r E) = 0$.

Step 4. We establish that $S_{\chi_E} \cap \Omega$ is countably $(N-1)$-rectifiable.

Step 5. We prove the generalized Gauss–Green formula.

Step 1. Contrary to the previous notions of boundary $\partial_M E$ and S_{χ_E}, the definition below is specifically defined for subsets E of Ω such that $D\chi_E$ belongs to $\mathbf{M}(\Omega, \mathbf{R}^N)$.

Definition 10.3.7. *Let E be a subset of finite perimeter in Ω. The generalized unit inner normal to E is the Radon–Nikodym derivative of $D\chi_E$ with respect to the measure $|D\chi_E|$. In other words,*

$$\text{for } |D\chi_E| \ a.e.\ x \text{ in } \Omega, \ \ \nu(x) = \lim_{\rho \to 0} \frac{\int_{B_\rho(x)} D\chi_E}{\int_{B_\rho(x)} |D\chi_E|}.$$

The reduced boundary $\partial_r E$ consists of all points x in Ω such that the limit above exists.

Let us remark that according to the trace theory, when E is a Lipschitz domain of $\mathbf{R}^N$ with boundary Γ, we have $D\chi_E = \nu \mathcal{H}^{N-1} \lfloor \Gamma$, where $\nu(x)$ is the inner normal to E at $\mathcal{H}^{N-1}$ a.e x in Γ. Consequently $\frac{D\chi_E}{|D\chi_E|}(x) = \nu(x)$ for $\mathcal{H}^{N-1}$ a.e. x in Γ and $\mathcal{H}^{N-1}(\Gamma \setminus \partial_r E) = 0$.

Step 2. We establish that $\partial_r E \subset S_{\chi_E} \cap \Omega$. The inclusion $S_{\chi_E} \cap \Omega \subset \partial_M E \cap \Omega$ has been proved in Proposition 10.3.4. The key of the proof is the following blow-up lemma.

Lemma 10.3.1. *Let $x_0 \in \partial_r E$, $\nu(x_0)$ be the generalized unit inner normal to E at x_0, E_ρ the homothetic subset $\{x \in \mathbf{R}^N : \rho(x - x_0) \in E\}$ of E, and $\pi_\nu(x_0)$ the half-space $\{x \in \mathbf{R}^N : \langle x - x_0, \nu(x_0)\rangle > 0\}$. Then χ_{E_ρ} weakly converges to $\chi_{\pi_\nu(x_0)}$ in $BV(B_1(x_0))$.*

SKETCH OF THE PROOF OF LEMMA 10.3.1. We admit the three following estimates: for each $x \in \partial_r E$ there exits a positive constant C such that for all sufficiently small $r > 0$,

$$C \le r^{N-1} |D\chi_E|(B_r(x)) \le C^{-1}, \tag{10.23}$$

$$C \le r^{-N} \mathcal{L}^N(B_r(x) \cap E), \tag{10.24}$$

$$C \le r^{-N} \mathcal{L}^N(B_r(x) \setminus E). \tag{10.25}$$

For a proof, we refer the reader to [229, Lemma 5.5.4].

Without loss of generality, one may assume $x_0 = 0$ and $\nu(0) = (0, \dots, 0, 1)$. Moreover, it is enough to establish that, for each sequence $(\rho_h)_{h\in\mathbf{N}}$, there exists a subsequence (not relabeled) satisfying

$$\chi_{E_{\rho_h}} \rightharpoonup \chi_{\pi_\nu(0)} \text{ in } BV(B_1(x_0)).$$

Let us fix $r > 0$. Reasoning with a smooth approximating sequence in the sense of the intermediate convergence (Theorem 10.1.2), and changing scale, a straightforward calculation gives

$$D\chi_{E_{\rho_h}}(B_r(0)) = \rho_h^{1-N} D\chi_E(B_{\rho_h r}(0)), \tag{10.26}$$

$$|D\chi_{E_{\rho_h}}|(B_r(0)) = \rho_h^{1-N} |D\chi_E|(B_{\rho_h r}(0)). \tag{10.27}$$

Collecting (10.23) and (10.27), we obtain, for ρ_h small enough,

$$Cr^{N-1} \le |D\chi_{E_{\rho_h}}|(B_r(0)) \le C^{-1} r^{N-1}. \tag{10.28}$$

Thanks to (10.28) with $r = 1$, applying Theorem 10.1.4 in $BV(B_1(0))$, there exist a subsequence (not relabeled) and a subset $F \subset \mathbf{R}^N$ of finite perimeter in $B_1(0)$ such that

$$\chi_{E_{\rho_h}} \to \chi_F \text{ strongly in } L^1(B_1(0)), \tag{10.29}$$

$$D\chi_{E_{\rho_h}} \rightharpoonup D\chi_F \text{ weakly in } \mathbf{M}(B_1(0), \mathbf{R}^N). \tag{10.30}$$

From now on we reason in $BV(B_1(0))$. It remains to establish that $F = \pi_\nu(0)$. According to Theorem 4.2.1 and Lemma 4.2.1, for all but countably many $0 < \rho < 1$

$$D\chi_{E_{\rho_h}}(B_\rho(0)) \to D\chi_F(B_\rho(0)). \tag{10.31}$$

Thus, from (10.27), (10.26), according to the definition of $\nu(0) = (0, \dots, 0, 1)$, and from (10.31) we deduce

$$\begin{aligned}
\lim_{h\to+\infty} |D\chi_{E_{\rho_h}}|(B_\rho(0)) &= \lim_{h\to+\infty} \frac{|D\chi_{E_{\rho_h}}|(B_\rho(0))}{\frac{\partial}{\partial x_N}\chi_{E_{\rho_h}}(B_\rho(0))} \lim_{h\to+\infty} \frac{\partial}{\partial x_N}\chi_{E_{\rho_h}}(B_\rho(0)) \\
&= \lim_{h\to+\infty} \frac{|D\chi_E|(B_{\rho_h\rho}(0))}{\frac{\partial}{\partial x_N}\chi_E(B_{\rho_h\rho}(0))} \lim_{h\to+\infty} \frac{\partial}{\partial x_N}\chi_{E_{\rho_h}}(B_\rho(0)) \\
&= \lim_{h\to+\infty} \frac{\partial}{\partial x_N}\chi_{E_{\rho_h}}(B_\rho(0)) = \frac{\partial}{\partial x_N}\chi_F(B_\rho(0)).
\end{aligned} \tag{10.32}$$

The lower semicontinuity of the total variation (Proposition 10.1.1) and (10.30), (10.32) finally yield

$$|D\chi_F|(B_\rho(0)) \le \frac{\partial}{\partial x_N}\chi_F(B_\rho(0)),$$

hence equality. Let us consider the density ν_N of $\frac{\partial}{\partial x_N}\chi_F$ with respect to $|D\chi_F|$ (cf. Radon–Nikodym theorem, Theorem 4.2.1). From above one has

$$|D\chi_F|(B_\rho(0)) = \frac{\partial}{\partial x_N}\chi_F(B_\rho(0)) = \int_{B_\rho(0)} \nu_N(x)\, d|D\chi_F|(x).$$

This shows that $\nu_N(x) = 1$ for $|D\chi_F|$ a.e. $x \in B_1(0)$; hence $\frac{\partial}{\partial x_N}\chi_F = |D\chi_F|$ is a nonnegative Radon measure. Note also that from (10.32) and (10.28), $\frac{\partial}{\partial x_N}\chi_F(B_\rho(0)) \geq C\rho^{N-1}$ so that $\frac{\partial}{\partial x_N}\chi_F \not\equiv 0$.

On the other hand, from (10.31), (10.26), (10.27), and the definition of $\nu(0)$, for $i = 1 \dots N-1$, and for all but countably many $0 < \rho < 1$, we have

$$\frac{\frac{\partial}{\partial x_i}\chi_F(B_\rho(0))}{|D\chi_F|(B_\rho(0))} = \lim_{h\to+\infty} \frac{\frac{\partial}{\partial x_i}\chi_{E_{\rho_h}}(B_\rho(0))}{|D\chi_{E_{\rho_h}}|(B_\rho(0))} = \lim_{h\to+\infty} \frac{\frac{\partial}{\partial x_i}\chi_E(B_{\rho_h\rho}(0))}{|D\chi_E|(B_{\rho_h\rho}(0))} = 0. \tag{10.33}$$

We deduce from (10.33) that $\frac{\partial}{\partial x_i}\chi_F(B_\rho(0)) = 0$, for $i = 1, \dots, N-1$. Let us consider the density ν_i of $\frac{\partial}{\partial x_i}\chi_F$ with respect to $|D\chi_F|$. Since

$$0 = \frac{\partial}{\partial x_i}\chi_F(B_\rho(0)) = \int_{B_\rho(0)} \nu_i(x)\, d|D\chi_F|(x),$$

$\nu_i(x) = 0$ for $|D\chi_F|$ a.e. $x \in B_1(0)$, and hence $\frac{\partial}{\partial x_i}\chi_F = 0$.

The function χ_F depends only on the variable x_N, is nondecreasing, and, from (10.29), takes only the two values 0 and 1. Now set $\alpha = \sup\{x_N : \chi_F(x_N) = 0\}$. Note that $\alpha \neq +\infty$; otherwise $D\chi_F \equiv 0$. For proving $F = \pi_\nu(0)$, it suffices to establish $\alpha = 0$. Assuming $\alpha > 0$ gives $B_\rho(0) \subset \mathbf{R}^N \setminus F$ for $\rho < \alpha$ so that, by (10.29),

$$\begin{aligned} 0 = \mathcal{L}^N(B_\rho(0) \cap F) &= \lim_{h\to+\infty} \mathcal{L}^N(B_\rho(0) \cap E_{\rho_h}) \\ &= \lim_{h\to+\infty} \rho_h^{-N} \mathcal{L}^N(B_{\rho\rho_h}(0) \cap E), \end{aligned}$$

which contradicts (10.24). If $\alpha < 0$, then $B_\rho(0) \subset F$ for $\rho < -\alpha$ and a similar argument gives

$$0 = \lim_{h\to+\infty} \rho_h^{-N} \mathcal{L}^N(B_{\rho\rho_h}(0) \setminus E),$$

which contradicts (10.25). The proof of Lemma 10.3.1 is complete.

We are going to establish $\partial_r E \subset S_{\chi_E} \cap \Omega$. Let $x_0 \in \partial_r E$. We have

$$\frac{\mathcal{L}^N\,(E \cap H_{\rho,-\nu}(x_0))}{\mathcal{L}^N(H_{\rho,-\nu}(x_0))} = \frac{\mathcal{L}^N\,(E_\rho \cap B_1(x_0) \cap \pi_{-\nu}(x_0))}{\mathcal{L}^N(B_1(x_0) \cap \pi_{-\nu}(x_0))}.$$

The convergence of χ_{E_ρ} to $\chi_{\pi_\nu(x_0)}$ in $L^1(B_1(x_0))$ yields

$$\lim_{\rho\to 0} \frac{\mathcal{L}^N\,(E \cap H_{\rho,-\nu}(x_0))}{\mathcal{L}^N(H_{\rho,-\nu}(x_0))} = \frac{\mathcal{L}^N(\pi_\nu(x_0) \cap B_1(x_0) \cap \pi_{-\nu}(x_0))}{\mathcal{L}^N(B_1(x_0) \cap \pi_{-\nu}(x_0))} = 0.$$

Similarly,

$$\lim_{\rho\to 0} \frac{\mathcal{L}^N\,(E \cap H_{\rho,\nu}(x_0))}{\mathcal{L}^N(H_{\rho,\nu}(x_0))} = \frac{\mathcal{L}^N(\pi_\nu(x_0) \cap B_1(x_0) \cap \pi_\nu(x_0))}{\mathcal{L}^N(B_1(x_0) \cap \pi_\nu(x_0))} = 1$$

and, according to Proposition 10.3.4, $x_0 \in S_{\chi_E}$.

Step 3. The proof of $\mathcal{H}^{N-1}(\partial_M E \cap \Omega \setminus \partial_r E) = 0$ consists first in proving $\mathcal{H}^{N-1}(S_{\chi_E} \cap \Omega \setminus \partial_r E) = 0$; next, $\mathcal{H}^{N-1}(\partial_M E \cap \Omega \setminus S_{\chi_E} \cap \Omega) = 0$. Keys of the proof are the relative isoperimetric inequality and Lemma 4.2.3. This is summarized in the next lemma.

Lemma 10.3.2. *Let E be a set of finite perimeter in Ω. Then the following assertions hold:*

(i) *Relative isoperimetric inequality: there exists a positive constant C such that*

$$\min\{\mathcal{L}^N(B_r \cap E), \mathcal{L}^N(B_r \setminus E)\}^{\frac{N-1}{N}} \leq C|D\chi_E|(B_r)$$

for all open ball B_r with radius R included in Ω.

(ii) *There exists a positive constant C such that for all $x \in S_{\chi_E} \cap \Omega$,*

$$\liminf_{\rho \to 0} \frac{|D\chi_E|(B_\rho(x))}{\rho^{N-1}} \geq C.$$

(iii) *For $\mathcal{H}^{N-1}$ almost all $x \in \Omega \setminus S_{\chi_E} \cap \Omega$,*

$$\limsup_{\rho \to 0} \frac{|D\chi_E|(B_\rho(x))}{\rho^{N-1}} = 0.$$

(iv) $\mathcal{H}^{N-1}(S_{\chi_E} \cap \Omega \setminus \partial_r E) = 0$.

(v) $\mathcal{H}^{N-1}(\partial_M E \cap \Omega \setminus S_{\chi_E} \cap \Omega) = 0$.

Proof of Lemma 10.3.2. Assertion (i) is a straightforward consequence of the Poincaré–Wirtinger inequality

$$\left(\int_{B_r} |u - \overline{u}|^{\frac{N}{N-1}} \right)^{\frac{N-1}{N}} \leq C \int_{B_r} |Du|, \text{ where } \overline{u} = \frac{1}{\mathcal{L}^N(B_r)} \int_{B_r} u(x)\, dx,$$

applied to $u = \chi_E$. Indeed, every function u in $BV(\Omega)$ satisfies the Poincaré–Wirtinger inequality: consider a smooth approximating sequence in the sense of the intermediate convergence, and apply Corollary 5.4.1 and Theorem 10.1.2 as in the proof of Proposition 10.1.3.

Let us prove (ii). Let $x \in S_{\chi_E} \cap \Omega$; according to Theorem 10.3.4, an easy computation yields

$$\liminf_{\rho \to 0} \frac{\mathcal{L}^N(B_\rho(x) \cap E)}{\mathcal{L}^N(B_\rho(x))} \geq \frac{1}{2} \text{ and } \liminf_{\rho \to 0} \frac{\mathcal{L}^N(B_\rho(x) \setminus E)}{\mathcal{L}^N(B_\rho(x))} \geq \frac{1}{2}$$

so that

$$\lim_{\rho \to 0} \frac{\mathcal{L}^N(B_\rho(x) \cap E)}{\mathcal{L}^N(B_\rho(x))} = \lim_{\rho \to 0} \frac{\mathcal{L}^N(B_\rho(x) \setminus E)}{\mathcal{L}^N(B_\rho(x))} = \frac{1}{2}.$$

The conclusion of assertion (ii) then follows from (i).

For proving (iii), it suffices to establish that for all $\delta > 0$, the set

$$A_\delta = (\Omega \setminus S_{\chi_E}) \cap \left\{ x \in \Omega : \limsup_{\rho \to 0} \frac{|D\chi_E|(B_\rho(x))}{\rho^{N-1}} > \delta \right\}$$

satisfies $\mathcal{H}^{N-1}(A_\delta) = 0$. From Lemma 4.2.3 applied to the Borel measure $\mu = |D\chi_E|$, one has $|D\chi_E|(A_\delta) \geq C\delta \mathcal{H}^{N-1}(A_\delta)$ and the conclusion follows from $|D\chi_E|(A_\delta) = 0$ which is a consequence of $\partial_r E \subset S_{\chi_E} \cap \Omega$ established in Step 2.

We establish assertion (iv). From assertion (ii) and Lemma 4.2.3 applied to the Borel measure $\mu = |D\chi_E|$, there exists a nonnegative constant C' depending only on N, such that, for all Borel set B included in S_{χ_E},

$$\mathcal{H}^{N-1}(B) \leq C' \, |D\chi_E|(B). \tag{10.34}$$

Obviously, by definition

$$|D\chi_E|(\Omega \setminus \partial_r E) = 0;$$

thus

$$|D\chi_E|(S_{\chi_E} \cap \Omega \setminus \partial_r E) = 0,$$

and the conclusion $\mathcal{H}^{N-1}(S_{\chi_E} \cap \Omega \setminus \partial_r E) = 0$ follows from (10.34).

We finally establish $\mathcal{H}^{N-1}(\partial_M E \cap \Omega \setminus S_{\chi_E} \cap \Omega) = 0$. According to (iii), we are reduced to prove that for each x in $\partial_M E \cap \Omega$,

$$\limsup_{\rho \to 0} \frac{|D\chi_E|(B_\rho(x))}{\rho^{N-1}} > 0.$$

From the definition of the measure theoretical boundary (Proposition 10.3.1), there exists $\delta > 0$ such that

$$\limsup_{\rho \to 0} \frac{\mathcal{L}^N(B_\rho(x) \cap E)}{\mathcal{L}^N(B_\rho(x))} > \delta \text{ and } \liminf_{\rho \to 0} \frac{\mathcal{L}^N(B_\rho(x) \cap E)}{\mathcal{L}^N(B_\rho(x))} < 1 - \delta.$$

Consequently, choosing $\delta < 1/2$, for ρ small enough,

$$\delta < \frac{\mathcal{L}^N(B_\rho(x) \cap E)}{\mathcal{L}^N(B_\rho(x))} < 1 - \delta,$$

and, for such ρ, the relative isoperimetric inequality (i) yields

$$\left(\delta \mathcal{L}^N(B_\rho(x))\right)^{\frac{N-1}{N}} \leq C |D\chi_E|(B_\rho(x)),$$

which completes the proof of Lemma 10.3.2.

Step 4. We admit that $S_{\chi_E} \cap \Omega$ is countably $(N-1)$-rectifiable. For a proof, see Ziemer [229, paragraph 5.7, pp. 243–246]. Actually, thanks to Rademacher's theorem, one may assume more precisely that $S_{\chi_E} \cap \Omega \subset \cup_{i=0}^{+\infty} A_i$, where $\mathcal{H}^{N-1} \lfloor \Omega(A_0) = 0$ and A_i is a $N-1$ manifold of class C^1 for $i = 1, \ldots, +\infty$ (see Ziemer [229, Lemma 5.7.2]).

Step 5. We establish the generalized Gauss–Green formula. We begin by establishing

$$|D\chi_E|(A) = \mathcal{H}^{N-1}(A) \tag{10.35}$$

for all Borel set A included in $\partial_r E$. From (10.27), (10.32), and Lemma 10.3.1, for all $x \in \partial_r E$, and all but countably many $0 < r < 1$ we have

$$\begin{aligned}\lim_{\rho\to 0} \frac{|D\chi_E|(B_\rho(x))}{\rho^{N-1}} &= r^{1-N} \lim_{\rho\to 0} |D\chi_{E_{\frac{\rho}{r}}}|(B_r(x)) \\ &= r^{1-N} |D\chi_{\pi_\nu(x)}|(B_r(x)) \\ &= |D\chi_{\pi_\nu(x)}|(B_1(x)) \\ &= \mathcal{H}^{N-1}(B_1(x) \cap \pi_\nu(x)) \\ &= \omega_{N-1},\end{aligned}$$

where ω_{N-1} is the volume of the unit ball of $\mathbf{R}^{N-1}$. Thus

$$\forall x \in \partial_r E,\ \lim_{\rho\to 0} \frac{|D\chi_E|(B_\rho(x))}{\mathcal{H}^{N-1}(B_\rho(x))} = 1. \tag{10.36}$$

On the other hand, from Steps 3 and 4, one may assume that

$$A \subset \cup_{i=0}^{+\infty} A_i,$$

where $\mathcal{H}^{N-1} \lfloor \Omega(A_0) = 0$ and A_i is a $N-1$ manifold of class C^1. Since each set A_i is regular, one has

$$\forall x \in A \cap A_i,\ \lim_{\rho\to 0} \frac{\mathcal{H}^{N-1} \lfloor A_i(B_\rho(x))}{\mathcal{H}^{N-1}(B_\rho(x))} = 1$$

so that, from (10.36),

$$\forall x \in A \cap A_i,\ \lim_{\rho\to 0} \frac{\mathcal{H}^{N-1} \lfloor A_i(B_\rho(x))}{|D\chi_E|(B_\rho(x))} = 1,$$

which yields

$$\mathcal{H}^{N-1}(A \cap A_i) = |D\chi_E|(A \cap A_i)$$

and finally gives $|D\chi_E|(A) = \mathcal{H}^{N-1}(A)$.

We are now in a position to prove the generalized Gauss–Green formula. Let $\varphi \in \mathbf{C}_c^1(\Omega, \mathbf{R}^N)$, from (10.35) and the definition of $\partial_r E$ and ν, one has

$$\begin{aligned}\int_\Omega \chi_E \ \text{div}\ \varphi\, dx &= -\int_\Omega \varphi\ D\chi_E \\ &= \int_{\partial_r E} \varphi.(-\nu)\ |D\chi_E| \\ &= \int_{\partial_r E} \varphi.(-\nu)\, d\mathcal{H}^{N-1}\end{aligned}$$

and the conclusion follows from Step 3. □

10.3.3 Structure of BV functions

The functions of $BV(\Omega)$ naturally inherit their properties from their level sets $[u > t] := \{x \in \Omega : u(x) > t\}$ when t varies in $\mathbf{R}$. The following property, established by Fleming and Richel in [136], generalizes the classical coarea formula (Theorem 4.2.5) to BV functions and states that for almost every t in $\mathbf{R}$, the level set $[u > t]$ of each BV function u has a finite perimeter in Ω. Consequently, we show that the jump set of $u \in BV(\Omega)$ inherits its structure from the one of the jump set of finite perimeter sets $[u > t]$, $t \in \mathbf{R}$, stated in Theorem 10.3.2.

Theorem 10.3.3 (coarea formula). *Let u be a given function in $BV(\Omega)$. Then, for a.e. t in $\mathbf{R}$, the level set $E_t = \{x \in \Omega : u(x) > t\}$ of u is a set of finite perimeter in Ω, and*

$$\begin{cases} Du = \displaystyle\int_{-\infty}^{+\infty} D\chi_{E_t}\, dt, \\ |Du|(\Omega) = \displaystyle\int_{-\infty}^{+\infty} \int_{\Omega} |D\chi_{E_t}| dt. \end{cases}$$

More generally for all Borel function $f : \Omega \to \mathbf{R}^+$,

$$\int_{\Omega} f|Du| = \int_{-\infty}^{+\infty} \int_{\Omega} f|D\chi_{E_t}| dt.$$

Proof. Let us assume for the moment that for a.e. t in $\mathbf{R}$, $D\chi_{E_t}$ belongs to $\mathbf{M}(\Omega, \mathbf{R}^N)$. For all t in $\mathbf{R}$, set

$$f_t = \begin{cases} \chi_{E_t} \text{ if } t \geq 0, \\ -\chi_{\Omega \setminus E_t} \text{ if } t < 0. \end{cases}$$

It is easily seen that for all x in Ω, $u(x) = \int_{-\infty}^{+\infty} f_t(x)\, dt$. For all φ in $\mathbf{C}_c^1(\Omega, \mathbf{R}^N)$ we have

$$\begin{aligned} < Du, \varphi > &= -\int_{\Omega} u \ \text{div} \ \varphi \, dx \\ &= -\int_{\Omega} dx \int_{-\infty}^{+\infty} f_t(x) \ \text{div} \ \varphi \, dt \\ &= \int_{-\infty}^{0} dt \int_{\Omega} \chi_{\Omega \setminus E_t} \text{div} \ \varphi \, dx - \int_{0}^{+\infty} dt \int_{\Omega} \chi_{E_t} \text{div} \ \varphi \, dx \\ &= -\int_{-\infty}^{0} dt \int_{\Omega} \chi_{E_t} \text{div} \ \varphi \, dx - \int_{0}^{+\infty} dt \int_{\Omega} \chi_{E_t} \text{div} \ \varphi \, dx \\ &= -\int_{-\infty}^{+\infty} dt \int_{\Omega} \chi_{E_t} \text{div} \ \varphi \, dx \\ &= \int_{-\infty}^{+\infty} < D\chi_{E_t}, \varphi > \, dt. \end{aligned}$$

Hence $Du = \int_{-\infty}^{+\infty} D\chi_{E_t}\, dt$ and $|Du|(\Omega) \leq \int_{-\infty}^{+\infty} \int_{\Omega} |D\chi_{E_t}| dt$.

We establish now the converse inequality $\int_{-\infty}^{+\infty} \int_\Omega |D\chi_{E_t}|\ dt \le |Du|(\Omega)$, which also proves that $D\chi_{E_t}$ belongs to $\mathbf{M}(\Omega, \mathbf{R}^N)$ for a.e. t in $\mathbf{R}$.

Step 1. We assume that u belongs to the space $\mathcal{A}(\Omega)$ of piecewise linear and continuous functions in Ω. By linearity, one can assume that u is the linear function $u = a.x + b$ with $a \in \mathbf{R}^N$ and $b \in \mathbf{R}$ so that

$$\begin{aligned}\int_\Omega |D\chi_{E_t}| &= \mathcal{H}^{N-1}(\Omega \cap \partial E_t)\\ &= \mathcal{H}^{N-1}(\Omega \cap [a.x + b = t]).\end{aligned}$$

Consequently, according to the classical coarea formula (Theorem 4.2.5),

$$\begin{aligned}\int_{-\infty}^{+\infty} \int_\Omega |D\chi_{E_t}| &= \int_{-\infty}^{+\infty} \int_{[a.x+b=t]} \chi_\Omega(x)\ d\mathcal{H}^{N-1}(x)\ dt\\ &= |a|\mathcal{L}^N(\Omega) = \int_\Omega |Du|.\end{aligned}$$

Step 2. We establish the inequality $\int_{-\infty}^{+\infty} \int_\Omega |D\chi_{E_t}|\ dt \le |Du|(\Omega)$ for all $u \in BV(\Omega)$.

Let $(u_n)_{n\in\mathbf{N}}$ be a sequence in $\mathcal{A}(\Omega)$ such that $u_n \rightharpoonup u$ for the intermediate convergence. Such a sequence exists from Theorem 10.1.2 and the well-known density of the space $\mathcal{A}(\Omega)$ in $W^{1,1}(\Omega)$ equipped with its strong topology. For another and direct proof of this assertion, consult Ziemer [229, Exercise 5.2]. Let us set $E_{n,t} := \{x \in \Omega : u_n(x) > t\}$. According to the first step and to Fatou's lemma, we have

$$\begin{aligned}\int_\Omega |Du| &= \lim_{n\to+\infty} \int_\Omega |Du_n|\\ &= \lim_{n\to+\infty} \int_{-\infty}^{+\infty} \int_\Omega |D\chi_{E_{n,t}}|\ dt\\ &\ge \int_{-\infty}^{+\infty} \liminf_{n\to+\infty} \int_\Omega |D\chi_{E_{n,t}}|\ dt. \qquad (10.37)\end{aligned}$$

On the other hand,

$$\int_\Omega |u_n - u|\ dx = \int_\Omega \int_{-\infty}^{+\infty} |\chi_{E_{n,t}} - \chi_{E_t}|\ dt\ dx = \int_{-\infty}^{+\infty} \left(\int_\Omega |\chi_{E_{n,t}} - \chi_{E_t}|\ dx \right) dt,$$

which converges to zero. Thus, for a subsequence (not relabeled), and for almost all t in $\mathbf{R}$,

$$\chi_{E_{n,t}} \to \chi_{E_t} \text{ strongly in } L^1(\Omega). \qquad (10.38)$$

The lower semicontinuity of the total variation with respect to the strong convergence in $L^1(\Omega)$ (Proposition 10.1.1) and (10.37), (10.38), finally yield

$$\int_\Omega |Du| \ge \int_{-\infty}^{+\infty} \int_\Omega |D\chi_{E_t}|\ dt.$$

It is now easy to adapt the proof above for obtaining the coarea formula with $f = \chi_E$ for any Borel subset E of Ω. The general coarea formula with a Borel function $f : \Omega \to \mathbf{R}^+$ is then obtained by a classical density argument. □

We are now in a position to establish that $\mathcal{H}^{N-1}$ almost all points of Ω are regular for function in $BV(\Omega)$, and that their jump set (cf. Theorem 10.3.1), is countably $(N-1)$-rectifiable. For each function u in $BV(\Omega)$ whose representative satisfies the convention of Remark 10.3.2, we set

$$S_u = \{\, x \in \Omega : u^-(x) < u^+(x) \,\},$$

where $u^-(x) = \operatorname{ap}\liminf_{y\to x} u(y)$ and $u^+(x) = \operatorname{ap}\limsup_{y\to x} u(y)$.

Theorem 10.3.4. *Let u be a given function in $BV(\Omega)$. Then, for $\mathcal{H}^{N-1}$-almost all x in Ω, $u^-(x)$ and $u^+(x)$ are finite and S_u is countably $(N-1)$-rectifiable. Moreover, S_u is, up to a set of $\mathcal{H}^{N-1}\lfloor\Omega$ measure zero, the jump set of u, and $\mathcal{H}^{N-1}$ almost all x in Ω are regular for u.*

Proof. We begin by proving that S_u is countably $(N-1)$-rectifiable. According to the coarea formula (Theorem 10.3.3), for almost all $t \in \mathbf{R}$, $E_t = \{x \in \Omega : u(x) > t\} := [u > t]$ is a set of finite perimeter in Ω. Now let D be a dense countable subset of $\{t \in \mathbf{R} : E_t \text{ is of finite perimeter}\}$ and set $S_{u,t} := \{\, x \in S_u : u^-(x) < t < u^+(x) \,\}$. We have $S_u = \bigcup_{t\in D} S_{u,t}$. On the other hand, from definitions of u^- and u^+, it is easy to establish that for all $x \in S_{u,t}$,

$$\begin{cases} \limsup\limits_{\rho\to 0} \dfrac{\mathcal{L}^N(B_\rho(x) \cap [u > t])}{\mathcal{L}^N(B_\rho(x))} > 0, \\ \limsup\limits_{\rho\to 0} \dfrac{\mathcal{L}^N(B_\rho(x) \cap [u < t])}{\mathcal{L}^N(B_\rho(x))} > 0, \end{cases}$$

so that $x \in \partial_M E_t$. Thus $S_u \subset \bigcup_{t\in D} \partial_M E_t$ and the conclusion follows from the structure theorem, Theorem 10.3.2.

We admit that $-\infty < u^-(x) \le u^+(x) < +\infty$ for $\mathcal{H}^{N-1}$ almost all x in Ω. For a proof, consult Evans and Gariepy [132, Theorem 2]. For establishing that up to an $\mathcal{H}^{N-1}\lfloor\Omega$-negligible set, S_u is the jump set of u and that $\mathcal{H}^{N-1}$ almost all x in Ω are regular for u, according to Proposition 10.3.3, Definition 10.3.5, and Theorem 10.3.1, it is enough to establish that for $\mathcal{H}^{N-1}$ almost all x in S_u, there exists $\nu(x)$ in S^{N-1} such that $u^-(x) = \operatorname{ap}\lim_{y\to x,\ y\in\pi_{-\nu(x)}(x)} u(y)$ and $u^+(x) = \operatorname{ap}\lim_{y\to x,\ y\in\pi_{\nu(x)}(x)} u(y)$. Let $x \in S_u$ such that $u^-(x)$ and $u^+(x)$ are finite and set $t = u^+(x) - \varepsilon$ with ε small enough so that $u^-(x) < t < u^+(x)$. Thus $x \in \partial_M E_t$ and, from Theorem 10.3.2, for $\mathcal{H}^{N-1}$-almost all such x in S_u, there exists $\nu(x)$ in S^{N-1} such that

$$\lim_{\rho\to 0} \frac{\mathcal{L}^N\big(H_{\rho,\nu(x)}(x) \cap [u > u^+(x) - \varepsilon]\big)}{\mathcal{L}^N\big(H_{\rho,\nu(x)}(x)\big)} = 1. \tag{10.39}$$

On the other hand, according to the definition of the approximate limsup,

$$\lim_{\rho\to 0} \frac{\mathcal{L}^N\big(B_\rho(x) \cap [u > u^+(x) + \varepsilon]\big)}{\mathcal{L}^N\big(B_\rho(x)\big)} = 0,$$

hence

$$\lim_{\rho\to 0}\frac{\mathcal{L}^N\big(H_{\rho,\nu(x)}(x)\cap[u>u^+(x)+\varepsilon]\big)}{\mathcal{L}^N\big(H_{\rho,\nu(x)}(x)\big)}=0. \tag{10.40}$$

Combining (10.39) and (10.40) we obtain

$$\lim_{\rho\to 0}\frac{\mathcal{L}^N\big(H_{\rho,\nu(x)}(x)\cap[|u-u^+(x)|<\varepsilon]\big)}{\mathcal{L}^N\big(H_{\rho,\nu(x)}(x)\big)}=1,$$

which proves that $u^+(x)=\operatorname{ap}\lim_{y\to x,\ y\in\pi_{\nu(x)}(x)}u(y)$. The proof of

$$u^-(x)=\operatorname{ap}\lim_{y\to x,\ y\in\pi_{-\nu(x)}(x)}u(y)$$

is similar. □

Remark 10.3.4. In the proof of Theorem 10.3.4 we have established that S_u possesses for $\mathcal{H}^{N-1}$-almost every x in Ω, a normal unit vector $\nu_u(x)$ and that

$$\mathcal{H}^{N-1}\big(S_u\setminus\{x\in\mathbf{R}^N : u_{\nu_u(x)}(x)\neq u_{-\nu_u(x)}(x)\}\big)=0.$$

Moreover, we have obtained that for almost every $t\in\mathbf{R}$ and for $\mathcal{H}^{N-1}$-almost every x in $\partial_M E_t\cap S_u$, $\nu_u(x)=\nu_{E_t}$, where ν_{E_t} is the inner measure theoretic normal to E_t at x.

Remark 10.3.5. According to Theorem 10.3.4 and Proposition 10.3.3, there exist two Borel sets E^+ and E^- such that for H^{N-1}-almost all x in S_u

$$u^+(x)=\lim_{y\to x,\ y\in E^+\cap\pi_{\nu_u}(x)}u(y)\text{ and }u^-(x)=\lim_{y\to x,\ y\in E^-\cap\pi_{-\nu_u}(x)}u(y).$$

Remark 10.3.6. According to our convention (Remark 10.3.2) on the representative of L^1-functions, for $\mathcal{H}^{N-1}$ a.e. x in $\Omega\setminus S_u$, $u^+(x)=u^-(x)=u(x)$.

In the following proposition, we give some piece of information on u when x belongs to S_u.

Proposition 10.3.5. *Let $u\in BV(\Omega)$. For $\mathcal{H}^{N-1}$-almost every x in S_u letting $E_t:=[u>t]$, one has*

$$(u^-(x),u^+(x))\subset\{t\in\mathbf{R}: x\in\partial_M E_t\cap\Omega\}\subset[u^-(x),u^+(x)],$$

and for $\mathcal{H}^{N-1}$-almost every x in $\partial_M E_t\cap\Omega\setminus S_u$, $u(x)=t$.

Proof. For the first inclusion $(u^-(x),u^+(x))\subset\{t\in\mathbf{R}: x\in\partial_M E_t\cap\Omega\}$, it suffices to note that $t\in(u^-(x),u^+(x))$ implies $x\in S_{u,t}\subset\partial_M E_t$. We establish now the second inclusion $\{t\in\mathbf{R}: x\in\partial_M E_t\cap\Omega\}\subset[u^-(x),u^+(x)]$. Let $t\in\mathbf{R}$ be such that $x\in\partial_M E_t\cap\Omega$, assume that $t>u^+(x)$, and take t_0 such that $t>t_0>u^+(x)$ and

$$\lim_{\rho\to 0}\frac{\mathcal{L}^N(B_\rho(x)\cap[u>t_0])}{\mathcal{L}^N(B_\rho(x))}=0.$$

Such a t_0 exists from the definition of the approximate limsup. We have

$$\lim_{\rho\to 0} \frac{\mathcal{L}^N(B_\rho(x) \cap [u > t])}{\mathcal{L}^N(B_\rho(x))} = 0,$$

which is in contradiction with $x \in \partial_M E_t$. Assuming $t < u^-(x)$ yields the same contradiction.

Since, up to a set of $\mathcal{H}^{N-1}\lfloor\Omega$ measure zero, S_u is the jump set of u, one has, for $\mathcal{H}^{N-1}$-almost every x in $\partial_M E_t \cap \Omega \setminus S_u$, $u(x) = u^+(x) = u^-(x)$, so that $u(x) = \text{ap } \lim_{y\to x} u(y)$. For such x, we establish that $u(x) = t$. Otherwise, assume that $t > u(x)$ and set $\varepsilon = t - u(x)$. According to the definition of the approximate limit at x, we have

$$\lim_{\rho\to 0} \frac{\mathcal{L}^N(B_\rho(x) \cap [|u - u(x)| > \varepsilon])}{\mathcal{L}^N(B_\rho(x))} = 0,$$

which yields

$$\lim_{\rho\to 0} \frac{\mathcal{L}^N(B_\rho(x) \cap [u > \varepsilon + u(x)])}{\mathcal{L}^N(B_\rho(x))} = 0,$$

that is,

$$\lim_{\rho\to 0} \frac{\mathcal{L}^N(B_\rho(x) \cap [u > t])}{\mathcal{L}^N(B_\rho(x))} = 0,$$

which contradicts the hypothesis $x \in \partial_M E_t$. Using the same arguments, assumption $u < t$ would give the same contradiction. □

10.4 Structure of the gradient of BV functions

Let u be a given function in $BV(\Omega)$ and $Du = D^a u + D^s u$ the Lebesgue–Nikodym decomposition of the measure Du with respect to the N-dimensional Lebesgue measure $\mathcal{L}^N\lfloor\Omega$ restricted to Ω. Let us recall that the measure $D^a u$ denotes the absolutely continuous part of Du with respect to the measure $\mathcal{L}^N\lfloor\Omega$ and $D^s u$ its singular part. We will denote the density of $D^a u$ with respect to $\mathcal{L}^N\lfloor\Omega$ by ∇u, so that $D^a u = \nabla u\ \mathcal{L}^N\lfloor\Omega$. The theorem below makes precise the structure of the singular part $D^s u$.

Theorem 10.4.1. *Let us denote the two measures $D^s u\lfloor S_u$ and $D^s u\lfloor\Omega \setminus S_u$ by Ju and Cu, respectively, called jump part and Cantor part of Du. Then Ju is absolutely continuous with respect to the restriction of the $(N-1)$-dimensional Hausdorff measure to S_u. More precisely,*

$$Ju = (u^+ - u^-)\ \nu_u\ \mathcal{H}^{N-1}\lfloor S_u.$$

Moreover, Ju and Cu are mutually singular: for all Borel sets E of Ω

$$\mathcal{H}^{N-1}(E) < +\infty \Longrightarrow |Cu|(E) = 0.$$

Consequently, the Hausdorff dimension of the support $spt(Cu)$ of the measure Cu satisfies

$$N - 1 \le dim_H(spt(Cu)) < N.$$

PROOF. According to the coarea formula, Theorem 10.3.3, and to Theorem 10.3.2(iii), for all Borel set $E \subset S_u$

$$\begin{aligned} Ju(E) = Du(E) &= \int_{-\infty}^{+\infty} D\chi_{E_t}(E)\, dt \\ &= \int_{-\infty}^{+\infty} \left(\int_{E \cap \partial_M E_t} \nu_u(x)\, d\mathcal{H}^{N-1}(x) \right) dt \\ &= \int_E \left(\int_{-\infty}^{+\infty} \chi_{\{t \in \mathbf{R} \,:\, x \in \partial_M E_t\}}\, dt \right) \nu_u(x)\, d\mathcal{H}^{N-1}(x). \end{aligned}$$

Since $E \subset S_u$, according to Proposition 10.3.5, for $\mathcal{H}^{N-1}$ a.e. x in E, one has

$$\mathcal{L}^1(\{t \in \mathbf{R} : x \in \partial_M E_t\}) = u^+(x) - u^-(x),$$

thus $Ju = (u^+ - u^-)\, \nu_u\, \mathcal{H}^{N-1} \lfloor S_u$.

Let now E be any Borel set included in $\Omega \setminus S_u$, satisfying $\mathcal{H}^{N-1}(E) < +\infty$. According to Theorems 10.3.3 and 10.3.2(iii), one has

$$|Cu|(E) = |Du|(E) = \int_{-\infty}^{+\infty} \mathcal{H}^{N-1}(E \cap \partial_M E_t)\, dt. \tag{10.41}$$

With our convention (Remark 10.3.2), the points of $\Omega \setminus S_u$ are all points of approximate continuity for u; thus, according to Proposition 10.3.5, one has $E \cap \partial_M E_t \subset \{y \in E : u(y) = t\}$. Moreover, $\mathcal{H}^{N-1}(E) < +\infty$, so that, from Lemma 4.2.1, the set of all t such that $\mathcal{H}^{N-1}(\{y \in E : u(y) = t\}) > 0$ is at most countable, and (10.41) yields $|Cu|(E) = 0$. □

The proposition below states that all functions u in $BV_{loc}(\mathbf{R}^N)$ possess an approximate derivative for almost all x in $\mathbf{R}^N$ in the following sense: there exists a linear function $L : \mathbf{R}^N \longrightarrow \mathbf{R}$ denoted by ap Du such that

$$\operatorname{ap}\lim_{y \to x} \frac{|u(y) - u(x) - L(y - x)|}{|y - x|} = 0.$$

For a proof, see [134, Theorem 4.5.9] or [132, Theorem 4].

Proposition 10.4.1. *Let u be a given function of $BV_{loc}(\mathbf{R}^N)$, i.e., $u \in BV(U)$ for all open bounded subset U of $\mathbf{R}^N$. Then for almost all x_0 in $\mathbf{R}^N$,*

$$\lim_{\rho \to 0} \frac{1}{\rho^d} \int_{B_\rho(x_0)} \frac{|\, u(x) - u(x_0) - \nabla u(x_0)\, .\, (x - x_0)\, |}{|\, x - x_0\, |} dx = 0.$$

Consequentely for almost all x_0 in $\mathbf{R}^N$ ap $\lim_{x \to x_0} Du = \nabla u(x_0)$.

Example 10.4.1. We construct a BV-function whose gradient is reduced to its Cantor part: the Cantor–Vitali function. Let $\Omega = (0, 1)$ and C be the classical triadic Cantor set

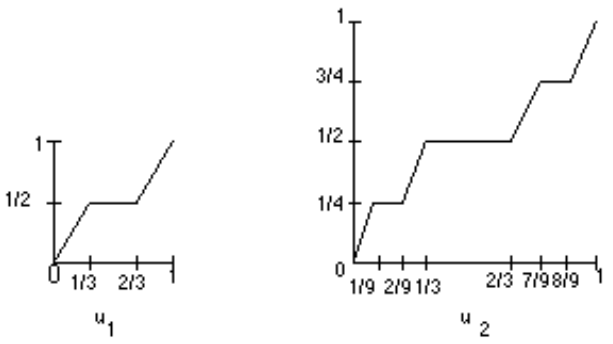


Figure 10.3. *Construction of the Cantor–Vitali function.*

$C = \bigcap_{n\in\mathbf{N}} C_n$, where C_n is the union of 2^n intervals of size 3^{-n}. We define

$$\begin{cases} f_n(x) := \left(\dfrac{2}{3}\right)^{-n} \chi_{C_n}, \\ \\ u_n(x) := \displaystyle\int_0^x f_n(t)\, dt \text{ (see Figure 10.3).} \end{cases}$$

All the functions u_n belong to $\mathbf{C}([0,1])$ and if I is any of the 2^n intervals of C_n,

$$\begin{cases} \displaystyle\int_I f_n(t)\, dt = \int_I f_{n+1}(t)\, dt = 2^{-n}, \\ \forall x \in (0,1) \setminus C_n,\ u_n(x) = u_{n+1}(x). \end{cases}$$

Indeed $\int_I f_n(t)\, dt = (\frac{3}{2})^n \text{mes}(I) = 2^{-n}$. We deduce that for all x in C_n,

$$|u_n(x) - u_{n+1}(x)| \leq 2^{-(n-1)},$$

so that u_n uniformly converges to a continuous function u. According to the lower semi-continuity of the total variation, we then obtain

$$\int_{(0,1)} |Du| \leq \liminf_{n\to+\infty} \int_{(0,1)} |Du_n|\, dx = 1,$$

which proves that u belongs to $BV(0,1)$ and, since u is continuous, that $Ju = 0$. Finally, since u is locally constant on $(0,1) \setminus C$ and $\mathcal{L}^1(C) = 0$, one has $\nabla u = 0$ and $Du = Cu$. Moreover, the support of Cu is the Cantor set C whose Hausdorff dimension is $\ln(2)/\ln(3) \sim 0.632$ (see Example 4.1.1).

10.5 The space $SBV(\Omega)$

In some problems arising in image segmentation, or in mechanics in the study of cracks and fissures (see chapters 12 and 14), the first distributional derivatives of the competing

functions which operate in the models are often measures without singular Cantor part. The solutions of these problems may be found in a special space of functions of bounded variation.

10.5.1 Definition

Definition 10.5.1. *The special set of functions of bounded variation is the subset $SBV(\Omega)$ of $BV(\Omega)$ made up of all the functions of $BV(\Omega)$ whose gradient measures have no Cantor part in their Lebesgue decomposition, i.e.,*

$$u \in SBV(\Omega) \iff u \in L^1(\Omega) \text{ and } Du = \nabla u \, \mathcal{L}^N \lfloor \Omega + (u^+ - u^-) \, \nu_u \, \mathcal{H}^{N-1} \lfloor S_u.$$

Remark 10.5.1. Arguing as in Remark 10.2.2, one may define the space $SBV(\Omega, \mathbf{R}^m)$ as the space of all functions $u : \Omega \to \mathbf{R}^m$ which belong to $L^1(\Omega, \mathbf{R}^m)$, and whose distributional derivative Du is a $M^{m \times N}$-valued measure of the form

$$Du = \nabla u \, \mathcal{L}^N \lfloor \Omega + (u^+ - u^-) \otimes \nu_u \, \mathcal{H}^{N-1} \lfloor S_u.$$

Example 10.5.1. Let Ω be an open bounded subset of $\mathbf{R}^N$, K a closed subset of Ω such that $\mathcal{H}^{N-1}(K) < +\infty$, and $u \in W^{1,1}(\Omega \setminus K) \cap L^\infty(\Omega)$. We claim that u belongs to $SBV(\Omega)$ and that $S_u \subset K$.

We first assume that K is regular in the following sense: there exist a $\mathbf{C}^1$ hypersurface Σ such that $K \subset \Sigma$ and two disjoint subsets Ω_1 and Ω_2 of Ω such that $\partial\Omega_1 \cap \partial\Omega_2 = \Sigma$ and $\mathcal{L}^N(\Omega \setminus \Omega_1 \cup \Omega_2) = 0$. Then the result is a straightforward consequence of the trace theory (see Subsection 10.2 and Example 10.2.1). It is worth pointing out that in this case, the hypothesis $u \in L^\infty(\Omega)$ is unnecessary.

We now consider the general case. If $N = 1$, the result follows from the previous argument. We assume $N \geq 2$. Since $\mathcal{H}^{N-1}(K) < +\infty$ and K is a compact set, for all $n \in \mathbf{N}^*$ there exists a finite family of closed balls $(B^n_{\rho_i}(x_i))_{i \in I_n}$, covering K, with $\rho_i \leq \frac{1}{n}$ and such that

$$\sum_{i \in I_n} c_{N-1} (2\rho_i)^{N-1} \leq \mathcal{H}^{N-1}(K) + 1.$$

Therefore

$$\sum_{i \in I_n} \mathcal{H}^{N-1}(\partial B^n_{\rho_i}(x_i)) \leq C(\mathcal{H}^{N-1}(K) + 1),$$

where C is a positive constant depending only on the dimension N. We now consider the following functions u_n:

$$u_n(x) = \begin{cases} u(x) \text{ for } x \in \Omega \setminus \bigcup_{i \in I_n} B^n_{\rho_i}(x_i), \\ 0 \text{ elsewhere.} \end{cases}$$

Since $\mathcal{L}^N(\cup_{i \in I_n} B^n_{\rho_i}(x_i))$ tends to zero, $u_n \to u$ strongly in $L^1(\Omega)$ when $n \to +\infty$. Since moreover $\partial B^n_{\rho_i}(x_i))$ is a finite union of $\mathbf{C}^1$ hypersurfaces, reasoning on a neighborhood

of each $\partial B^n_{\rho_i}(x_i)$), from the trace theory and the estimate $|u_n^+(x)| \leq ||u||_{L^\infty(\Omega)}$ (note that according to Remark 10.3.4, u^+ is a classical limit in a Borel set E^+ of Ω), one has

$$\int_\Omega |Du_n| \leq \int_{\Omega\setminus K} |\nabla u|\, dx + C||u||_{L^\infty(\Omega)}(\mathcal{H}^{N-1}(K)+1).$$

The semicontinuity of the total variation (Proposition 10.1.1) yields $u \in BV(\Omega)$. On the other hand, it is easily seen that $S_u \subset K$. Since $\mathcal{H}^{N-1}(K \setminus S_u) \leq \mathcal{H}^{N-1}(K) < +\infty$, according to Theorem 10.4.1, $Cu(K \setminus S_u) = 0$. Finally $Cu = 0$ because $Cu(S_u) = 0$.

10.5.2 Properties

The following chain-rule for the derivatives in $BV(\Omega)$ was established in Ambrosio [20].

Proposition 10.5.1. *Let u be a given function in $BV(\Omega)$ and φ in $\mathbf{C}^1_0(\mathbf{R})$. Then $v := \varphi \circ u$ belongs to $BV(\Omega)$ and, even if it means changing ν_u by $-\nu_u$,*

$$\begin{cases} Jv = \big(\varphi(u^+) - \varphi(u^-)\big)\nu_u\, \mathcal{H}^{N-1}\lfloor S_u, \\ \nabla v = \varphi'(u)\nabla u,\ Cv = \varphi'(u)Cu. \end{cases}$$

Proof. Consider $u_n \in \mathbf{C}^\infty(\Omega) \cap BV(\Omega)$ converging to u for the intermediate convergence. Then $v_n := \varphi \circ u_n \to v := \varphi \circ u$ in $L^1(\Omega)$. On the other hand,

$$\begin{aligned} |Dv|(\Omega) &\leq \liminf_{n\to+\infty} |Dv_n|(\Omega) \\ &\leq ||\varphi'||_\infty \liminf_{n\to+\infty} |Du_n|(\Omega) \\ &= ||\varphi'||_\infty |Du|(\Omega) < +\infty. \end{aligned}$$

This proves that $v \in BV(\Omega)$.

We now show that $Jv = \big(\varphi(u^+) - \varphi(u^-)\big)\nu_u\, \mathcal{H}^{N-1}\lfloor S_u$. Since $\varphi \in \mathbf{C}^1_0(\mathbf{R})$, φ is the difference of two nondecreasing functions in $\mathbf{C}^1(\mathbf{R})$. One may then assume φ nondecreasing so that $v^+ = \varphi \circ u^+,\ v^- = \varphi \circ u^-$, $S_v = S_u$, and $\nu_v = \nu_u$.

It remains to establish $\nabla v = \varphi'(u)\nabla u$ and $Cv = \varphi'(u)Cu$ or equivalently, $Dv\lfloor_{\Omega\setminus S_v} = \varphi'(u)Du\lfloor_{\Omega\setminus S_v}$. Consider a Borel set E of Ω included in $\Omega \setminus S_v$. From the coarea formula (Theorem 10.3.3) and the structure of simple functions of $BV(\Omega)$ (Theorem 10.3.2(iii)), one has

$$\begin{aligned} Dv(E) = \int_{-\infty}^{+\infty} D\chi_{[v>t]}(E)\, dt &= \int_{-\infty}^{+\infty} D\chi_{[u>\varphi^{-1}(t)]}(E)\, dt \\ &= \int_{-\infty}^{+\infty} D\chi_{[u>t]}(E)\, \varphi'(t)\, dt \\ &= \int_{-\infty}^{+\infty} \int_E D\chi_{[u>t]}\, \varphi'(t)\, dt \\ &= \int_{-\infty}^{+\infty} \mathcal{H}^{N-1}(\partial_M([u>t]) \cap E)\, \varphi'(t)\, dt. \end{aligned} \tag{10.42}$$

But, according to Proposition 10.3.5, for $\mathcal{H}^{N-1}$ a.e. x in Ω, $x \in \partial_M([u > t]) \cap E \Longrightarrow u(x) = t$ so that (10.42) yields

$$\begin{aligned} Dv(E) &= \int_{-\infty}^{+\infty} \left(\int_E \varphi'(u(x))\, d\mathcal{H}^{N-1} \lfloor \partial_M([u > t])(x) \right) dt \\ &= \int_{-\infty}^{+\infty} \left(\int_E \varphi'(u) D\chi_{[u>t]} \right) dt \\ &= \varphi'(u)\ Du(E), \end{aligned}$$

which completes the proof. □

The following criterion for a function u in $BV(\Omega)$ to belong to $SBV(\Omega)$ was established by Ambrosio in [14].

Theorem 10.5.1. *Let u be a given function in $BV(\Omega)$. Then u belongs to $SBV(\Omega)$ iff there exists a Borel measure μ in $\mathbf{M}(\Omega \times \mathbf{R}, \mathbf{R}^N)$ and a in $L^1(\Omega, \mathbf{R}^N)$ such that for all Φ in $\mathbf{C}_c^1(\Omega, \mathbf{R}^N)$ and all φ in $\mathbf{C}_0^1(\mathbf{R})$,*

$$\int_{\Omega \times \mathbf{R}} \varphi(s)\Phi(x)\ \mu(dx, ds) = -\int_\Omega \Big(\varphi'(u)\ a\ .\ \Phi(x) + \varphi(u)\ \text{div}\ \Phi(x) \Big)\, dx. \qquad (10.43)$$

Moreover, $a = \nabla u$ a.e. and

$$\begin{cases} \mu = \Lambda_\#^+(\nu_u\ \mathcal{H}^{N-1} \lfloor S_u) - \Lambda_\#^-(\nu_u\ \mathcal{H}^{N-1} \lfloor S_u), \\ |\mu|(\Omega \times \mathbf{R}) = 2\ \mathcal{H}^{N-1}(S_u), \end{cases}$$

where $\Lambda^+ : \Omega \longrightarrow \Omega \times \mathbf{R},\ x \mapsto (x, u^+(x))$ and $\Lambda^- : \Omega \longrightarrow \Omega \times \mathbf{R},\ x \mapsto (x, u^-(x))$.

Proof. Let us assume that u belongs to $SBV(\Omega)$. According to Proposition 10.5.1, for all $\varphi \in \mathbf{C}_0^1(\mathbf{R})$, $\varphi(u)$ belongs to $SBV(\Omega)$ and $D(\varphi(u)) = \varphi'(u)\nabla u \mathcal{L} \lfloor \Omega + \big(\varphi(u^+) - \varphi(u^-)\big)\nu_u\ \mathcal{H}^{N-1} \lfloor S_u$. Consequently, for all $\Phi \in \mathbf{C}_c^1(\Omega, \mathbf{R}^N)$

$$\begin{aligned} \int_\Omega \varphi(u)\ \text{div}\ \Phi\, dx &= -\langle D(\varphi(u)), \Phi \rangle \\ &= -\int_\Omega \varphi'(u)\nabla u.\Phi\, dx - \int_\Omega \big(\varphi(u^+) - \varphi(u^-)\big)\Phi.\nu_u\, d\mathcal{H}^{N-1} \lfloor S_u, \end{aligned}$$

which we write

$$\int_\Omega \big(\varphi(u^+) - \varphi(u^-)\big)\Phi.\nu_u\, d\mathcal{H}^{N-1} \lfloor S_u = -\int_\Omega \Big(\varphi'(u)\ \nabla u\ .\ \Phi + \varphi(u)\ \text{div}\ \Phi \Big)\, dx.$$

According to the definition of the image of a measure, the left-hand side above is nothing but

$$\int_{\Omega \times \mathbf{R}} \varphi(s)\Phi(x)\, d\mu(x, s).$$

Finally, since $\Lambda^+(\Omega)$ and $\Lambda^-(\Omega)$ are disjoint sets, and by injectivity of Λ^+ and Λ^-, one has

$$\begin{aligned} |\mu| &= |\Lambda_\#^+(\nu_u\ \mathcal{H}^{N-1} \lfloor S_u) - \Lambda_\#^-(\nu_u\ \mathcal{H}^{N-1} \lfloor S_u)| \\ &= |\Lambda_\#^+(\nu_u\ \mathcal{H}^{N-1} \lfloor S_u)| + |\Lambda_\#^-(\nu_u\ \mathcal{H}^{N-1} \lfloor S_u)| \\ &= \Lambda_\#^+(\mathcal{H}^{N-1} \lfloor S_u) + \Lambda_\#^-(\mathcal{H}^{N-1} \lfloor S_u). \end{aligned}$$

Hence $|\mu|(\Omega \times \mathbf{R}) = 2\mathcal{H}^{N-1}(S_u)$. The proof of the converse condition proceeds in three steps.

First step. We establish $a(x_0) = \nabla u(x_0)$ for a.e. x_0 in Ω. Let us fix $x_0 \in \Omega$ such that

$$\begin{cases} a_\rho(y) := a(x_0 + \rho y) \to a(x_0) \text{ strongly in } L^1(B), \\ u_\rho \to \nabla u(x_0).y \text{ strongly in } L^1(B), \\ \lim\limits_{\rho \longrightarrow 0} \dfrac{1}{\rho^{N-1}} |\mu|(B_\rho(x_0) \times \mathbf{R}) = 0, \end{cases}$$

where u_ρ denotes the rescaled function

$$u_\rho(y) := \frac{1}{\rho}\big(u(x_0 + \rho y) - u(x_0)\big),$$

and $B_\rho(x_0)$, B, respectively, the open ball in $\mathbf{R}^N$ with radius ρ, centered at x_0, and the unit open ball in $\mathbf{R}^N$ centered at 0. Let us justify the possible choice of such x_0. Actually x_0 is chosen to be a Lebesgue point of $y \mapsto a_\rho(y) := a(x_0 + \rho y)$. On the other hand, according to Proposition 10.4.1, the second property is satisfied a.e. in Ω. Finally, denoting by π the projection from $\Omega \times \mathbf{R}$ onto Ω and by $\pi\#|\mu|$ the image of the measure $|\mu|$ by π, the limit

$$\lim_{\rho \longrightarrow 0} \frac{1}{\rho^N} |\mu|(B_\rho(x_0) \times \mathbf{R}) = \lim_{\rho \longrightarrow 0} \frac{1}{\rho^N} \pi_\# |\mu|(B_\rho(x_0))$$

exists for almost every x_0 in Ω, and, up to a positive multiplicative constant, is equal to the density of the regular part of the measure $\pi_\#|\mu|$ in its Lebesgue–Nikodym decomposition.

In what follows, x_0 is a fixed element in Ω where these three properties are satisfied. Applying condition (10.43) to the function Φ defined by $\Phi(x) := \tilde{\Phi}(\frac{x - x_0}{\rho})$, where $\tilde{\Phi} \in \mathbf{C}^1_c(B, \mathbf{R}^N)$, one obtains

$$-\frac{1}{\rho}\int_{B_\rho(x_0)} \operatorname{div}\, \tilde{\Phi}\left(\frac{x - x_0}{\rho}\right)\varphi(u)\,dx - \int_{B_\rho(x_0)} \tilde{\Phi}\left(\frac{x - x_0}{\rho}\right)\varphi'(u).a\,dx = \int_{\Omega \times \mathbf{R}} \varphi(s)\Phi(x)\,d\mu,$$

and the change of scale $y = \frac{x - x_0}{\rho}$ gives

$$-\int_B \Big(\operatorname{div}\, \tilde{\Phi}\,\varphi(u(x_0 + \rho y)) + \rho\varphi'(u(x_0 + \rho y))\tilde{\Phi}(y).a_\rho(y)\Big)dy = \frac{1}{\rho^{N-1}}\int_{\Omega \times \mathbf{R}} \varphi(s)\Phi(x)\,d\mu.$$

Testing this equality with the function φ defined by $\varphi(s) := \gamma(\frac{s - u(x_0)}{\rho})$, one obtains

$$-\int_B \Big(\operatorname{div}\, \tilde{\Phi}\,\gamma(u_\rho) + \gamma'(u_\rho)\tilde{\Phi}(y).a_\rho(y)\Big)dy = \frac{1}{\rho^{N-1}}\int_{\Omega \times \mathbf{R}} \varphi(s)\Phi(x)\,d\mu$$

and letting $\rho \to 0$,

$$\int_B \operatorname{div}\, \tilde{\Phi}\,\gamma(\nabla u(x_0).y) + \gamma'(\nabla u(x_0).y)\tilde{\Phi}(y).a(x_0)\,dy = 0.$$

Since

$$\begin{aligned}\int_B \Big(\operatorname{div}\, \tilde{\Phi}\,\gamma(\nabla u(x_0).y) + \gamma'(\nabla u(x_0).y)\tilde{\Phi}(y).\nabla u(x_0)\Big)dy &= \int_B \operatorname{div}\, \big(\gamma(\nabla u(x_0).y)\tilde{\Phi}\big)dy \\ &= 0,\end{aligned}$$

we deduce

$$\big(a(x_0) - \nabla u(x_0)\big).\int_B \gamma'(\nabla u(x_0).y)\tilde{\Phi}(y)\,dy = 0.$$

The choice of $\tilde{\Phi}$ and γ being arbitrary, we deduce $a(x_0) = \nabla u(x_0)$ and the proof of the first step is complete.

Second step. We establish $Cu = 0$. From Proposition 10.5.1, for all φ in $\mathbf{C}^1_0(\mathbf{R})$, $\varphi \circ u \in BV(\Omega)$ so that for all Φ in $\mathbf{C}^1_c(\Omega, \mathbf{R}^N)$, one has

$$\int_\Omega \Big(\text{ div } \Phi\,\varphi(u)+\varphi'(u)\Phi.\nabla u\Big)dx = -\int_{S_u}\big(\varphi(u^+)-\varphi(u^-)\big)\Phi.\nu_u\,d\mathcal{H}^{N-1}-\int_\Omega \varphi'(u)\Phi\,dCu,$$

and condition (10.43) yields

$$\int_\Omega \varphi'(u)\Phi\,dCu = \int_{\Omega\times\mathbf{R}} \varphi(s)\Phi(x)\,\mu(dx,ds) - \int_{S_u}\big(\varphi(u^+)-\varphi(u^-)\big)\Phi.\nu_u\,d\mathcal{H}^{N-1}. \tag{10.44}$$

We now focus on a careful analysis of the measure $\varphi'(u)\,Cu$. Let us apply the slicing Theorem 4.2.4 for the measure μ. Let τ_i denote the density of μ_i with respect to $|\mu_i|$, where $(\mu_i)_{1\le i\le N}$ is the family of components of the measure μ. According to Theorem 4.2.4 one has

$$\int_{\Omega\times\mathbf{R}} \varphi(s)\Phi_i(x)\,\mu_i(dx,ds) = \int_\Omega \Phi_i(x)\Big(\int_{\mathbf{R}} \varphi(s)\tau_i(x,s)\,\theta_x(ds)\Big)\,\sigma_i(dx),$$

where σ_i is the image of the measure $|\mu_i|$ by the projection of $\Omega\times\mathbf{R}$ on Ω and $(\theta_x)_{x\in\mathbf{R}}$ is a family of probability measures on $\mathbf{R}$. Then (10.44) yields, for $i = 1\ldots N$,

$$\varphi'(u)C_iu = \Big(\int_{\mathbf{R}} \varphi(s)\tau_i(x,s)\,\theta_x(ds)\Big)\sigma_i - \big(\varphi(u^+)-\varphi(u^-)\big)\nu_{u,i}\,\mathcal{H}^{N-1}\lfloor S_u, \tag{10.45}$$

and finally, since Cu and $\mathcal{H}^{N-1}\lfloor S_u$ are mutually singular,

$$\varphi'(u)C_iu = \Big(\int_{\mathbf{R}} \varphi(s)\tau_i(.,s)\,\theta_.(ds)\Big)\lambda_i \tag{10.46}$$

with $\lambda_i := \sigma_i\lfloor_{\Omega\setminus S_u}$. To complete the proof, the idea is to express (10.46) in terms of functional identity. Consider the densities b and c of, respectively, C_iu and λ_i with respect to the measure $\alpha := |C_iu| + \lambda_i$, equality (10.46) yields, for α a.e. x in Ω,

$$b(x)(\varphi'\circ u)(x) = c(x)\int_{\mathbf{R}} \varphi(s)\tau_i(x,s)\,\theta_x(ds).$$

Indeed, there exists a Borel set N_φ with $\mathcal{L}^N(N_\varphi) = 0$ such that above equality holds true for all x in $\Omega\setminus N_\varphi$. Since $\mathbf{C}^1_0(\Omega)$ possesses a dense countable subset D, it also holds for all x in $\Omega' = \Omega\setminus\bigcup_{\varphi\in D} N_\varphi$. Let then x in Ω' and assume that $b(x)\neq 0$. We deduce

$$\varphi'\circ u(x) = \frac{c(x)}{b(x)}\int_{\mathbf{R}} \varphi(s)\tau_i(x,s)\,\theta_x(ds) \quad \forall\varphi\in\mathbf{C}^1_0(\mathbf{R}).$$

Equality between the two linear forms

$$\varphi \mapsto (\varphi' \circ u)(x) \text{ and } \varphi \mapsto \frac{c(x)}{b(x)} \int_{\mathbf{R}} \varphi(s)\tau_i(x,s)\ \theta_x(ds)$$

brings a contradiction. (The first is not continuous in $\mathbf{C}_0^1(\mathbf{R})$.) Consequently, $b(x) = 0$ and $C_i u = b\alpha = 0$ which ends the proof of the second step.

Last step. It remains to establish $\mu = \Lambda_{\#}^+(v_u\ \mathcal{H}^{N-1}\lfloor S_u) - \Lambda_{\#}^-(v_u\ \mathcal{H}^{N-1}\lfloor S_u)$. According to the previous step, equality (10.45) now becomes

$$\left(\int_{\mathbf{R}} \varphi(s)\tau_i(x,s)\ \theta_x(ds)\right)\sigma_i = \big(\varphi(u^+) - \varphi(u^-)\big)v_{u,i}\ \mathcal{H}^{N-1}\lfloor S_u.$$

Let $\beta := \sigma_i + \mathcal{H}^{N-1}\lfloor S_u$ and let b, c denote now the densities of $\mathcal{H}^{N-1}\lfloor S_u$ and σ_i with respect to the measure β. We have for all $x \in \Omega \setminus N$ satisfying $\beta(N) = 0$, and for all $\varphi \in \mathbf{C}_0^1(\mathbf{R})$,

$$c(x)\int_{\mathbf{R}} \varphi(s)\tau_i(x,s)\ \theta_x(ds) = b(x)\big(\varphi(u^+) - \varphi(u^-)\big)v_{u,i}.$$

Thus, for all x in $\Omega \setminus N$

$$c(x)\tau_i(x,.)\ \theta_x = b(x)v_{u,i}(x)\big(\delta_{u^+(x)} - \delta_{u^-(x)}\big).$$

For all bounded Borel function $f : \Omega \times \mathbf{R} \longrightarrow \mathbf{R}$, we now obtain

$$\begin{aligned}
\int_{\Omega\times\mathbf{R}} f(x,s)\ d\mu_i(x,s) &= \int_\Omega \left(\int_{\mathbf{R}} f(x,s)\tau_i(x,s)\ \theta_x(ds)\right)\sigma_i(dx) \\
&= \int_\Omega \left(\int_{\mathbf{R}} f(x,s)\tau_i(x,s)\ \theta_x(ds)\right)c(x)\beta(dx) \\
&= \int_{S_u} \big(f(x,u^+(x)) - f(x,u^-(x))\big)v_{u,i}(x)\ d\mathcal{H}^{N-1}(x)
\end{aligned}$$

and the proof is complete. □

We now state, without proof, another criterion for a function u in $L^\infty(\Omega)$ to belong to $SBV(\Omega)$. This criterion concerns the restrictions of u to the one dimensional slices of Ω. Let us define for all $\nu \in S^{N-1}$

$$\begin{cases} \pi_\nu = \{x \in \mathbf{R}^N : x.\nu = 0\}, \\ \Omega_x = \{t \in \mathbf{R} : x + t\nu \in \Omega\},\ x \in \pi_\nu, \\ \Omega_\nu = \{x \in \pi_\nu : \Omega_x \neq \emptyset\}. \end{cases}$$

On the other hand, for all Borel functions $u : \Omega \to \mathbf{R}$ and x in Ω_ν, we define the Borel function u_x for all t in Ω_x by : $u_x(t) = u(x + t\nu)$. For a proof of the theorem below, consult Braides [82].

Theorem 10.5.2. *Let u be a given function in $L^\infty(\Omega)$ such that for all $\nu \in S^{N-1}$*

(i) $u_x \in SBV(\Omega_x)$ *for $\mathcal{H}^{N-1}$ a.e. $x \in \Omega_\nu$;*

(ii) $\displaystyle\int_{\Omega_\nu} \Big(\int_{\Omega_x} |\nabla u_x|\, dt + \mathcal{H}^0(S_{u_x}) \Big)\, \mathcal{H}^{N-1}(dx) < +\infty.$

Then u belongs to $SBV(\Omega)$. Conversely, if u belongs to $SBV(\Omega) \cap L^\infty(\Omega)$, conditions (i) *and* (ii) *are satisfied for all ν in S^{N-1}. Moreover, for $\mathcal{H}^{N-1}$ a.e. x in Ω_ν,*

$$\nabla u(x + t\nu).\nu = \nabla u_x(t)$$

and

$$\int_{\Omega_\nu} H_0(S_{u_x})\, \mathcal{H}^{N-1}(dx) = \int_{S_u} |\nu_u \cdot \nu|\, d\mathcal{H}^{N-1}.$$

Chapter 11

Relaxation in Sobolev, BV, and Young measures spaces

11.1 Relaxation in abstract metrizable spaces

This section is devoted to the description of the relaxation principle in a general metrizable space or, more generally, in a first countable topological space X. Roughly speaking, given an extended real-valued function $F : X \longrightarrow \mathbf{R} \cup \{+\infty\}$, we wish to apply the direct method in the calculus of variations to the lower semicontinuous envelope $\mathrm{cl}(F)$ of the function F so that $\inf_X F = min_X \mathrm{cl}(F)$. Such a procedure is very important in various applications and leads to the concept of generalized solutions for the optimization problem $\inf_X F$. We begin by giving some complements on the sequential version, denoted by $\overline{F}$, of the general notion of lower semicontinuous envelope $\mathrm{cl}(F)$ introduced in chapter 3.

Proposition 11.1.1. *Let $F : X \longrightarrow \mathbf{R} \cup \{+\infty\}$ be a proper extended real-valued function defined on a metrizable space (X, d) or, more generally, on a first countable topological space, and let us define the extended real-valued function $\overline{F} : X \longrightarrow \overline{\mathbf{R}}$ by*

$$\overline{F}(x) := \inf \Big\{ \liminf_{n\to+\infty} F(x_n) \ : \ (x_n)_{n\in\mathbf{N}},\ x = \lim_{n\to+\infty} x_n \Big\}. \tag{11.1}$$

Then the function $\overline{F}$ is characterized for every x in X by the two following assertions:

(i) $\forall (x_n)_{n\in\mathbf{N}}$ *such that* $x_n \to x$, $\overline{F}(x) \leq \liminf_{n\to+\infty} F(x_n)$;

(ii) *there exists a sequence $(y_n)_{n\in\mathbf{N}}$ in X such that $y_n \to x$ and* $\overline{F}(x) \geq \limsup_{n\to+\infty} F(y_n)$.

Proof. Note that trivially the system of assertions (i) and (ii) is equivalent to (i) and (ii)′:

(i) $\forall (x_n)_{n\in\mathbf{N}}$ such that $x_n \to x$, $\overline{F}(x) \leq \liminf_{n\to+\infty} F(x_n)$;

(ii)′ there exists a sequence $(y_n)_{n\in\mathbf{N}}$ in X such that $y_n \to x$ and $\overline{F}(x) = \lim_{n\to+\infty} F(y_n)$;

and each function $\overline{F}$ satisfying (i) and (ii)′ automatically satisfies

$$\overline{F}(x) := \inf\Big\{ \liminf_{n\to+\infty} F(x_n) : (x_n)_{n\in\mathbf{N}},\ x = \lim_{n\to+\infty} x_n \Big\}.$$

We are reduced to establishing that the function $\overline{F}$ defined by formula (11.1) satisfies (i) and (ii)′. We establish only the nontrivial assertion (ii)′. Its proof is based on the following diagonalization lemma.

Lemma 11.1.1. *Let $(a_{m,n})_{(m,n)\in\mathbf{N}\times\mathbf{N}}$ be a sequence in a first countable topological space X such that*

(i) $\lim_{n\to+\infty} a_{m,n} = a_m$;

(ii) $\lim_{m\to+\infty} a_m = a$.

Then there exists a nondecreasing map $n \mapsto m(n)$ from $\mathbf{N}$ into $\mathbf{N}$ such that

$$\lim_{n\to+\infty} a_{m(n),n} = a.$$

For a proof and other diagonalization results, see [28]. Note that under the same conditions, we have the following more classical diagonalization procedure: there exists an increasing map $m \mapsto n(m)$ such that $\lim_{m\to+\infty} a_{m,m(n)} = a$. This second result could be applied for proving Proposition 11.1.1 but we prefer using Lemma 11.1.1, which will turn out to be the fitting tool for establishing a similar proposition in the context of the Γ-convergence (see the next chapter).

Let us go back to the proof of Proposition 11.1.1. By definition of the infima, for all $m \in \mathbf{N}^*$ there exists a sequence $(x_{m,n})_{n\in\mathbf{N}}$ in X satisfying $\lim_{n\to+\infty} x_{m,n} = x$ and such that

$$\text{if } \overline{F}(x) \neq -\infty, \quad \overline{F}(x) \leq \liminf_{n\to+\infty} F(x_{m,n}) \leq \overline{F}(x) + \frac{1}{m},$$
$$\text{if } \overline{F}(x) = -\infty, \quad \liminf_{n\to+\infty} F(x_{m,n}) \leq -m.$$

Therefore $\lim_{m\to+\infty} \lim_{n\to+\infty} F(x_{m,\sigma_m(n)}) = \overline{F}(x)$, where $\sigma_m : \mathbf{N} \longrightarrow \mathbf{N}$ is an increasing map, possibly depending on m, such that

$$\liminf_{n\to+\infty} F(x_{m,n}) = \lim_{n\to+\infty} F(x_{m,\sigma_m(n)}).$$

We end the proof by applying Lemma 11.1.1 to the sequence $((x_{m,\sigma_m(n)}, F(x_{m,\sigma_m(n)}))_{(m,n)\in\mathbf{N}^2}$ in the metrizable space $X \times \overline{\mathbf{R}}$: there exists $n \mapsto m(n)$ mapping $\mathbf{N}$ into $\mathbf{N}$ such that

$$\begin{cases} \lim_{n\to+\infty} F(x_{m(n),\sigma_{m(n)}(n)}) = \lim_{m\to+\infty} \lim_{n\to+\infty} F(x_{m,\sigma_m(n)}) = \overline{F}(x), \\ \lim_{n\to+\infty} x_{m(n),\sigma_{m(n)}(n)} = \lim_{m\to+\infty} \lim_{n\to+\infty} x_{m,\sigma_m(n)} = x. \end{cases}$$

The sequence $(y_n)_{n\in\mathbf{N}}$ defined by $y_n = x_{m(n),\sigma_{m(n)}(n)}$ fulfills assertion (ii)′. □

Theorem 11.1.1. *The function $\overline{F}$ defined in Proposition* 11.1.1 *is the lower semicontinuous (lsc) envelope cl(F) of the function F, i.e., the greatest lower semicontinuous function less than F.*

PROOF. We must establish

$$\begin{cases} \overline{F} \le F; \\ \overline{F} \text{ lsc}; \\ G : X \longrightarrow \overline{\mathbf{R}},\ G \text{ lsc and } G \le F \Longrightarrow G \le \overline{F}. \end{cases}$$

For the first assertion, take the constant sequence $(x_n)_{n\in\mathbf{N}} = (x)_{n\in\mathbf{N}}$ in formula (11.1).

Let us prove the second assertion. Let $(y_m)_{m\in\mathbf{N}}$ be a sequence in X converging to $y \in X$ and consider a subsequence $(y_{\sigma(m)})_{m\in\mathbf{N}}$ satisfying $\lim_{m\to+\infty} \overline{F}(y_{\sigma(m)}) = \liminf_{m\to+\infty} \overline{F}(y_m)$. According to Proposition 11.1.1, there exists a sequence $(y_{\sigma(m),n})_{n\in\mathbf{N}}$ in X satisfying $\lim_{n\to+\infty} y_{\sigma(m),n} = y_{\sigma(m)}$ and such that

$$\begin{aligned} \liminf_{m\to+\infty} \overline{F}(y_m) &= \lim_{m\to+\infty} \overline{F}(y_{\sigma(m)}) \\ &= \lim_{m\to+\infty} \lim_{n\to+\infty} F(y_{\sigma(m),n}). \end{aligned}$$

On the other hand, we have $\lim_{m\to+\infty} \lim_{n\to+\infty} y_{\sigma(m),n} = y$. Applying the diagonalization lemma, Lemma 11.1.1, to the sequence $(y_{\sigma(m),n},\ F(y_{\sigma(m),n}))_{(m,n)\in\mathbf{N}^2}$ in the metrizable space $X \times \overline{\mathbf{R}}$, there exists $n \mapsto m(n)$ mapping $\mathbf{N}$ to $\mathbf{N}$ such that

$$\begin{cases} \lim_{n\to+\infty} F(y_{\sigma(m(n)),n}) = \liminf_{m\to+\infty} \overline{F}(y_m), \\ \lim_{n\to+\infty} y_{\sigma(m(n)),n} = y. \end{cases}$$

Hence

$$\begin{aligned} \liminf_{m\to+\infty} \overline{F}(y_m) &= \lim_{n\to+\infty} F(y_{\sigma(m(n)),n}) \\ &\ge \liminf_{n\to+\infty} F(y_{\sigma(m(n)),n}) \\ &\ge \inf \Big\{ \liminf_{n\to+\infty} F(x_n) \ :\ (x_n)_{n\in\mathbf{N}},\ y = \lim_{n\to+\infty} x_n \Big\} = \overline{F}(y). \end{aligned}$$

We establish now the third assertion. Let $G \le F$ be an lsc function mapping X into $\overline{\mathbf{R}}$ and, for every x in X, consider any sequence $(x_n)_{n\in\mathbf{N}}$ in X converging to x. We have

$$\begin{aligned} G(x) &\le \liminf_{n\to+\infty} G(x_n) \\ &\le \liminf_{n\to+\infty} F(x_n). \end{aligned}$$

Taking the infimum over all the sequences $(x_n)_{n\in\mathbf{N}}$ converging to x, we finally obtain the inequality $G(x) \le \overline{F}(x)$. □

From now on, we write indifferently $\overline{F}$ or cl(F) when X is metrizable or first countable. But note that the two notions differ when X is a general topological space.

In the following theorem, we state the abstract relaxation principle in countable topological spaces. (One could also apply Theorem 3.2.3 and Theorem 11.1.1 to obtain it.)

Theorem 11.1.2. *Let $F : X \longrightarrow \mathbf{R} \cup \{+\infty\}$ be a proper extended real-valued function defined on a metrizable space (X, d) or, more generally, on a first countable topological space, and assume that there exists a minimizing sequence $(x_n)_{n\in\mathbf{N}}$ (i.e., $\lim_{n\to+\infty} F(x_n) = \inf_X F$) such that $S = \{x_n : n \in \mathbf{N}\}$ is relatively compact in X. Then*

(i) $\inf_X F = \min_X cl(F)$;

(ii) *every cluster point $\overline{x}$ of S is a solution of* $\min_X cl(F)$, *i.e.*, $cl(F)(\overline{x}) = \min_X cl(F)$.

Proof. Let $\overline{x}$ be any cluster point of S and $(x_{\sigma(n)})_{n\in\mathbf{N}}$ a subsequence of $(x_n)_{n\in\mathbf{N}}$ converging to $\overline{x}$. According to Proposition 11.1.1(i), we have

$$\mathrm{cl}(F)(\overline{x}) \le \liminf_{n\to+\infty} F(x_{\sigma(n)}) = \inf_X F. \tag{11.2}$$

Let now x be any element of X. From Proposition 11.1.1(ii), there exists a sequence $(y_n)_{n\in\mathbf{N}}$ converging to x and satisfying

$$\mathrm{cl}(F)(x) \ge \limsup_{n\to+\infty} F(y_n). \tag{11.3}$$

Combining (11.2) and (11.3), we obtain

$$\mathrm{cl}(F)(\overline{x}) \le \liminf_{n\to+\infty} F(x_{\sigma(n)}) = \inf_X F \le \limsup_{n\to+\infty} F(y_n) \le \mathrm{cl}(F)(x). \tag{11.4}$$

This proves that $\mathrm{cl}(F)(\overline{x}) = \min_X \mathrm{cl}(F)$. Taking now $x = \overline{x}$ in (11.4), we obtain

$$\inf_X F = \min_X \mathrm{cl}(F) = \mathrm{cl}(F)(\overline{x})$$

and the proof is complete. □

In the terminology of the relaxation theory, the problem

$$(\overline{\mathcal{P}}): \qquad \min \mathrm{cl}(F)$$

is called the *relaxed problem* of the optimization problem

$$(\mathcal{P}): \qquad \inf_X F.$$

A solution of $(\overline{\mathcal{P}})$ is sometimes called a *generalized solution* of the initial problem $(\mathcal{P})$. The relaxation procedure consists in making explicit the lsc envelope of the functional F for a suitable topology on the space X, to obtain a well-posed problem $(\overline{\mathcal{P}})$ in the sense of Theorem 11.1.2 (i.e., the existence of an optimal solution holds for $(\overline{\mathcal{P}})$).

11.2 Relaxation of integral functionals with domain $W^{1,p}(\Omega, \mathbf{R}^m)$, $p > 1$

One of the fundamental hypotheses in elasticity theory is that the total free energy F associated with many materials is of local nature. From a mathematical point of view, the functional F can be represented as the integral over the reference configuration $\Omega \subset \mathbf{R}^N$ ($N = 3$), of a density associated with the possible deformation gradients of the body, which account for the local deformations. The other basic principle is that equilibrium configurations correspond to minimizers of F under prescribed conditions in a Sobolev space $W^{1,p}(\Omega, \mathbf{R}^m)$ ($m = 3$). The functional F may fail to be lower semicontinuous. Indeed, to model the various solid/solid phase transformations in the microstructure, the density energy possesses in general a multiwell structure, and corresponding optimization problems have no solutions.

According to Section 11.1, a classical procedure is to replace F by its lower semicontinuous envelope with respect to the weak topology of $W^{1,p}(\Omega, \mathbf{R}^m)$. The relaxed problem possesses now at least a solution giving the same initial energy.

The case $p = 1$ will be treated in the next section. We will see that contrary to the case $p > 1$, the domain of the lower semicontinuous envelope $\mathrm{cl}(F)$ of the functional F, strictly contains the domain $W^{1,1}(\Omega, \mathbf{R}^m)$ of F. This domain is indeed the space $BV(\Omega, \mathbf{R}^m)$ of functions of bounded variation introduced in chapter 10. This phenomena is due to the lack of reflexivity of the Sobolev space $W^{1,1}(\Omega, \mathbf{R}^m)$.

This approach, which consists in relaxing the functional F with respect to the weak topology of $W^{1,p}(\Omega, \mathbf{R}^m)$ (or to the strong topology of $L^p(\Omega, \mathbf{R}^m)$), is not the only one. Another idea consists in "enlarging" the space $W^{1,p}(\Omega, \mathbf{R}^m)$ of admissible functions and treating the problem in the space $\mathcal{Y}(\Omega; \mathbf{M}^{m\times N})$ of Young measures introduced in Section 4.3. Actually this is the same general procedure: the integral functionals are considered as living on the space $X = \mathcal{Y}(\Omega; \mathbf{M}^{m\times N})$ equipped with a metrizable topology for which compactness of minimizing sequences also holds. One computes the lsc envelope of F relative to this new space. The two relaxed problems are different but there are some important connections between them. This procedure will be described in detail in Section 11.4.

Let us now make precise the structure of the functional F. We consider a bounded open subset Ω of $\mathbf{R}^N$, sufficiently regular so that the trace theory, the Rellich–Kondrakov theorem, Theorem 5.4.2, and density arguments apply (take, for example, Ω of class $\mathbf{C}^1$). We denote the space of $m \times N$ matrices with entries in $\mathbf{R}$ by $\mathbf{M}^{m\times N}$ and consider a function

$$f : \mathbf{M}^{m\times N} \longrightarrow \mathbf{R}$$

such that there exist three positive constants $\alpha,\ \beta,\ L$ satisfying

$$\forall a \in \mathbf{M}^{m\times N}, \quad \alpha|a|^p \le f(a) \le \beta(1 + |a|^p), \tag{11.5}$$

$$\forall a,\ b \in \mathbf{M}^{m\times N}, \quad |f(a) - f(b)| \le L|b - a|(1 + |a|^{p-1} + |b|^{p-1}). \tag{11.6}$$

Let $W^{1,p}(\Omega, \mathbf{R}^m)$ be the space (isomorphic to $W^{1,p}(\Omega)^m$; see chapter 5), made up of all functions $u : \Omega \longrightarrow \mathbf{R}^m$ whose distributional gradient $\nabla u = (\frac{\partial u_i}{\partial x_j})_{i=1\ldots m,\ j=1\ldots N}$ belongs to

$L^p(\Omega, \mathbf{M}^{m\times N})$. We define the functional $F : L^p(\Omega, \mathbf{R}^m) \longrightarrow \mathbf{R}^+ \cup \{+\infty\}$ by

$$F(u) = \begin{cases} \displaystyle\int_\Omega f(\nabla u)\, dx & \text{if } u \in W^{1,p}(\Omega, \mathbf{R}^m), \\ +\infty & \text{otherwise,} \end{cases} \tag{11.7}$$

and we intend to compute its lsc envelope in the space $X = L^p(\Omega, \mathbf{R}^m)$ equipped with its strong topology. Actually, by classical arguments (the compactness of the embedding of $W^{1,p}(\Omega, \mathbf{R}^m)$ into $L^p(\Omega, \mathbf{R}^m)$ and the lower bound in (11.5)), one can easily establish that the lsc envelope of F considered as living on $W^{1,p}(\Omega, \mathbf{R}^m)$ equipped with its weak topology, coincides with the restriction to $W^{1,p}(\Omega, \mathbf{R}^m)$ of the lsc envelope of F considered here. Proposition 11.2.1 states that the domain of F is not relaxed.

Proposition 11.2.1. *The domain of the functional $cl(F)$ is the space $W^{1,p}(\Omega, \mathbf{R}^m)$.*

Proof. From the inequality $\mathrm{cl}(F) \le F$, we obviously obtain $W^{1,p}(\Omega, \mathbf{R}^m) \subset \mathrm{dom}(\mathrm{cl}(F))$. For the converse inclusion, let $u \in L^p(\Omega, \mathbf{R}^m)$ such that $\mathrm{cl}(F)(u) < +\infty$ and consider a sequence $(u_n)_{n\in\mathbf{N}}$ strongly converging to u in $L^p(\Omega, \mathbf{R}^m)$ and satisfying $\mathrm{cl}(F)(u) = \lim_{n\to+\infty} F(u_n)$. Such a sequence exists from Proposition 11.1.1. According to the lower bound (11.5) and to the equality $\mathrm{cl}(F)(u) = \lim_{n\to+\infty} F(u_n) < +\infty$ one obtains

$$\sup_{n\in\mathbf{N}} \int_\Omega |\nabla u_n|^p\, dx < +\infty,$$

so that $(u_n)_{n\in\mathbf{N}}$ is bounded in $W^{1,p}(\Omega, \mathbf{R}^m)$. According to the Rellich–Kondrakov theorem, Theorem 5.4.2, there exists a subsequence (not relabeled) and $v \in W^{1,p}(\Omega, \mathbf{R}^m)$ such that

$$\begin{aligned} u_n &\rightharpoonup v \text{ weakly in } W^{1,p}(\Omega, \mathbf{R}^m), \\ u_n &\to v \text{ strongly in } L^p(\Omega, \mathbf{R}^m). \end{aligned}$$

Consequently $u = v \in W^{1,p}(\Omega, \mathbf{R}^m)$. □

In Theorem 11.2.1, we establish that the functional $\mathrm{cl}(F)$ possesses an integral representation. In the following proposition, also valid for $p = 1$, we characterize the density of this integral functional. For every bounded Borel set A of $\mathbf{R}^N$, we will sometimes denote its N-dimensional Lebesgue measure by $|A|$ rather than $\mathcal{L}^N(A)$.

Proposition 11.2.2 (quasi-convex envelope of f). *Let us consider a function $f : \mathbf{M}^{m\times N} \longrightarrow \mathbf{R}^+$ satisfying for $p \ge 1$ and all $a \in \mathbf{M}^{m\times N}$ the upper growth condition $0 \le f(a) \le \beta(1 + |a|^p)$, and the continuity assumption* (11.6)*. Then for each fixed a in $\mathbf{M}^{m\times N}$, the map*

$$D \mapsto I_D := \inf\left\{ \frac{1}{|D|} \int_D f(a + \nabla\phi(x))\, dx \;:\; \phi \in W_0^{1,p}(D, \mathbf{R}^m) \right\}$$

is constant on the family of all open bounded subsets of $\mathbf{R}^N$ whose boundary satisfies $|\partial D| = 0$; we denote it by $Qf(a)$. If f satisfies (11.5)*, the function $Qf : \mathbf{M}^{m\times N} \to \mathbf{R}^+$, defined for all $a \in \mathbf{M}^{m\times N}$ by*

$$Qf(a) = \inf\left\{ \frac{1}{|D|} \int_D f(a + \nabla\phi(x))\, dx \;:\; \phi \in W_0^{1,p}(D, \mathbf{R}^m) \right\}$$

satisfies the same condition (11.5) *and* (11.6) *with a new constant L' depending only on α, β, and p. Moreover, Qf is $W^{1,p}$-quasi-convex in the sense of Morrey (quasi-convex in short), namely, it satisfies the so-called quasi-convexity inequality: for all open bounded subset D of $\mathbf{R}^N$ with $|\partial D| = 0$,*

$$\forall a \in \mathbf{M}^{m\times N},\ \forall \phi \in W_0^{1,p}(D, \mathbf{R}^m), \quad Qf(a) \le \frac{1}{|D|}\int_D Qf(a + \nabla\phi(x))\, dx. \tag{11.8}$$

Furthermore, the function Qf is the greatest quasi-convex function less than or equal to f, also called the quasi-convexification or quasi-convex envelope of f.

Proof. (a) Let D and D' be two open bounded subsets of $\mathbf{R}^N$ with $|\partial D| = |\partial D'| = 0$. For proving the first assertion, it suffices to establish $I_D \le I_{D'}$ and to invert the roles of D and D'.

For every $\varepsilon > 0$ there exists a finite family $(x_i + \varepsilon_i D')_{i\in I_\varepsilon}$ of pairwise disjoint sets $x_i + \varepsilon_i D' \subset D$, $\varepsilon_i > 0$, satisfying

$$\left| D \setminus \bigcup_{i\in I_\varepsilon} (x_i + \varepsilon_i D') \right| < \varepsilon. \tag{11.9}$$

First, we claim that the map I verifies the following subadditivity property: if A and B are two disjoint bounded open subsets of $\mathbf{R}^N$, then

$$|A \cup B| I_{A\cup B} \le |A| I_A + |B| I_B. \tag{11.10}$$

Indeed, let ϕ_A and ϕ_B be two η-minimizers of $|A|I_A$ and $|B|I_B$ in $\mathcal{D}(A, \mathbf{R}^m)$ and $\mathcal{D}(B, \mathbf{R}^m)$, respectively, extended by 0 on $\mathbf{R}^N \setminus A$ and $\mathbf{R}^N \setminus B$. We have

$$|A| I_A \ge \int_A f(a + \nabla\phi_A)\, dx - \eta,$$

$$|B| I_B \ge \int_B f(a + \nabla\phi_B)\, dx - \eta.$$

Such η-minimizers exist thanks to (11.6) and by a density argument. The function ϕ which coincides with ϕ_A and ϕ_B on A and B, respectively, belongs to $W_0^{1,p}(A \cup B, \mathbf{R}^m)$ so that

$$\begin{aligned} |A \cup B| I_{A\cup B} &\le \int_{A\cup B} f(a + \nabla\phi)\, dx \\ &= \int_A f(a + \nabla\phi_A)\, dx + \int_B f(a + \nabla\phi_B)\, dx \\ &\le |A| I_A + |B| I_B + 2\eta. \end{aligned}$$

The thesis is obtained after making $\eta \to 0$. Using quite similar arguments, one also obtains

$$|A| I_A \le |A \setminus B| I_{A\setminus\overline{B}} + |B| I_B \tag{11.11}$$

whenever A and B are two open bounded subsets of $\mathbf{R}^N$ with $B \subset A$ and $|\partial B| = 0$.

Applying, respectively, (11.10) and (11.11) to the finite union $\bigcup_{i\in I_\varepsilon}(x_i+\varepsilon_i D')$ and to $A=D$, $B=\bigcup_{i\in I_\varepsilon}(x_i+\varepsilon_i D')$, according to (11.9) and to the growth condition (11.5), we obtain

$$\left|\bigcup_{i\in I_\varepsilon}(x_i+\varepsilon_i D')\right| I_{\bigcup_{i\in I_\varepsilon}(x_i+\varepsilon_i D')} \le \sum_{i\in I_\varepsilon} |\varepsilon_i D'| I_{x_i+\varepsilon_i D'}, \tag{11.12}$$

$$|D| I_D \le \beta(1+|a|^p)\varepsilon + \left|\bigcup_{i\in I_\varepsilon}(x_i+\varepsilon_i D')\right| I_{\bigcup_{i\in I_\varepsilon}(x_i+\varepsilon_i D')}\ . \tag{11.13}$$

A change of scale easily gives $I_{x_i+\varepsilon_i D'} = I_{D'}$. Combining now (11.12) and (11.13), one obtains

$$I_D \le \frac{\beta(1+|a|^p)\varepsilon}{|D|} + I_{D'}.$$

Since ε is arbitrary we have indeed established $I_D \le I_{D'}$. It is worth noticing that the above subadditivity argument is a particular case of a more general result related to subadditive ergodic processes (see Krengel [165], Dal Maso and Modica [119], or Licht and Michaille [170]). Such a general argument will be used in chapter 12.

(b) We assume that f satisfies (11.5) and show that Qf satisfies the same conditions. Taking $D=Y=(0,1)^N$ in the definition of Qf, from (11.5), we have

$$\begin{aligned} Qf(a) &\ge \alpha \inf\left\{\int_Y |a+\nabla\phi|^p\, dx \ :\ \phi\in W_0^{1,p}(Y,\mathbf{R}^m)\right\} \\ &\ge \alpha \inf\left\{\left|\int_Y (a+\nabla\phi)\, dx\right|^p : \phi\in W_0^{1,p}(Y,\mathbf{R}^m)\right\} \\ &= \alpha|a|^p. \end{aligned}$$

We have used Jensen's inequality satisfied by the convex function $a\mapsto |a|^p$ in the second inequality above. The upper bound of (11.5) is trivially obtained by taking $\phi=0$ as admissible function in the expression of the infima and by using the upper bound condition satisfied by f.

Let us now establish (11.6). Given arbitrary $\eta>0$, let $\phi_\eta\in W_0^{1,p}(Y,\mathbf{R}^m)$ be such that

$$Qf(b) \ge \int_Y f(b+\nabla\phi_\eta)\, dx - \eta.$$

From (11.6) and Hölder's inequality, we obtain

$$\begin{aligned} &Qf(a)-Qf(b) \\ &\le \int_Y f(a+\nabla\phi_\eta)\, dx - \int_Y f(b+\nabla\phi_\eta)\, dx + \eta \\ &\le \int_Y |\ f(a+\nabla\phi_\eta)-f(b+\nabla\phi_\eta)\ |\, dx + \eta \\ &\le L\ |\ a-b\ |\left(\int_Y (1+|\ a+\nabla\phi_\eta(x)\ |^{p-1}+|\ b+\nabla\phi_\eta(x)\ |^{p-1})^{\frac{p}{p-1}} dx\right)^{\frac{p-1}{p}} + \eta \\ &\le CL\ |\ a-b\ |\left(\int_Y (1+|\ a\ |^p+|\ b\ |^p+|\ b+\nabla\phi_\eta(x)\ |^p) dx\right)^{\frac{p-1}{p}} + \eta, \end{aligned} \tag{11.14}$$

where C is a constant depending only on p. On the other hand, from (11.5),

$$\begin{aligned}\int_Y \mid b + \nabla\phi_\eta(x) \mid^p dx &\leq \frac{1}{\alpha}\int_Y f(b + \nabla\phi_\eta)\, dx \\ &\leq \frac{1}{\alpha}(Qf(b) + \eta) \\ &\leq \frac{\beta}{\alpha}(1 + \mid b \mid^p) + \frac{\eta}{\alpha}. \end{aligned} \tag{11.15}$$

Combining (11.14) and (11.15) and letting $\eta \to 0$, we obtain

$$Qf(a) - Qf(b) \leq L' \mid b - a \mid (1 + \mid a \mid^{p-1} + \mid b \mid^{p-1}),$$

where L' is a constant which depends only on $p,\ \alpha\ \beta$. We end the proof by interchanging the roles of a and b.

(c) We establish the quasi-convex inequality. Let us set

$$\mathrm{Aff}_0(D, \mathbf{R}^m) := \left\{\phi \in W_0^{1,p}(D, \mathbf{R}^m) \ :\ \phi \text{ piecewise affine } \right\}.$$

Note that according to the density of $\mathrm{Aff}_0(D, \mathbf{R}^m)$ in $W_0^{1,p}(D, \mathbf{R}^m)$ equipped with its strong topology, and to (11.6), we have

$$\forall a \in \mathbf{M}^{m\times N}, \quad Qf(a) = \inf\left\{\frac{1}{|D|}\int_D f(a + \nabla\phi)\, dx \ :\ \phi \in \mathrm{Aff}_0(D, \mathbf{R}^m)\right\}.$$

It is now easily seen that the quasi-convex inequality is satisfied with every $\phi \in \mathrm{Aff}_0(D, \mathbf{R}^m)$. Indeed, since $\phi \in \mathrm{Aff}_0(D, \mathbf{R}^m)$, there exist some open bounded and pairwise disjoint sets $D_i \subset D,\ i = 1, \ldots, r$, with $\bar{D} = \cup_{i=1}^r \bar{D}_i$, $|\partial D_i| = 0$, and $a_i \in \mathbf{M}^{m\times N},\ i = 1, \ldots, r$, such that $\nabla\phi \equiv a_i$ on D_i, which clearly implies

$$\int_D Qf(a + \nabla\phi)\, dx = \sum_{i=1}^r Qf(a + a_i)|D_i|. \tag{11.16}$$

On the other hand, for $\eta > 0$ and $i = 1, \ldots, r$, there exists $\phi_{i,\eta} \in \mathrm{Aff}_0(D_i, \mathbf{R}^m)$ such that

$$Qf(a + a_i) \geq \frac{1}{|D_i|}\int_{D_i} f(a + a_i + \nabla\phi_{i,\eta})\, dx - \eta. \tag{11.17}$$

Let us consider the function $\tilde{\phi}$ defined on D by

$$\tilde{\phi}(x) = \phi(x) + \phi_{i,\eta}(x) \text{ when } x \in D_i.$$

Clearly $\tilde{\phi}$ belongs to $W_0^{1,p}(D, \mathbf{R}^m)$ (actually to $\mathrm{Aff}_0(D, \mathbf{R}^m)$). Summing inequalities (11.17) for $i = 1, \ldots, r$, equality (11.16) yields

$$\begin{aligned}\int_D Qf(a + \nabla\phi)\, dx &\geq \int_D f(a + \nabla\tilde{\phi})\, dx - \eta|D| \\ &\geq Qf(a)|D| - \eta|D|. \end{aligned}$$

Letting $\eta \to 0$, we obtain the quasi-convex inequality for every $\phi \in \mathrm{Aff}_0(D, \mathbf{R}^m)$. Thanks to the density of $\mathrm{Aff}_0(D, \mathbf{R}^m)$ in $W_0^{1,p}(D, \mathbf{R}^m)$ and to (11.6) satisfied by Qf, the quasi-convex inequality is now satisfied for any $\phi \in W_0^{1,p}(D, \mathbf{R}^m)$.

It remains to establish that Qf is the greatest quasi-convex function less than or equal to f. First notice that $g \leq f$ yields $Qg \leq Qf$. On the other hand, if g is quasi-convex, $Qg = g$. Indeed $Qg \leq g$ by definition and inequality (11.8) satisfied by g gives the converse inequality. We then obtain from $g \leq f$ and the quasi-convexity of g, that $g = Qg \leq Qf$. □

Remark 11.2.1. One can prove that Qf is the quasi-convex envelope of f under less restrictive assumptions on f, as for example, without continuity condition (see Dacorogna [116, Theorem 1.1] and [57]). For the relationship between the notions of convexity, polyconvexity, rank-one convexity, and various examples, consult Dacorogna [116] and Sverak [211], [212], [213].

One can establish that Qf is convex on each interval $[a, b]$ in $\mathbf{M}^{m\times N}$ satisfying $\mathrm{rank}(a - b) = 1$, i.e., is rank-one convex. Consequently, when $m = 1$ or $N = 1$, Qf is a convex function and actually $Qf = f^{**}$. For a proof, consult Step 3′ in the proof of Theorem 1.1 in Dacorogna [116]. Nevertheless, when $m > 1$ or $N > 1$, Qf is not in general the convexification of f. Consequently, the optimization problems related to integral functionals F with a convex density f are not well posed. One needs to use a relaxation procedure in the sense of the relaxation theorem, Theorem 11.1.2.

Let $f : \mathbf{M}^{m\times N} \longrightarrow \mathbf{R}$ satisfying (11.5) and (11.6). In Theorem 13.2.1, we will establish, in a more general situation, that the quasi-convex inequality (11.8) is a necessary and sufficient condition to ensure the lower semicontinuity of the integral functional $u \mapsto \int_\Omega f(\nabla u)\, dx$ defined on $W^{1,p}(\Omega, \mathbf{R}^m)$ equipped with its weak topology.

We now state the main result of this section.

Theorem 11.2.1. *Let us consider a function $f : \mathbf{R}^N \longrightarrow \mathbf{R}$ satisfying* (11.5) *and* (11.6) *with $p > 1$, and F the associated integral functional* (11.7) *defined in $L^p(\Omega, \mathbf{R}^m)$ equipped with its strong topology. Then the lsc envelope of F is given, for every u in $L^p(\Omega, \mathbf{R}^m)$, by*

$$cl(F)(u) = \begin{cases} \displaystyle\int_\Omega Qf(\nabla u)\, dx \text{ if } u \in W^{1,p}(\Omega, \mathbf{R}^m), \\ +\infty \quad \textit{otherwise.} \end{cases}$$

The proof of Theorem 11.2.1 is the straightforward consequence of Propositions 11.2.3 and 11.2.4. We adopt the following notation:

$$QF(u) = \begin{cases} \displaystyle\int_\Omega Qf(\nabla u)\, dx \text{ if } u \in W^{1,p}(\Omega, \mathbf{R}^m), \\ +\infty \quad \text{otherwise.} \end{cases}$$

The proofs given here have the advantage of being easily adapted to the theory of homogenization (see the next chapter).

Proposition 11.2.3. *For every u in $L^p(\Omega, \mathbf{R}^m)$, $p > 1$, and every sequence $(u_n)_{n\in\mathbf{N}}$ strongly converging to u in $L^p(\Omega, \mathbf{R}^m)$, one has*

$$QF(u) \le \liminf_{n\to+\infty} F(u_n). \tag{11.18}$$

Assume that f satisfies (11.6) *and only the upper growth condition $0 \le f(a) \le \beta(1+|a|^p)$ for all $a \in \mathbf{M}^{m\times N}$. Then, for every u in $W^{1,p}(\Omega, \mathbf{R}^m)$ and every sequence $(u_n)_{n\in\mathbf{N}}$ weakly converging to u in $W^{1,p}(\Omega, \mathbf{R}^m)$, one has*

$$QF(u) \le \liminf_{n\to+\infty} F(u_n).$$

PROOF. We assume that f satisfies (11.5) and (11.6) and we establish the first assertion. The second assertion will be obtained at the end of the proof as a straightforward consequence. Obviously, one can assume $\liminf_{n\to+\infty} F(u_n) < +\infty$ so that u belongs to $W^{1,p}(\Omega, \mathbf{R}^m)$. For a nonrelabeled subsequence, consider the nonnegative Borel measure $\mu_n := f(\nabla u_n(.))\mathcal{L}^N \lfloor \Omega$. We have

$$\sup_{n\in\mathbf{N}} \mu_n(\Omega) < +\infty.$$

Consequently there exists a further subsequence (not relabeled) and a nonnegative Borel measure $\mu \in \mathbf{M}(\Omega)$ such that

$$\mu_n \rightharpoonup \mu \text{ weakly in } \mathbf{M}(\Omega).$$

Let $\mu = g\mathcal{L}^N \lfloor \Omega + \mu_s$ be the Lebesgue–Nikodym decomposition of μ, where μ_s is a nonnegative Borel measure, singular with respect to the Lebesgue measure $\mathcal{L}^N \lfloor \Omega$. For establishing (11.18) it is enough to prove that for a.e. $x \in \Omega$,

$$g(x) \ge Qf(\nabla u(x)).$$

Indeed, according to Alexandrov's theorem, Proposition 4.2.3, we will obtain

$$\begin{aligned}\liminf_{n\to+\infty} F(u_n) = \liminf_{n\to+\infty} \mu_n(\Omega) \ge \mu(\Omega) &= \int_\Omega g(x)\,dx + \mu_s(\Omega) \\ &\ge \int_\Omega g(x)\,dx \\ &\ge \int_\Omega Qf(\nabla u(x))\,dx.\end{aligned}$$

Let $\rho > 0$ intended to tend to 0 and denote the open ball of radius ρ centered at x_0 by $B_\rho(x_0)$. According to the theory of differentiation of measures (see Theorem 4.2.1), there exists a negligible set N for the measure $\mathcal{L}^N \lfloor \Omega$, such that for all $x_0 \in \Omega \setminus N$,

$$g(x_0) = \lim_{\rho\to 0} \frac{\mu(B_\rho(x_0))}{|B_\rho(x_0)|}.$$

Applying Lemma 4.2.1, for all but countably many $\rho > 0$, one may assume $\mu(\partial B_\rho(x_0)) = 0$. From Alexandrov's theorem, Proposition 4.2.3, we then obtain $\mu(B_\rho(x_0)) = \lim_{n\to+\infty} \mu_n(B_\rho(x_0))$ and we finally are reduced to establishing

$$\lim_{\rho\to 0} \lim_{n\to+\infty} \frac{\mu_n(B_\rho(x_0))}{|B_\rho(x_0)|} \geq Qf(\nabla u(x_0)) \quad \text{for } x_0 \in \Omega \setminus N. \tag{11.19}$$

Let us assume for the moment that the trace of u_n on $\partial B_\rho(x_0)$ is equal to the affine function u_0 defined by $u_0(x) := u(x_0) + \langle \nabla u(x_0), x - x_0 \rangle$. It follows that

$$\begin{aligned}
\lim_{n\to+\infty} \frac{\mu_n(B_\rho(x_0))}{|B_\rho(x_0)|} &= \lim_{n\to+\infty} \frac{1}{|B_\rho(x_0)|} \int_{B_\rho(x_0)} f(\nabla u(x_0) + \nabla(u_n - u_0))\, dx \\
&\geq \inf \left\{ \frac{1}{|B_\rho(x_0)|} \int_{B_\rho(x_0)} f(\nabla u(x_0) + \nabla\phi)\, dx : \phi \in W_0^{1,p}(B_\rho(x_0), \mathbf{R}^m) \right\} \\
&= Qf(\nabla u(x_0))
\end{aligned}$$

and the proof would be complete. Note that the small size of the radius ρ of $B_\rho(x_0)$ does not contribute to the estimate above. The idea now consists in modifying u_n by a function of $W^{1,p}(B_\rho(x_0), \mathbf{R}^m)$ which coincides with u_0 on $\partial B_\rho(x_0)$ in the trace sense, to follow the previous procedure and to control additional terms, when ρ goes to zero, thanks to the following classical estimate.

Lemma 11.2.1. *For every u in $W^{1,p}(\Omega, \mathbf{R}^m)$, $p \geq 1$, there exists a negligible set N' such that for all x_0 in $\Omega \setminus N'$ we have*

$$\left[\frac{1}{|B_\rho(x_0)|} \int_{B_\rho(x_0)} |u(x) - \big(u(x_0) + \nabla u(x_0)(x - x_0)\big)|^p\, dx \right]^{1/p} = o(\rho). \tag{11.20}$$

Proof. For the proof we refer the reader to Ziemer [229, Theorem 3.4.2]. □

From now on, we fix x_0 in the set $\Omega \setminus N \cup N'$. To suitably modify u_n on the boundary of $B_\rho(x_0)$, we use a well-known method due to De Giorgi. A neighborhood of the boundary of $B_\rho(x_0)$ is sliced as follows: let $\nu \in \mathbf{N}^*$, $\rho_0 := \lambda\rho$, where $0 < \lambda < 1$, and set

$$B_i := B_{\rho_0 + i\frac{\rho-\rho_0}{\nu}}(x_0) \text{ for } i = 0, \ldots, \nu.$$

On the other hand, consider for $i = 1, \ldots, \nu$,

$$\varphi_i \in \mathbf{C}_0^\infty(B_i),\ 0 \leq \varphi_i \leq 1,\ \varphi_i = 1 \text{ in } B_{i-1},\ |grad\varphi_i|_{L^\infty(B_i)} \leq \frac{\nu}{\rho - \rho_0} = \frac{\nu}{\rho(1-\lambda)}$$

and define $u_{n,i} \in W^{1,p}(B_\rho(x_0), \mathbf{R}^m)$ by

$$u_{n,i} := u_0 + \varphi_i(u_n - u_0).$$

For $i = 1, \ldots, \nu$, we have

$$
\begin{aligned}
&Qf(\nabla u(x_0))\\
&=\inf\left\{\frac{1}{|B_\rho(x_0)|}\int_{B_\rho(x_0)} f(\nabla u(x_0)+\nabla\phi)\,dx : \phi\in W_0^{1,p}(B_\rho(x_0),\mathbf{R}^m)\right\}\\
&\le \frac{1}{|B_\rho(x_0)|}\int_{B_\rho(x_0)} f(\nabla u_{n,i})\,dx\\
&=\frac{1}{|B_\rho(x_0)|}\int_{B_{i-1}} f(\nabla u_n)\,dx+\frac{1}{|B_\rho(x_0)|}\int_{B_i\setminus B_{i-1}} f(\nabla u_{n,i})\,dx\\
&\qquad\qquad+\frac{1}{|B_\rho(x_0)|}\int_{B_\rho(x_0)\setminus B_i} f(\nabla u(x_0))\,dx\\
&\le \frac{1}{|B_\rho(x_0)|}\int_{B_\rho(x_0)} f(\nabla u_n)\,dx+\frac{1}{|B_\rho(x_0)|}\int_{B_i\setminus B_{i-1}} f(\nabla u_{n,i})\,dx\\
&\qquad\qquad+\beta\big(1+|\nabla u(x_0)|^p\big)(1-\lambda)^N. \qquad (11.21)
\end{aligned}
$$

Let us estimate the second term of the right-hand side of (11.21). From the growth condition (11.5) we obtain

$$
\begin{aligned}
&\frac{1}{|B_\rho(x_0)|}\int_{B_i\setminus B_{i-1}} f(\nabla u_{n,i})dx \le C\big(1+|\nabla u(x_0)|^p\big)(1-\lambda)^N\\
&+\frac{C}{|B_\rho(x_0)|}\int_{B_i\setminus B_{i-1}}|\nabla(u_n-u_0)|^p\,dx+\frac{C}{|B_\rho(x_0)|}\frac{\nu^p}{\rho^p(1-\lambda)^p}\int_{B_i\setminus B_{i-1}}|u_n-u_0|^p\,dx,
\end{aligned}
$$

where C is a positive constant depending only on p and β. Then, averaging inequalities (11.21), we obtain

$$
\begin{aligned}
Qf(\nabla u(x_0)) &= \frac{1}{\nu}\sum_{i=1}^{\nu} Qf(\nabla u(x_0))\\
&\le \frac{1}{|B_\rho(x_0)|}\int_{B_\rho(x_0)} f(\nabla u_n)\,dx+\frac{C\nu^{p-1}}{|B_\rho(x_0)|\rho^p(1-\lambda)^p}\int_{B_\rho(x_0)}|u_n-u_0|^p\,dx\\
&+C\big(1+|\nabla u(x_0)|^p\big)(1-\lambda)^N+\frac{1}{\nu}\frac{C}{|B_\rho(x_0)|}\int_{B_\rho(x_0)}|\nabla u_n|^p\,dx\\
&\le \frac{\frac{C}{\alpha\nu}+1}{|B_\rho(x_0)|}\int_{B_\rho(x_0)} f(\nabla u_n)\,dx+\frac{C\nu^{p-1}}{|B_\rho(x_0)|\rho^p(1-\lambda)^p}\int_{B_\rho(x_0)}|u_n-u_0|^p\,dx\\
&+C\big(1+|\nabla u(x_0)|^p\big)(1-\lambda)^N, \qquad (11.22)
\end{aligned}
$$

where we have used the lower bound (11.5) in the last inequality. Letting $n\to+\infty$ and $\rho\to 0$, from Lemma 11.2.1 we obtain

$$
Qf(\nabla u(x_0))\le\left(\frac{C}{\alpha\nu}+1\right)\lim_{\rho\to 0}\lim_{n\to+\infty}\frac{\mu_n(B_\rho(x_0))}{|B_\rho(x_0)|}+C\big(1+|\nabla u(x_0)|^p\big)(1-\lambda)^N
$$

and (11.19) follows after letting $\lambda\to 1$ and $\nu\to+\infty$. It is worth noticing that we have used the slicing method to control the term $\frac{1}{|B_\rho(x_0)|}\int_{B_\rho(x_0)}|\nabla u_n|^p\,dx$ by letting the slices become increasingly thinner (i.e., $\nu\to+\infty$).

If now f does not verify the lower bound in (11.5), we first note that the weak convergence of u_n to u in $W^{1,p}(\Omega, \mathbf{R}^m)$ yields the boundedness of

$$\sup_{n\in\mathbf{N}} \int_{B_\rho(x_0)} |\nabla u_n|^p \, dx < +\infty.$$

Moreover, according to the Rellich–Kondrakov compact embedding theorem, Theorem 5.4.2, $u_n \to u$ strongly in $L^p(\Omega, \mathbf{R}^m)$. Then, to conclude, it suffices going to the limit respectively on n and ρ in (11.22) and to let $\lambda \to 1$, $\nu \to +\infty$ as above. □

Proposition 11.2.4. *For every u in $L^p(\Omega, \mathbf{R}^m)$, $p > 1$, there exists a sequence $(u_n)_{n\in\mathbf{N}}$ strongly converging to u in $L^p(\Omega, \mathbf{R}^m)$ such that*

$$QF(u) \geq \limsup_{n\to+\infty} F(u_n).$$

Proof. One can assume $QF(u) < +\infty$. Therefore, taking $D = Y := (0,1)^N$ in the definition of Qf, and $\eta = 1/k$, $k \in \mathbf{N}^*$

$$\begin{aligned} QF(u) &= \int_\Omega Qf(\nabla u)\, dx \\ &= \int_\Omega \inf\left\{ \int_Y f(\nabla u(x) + \nabla_y \phi(y))\, dy \ : \ \phi \in W_0^{1,p}(Y, \mathbf{R}^m) \right\} dx \\ &\geq \int_{\Omega\times Y} f(\nabla u(x) + \nabla_y \phi_\eta(x, y))\, dxdy - \eta|\Omega|, \end{aligned} \tag{11.23}$$

where $\phi_\eta(x, .)$ is a η-minimizer of $\inf\left\{ \int_Y f(\nabla u(x) + \nabla_y \phi(y))\, dy \ : \ \phi \in W_0^{1,p}(Y, \mathbf{R}^m)\right\}$. We admit that we can select a measurable map $x \mapsto \phi_\eta(x, .)$ from Ω into $W_0^{1,p}(Y, \mathbf{R}^m)$. For a proof, consult Castaing and Valadier [105]. We claim that ϕ_η belongs to $L^p(\Omega, W_0^{1,p}(Y, \mathbf{R}^m)$. Indeed,

$$\begin{aligned} \int_\Omega \|\nabla_y \phi_\eta(x, .)\|^p_{L^p(Y, \mathbf{M}^{m\times N})}\, dx &= \int_{\Omega\times Y} |\nabla_y \phi_\eta(x, y)|^p \, dxdy \\ &\leq C\left(\int_{\Omega\times Y} |\nabla u(x) + \nabla_y \phi_\eta(x, y)|^p \, dxdy + \int_\Omega |\nabla u|^p \, dx \right) \\ &\leq C\left(\frac{1}{\alpha} \int_{\Omega\times Y} f(\nabla u(x) + \nabla_y \phi_\eta(x, y))\, dxdy + \int_\Omega |\nabla u|^p \, dx \right) \\ &\leq C\left(\frac{1}{\alpha} QF(u) + \frac{\eta|\Omega|}{\alpha} + \int_\Omega |\nabla u|^p \, dx \right) < +\infty, \end{aligned}$$

where C is a positive constant depending only on p.

Classically $\mathbf{C}_c(\Omega, \mathcal{D}(Y, \mathbf{R}^m))$ is dense in $L^p(\Omega, W_0^{1,p}(Y, \mathbf{R}^m)$. Consequently, from the continuity assumption (11.6) fulfilled by f, it is easily seen that (11.23) yields

$$QF(u) = \int_\Omega Qf(\nabla u)\, dx \geq \int_{\Omega\times Y} f(\nabla u(x) + \nabla_y \tilde{\phi}_\eta(x, y))\, dxdy - 2\eta|\Omega| \tag{11.24}$$

for some $\tilde{\phi}_\eta$ in $\mathbf{C}_c(\Omega, \mathcal{D}(Y, \mathbf{R}^m))$. We have actually established the following interchange result between infimum and integral:

$$\begin{aligned}
&\int_\Omega \left(\inf_{\phi \in W_0^{1,p}(Y,\mathbf{R}^m)} \int_Y f(\nabla u(x) + \nabla_y \phi(y))\, dy \right) dx \\
&= \inf_{\Phi \in L^p(\Omega, W_0^{1,p}(Y,\mathbf{R}^m))} \int_{\Omega \times Y} f(\nabla u(x) + \nabla_y \Phi(x, y))\, dxdy \\
&= \inf_{\Phi \in \mathbf{C}_c(\Omega, \mathcal{D}(Y,\mathbf{R}^m))} \int_{\Omega \times Y} f(\nabla u(x) + \nabla_y \Phi(x, y))\, dxdy.
\end{aligned}$$

Indeed from (11.23)

$$\begin{aligned}
&\int_\Omega \left(\inf_{\phi \in W_0^{1,p}(Y,\mathbf{R}^m)} \int_Y f(\nabla u(x) + \nabla_y \phi(y))\, dy \right) dx \\
&\geq \inf_{\Phi \in L^p(\Omega, W_0^{1,p}(Y,\mathbf{R}^m))} \int_{\Omega \times Y} f(\nabla u(x) + \nabla_y \Phi(x, y))\, dxdy
\end{aligned}$$

and the converse inequality is trivial. Note that this result also holds for $p = 1$. Because of its importance, we state it in a slightly more general form.

Lemma 11.2.2. *Let $f : \mathbf{M}^{m\times N} \to \mathbf{R}^+$ be any function satisfying* (11.5) *and* (11.6) *with $p \geq 1$. Let moreover ξ be any element of $L^p(\Omega, \mathbf{M}^{m\times N})$. Then*

$$\begin{aligned}
&\int_\Omega \left(\inf_{\phi \in W_0^{1,p}(Y,\mathbf{R}^m)} \int_Y f(\xi(x) + \nabla_y \phi(y))\, dy \right) dx \\
&= \inf_{\Phi \in L^p(\Omega, W_0^{1,p}(Y,\mathbf{R}^m))} \int_{\Omega \times Y} f(\xi(x) + \nabla_y \Phi(x, y))\, dxdy \\
&= \inf_{\Phi \in \mathbf{C}_c(\Omega, \mathcal{D}(Y,\mathbf{R}^m))} \int_{\Omega \times Y} f(\xi(x) + \nabla_y \Phi(x, y))\, dxdy.
\end{aligned}$$

For more about interchange theorems, consult, for instance, [26].

PROOF OF PROPOSITION 11.2.4 CONTINUED. Let us go back to (11.24). To shorten the notation we denote the previous η-minimizer $\tilde{\phi}_\eta$ in $\mathbf{C}_c(\Omega, \mathcal{D}(Y, \mathbf{R}^m))$ by ϕ_η and extend $y \mapsto \phi_\eta(x, y)$ by Y-periodicity on $\mathbf{R}^N$. Consider now the function $u_{\eta,n}$ defined by

$$u_{\eta,n}(x) = u(x) + \frac{1}{n}\phi_\eta(x, nx).$$

Note that ϕ_η is a Carathéodory function so that $x \mapsto \phi_\eta(x, nx)$ is measurable. Clearly $u_{\eta,n}$ belongs to $W^{1,p}(\Omega, \mathbf{R}^m)$ and $u_{\eta,n} \to u$ strongly in $L^p(\Omega, \mathbf{R}^m)$ when $n \to +\infty$. It's indeed a straightforward consequence of

$$\begin{aligned}
\int_\Omega |\phi_\eta(x, nx)|^p\, dx &\leq \int_\Omega \sup_{y\in Y} |\phi_\eta(x, y)|^p\, dx \\
&\leq |\Omega| \sup_{x\in\Omega} \sup_{y\in Y} |\phi_\eta(x, y)|^p < +\infty.
\end{aligned}$$

On the other hand,

$$\begin{aligned}\lim_{n\to+\infty}\int_\Omega f(\nabla u_{\eta,n})\,dx &= \lim_{n\to+\infty}\int_\Omega f(\nabla u(x)+(\nabla_y\phi_\eta)(x,nx)+\frac{1}{n}\nabla_x\phi_\eta(x,nx))\,dx\\ &= \lim_{n\to+\infty}\int_\Omega f(\nabla u(x)+(\nabla_y\phi_\eta)(x,nx))\,dx\\ &= \int_{\Omega\times Y} f(\nabla u(x)+\nabla_y\phi_\eta(x,y))\,dxdy.\end{aligned}$$

For passing from the first to the second equality we have used the continuity assumption (11.6) on f. The last equality is a consequence of Lemma 11.2.3 stated at the end of the proof and applied to $g(x,y)=f(\nabla u(x)+\nabla_y\phi_\eta(x,y))$ (note that $y\mapsto f(\nabla u(x)+\nabla_y\phi_\eta(x,y))$ belongs to the set $\mathbf{C}_\#(Y)$ of all the restrictions to Y of continuous and Y-periodic functions on $\mathbf{R}^N$ and that $g\in L^1(\Omega,\mathbf{C}_\#(Y))$). Consequently, from (11.24)

$$\begin{aligned}\lim_{n\to+\infty}F(u_{\eta,n}) &= \lim_{n\to+\infty}\int_\Omega f(\nabla u(x)+(\nabla_y\phi_\eta)(x,nx))\,dx\\ &= \int_{\Omega\times Y} f(\nabla u(x)+\nabla_y\phi_\eta(x,y))\,dxdy\\ &\le QF(u)+2\eta.\end{aligned}$$

Letting $\eta\to 0$ (i.e., $k\to+\infty$), up to a subsequence, one obtains

$$\lim_{\eta\to+0}\lim_{n\to+\infty}F(u_{\eta,n})\le QF(u).$$

We apply now the diagonalization Lemma 11.1.1 for the sequence $\big(F(u_{\eta,n}),u_{\eta,n}\big)_{\eta,n}$ in the metric space $\mathbf{R}\times L^p(\Omega,\mathbf{R}^m)$: there exists a map $n\mapsto\eta(n):=\frac{1}{k(n)}$ such that

$$\begin{aligned}&\lim_{n\to+\infty}F(u_{\eta(n),n})\le QF(u),\\ &\lim_{n\to+\infty}u_{\eta(n),n}=u \text{ strongly in } L^p(\Omega,\mathbf{R}^m).\end{aligned}$$

The sequence $(u_n)_{n\in\mathbf{N}}$ where $u_n=u_{\eta(n),n}$ then verifies the assertion of Proposition 11.2.4.

We now state and prove Lemma 11.2.3 invoked above. Let D be any open cube of $\mathbf{R}^N$. We recall that $\mathbf{C}_\#(D)$ denotes the set of all the restrictions to D of continuous and D-periodic functions on $\mathbf{R}^N$, equipped with the uniform norm on D, and that $L^1(\Omega,\mathbf{C}_\#(D))$ denotes the space of all measurable functions h from Ω into $\mathbf{C}_\#(D)$ satisfying

$$\int_\Omega\sup_{y\in D}|h(x,y)|\,dx<+\infty.$$

Lemma 11.2.3. *For every function g in $L^1(\Omega,\mathbf{C}_\#(D))$,*

$$\lim_{n\to+\infty}\int_\Omega g(x,nx)\,dx=\frac{1}{|D|}\int_{\Omega\times D}g(x,y)\,dxdy. \tag{11.25}$$

Proof. It is well known (see, for instance, Yosida [225]) that if g belongs to $L^1(\Omega,\mathbf{C}_\#(D))$, then g is a Carathéodory function satisfying $\sup_{y\in D}|g(.,y)|\in L^1(\Omega)$ so that $\int_\Omega g(x,nx)\,dx$

is well defined. For $k \in \mathbf{N}^*$, let us decompose the cube D as follows:

$$\overline{D} = \bigcup_{i=1}^{k^N} \overline{D}_i,$$

where D_i are small pairwise disjoint open cubes $1/k$-homothetic of D. We approximate g in $L^1(\Omega, \mathbf{C}_{\#}(D))$ by the following step function g_k:

$$g_k(x, y) = \sum_{i=1}^{k^N} g(x, y_i) 1_{D_i}(y),$$

where 1_{D_i} is the characteristic function of the set D_i extended by D-periodicity on $\mathbf{R}^N$ and y_i is any fixed element of D_i. Due to the periodicity of 1_{D_i}, classically $x \mapsto 1_{D_i}(nx)$ $\sigma(L^\infty, L^1)$ weakly converges to $|D_i|/|D|$ so that

$$\lim_{n \to +\infty} \int_\Omega g(x, y_i) 1_{D_i}(nx)\, dx = \frac{|D_i|}{|D|} \int_\Omega g(x, y_i)\, dx.$$

For a proof, see Example 2.4.2 or Proposition 13.2.1 and the proof of Theorem 13.2.1. Summing these equalities over $i = 1, \ldots, k^N$, one obtains that (11.25) is satisfied for g_k.

To conclude by going to the limit on k, we use the uniform bound with respect to n:

$$\int_\Omega |g(x, nx)\, dx - g_k(x, nx)|\, dx \le \|g - g_k\|_{L^1(\Omega, \mathbf{C}_{\#}(D))}. \tag{11.26}$$

Moreover, since

$$\sup_{y \in D} |g_k(x, y) - g(x, y)| \le 2 \sup_{y \in D} |g(x, y)| \in L^1(\Omega)$$

and $\lim_{k \to +\infty} \sup_{y \in D} |g_k(x, y) - g(x, y)| = 0$, according to the Lebesgue dominated convergence theorem, we obtain

$$\lim_{k \to +\infty} \|g - g_k\|_{L^1(\Omega, \mathbf{C}_{\#}(D))} = 0. \tag{11.27}$$

Thus, from (11.26)

$$\begin{aligned}
\left| \int_\Omega g(x, nx)\, dx - \frac{1}{|D|} \int_{\Omega \times D} g(x, y)\, dxdy \right| &\le \left| \int_\Omega g(x, nx)\, dx - \int_\Omega g_k(x, nx)\, dx \right| \\
&+ \left| \int_\Omega g_k(x, nx)\, dx - \frac{1}{|D|} \int_{\Omega \times D} g_k(x, y)\, dxdy \right| \\
&+ \left| \frac{1}{|D|} \int_{\Omega \times D} g_k(x, y)\, dxdy - \frac{1}{|D|} \int_{\Omega \times D} g(x, y)\, dxdy \right| \\
&\le \max\left(1, \frac{1}{|D|}\right) \|g - g_k\|_{L^1(\Omega, \mathbf{C}_{\#}(D))} \\
&+ \left| \int_\Omega g_k(x, nx)\, dx - \frac{1}{|D|} \int_{\Omega \times D} g_k(x, y)\, dxdy \right|.
\end{aligned}$$

We conclude first by letting $n \to +\infty$ and using (11.25) satisfied by g_k and then by letting $k \to +\infty$ and using (11.27). □

Now, we would like to compute the lower semicontinuous envelope of the integral functional in (11.7), by taking into account a boundary condition on a part Γ_0 of the boundary $\partial\Omega$ of Ω. More precisely, we aim to describe the lower semicontinuous envelope of the integral functional $F : L^p(\Omega, \mathbf{R}^m) \longrightarrow \mathbf{R}^+ \cup \{+\infty\}$ defined by

$$F(u) = \begin{cases} \displaystyle\int_\Omega f(\nabla u)\, dx \text{ if } u \in W^{1,p}_{\Gamma_0}(\Omega, \mathbf{R}^m), \\ +\infty \text{ otherwise,} \end{cases} \tag{11.28}$$

where $W^{1,p}_{\Gamma_0}(\Omega, \mathbf{R}^m)$ denotes the subspace of all the functions u in $W^{1,p}(\Omega, \mathbf{R}^m)$ such that $u = 0$ on Γ_0 in the sense of traces. In the following corollary, we state that the boundary condition is "not relaxed." We will see that the preservation of the boundary condition is not satisfied in the case $p = 1$.

Corollary 11.2.1. *The lower semicontinuous envelope of the integral functional F defined in* (11.28) *is given by*

$$cl(F)(u) = \begin{cases} \displaystyle\int_\Omega Qf(\nabla u)\, dx \text{ if } u \in W^{1,p}_{\Gamma_0}(\Omega, \mathbf{R}^m), \\ +\infty \text{ otherwise.} \end{cases}$$

Proof. We only have to establish the existence of a sequence $(u_n)_{n\in\mathbf{N}}$ strongly converging to u in $L^p(\Omega, \mathbf{R}^m)$ such that $\mathrm{cl}(F)(u) \geq \limsup_{n\to+\infty} F(u_n)$. Assuming $\mathrm{cl}(F)(u) < +\infty$, we must construct a sequence $(u_n)_{n\in\mathbf{N}}$ in $W^{1,p}_{\Gamma_0}(\Omega, \mathbf{R}^m)$, strongly converging to u in $L^p(\Omega, \mathbf{R}^m)$ and satisfying $\mathrm{cl}(F)(u) \geq \limsup_{n\to+\infty} F(u_n)$.

According to Theorem 11.2.1, there exists a sequence $(v_n)_{n\in\mathbf{N}}$ in $W^{1,p}(\Omega, \mathbf{R}^m)$ strongly converging to u in $L^p(\Omega, \mathbf{R}^m)$ such that

$$\int_\Omega Qf(\nabla u)\, dx \geq \limsup_{n\to+\infty} \int_\Omega f(\nabla v_n)\, dx. \tag{11.29}$$

The idea is now to modify v_n on a neighborhood of $\partial\Omega$ so that the new function belongs to $W^{1,p}_{\Gamma_0}(\Omega, \mathbf{R}^m)$ and in such a way to decrease the energy. We use again the slicing method of De Giorgi. Let $\nu \in \mathbf{N}^*$ and $\Omega_0 \subset\subset \Omega$ such that

$$\int_{\Omega\setminus\Omega_0} \left(1 + |\nabla u|^p\right) dx \leq \frac{1}{\nu} \tag{11.30}$$

and $(\Omega_i)_{i=0,\ldots,\nu}$ an increasing sequence of open subsets strictly included in Ω, $\Omega_i \subset\subset \Omega_{i+1} \subset\subset \Omega$. Let $(\varphi_i)_{i=0,\ldots,\nu-1}$ be a sequence of functions in $\mathcal{D}(\mathbf{R}^N)$ satisfying

$$\begin{aligned} &\varphi_i = 1 \text{ on } \Omega_i,\ \varphi_i = 0 \text{ on } \mathbf{R}^N \setminus \Omega_{i+1},\ 0 \leq \varphi_i \leq 1, \\ &|\nabla\varphi_i| \leq \frac{\nu}{d}, \end{aligned}$$

where $d = \mathrm{dist}(\Omega_0, \mathbf{R}^N \setminus \overline{\Omega})$, and define

$$u_{n,i} = \varphi_i(v_n - u) + u.$$

Clearly $u_{n,i}$ belongs to $W^{1,p}_{\Gamma_0}(\Omega, \mathbf{R}^m)$ and

$$\int_\Omega f(\nabla u_{n,i})\,dx = \int_{\Omega\setminus\Omega_{i+1}} f(\nabla u_{n,i})\,dx + \int_{\Omega_{i+1}\setminus\Omega_i} f(\nabla u_{n,i})\,dx + \int_{\Omega_i} f(\nabla u_{n,i})\,dx$$
$$\leq \int_{\Omega\setminus\Omega_0} f(\nabla u)\,dx + \int_{\Omega_{i+1}\setminus\Omega_i} f(\nabla u_{n,i})\,dx + \int_\Omega f(\nabla v_n)\,dx.$$

Then, from (11.30) and the growth condition in (11.5), we obtain

$$\int_\Omega f(\nabla u_{n,i})\,dx \leq C\left(\frac{1}{\nu} + \left(\frac{\nu}{d}\right)^p \int_\Omega |v_n - u|^p\,dx + \int_{\Omega_{i+1}\setminus\Omega_i} (|\nabla v_n|^p)\,dx\right) + \int_\Omega f(\nabla v_n)\,dx,$$

where, from now on, C denotes various positive constants depending only on β, p, and Ω. By averaging these ν inequalities, we obtain

$$\frac{1}{\nu}\sum_{i=0}^{\nu-1} \int_\Omega f(\nabla u_{n,i})\,dx \leq C\left(\frac{1}{\nu} + \left(\frac{\nu}{d}\right)^p \int_\Omega |v_n - u|^p\,dx + \frac{1}{\nu}\int_\Omega |\nabla v_n|^p\,dx\right) + \int_\Omega f(\nabla v_n)\,dx.$$

As already said in the proof of Proposition 11.2.3, we used a slicing method to control the term $\int_\Omega |\nabla v_n|^p\,dx$ by taking increasingly thin slices (i.e, $\nu \to +\infty$). We could not conclude by using a simple troncature.

From the coercivity condition (11.5), $\int_\Omega |\nabla v_n|^p\,dx$ is bounded, hence

$$\frac{1}{\nu}\sum_{i=0}^{\nu-1} \int_\Omega f(\nabla u_{n,i})\,dx \leq C\left(\frac{1}{\nu} + \left(\frac{\nu}{d}\right)^p \int_\Omega |v_n - u|^p\,dx\right) + \int_\Omega f(\nabla v_n)\,dx. \quad (11.31)$$

Let $i(n, \nu)$ be the indice i such that

$$\int_\Omega f(\nabla u_{n,i(n,\nu)})\,dx = \min_{i=0,\ldots,\nu-1} \int_\Omega f(\nabla u_{n,i})\,dx.$$

Inequality (11.31) yields

$$\int_\Omega f(\nabla u_{n,i(n,\nu)})\,dx \leq C\left(\frac{1}{\nu} + \left(\frac{\nu}{d}\right)^p \int_\Omega |v_n - u|^p\,dx\right) + \int_\Omega f(\nabla v_n)\,dx$$

so that from (11.29) and since $v_n \to u$ in $L^p(\Omega, \mathbf{R}^m)$

$$\limsup_{\nu\to+\infty} \limsup_{n\to+\infty} F(u_{n,i(n,\nu)}) \leq \limsup_{n\to+\infty} F(v_n) \leq \int_\Omega Qf(\nabla u)\,dx.$$

We conclude by a classical diagonalization argument: there exists $n \mapsto \nu(n)$ mapping $\mathbf{N}$ into $\mathbf{N}$ such that

$$\lim_{n\to+\infty} F(u_{n,i(n,\nu(n))}) \leq \int_\Omega Qf(\nabla u)\,dx.$$

Obviously

$$\lim_{n\to+\infty} u_{n,i(n,\nu(n))} = u \text{ strongly in } L^p(\Omega, \mathbf{R}^m).$$

The sequence $(u_n)_{n\in\mathbf{N}}$ defined by $u_n = u_{n,i(n,\nu(n))}$ then tends to u in $L^p(\Omega,\mathbf{R}^m)$ and satisfies $\int_\Omega Qf(\nabla u)\,dx \geq \limsup_{n\to+\infty} F(u_n)$. □

As a consequence we obtain the following relaxation theorem in the case $p > 1$.

Theorem 11.2.2 (relaxation theorem, $p > 1$). *Let us consider a function $f : \mathbf{M}^{m\times N} \longrightarrow \mathbf{R}$ satisfying* (11.5), (11.6), *a function g in $L^q(\Omega,\mathbf{R}^m)$, where $\frac{1}{p}+\frac{1}{q}=1$, and consider the following problem:*

$$\inf\left\{\int_\Omega f(\nabla u)\,dx - \int_\Omega g.u\,dx \;:\; u\in W^{1,p}_{\Gamma_0}(\Omega,\mathbf{R}^m)\right\}. \qquad (\mathcal{P})$$

Then, the relaxed problem of $(\mathcal{P})$ in the sense of Theorem 11.1.2 *is*

$$\min\left\{\int_\Omega Qf(\nabla u)\,dx - \int_\Omega g.u\,dx \;:\; u\in W^{1,p}_{\Gamma_0}(\Omega,\mathbf{R}^m)\right\}. \qquad (\overline{\mathcal{P}})$$

Proof. With the notation of Theorem 11.2.1 one has

$$\inf\left\{\int_\Omega f(\nabla u)\,dx - \int_\Omega g.u\,dx \;:\; u\in W^{1,p}_{\Gamma_0}(\Omega,\mathbf{R}^m)\right\} = \inf_{u\in L^p(\Omega,\mathbf{R}^m)}\left\{F(u) - \int_\Omega g.u\,dx\right\}$$

and

$$\inf\left\{\int_\Omega Qf(\nabla u)\,dx - \int_\Omega g.u\,dx \;:\; u\in W^{1,p}_{\Gamma_0}(\Omega,\mathbf{R}^m)\right\} = \inf_{u\in L^p(\Omega,\mathbf{R}^m)}\left\{\mathrm{cl}(F)(u) - \int_\Omega g.u\,dx\right\}.$$

Since $G : u \mapsto \int_\Omega g.u\,dx$ is a continuous perturbation of F, one has $\mathrm{cl}(F+G) = \mathrm{cl}(F)+G$ in $L^p(\Omega,\mathbf{R}^m)$. Then, according to Theorems 11.1.2, 11.2.1, it suffices to establish the inf-compactness of

$$u \mapsto F(u) - \int_\Omega g.u\,dx$$

in $L^p(\Omega,\mathbf{R}^m)$ equipped with its strong topology. Let $u \in L^p(\Omega,\mathbf{R}^m)$ such that $F(u) - \int_\Omega g.u\,dx \leq C$, where C is any positive constant. Then $u \in W^{1,p}(\Omega,\mathbf{R}^m)$ and

$$F(u) - \int_\Omega g.u\,dx = \int_\Omega f(\nabla u)\,dx - \int_\Omega g.u\,dx \leq C.$$

From (11.5) and Hölder's inequality, we obtain

$$\alpha\int_\Omega |\nabla u|^p\,dx \leq \|g\|_{L^q(\Omega,\mathbf{R}^m)}\|u\|_{L^p(\Omega,\mathbf{R}^m)} + C. \qquad (11.32)$$

Applying Young's inequality $ab \leq \frac{\lambda^p a^p}{p} + \frac{1}{\lambda^q}\frac{b^q}{q}$ with $a = \|u\|_{L^p(\Omega,\mathbf{R}^m)}$, $b = \|g\|_{L^q(\Omega,\mathbf{R}^m)}$ where λ is chosen so that $\frac{\lambda^p}{p}C_P < \alpha$ and C_P denotes the Poincaré constant, i.e., the best constant satisfying

$$\int_\Omega |u|^p\,dx \leq C_P\int_\Omega |\nabla u|^p\,dx,$$

the estimate (11.32) yields

$$\int_\Omega |\nabla u|^p \, dx \leq C,$$

where C is a positive constant depending only on Ω, p, α, and $\|g\|_{\mathrm{L}^q(\Omega,\mathbf{R}^m)}$. Therefore u belongs to the closed ball with radius C of $W^{1,p}_{\Gamma_0}(\Omega, \mathbf{R}^m)$ which, according to the Rellich–Kondrakov theorem, Theorem 5.4.2, is compact in $L^p(\Omega, \mathbf{R}^m)$. □

Remark 11.2.2. In the case $p = 1$, Theorem 11.2.1 obviously remains valid as long as one considers the restrictions of F and $\mathrm{cl}(F)$ to $W^{1,1}(\Omega, \mathbf{R}^m)$. More precisely, for all $u \in W^{1,1}(\Omega, \mathbf{R}^m)$,

$$\mathrm{cl}(F)(u) = \int_\Omega Qf(\nabla u) \, dx.$$

Nevertheless, we do not have a complete description of $\mathrm{cl}(F)$, which is indeed given in the next section.

11.3 Relaxation of integral functionals with domain $W^{1,1}(\Omega, \mathbf{R}^m)$

We show how the space $BV(\Omega, \mathbf{R}^m)$ and the notion of trace (see Remark 10.2.2) take place in the relaxation theory. For simplicity of the exposition, in a first approach we limit our study to the case $m = 1$ and $f = |.|$. The general case is treated at the end of this section. From now on Ω is a Lipschitz open bounded subset of $\mathbf{R}^N$.

It is well known that the integral functionals defined on $L^p(\Omega)$, $p > 1$ by

$$F(u) = \begin{cases} \displaystyle\int_\Omega |\nabla u|^p \, dx \text{ if } u \in W^{1,p}(\Omega), \\ +\infty \text{ otherwise,} \end{cases} \qquad G(u) = \begin{cases} \displaystyle\int_\Omega |\nabla u|^p \, dx \text{ if } u \in W^{1,p}_0(\Omega), \\ +\infty \text{ otherwise,} \end{cases}$$

are lower semicontinuous for the strong topology of $L^P(\Omega)$ or the weak topology of $W^{1,p}(\Omega)$. For instance, one may argue directly by using the convexity of these two functionals or one may apply the previous section by noticing that $Qf = f$ when $f = |.|^p$.

The case $p = 1$ where $L^1(\Omega)$ is equipped with its strong topology and the functionals F, G are given by

$$F(u) = \begin{cases} \displaystyle\int_\Omega |\nabla u| \, dx \text{ if } u \in W^{1,1}(\Omega), \\ +\infty \text{ otherwise,} \end{cases} \qquad G(u) = \begin{cases} \displaystyle\int_\Omega |\nabla u| \, dx \text{ if } u \in W^{1,1}_0(\Omega), \\ +\infty \text{ otherwise,} \end{cases}$$

is more involved. Indeed, we have seen in Section 10.4 that the sequence $(u_n)_{n\in\mathbf{N}}$ which generates the Cantor–Vitali function u satisfies $u_n \to u$ strongly in $L^1(0, 1)$, $\sup_{n\in\mathbf{N}} F(u_n) < +\infty$ but $u \notin W^{1,1}(\Omega)$. Consequently, the domain of the lower closure of F strictly contains the space $W^{1,1}(\Omega)$. We see below that this domain is included in the space $BV(\Omega)$. Actually, as a consequence of Proposition 11.3.2, the domain is exactly $BV(\Omega)$. Concerning the second functional G, we will see that the boundary condition $u = 0$ is relaxed by a surface energy.

Proposition 11.3.1. *The domain of $cl(F)$ and $cl(G)$ is included in $BV(\Omega)$.*

PROOF. Let $u \in L^1(\Omega, \mathbf{R}^m)$ be such that $\mathrm{cl}(F)(u) < +\infty$ and consider a sequence $(u_n)_{n\in\mathbf{N}}$ strongly converging to u in $L^1(\Omega, \mathbf{R}^m)$ and satisfying $\mathrm{cl}(F)(u) = \lim_{n\to+\infty} F(u_n)$. Such a sequence exists from Proposition 11.1.1. According to $\mathrm{cl}(F)(u) = \lim_{n\to+\infty} F(u_n) < +\infty$, one obtains, for a not relabeled subsequence of $(u_n)_{n\in\mathbf{N}}$,

$$\sup_{n\in\mathbf{N}} \int_\Omega |\nabla u_n|\, dx < +\infty.$$

Thus, from Proposition 10.1.1(i), $u \in BV(\Omega)$. □

Proposition 11.3.2. *Let Ω be a Lipschitz bounded open subset of $\mathbf{R}^N$ and Γ its boundary. The lsc envelopes $cl(F)$ and $cl(G)$ of the functionals F and G defined on $L^1(\Omega)$ equipped with its strong topology are given by*

$$cl(F)(u) = \begin{cases} \displaystyle\int_\Omega |Du| \text{ if } u \in BV(\Omega), \\ +\infty \text{ otherwise}, \end{cases}$$

$$cl(G)(u) = \begin{cases} \displaystyle\int_\Omega |Du| + \int_\Gamma |\gamma_0(u)|\, d\mathcal{H}^{N-1} \text{ if } u \in BV(\Omega), \\ +\infty \text{ otherwise,} \end{cases}$$

where γ_0 is the trace operator from $BV(\Omega)$ into $L^1(\Gamma)$, defined in Section 10.2.

PROOF. Let us set

$$QF(u) = \begin{cases} \displaystyle\int_\Omega |Du| \text{ if } u \in BV(\Omega), \\ +\infty \text{ otherwise}, \end{cases}$$

$$QG(u) = \begin{cases} \displaystyle\int_\Omega |Du| + \int_\Gamma |\gamma_0(u)|\, d\mathcal{H}^{N-1} \text{ if } u \in BV(\Omega), \\ +\infty \text{ otherwise.} \end{cases}$$

We first establish that for all u in $L^1(\Omega)$:

$$\begin{cases} \text{if } u_n \to u \text{ in } L^1(\Omega), \text{ then } QF(u) \le \liminf\limits_{n\to+\infty} F(u_n), \\ \text{there exists a sequence } (u_n)_{n\in\mathbf{N}} \text{ converging to } u \text{ in } L^1(\Omega) \text{ such that } QF(u) \ge \limsup\limits_{n\to+\infty} F(u_n). \end{cases}$$

These two assertions are straightforward consequences of Proposition 10.1.1 and Theorem 10.1.2. We now deal with the functional G and establish for every $u \in L^1(\Omega)$,

$$\begin{cases} \text{if } u_n \to u \text{ in } L^1(\Omega), \text{ then } QG(u) \le \liminf\limits_{n\to+\infty} G(u_n), \\ \text{there exists a sequence } (u_n)_{n\in\mathbf{N}} \text{ converging to } u \text{ in } L^1(\Omega) \text{ such that } QG(u) \ge \limsup\limits_{n\to+\infty} G(u_n). \end{cases}$$

Proof of the first assertion. Consider a sequence $(u_n)_{n\in\mathbf{N}}$ strongly converging to u in $L^1(\Omega)$ such that $\liminf_{n\to+\infty} G(u_n) < +\infty$. For a subsequence (not relabeled), we have $u_n \in W_0^{1,1}(\Omega)$. Let $\tilde{\Omega}$ denote a bounded open subset of $\mathbf{R}^N$ strongly containing Ω, i.e., $\overline{\Omega} \subset \tilde{\Omega}$, and define, for every function v in $L^1(\Omega)$, the function $\tilde{v}$ in $L^1(\tilde{\Omega})$ by

$$\tilde{v}(x) = \begin{cases} v(x) \text{ if } x \in \Omega, \\ 0 \text{ if } x \in \tilde{\Omega} \setminus \overline{\Omega}. \end{cases}$$

It is easily seen that $\tilde{u}_n \to \tilde{u}$ in $L^1(\tilde{\Omega})$, $\tilde{u}_n \in W^{1,1}(\tilde{\Omega})$, and $\tilde{u} \in BV(\tilde{\Omega})$ with $D\tilde{u} = Du\lfloor\Omega + \gamma_0(u)\ \nu\ \mathcal{H}^{N-1}\lfloor\Gamma$, where ν denotes the inner unit normal at $\mathcal{H}^{N-1}$ a.e. x in Γ (see Examples 10.2.1 and 10.2.2).

Since $Du\lfloor\Omega$ and $\gamma_0(u)\ \nu\ \mathcal{H}^{N-1}\lfloor\Gamma$ are mutually singular, one has $|D\tilde{u}| = |Du|\lfloor\Omega + |\gamma_0(u)|\mathcal{H}^{N-1}\lfloor\Gamma$. Thus, according to the previous result for F, one has

$$\int_\Omega |Du| + \int_\Gamma |\gamma_0(u)|\, d\mathcal{H}^{N-1} = \int_{\tilde{\Omega}} |D\tilde{u}| \le \liminf_{n\to+\infty} \int_{\tilde{\Omega}} |D\tilde{u}_n| = \liminf_{n\to+\infty} \int_\Omega |Du_n|.$$

Proof of the second assertion. For $t > 0$ let us consider the open subset $\Omega_t = \{x \in \Omega : \operatorname{dist}(x, \mathbf{R}^N \setminus \Omega) > t\}$ of Ω, and define for every u in $BV(\Omega)$, the function u_t by

$$u_t(x) = \begin{cases} u(x) \text{ if } x \in \Omega_t, \\ 0 \text{ if } x \in \Omega \setminus \overline{\Omega}_t. \end{cases}$$

It is easy to see that $u_t \in BV(\Omega)$ and that

$$Du_t\lfloor\Omega = Du\lfloor\Omega_t + \gamma_t(u)\ \nu_t\ \mathcal{H}^{N-1}\lfloor_{\Gamma_t},$$

where Γ_t denotes the boundary of Ω_t, ν_t the inner unit normal at $\mathcal{H}^{N-1}$ a.e. x in Γ_t, and γ_t the trace operator from $\mathrm{BV}(\Omega_t)$ into $L^1(\Gamma_t)$ (see Examples 10.2.1 and 10.2.2). According to Theorem 10.1.2 and Remark 10.2.1, there exists $u_{t,n} \in \mathbf{C}^\infty(\Omega) \cap BV(\Omega)$ satisfying $u_{t,n} = 0$ on Γ in the trace sense and converging to u_t for the intermediate convergence in $BV(\Omega)$. We obtain, for a.e. t,

$$\begin{aligned} \lim_{n\to+\infty} \int_\Omega |Du_{t,n}| &= \int_\Omega |Du_t| \\ &= \int_{\Omega_t} |Du| + \int_{\Gamma_t} |u|\, d\mathcal{H}^{N-1}, \end{aligned}$$

where we have used that for a.e. t, $\gamma_t(u) = u$ on Γ_t (cf. Example 10.2.4). Letting $t \to 0^+$, we claim that

$$\lim_{t\to0^+} \lim_{n\to+\infty} \int_\Omega |Du_{t,n}| = \int_\Omega |Du| + \int_\Gamma |\gamma_0(u)|\, d\mathcal{H}^{N-1}. \tag{11.33}$$

To justify the nontrivial limit

$$\lim_{t\to0^+} \int_{\Gamma_t} |u|\, d\mathcal{H}^{N-1} = \int_\Gamma |\gamma_0(u)|\, d\mathcal{H}^{N-1},$$

we argue with local coordinates and make use of estimate (10.13) in the proof of Theorem 10.2.1. With the notations of this proof, we have

$$\int_\Gamma |u(\tilde{x}, t) - u(\tilde{x}, t')| \le \int_{C_{R,t',t}} |Du|,$$

and letting $t' \to 0$,

$$\int_\Gamma |\gamma_0(u) - u(\tilde{x}, t)| \le \int_{C_{R,t}} |Du|.$$

The result follows after letting $t \to 0$.

Now t denotes a sequence converging to 0. Going back to (11.33) and using the diagonalization Lemma 11.1.1 for the sequence $\left(\int_\Omega |Du_{t,n}|, u_{t,n}\right)$ in the metrizable space $\mathbf{R} \times L^1(\Omega)$, we conclude that there exists a sequence $n \mapsto t(n)$ such that

$$\begin{cases} \lim\limits_{n\to+\infty} \int_\Omega |Du_{t(n),n}| = \int_\Omega |Du| + \int_\Gamma |\gamma_0(u)|\, d\mathcal{H}^{N-1}, \\ \lim\limits_{n\to+\infty} u_{t(n),n} = u \text{ in } L^1(\Omega), \end{cases}$$

and the proof is complete. □

Let us now consider the general case and state the following theorem, analogous to Theorem 11.2.1. Given a function $f : \mathbf{M}^{m\times N} \to \mathbf{R}$ satisfying (11.5) and (11.6) with $p = 1$, we intend to compute the lsc envelope of the functional $F : L^1(\Omega, \mathbf{R}^m) \longrightarrow \mathbf{R}^+ \cup \{+\infty\}$ defined by

$$F(u) = \begin{cases} \int_\Omega f(\nabla u)\, dx \text{ if } u \in W^{1,1}(\Omega, \mathbf{R}^m), \\ +\infty \text{ otherwise.} \end{cases} \tag{11.34}$$

Let us recall that when $u \in BV(\Omega)$, according to the Radon–Nikodym theorem, Theorem 4.2.1, one has $Du = \nabla u \mathcal{L}^N \lfloor \Omega + D^s u$, where $\nabla u \mathcal{L}^N \lfloor \Omega$ and $D^s u$ are two mutually singular measures in $\mathbf{M}(\Omega, \mathbf{R}^N)$.

Theorem 11.3.1. *The lsc envelope of the functional F defined in* (11.34) *is given, for every $u \in L^1(\Omega, \mathbf{R}^m)$, by*

$$cl(F)(u) = \begin{cases} \int_\Omega Qf(\nabla u)\, dx + \int_\Omega (Qf)^\infty \left(\dfrac{D^s u}{|D^s u|}\right) |D^s u| \text{ if } u \in BV(\Omega, \mathbf{R}^m), \\ +\infty \text{ otherwise,} \end{cases}$$

where Qf is the quasiconvex envelope of f defined in Proposition 11.2.2, *and $(Qf)^\infty$ is the recession function of Qf defined for every a in $\mathbf{M}^{m\times N}$ by $(Qf)^\infty(a) = \limsup_{t\to+\infty} (Qf(ta)/t)$.*

The proof of Theorem 11.3.1 is a consequence of Propositions 11.3.3 and 11.3.4. We set

$$QF(u) = \begin{cases} \int_\Omega Qf(\nabla u)\, dx + \int_\Omega (Qf)^\infty \left(\dfrac{D^s u}{|D^s u|}\right) |D^s u| \text{ if } u \in \mathrm{BV}(\Omega, \mathbf{R}^m), \\ +\infty \text{ otherwise.} \end{cases}$$

Proposition 11.3.3. *For every u in $L^1(\Omega, \mathbf{R}^m)$ and every sequence $(u_n)_{n\in\mathbf{N}}$ strongly converging to u in $L^1(\Omega, \mathbf{R}^m)$, one has*

$$QF(u) \le \liminf_{n\to+\infty} F(u_n). \tag{11.35}$$

PROOF. Our strategy is exactly the one of Proposition 11.2.3. Obviously, one may assume $\liminf_{n\to+\infty} F(u_n) < +\infty$. For a nonrelabeled subsequence, let us consider the nonnegative Borel measure $\mu_n := f(\nabla u_n(.))\mathcal{L}^N \lfloor \Omega$. Since

$$\sup_{n\in\mathbf{N}} \mu_n(\Omega) < +\infty,$$

there exists a further subsequence (not relabeled) and a nonnegative Borel measure $\mu \in \mathbf{M}(\Omega)$ such that

$$\mu_n \rightharpoonup \mu \text{ weakly in } \mathbf{M}(\Omega).$$

Let $\mu = g\,\mathcal{L}^N \lfloor \Omega + \mu^s$ be the Lebesgue–Nikodym decomposition of μ where μ^s is a nonnegative Borel measure, singular with respect to the N-dimensional Lebesgue measure $\mathcal{L}^N \lfloor \Omega$. For establishing (11.35) it suffices to prove that

$$\begin{aligned} &g(x) \geq Qf(\nabla u(x)) \qquad \text{for a.e } x \in \Omega, \\ &\mu^s \geq (Qf)^\infty\left(\frac{D^s u}{|D^s u|}\right) |D^s u|. \end{aligned}$$

Indeed, according to Alexandrov's theorem, Proposition 4.2.3, we will obtain

$$\begin{aligned} \liminf_{n\to+\infty} F(u_n) = \liminf_{n\to+\infty} \mu_n(\Omega) \geq \mu(\Omega) &= \int_\Omega g(x)\,dx + \mu_s(\Omega) \\ &\geq \int_\Omega Qf(\nabla u(x))\,dx + \int_\Omega (Qf)^\infty\left(\frac{D^s u}{|D^s u|}\right) |D^s u|. \end{aligned}$$

(a) *Proof of $g(x) \geq Qf(\nabla u(x))$ for a.e. $x \in \Omega$.* It suffices to reproduce the proof of Proposition 11.2.3 which obviously holds true for $p = 1$ (indeed, Lemma 11.2.1 holds true from Proposition 10.4.1).

(b) *Proof of $\mu^s \geq (Qf)^\infty(\frac{D^s u}{|D^s u|})\,|D^s u|$.* The proof is based on various arguments of Ambrosio–Dal Maso [21]. The density $\frac{Du}{|Du|}$ satisfies the following property (see Alberti [8]).

Lemma 11.3.1. *The density $Du/|Du|$ is for $|D^s u|$-a.e. $x_0 \in \Omega$ a rank-one matrix, i.e., for $|D^s u|$-a.e. $x_0 \in \Omega$*

$$\frac{Du}{|Du|}(x_0) = a(x_0) \otimes b(x_0)$$

with $a(x_0) \in \mathbf{R}^m,\ b(x_0) \in \mathbf{R}^N,\ |a(x_0)| = |b(x_0)| = 1$.

The rank-one property of the jump part of the singular measure $D^s u$ is indeed trivial because of its structure (see Section 10.3). De Giorgi conjectured that the diffuse singular part is also a rank-one matrix valued measure. The proof was later given by Alberti in [8]. Let us give now some notation. Let x_0 be an element of the complementary of the $|D^s u|$-nullset invoked in Lemma 11.3.1 and let Q denote the unit cube of $\mathbf{R}^N$ centered at the origin, whose sides are either orthogonal or parallel to $b(x_0)$. We set $Q_\rho(x_0) := \{x_0 + \rho x\ :\ x \in Q\}$, where ρ is a positive parameter intended to tend to zero.

According to the theory of differentiation of measures (see Theorem 4.2.1), for establishing

$$\mu^s \geq (Qf)^\infty(D^s u),$$

it is enough to prove

$$\lim_{\rho\to 0} \frac{\mu(Q_\rho(x_0))}{|Du|(Q_\rho(x_0))} \geq (Qf)^\infty(a(x_0)\otimes b(x_0)). \tag{11.36}$$

We will make use of the following two estimates: for $|D^s u|$-a.e. $x_0 \in \Omega$, we have

$$\lim_{\rho\to 0} \frac{|Du|(Q_\rho(x_0))}{\rho^N} = +\infty, \tag{11.37}$$

$$\limsup_{\rho\to 0} \frac{|Du|(Q_{t\rho}(x_0))}{|Du|(Q_\rho(x_0))} \geq t^N \ \forall t \in]0,1[^d. \tag{11.38}$$

Assertion (11.37) easily follows from the theory of differentiation of measures (see Theorem 4.2.1). For a proof of (11.38), consult Ambrosio–Dal Maso [21, Theorem 2.3].

From now on, x_0 will be a fixed element of the complementary of the $|D^s u|$-null set invoked in Lemma 11.3.1, for which moreover estimates (11.37), (11.38) hold true and the two limits

$$\lim_{\rho\to 0} \frac{\mu(Q_\rho(x_0))}{|Du|(Q_\rho(x_0))}, \quad \lim_{\rho\to 0} \frac{Du(Q_\rho(x_0))}{|Du|(Q_\rho(x_0))}$$

exist (cf. Theorem 4.2.1) and make use of a blow up procedure.

We set $t_\rho := |Du|(Q_\rho(x_0))/\rho^N$. According to (11.37), t_ρ will play the role of the parameter t in the definition of $(Qf)^\infty$:

$$(Qf)^\infty(a) = \limsup_{t\to+\infty} \frac{Qf(ta)}{t}.$$

Let us now define the following rescaled functions of $BV(Q,\mathbf{R}^m)$:

$$v_\rho(x) := \frac{\rho^{N-1}}{|Du|(Q_\rho(x_0))}\left(u(x_0+\rho x) - \frac{1}{|Q_\rho(x_0)|}\int_{Q_\rho(x_0)} u(y)dy \right),$$

$$v_{\rho,n}(x) := \frac{\rho^{N-1}}{|Du|(Q_\rho(x_0))}\left(u_n(x_0+\rho x) - \frac{1}{|Q_\rho(x_0)|}\int_{Q_\rho(x_0)} u_n(y)dy \right),$$

which satisfy

$$\int_Q v_\rho(y)dy = 0, \ \lim_{n\to+\infty} |v_{\rho,n} - v_\rho|_{L^1(Q,\mathbf{R}^m)} = 0$$

and, from Lemma 11.3.1,

$$Dv_\rho(Q) = \frac{Du(Q_\rho(x_0))}{|Du|(Q_\rho(x_0)} \to a(x_0)\otimes b(x_0) \text{ when } \rho\to 0.$$

The sequence $(v_\rho)_{\rho>0}$ fulfills the following properties (for a proof, consult Ambrosio and Dal Maso [21, Theorem 2.3]).

Lemma 11.3.2. *There exists a subsequence of $(v_\rho)_{\rho>0}$, not relabeled, which weakly converges in $BV(Q, \mathbf{R}^m)$ to a function v of the form $v(x) = \overline{v}(\langle b(x_0), x\rangle)a(x_0)$, where $\overline{v}$ is nondecreasing and belongs to $BV(]-1/2, 1/2[)$. Moreover, for a.e. δ in $(0,1)$ one has*

$$Dv_\rho(\delta Q) \to Dv(\delta Q).$$

We are now in a position to establish $\mu^s \geq (Qf)^\infty(\frac{D^s u}{|D^s u|})\,|D^s u|$. The proof will be divided into three steps. We fix δ in $(0,1)$ outside of a set of null measure so that the last assertion of Lemma 11.3.2 holds. We will make δ tend to 1 at the end of the proof.

First step (truncation). Let $u_{\rho,n}$ in $W^{1,1}(Q, \mathbf{R}^m)$ defined by

$$u_{\rho,n} := \frac{1}{\rho}\left(u_n(x_0 + \rho y) - \frac{1}{|Q_\rho(x_0)|}\int_{Q_\rho(x_0)} u_n(y)dy\right).$$

Note that $\frac{1}{t_\rho}u_{\rho,n}$ is the function $v_{\rho,n}$ previously defined. A change of scale gives

$$\frac{1}{|Du|(Q_\rho(x_0))}\int_{Q_{\delta\rho}(x_0)} f(\nabla u_n)dx = \frac{1}{t_\rho}\int_{\delta Q} f(\nabla u_{\rho,n})dx.$$

We want to modify the function $u_{\rho,n}$ in a neighborhood of δQ so that it coincides with an affine function of gradient $t_\rho Dv(\delta Q)$ on $\partial\delta Q$. The basic idea, similar to the one used in the proof of Proposition 11.2.3, consists in slicing a neighborhood of the boundary of δQ by thin slices whose size is of order $\|v_\rho - v\|^{1/2}_{L^1(Q,\mathbf{R}^m)}$. (Recall that $|v_\rho - v|_{L^1(Q,\mathbf{R}^m)}$ goes to zero when $\rho \to 0$.)

More precisely, we set $\alpha_\rho := \|v_\rho - v\|^{\frac{1}{2}}_{L^1(Q,\mathbf{R}^m)}$ and, for $\nu \in \mathbf{N}^*$ intended to tend to $+\infty$,

$$Q_0 := (1-\alpha_\rho)\delta Q,\ Q_i := \left(1 - \alpha_\rho + i\frac{\alpha_\rho}{\nu}\right)\delta Q \text{ for } i = 1, \ldots, \nu.$$

On the other hand, for $i = 1, \ldots, \nu$, we consider

$$\varphi_i \in \mathbf{C}_0^\infty(Q_i),\ 0 \leq \varphi_i \leq 1,\ \varphi_i = 1 \text{ in } Q_{i-1},\ |\nabla\varphi_i| \leq \frac{\nu}{\alpha_\rho}$$

and define $u_{\rho,n,i} \in t_\rho\Theta_\delta + W_0^{1,1}(\delta Q, \mathbf{R}^m)$ by

$$u_{\rho,n,i} := t_\rho\Theta_\delta + \varphi_i(u_{\rho,n} - t_\rho\Theta_\delta),$$

where Θ_δ is the affine function:

$$\Theta_\delta(x) := \frac{Dv(\delta Q)}{\delta^N}.x + \frac{\overline{v}((\frac{\delta}{2})^-) + \overline{v}((-\frac{\delta}{2})^+)}{2}a(x_0).$$

The constant $\frac{\overline{v}((\delta/2)^-)+\overline{v}(-(\delta/2)^+)}{2}a(x_0)$ has been chosen so that the traces of v and Θ agree on the faces of δQ orthogonal to $b(x_0)$ and, consequently, so that $v - \theta$ fulfills the Poincaré inequality on δQ. From (11.5), an easy computation gives

$$\frac{1}{t_\rho}\int_{\delta Q} f(\nabla u_{\rho,n})dx \geq \frac{1}{t_\rho}\int_{\delta Q} f(\nabla u_{\rho,n,i})dx - R_{\rho,n,\nu,\delta,i}, \tag{11.39}$$

where

$$R_{\rho,n,\nu,\delta,i} := O_\rho + \frac{\beta}{t_\rho}\int_{Q_i\setminus Q_{i-1}} |D(u_{\rho,n} - t_\rho\Theta_\delta)|dx + \frac{\beta\nu}{\alpha_\rho}\int_{Q_i\setminus Q_{i-1}} |v_{\rho,n} - \Theta_\delta|dx$$

and O_ρ does not depend on n and tends to 0 when $\rho \to +\infty$.

Second step (averaging). According to (11.39) and to the definition of Qf, we have

$$\begin{aligned}\frac{1}{|Du|(Q_\rho(x_0))}\int_{Q_{\delta\rho}(x_0)} f(\nabla u_n)dx \geq{}& \frac{\delta^N}{t_\rho}Qf\left(\frac{t_\rho}{\delta^N}Dv(\delta Q)\right)\\ &- O_\rho - \frac{\beta}{t_\rho}\int_{Q_i\setminus Q_{i-1}} |D(u_{\rho,n} - t_\rho\Theta_\delta)|dx\\ &- \frac{\beta\nu}{\alpha_\rho}\int_{Q_i\setminus Q_{i-1}} |v_{\rho,n} - \Theta_\delta|dx.\end{aligned}$$

Averaging these ν inequalities, we obtain

$$\begin{aligned}&\frac{1}{|Du|(Q_\rho(x_0))}\int_{Q_{\delta\rho}(x_0)} f(\nabla u_n)dx \geq \frac{\delta^N}{t_\rho}Qf\left(\frac{t_\rho}{\delta^N}Dv(\delta Q)\right)\\ &- O_\rho - \frac{\beta}{\nu t_\rho}\int_Q |D(u_{\rho,n} - t_\rho\Theta_\delta)|dx - \frac{\beta}{\alpha_\rho}\int_{\delta Q\setminus(1-\alpha_\rho)\delta Q} |v_{\rho,n} - \Theta_\delta|dx\\ &\geq \frac{\delta^N}{t_\rho}Qf\left(\frac{t_\rho}{\delta^N}Dv(\delta Q)\right) - O_\rho - \frac{\beta}{\nu t_\rho}\int_Q |D(u_{\rho,n} - t_\rho\Theta_\delta)|dx\\ &- \beta\left(\int_Q |v_{\rho,n} - v|dx\right)^{\frac{1}{2}} - \frac{\beta}{\alpha_\rho}\int_{\delta Q\setminus(1-\alpha_\rho)\delta Q} |v - \Theta_\delta|dx.\end{aligned}$$

Letting successively $\nu \to +\infty$ and $n \to +\infty$, we obtain

$$\begin{aligned}&\limsup_{n\to+\infty}\frac{1}{|Du|(Q_\rho(x_0))}\int_{Q_{\delta\rho}(x_0)} f(\nabla u_n)dx\\ &\geq \frac{\delta^N}{t\rho}Qf\left(\frac{t_\rho}{\delta^N}Dv(\delta Q)\right) - O_\rho - \beta\left(\int_Q |v_m - v|dx\right)^{\frac{1}{2}}\\ &- \frac{\beta}{\alpha_\rho}\int_{\delta Q\setminus(1-\alpha_\rho)\delta Q} |v - \Theta_\delta|dx.\end{aligned} \tag{11.40}$$

Last step. From Lipschitz continuity of Qf (cf Proposition 11.2.2) and according to Lemmas 11.3.1 and 11.3.2 and estimates (11.37) and (11.38), we obtain

$$\begin{aligned}\limsup_{\rho\to0}\frac{\delta^N}{t_\rho}Qf\left(\frac{t_\rho}{\delta^N}Dv(\delta Q)\right) &= \limsup_{\rho\to0}\frac{\delta^N}{t_\rho}Qf\left(\frac{t_\rho}{\delta^N}Dv(\delta Q)\right)\\ &\geq \limsup_{\rho\to0}\frac{\delta^N}{t_\rho}Qf\left(\frac{t_\rho}{\delta^N}a(x_0)\otimes b(x_0)\right)\\ &\quad - \liminf_{\rho\to0} L'\left(1 - \frac{|Du|(C_{\delta\rho}(x_0))}{|Du|(C_\rho(x_0))}\right)\\ &\geq (Qf)^\infty(a(x_0)\otimes b(x_0)) - L'(1-\delta^N).\end{aligned} \tag{11.41}$$

On the other hand,

$$\lim_{\rho \to 0} \frac{1}{\alpha_\rho} \int_{\delta Q \setminus (1-\alpha_\rho)\delta Q} |v - \Theta_\delta| dx = 0. \tag{11.42}$$

Indeed, from Poincaré's inequality

$$\frac{1}{\alpha_\rho} \int_{\delta Q \setminus (1-\alpha_\rho)\delta Q} |v - \Theta_\delta| dx \leq C \int_{\delta Q \setminus (1-\alpha_\rho)\delta Q} \left| Dv - \frac{Dv(\delta Q)}{\delta^N} \right|$$

which tends to zero when ρ goes to zero. Combining (11.40), (11.41), and (11.42), we obtain

$$\limsup_{\rho \to 0} \limsup_{n \to +\infty} \frac{1}{|Du|(Q_\rho(x_0))} \int_{\delta Q_\rho(x_0)} f(\nabla u_n) dx \geq (Qf)^\infty(a(x_0) \otimes b(x_0)) - L'(1 - \delta^N).$$

The lower bound is then established after letting $\delta \to 1$. □

Proposition 11.3.4. *For every u in $L^p(\Omega, \mathbf{R}^m)$, there exists a sequence $(u_n)_{n \in \mathbf{N}}$ strongly converging to u in $L^1(\Omega, \mathbf{R}^m)$ such that*

$$QF(u) \geq \limsup_{n \to +\infty} F(u_n).$$

Proof. The proof proceeds in two steps.

First step (computation of $cl(F)\lfloor W^{1,1}(\Omega, \mathbf{R}^m)$). According to Remark 11.2.2 and Theorem 11.2.1 for every $u \in W^{1,1}(\Omega, \mathbf{R}^m)$,

$$\mathrm{cl}(F)(u) = \int_\Omega Qf(\nabla u)\, dx. \tag{11.43}$$

Second step. Let us consider the functional $\tilde{F} : L^1(\Omega, \mathbf{R}^m) \longrightarrow \mathbf{R}^+ \cup \{+\infty\}$ defined by

$$\tilde{F}(u) = \begin{cases} \displaystyle\int_\Omega Qf(\nabla u)\, dx \text{ if } u \in W^{1,1}(\Omega, \mathbf{R}^m), \\ +\infty \text{ otherwise.} \end{cases}$$

We claim that it is enough to establish that for all $u \in \mathrm{BV}(\Omega, \mathbf{R}^m)$,

$$\mathrm{cl}(\tilde{F})(u) \leq \int_\Omega Qf(\nabla u)\, dx + \int_\Omega (Qf)^\infty\left(\frac{D^s u}{|D^s u|}\right) |D^s u|. \tag{11.44}$$

Indeed, let us assume (11.44). For every $u \in \mathrm{BV}(\Omega, \mathbf{R}^m)$, according to Proposition 11.1.1 and Theorem 11.1.1, we obtain the existence of a sequence $(u_k)_{k \in \mathbf{N}}$ in $W^{1,1}(\Omega, \mathbf{R}^m)$ strongly converging to u in $L^1(\Omega, \mathbf{R}^m)$ and satisfying

$$\begin{aligned} \int_\Omega Qf(\nabla u)\, dx + \int_\Omega (Qf)^\infty\left(\frac{D^s u}{|D^s u|}\right) |D^s u| &\geq \mathrm{cl}(\tilde{F})(u) \\ &= \lim_{k \to +\infty} \tilde{F}(u_k) \\ &= \lim_{k \to +\infty} \int_\Omega Qf(\nabla u_k)\, dx. \end{aligned}$$

On the other hand, from (11.43), there exists a sequence $(u_{k,n})_{n\in\mathbf{N}}$ in $W^{1,1}(\Omega,\mathbf{R}^m)$ strongly converging to u_k in $L^1(\Omega,\mathbf{R}^m)$ and satisfying

$$\int_\Omega Qf(\nabla u_k)\,dx = \lim_{n\to+\infty}\int_\Omega f(\nabla u_{k,n})\,dx.$$

Consequently,

$$\begin{cases} \displaystyle\int_\Omega Qf(\nabla u)\,dx + \int_\Omega (Qf)^\infty\left(\frac{D^s u}{|D^s u|}\right)|D^s u| \geq \lim_{k\to+\infty}\lim_{n\to+\infty}\int_\Omega f(\nabla u_{k,n})\,dx, \\ \displaystyle\lim_{k\to+\infty}\lim_{n\to+\infty} u_{k,n} = u \text{ strongly in } L^1(\Omega,\mathbf{R}^m), \end{cases}$$

and we conclude by a diagonalization argument: there exists $n\mapsto k(n)$ mapping $\mathbf{N}$ to $\mathbf{N}$ such that

$$\begin{cases} \displaystyle\lim_{n\to+\infty}\int_\Omega f(\nabla u_{k(n),n})\,dx \leq \int_\Omega Qf(\nabla u)\,dx + \int_\Omega (Qf)^\infty\left(\frac{D^s u}{|D^s u|}\right)|D^s u|, \\ \displaystyle\lim_{n\to+\infty} u_{k(n),n} = u \text{ strongly in } L^1(\Omega,\mathbf{R}^m). \end{cases}$$

We now establish (11.44). We use some arguments of Ambrosio and Dal Maso [21]. Let u be a fixed element of $\mathrm{BV}(\Omega,\mathbf{R}^m)$ and consider the set function

$$\mu : A \mapsto \mathrm{cl}(\tilde{F})(u,A)$$

defined for all bounded open subset A of Ω. The notation $\mathrm{cl}(\tilde{F})(.,A)$ means that we consider the lower semicontinuous envelope of the functional $\tilde{F}$ localized on A, the space $L^1(A,\mathbf{R}^m)$ being equipped with its strong topology. Note that μ satisfies the following estimate for every open bounded subset A of Ω:

$$\mu(A) \leq \beta|A| + \beta\int_A |Du|. \tag{11.45}$$

Indeed, applying the approximation theorem, Theorem 10.1.2, there exists a sequence $(v_n)_{n\in\mathbf{N}}$ in $\mathrm{BV}(A,\mathbf{R}^m)\cap\mathbf{C}^\infty(A,\mathbf{R}^m)$ such that

$$v_n \to u \text{ strongly in } L^1(A,\mathbf{R}^m),$$

$$\int_A |\nabla v_n|\,dx \to \int_A |Du|.$$

Thus, from the growth condition (11.5)

$$\begin{aligned} \mu(A) = \mathrm{cl}(\tilde{F})(u,A) &\leq \liminf_{n\to+\infty} F(v_n,A) \\ &\leq \beta|A| + \liminf_{n\to+\infty}\int_A |\nabla v_n|\,dx \\ &= \beta|A| + \int_A |Du|. \end{aligned}$$

According to the definition of μ and to (11.45), one can now easily establish that for all bounded open subsets A and A' of Ω, μ satisfies

$$A \subset A' \Longrightarrow \mu(A) \leq \mu(A');$$

$$A \cap A' = \emptyset \Longrightarrow \mu(A \cup A') \geq \mu(A) + \mu(A');$$

$$\mu(A) \leq \sup\{\mu(A') \ : \ A' \subset\subset A\};$$

$$\mu(A \cup A') \leq \mu(A) + \mu(A').$$

For a complete proof, consult [21]. Consequently (consult, for instance, the book [117]), one can extend μ to a Borel measure on Ω, still denoted by μ, defined for every Borel set B of Ω by

$$\mu(B) = \inf\{\mu(A) \ : \ A \text{ open}, \ B \subset A\}.$$

By considering the Lebesgue–Nikodym decomposition (cf. Theorem 4.2.1) $\mu = \mu^a + \mu^s$ of μ, it suffices now to establish, for every Borel set B of Ω,

$$\mu^a(B) \leq \int_B Qf(\nabla u)\, dx, \tag{11.46}$$

$$\mu^s(B) \leq \int_B (Qf)^\infty \left(\frac{D^s u}{|D^s u|}\right) |D^s u|. \tag{11.47}$$

To estimate the singular part μ^s from above, i.e., (11.47), we will need the following continuity result.

Lemma 11.3.3. *Let $h : \mathbf{M}^{m\times N} \to \mathbf{R}$ be a continuous function and $(u_n)_{n\in\mathbf{N}}$ a sequence in $BV(\Omega, \mathbf{R}^m)$ converging to u for the intermediate convergence. Then*

$$\lim_{n\to+\infty} \int_\Omega h\left(\frac{Du_n}{|Du_n|}\right) |Du_n| = \int_\Omega h\left(\frac{Du}{|Du|}\right) |Du|.$$

For a proof, we refer the reader to Luckhaus and Modica [174] or, in the convex case, to Demengel and Temam [126]. Let us set, for every $a \in \mathbf{M}^{m\times N}$,

$$g(a) = \sup_{t>0} \frac{Qf(ta) - Qf(0)}{t}.$$

According to the rank-one convexity of Qf (cf. Remark 11.2.1), we note that if $\operatorname{rank}(a) = 1$, one has $g(a) = (Qf)^\infty(a)$. For every open subset A of Ω, consider $u_n \in \mathrm{BV}(A, \mathbf{R}^m) \cap \mathbf{C}^\infty(A, \mathbf{R}^m)$, strongly converging to u in $L^1(A, \mathbf{R}^m)$, and such that

$$\lim_{n\to+\infty} |Du_n|(A) = |Du|(A).$$

Such a sequence exists by Theorem 10.1.2. From $f \leq f(0) + g$ and Lemma 11.3.3, one has

$$\begin{aligned}\mu(A) &\leq Qf(0)|A| + \lim_{n\to+\infty} \int_A g\left(\frac{Du_n}{|Du_n|}\right) |Du_n| \\ &= Qf(0)|A| + \int_A g\left(\frac{Du}{|Du|}\right) |Du|.\end{aligned}$$

The same inequality now holds true for any Borel set B of Ω. Taking the singular part of the two members and according to the rank-one property of the singular part of Du, we obtain

$$\mu^s(B) \leq \int_B (Qf)^\infty\left(\frac{D^s u}{|D^s u|}\right) |D^s u|.$$

We are going to estimate from above the absolutely continuous part μ^a, i.e., (11.46). Let ρ_n be a regularizing kernel (see Theorem 4.2.2) and set $u_n = \rho_n * u$. From

$$\nabla u_n = \rho_n * Du = \rho_n * \nabla u + \rho_n * D^s u$$

and the local Lipschitz continuity of Qf, one has, for every open set $A' \subset\subset A$,

$$\begin{aligned}\int_{A'} Qf(\nabla u_n)\, dx &\leq \int_{A'} Qf(\rho_n * \nabla u)\, dx + L'|\rho_n * D^s u|(A') \\ &\leq \int_{A'} Qf(\rho_n * \nabla u)\, dx + L'|D^s u|(A' + \mathrm{spt}(\rho_n)).\end{aligned}$$

Letting $n \to +\infty$ yields

$$\mu(A') \leq \int_{A'} Qf(\nabla u)\, dx + L'|D^s u|(A)$$

and, since $A' \subset\subset A$ is arbitrary,

$$\mu(A) \leq \int_A Qf(\nabla u)\, dx + L'|D^s u|(A)$$

for every open subset A of Ω. Since the Lebesgue measure and the nonnegative Borel measure $|D^s u|$ are regular, the same inequality now holds true for every Borel subset B of Ω. Taking the absolutely continuous part of each member, we finally obtain

$$\mu^a(B) \leq \int_B Qf(\nabla u)\, dx$$

and the proof is complete. □

We now compute the lsc envelope of the integral functional in (11.34) by taking into account a boundary condition on a part Γ_0 of the boundary $\partial\Omega$ of Ω. More precisely, we aim to describe the lsc envelope of the integral functional $F : L^1(\Omega, \mathbf{R}^m) \longrightarrow \mathbf{R}^+ \cup \{+\infty\}$ defined by

$$F(u) = \begin{cases} \displaystyle\int_\Omega f(\nabla u)\, dx \text{ if } u \in W^{1,1}_{\Gamma_0}(\Omega, \mathbf{R}^m), \\ +\infty \text{ otherwise.} \end{cases}$$

Corollary 11.3.1. *The lsc envelope of the integral functional F is given by*

$$cl(F)(u) = \begin{cases} \displaystyle\int_\Omega Qf(\nabla u)\,dx + \int_\Omega (Qf)^\infty\left(\frac{D^s u}{|D^s u|}\right)|D^s u| \\ \qquad\qquad + \displaystyle\int_{\Gamma_0} (Qf)^\infty(\gamma_0(u)\otimes \nu)\,d\mathcal{H}^{N-1} \text{ if } u \in BV(\Omega,\mathbf{R}^m), \\ +\infty \text{ otherwise}, \end{cases}$$

where ν denotes the outer unit normal to Γ_0 and γ_0 the trace operator.

In other words, in this relaxation process, the boundary condition is translated in terms of surface energy $\int_\Omega (Qf)^\infty([u]\otimes\nu)\,d\mathcal{H}^{N-1}\lfloor\Gamma_0$. The proof is very similar to that of Proposition 11.3.2. For details, consult [1] and the references therein. As a consequence we obtain the relaxation theorem in the case $p = 1$.

Theorem 11.3.2 (relaxation theorem, $p = 1$). *Let us consider a function $f : \mathbf{M}^{m\times N} \longrightarrow \mathbf{R}$ satisfying* (11.5) *and* (11.6) *and g in $L^\infty(\Omega,\mathbf{R}^m)$ satisfying $\|g\|_\infty < \frac{\alpha}{C_P}$, where C_P is the Poincaré constant in Ω. Then the relaxed problem of*

$$\inf\left\{\int_\Omega f(\nabla u)\,dx - \int_\Omega g.u\,dx \;:\; u \in W^{1,1}_{\Gamma_0}(\Omega,\mathbf{R}^m)\right\} \qquad (\mathcal{P})$$

in the sense of Theorem 11.1.2 *is given by*

$$\min\left\{\int_\Omega Qf(\nabla u)\,dx - \int_\Omega (Qf)^\infty\left(\frac{D^s u}{|D^s u|}\right)|D^s u| + \int_{\Gamma_0}(Qf)^\infty(\gamma_0(u)\otimes\nu)\,d\mathcal{H}^{N-1} - \int_\Omega g.u\,dx \;:\; u\in BV(\Omega,\mathbf{R}^m)\right\}. \quad (\overline{\mathcal{P}})$$

Proof. Note that from the estimate $\|g\|_\infty < \frac{\alpha}{C_P}$, we easily see that the infimum is finite. Arguing as in the proof of Theorem 11.2.2, according to Theorem 11.1.2, it suffices to establish the inf-compactness of

$$u \mapsto F(u) - \int_\Omega g.u\,dx$$

in $L^1(\Omega,\mathbf{M}^{m\times N})$. This property is a straightforward consequence of the estimate $\|g\|_\infty < \frac{\alpha}{C_P}$ and the compactness of the embedding of $W^{1,1}_{\Gamma_0}(\Omega,\mathbf{R}^m)$ into $L^1(\Omega,\mathbf{M}^{m\times N})$. □

11.4 Relaxation in the space of Young measures in nonlinear elasticity

The strategy described in Sections 11.2 and 11.3 has the disadvantage to quasi-convexify the density function f so that the relaxed functional, with density Qf, does not provide information on the oscillations of the gradient minimizing sequences. In this section, we describe an alternative way for relaxing the free energy by using the notion of Young measure

introduced in Section 4.3, well adapted for capturing oscillations of minimizing sequences (see Subsection 4.3.6). According to the point of view of Ball and James in [56] or Bhattacharya and Kohn in [69], the density of the relaxed free-energy obtained this way is the microscopic free-energy density corresponding to the macroscopic free-energy density Qf.

11.4.1 Young measures generated by gradients

To shorten notations, we denote the N-dimensional Lebesgue measure restricted to the open bounded subset $\Omega \subset \mathbf{R}^N$ by $\mathcal{L}$.

Definition 11.4.1. *Let us denote by $E := \mathbf{R}^{mN} \sim \mathbf{M}^{m\times N}$, the set of $m \times N$ matrices. A Young measure μ in $\mathcal{Y}(\Omega; E)$ is called a $W^{1,p}$-Young measure if there exists a bounded sequence $(u_n)_{n\in\mathbf{N}}$ in $W^{1,p}(\Omega, \mathbf{R}^m)$, $p \geq 1$, such that μ is generated by the sequence of gradients $(\nabla u_n)_{n\in\mathbf{N}}$, i.e.,*

$$(\delta_{\nabla u_n(x)})_{x\in\Omega} \otimes \mathcal{L} \overset{nar}{\rightharpoonup} \mu,$$

or, equivalently, from Theorem 4.3.1,

$$(\delta_{\nabla u_n(x)})_{x\in\Omega} \overset{L_w}{\rightharpoonup} (\mu_x)_{x\in\Omega}.$$

We admit the following important technical lemma. For a proof, consult Kinderlehrer and Pedregal [162], Pedregal [195], or Fonseca, Müller, and Pedregal [137].

Lemma 11.4.1. *Let $(u_n)_{n\in\mathbf{N}}$ be a sequence in $W^{1,p}(\Omega, \mathbf{R}^m)$ weakly converging to some u in $W^{1,p}(\Omega, \mathbf{R}^m)$. Then, there exists a sequence $(v_n)_{n\in\mathbf{N}}$ in $W^{1,p}(\Omega, \mathbf{R}^m)$ satisfying*

(i) $v_n \in u + W_0^{1,p}(\Omega, \mathbf{R}^m)$;

(iii) $(|\nabla v_n|^p)_{n\in\mathbf{N}}$ *is uniformly integrable;*

(iii) $v_n - u_n \to 0$ *and* $\nabla(v_n - u_n) \to 0$ *in measure.*

Let us point out that according to item (iii) and to Proposition 4.3.8, the two sequences $(\nabla u_n)_{n\in\mathbf{N}}$ and $(\nabla v_n)_{n\in\mathbf{N}}$ generate the same $W^{1,p}$-Young measure.

The main result of this section is the following characterization of $W^{1,p}$-Young measures, established by Kinderlehrer and Pedregal (see Kinderlehrer and Pedregal [162], Pedregal [195], Sychev [215]).

Theorem 11.4.1. *Let $p > 1$; then $\mu \in \mathcal{Y}(\Omega; E)$ is a $W^{1,p}$-Young measure iff there exists $u \in W^{1,p}(\Omega, \mathbf{R}^m)$ such that the three following assertions hold:*

(i) $\nabla u(x) = \int_E \lambda \, d\mu_x(\lambda)$ *for a.e. x in Ω;*

(ii) *for all quasi-convex function $\phi : E \to \mathbf{R}$ for which there exist some $\gamma \in \mathbf{R}$ and $\beta > 0$ such that $\gamma \leq \phi(\lambda) \leq \beta(1 + |\lambda|^p)$ for all $\lambda \in E$, one has*

$$\phi(\nabla u(x)) \leq \int_E \phi(\lambda) \, d\mu_x(\lambda) \text{ for a.e. } x \text{ in } \Omega;$$

(iii) $\int_{\Omega\times E} |\lambda|^p \, d\mu(x,\lambda) < +\infty$.

The function u will be refered as the underlying deformation of the Young measure μ.

PROOF. We split the proof into several steps.

PROOF OF THE NECESSARY CONDITIONS.

First step: Necessity of (i) *and* (iii). By definition, there exists a bounded sequence $(u_n)_{n\in\mathbf{N}}$ in $W^{1,p}(\Omega,\mathbf{R}^m)$ weakly converging to some $u \in W^{1,p}(\Omega,\mathbf{R}^m)$, whose sequence of gradients generates μ. Since $\nabla u_n \rightharpoonup \nabla u$ weakly in $L^p(\Omega,E)$, one obtains (i) by applying Proposition 4.3.6. Take now $\varphi(x,\lambda) = |\lambda|^p$. Since $\lambda \mapsto \varphi(x,\lambda)$ is lsc (actually continuous), according to Proposition 4.3.3, we deduce

$$\int_{\Omega\times E} |\lambda|^p \, d\mu(x,\lambda) \le \liminf_{n\to+\infty} \int_\Omega |\nabla u_n|^p \, dx < +\infty.$$

Second step: Necessity of (ii). Let ϕ be a quasi-convex function satisfying the growth condition in (ii) and x_0 a fixed element in Ω such that the two following limits exist:

$$\phi(\nabla u(x_0)) = \lim_{\rho\to 0} \frac{1}{\mathcal{L}(B_\rho(x_0))} \int_{B_\rho(x_0)} \phi(\nabla u(x)) \, dx,$$

$$\int_E \phi(\lambda) \, d\mu_{x_0}(\lambda) = \lim_{\rho\to 0} \frac{1}{\mathcal{L}(B_\rho(x_0))} \int_{B_\rho(x_0)} \left(\int_E \phi(\lambda) \, d\mu_x(\lambda) \right) dx.$$

Such x_0 exists outside a negligible set, from the Lebesgue differentiation theorem. Let us set $\varphi(x,\lambda) = \frac{1}{\mathcal{L}(B_\rho(x_0))} 1_{B_\rho(x_0)}(x)\phi(\lambda)$ which defines a $\mathcal{B}(\Omega)\otimes\mathcal{B}(E)$-measurable function, continuous with respect to λ. Consider now the sequence $(v_n)_{n\in\mathbf{N}}$ given by Lemma 11.4.1 with $\Omega = B_\rho(x_0)$. From the growth condition fulfilled by ϕ, the sequence $(\varphi(x,\nabla v_n(x)))_{n\in\mathbf{N}}$ is uniformly integrable. Therefore, by applying Theorem 4.3.3, we obtain

$$\begin{aligned} \int_{\Omega\times E} \varphi(x,\lambda) \, d\mu(x,\lambda) &= \lim_{n\to+\infty} \frac{1}{\mathcal{L}(B_\rho(x_0))} \int_{B_\rho(x_0)} \phi(\nabla v_n(x)) \, dx \\ &\ge \frac{1}{\mathcal{L}(B_\rho(x_0))} \int_{B_\rho(x_0)} \phi(\nabla u(x)) \, dx. \end{aligned} \tag{11.48}$$

The last inequality is a consequence of lower semicontinuity of the integral functional

$$v \mapsto \int_{B_\rho(x_0)} \phi(\nabla v(x)) \, dx$$

when $W^{1,p}(B_\rho(x_0),\mathbf{R}^m)$ is equipped with its weak convergence, due to the quasi-convexity assumption on ϕ (see Theorem 13.2.1). But, according to the slicing theorem, Theorem 4.2.4,

$$\int_{\Omega\times E} \varphi(x,\lambda) \, d\mu(x,\lambda) = \frac{1}{\mathcal{L}(B_\rho(x_0))} \int_{B_\rho(x_0)} \left(\int_E \phi(\lambda) \, d\mu_x(\lambda) \right) dx$$

so that (11.48) yields

$$\frac{1}{\mathcal{L}(B_\rho(x_0))}\int_{B_\rho(x_0)}\left(\int_E \phi(\lambda)\,d\mu_x(\lambda)\right)dx \geq \frac{1}{\mathcal{L}(B_\rho(x_0))}\int_{B_\rho(x_0)}\phi(\nabla u(x))\,dx.$$

The conclusion then follows by letting $\rho \to 0$.

PROOF OF THE SUFFICENT CONDITIONS. The proof of the sufficient conditions is more involved. Furthermore, we prove that any Young measure satisfying conditions (i), (ii), and (iii) is generated by the sequence of gradients of a bounded sequence in $u + W_0^{1,p}(\Omega, \mathbf{R}^m)$. For the convenience of the reader we divide the proof into several steps.

First step: The Young measure μ is assumed to be homogeneous. Let a be a fixed matrix in E, Q a fixed open bounded cube of $\mathbf{R}^N$ and let l_a be the linear function defined for every $x \in Q$ by $l_a(x) = a.x$. We consider the closed set $\mathcal{H}_a(E)$ of probability measures μ on E satisfying the three following conditions:

(i) $\int_E \lambda\,d\mu = a$;

(ii) for all quasi-convex function ϕ satisfying $\forall\lambda \in E, \gamma \leq \phi(\lambda) \leq \beta(1+|\lambda|^p)$ for some $\gamma \in \mathbf{R}$ and $\beta > 0$, one has $\phi(a) \leq \int_E \phi(\lambda)\,d\mu(\lambda)$;

(iii) $\int_E |\lambda|^p\,d\mu < +\infty$.

Note that these three conditions are exactly the one stated in Theorem 11.4.1, fulfilled by the homogeneous Young measure $\mu \otimes \mathcal{L}\lfloor\Omega$. In this step we set $\Omega = Q$ and we aim to establish the existence of a bounded sequence $(u_n)_{n\in\mathbf{N}}$ in $l_a + W_0^{1,p}(Q, \mathbf{R}^m)$ such that the sequence $(\nabla u_n)_{n\in\mathbf{N}}$ generates the Young measure $\mu \otimes \mathcal{L}\lfloor Q$. In the second step, we will apply this result in the case when $a = 0$.

For every $v \in l_a + W^{1,p}(Q, \mathbf{R}^m)$, let us consider the following probability measure on E

$$\mu_v := \frac{1}{\mathcal{L}(Q)}\int_Q \delta_{\nabla v(x)}\,dx,$$

which acts on every $\varphi \in \mathbf{C}_0(E)$ as follows:

$$\langle \mu_v, \varphi\rangle = \frac{1}{\mathcal{L}(Q)}\int_Q \varphi(\nabla v(x))\,dx.$$

Note that $\int_E |\lambda|^p\,d\mu_v$ is well defined and precisely $\int_E |\lambda|^p\,d\mu_v = \frac{1}{\mathcal{L}(Q)}\int_Q |\nabla v(x)|^p\,dx$. We finally consider the following convex subset of $\mathcal{H}_a(E)$:

$$\mathcal{C}_a(E) := \left\{\mu_v \,:\, v \in l_a + W^{1,p}(Q, \mathbf{R}^m)\right\}.$$

Let now v be a fixed element of $l_a + W_0^{1,p}(Q, \mathbf{R}^m)$, extended by Q-periodicity on $\mathbf{R}^N$, and, for every $n \in \mathbf{N}^*$, let us define the function $v_n : x \mapsto v_n(x) := \frac{1}{n}v(nx)$ in

$l_a + W^{1,p}(Q, \mathbf{R}^m)$. According to a classical result on oscillating functions, rephrased in terms of narrow convergence of Young measures, one has

$$(\delta_{\nabla v_n(x)})_{x\in\mathbf{Q}} \otimes \mathcal{L}\lfloor Q \overset{nar}{\rightharpoonup} \mu_v \otimes \mathcal{L}\lfloor Q \quad \text{when } n \to +\infty. \tag{11.49}$$

Note that the norm of ∇v_n in $L^p(Q, E)$ is exactly the one of ∇v. To conclude, it is enough to show that for every $\mu \in \mathcal{H}_a(E)$, there exists a sequence $(\mu_{w_k})_{k\in\mathbf{N}}$ in $\mathcal{C}_a(E)$, $(w_k)_{k\in\mathbf{N}}$ bounded in $l_a + W_0^{1,p}(Q, \mathbf{R}^m)$, such that

$$\mu_{w_k} \rightharpoonup \mu \text{ weakly in } \mathbf{M}(E), \text{ i.e. } \sigma(\mathbf{C}_0'(E), \mathbf{C}_0(E));$$

hence, according to Theorem 4.3.1, and since μ_{w_k} and μ are homogeneous,

$$\mu_{w_k} \otimes \mathcal{L}\lfloor Q \overset{nar}{\rightharpoonup} \mu \otimes \mathcal{L}\lfloor Q \quad \text{when } k \to +\infty. \tag{11.50}$$

Indeed, since the space $\mathcal{Y}(Q; E)$ endowed with the topology of the narrow convergence is metrizable (see [104, Proposition 2.3.1]), combining (11.49), (11.50), and by using a diagonalization argument (cf. Lemma 11.1.1), we will show that there exists a bounded sequence $(u_n)_{n\in\mathbf{N}^*}$ in $l_a + W^{1,p}(Q, \mathbf{R}^m)$, whose sequence of gradients generates the Young measure $\mu \otimes \mathcal{L}\lfloor Q$.

We are going to establish (11.50) or, equivalently, the density of $\mathcal{C}_a(E)$ in $\mathcal{H}_a(E)$ for the $\sigma(\mathbf{C}_0'(E), \mathbf{C}_0(E))$ topology. Let us point out that we want to prove the existence of a sequence $(\mu_{w_m})_{m\in\mathbf{N}}$ weakly converging to μ, such that moreover $(w_m)_{m\in\mathbf{N}}$ be bounded in $l_a + W_0^{1,p}(Q, \mathbf{R}^m)$ or, equivalently, such that $(\nabla w_m)_{m\in\mathbf{N}}$ be bounded in $L^p(Q, E)$. To take into account this condition, we establish $\overline{\mathcal{C}_a(E)} = \mathcal{H}_a(E)$ for a metric d' finer than the classical metric d inducing the $\sigma(\mathbf{C}_0'(E), \mathbf{C}_0(E))$ topology in the set $\mathbf{P}(E)$ of probability measures on E. Let us recall that given a dense countable family $(\varphi_i)_{i\in\mathbf{N}^*}$ in $\mathbf{C}_0(E)$, the distance d is given by

$$\forall (\mu, \nu) \in \mathbf{P}(E) \times \mathbf{P}(E),\ d(\mu, \nu) := \sum_{i=1}^{+\infty} \frac{1}{2^i \|\varphi_i\|_\infty} |\langle \mu, \varphi_i\rangle - \langle \nu, \varphi_i\rangle|.$$

We define now the distance d' by setting, for every $(\mu, \nu) \in \mathbf{P}(E) \times \mathbf{P}(E)$,

$$d'(\mu, \nu) := \Big| \int_E |\lambda|^p \, d\mu - \int_E |\lambda|^p \, d\nu \Big| + \sum_{i=1}^{+\infty} \frac{1}{2^i \|\varphi_i\|_\infty} |\langle \mu, \varphi_i\rangle - \langle \nu, \varphi_i\rangle|,$$

and we argue by contradiction. Let us assume that $\mathcal{C}_a(E)$ is not dense in $\mathcal{H}_a(E)$ for the metric associated with the distance d'. Then, there exists $\mu_0 \in \mathcal{H}_a(E)$, $k \in \mathbf{N}^*$ and $\eta > 0$ such that

$$\forall \nu \in \mathcal{C}_a(E),\ \Big| \int_E |\lambda|^p \, d\mu_0 - \int_E |\lambda|^p \, d\nu \Big| + \sum_{i=1}^{k} \frac{1}{2^i \|\varphi_i\|_\infty} |\langle \mu_0, \varphi_i\rangle - \langle \nu, \varphi_i\rangle| > \eta. \tag{11.51}$$

We reason now in the finite dimensional space $\mathbf{R}^{k+1}$. From (11.51), the vector

$$\left(\int_E |\lambda|^p \, d\mu_0,\ \frac{1}{2^i \|\varphi_i\|_\infty} \langle \mu_0, \varphi_i\rangle \right)_{i=1,\dots k}$$

does not belong to the closure of the following convex set of $\mathbf{R}^{k+1}$

$$\mathbf{C}_a := \left\{ \left(\int_E |\lambda|^p \, d\nu, \frac{1}{2^i \|\varphi_i\|_\infty} \langle \nu, \varphi_i \rangle \right)_{i=1,\dots k} : \nu \in \mathcal{C}_a(E) \right\}.$$

Consequently, according to the Hahn–Banach separation Theorem 9.1.1, there exist $(c_i)_{i=0,\dots k}$ in $\mathbf{R}^{k+1}$ and $\eta' > 0$ such that for all $\nu \in \mathcal{C}_a(E)$,

$$\begin{aligned} & c_0 \int_E |\lambda|^p \, d\nu + \sum_{i=1}^k \frac{1}{2^i \|\varphi_i\|_\infty} c_i \langle \nu, \varphi_i \rangle \\ & \geq \eta' + c_0 \int_E |\lambda|^p \, d\mu_0 + \sum_{i=1}^k \frac{1}{2^i \|\varphi_i\|_\infty} c_i \langle \mu_0, \varphi_i \rangle. \end{aligned} \tag{11.52}$$

Let us set

$$\phi := c_0 |.|^p + \sum_{i=1}^k \frac{c_i}{\|\varphi_i\|_\infty} \varphi_i.$$

We claim that $c_0 \geq 0$. Otherwise, taking in (11.52) $\nu = \mu_{v_t}$ with $v_t := tw + l_a$, $t > 0$, where w is a fixed function in $W_0^{1,p}(Q, \mathbf{R}^m)$, and letting $t \to +\infty$, the right-hand side in (11.52) would be $-\infty$. Replacing if necessary (when $c_0 = 0$), the function ϕ by the function $\phi + \delta |.|^p$ with $\delta > 0$ small enough (choose precisely δ satisfying $\eta' - \delta \int_E |\lambda|^p \, d\mu_0 > 0$), one may assume $c_0 > 0$. Thus the function ϕ still satisfies (11.52), for $\eta' > 0$ replaced by $\eta' - \delta \int_E |\lambda|^p \, d\mu_0 > 0$ if necessary, together with the growth conditions $c_0 |\lambda|^p + \gamma \leq \phi(\lambda) \leq \beta(1 + |\lambda|^p)$ for some constants $\gamma \in \mathbf{R}$ and $\beta > 0$. Finally, replacing ϕ by $\phi - \gamma$, one may assume that ϕ satisfies (11.52) and the growth conditions $c_0 |\lambda|^p \leq \phi(\lambda) \leq \beta(1 + |\lambda|^p)$ of Proposition 11.2.2. Therefore, (11.52) and condition (ii) satisfied by μ_0 yield

$$\inf \left\{ \frac{1}{\mathcal{L}(Q)} \int_Q \phi(\nabla v(x)) \, dx : v \in l_a + W_0^{1,p}(Q, \mathbf{R}^m) \right\} > \int_E \phi \, d\mu_0 \geq \int_E Q\phi \, d\mu_0 \geq Q\phi(a).$$

But, according to the classical variational principle (Proposition 11.2.2), since ϕ satisfies appropriate growth conditions,

$$\inf \left\{ \frac{1}{\mathcal{L}(Q)} \int_Q \phi(\nabla v(x)) \, dx : v \in l_a + W_0^{1,p}(Q, \mathbf{R}^m) \right\} = Q\phi(a),$$

a contradiction.

Second step: The Young measure μ satisfies the conditions (i), (ii), *and* (iii) *of Theorem* 11.4.1 *and $u = 0$.* We want to construct a bounded sequence $(u_n)_{n \in \mathbf{N}}$ in $W^{1,p}(\Omega, \mathbf{R}^m)$ whose gradients generate μ. The idea of the proof consists in "localizing" $(\mu_x)_{x \in \Omega}$ thanks to Vitali's covering theorem, to apply the previous step for each localization, then to stick together the sequences of functions whose gradients generate each localized Young measures.

According to Vitali's covering Lemma 4.1.2 and Remark 4.1.4, for every $k \in \mathbf{N}^*$, there exists a finite family $(Q_{i,k})_{i \in I_k}$ of pairwise disjoint open cubes included in Ω satisfying

$$\mathcal{L}\left(\Omega \setminus \bigcup_{i \in I_k} Q_{i,k} \right) \leq \frac{1}{k}, \quad \operatorname{diam}(Q_{i,k}) < \frac{1}{k}. \tag{11.53}$$

For every $i \in I_k$, let us define the probability measure $\mu_{i,k}$ on E by

$$\mu_{i,k} := \frac{1}{\mathcal{L}(Q_{i,k})} \int_{Q_{i,k}} \mu_x \, dx,$$

which acts on every $\varphi \in \mathbf{C}_0(E)$ as follows:

$$\langle \mu_{i,k}, \varphi \rangle = \frac{1}{\mathcal{L}(Q_{i,k})} \int_{Q_{i,k}} \left(\int_E \varphi(\lambda) \, d\mu_x(\lambda) \right) dx.$$

With the notations of the first step, it is easy to show that $\mu_{i,k}$ belongs to $\mathcal{H}_0(E)$. Consequently, according to step 1, for every $i \in I_k$, there exists a bounded sequence $(v_{i,k,n})_{n\in\mathbf{N}}$ in $W_0^{1,p}(Q_{i,k}, \mathbf{R}^m)$ such that

$$(\delta_{\nabla v_{i,k,n}(x)})_{x\in Q_{i,k}} \otimes \mathcal{L}\lfloor Q_{i,k} \overset{nar}{\rightharpoonup} \mu_{i,k} \otimes \mathcal{L}\lfloor Q_{i,k} \quad \text{when } n \to +\infty. \tag{11.54}$$

By using Lemma 11.4.1, one may furthermore assume the sequence $(|\nabla v_{i,k,n}|^P)_{n\in\mathbf{N}}$ uniformly integrable so that, from Theorem 4.3.3, (11.54) yields

$$\begin{aligned} \lim_{n\to+\infty} \int_{Q_{i,k}} |\nabla v_{i,k,n}|^p \, dx &= \int_{Q_{i,k}} \int_E |\lambda|^p \, d\mu_{i,k} \, dx \\ &= \frac{1}{\mathcal{L}(Q_{i,k})} \int_{Q_{i,k}} \left(\int_E |\lambda|^p \, d\mu_x \right) dx. \end{aligned} \tag{11.55}$$

We now stick together the functions $v_{i,k,n}$, $i \in I_k$, by setting

$$v_{k,n}(x) := \begin{cases} v_{i,k,n} \text{ if } x \in Q_{i,k}, \\ \\ 0 \text{ if } x \in \Omega \setminus \bigcup_{i\in I_k} Q_{i,k}. \end{cases}$$

Take now θ in a dense subset of regular functions of $L^1(\Omega)$, $\varphi \in C_0(E)$ and set $R_k := \int_{\Omega\setminus\bigcup_{i\in I_k} Q_{i,k}} \theta(x)\varphi(0) \, dx$. Note that from (11.53), $\lim_{k\to+\infty} R_k = 0$. From (11.54), the definition of $\mu_{i,k}$, and according to the mean value theorem, one has

$$\begin{aligned} \lim_{n\to+\infty} \int_\Omega \theta(x)\varphi(\nabla v_{k,n}(x)) \, dx &= \lim_{n\to+\infty} \sum_{i\in I_k} \int_{Q_{i,k}} \theta(x)\varphi(\nabla v_{k,n}(x)) \, dx + R_k \\ &= \sum_{i\in I_k} \int_{Q_{i,k}} \theta(x)\langle \mu_{i,k}, \varphi \rangle \, dx + R_k \\ &= \sum_{i\in I_k} \int_{Q_{i,k}} \left(\int_E \varphi \, d\mu_y \right) dy \; \frac{1}{\mathcal{L}(Q_{i,k})} \int_{Q_{i,k}} \theta(x) \, dx + R_k \\ &= \sum_{i\in I_k} \int_{Q_{i,k}} \theta(x_{i,k}) \left(\int_E \varphi \, d\mu_y \right) dy + R_k \end{aligned} \tag{11.56}$$

for some $x_{i,k} \in Q_{i,k}$. We stress the fact that the convergence in (11.56) may be taken in the narrow convergence sense. Indeed, setting

$$\tilde{\mu}_x = \begin{cases} \mu_{i,k} \text{ if } x \in Q_{i,k}, \\ \delta_0 \text{ if } x \in \Omega \setminus \bigcup_{i \in I_k} Q_{i,k}, \end{cases}$$

one defines a Young measures $(\tilde{\mu}_x)_{x\in\Omega} \otimes \mathcal{L}$ and estimate (11.56) yields $\lim_{n\to+\infty} (\delta_{\nabla v_{k,n}(x)})_{x\in\Omega}$ $\otimes \mathcal{L} = (\tilde{\mu}_x)_{x\in\Omega} \otimes \mathcal{L}$ in the narrow convergence sense in $\mathcal{Y}(\Omega; E)$. Letting $k \to +\infty$ in (11.56), from (11.53), one obtains

$$\begin{aligned} \lim_{k\to+\infty} \lim_{n\to+\infty} \int_\Omega \theta(x)\varphi(\nabla v_{k,n}(x))\, dx &= \lim_{k\to+\infty} \sum_{i\in I_k} \int_{Q_{i,k}} \theta(x_{i,k}) \Big(\int_E \varphi \, d\mu_y \Big)\, dy \\ &= \lim_{k\to+\infty} \sum_{i\in I_k} \int_{Q_{i,k}} \theta(y) \Big(\int_E \varphi \, d\mu_y \Big)\, dy \\ &= \int_\Omega \theta(y) \Big(\int_E \varphi \, d\mu_y \Big)\, dy. \end{aligned}$$

This shows that $\lim_{k\to+\infty} \lim_{n\to+\infty} (\delta_{\nabla v_{k,n}(x)})_{x\in\Omega} \otimes \mathcal{L} = (\mu_x)_{x\in\Omega} \otimes \mathcal{L}$, where each limit must be taken in the narrow convergence sense in $\mathcal{Y}(\Omega; E)$. According to the diagonalization Lemma 11.1.1, there exists a map $n \mapsto k(n)$ such that

$$(\delta_{\nabla v_{k(n),n}(x)})_{x\in\Omega} \otimes \mathcal{L} \overset{nar}{\rightharpoonup} (\mu_x)_{x\in\Omega} \otimes \mathcal{L}$$

when $n \to +\infty$. Furthermore, from (11.55), the sequence $(v_{k(n),n})_{n\in\mathbf{N}}$ is bounded in $W^{1,p}(\Omega, \mathbf{R}^m)$. Setting $u_n := v_{k(n),n}$ shows what is required.

Last step: The Young measure μ satisfies the conditions (i), (ii), *and* (iii) *of Theorem* 11.4.1 *without condition on u in $W^{1,p}(\Omega, \mathbf{R}^m)$.* Given $\mu \in \mathcal{Y}(\Omega; E)$ satisfying (i), (ii), and (iii), let us consider $\tilde{\mu} = (\tilde{\mu}_x)_{x\in\Omega} \otimes \mathcal{L}$ in $\mathcal{Y}(\Omega; E)$ defined by $\langle \tilde{\mu}_x, \varphi \rangle = \langle \mu_x, \varphi(. - \nabla u(x)) \rangle$ for every $\varphi \in L^p_{\mu_x}(E)$ and for a.e. x in Ω. It is straightforward to check that $\tilde{\mu}$ satisfies the conditions of step 2. Thus, there exists a sequence $(v_n)_{n\in\mathbf{N}}$ in $W^{1,p}_0(\Omega, \mathbf{R}^m)$ such that $(\nabla v_n)_{n\in\mathbf{N}}$ generates $\tilde{\mu}$. Consider $u_n := v_n + u$ in $W^{1,p}(\Omega, \mathbf{R}^m)$. For each $\psi \in \mathbf{C}_b(\Omega; E)$, let us define $\tilde{\psi}$ in $\mathbf{C}_b(\Omega; E)$ by setting $\tilde{\psi}(x, \lambda) := \psi(x, \lambda + \nabla u(x))$. Thus, for every $\psi \in \mathbf{C}_b(\Omega; E)$ we have

$$\begin{aligned} \lim_{n\to+\infty} \int_\Omega \psi(x, \nabla u_n(x))\, dx &= \lim_{n\to+\infty} \int_\Omega \tilde{\psi}(x, \nabla v_n(x))\, dx \\ &= \int_{\Omega\times E} \tilde{\psi}(x, \lambda)\, d\tilde{\mu}(x, \lambda) \\ &= \int_{\Omega\times E} \psi(x, \lambda) d\mu(x, \lambda), \end{aligned}$$

which proves that $(\nabla u_n)_{n\in\mathbf{N}}$ generates μ. □

Remark 11.4.1. When considering functions depending on x and u, the necessary condition (ii) must be replaced by condition (ii)′ below: for every $\Phi : \Omega \times \mathbf{R}^m \times E \longrightarrow \mathbf{R}$ such that

for a.e. x in Ω, $\phi(x, u(x), .)$ is quasi-convex and satisfies

$$\gamma \leq \phi(x, \xi, \lambda) \leq \beta(1 + |\xi|^p + |\lambda|^p)$$

for some constants γ and $\beta > 0$, one has

$$\phi(x, u(x), \nabla u(x)) \leq \int_E \phi(x, u(x), \lambda)\, d\mu_x(\lambda) \text{ for a.e. } x \in \Omega.$$

Indeed, it suffices to argue as in the second step of the proof of the necessary conditions by setting

$$\varphi(x, \lambda) = \frac{1}{\mathcal{L}(B_\rho(x_0))} 1_{B_\rho(x_0)}(x)\phi(x, u(x), \lambda),$$

where x_0 is such that the two following limits exist:

$$\phi(x_0, u(x_0), \nabla u(x_0)) = \lim_{\rho \to 0} \frac{1}{\mathcal{L}(B_\rho(x_0))} \int_{B_\rho(x_0)} \phi(x, u(x), \nabla u(x))\, dx,$$
$$\int_E \phi(x_0, u(x_0), \lambda)\, d\mu_{x_0}(\lambda) = \lim_{\rho \to 0} \frac{1}{\mathcal{L}(B_\rho(x_0))} \int_{B_\rho(x_0)} \left(\int_E \phi(x, u(x), \lambda)\, d\mu_x \right) dx.$$

11.4.2 Relaxation of classical integral functionals in $\mathcal{Y}(\Omega; E)$

We intend to apply the relaxation procedure for the integral functionals of Section 11.2, but considered as living in the space $X = \mathcal{Y}(\Omega; E)$, equipped with the topology of the narrow convergence. The generalized solutions of the relaxed problem may be interpreted as microstructures: they capture highly oscillatory minimizing sequences on smaller and smaller spatial scales and describe fine phase mixtures in elastic crystals. Fundamental applications to real materials and polycrystals may be found in [66], [67], [68], [69], and [110]. For papers dealing with laminates and multiwell problems, see [56], [190], [214], [196], and [228].

We consider the problem

$$\inf \left\{ F(u) \ : \ u \in u_0 + W_0^{1,p}(\Omega, \mathbf{R}^m) \right\} \qquad (:= \inf(\mathcal{P})), \qquad (\mathcal{P})$$

where u_0 is a given function in $W^{1,p}(\Omega, \mathbf{R}^m)$, $p > 1$, and $F : W^{1,p}(\Omega, \mathbf{R}^m) \longrightarrow \mathbf{R}$ is the integral functional defined by

$$F(u) = \int_\Omega f(x, u, \nabla u)\, dx.$$

The density $f : \Omega \times \mathbf{R}^N \times E \longrightarrow \mathbf{R}$ is assumed to be $\mathcal{B}(\Omega) \otimes \mathcal{B}(\mathbf{R}^N) \otimes \mathcal{B}(E)$-measurable, continuous with respect to the third variable, to satisfy the continuity assumption

$$|f(x, \xi, \lambda) - f(x, \xi', \lambda)| \leq L|\xi - \xi'|(1 + |\xi|^{p-1} + |\xi'|^{p-1}) \qquad (11.57)$$

with respect to the second variable, and the usual bounds

$$\alpha(|\lambda|^p - 1) \leq f(x, \xi, \lambda) \leq \beta(1 + |\xi|^p + |\lambda|^p),$$

where $L > 0, 0 < \alpha < \beta$ are three given constants. For example, f may possess the following structure: $f(x,\xi,\lambda) = f_0(x,\xi) + f_1(x,\lambda)$. For less restrictive continuity assumptions on f, see Acerbi and Fusco [2], Dacorogna [116], and Pedregal [195].

In general $(\mathcal{P})$ has no solution and the relaxation procedure applied to this problem with $X = W^{1,p}(\Omega, \mathbf{R}^m)$ equipped with its weak convergence leads to the classical relaxed problem

$$\min\left\{\overline{F}(u) \ : \ u \in u_0 + W_0^{1,p}(\Omega, \mathbf{R}^m)\right\} \qquad (:= \min(\overline{\mathcal{P}})), \qquad (\overline{\mathcal{P}})$$

where $\overline{F}$ is the integral functional defined on $W^{1,p}(\Omega, \mathbf{R}^m)$ by

$$\overline{F}(u) = \int_\Omega Qf(x,u,\nabla u)\, dx,$$

and, for a.e. x in Ω, $Qf(x,\xi,.)$ is the quasi-convex envelope of $f(x,\xi,.)$. According to Theorem 11.1.2 and arguing as in the proof of Theorem 11.2.2, one can establish $\inf(\mathcal{P}) = \min(\overline{\mathcal{P}})$.

Because of the quasiconvexification, this procedure has the disadvantage of erasing the possible potential wells of $f(x,\xi,.)$ so that the relaxed problem does not provide much information on the behavior of minimizing sequences (see Example 11.4.1 or examples given in [142]). An alternative way, as we will see now, is to relaxe $(\mathcal{P})$ in the space $\mathcal{Y}(\Omega; E)$ equipped with the narrow convergence.

Let us firstly give some definitions and notations. For a fixed u_0 in $W^{1,p}(\Omega, \mathbf{R}^m)$, let $G\mathcal{Y}_0(\Omega; E)$ be the set of $W^{1,p}$-Young measures such that the underlying deformation u defined in Theorem 11.4.1 belongs to $u_0 + W_0^{1,p}(\Omega, \mathbf{R}^m)$. We define its subset of elementary $W^{1,p}$-Young measures by

$$EG\mathcal{Y}_0(\Omega; E) := \left\{\mu \in \mathcal{Y}(\Omega; E) : \exists u \in u_0 + W_0^{1,p}(\Omega, \mathbf{R}^m),\ \mu = (\delta_{\nabla u(x)})_{x\in\Omega} \otimes \mathcal{L}\right\}.$$

Note that $G\mathcal{Y}_0(\Omega; E)$ is nothing but the closure of $EG\mathcal{Y}_0(\Omega; E)$ in $\mathcal{Y}(\Omega; E)$. Since $u \in u_0 + W_0^{1,p}(\Omega, \mathbf{R}^m)$ is uniquely defined by its gradient $\nabla u \in L^p(\Omega, E)$, the operator

$$T : G\mathcal{Y}_0(\Omega; E) \longrightarrow u_0 + W_0^{1,p}(\Omega, \mathbf{R}^m), \quad \mu \mapsto T\mu := u,\ \nabla u(x) = \int_E \lambda\, d\mu_x(\lambda),$$

is well defined.

We now reformulate the problem $(\mathcal{P})$ in terms of Young measures by considering the integral functional $G : \mathcal{Y}(\Omega; E) \longrightarrow \mathbf{R} \cup \{+\infty\}$, defined by

$$G(\mu) = \begin{cases} \displaystyle\int_{\Omega\times E} f(x, T\mu(x), \lambda)\, d\mu(x,\lambda) \text{ if } \mu \in EG\mathcal{Y}_0(\Omega; E), \\ +\infty \text{ otherwise.} \end{cases}$$

Note that, in its domain, the functional G is nothing but the functional F. More precisely,

$$G(\mu) = \int_\Omega f(x, u(x), \nabla u(x))\, dx = F(u),$$

when $u \in u_0 + W_0^{1,p}(\Omega, \mathbf{R}^m)$ and $\mu = (\delta_{\nabla u(x)})_{x\in\Omega} \otimes \mathcal{L}$. The problem $(\mathcal{P})$ is then equivalent to

$$\inf\big\{G(\mu) \ : \ \mu \in \mathcal{Y}(\Omega; E)\big\}.$$

On the other hand, let us define the integral functional $\overline{G} : \mathcal{Y}(\Omega; E) \longrightarrow \mathbf{R} \cup \{+\infty\}$ by setting

$$\overline{G}(\mu) = \begin{cases} \displaystyle\int_{\Omega\times E} f(x, T\mu(x), \lambda)\, d\mu(x, \lambda) \text{ if } \mu \in G\mathcal{Y}_0(\Omega; E), \\ +\infty \text{ otherwise.} \end{cases}$$

According to Theorem 11.4.1(iii) and to the growth conditions fulfilled by f, the domain of $\overline{G}$ is $G\mathcal{Y}_0(\Omega; E)$. The functional $\overline{G}$ is nothing but the natural extension of G to $G\mathcal{Y}_0(\Omega; E)$. In the following theorem, we show that $\overline{G}$ is the sequential lower semicontinuous envelope of G when $\mathcal{Y}(\Omega; E)$ is equipped with the narrow convergence and we make precise the relationship between the two relaxed problems in $W^{1,p}(\Omega, \mathbf{R}^m)$ and $\mathcal{Y}(\Omega; E)$.

Theorem 11.4.2. *The sequential lower semicontinuous envelope of G in $\mathcal{Y}(\Omega; E)$ equipped with the narrow convergence is the extended functional $\overline{G}$. Moreover, we have*

$$\inf(\mathcal{P}) = \min(\overline{\mathcal{P}}) = \min(\overline{\mathcal{P}}^{young}),$$

where

$$\min(\overline{\mathcal{P}}^{young}) := \min\big\{\overline{G}(\mu) \ : \ \mu \in \mathcal{Y}(\Omega; E)\big\},$$

and if $\overline{\mu}$ is a solution of $\min(\overline{\mathcal{P}}^{young})$*, then $\overline{u} = T\overline{\mu}$ is a solution of* $\min(\overline{\mathcal{P}})$*. Furthermore, if $(u_n)_{n\in\mathbf{N}}$ is a $1/n$-minimizing sequence of the problem* $\inf(\mathcal{P})$*, then every cluster point of $(\delta_{\nabla u_n(x)})_{x\in\Omega} \otimes \mathcal{L}$ converges narrowly to a solution of* $\min(\overline{\mathcal{P}}^{young})$*.*

Proof. We begin by proving that $\overline{G}$ is the lsc envelope of G. According to Proposition 11.1.1 and Theorem 11.1.1 we must establish the two following assertions: for every $\mu \in \mathcal{Y}(\Omega; E)$,

$$\forall \mu_n \overset{nar}{\rightharpoonup} \mu,\ \overline{G}(\mu) \le \liminf_{n\to+\infty} G(\mu_n), \tag{11.58}$$

$$\text{there exists a sequence } (\nu_n)_{n\in\mathbf{N}} \text{ in } \mathcal{Y}(\Omega; E),\ \nu_n \overset{nar}{\rightharpoonup} \mu, \text{ such that } \overline{G}(\mu) \ge \limsup_{n\to+\infty} G(\nu_n). \tag{11.59}$$

In a second step, to apply Theorem 11.1.1, we will establish the compactness of every minimizing sequence of $\inf(\mathcal{P})$, in terms of Young measures. We will conclude by giving the relations between a solution of $\min(\overline{\mathcal{P}}^{young})$ and the corresponding underlying deformation.

First step: Proof of (11.58). One may assume $\liminf_{n\to+\infty} G(\mu_n) < +\infty$, so that, for a nonrelabeled subsequence, one has $G(\mu_n) < +\infty$, $\mu_n \in EG\mathcal{Y}_0(\Omega; E)$, $\mu \in G\mathcal{Y}_0(\Omega; E)$, and

$$G(\mu_n) = \int_\Omega f(x, u_n(x), \nabla u_n(x))\, dx.$$

According to the coerciveness assumption on f, there exists $u \in u_0 + W_0^{1,p}(\Omega, \mathbf{R}^m)$ such that $u_n \rightharpoonup u$ weakly in $W^{1,p}(\Omega, \mathbf{R}^m)$ and strongly in $L^p(\Omega, \mathbf{R}^m)$. Let us write

$$
\begin{aligned}
G(\mu_n) &= \int_\Omega f(x, u_n(x), \nabla u_n(x))\, dx \\
&= \int_{\Omega\times E} f(x, u(x), \lambda)\, d\mu_n(x,\lambda) \\
&\quad + \left(\int_\Omega f(x, u_n(x), \nabla u_n(x))\, dx - \int_\Omega f(x, u(x), \nabla u_n(x))\, dx \right).
\end{aligned}
$$

From continuity assumption (11.57), and the boundedness of the sequence $(u_n)_{n\in\mathbf{N}}$ in $W^{1,p}(\Omega, \mathbf{R}^m)$, the second term in the right-hand side tends to zero. Since, for a.e. $x \in \Omega$, $\lambda \mapsto f(x, u(x), \lambda)$ is lsc (actually continuous), the conclusion then follows from Proposition 4.3.3.

Second step: Proof of (11.59). One may assume $\overline{G}(\mu) < +\infty$ so that $\mu \in G\mathcal{Y}_0(\Omega; E)$. According to the definition of $G\mathcal{Y}_0(\Omega; E)$, there exists $u \in u_0 + W_0^{1,p}(\Omega, \mathbf{R}^m)$ and a bounded sequence $(u_n)_{n\in\mathbf{N}}$ in $W^{1,p}(\Omega, \mathbf{R}^m)$ such that $\mu_n = (\delta_{\nabla u_n(x)})_{x\in\Omega} \otimes \mathcal{L}$ narrowly converges to μ. But, from Lemma 11.4.1, there exists a sequence $(v_n)_{n\in\mathbf{N}}$ in $W^{1,p}(\Omega, \mathbf{R}^m)$ satisfying

(i) $v_n \in u + W_0^{1,p}(\Omega, \mathbf{R}^m)$;

(ii) $(|\nabla v_n|^p)_{n\in\mathbf{N}}$ is uniformly integrable;

(iii) $v_n - u_n \to 0$ and $\nabla(v_n - u_n) \to 0$ in measure.

Hence by (iii) and Proposition 4.3.8, $\nu_n := (\delta_{\nabla v_n(x)})_{x\in\Omega} \otimes \mathcal{L}$ narrowly converges to μ. Moreover, it is easily seen that up to a subsequence, v_n strongly converges to u in $L^p(\Omega, \mathbf{R}^m)$. From now on we consider a nonrelabeled subsequence of $(v_n)_{n\in\mathbf{N}}$ such that v_n strongly converges to u in $L^p(\Omega, \mathbf{R}^m)$. Let us write

$$
\begin{aligned}
\int_\Omega f(x, v_n, \nabla v_n)\, dx &= \int_\Omega f(x, u, \nabla v_n)\, dx \\
&\quad + \left(\int_\Omega f(x, v_n, \nabla v_n)\, dx - \int_\Omega f(x, u, \nabla v_n)\, dx \right). \quad (11.60)
\end{aligned}
$$

According to the growth condition satisfied by f, the sequence $(f(x, u, \nabla v_n))_{n\in\mathbf{N}}$ is also uniformly integrable, so that, by Theorem 4.3.3,

$$
\lim_{n\to+\infty} \int_\Omega f(x, u, \nabla v_n)\, dx = \int_{\Omega\times E} f(x, u, \lambda)\, d\mu(x,\lambda). \quad (11.61)
$$

On the other hand, from (11.57) and the boundedness of the sequence $(v_n)_{n\in\mathbf{N}}$ in $L^p(\Omega, \mathbf{R}^m)$, the second term in the right-hand side of (11.60) tends to zero. Collecting (11.60) and (11.61), we obtain

$$
\lim_{n\to+\infty} G(v_n) = \overline{G}(\mu),
$$

which completes the proof of the second step.

Third step: Compactness of minimizing sequences. Let $(\mu_n)_{n\in\mathbf{N}}$ be a minimizing sequence of $\inf(\mathcal{P})$. Since $G(\mu_n) < +\infty$, there exists $(u_n)_{n\in\mathbf{N}}$ in $u_0 + W_0^{1,p}(\Omega, \mathbf{R}^m)$ such that $\mu_n = (\delta_{\nabla u_n(x)})_{x\in\Omega} \otimes \mathcal{L}$ and $G(\mu_n) = F(u_n)$. According to the coerciveness

assumption on f, the sequence $(\nabla u_n)_{n\in\mathbf{N}}$ is bounded in $L^p(\Omega, E)$; thus, from Remark 4.3.3, the sequence $(\mu_n)_{n\in\mathbf{N}}$ is tight. The conclusion then follows by applying Prokhorov's theorem, Theorem 4.3.2.

Last step. Let $\overline{\mu}$ be a solution of $\min(\overline{\mathcal{P}}^{young})$ and $\overline{u} = T\overline{\mu}$. According to Theorem 11.4.1(ii) and Remark 11.4.1, we have for a.e. $x \in \Omega$,

$$\begin{aligned} Qf(x, \overline{u}(x), \nabla\overline{u}(x)) &\leq \int_E Qf(x, \overline{u}(x), \lambda)\, d\overline{\mu}_x(\lambda) \\ &\leq \int_E f(x, \overline{u}(x), \lambda)\, d\overline{\mu}_x(\lambda). \end{aligned}$$

Therefore,

$$\begin{aligned} \int_\Omega Qf(x, \overline{u}(x), \nabla\overline{u}(x))\, dx &\leq \int_{\Omega\times E} f(x, T\overline{\mu}(x), \lambda)\, d\overline{\mu}(x, \lambda) \\ &= \min(\overline{\mathcal{P}}^{young}) = \inf(\mathcal{P}) = \min(\overline{\mathcal{P}}). \end{aligned}$$

This shows that $\overline{u} = T\overline{\mu}$ is a solution of $\min(\overline{\mathcal{P}})$. The last assertion is a straightforward consequence of Theorem 11.1.2. □

Example 11.4.1. Let us illustrate Theorem 11.4.2 with the classical example: take $\Omega = (0, 1)$, $p = 4$, $u_0 = 0$, and $\varphi(\xi, \lambda) = (\lambda^2 - 1)^2 + \xi^2$; see Figure 11.1. The map $\lambda \mapsto \varphi(\xi, \lambda)$ possesses the two potential wells ± 1 and $\inf(\mathcal{P}) = 0$ has no solution.

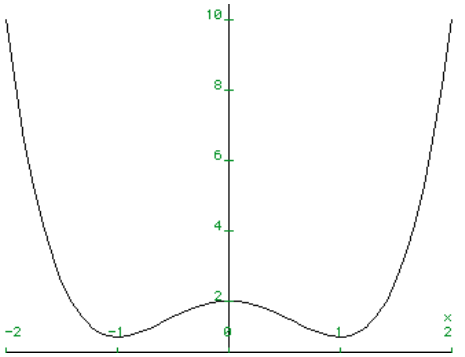


Figure 11.1. *The graph of* $\lambda \mapsto \varphi(\xi, \lambda)$ *with* $\xi = 1$.

For each fixed ξ, the quasi-convex envelope of $\lambda \mapsto (\lambda^2 - 1)^2 + \xi^2$ (equivalently, its convex envelope) is given by

$$Q\varphi(x, \xi, \lambda) = \begin{cases} (\lambda^2 - 1)^2 + \xi^2 \text{ if } |\lambda| \geq 1, \\ \xi^2 \text{ otherwise,} \end{cases}$$

so that $\min(\overline{\mathcal{P}}) = 0$ possesses a unique solution $\overline{u} = 0$. Therefore, from Theorem 11.4.2 we have $\min(\overline{\mathcal{P}}^{young}) = 0$ and if $\overline{\mu}$ is a solution

$$\int_{(0,1)} \left(\int_{\mathbf{R}} ((\lambda^2 - 1)^2 + (T\overline{\mu})^2) d\overline{\mu}_x(\lambda) \right) dx = 0, \quad x \ a.e. \text{ on } (0,1),$$
$$0 = \int_{\mathbf{R}} \lambda \, d\overline{\mu}_x(\lambda).$$

From the first equality, we deduce that the support of $\overline{\mu}_x$ is include in $\{-1, 1\}$. Therefore, the second equality gives $\overline{\mu}_x = \frac{1}{2}\delta_{-1} + \frac{1}{2}\delta_1$ and $\overline{\mu} = (\frac{1}{2}\delta_{-1} + \frac{1}{2}\delta_1) \otimes \mathcal{L}$.

Chapter 12

Γ-convergence and applications

12.1 Γ-convergence in abstract metrizable spaces

Given a metrizable space, or more generally a first countable topological space, we would like to define a convergence notion on the space of extended real-valued functions $F : X \longrightarrow \mathbf{R} \cup \{+\infty\}$ so that the maps

$$F \mapsto \inf_X F, \quad F \mapsto \operatorname{argmin}_X F$$

are sequentially continuous. More precisely, given $F_n, F : X \to \mathbf{R} \cup \{+\infty\}$, under some compactness hypotheses, we wish that the following implications hold true when $n \to +\infty$:

$$F_n \to F \Longrightarrow \inf_X F_n \to \min_X F;$$

$$F_n \to F,\ u_n \in \operatorname{argmin}_X F_n \Longrightarrow u_n \to u \in \operatorname{argmin}_X F \text{ at least for a subsequence.}$$

It is worth noticing that such convergence theory contains in some sense the theory of relaxation of chapter 11. Indeed, according to the relaxation theorem, Theorem 11.1.2, one has

$$F_n \equiv F \to \operatorname{cl}(F) \Longrightarrow \inf_X = \min_X \operatorname{cl}(F)$$

and every relatively compact minimizing sequence possesses a subsequence converging to $u \in \operatorname{argmin}_X F$. Therefore, constant sequences $F_n \equiv F$ must converge to $\operatorname{cl}(F)$ in the sense described above (see Remark 12.1.1).

Such an issue is of central importance in the calculus of variations. Indeed, many problems arising from physics, mechanics, economics, and approximation methods in numerical analysis are modeled by means of minimization of functionals depending on some parameter, here formally denoted by n. For instance, we write F_n for F_ε, where ε is a small parameter associated to a thickness, a stiffness in mechanics, or to a size of small discontinuities. Then, if the model associated with F_n possesses a variational formulation, the problem of finding a functional F asymptotically equivalent to F_n, formally written $F \sim F_n$, must be posed in terms of variational analysis: $F \sim F_n$ means that when n tends

to infinity (or ε tends to zero),

$$\inf_X F \sim \min_X F_n;$$

$$\overline{x} \in \operatorname{argmin}_X F \sim x_n \in \varepsilon_n\text{-}\operatorname{argmin}_X F_n,\ \varepsilon_n \to 0,$$

in the sense of some suitable topology on X.

The notion of Γ-convergence, introduced by De Giorgi and Franzoni [124] and studied in this section, corresponds to that issue.

Definition 12.1.1. *Let (X, d) be a metrizable space, or more generally a first countable topological space, $(F_n)_{n\in\mathbf{N}}$ a sequence of extended real-valued functions $F_n : X \longrightarrow \mathbf{R} \cup \{+\infty\}$, and $F : X \longrightarrow \mathbf{R} \cup \{+\infty\}$. The sequence $(F_n)_{n\in\mathbf{N}}$ (sequentially) Γ-converges to F at $x \in X$ iff both the following assertions hold:*

(i) *for all sequences $(x_n)_{n\in\mathbf{N}}$ converging to x in X, one has*

$$F(x) \leq \liminf_{n\to+\infty} F_n(x_n);$$

(ii) *there exists a sequence $(y_n)_{n\in\mathbf{N}}$ converging to x in X such that*

$$F(x) \geq \limsup_{n\to+\infty} F_n(y_n).$$

When (i) *and* (ii) *hold for all x in X, we say that $(F_n)_{n\in\mathbf{N}}$ Γ-converges to F in (X, d) and we write $F = \Gamma - \lim_{n\to+\infty} F_n$.*

Note that trivially the system of assertions (i) and (ii) is equivalent to (i) and (ii)′:

(i) for all sequences $(x_n)_{n\in\mathbf{N}}$ converging to x in X, one has

$$F(x) \leq \liminf_{n\to+\infty} F_n(x_n);$$

(ii)′ there exists a sequence $(y_n)_{n\in\mathbf{N}}$ converging to x in X such that

$$F(x) = \lim_{n\to+\infty} F_n(y_n).$$

Remark 12.1.1. Let us consider the constant sequence $(F_n)_{n\in\mathbf{N}}$ where $F_n = F : X \longrightarrow \mathbf{R} \cup \{+\infty\}$ is a given function. From chapter 11, this sequence does not Γ-converge to F but to the lsc envelope $\mathrm{cl}(F)$ of F. Consequently, the Γ-convergence is not in general associated with a topology on the family of all functions $F : X \longrightarrow \mathbf{R} \cup \{+\infty\}$. For a detailed analysis of subfamilies of functions on which the Γ-convergence is endowed by a topology, we refer the interested reader to [117].

Remark 12.1.2. Let us recall the definition of the set convergence. Let $(C_n)_{n\in\mathbf{N}}$ be a sequence of subsets of a metric space (X, d), or more generally of any topological space.

The lower limit of the sequence $(C_n)_{n\in\mathbf{N}}$ is the subset of X denoted by $\liminf_{n\to+\infty} C_n$ and defined by

$$\liminf_{n\to+\infty} C_n = \left\{x \in X : \exists x_n \to x,\ x_n \in C_n \text{ for all } n \in \mathbf{N}\right\}.$$

The upper limit of the sequence $(C_n)_{n\in\mathbf{N}}$ is the subset of X denoted by $\limsup_{n\to+\infty} C_n$ and defined by

$$\limsup_{n\to+\infty} C_n = \left\{x \in X : \exists (n_k)_{k\in\mathbf{N}},\ \exists (x_k)_{k\in\mathbf{N}},\ \forall k, x_k \in C_{n_k},\ x_k \to x\right\}.$$

The sequence $(C_n)_{n\in\mathbf{N}}$ is said to be convergent if the following equality holds:

$$\liminf_{n\to+\infty} C_n = \limsup_{n\to+\infty} C_n.$$

The common value is called the limit of $(C_n)_{n\in\mathbf{N}}$ in the Painlevé–Kuratowski sense and denoted by $\lim_{n\to+\infty} C_n$. Therefore, by definition $x \in C := \lim_{n\to+\infty} C_n$ iff the two following assertions hold:

$$\forall x \in C,\ \exists (x_n)_{n\in\mathbf{N}} \text{ such that } \forall n \in \mathbf{N}, x_n \in C_n \text{ and } x_n \to x;$$

$$\forall (n_k)_{k\in\mathbf{N}},\ \forall (x_k)_{k\in\mathbf{N}} \text{ such that } \forall k \in \mathbf{N},\ x_k \in C_{n_k},\ x_k \to x \Longrightarrow x \in C.$$

One can prove that the Γ-convergence of a sequence $(F_n)_{n\in\mathbf{N}}$ to a function F is equivalent to the convergence of the sequence of epigraphs of the functions F_n to the epigraph of F when the class of subsets of $X \times \mathbf{R}$ is equipped with the set convergence previously defined (see Attouch [28]). This is why Γ-convergence is sometimes called epiconvergence.

We define the extended real-valued functions $\Gamma-\limsup_{n\to+\infty} F_n$ and $\Gamma-\liminf_{n\to+\infty} F_n$ by setting for all $x \in X$

$$\Gamma - \limsup_{n\to+\infty} F_n(x) := \min\left\{\limsup_{n\to+\infty} F_n(x_n) : x_n \to x\right\},$$
$$\Gamma - \liminf_{n\to+\infty} F_n(x) := \min\left\{\liminf_{n\to+\infty} F_n(x_n) : x_n \to x\right\}.$$

The following proposition is a straightforward consequence of the definitions above.

Proposition 12.1.1. *Let (X, d) be a metrizable space, or more generally a first countable topological space, $(F_n)_{n\in\mathbf{N}}$ a sequence of functions $F_n : X \longrightarrow \mathbf{R} \cup \{+\infty\}$ and $F : X \longrightarrow \mathbf{R} \cup \{+\infty\}$. Then*

(i) *the sequence $(F_n)_{n\in\mathbf{N}}$ (sequentially) Γ-converges to F iff*

$$\Gamma - \limsup_{n\to+\infty} F_n \leq F \leq \Gamma - \liminf_{n\to+\infty} F_n;$$

(ii) *the functions $\Gamma-\limsup_{n\to+\infty} F_n$ and $\Gamma-\liminf_{n\to+\infty} F_n$ are lower semicontinuous.*

The main interest of the concept of Γ-convergence is its variational nature made precise in item (i) below. For more precise details about epiconvergence or Γ-convergence, we refer the reader to Attouch [28], Braides [83], and Dal Maso [117].

Theorem 12.1.1. *Let $(F_n)_{n\in\mathbf{N}}$ be a sequence of functions $F_n : X \longrightarrow \mathbf{R}\cup\{+\infty\}$ which Γ-converges to some function $F : X \longrightarrow \mathbf{R}\cup\{+\infty\}$. Then the following assertions hold:*

(i) *Let $x_n \in X$ be such that $F_n(x_n) \le \inf\{ F_n(x) : x \in X \} + \varepsilon_n$, where $\varepsilon_n > 0$, $\varepsilon_n \to 0$ when $n \to +\infty$. Assume that $\{x_n,\ n \in \mathbf{N}\}$ is relatively compact; then every cluster point $\overline{x}$ of $\{x_n : n \in \mathbf{N}\}$ is a minimizer of F and*

$$\lim_{n\to+\infty} \inf\{ F_n(x) \ : x \in X \} = F(\overline{x}).$$

(ii) *If $G : X \to \mathbf{R}$ is continuous, then $(F_n + G)_{n\in\mathbf{N}}$ Γ-converges to $F+G$.*

Let $(F_n)_{n\in\mathbf{N}}$ be a sequence of functions $F_n : X \longrightarrow \mathbf{R}\cup\{+\infty\}$. If there exists a function $F : X \longrightarrow \mathbf{R}\cup\{+\infty\}$ such that each subsequence of $(F_n)_{n\in\mathbf{N}}$ possesses a subsequence which Γ-converges to F, then all the sequence Γ-converges to F.

Proof. Assertion (ii) is easy to establish and is left to the reader. For a proof of the last assertion, consult Attouch [28, Proposition 2.72]. The proof of assertion (i) is very close to that of Theorem 11.1.2. Let $\overline{x}$ be a cluster point of $\{x_n : n \in \mathbf{N}\}$, $(x_{\sigma(n)})_{n\in\mathbf{N}}$ a subsequence of $\{x_n : n \in \mathbf{N}\}$ converging to $\overline{x}$ and set

$$\tilde{x}_m = \begin{cases} x_{\sigma(n)} \text{ if there exists } n \text{ such that } m = \sigma(n), \\ \bar{x} \text{ otherwise.} \end{cases}$$

Then $\tilde{x}_m \to \bar{x}$ when $m \to +\infty$ and, according to (i) of Definition 12.1.1, we have

$$F(\overline{x}) \le \liminf_{n\to+\infty} F_n(\tilde{x}_n) \le \liminf_{n\to+\infty} F_{\sigma(n)}(x_{\sigma(n)}) = \liminf_{n\to+\infty} \inf_X F_{\sigma(n)}. \tag{12.1}$$

Let now x be any element of X. According to (ii) of Definition 12.1.1, there exists a sequence $(y_n)_{n\in\mathbf{N}}$ converging to x and satisfying

$$F(x) \ge \limsup_{n\to+\infty} F_n(y_n) \ge \limsup_{n\to+\infty} F_{\sigma(n)}(y_{\sigma(n)}). \tag{12.2}$$

Combining (12.1) and (12.2), we obtain

$$F(\overline{x}) \le \liminf_{n\to+\infty} \inf_X F_{\sigma(n)} \le \limsup_{n\to+\infty} \inf_X F_{\sigma(n)} \le \limsup_{n\to+\infty} F_{\sigma(n)}(y_{\sigma(n)}) \le F(x). \tag{12.3}$$

This proves that $F(\overline{x}) = \min_X F$.

Taking $x = \overline{x}$ in (12.3) we also obtain $\lim_{n\to+\infty} \inf_X F_{\sigma(n)} = \min_X F$. Since all subsequence of $\inf_X F_n$ possesses a subsequence converging to $\min_X F$, one has $\lim_{n\to+\infty} \inf_X F_n = \min_X F$ as required. □

In the following sections, we give three applications of the Γ-convergence. In Section 14.2, we also show how the Γ-convergence allows us to justify some one-dimensional models in the framework of fracture mechanics.

12.2 Application to the nonlinear membrane model

Let ω be an open bounded subset of $\mathbf{R}^2$ with boundary γ and consider $\Omega_\varepsilon = \omega \times (0, \varepsilon)$, the reference configuration filled up by some elastic material. This three-dimensional thin structure is clamped on a part $\Gamma_{0,\varepsilon} = \gamma_0 \times (0, \varepsilon)$ of the boundary $\partial\Omega_\varepsilon$ of Ω_ε (see Figure 12.1).

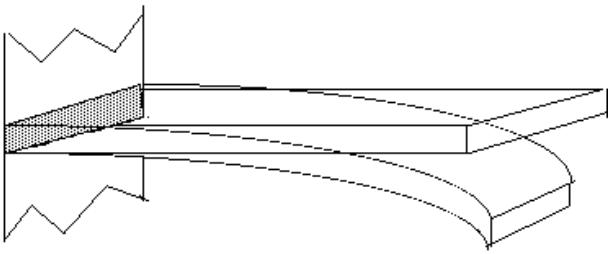


Figure 12.1. *The deformation of a thin layer Ω_ε of size ε.*

To take into account large purely elastic deformation, the constitutive law of the deformable body is associated with a nonconvex elastic density f satisfying a growth condition of order $p > 1$. The stored strain energy associated with a displacement field $u : \Omega_\varepsilon \to \mathbf{R}^3$ is given by the integral functional $F_\varepsilon : L^p(\Omega_\varepsilon, \mathbf{R}^3) \longrightarrow \mathbf{R}^+ \cup \{+\infty\}$ defined by

$$F_\varepsilon(u) = \begin{cases} \dfrac{1}{\varepsilon} \displaystyle\int_{\Omega_\varepsilon} f(\nabla u)\, dx \text{ if } u \in W^{1,p}_{\Gamma_{0,\varepsilon}}(\Omega_\varepsilon, \mathbf{R}^3), \\ +\infty \text{ otherwise,} \end{cases}$$

where the density f satisfies the conditions (11.5) and (11.6), namely, there exist three positive constants $\alpha,\ \beta,\ L$ such that

$$\forall a \in \mathbf{M}^{3\times 3}, \quad \alpha|a|^p \le f(a) \le \beta(1 + |a|^p), \tag{12.4}$$

$$\forall a,\ b \in \mathbf{M}^{3\times 3}, \quad |f(a) - f(b)| \le L|b - a|(1 + |a|^{p-1} + |b|^{p-1}). \tag{12.5}$$

The scaling parameter ε^{-1} accounts for the stiffness of the material. In the linearized elasticity framework, it corresponds to Lamé coefficients of order ε^{-1}.

The structure is subjected to applied body forces $g_\varepsilon : \Omega_\varepsilon \longrightarrow \mathbf{R}^3$ for which we make the following assumption: there exists a vector valued function $g : \Omega = \omega \times (0, 1) \longrightarrow \mathbf{R}^3$, $g \in L^q(\Omega, \mathbf{R}^3)$ $(1/p + 1/q = 1)$, such that $\varepsilon g_\varepsilon(\hat{x}, \varepsilon x_3) = g(x),\ x = (\hat{x}, x_3)$. The exterior loading is

$$L_\varepsilon(u) = \int_{\Omega_\varepsilon} g_\varepsilon . u\, dx$$

and the equilibrium configuration is given by displacement vector fields $\overline{u}_\varepsilon$, solutions of the problem

$$\inf \left\{ F_\varepsilon(u) - L_\varepsilon(u) : u \in L^p(\Omega_\varepsilon, \mathbf{R}^3) \right\}.$$

Due to the very small thickness ε of the layer Ω_ε, for computing an approximate equilibrium displacement field, it is illusory to make a direct use of the finite element

method described in chapter 7. The variational property of Γ-convergence (Theorem 12.1.1) provides a new procedure: by letting ε go to zero, we aim at finding the elastic energy of a (fictitious) material occupying the two-dimensional membrane ω. We will finally compute an approximate equilibrium displacement field for the corresponding minimization problem by means of a two-dimensional finite element method.

To work in the fixed space $L^p(\Omega, \mathbf{R}^3)$, $\Omega = \omega \times (0, 1)$, the change of scale $(\hat{x}, x_3) = (\hat{x}, \varepsilon x_3')$ transforming $(\hat{x}, x_3) \in \Omega_\varepsilon$ into $(\hat{x}, x_3') \in \Omega$ leads to the following equivalent optimization problem: find $\tilde{u}_\varepsilon$ solution of

$$\inf \left\{ \tilde{F}_\varepsilon(u) - \int_\Omega g.u \, dx : u \in L^p(\Omega, \mathbf{R}^3) \right\},$$

where

$$\tilde{F}_\varepsilon(v) = \begin{cases} \displaystyle\int_\Omega f\left(\widehat{\nabla} v, \frac{1}{\varepsilon}\frac{\partial v}{\partial x_3}\right) dx \text{ if } v \in W^{1,p}_{\Gamma_0}(\Omega, \mathbf{R}^3), \\ +\infty \text{ otherwise,} \end{cases}$$

$\Gamma_0 = \gamma_0 \times (0, 1)$, and $\widehat{\nabla} v$ denotes the tangential gradient of v (i.e., $\widehat{\nabla} v = (\frac{\partial v_i}{\partial x_j})_{i=1,2,3,\ j=1,2}$).

As suggested above, we first establish the existence of the Γ-limit of the sequence $(\tilde{F}_\varepsilon)_{\varepsilon>0}$ when ε goes to zero. To make precise its domain, we establish the following compactness result.

Proposition 12.2.1 (compactness). *Let $(u_\varepsilon)_{\varepsilon>0}$ be a sequence in $L^p(\Omega, \mathbf{R}^3)$ satisfying*

$$\sup_{\varepsilon>0} \tilde{F}_\varepsilon(u_\varepsilon) < +\infty.$$

Then, there exist a nonrelabeled subsequence and u in $V = \{v \in W^{1,p}_{\Gamma_0}(\Omega, \mathbf{R}^m) : \frac{\partial v}{\partial x_3} = 0\}$, such that u_ε converges to u, weakly in $W^{1,p}_{\Gamma_0}(\Omega, \mathbf{R}^3)$, and strongly in $L^p(\Omega, \mathbf{R}^3)$.

Proof. Since $\sup_{\varepsilon>0} \tilde{F}_\varepsilon(u_\varepsilon) < +\infty$, by using the lower bound in (12.4), we obtain

$$\tilde{F}_\varepsilon(u_\varepsilon) = \int_\Omega f\left(\widehat{\nabla} u_\varepsilon, \frac{1}{\varepsilon}\frac{\partial u_\varepsilon}{\partial x_3}\right) dx;$$

$$(\nabla u_\varepsilon)_{\varepsilon>0} \text{ is bounded in } L^p(\Omega, \mathbf{M}^{3\times 3});$$

$$\left(\frac{1}{\varepsilon}\frac{\partial u_\varepsilon}{\partial x_3}\right)_{\varepsilon>0} \text{ is bounded in } L^p(\Omega, \mathbf{R}^3).$$

Consequently, according to the Rellich–Kondrakov theorem, Theorem 5.4.2, and the Poincaré inequality, Theorem 5.3.1, there exist some $u \in W^{1,p}_{\Gamma_0}(\Omega, \mathbf{R}^3)$ and a nonrelabeled subsequence of $(u_\varepsilon)_{\varepsilon>0}$ such that

$$u_\varepsilon \rightharpoonup u \text{ weakly in } W^{1,p}_{\Gamma_0}(\Omega, \mathbf{R}^3);$$

$$u_\varepsilon \to u \text{ strongly in } L^p(\Omega, \mathbf{R}^3);$$

$$\frac{\partial u_\varepsilon}{\partial x_3} \to 0 \text{ strongly in } L^p(\Omega, \mathbf{R}^3).$$

Therefore u belongs to V. □

Note that V is canonically isomorphic to $W^{1,p}_{\gamma_0}(\omega, \mathbf{R}^3)$. In the sequel, we will use the same notation for $v \in V$ and its canonical representant in $W^{1,p}_{\gamma_0}(\omega, \mathbf{R}^3)$. The following theorem was established by Ledret-Raoult [168]. For more general and recent variational models related to thin elastic plates, see [25], [27], [77], [175], and [176]. For a variational model taking into account oscillation-concentration effects, see [169].

Theorem 12.2.1 (Le Dret and Raoult [168]). *Let us equip $L^p(\Omega, \mathbf{R}^3)$ with its strong topology. The sequence of integral functionals $(\tilde{F}_\varepsilon)_{\varepsilon>0}$ Γ-converges to the integral functional F defined in $L^p(\Omega, \mathbf{R}^3)$ by:*

$$F(u) = \begin{cases} \displaystyle\int_\omega Qf_0(\widehat{\nabla} u)\, d\hat{x} \text{ if } u \in V, \\ +\infty \text{ otherwise.} \end{cases}$$

The energy density $f_0 : \mathbf{M}^{3\times 2} \longrightarrow \mathbf{R}$ is defined for all m in $\mathbf{M}^{3\times 2}$ by

$$f_0(m) = \inf \left\{ f((m|\xi)) : \xi \in \mathbf{R}^3 \right\}$$

and Qf_0 denotes the quasi-convex envelope of f_0. We write (m/ξ) to denote the matrix $\mathbf{M}^{3\times 3}$ obtained by completing the matrix m with the column ξ.

Since the map $u \mapsto \int_\Omega g.u\, dx$ is continuous on $L^p(\Omega, \mathbf{R}^3)$, from Theorem 12.1.1 we deduce the following corollary.

Corollary 12.2.1. *The sequence of optimization problems*

$$\inf \left\{ \tilde{F}_\varepsilon(u) - \int_{\Omega_\varepsilon} g.u\, dx : u \in L^p(\Omega_\varepsilon, \mathbf{R}^3) \right\} \qquad (\mathcal{P}_\varepsilon)$$

converges to the limit problem

$$\min \left\{ F(u) - \int_\omega \overline{g}.u\, d\hat{x} : u \in V \right\}, \qquad (\mathcal{P})$$

where $\overline{g}$ is defined for all $\hat{x} \in \omega$, by $\overline{g}(\hat{x}) = \int_0^1 g(\hat{x}, s)\, ds$.

Moreover, if $\overline{u}_\varepsilon$ is a solution or an ε-minimizer of $(\mathcal{P}_\varepsilon)$, then $\tilde{u}_\varepsilon$ defined by $\tilde{u}_\varepsilon(x) = \overline{u}_\varepsilon(\hat{x}, \varepsilon x_3)$ strongly converges in $L^p(\Omega, \mathbf{R}^3)$ to a solution $\overline{u}$ of the limit problem $(\mathcal{P})$.

Roughly speaking, for very small thickness of the layer Ω_ε, an equilibrium configuration $\overline{u}_\varepsilon$ of $(\mathcal{P}_\varepsilon)$, living in $W^{1,p}_{\Gamma_{0,\varepsilon}}(\Omega_\varepsilon, \mathbf{R}^3)$, is close to $\overline{u}$, living in $W^{1,p}_{\gamma_0}(\omega, \mathbf{R}^3)$, and the layer Ω_ε may be considered as a two-dimensional membrane ω, reference configuration filled up by some elastic material whose strain energy density is the function Qf_0.

Proof of Theorem 12.2.1. For all bounded Borel sets A of $\mathbf{R}^N$, we will sometimes denote its N-dimensional Lebesgue measure by $|A|$ rather than $\mathcal{L}^N(A)$. The proof proceeds in two

steps, corresponding to each inequality in the definition of the Γ-convergence. Let us first notice that f_0 satisfies

$$\forall m \in \mathbf{M}^{3\times 2}, \quad \alpha|m|^p \le f_0(m) \le \beta(1+|m|^p), \tag{12.6}$$

$$\forall m,\ m' \in \mathbf{M}^{3\times 2}, \quad |f_0(m) - f_0(m')| \le L'|m-m'|(1+|m|^{p-1}+|m'|^{p-1}), \tag{12.7}$$

where L' is a positive constant depending only on p and L. These estimates are obtained by easy calculations from (12.4) and (12.5) and left to the reader. Consequently f_0 fulfills all conditions of Proposition 11.2.2.

First step. Let u_ε strongly converge to u in $L^p(\Omega, \mathbf{R}^3)$. We are going to establish

$$F(u) \le \liminf_{\varepsilon\to 0} \tilde{F}_\varepsilon(u_\varepsilon).$$

Obviously, one may assume $\liminf_{\varepsilon\to 0} \tilde{F}_\varepsilon(u_\varepsilon) < +\infty$ so that for a nonrelabeled subsequence,

$$\tilde{F}_\varepsilon(u_\varepsilon) = \int_\Omega f\left(\widehat{\nabla} u_\varepsilon, \frac{1}{\varepsilon}\frac{\partial u_\varepsilon}{\partial x_3}\right) dx$$

and from Proposition 12.2.1, u belongs to V. Trivially we have

$$\begin{aligned}\int_\Omega f\left(\widehat{\nabla} u_\varepsilon, \frac{1}{\varepsilon}\frac{\partial u_\varepsilon}{\partial x_3}\right) dx &\ge \int_\Omega f_0(\widehat{\nabla} u_\varepsilon)\, dx \\ &\ge \int_\Omega Qf_0(\widehat{\nabla} u_\varepsilon)\, dx. \end{aligned}\tag{12.8}$$

Let us now consider the function $h : \mathbf{M}^{3\times 3} \longrightarrow \mathbf{R}$ defined by $h(a) = Qf_0(a_1|a_2)$, where a_1 and a_2 denote the two first columns of the matrix $a = (a_1, a_2, a_3)$. We claim that

$$\forall a \in \mathbf{M}^{3\times 3}, \quad 0 \le h(a) \le \beta(1+|a|^p), \tag{12.9}$$

$$\forall a,\ b \in \mathbf{M}^{3\times 3}, \quad |h(a)-h(b)| \le L''|b-a|(1+|a|^{p-1}+|b|^{p-1}), \tag{12.10}$$

where L'' is some positive constant depending only on p and L'. We also claim that $h = Qh$ where, D denoting any bounded open subset of $\mathbf{R}^3$ with $|\partial D| = 0$, the function Qh is defined by

$$\forall a \in \mathbf{M}^{3\times 3},\ Qh(a) = \inf\left\{\frac{1}{|D|}\int_D h(a+\nabla\phi)\, dx : \phi \in W_0^{1,p}(D, \mathbf{R}^3)\right\}$$

(see Proposition 11.2.2). Inequalities (12.9) and (12.10) are straightforward consequences of inequalities (12.6) and (12.7). Let us show that $h = Qh$. Indeed, let $Y = \hat{Y} \times (0,1)$, where $\hat{Y} = (0,1)^2$, $\phi \in W_0^{1,p}(Y, \mathbf{R}^3)$, and set $\phi_y(\hat{x}) = \phi(\hat{x}, y)$ for a.e. y in $(0,1)$. Clearly ϕ_y belongs to $W_0^{1,p}(\hat{Y}, \mathbf{R}^3)$ and we have

$$\begin{aligned}\int_Y h(a+\nabla\phi)\, dx &= \int_Y Qf_0((a_1|a_2) + \widehat{\nabla}\phi)\, dx \\ &= \int_0^1 \left(\int_{\hat{Y}} Qf_0((a_1|a_2) + \nabla\phi_y(\hat{x}))\, d\hat{x}\right) dy \\ &\ge \int_0^1 Qf_0((a_1|a_2))dy = h(a),\end{aligned}$$

where we have used the quasi-convexity inequality satisfied by Qf_0 in the last inequality (see Proposition 11.2.2). Consequently,

$$h(a) = \inf\left\{\int_Y h(a+\nabla\phi)\,dx : \phi \in W_0^{1,p}(Y,\mathbf{R}^3)\right\}$$

and, according to Proposition 11.2.2, the equality $h = Qh$ follows from

$$\begin{aligned}&\inf\left\{\frac{1}{|D|}\int_D h(a+\nabla\phi)\,dx : \phi \in W_0^{1,p}(D,\mathbf{R}^3)\right\}\\ &= \inf\left\{\int_Y h(a+\nabla\phi)\,dx : \phi \in W_0^{1,p}(Y,\mathbf{R}^3)\right\}.\end{aligned}$$

Letting $\varepsilon \to 0$ in (12.8) and according to Remark 11.2.1 and Theorem 13.2.1, we obtain

$$\begin{aligned}\liminf_{\varepsilon\to 0}\int_\Omega f\left(\widehat{\nabla}u_\varepsilon, \frac{1}{\varepsilon}\frac{\partial u_\varepsilon}{\partial x_3}\right)dx &\geq \liminf_{\varepsilon\to 0}\int_\Omega Qf_0(\widehat{\nabla}u_\varepsilon)\,dx\\ &= \liminf_{\varepsilon\to 0}\int_\Omega h(\nabla u_\varepsilon)\,dx\\ &\geq \int_\Omega h(\nabla u)\,dx = \int_\Omega Qf_0(\widehat{\nabla}u)\,dx.\end{aligned}$$

Since u belongs to V, with our convention the last integral is also equal to $\int_\omega Qf_0(\nabla u)\,d\hat{x}$ and the proof of the first step is complete.

Second step. We are going to establish $(\Gamma - \limsup_{\varepsilon\to 0}\tilde{F}_\varepsilon) \leq F$. Let us assume $F(u) < +\infty$ so that $u \in W_{\gamma_0}^{1,p}(\omega,\mathbf{R}^3)$ and

$$F(u) = \int_\omega Qf_0(\nabla u(\hat{x}))\,d\hat{x}.$$

Following the proof of Lemma 11.2.2 about interchange between infimum and integral, one may easily establish

$$\int_\omega f_0(\nabla u)\,d\hat{x} = \inf_{\xi\in\mathcal{D}(\omega,\mathbf{R}^3)}\int_\omega f(\nabla u(\hat{x}),\xi(\hat{x}))\,d\hat{x}. \tag{12.11}$$

Let now ξ be some fixed element in $\mathcal{D}(\omega,\mathbf{R}^3)$ and define in $W_{\Gamma_0}^{1,p}(\Omega,\mathbf{R}^3)$ the following function:

$$w_\varepsilon(\hat{x},x_3) = u(\hat{x}) + \varepsilon x_3\xi(\hat{x}).$$

It is easy to see that w_ε strongly converges to u in $L^p(\Omega,\mathbf{R}^3)$. On the other hand, from (12.5), an easy computation gives

$$\lim_{\varepsilon\to 0}\tilde{F}_\varepsilon(w_\varepsilon) = \int_\omega f(\nabla u(\hat{x}),\xi(\hat{x}))\,d\hat{x}. \tag{12.12}$$

Now, from (12.11), taking the infimum over $\xi \in \mathcal{D}(\omega, \mathbf{R}^3)$, (12.12) gives

$$\lim_{\varepsilon \to 0} \tilde{F}_\varepsilon(w_\varepsilon) \leq \int_\omega f_0(\nabla u)\, d\hat{x}$$

so that

$$\inf \left\{ \limsup_{\varepsilon \to 0} \tilde{F}_\varepsilon(v_\varepsilon) : v_\varepsilon \to u \text{ strongly in } L^p(\Omega, \mathbf{R}^3) \right\} \leq \int_\omega f_0(\nabla u)\, d\hat{x},$$

that is to say,

$$\left(\Gamma - \limsup_{\varepsilon \to 0} \tilde{F}_\varepsilon \right)(u) \leq \tilde{F}(u), \tag{12.13}$$

where $\tilde{F}$ is the functional defined in $L^p(\Omega, \mathbf{R}^3)$ by

$$\tilde{F}(v) = \begin{cases} \displaystyle\int_\omega f_0(\nabla v(\hat{x}))\, d\hat{x} \text{ if } v \in V, \\ +\infty \text{ otherwise.} \end{cases}$$

Obviously, (12.13) also holds for any function $u \in L^p(\Omega, \mathbf{R}^3)$. Taking the lower semicontinuous envelope of each of the two members when $L^p(\Omega, \mathbf{R}^3)$ is equipped with its strong topology, and according to Proposition 12.1.1(ii) and to an easy adaptation of Corollary 11.2.1, we obtain

$$\left(\Gamma - \limsup_{\varepsilon \to 0} \tilde{F}_\varepsilon \right)(u) \leq F(u)$$

for all $u \in L^P(\Omega, \mathbf{R}^3)$.

Last step. Collecting the two previous steps gives $\Gamma - \limsup_{\varepsilon \to 0} \tilde{F}_\varepsilon \leq F \leq \Gamma - \liminf_{\varepsilon \to 0} \tilde{F}_\varepsilon$ so that $\Gamma - \lim_{\varepsilon \to 0} \tilde{F}_\varepsilon = F$. □

12.3 Application to homogenization of composite media

12.3.1 The quadratic case in one dimension

Before giving the general result concerning homogenization of composite media in Subsection 12.3.2, we establish a complete description of Γ-limits of integral functionals with quadratic density in the one-dimensional case. More precisely, given $a_\varepsilon : \mathbf{R} \to \mathbf{R}$ satisfying that there exist $\alpha > 0$ and $\beta > 0$ such that, for all $x \in \mathbf{R}$,

$$\alpha \leq a_\varepsilon \leq \beta, \tag{12.14}$$

we would like to establish the existence of a Γ-limit for the sequence of integral functionals $F_\varepsilon : L^2(0,1) \to \mathbf{R}^+ \cup \{+\infty\}$ defined by

$$F_\varepsilon(u) = \begin{cases} \displaystyle\int_{(0,1)} a_\varepsilon(x)\, u'^2(x)\, dx \text{ if } u \in H^1(0,1), \\ +\infty \text{ otherwise,} \end{cases}$$

when $L^2(0,1)$ is equipped with its strong topology.

Theorem 12.3.1. *Assume that a_ε fulfills condition* (12.14)*. Then the following assertions hold:*

(i) *If $\frac{1}{a_\varepsilon} \rightharpoonup \frac{1}{a}$ for the $\sigma(L^\infty, L^1)$ topology, then $(F_\varepsilon)_{\varepsilon>0}$ Γ-converges to the integral functional F defined on $L^2(0,1)$ by*

$$F(u) = \begin{cases} \displaystyle\int_{(0,1)} a(x)\, u'^2(x)\, dx \text{ if } u \in H^1(0,1), \\ +\infty \text{ otherwise.} \end{cases}$$

(ii) *Conversely, if $(F_\varepsilon)_{\varepsilon>0}$ Γ-converges to some functional F, then $(\frac{1}{a_\varepsilon})_{\varepsilon>0}$ $\sigma(L^\infty, L^1)$-converges to some b with $a = \frac{1}{b}$ satisfying* (12.14)*, and F has the integral representation*

$$F(u) = \begin{cases} \displaystyle\int_{(0,1)} a(x)\, u'^2(x)\, dx \text{ if } u \in H^1(0,1), \\ +\infty \text{ otherwise.} \end{cases}$$

PROOF OF (i). Let $(u_\varepsilon)_{\varepsilon>0}$ be a sequence strongly converging to some u in $L^2(0,1)$. We want to establish

$$F(u) \le \liminf_{\varepsilon\to 0} F_\varepsilon(u_\varepsilon). \tag{12.15}$$

Obviously, one may assume $\liminf_{\varepsilon\to 0} F_\varepsilon(u_\varepsilon) < +\infty$, so that, for a nonrelabeled subsequence, u_ε belongs to $H^1(0,1)$. From the equiboundedness of u'_ε in $L^2(0,1)$, we deduce that u_ε is bounded in $H^1(0,1)$ so that u_ε weakly converges to u in $H^1(0,1)$. We now take into account the quadratic expression of F_ε and write $F_\varepsilon(u_\varepsilon)$ as follows:

$$\begin{aligned} F_\varepsilon(u_\varepsilon) = \int_{(0,1)} a_\varepsilon\, u'^2_\varepsilon\, dx &= \int_{(0,1)} a_\varepsilon (u'_\varepsilon - a u'/a_\varepsilon)^2\, dx \\ &\quad + 2\int_{(0,1)} u'_\varepsilon u' a\, dx - \int_{(0,1)} u'^2 a^2/a_\varepsilon\, dx \\ &\ge 2\int_{(0,1)} u'_\varepsilon u' a\, dx - \int_{(0,1)} u'^2 a^2/a_\varepsilon\, dx. \end{aligned}$$

Letting $\varepsilon \to 0$ gives (12.15).

Given $u \in L^2(0,1)$, we now must construct a sequence $(v_\varepsilon)_{\varepsilon>0}$ strongly converging to u and satisfying

$$F(u) \ge \limsup_{\varepsilon\to 0} F_\varepsilon(v_\varepsilon). \tag{12.16}$$

One may assume $u \in H^1(0,1)$. Let us set

$$v_\varepsilon(x) = u(0) + \int_0^x a(t)u'(t)/a_\varepsilon(t)\, dt.$$

We recall that u belongs to $\mathbf{C}([0,1])$ so that the previous expression is well defined. Then $v'_\varepsilon = au'/a_\varepsilon$ weakly converges to u' in $L^2(0,1)$. Since v_ε is bounded in $L^\infty(0,1)$, we deduce

that v_ε is bounded in $H^1(0,1)$; thus, from the Rellich–Kondrakov theorem, Theorem 5.4.2, it strongly converges to some θ in $L^2(0,1)$ with $\theta' = u'$, thus $\theta = u + c$. According to the continuity of the trace, one has $\theta(0) = u(0)$, so that $\theta = u$ and $v_\varepsilon \to u$ strongly in $L^2(0,1)$. On the other hand,

$$F_\varepsilon(v_\varepsilon) = \int_{(0,1)} a^2 u'^2 / a_\varepsilon \, dx,$$

and letting $\varepsilon \to 0$ yields

$$\lim_{\varepsilon \to 0} F_\varepsilon(v_\varepsilon) = F(u),$$

hence (12.16).

Proof of (ii). From (12.14), $1/a_\varepsilon$ is bounded in $L^\infty(0,1)$. Therefore, for a nonrelabeled subsequence, it $\sigma(L^\infty, L^1)$-converges to some b with $a = \frac{1}{b}$ satisfying (12.14). According to (i), the corresponding subsequence of $(F_\varepsilon)_{\varepsilon>0}$ Γ-converges to the functional G defined by

$$G(u) = \begin{cases} \displaystyle\int_{(0,1)} a(x)\, u'^2(x)\, dx \text{ if } u \in H^1(0,1), \\ +\infty \text{ otherwise.} \end{cases}$$

Consequently $F = G$ and a is uniquely defined by F. Since all subsequence of $(1/a_\varepsilon)_{\varepsilon>0}$ possesses a subsequence which $\sigma(L^\infty, L^1)$-converges to the same limit $1/a$, all the sequence $(1/a_\varepsilon)_{\varepsilon>0}$ $\sigma(L^\infty, L^1)$-converges to $1/a$. □

Example 12.3.1. Let us consider a_ε defined by $a_\varepsilon(x) = a(x/\varepsilon)$, where a is a $(0,1)$-periodic function taking two positive values α and β on $(0, 1/2)$ and $(1/2, 1)$, respectively. Then $1/a_\varepsilon$ weakly converges for the $\sigma(L^\infty, L^1)$ topology to $(\alpha^{-1} + \beta^{-1})/2$. For a proof, see Example 2.4.2 or Proposition 13.2.1 and the proof of Theorem 13.2.1. Therefore

$$F(u) = \begin{cases} \displaystyle\int_{(0,1)} \left(\frac{\alpha^{-1} + \beta^{-1}}{2} \right)^{-1} u'^2(x)\, dx \text{ if } u \in H^1(0,1), \\ +\infty \text{ otherwise.} \end{cases}$$

The functional F_ε may be interpreted for example as the elastic energy of a system of two kinds of small periodically distributed springs with size ε. Theorem 12.3.1 shows that the mechanical behavior of such a system is equivalent to a homogeneous string in the sense of Γ-convergence. The equivalent density is associated with the strain tensor $\left(\frac{\alpha^{-1}+\beta^{-1}}{2}\right)^{-1}$ strictly smaller than the mean value $\frac{\alpha+\beta}{2}$ of the tensors associated with the two kinds of material.

Remark 12.3.1. The same problem in the two-dimensional case, describing, for example, a system of two kinds of small elastic pieces in $\Omega = (0,1)^2$ in a chessboard structure, may be treated as above by using the concept of Γ-convergence. One can show that the equivalent density is quadratic and associated with the strain tensor $\sqrt{\alpha\beta}$. In the three-dimensional case, there is no explicit formula for the strain tensor limit. More generally, when working with general quadratic densities of the form $f_\varepsilon(\xi) = \langle A_\varepsilon \xi, \xi \rangle$, $A_\varepsilon \in \mathcal{M}^{3\times 3}$ satisfying, for all $\xi \in \mathbf{R}^3$,

$$\alpha|\xi|^2 \le \langle A_\varepsilon \xi, \xi \rangle \le \beta|\xi|^2,$$

one can show that the Γ-limit of the associated integral functional possesses a density of the form $\langle A\xi, \xi\rangle$. The strategy is then to derive optimal bounds for the constant limit matrix A (see Murat and Tartar [192]).

12.3.2 Periodic homogenization in the general case

Let Ω be an open bounded subset of $\mathbf{R}^3$ which represents the interior of the reference configuration filled up by some elastic ($p > 1$) or pseudoplastic ($p = 1$) material which is clamped on a part Γ_0 of the boundary $\partial\Omega$ of Ω. We assume that this material is heterogeneous with a periodic distribution of small heterogeneities of size of order $\varepsilon > 0$, so that the stored strain energy density is of the form

$$(x, a) \mapsto f\left(\frac{x}{\varepsilon}, a\right),$$

where $f(., a)$ is Y-periodic, $Y = (0, 1)^3$. We assume that f satisfies conditions (12.4) and (12.5) of the previous section and, to take into account large purely elastic deformations, f is not assumed to be convex but possibly quasiconvex. With the notations of Sections 11.2 and 11.3, the stored strain energy associated with a displacement field $u : \Omega \to \mathbf{R}^3$ is given by the integral functional $F_\varepsilon : L^p(\Omega, \mathbf{R}^3) \longrightarrow \mathbf{R}^+ \cup \{+\infty\}$ defined by

$$F_\varepsilon(u) = \begin{cases} \displaystyle\int_\Omega f\left(\frac{x}{\varepsilon}, \nabla u\right) dx \text{ if } u \in W^{1,p}_{\Gamma_0}(\Omega, \mathbf{R}^3), \\ +\infty \text{ otherwise.} \end{cases}$$

The structure is subjected to applied body forces $g : \Omega \longrightarrow \mathbf{R}^3$, $g \in L^q(\Omega, \mathbf{R}^3)$, ($1/p + 1/q = 1$ if $p > 1$; $q = +\infty$ if $p = 1$) and the exterior loading is defined by

$$L(u) = \int_\Omega g.u \, dx.$$

The equilibrium configuration is then given by the displacement field $\overline{u}_\varepsilon$ solution of the problem

$$\inf\left\{F_\varepsilon(u) - \int_\Omega L(u) : u \in L^p(\Omega, \mathbf{R}^3)\right\}.$$

Due to the very small size ε of heterogeneousness, for computing an approximate equilibrium displacement field, it is illusory to make direct use of the finite element method. The variational property of Γ-convergence (Theorem 12.1.1) would again provide a new procedure: to find a fictitious material occupying Ω, which appears to be homogeneous when ε goes to zero, and to compute the approximate equilibrium displacement field by means of a finite element method related to a discretization of the new model.

To treat more general situations, we deal with functionals $F_\varepsilon : L^p(\Omega, \mathbf{R}^m) \longrightarrow \mathbf{R}^+ \cup \{+\infty\}$ defined by

$$F_\varepsilon(u) = \begin{cases} \displaystyle\int_\Omega f\left(\frac{x}{\varepsilon}, \nabla u\right) dx \text{ if } u \in W^{1,p}_{\Gamma_0}(\Omega, \mathbf{R}^m), \\ +\infty \text{ otherwise,} \end{cases}$$

where Ω is an open bounded subset of $\mathbf{R}^N$, m is any positive integer, f satisfies the growth conditions (12.4) and (12.5), and, for all a in the set $\mathbf{M}^{m\times N}$ of $m \times N$-matrices, the Borel function $f(.,a)$ is assumed to be Y-periodic, $Y = (0,1)^N$.

Following the strategy of the previous subsection, we are going to establish the Γ-convergence of the sequence $(F_\varepsilon)_{\varepsilon>0}$ when $L^p(\Omega,\mathbf{R}^m)$ is equipped with its strong topology. In Theorem 12.3.2, we will establish that the Γ-limit of F_ε possesses an integral representation. In the following proposition, we characterize its density ($p > 1$) or its regular part ($p = 1$). It is worth noticing the similarity between this proposition and Proposition 11.2.2, where we defined the relaxed density of the integral functional on Sobolev or BV spaces.

Proposition 12.3.1. *For all open bounded convex set A in $\mathbf{R}^N$ the following limit exists:*

$$f^{hom}(a) = \lim_{\varepsilon\to 0}\left[\inf\left\{\frac{1}{|A/\varepsilon|}\int_{A/\varepsilon} f(x,a+\nabla u(x))\,dx : u \in W_0^{1,p}(A/\varepsilon,\mathbf{R}^m)\right\}\right].$$

Moreover, this limit does not depend on the choice of the open bounded convex set A and is equal to

$$\inf_{n\in\mathbf{N}^*}\inf\left\{\frac{1}{n^N}\int_{nY} f(y,a+\nabla u(y))\,dy : u \in W_0^{1,p}(Y,\mathbf{R}^m)\right\}.$$

The proof is based on a convergence result related to subadditive processes. To go further, we first give some notations and definitions. Let us denote the family of all the bounded Borel sets of $\mathbf{R}^N$ by $\mathcal{B}_b(\mathbf{R}^N)$.

A sequence $(B_n)_{n\in\mathbf{N}}$ of sets of $\mathcal{B}_b(\mathbf{R}^N)$ is said to be regular if there exists an increasing sequence of intervals I_n with vertices in $\mathbf{Z}^N$ and a positive constant C independent of n such that $B_n \subset I_n$ and $|I_n| \le C|B_n|$ for all $n \in \mathbf{N}$.

A subadditive $\mathbf{Z}^N$-invariant set function indexed on $\mathcal{B}_b(\mathbf{R}^N)$ is a map $\mathcal{S} : \mathcal{B}_b(\mathbf{R}^N) \longrightarrow \mathbf{R},\ A \mapsto \mathcal{S}_A$, such that

(i) $\forall A,\ B \in \mathcal{B}_b(\mathbf{R}^N)$ with $A \cap B = \emptyset$, $\mathcal{S}_{A\cup B} \le \mathcal{S}_A + \mathcal{S}_B$;

(ii) $\forall A \in \mathcal{B}_b(\mathbf{R}^N),\ \forall z \in \mathbf{Z}^N$, $\mathcal{S}_{z+A} = \mathcal{S}_A$.

Finally, for all A in $\mathcal{B}_b(\mathbf{R}^N)$, we define the positive number $\rho(A) := \sup\{r \ge 0 : \exists \bar{B}_r(x) \subset A\}$, where $\bar{B}_r(x)$ is the closed ball with radius $r > 0$ centered at x. The following lemma generalizes the classical limit theorem related to subadditive processes indexed by cubes. For a proof, we refer the reader to Ackoglu and Krengel [3], to Licht and Michaille [170], or to Dal Maso and Modica [119].

Lemma 12.3.1. *Let $\mathcal{S}$ be a subadditive $\mathbf{Z}^N$-invariant set function such that*

$$\gamma(\mathcal{S}) := \inf\left\{\frac{\mathcal{S}_I}{|I|} : I = [a,b[,\ a = (a_i)_{i=1\ldots4},\ b = (b_i)_{i=1\ldots4} \in \mathbf{Z}^N,\right.$$
$$\left.\forall i = 1,\ldots,N,\ a_i < b_i\right\} > -\infty$$

and which satisfies the following domination property: there exists a positive constant $C(\mathcal{S}) < +\infty$ such that $|\mathcal{S}_A| \le C(\mathcal{S})$ for all Borel sets A included in $[0,1[^N$. Let $(A_n)_{n\in\mathbf{N}}$

be a regular sequence of Borel convex sets of $\mathcal{B}_b(\mathbf{R}^N)$ *satisfying* $\lim_{n\to+\infty}\rho(A_n)=+\infty$. *Then*

$$\lim_{n\to+\infty}\frac{\mathcal{S}_{A_n}}{|A_n|}=\inf_{m\in\mathbf{N}^*}\left\{\frac{\mathcal{S}_{[0,m[^N}}{m^N}\right\}=\gamma(\mathcal{S}).$$

Proof of Proposition 12.3.1. Let us notice that in the definition of subadditivity, assertion (i) may be replaced by the following: for all $A,\ B\in\mathcal{B}_b(\mathbf{R}^N)$ with $A\cap B=\emptyset$ and $|\partial A|=|\partial B|=0,\ \mathcal{S}_{A\cup B}\le\mathcal{S}_A+\mathcal{S}_B$ (see [170, Remark of Theorem 2.1] or [119]). Then we claim that

$$\mathcal{S}:A\mapsto\inf\left\{\int_{\overset{0}{A}}f(x,a+\nabla u(x))\,dx:u\in W_0^{1,p}(\overset{0}{A},\mathbf{R}^m)\right\}$$

is a subadditive $\mathbf{Z}^N$-invariant process. Indeed, Y-periodicity of $f(.,a)$ yields $\mathcal{S}_{A+z}=\mathcal{S}_A$ for all $A\in\mathcal{B}_b(\mathbf{R}^N)$ and all $z\in\mathbf{Z}^N$. Let now A, B in $\mathcal{B}_b(\mathbf{R}^N)$ such that $A\cap B=\emptyset$ and $|\partial A|=|\partial B|=0$. For arbitrary fixed $\eta>0$, consider $\varphi_A\in\mathcal{D}(\overset{0}{A},\mathbf{R}^m)$ and $\varphi_B\in\mathcal{D}(\overset{0}{B},\mathbf{R}^m)$ satisfying

$$\int_{\overset{0}{A}}f(x,a+\nabla\varphi_A)\,dx\le\mathcal{S}_A+\eta,$$

$$\int_{\overset{0}{B}}f(x,a+\nabla\varphi_B)\,dx\le\mathcal{S}_B+\eta.$$

Extending φ_A and φ_B by zero, respectively, on $\mathbf{R}^N\setminus\overset{0}{A}$ and $\mathbf{R}^N\setminus\overset{0}{B}$, the function φ which coincides with φ_A on $\overset{0}{A}$ and φ_B in $\overset{0}{B}$ belongs to $W_0^{1,p}(\overset{0}{A\cup B},\mathbf{R}^m)$. Since $|\partial A|=|\partial B|=0$, $|\overset{0}{A\cup B}\setminus\overset{0}{A}\cup\overset{0}{B}|=0$, thus

$$\begin{aligned}\int_{\overset{0}{A\cup B}}f(x,a+\nabla\varphi)\,dx&=\int_{\overset{0}{A}}f(x,a+\nabla\varphi_A)\,dx+\int_{\overset{0}{B}}f(x,a+\nabla\varphi_B)\,dx\\&\quad+\int_{\overset{0}{A\cup B}\setminus\overset{0}{A}\cup\overset{0}{B}}f(x,a)\,dx\\&=\int_{\overset{0}{A}}f(x,a+\nabla\varphi_A)\,dx+\int_{\overset{0}{B}}f(x,a+\nabla\varphi_B)\,dx\\&\le\mathcal{S}_A+\mathcal{S}_B+2\eta.\end{aligned}$$

Consequently, $\mathcal{S}_{A\cup B}\le\mathcal{S}_A+\mathcal{S}_B+2\eta$ and the subadditivity of $\mathcal{S}$ follows by letting $\eta\to0$. We conclude the proof by applying Lemma 12.3.1 to this process. □

Remark 12.3.2. Lemma 12.3.1 may be generalized for ergodic subadditive processes, i.e., for subadditive processes with value in $L^1(\Sigma,\mathcal{T},\mathbf{P})$ and whose probability law is, roughly speaking, invariant under a group $(T_z)_{z\in\mathbf{Z}^N}$ of measure preserving transformations on the probability space $(\Sigma,\mathcal{T},\mathbf{P})$. This probabilistic version allows to treat stochastic homogenization (see, for instance, [119], [170], or Section 14.2).

We can now establish the main convergence result, a generalization of Theorem 11.2.1 and Theorem 11.3.1. In what follows, ε actually denotes a sequence $(\varepsilon_n)_{n\in\mathbf{N}}$ of positive numbers ε_n going to zero when $n \to +\infty$ and we adopt the notation of Section 11.2. For more general problems involving multiple small parameters, see [10], [11]. For problems concerned with nonlocal effects, see [62], [63].

Theorem 12.3.2. *Let f satisfying* (12.4) *and* (12.5) *with $p \geq 1$ and assume that the Borel function $f(.,a)$ is Y-periodic for all a in $\mathbf{M}^{m\times N}$. Let us consider the integral functional F_ε defined in $L^p(\Omega,\mathbf{R}^m)$ by*

$$F_\varepsilon(u) = \begin{cases} \displaystyle\int_\Omega f\left(\frac{x}{\varepsilon}, \nabla u\right)\, dx \text{ if } u \in W^{1,p}_{\Gamma_0}(\Omega,\mathbf{R}^m), \\ +\infty \text{ otherwise,} \end{cases}$$

where $L^p(\Omega,\mathbf{R}^m)$ is equipped with its strong topology. Then $(F_\varepsilon)_{\varepsilon>0}$ Γ-converges to the integral functional F^{hom} defined by

(i) *case $p > 1$*

$$F^{hom}(u) = \begin{cases} \displaystyle\int_\Omega f^{hom}(\nabla u)\, dx \text{ if } u \in W^{1,p}_{\Gamma_0}(\Omega,\mathbf{R}^m), \\ +\infty \text{ otherwise;} \end{cases}$$

(ii) *case $p = 1$*

$$F^{hom}(u) = \begin{cases} \displaystyle\int_\Omega f^{hom}(\nabla u)\, dx + \int_\Omega (f)^{hom,\infty}\left(\frac{D^s u}{|D^s u|}\right)|D^s u| \\ \qquad\qquad + \displaystyle\int_{\Gamma_0} (f)^{hom,\infty}(\gamma_0(u)\otimes \nu)\, d\mathcal{H}^{N-1} \text{ if } u \in BV(\Omega,\mathbf{R}^m), \\ +\infty \text{ otherwise,} \end{cases}$$

where ν denotes the outer unit normal to Γ_0, γ_0 the trace operator, and $(f)^{hom,\infty}$ the recession function of f^{hom} defined by

$$(f)^{hom,\infty}(a) = \limsup_{t\to+\infty} \frac{(f)^{hom}(ta)}{t}.$$

The proof of Theorem 12.3.2 is the consequence of Propositions 12.3.2 and 12.3.3. To shorten the proofs we do not take into account the boundary condition, i.e., the domain of F_ε is $W^{1,p}(\Omega,\mathbf{R}^m)$. For treating the general case, it suffices to reproduce exactly the proofs of Corollary 11.2.1 when $p > 1$ and Corollary 11.3.1 when $p = 1$.

Proposition 12.3.2. *For all u in $L^p(\Omega,\mathbf{R}^m)$ and all sequence $(u_n)_{n\in\mathbf{N}}$ strongly converging to u in $L^p(\Omega,\mathbf{R}^m)$, one has*

$$F^{hom}(u) \leq \liminf_{n\to+\infty} F_{\varepsilon_n}(u_n). \tag{12.17}$$

PROOF. Our strategy is exactly the one of Proposition 11.2.3 or 11.3.3. Obviously, one may assume $\liminf_{n\to+\infty} F_{\varepsilon_n}(u_n) < +\infty$. For a nonrelabeled subsequence, consider the nonnegative Borel measure $\mu_n := f(\frac{\cdot}{\varepsilon_n}, \nabla u_n(.))\mathcal{L}\lfloor\Omega$; we have

$$\sup_{n\in\mathbf{N}} \mu_n(\Omega) < +\infty.$$

Consequently, there exists a further subsequence (not relabeled) and a nonnegative Borel measure $\mu \in \mathbf{M}(\Omega)$ such that

$$\mu_n \rightharpoonup \mu \text{ weakly in } \mathbf{M}(\Omega).$$

Let $\mu = g\mathcal{L}^N\lfloor\Omega + \mu^s$ be the Lebesgue–Nikodym decomposition of μ, where μ^s is a nonnegative Borel measure, singular with respect to the N-dimensional Lebesgue measure $\mathcal{L}\lfloor\Omega$ restricted to Ω. For establishing (12.17) it is enough to prove that

$$g(x) \geq f^{hom}(\nabla u(x)) \quad x \text{ a.e.}, \tag{12.18}$$

$$\mu^s \geq f^{hom,\infty}\left(\frac{D^s u}{|D^s u|}\right) |D^s u| \text{ when } p = 1. \tag{12.19}$$

(a) *Proof of* (12.18). Let $\rho > 0$ intended to tend to 0 and let $B_\rho(x_0)$ be the open ball of radius ρ centered at x_0. According to the theory of differentiation of measures, for a.e. $x_0 \in \Omega$

$$g(x_0) = \lim_{\rho\to 0} \frac{\mu(B_\rho(x_0))}{|B_\rho(x_0)|}.$$

Applying Lemma 4.2.1, one may assume $\mu(\partial B_\rho(x_0)) = 0$ for all but countably many $\rho > 0$, so that, from Alexandrov's theorem, Proposition 4.2.3, we have $\mu(B_\rho(x_0) = \lim_{n\to+\infty} \mu_n$ $(B_\rho(x_0))$ and we finally must establish

$$\lim_{\rho\to 0}\lim_{n\to+\infty} \frac{\mu_n(B_\rho(x_0))}{|B_\rho(x_0)|} \geq f^{hom}(\nabla u(x_0)) \quad \text{for } a.e.\ x_0 \in \Omega. \tag{12.20}$$

Let us assume for the moment that the trace of u_n on $\partial B_\rho(x_0)$ coincides with the affine function u_0 defined by $u_0(x) := u(x_0) + \langle \nabla u(x_0), x - x_0\rangle$. It follows from Proposition 12.3.1 that

$$\begin{aligned}
&\lim_{n\to+\infty} \frac{\mu_n(B_\rho(x_0))}{|B_\rho(x_0)|} \\
&= \lim_{n\to+\infty} \frac{1}{|B_\rho(x_0)|}\int_{B_\rho(x_0)} f\left(\frac{x}{\varepsilon_n}, \nabla u(x_0) + \nabla(u_n - u_0)\right) dx \\
&\geq \limsup_{n\to+\infty} \inf\left\{\frac{1}{|B_\rho(x_0)|}\int_{B_\rho(x_0)} f\left(\frac{x}{\varepsilon_n}, \nabla u(x_0) + \nabla\phi\right) dx : \phi \in W_0^{1,p}(B_\rho(x_0), \mathbf{R}^m)\right\} \\
&= \lim_{n\to+\infty} \inf\left\{\frac{1}{|\frac{1}{\varepsilon_n}B_\rho(x_0)|}\int_{\frac{1}{\varepsilon_n}B_\rho(x_0)} f(x, \nabla u(x_0) + \nabla\phi)\, dx : \phi \in W_0^{1,p}\left(\frac{1}{\varepsilon_n}B_\rho(x_0), \mathbf{R}^m\right)\right\} \\
&= f^{hom}(\nabla u(x_0)),
\end{aligned}$$

and the proof would be complete. The idea now consists in modifying u_n into a function of $W^{1,p}(B_\rho(x_0), \mathbf{R}^m)$ which coincides with u_0 on $\partial B_\rho(x_0)$ in the trace sense, to follow the

previous procedure and to control the additional terms, when ρ goes to zero, thanks to the estimate (see Lemma 11.2.1 and Proposition 10.4.1): for a.e. $x \in \Omega$,

$$\left[\frac{1}{|B_\rho(x_0)|}\int_{B_\rho(x_0)} |u(x) - \big(u(x_0) + \nabla u(x_0)(x - x_0)\big)|^p \, dx\right]^{1/p} = o(\rho).$$

The suitable modification of u_n is exactly the one of Proposition 11.2.3 because of the conditions (12.4) and (12.5) satisfied by f. The proof of (12.20) is then complete.

(b) *Proof of* (12.19). It suffices to reproduce the proof of inequality

$$\mu^s \geq (Qf)^\infty\left(\frac{D^s u}{|D^s u|}\right)|D^s u|$$

obtained in the proof of Proposition 11.3.3 after substituting f by $f(\frac{x}{\varepsilon_n}, .)$ and, according to Proposition 12.3.1, after substituting Qf by f^{hom}. □

Proposition 12.3.3. *For all u in $L^p(\Omega, \mathbf{R}^m)$, $p \geq 1$, there exists a sequence $(u_n)_{n\in\mathbf{N}}$ strongly converging to u in $L^p(\Omega, \mathbf{R}^m)$ such that*

$$F^{hom}(u) \geq \limsup_{n\to+\infty} F_{\varepsilon_n}(u_n).$$

Proof. The proof will be obtained in two steps.

First step. We assume $u \in W^{1,p}(\Omega, \mathbf{R}^m)$. We reproduce, with minor modifications, the outline of the proof of Proposition 11.2.4. According to Proposition 12.3.1 and to the Lebesgue dominated convergence theorem,

$$F^{hom}(u) = \int_\Omega f^{hom}(\nabla u)\, dx = \lim_{k\to+\infty}\int_\Omega f_k^{hom}(\nabla u)\, dx, \tag{12.21}$$

where

$$f_k^{hom}(a) = \inf\left\{\frac{1}{|kY|}\int_{kY} f(y, a + \nabla v)\, dy : v \in W_0^{1,p}(kY, \mathbf{R}^m)\right\}.$$

Let us fix $k \in \mathbf{N}^*$. Applying the interchange lemma, Lemma 11.2.2, we have for all $\eta > 0$ (of the form $1/h$ with h integer) and for some $\phi_{k,\eta}$ in $\mathbf{C}_c(\Omega, \mathcal{D}(kY, \mathbf{R}^m))$,

$$\begin{aligned}\frac{1}{|kY|}\int_{\Omega\times kY} f(y, \nabla u(x) + \nabla_y\phi_{k,\eta}(x, y))\, dxdy &\geq \int_\Omega f_k^{hom}(\nabla u)\, dx \\ > \frac{1}{|kY|}\int_{\Omega\times kY} f(y, \nabla u(x) + \nabla_y\phi_{k,\eta}(x, y))\, dxdy &- \eta.\end{aligned} \tag{12.22}$$

Let us extend $y \mapsto \phi_{k,\eta}(x, y)$ by kY-periodicity on $\mathbf{R}^N$ and consider the function $u_{k,\eta,n}$ defined by

$$u_{k,\eta,n}(x) = u(x) + \varepsilon_n\phi_{k,\eta}\left(x, \frac{x}{\varepsilon_n}\right).$$

Note that $\phi_{k,\eta}$ is a Carathéodory function so that $x \mapsto \phi_{k,\eta}(x, \frac{x}{\varepsilon_n})$ is measurable. Clearly $u_{k,\eta,n}$ belongs to $W^{1,p}(\Omega, \mathbf{R}^m)$ and

$$u_{k,\eta,n} \to u \text{ strongly in } L^p(\Omega, \mathbf{R}^m)$$

when n goes to ∞. On the other hand, according to the continuity assumption (12.5) on f and to Lemma 11.2.3,

$$\begin{aligned}
&\lim_{n\to+\infty}\int_\Omega f\left(\frac{x}{\varepsilon_n},\nabla u_{k,\eta,n}\right)dx\\
&=\lim_{n\to+\infty}\int_\Omega f\left(\frac{x}{\varepsilon_n},\nabla u(x)+\nabla_y\phi_{k,\eta}\left(x,\frac{x}{\varepsilon_n}\right)+\varepsilon_n\nabla\phi_{k,\eta}\left(x,\frac{x}{\varepsilon_n}\right)\right)dx\\
&=\lim_{n\to+\infty}\int_\Omega f\left(\frac{x}{\varepsilon_n},\nabla u(x)+(\nabla_y\phi_{k,\eta})\left(x,\frac{x}{\varepsilon_n}\right)\right)dx\\
&=\frac{1}{|kY|}\int_{\Omega\times kY} f(y,\nabla u(x)+\nabla_y\phi_{k,\eta}(x,y))\,dxdy.
\end{aligned}$$

Consequently, from (12.22)

$$\lim_{n\to+\infty}F_{\varepsilon_n}(u_{k,\eta,n})=\frac{1}{|kY|}\int_{\Omega\times kY}f(y,\nabla u(x)+\nabla_y\phi_{k,\eta}(x,y))\,dxdy\le\int_\Omega f_k^{hom}(\nabla u)\,dx+\eta.$$

The inequality above and (12.21), letting $\eta\to 0$ (i.e., $h\to+\infty$) and $k\to+\infty$, yield

$$\lim_{k\to+\infty}\lim_{\eta\to+0}\lim_{n\to+\infty}F_{\varepsilon_n}(u_{k,\eta,n})=F^{hom}(u).$$

Let us now apply the diagonalization Lemma 11.1.1 to the sequence $\left(F_{\varepsilon_n}(u_{k,\eta,n}),u_{k,\eta,n}\right)_{k,\eta,n}$ in the metric space $\mathbf{R}\times L^p(\Omega,\mathbf{R}^m)$: there exists a map $n\mapsto(k,\eta)(n)$ such that

$$\begin{aligned}
&\lim_{n\to+\infty}F_{\varepsilon_n}(u_{(k,\eta)(n),n})=F^{hom}(u),\\
&\lim_{n\to+\infty}u_{(k,\eta)(n),n}=u\text{ strongly in }L^p(\Omega,\mathbf{R}^m).
\end{aligned}$$

We have proved that $F^{hom}(u)\ge\limsup_{n\to+\infty}F_{\varepsilon_n}(u_n)$ for $u_n=u_{(k,\eta)(n),n}$ converging to u in $W^{1,p}(\Omega,\mathbf{R}^m)$ equipped with the strong convergence of $L^p(\Omega,\mathbf{R}^m)$. If $p>1$, the proof is complete because the domain of F^{hom} is $W^{1,p}(\Omega,\mathbf{R}^m)$.

Second step ($p=1$). Let us consider the functional G defined on $L^1(\Omega,\mathbf{R}^m)$ by

$$G(u)=\begin{cases}\displaystyle\int_\Omega f^{hom}(\nabla u)\,dx\text{ if }u\in W^{1,1}(\Omega,\mathbf{R}^m),\\ +\infty\text{ otherwise.}\end{cases}$$

According to Theorem 11.3.1, F^{hom} is nothing but the lower semicontinuous envelope $\mathrm{cl}(G)$ of G. Therefore, for all $u\in L^1(\Omega,\mathbf{R}^m)$, there exists a sequence $(u_l)_{l\in\mathbf{N}}$ in $L^1(\Omega,\mathbf{R}^m)$, strongly converging to u in $L^1(\Omega,\mathbf{R}^m)$ such that

$$F^{hom}(u)=\lim_{l\to+\infty}G(u_l).$$

One may assume $F^{hom}(u)<+\infty$ so that according to the first step, there exists a sequence $(u_{l,n})_{n\in\mathbf{N}}$ strongly converging to u_l in $L^1(\Omega,\mathbf{R}^m)$ when $n\to+\infty$, such that

$$G(u_l)=\lim_{n\to+\infty}F_{\varepsilon_n}(u_{l,n}).$$

Combining these two equalities, we obtain

$$F^{hom}(u) = \limsup_{l\to+\infty} \limsup_{n\to+\infty} F_{\varepsilon_n}(u_{l,n})$$

and

$$\lim_{l\to+\infty} \lim_{l\to+\infty} u_{l,n} = u \text{ strongly in } L^1(\Omega, \mathbf{R}^m).$$

We end the proof by applying the diagonalization Lemma 11.1.1. □

12.4 Application to image segmentation and phase transitions

12.4.1 The Mumford–Shah model

Let Ω be a bounded open subset of $\mathbf{R}^N$ and g a given function in $L^\infty(\Omega)$. Denoting by $\mathcal{F}$ the class of the closed sets of Ω, for all K in $\mathcal{F}$ and all u in $\mathbf{C}^1(\Omega \setminus K)$ we deal with the functional

$$E(u, K) := \int_\Omega |u - g|^2 \, dx + \int_{\Omega\setminus K} |\nabla u|^2 \, dx + \mathcal{H}^{N-1}(K)$$

and the associated optimization problem

$$\inf\{E(u, K) : (u, K) \in \mathbf{C}^1(\Omega \setminus K) \times \mathcal{F}\}. \tag{12.23}$$

When Ω is a rectangle in $\mathbf{R}^2$ and $g(x)$ is the light signal striking Ω at a point x, (12.23) is the Mumford–Shah model of image segmentation: K may be considered as the outline of the given light image in computer vision. If it exists, a solution (u^*, K^*) of (12.23) fulfills the three following properties:

(i) the first term in $E(u, K)$ asks that u^* approximates the light signal g in $L^2(\Omega)$;

(ii) in $\Omega \setminus K^*$, u^* does not vary very much (because of the term $\int_{\Omega\setminus K} |\nabla u^*|^2 \, dx$);

(iii) the third term asks that the boundaries K^* be as short as possible.

Let us remark that dropping one of the three terms makes the problem trivial, i.e.,

$$\inf\{E(u, K) \ : (u, K) \in \mathbf{C}^1(\Omega \setminus K) \times \mathcal{F}\} = 0.$$

Indeed, when $E(u, K) = \int_{\Omega\setminus K} |\nabla u|^2 \, dx + \mathcal{H}^{N-1}(K)$, take $u^* = 0$ and $K^* = \emptyset$.

When $E(u, K) := \int_\Omega |u - g|^2 \, dx + \mathcal{H}^{N-1}(K)$, take $u^* = g$ and $K^* = \emptyset$.

When $E(u, K) := \int_\Omega |u - g|^2 \, dx + \int_{\Omega\setminus K} |\nabla u|^2 \, dx$, let us decompose Ω by a finite union of open cubes $Q_{i,\eta}$ with diameter η and boundary $K_{i,\eta}$:

$$\mathcal{L}^N\left(\Omega \setminus \bigcup_{i\in I(\eta)} Q_{i,\eta}\right) = 0, \ K_\eta = \bigcup_{i\in I(\eta)} K_{i,\eta}$$

and set

$$u_{i,\eta} := \frac{1}{|Q_{i,\eta}|} \int_{Q_{i,\eta}} g(x) \, dx, \ u_\eta := \sum_{i\in I(\eta)} u_{i,\eta} 1_{Q_{i,\eta}}.$$

Then $E(u_\eta, K_\eta)$ tends to zero when η goes to zero. In this last case (14.14) has obviously no solution if g is not constant.

To fit several applications to computer vision problems, one can adjust the functional E by suitable positive constants α, β, and γ and consider

$$E(u, K) := \alpha \int_\Omega |u - g|^2 \, dx + \beta \int_{\Omega \setminus K} |\nabla u|^2 \, dx + \gamma \mathcal{H}^{N-1}(K).$$

In what follows, to shorten notations, we set $\alpha = \beta = \gamma = 1$.

The existence of a solution for the optimization problem (12.23) was conjectured in [191] and has been established in [123] by using the semicontinuity and the compactness results of Ambrosio [13] related to functionals defined in SBV spaces (see chapter 13). They defined a weak formulation of (12.23) as follows. If (12.23) has a solution (u^*, K^*), the closed set K^* must contain the jump set of u^*. Then, it is natural to solve the problem in $SBV(\Omega)$ and to consider K^* as the closure of the set S_{u*}. That leads us to consider the following weak formulation of the problem (12.23):

$$\inf \left\{ \int_\Omega |\nabla u|^2 \, dx + \mathcal{H}^{N-1}(S_u) + \int_\Omega |u - g|^2 \, dx : u \in SBV(\Omega) \right\}, \tag{12.24}$$

where ∇u denotes the density of the Lebesgue part of Du and S_u the jump set of u (see Section 10). The functional E defined on $SBV(\Omega)$ by $E(u) = \int_\Omega |\nabla u|^2 \, dx + \mathcal{H}^{N-2}(S_u) + \int_\Omega |u - g|^2 \, dx$ will be referred to as the Mumford–Shah energy functional. In Section 14.3 we will establish the following existence result.

Theorem 12.4.1. *There exists at least a solution of the weak problem* (12.24).

12.4.2 Variational approximation of a more elementary problem: A phase transitions model

To describe a numerical processing of the weak formulation (problem (12.24)), a natural way consists in approximating in a variational sense the Mumford–Shah energy functional by classical integral functionals. Before treating the Mumford–Shah energy, we begin by showing how the Van Der Waals–Cahn–Hiliard thermodynamical model of phase transitions allows us to define a good approximation of the term $\mathcal{H}^{N-1}(S_u)$. For another and more direct method in one or two dimensions (i.e., $N = 1, \ 2$), we refer the reader to Chambolle [107]. For nonlocal variational approximations of the Mumford–Shah functional we refer the reader to Braides–Dal Maso [86], Gobbino [148], Cortesani and Toader [114], and references therein.

Let Ω be an open bounded subset of $\mathbf{R}^N$, $m > 0$, $0 < \alpha < \beta$ be such that $\alpha \operatorname{meas}(\Omega) \leq m \leq \beta \operatorname{meas}(\Omega)$ and let $SBV(\Omega : \{\alpha, \beta\})$ be the subspace of all functions of $SBV(\Omega)$ taking only the two values α or β. We consider the following problem:

$$\inf \left\{ \mathcal{H}^{N-1}(S_u) : u \in SBV(\Omega : \alpha, \beta), \ \int_\Omega u \, dx = m \right\}.$$

As said above, the thermodynamical model of phase transition provides an analogous estimate of this problem. Indeed, consider the functional F_ε defined in $L^1(\Omega)$ by

$$F_\varepsilon(u) = \begin{cases} c_0 \displaystyle\int_\Omega \left(\sqrt{\varepsilon}\, |Du|^2 + \frac{1}{\sqrt{\varepsilon}} W(u) \right) dx \text{ if } u \in H^1(\Omega),\ u \geq 0,\ \int_\Omega u\, dx = m, \\ +\infty \text{ otherwise,} \end{cases}$$

where

$$c_0 = \left(2 \int_\alpha^\beta \sqrt{W(t)}\, dt \right)^{-1},$$

ε is a positive parameter intended to go to zero, and $W : [0, +\infty[\to \mathbf{R}$ is a nonnegative, continuous function with exactly two zeros α, β $(0 < \alpha < \beta)$. The integral functional $c_0^{-1} F_\varepsilon$ is the rescaled Van Der Waals–Cahn–Hiliard energy functional by the ratio $1/\sqrt{\varepsilon}$ and W is a thermodynamical potential of a liquid with constant mass m confined to a bounded container Ω under isothermal conditions and whose density distribution u presents two phases α and β. More precisely, the Van Der Waals–Cahn–Hiliard energy is the functional

$$u \mapsto \int_\Omega W(u)\, dx + \mathrm{Ind}_{\{\int_\Omega v(x)\, dx = m,\ v \geq 0\}}(u) + \varepsilon \int_\Omega |Du|^2 dx,$$

where

$$\mathrm{Ind}_{\{\int_\Omega v(x)\, dx = m,\ v \geq 0\}}(u) = \begin{cases} 0 \text{ if } \displaystyle\int_\Omega u(x)\, dx = m,\ u \geq 0, \\ +\infty \text{ otherwise.} \end{cases}$$

The thickness L of the transition between the two phases is given by

$$L = \sqrt{\varepsilon} \frac{\beta - \alpha}{2 \int_\alpha^\beta \sqrt{W(\tau)}\, d\tau}.$$

Since L is very small, so is ε and the Van Der Waals–Cahn–Hiliard free energy is nothing but a perturbation of the Gibbs free energy functional

$$G : u \mapsto \int_\Omega W(u)\, dx + \mathrm{Ind}_{\{\int_\Omega v(x)\, dx = m,\ v \geq 0\}}(u)$$

by the functional $H : u \mapsto \varepsilon \int_\Omega |Du|^2\, dx$.

The first Gibbs model, which consists in minimizing G, is unsatisfactory. It is indeed easily seen that the set argmin (G) is made by infinitely many piecewise constant functions u taking the value α in an arbitrary subset A of Ω with measure $(\beta \mathrm{meas}(\Omega) - m)/(\beta - \alpha)$, and the value β in $\Omega \setminus A$, with no restriction on the shape of the interface between $[u = \alpha]$ and $[u = \beta]$. In particular, there is no way to recover the physical criterion: the interface has minimal area. This criterion may be recovered by the new model, consisting in minimizing the functional F_ε. We point out that because of argmin$(G) \cap$ dom $(H) = \emptyset$, this last model is a (viscosity) singular perturbation of the first one. For a general study of viscosity perturbations consult Attouch [29].

Modica proved in [180] the following result previously established in the special case $N = 1$ by Gurtin [152].

Theorem 12.4.2. *The sequence $(F_\varepsilon)_{\varepsilon>0}$ Γ-converges to the functional F defined by*

$$F(u) = \begin{cases} \mathcal{H}^{N-1}(S_u) \text{ if } u \in SBV(\Omega : \alpha, \beta), \text{ and } \displaystyle\int_\Omega u\, dx = m, \\ +\infty \text{ otherwise} \end{cases}$$

in $L^1(\Omega)$ equipped with its strong topology.

Assume moreover that W satisfies the following polynomial behavior at infinity: there exist $t_0 > 0, c_1 > 0,\ c_2 > 0,\ k \geq 2$ such that for all $t \geq t_0$

$$c_1 t^k \leq W(t) \leq c_2 t^k.$$

Then, the set $\{u_\varepsilon : \varepsilon \to 0\}$ of minimum points of F_ε has a compact closure in $L^1(\Omega)$, and any cluster point u is a minimum point of F.

Proof. We only give the proof of the lower bound in the definition of Γ-convergence and establish the compactness result. For a complete proof, consult [180] or the proof of Proposition 12.4.2. We begin by substituting $\sqrt{\varepsilon}$ by ε and we omit the constant c_0 in the definition of F_ε. The expected Γ-limit must be

$$F(u) = \begin{cases} c_0^{-1}\mathcal{H}^{N-1}(S_u) \text{ if } u \in SBV(\Omega : \alpha, \beta), \text{ and } \displaystyle\int_\Omega u\, dx = m, \\ +\infty \text{ otherwise,} \end{cases}$$

which is actually the asymptotic model of Van Der Waals–Cahn–Hiliard.

First step. We begin by proving that for all v in $L^1(\Omega)$ and all sequence $(v_\varepsilon)_{\varepsilon>0}$ strongly converging to v in $L^1(\Omega)$, one has

$$\liminf_{\varepsilon\to 0} F_\varepsilon(v_\varepsilon) \geq F(v).$$

The proof given here is based on a general method described in [221]. One may assume, for a subsequence not relabeled, that $\liminf_{\varepsilon\to 0} F_\varepsilon(v_\varepsilon) = \lim_{\varepsilon\to 0} F_\varepsilon(v_\varepsilon) = C < +\infty$ where C is a nonnegative constant which does not depend on ε. We then deduce

$$\begin{cases} v \geq 0, \ \displaystyle\int_\Omega v\, dx = m, \\ \displaystyle\int_\Omega W(v_\varepsilon)\, dx \leq C\varepsilon. \end{cases}$$

According to the continuity of W and Fatou's lemma, the last inequality yields

$$\int_\Omega W(v)\, dx \leq \liminf_{\varepsilon\to 0} \int_\Omega W(v_\varepsilon)\, dx \leq 0,$$

so that $W(v(x)) = 0$ a.e. and v takes only the two values α and β. Note that since truncatures operate on $H^1(\Omega)$, v is also the strong limit of the truncated functions $\tilde{v}_\varepsilon = \alpha \vee v_\varepsilon \wedge \beta$. Moreover, from the definition of W which achieves its infimum at α and β, $F_\varepsilon(v_\varepsilon) \geq F_\varepsilon(\tilde{v}_\varepsilon)$. According to these remarks, keeping the same notation, we will replace v_ε by $\tilde{v}_\varepsilon$. The elementary Young inequality yields

$$F_\varepsilon(v_\varepsilon) \geq 2\left(\int_\Omega W(v_\varepsilon)\, dx\right)^{1/2} \left(\int_\Omega |Dv_\varepsilon|^2\, dx\right)^{1/2}.$$

This estimate is optimal and may be recovered by studying the map $\varepsilon \mapsto F_\varepsilon(u)$ for a fixed u in $H^1(\Omega)$ whose minimum point is

$$\varepsilon = \left(\frac{\int_\Omega W(u)\, dx}{\int_\Omega |Du|^2\, dx} \right)^{1/2}$$

and for which the minimal value is

$$2 \left(\int_\Omega W(u)\, dx \right)^{1/2} \left(\int_\Omega |Du|^2\, dx \right)^{1/2}.$$

By the Cauchy–Schwarz inequality we obtain

$$\begin{aligned} \liminf_{\varepsilon \to 0} F_\varepsilon(v_\varepsilon) &\geq 2 \liminf_{\varepsilon \to 0} \int_\Omega \sqrt{W(v_\varepsilon)} |Dv_\varepsilon|\, dx \\ &= 2 \liminf_{\varepsilon \to 0} \int_\Omega |D\big(\psi(v_\varepsilon)\big)|\, dx, \end{aligned}$$

where $\psi(t) = \int_\alpha^t \sqrt{W(s)}\, ds$. Since $v_\varepsilon \to v$ strongly in $L^1(\Omega)$, $\alpha \leq v_\varepsilon \leq \beta$, and ψ is continuous, we deduce that $\psi(v_\varepsilon)$ strongly converges to $\psi(v)$ in $L^1(\Omega)$. According to Proposition 10.1.1, we finally deduce that $\psi(v)$ belong to $BV(\Omega)$ and

$$\liminf_{\varepsilon \to 0} F_\varepsilon(v_\varepsilon) \geq 2 \int_\Omega |D\psi(v)|. \tag{12.25}$$

Let us compute this last integral. Since

$$\psi(v) = \begin{cases} 0 \text{ on } [v = \alpha], \\ \displaystyle\int_\alpha^\beta \sqrt{W(s)}\, ds \text{ on } [v = \beta], \end{cases}$$

the function $\psi(v)$ is a simple function of $BV(\Omega)$ and

$$\int_\Omega |D\psi(v)| = \left(\int_\alpha^\beta \sqrt{W(s)}\, ds \right) \mathcal{H}^{N-1}(S_v).$$

Inequality (12.25) finally gives

$$\liminf_{\varepsilon \to 0} F_\varepsilon(v_\varepsilon) \geq \left(2 \int_\alpha^\beta \sqrt{W(s)}\, ds \right) \mathcal{H}^{N-1}(S_v), = F(v),$$

concluding the first step.

Second step. Let us now establish the relative compactness in $L^1(\Omega)$ of the set $\{u_\varepsilon : \varepsilon \to 0\}$ of minimum points of F_ε. The letter C will denote various positive constants. Consider $v_\varepsilon = \psi \circ u_\varepsilon$, where ψ is the primitive of the function $W^{1/2}$ defined above. Let us first prove the relative compactness of the set $\{v_\varepsilon : \varepsilon \to 0\}$. From the polynomial behavior of W and the fact that $k/2 + 1 \leq k$, we have for all $t \geq t_0$,

$$\begin{aligned} \psi(t) &\leq \int_\alpha^{t_0} W^{1/2}(s)\, ds + \int_{t_0}^t W^{1/2}(s)\, ds \\ &\leq C(1 + W(t)), \end{aligned}$$

which yields

$$\int_\Omega v_\varepsilon \, dx \le C(1 + \sqrt{\varepsilon} F_\varepsilon(u_\varepsilon))$$

and gives the boundedness in $L^1(\Omega)$ of v_ε. On the other hand, from the proof of the lower bound above

$$\int_\Omega |Dv_\varepsilon| \, dx \le \frac{1}{2} F_\varepsilon(v_\varepsilon),$$

which finally gives the boundedness of v_ε in $BV(\Omega)$. The relative compactness of $\{v_\varepsilon : \varepsilon \to 0\}$ is a consequence of the compactness of the embedding $BV(\Omega) \hookrightarrow L^1(\Omega)$.

Let us now go back to the functions u_ε. Let v be a strong limit in $L^1(\Omega)$ of a nonrelabeled subsequence of v_ε, consider the inverse function ψ^{-1} of ψ, and set $u = \psi^{-1} \circ v$. We establish the strong convergence of u_ε to u in $L^1(\Omega)$. We proceed as follows: we prove the equi-integrability of u_ε and the convergence in measure of u_ε to u (see, for instance, Marle [178]). From the polynomial behavior of W, we have

$$\begin{aligned}\int_\Omega |u_\varepsilon|^k \, dx &\le t_0^k \mathrm{meas}(\Omega) + C \int_\Omega W(u_\varepsilon) \, dx \\ &\le C(1 + \sqrt{\varepsilon} F_\varepsilon(u_\varepsilon)) \le C\end{aligned}$$

and equi-integrability follows from $k \ge 2$. On the other hand, since $\psi'(t) \ge \sqrt{c_1} t_0^{k/2}$ for all $t > t_0$, ψ^{-1} is a Lipschitz function on $[\psi(t_0), +\infty)$, hence uniformly continuous on $\mathbf{R}^+$. Therefore, u_ε converges in measure to u. □

For a numerical approach, it suffices now to establish the Γ-convergence of the discretization $F_{\varepsilon,h(\varepsilon)}$ of the functional F_ε by finite elements, to the functional F, with a suitable choice of the size h of discretization. For this study, consult Bellettini [61].

12.4.3 Variational approximation of the Mumford–Shah functional energy

When neglecting the functional $u \mapsto \int_\Omega |u - g|^2 \, dx$ in the expression of the Mumford–Shah functional, to control the jumps of admissible functions, a natural domain is the space $GSBV(\Omega)$ of generalized special functions of bounded variation defined by

$$GSBV(\Omega) := \big\{u : \Omega \to \mathbf{R} : u \text{ Borel function }, k \wedge u \vee (-k) \in SBV(\Omega) \ \forall k \in \mathbf{N}\big\}.$$

It can be shown (see Ambrosio–Tortorelli [24]) that to each function $u \in GSBV(\Omega)$ there corresponds a Borel function $\nabla u : \Omega \to \mathbf{R}^N$ and $S_u \subset \Omega$ such that $\nabla u = \nabla(k \wedge u \vee (-k))$ a.e. on $[|u| \le k]$ for all $k \in \mathbf{N}$ and $\mathcal{H}^{N-1}(S_{k \wedge u \vee (-k)}) \to \mathcal{H}^{N-1}(S_u)$ when $k \to +\infty$.

Following the strategy of the previous subsection, as in Ambrosio–Tortorelli [24], we establish that the functional F defined in $X := L^1(\Omega) \times L^1(\Omega, [0, 1])$ equipped with its strong topology by

$$F(u, s) = \begin{cases} \displaystyle\int_\Omega |\nabla u|^2 \, dx + \mathcal{H}^{N-1}(S_u) \text{ if } u \in GSBV(\Omega) \text{ and } s = 1, \\ +\infty \text{ otherwise} \end{cases}$$

can be approximated, in the sense of Γ-convergence, by the functionals defined in X by

$$F_\varepsilon(u,s) = \begin{cases} \displaystyle\int_\Omega (s^2+\varepsilon^2)|\nabla u|^2\,dx + M_\varepsilon(s,\Omega) \text{ if } (u,s) \in \mathbf{C}^1(\Omega)\times \mathbf{C}^1(\Omega,[0,1]) \cap X, \\ +\infty \text{ otherwise.} \end{cases}$$

For all open subsets A of Ω and all s in $\mathbf{C}^1(\Omega,[0,1])$, $M_\varepsilon(.,A)$ denotes the integral functional

$$M_\varepsilon(s,A) := \int_A \left(\varepsilon|\nabla s|^2 + \frac{1}{4\varepsilon}(1-s)^2\right) dx.$$

The second argument s is, as we will see in the proof, a control parameter on the gradient. The approximation of the Mumford–Shah energy will be the functional G_ε defined by $G_\varepsilon(u,s) = F_\varepsilon(u,s) + \int_\Omega |u-g|^2\,dx$. Indeed, $u \mapsto \int_\Omega |u-g|^2\,dx$ is a continuous perturbation of F_ε and the conclusion will follow from Theorem 12.1.1(ii).

We assume that Ω satisfies the following "reflection condition" $(\mathcal{R})$ on $\partial\Omega$: there exists an open neighborhood U of $\partial\Omega$ in $\mathbf{R}^N$ and a one-to-one Lipschitz function $\varphi : U\cap\Omega \longrightarrow U \setminus \overline{\Omega}$ such that φ^{-1} is Lipschitz.

Theorem 12.4.3. *Assume that Ω satisfies condition $(\mathcal{R})$. Then, the sequence of functionals $(F_\varepsilon)_{\varepsilon>0}$ Γ-converges to the functional F.*

The proof proceeds in the two Propositions 12.4.1 and 12.4.2. We denote the strong topology of $L^1(\Omega)\times L^1(\Omega,[0,1])$ by τ, and the letter C will denote various positive constants which do not depend on ε. We point out that condition $(\mathcal{R})$ is not necessary for obtaining the lower bound in Proposition 12.4.1.

Proposition 12.4.1. *For all $(u,s)\in X$ and all sequences $((u_\varepsilon,s_\varepsilon))_\varepsilon$ τ-converging to (u,s), we have $F(u,s) \le \liminf_{\varepsilon\to 0} F_\varepsilon(u_\varepsilon,s_\varepsilon)$, or equivalently, $F \le \Gamma - \liminf_{\varepsilon\to 0} F_\varepsilon$.*

Proof. Obviously, one may assume $\Gamma - \liminf_{\varepsilon\to 0} F_\varepsilon(u,s) < +\infty$ and $s=1$.

First step. We assume $u\in L^\infty(\Omega)$ and establish the proposition in the one-dimensional case $N=1$ when Ω is a bounded interval I in $\mathbf{R}$. When I is not an interval, it suffices to argue on each connected component of I and to conclude thanks to the superadditivity of $\Gamma - \liminf_{\varepsilon\to 0} F_\varepsilon$. When working on a bounded open subset A of $\mathbf{R}$, we will denote F and $\Gamma - \liminf_{\varepsilon\to 0} F_\varepsilon$ by $F(.,A)$ and $\Gamma - \liminf_{\varepsilon\to 0} F_\varepsilon(.,A)$, respectively. The key point of the proof is the following lemma.

Lemma 12.4.1. *Assume u in $L^\infty(I)$ and fix $x_0\in I$.*

(i) *If there exists $\eta>0$ such that for all $\rho<\eta$, $u\notin W^{1,2}(B_\rho(x_0))$, then for all $\rho<\eta$*

$$\Gamma - \liminf_{\varepsilon\to 0} F_\varepsilon((u,1),B_\rho(x_0)) \ge 1.$$

(ii) *If there exists $\rho>0$ such that $u\in W^{1,2}(B_\rho(x_0))$, then*

$$\Gamma - \liminf_{\varepsilon\to 0} F_\varepsilon((u,1),B_\rho(x_0)) \ge \int_{B_\rho(x_0)} |\nabla u|^2\,dx.$$

Assume for the moment that the proof of Lemma 12.4.1 is established. We claim that the set $E := \{x \in I : \exists \eta > 0 \ \forall \rho < \eta,\ u \notin W^{1,2}(B_\rho(x_0))\}$ is finite. Indeed, otherwise E would contain an infinite countable subset $D = \{x_i, i \in \mathbf{N}\}$. For all n in $\mathbf{N}$ and ρ small enough such that $B_\rho(x_i),\ i = 0,\ \ldots,\ n$, are pairwise disjoint sets, we would have from (i), superadditivity, and nondecreasing properties of $A \mapsto \Gamma - \liminf_{\varepsilon\to 0}\ F_\varepsilon(., A)$,

$$+\infty > \Gamma - \liminf_{\varepsilon\to 0} F_\varepsilon((u,1), I) \geq \sum_{i=0}^{n} \Gamma - \liminf_{\varepsilon\to 0}\ F_\varepsilon((u,1), B_\rho(x_i)) \geq n.$$

This being true for all $n \in \mathbf{N}$, we obtain a contradiction.

The set E is then made up of a finite number of points $x_0,\ \ldots,\ x_n$ and it is easily seen that $u \in W^{1,2}(I \setminus E)$. From $\mathcal{H}^0(E) < +\infty$, we deduce $u \in SBV(I)$ and $E = S_u$. For ρ small enough as previously, according to (ii) of Lemma 12.4.1, we have

$$\begin{aligned}
\Gamma - \liminf_{\varepsilon\to 0} F_\varepsilon((u,1), I) &\geq \Gamma - \liminf_{\varepsilon\to 0} F_\varepsilon\left((u,1), I \setminus \overline{\bigcup_{x\in S_u} B_\rho(x_0)}\right) \\
&\quad + \Gamma - \liminf_{\varepsilon\to 0} F_\varepsilon\left((u,1), \bigcup_{x\in S_u} B_\rho(x_0)\right) \\
&\geq \int_{I\setminus\overline{\bigcup_{x\in S_u} B_\rho(x_0)}} |\nabla u|^2\, dx + \mathcal{H}^0(S_u).
\end{aligned}$$

We conclude the step by letting $\rho \to 0$ in the above inequality.

It remains to establish assertions (i) and (ii) of Lemma 12.4.1. Let $(u_\varepsilon, s_\varepsilon) \in X \cap \mathbf{C}^1(I) \times \mathbf{C}^1(I, [0,1])$ τ-converging to $(u, 1)$ and satisfying

$$\liminf_{\varepsilon\to 0} F_\varepsilon((u_\varepsilon, s_\varepsilon), B_\rho(x_0)) < +\infty.$$

For proving (i), we establish the existence of $x_\varepsilon,\ x'_\varepsilon,\ x''_\varepsilon$ in $B_\rho(x_0)$ such that $x'_\varepsilon < x_\varepsilon < x''_\varepsilon$ and satisfying $\lim_{\varepsilon\to 0} s_\varepsilon(x_\varepsilon) = 0$, $\lim_{\varepsilon\to 0} s_\varepsilon(x'_\varepsilon) = \lim_{\varepsilon\to 0} s_\varepsilon(x''_\varepsilon) = 1$, for a nonrelabeled subsequence. Let us assume this result for the moment. We conclude as follows: by convexity inequality (precisely $a^2 + b^2 \geq 2ab$),

$$\begin{aligned}
F_\varepsilon(u_\varepsilon, s_\varepsilon) &\geq M_\varepsilon(s_\varepsilon, B_\rho(x_0)) \\
&\geq \int_{B_\rho(x_0)} (1 - s_\varepsilon)|\nabla s_\varepsilon|\, dx \\
&\geq \int_{x'_\varepsilon}^{x_\varepsilon} (1 - s_\varepsilon)|\nabla s_\varepsilon|\, dx + \int_{x_\varepsilon}^{x''_\varepsilon} (1 - s_\varepsilon)|\nabla s_\varepsilon|\, dx \\
&\geq \left|\int_{x'_\varepsilon}^{x_\varepsilon} (1 - s_\varepsilon)\nabla s_\varepsilon\, dx\right| + \left|\int_{x_\varepsilon}^{x''_\varepsilon} (1 - s_\varepsilon)\nabla s_\varepsilon\, dx\right| \\
&= \left|\left[-\frac{(1-s_\varepsilon)^2}{2}\right]_{x'_\varepsilon}^{x_\varepsilon}\right| + \left|\left[-\frac{(1-s_\varepsilon)^2}{2}\right]_{x_\varepsilon}^{x''_\varepsilon}\right|,
\end{aligned}$$

which tends to 1 when $\varepsilon \to 0$.

We are going to establish the existence of x_ε, x'_ε, and x''_ε. In what follows, we argue with various nonrelabeled subsequences and C denotes various positive constants independent of ε. Let $\sigma < \rho$ and set $m_\varepsilon := \inf_{B_\sigma(x_0)} s_\varepsilon$. From $F_\varepsilon((u_\varepsilon, s_\varepsilon), B_\rho(x_0)) \le C$, we derive

$$m_\varepsilon^2 \int_{B_\sigma(x_0)} |\nabla u_\varepsilon|^2 \, dx \le C.$$

Up to a subsequence, m_ε converges to some l, $0 \le l \le 1$. We claim that $l = 0$. Otherwise,

$$\lim_{\varepsilon\to 0} \int_{B_\sigma(x_0)} |\nabla u_\varepsilon|^2 \, dx \le \frac{C}{l^2}$$

and u_ε would weakly converge to u in $W^{1,2}(B_\sigma(x_0))$, which is in contradiction with $u \notin W^{1,2}(B_\rho(x_0))$ for all $\rho < \eta$. Consequently, there exists $x_\varepsilon \in \overline{B_\sigma(x)}$, satisfying $\lim_{\varepsilon\to 0} s_\varepsilon(x_\varepsilon) = \lim_{\varepsilon\to 0} m_\varepsilon = 0$. On the other hand, estimates

$$\int_{x_0-\rho}^{x_0-\sigma} \frac{(1-s_\varepsilon)^2}{4\varepsilon} \, dx \le C, \quad \int_{x_0+\sigma}^{x_0+\rho} \frac{(1-s_\varepsilon)^2}{4\varepsilon} \, dx \le C,$$

and the mean value theorem yield, for a subsequence, the existence of x'_ε and x''_ε satisfying the required assertions.

Let us show (ii). According to

$$\int_{B_\rho(x_0)} s_\varepsilon^2 |\nabla u_\varepsilon|^2 \, dx \le F_\varepsilon(u_\varepsilon, s_\varepsilon),$$

it is enough to establish the following inequality

$$\int_{B_\rho(x_0)} |\nabla u|^2 \, dx \le \liminf_{\varepsilon\to 0} \int_{B_\rho(x_0)} s_\varepsilon^2 |\nabla u_\varepsilon|^2 \, dx$$

when $F_\varepsilon(u_\varepsilon, s_\varepsilon)$ is equibounded. Set $v_\varepsilon = (1-s_\varepsilon)^2$. The equiboundedness of ∇v_ε in $L^1(I)$ will provide the following uniform control on v_ε: for all $\delta > 0$, there exists a finite part J_δ of I such that

$$\text{for all compact subset } K,\ K \subset I \setminus J_\delta,\ \text{one has } \limsup_{\varepsilon\to 0} \left(\sup_K v_\varepsilon\right) < \delta. \tag{12.26}$$

We will then deduce $\liminf_{\varepsilon\to 0} \left(\inf_K s_\varepsilon\right) > 1 - \delta^{1/2}$.

Let us assume for the moment estimate (12.26). We have for all $\delta > 0$ and all compact subset K with $K \subset I \setminus J_\delta$,

$$\begin{aligned} C \ge \int_{B_\rho(x_0)} s_\varepsilon^2 |\nabla u_\varepsilon|^2 \, dx &\ge \int_{B_\rho(x_0)\cap K} s_\varepsilon^2 |\nabla u_\varepsilon|^2 \, dx \\ &\ge \inf_K (s_\varepsilon^2) \int_{B_\rho(x_0)\cap K} |\nabla u_\varepsilon|^2 \, dx. \end{aligned}$$

Therefore

$$C \ge \liminf_{\varepsilon\to 0} \int_{B_\rho(x_0)} s_\varepsilon^2 |\nabla u_\varepsilon|^2 \, dx \ge (1-\delta^{\frac{1}{2}})^2 \liminf_{\varepsilon\to 0} \int_{B_\rho(x_0)\cap K} |\nabla u_\varepsilon|^2 \, dx,$$

and the weak convergence of u_ε to u in $W^{1,2}(B_\rho(x_0)\cap K)$ yields, by lower semicontinuity,

$$\liminf_{\varepsilon\to 0}\int_{B_\rho(x_0)} s_\varepsilon^2|\nabla u_\varepsilon|^2\,dx \geq (1-\delta^{\frac{1}{2}})^2\int_{B_\rho(x_0)\cap K}|\nabla u|^2\,dx.$$

The conclusion (ii) follows after letting K to I and $\delta\to 0$.

We are now going to establish (12.26). We claim that $\sup_\varepsilon \int_I |\nabla v_\varepsilon|\,dx < +\infty$. Indeed, by convexity

$$\begin{aligned}+\infty \geq 2M_\varepsilon(s_\varepsilon, I) &\geq \int_I 2(1-s_\varepsilon)|\nabla s_\varepsilon|\,dx\\ &= \int_I |\nabla v_\varepsilon|\,dx.\end{aligned}$$

Let $\sigma > 0$ satisfying $\delta > \sigma$ and consider for all t in $\mathbf{R}$, the sets $A_\varepsilon^t := [v_\varepsilon \leq t]$. According to the classical coarea formula, more precisely to Corollary 4.2.2, we have

$$\begin{aligned}C \geq \int_I |\nabla v_\varepsilon|\,dx &= \int_{-\infty}^{+\infty}\mathcal{H}^0([v_\varepsilon = t])\,dt\\ &\geq \int_\sigma^\delta \mathcal{H}^0([v_\varepsilon = t])\,dt.\end{aligned}$$

Therefore, there exists $t_\varepsilon \in]\sigma,\delta[$ such that $\mathcal{H}^0([v_\varepsilon = t_\varepsilon]) \leq \frac{C}{\delta-\sigma}$. The set $A_\varepsilon^{t_\varepsilon}$ has then at most $k = [\frac{C}{\delta-\sigma}]$ connected components with k independent of ε: more precisely, there exists a family $(I_\varepsilon^i)_{i=1,\dots,k}$ of intervals (possibly empty) such that $A_\varepsilon^{t_\varepsilon} = \bigcup_{i=1}^k I_\varepsilon^i$. For every $i = 1,\dots,k$, consider the interval $I_\infty^i = \bigcup_N \bigcap_{n\geq N} I_{\varepsilon_n}^i$. The complementary of the union of k intervals $I_\infty := \bigcup_{i=1}^k I_\infty^i$ is the required finite part I_δ of I. Indeed, since v_ε converges a.e. to zero,

$$\begin{aligned}\operatorname{meas}(I_\infty) = \operatorname{meas}\left(\bigcup_{i=1}^k\bigcup_N\bigcap_{n\geq N} I_{\varepsilon_n}^i\right) &= \operatorname{meas}\left(\bigcup_N\bigcap_{n\geq N} A_{\varepsilon_n}^{t_n}\right)\\ &\geq \operatorname{meas}\left(\bigcup_N\bigcap_{n\geq N}[v_{\varepsilon_n}\leq\sigma]\right)\\ &= \operatorname{meas}(I)\end{aligned}$$

so that $I_\delta = I\setminus I_\infty$ possesses k elements. Finally, if K is a compact set included in I_δ, arguing on each interval $\overset{0}{I}{}_\infty^i$, we have $K\cap \overset{0}{I}{}_\infty^i \subset\subset I_\infty^i$ and, for N large enough

$$K\cap \overset{0}{I}{}_\infty^i \subset \bigcap_{n\geq N} I_{\varepsilon_n}^i \subset \bigcap_{n\geq N}[v_{\varepsilon_n}\leq\delta].$$

Second step. We establish Proposition 12.4.1 in the N-dimensional case, $N > 1$. We will use the same notation for the functionals considered in the one-dimensional and the N-dimensional case.

We begin by assuming $u \in L^\infty(\Omega)$. Let $(u_\varepsilon, s_\varepsilon)$ be a sequence in X converging to $(u,1)$ such that $\liminf_{\varepsilon\to 0} F_\varepsilon((u_\varepsilon, s_\varepsilon), \Omega) < +\infty$ and A any open subset of Ω. With the notation and definitions of Theorem 10.5.2 for all $\nu \in S^{N-1}$ and for a subsequence not relabeled, $(u_{\varepsilon,x}, s_{\varepsilon,x})$ strongly converges in $L^1(A_x) \times L^1(A_x,[0,1])$ for $\mathcal{H}^{N-1}$ a.e. x in A_ν (it's an easy consequence of Fubini's theorem). On the other hand,

$$\begin{aligned}\int_{A_\nu} \liminf_{\varepsilon\to 0} F_\varepsilon((u_{\varepsilon,x}, s_{\varepsilon,x}), A_x)\, d\mathcal{H}^{N-1} &\le \liminf_{\varepsilon\to 0} \int_{A_\nu} F_\varepsilon((u_{\varepsilon,x}, s_{\varepsilon,x}), A_x)\, d\mathcal{H}^{N-1} \\ &= \liminf_{\varepsilon\to 0} F_\varepsilon((u_\varepsilon, s_\varepsilon), A) < +\infty.\end{aligned}$$

Thus, for $\mathcal{H}^{N-1}$ a.e. x in A_ν, $\liminf_{\varepsilon\to 0} F_\varepsilon((u_{\varepsilon,x}, s_{\varepsilon,x}), A_x) < +\infty$. One may apply the result of the first step: for $\mathcal{H}^{N-1}$ a.e. x in A_ν, u_x belongs to $SBV(A_x) \cap L^\infty(A_x)$ and

$$\liminf_{\varepsilon\to 0} F_\varepsilon((u_{\varepsilon,x}, s_{\varepsilon,x}), A_x) \ge \int_{A_x} |\nabla u_x|^2 + \mathcal{H}^0(S_{u_x} \cap A_x).$$

Integrating this inequality over A_ν, according to Theorem 10.5.2, we deduce that for all open subset A of Ω, $u \in SBV(A)$ and

$$\begin{aligned}\liminf_{\varepsilon\to 0} F_\varepsilon((u_\varepsilon, s_\varepsilon), A) &\ge \int_{A_\nu} \liminf_{\varepsilon\to 0} F_\varepsilon((u_{\varepsilon,x}, s_{\varepsilon,x}), A_x)\, d\mathcal{H}^{N-1} \\ &\ge \int_{A_\nu}\int_{A_x} |\nabla u_x|^2\, dt\, d\mathcal{H}^{N-1}(x) + \int_{A_\nu} \mathcal{H}^0(S_{u_x} \cap A_x)\, d\mathcal{H}^{N-1}(x) \\ &= \int_A |\nabla u.\nu|^2\, dx + \int_A |\nu_u.\nu|\, d\mathcal{H}^{N-1}\lfloor_{S_u}.\end{aligned}$$

We conclude thanks to Lemma 4.2.2 and Example 4.2.2.

If now u is not assumed to belong to $L^\infty(\Omega)$, by a truncation argument we have

$$\begin{aligned}\Gamma - \liminf_{\varepsilon\to 0} F_\varepsilon((u,s),\Omega) &\ge \Gamma - \liminf_{\varepsilon\to 0} F_\varepsilon((N \wedge u \vee (-N)), s), \Omega) \\ &\ge F((N \wedge u \vee (-N)), s), \Omega).\end{aligned}$$

Letting $N \to +\infty$ gives the thesis. □

Proposition 12.4.2. *For all $(u,s) \in X$ there exists a subsequence $((u_\varepsilon, s_\varepsilon))_\varepsilon$ τ-converging to (u,s) such that $F(u,s) \ge \limsup_{\varepsilon\to 0} F_\varepsilon(u_\varepsilon, s_\varepsilon)$ or, equivalently, $F \ge \Gamma - \limsup_{\varepsilon\to 0} F_\varepsilon$.*

Proof. One may assume u in $SBV(\Omega) \cap L^\infty(\Omega)$. Indeed, if $u \in GSBV(\Omega)$, an easy truncation argument gives the thesis. For a given $u \in SBV(\Omega) \cap L^\infty(\Omega)$, it suffices to construct $(u_\varepsilon, s_\varepsilon)$ in $H^1(\Omega) \times H^1(\Omega,[0,1]) \cap X$, τ-converging to $(u,1)$ in X and satisfying

$$F(u,s) \ge \limsup_{\varepsilon\to 0} F_\varepsilon(u_\varepsilon, s_\varepsilon).$$

The expression of F_ε has indeed a sense in $H^1(\Omega) \times H^1(\Omega,[0,1]) \cap X$. Moreover, $\mathbf{C}^1(\Omega) \times \mathbf{C}^1(\Omega,[0,1])$ is dense in $H^1(\Omega) \times H^1(\Omega,[0,1])$ equipped with its strong topology and F_ε is continuous for this topology which is stronger than τ. The conclusion will follow by a diagonalization argument.

First step. Let a_ε, b_ε, c_ε be three sequences in $\mathbf{R}^+$ going to zero which will be adjusted later in a suitable way. The idea consists in modifying $(u, 1)$ in a neighborhood of S_u to obtain, from the expression of $F_\varepsilon(u_\varepsilon, s_\varepsilon)$, an equivalent of $\mathcal{H}^{N-1}(S_u)$. We begin by assuming the following regularity condition on S_u:

$$\lim_{\rho\to 0} \frac{\mathrm{mes}(\Omega \cap (S_u)_\rho)}{2\rho} = H_{N-1}(S_u), \tag{12.27}$$

where $(S_u)_\rho$ is the tubular neighborhood $\{x \in \mathbf{R}^N : \mathrm{d}(x, S_u) < \rho\}$ of order ρ of S_u. In what follows, for any $t \in \mathbf{R}^+$, $(S_u)_t$ denotes a tubular neighborhood of order $t > 0$ of S_u. We construct u_ε in $H^1(\Omega)$ satisfying $u_\varepsilon = u$ in $\Omega \setminus (S_u)_{a_\varepsilon}$ and such that its gradient satisfies

$$|\nabla u_\varepsilon(x)| \le \frac{C}{a_\varepsilon} \tag{12.28}$$

a.e. in $(S_u)_{a_\varepsilon}$. Consider now s_ε in $H^1(\Omega, [0, 1])$ such that

$$s_\varepsilon = \begin{cases} 0 \text{ on } (S_u)_{a_\varepsilon}, \\ 1 - c_\varepsilon \text{ on } \Omega \setminus (S_u)_{a_\varepsilon + b_\varepsilon}. \end{cases}$$

The positive constant c_ε is introduced for technical reasons and, as said before, will be adjusted later. We have $(u_\varepsilon, s_\varepsilon) \in H^1(\Omega) \times H^1(\Omega, [0, 1]) \cap X$ and

$$\begin{aligned} F_\varepsilon((u_\varepsilon, s_\varepsilon), \Omega) = &\int_{\Omega\setminus(S_u)_{a_\varepsilon}} (s_\varepsilon^2 + \varepsilon^2)|\nabla u|^2\, dx + \int_{(S_u)_{a_\varepsilon}} \varepsilon^2 |\nabla u_\varepsilon|^2\, dx \\ &+ \frac{c_\varepsilon^2}{4\varepsilon}\mathrm{meas}(\Omega \setminus (S_u)_{a_\varepsilon + b_\varepsilon}) + \frac{1}{4\varepsilon}\mathrm{meas}((S_u)_{a_\varepsilon}) \\ &+ M_\varepsilon(s_\varepsilon, (S_u)_{a_\varepsilon + b_\varepsilon} \setminus (S_u)_{a_\varepsilon}). \end{aligned}$$

The first term trivially goes to $\int_\Omega |\nabla u|^2\, dx$ when ε goes to zero. We adjust a_ε so that the second and fourth terms go to zero. For this, it suffices, thanks to (12.28), to select an intermediate power of ε between ε^2 and ε, for instance, $a_\varepsilon = \varepsilon^{3/2}$. To make the third term vanish, it suffices to choose $c_\varepsilon \le \sqrt{\varepsilon}$, for instance, $c_\varepsilon = \varepsilon^{3/4}$. We are reduced to finding s_ε satisfying

$$\limsup_{\varepsilon\to 0} M_\varepsilon(s_\varepsilon, (S_u)_{a_\varepsilon + b_\varepsilon} \setminus (S_u)_{a_\varepsilon}) \le H_{N-1}(S_u).$$

Let us denote the map $d(.) = \mathrm{dist}(., S_u)$ by d. We try to find s_ε of the form $\sigma_\varepsilon \circ d$. Applying the coarea formula Theorem 4.2.5 to the function $g = \varepsilon|\sigma_\varepsilon' \circ d|^2 + \frac{(1-\sigma_\varepsilon \circ d)^2}{4\varepsilon}$ and to the truncated function $f = a_\varepsilon \vee d \wedge (a_\varepsilon + b_\varepsilon)$ of d, we obtain

$$M_\varepsilon(s_\varepsilon, (S_u)_{a_\varepsilon + b_\varepsilon} \setminus (S_u)_{a_\varepsilon}) = \int_{a_\varepsilon}^{a_\varepsilon + b_\varepsilon} \left(\varepsilon|\sigma_\varepsilon'(t)|^2 + \frac{(1-\sigma_\varepsilon(t))^2}{4\varepsilon}\right) H_{N-1}([d = t])\, dt.$$

Consider $h(t) = \mathrm{meas}([d < t])$. Then, according to Corollary 4.2.3, $h'(t) = H_{N-1}([d = t])$, and

$$M_\varepsilon(s_\varepsilon, (S_u)_{a_\varepsilon + b_\varepsilon} \setminus (S_u)_{a_\varepsilon}) = \int_{a_\varepsilon}^{a_\varepsilon + b_\varepsilon} \left(\varepsilon|\sigma_\varepsilon'(t)|^2 + \frac{(1-\sigma_\varepsilon(t))^2}{4\varepsilon}\right) h'(t)\, dt.$$

The function σ_ε is chosen as the solution of the ordinary boundary value problem

$$\begin{cases} \sigma_\varepsilon' = \frac{1-\sigma_\varepsilon}{2\varepsilon}, \\ \sigma_\varepsilon(a_\varepsilon) = 0,\ \sigma_\varepsilon(a_\varepsilon + b_\varepsilon) = 1 - c_\varepsilon, \end{cases}$$

that is, $\sigma_\varepsilon(t) = 1 - \exp(\frac{a_\varepsilon - t}{2\varepsilon})$ when we choose $b_\varepsilon = -\varepsilon \ln(\varepsilon^{3/2})$. On the other hand, the regularity assumption on S_u yields for all $\eta > 0$ the existence of ε_0 such that for all $\varepsilon < \varepsilon_0$ and all $t < a_\varepsilon + b_\varepsilon$, one has $h(t) \le 2t(H_{N-1}(S_u) + \eta)$. Thanks to this estimate, the conclusion then follows by integrating by parts the expression $M_\varepsilon(s_\varepsilon, (S_u)_{a_\varepsilon + b_\varepsilon} \setminus (S_u)_{a_\varepsilon})$ (for details see [24, Proposition 5.1]).

Second step. We do not assume hypothesis (12.27). To apply the first step, we construct a sequence u_η converging to u in $L^2(\Omega)$ such that S_{u_η} satisfies (12.27) and such that $F(u) = \lim_{\eta \to 0} F(u_\eta)$. Afterwards, it will suffice to apply the procedure of the first step to the function u_η and to conclude by a diagonalization argument.

For constructing u_η, the idea consists in finding u_η as a solution of the Mumford–Shah problem

$$\inf \left\{ \int_{\Omega'} |\nabla v|^2 + \mathcal{H}^{N-1}(S_v) + \frac{1}{\eta} \int_{\Omega'} |v - \overline{u}|^2 \, dx : v \in SBV(\Omega') \right\}, \qquad (\mathcal{P}_\eta)$$

where $\Omega' = \Omega \cup U$ and $\overline{u}$ is the extension of u on Ω' defined by

$$\overline{u}(x) = \begin{cases} u(\varphi^{-1}(x)) \text{ if } x \in U \setminus \overline{\Omega}, \\ \gamma_0(u) \text{ if } x \in \partial\Omega, \\ u(x) \text{ if } x \in \Omega. \end{cases}$$

U and φ are given by the regularity condition $(\mathcal{R})$ fulfilled by Ω and γ_0 denotes the trace operator.

We next use the following regularity property related to Mumford–Shah solutions (see Ambrosio and Tortorelli [24], De Giorgi, Carriero, and Leaci [123]): $\mathcal{H}^{N-1}(\overline{S}_{u_\eta} \cap \Omega' \setminus S_{u_\eta}) = 0$ and, for all compact set K included in $\overline{S}_{u_\eta} \cap \Omega'$,

$$\lim_{\rho \to 0} \frac{\text{meas}((K)_\rho)}{2\rho} = \mathcal{H}^{N-1}(K).$$

Taking $K = \overline{\Omega} \cap \overline{S_{u_\eta}}$, we obtain the required regularity on u_η in Ω.

Obviously, u_η converges to u in $L^2(\Omega)$ thanks to the penalization parameter $1/\eta$ in $(\mathcal{P}_\eta)$. It remains to establish the convergence of $F(u_\eta)$ to $F(u)$. Consider the two Borel measures μ_η and μ in $\mathbf{M}^+(\Omega')$ defined for all Borel set B in Ω', by

$$\begin{cases} \mu_\eta(B) := \displaystyle\int_B |\nabla u_\eta|^2 \, dx + \mathcal{H}^{N-1}(B \cap S_{u_\eta}), \\ \mu(B) := \displaystyle\int_B |\nabla \overline{u}|^2 \, dx + \mathcal{H}^{N-1}(B \cap S_{\overline{u}}). \end{cases}$$

It is worth noticing that $\overline{u}$ has no jump through $\partial\Omega$ so that $\mu(\partial\Omega) = 0$. Taking $v = \overline{u}$ as a test function in $(\mathcal{P}_\eta)$, we obtain

$$\limsup_{\eta \to 0} \mu_\eta(\Omega') \le \mu(\Omega').$$

On the other hand, according to Theorem 13.4.3, which we will establish in the next chapter, we have for all open subset A of Ω'

$$\mu(A) \leq \liminf_{\eta \to 0} \mu_\eta(A).$$

According to Proposition 4.2.5, we deduce that μ_η narrow converges to μ. Since $\mu(\partial\Omega) = 0$, we have $\mu_\eta(\Omega) \to \mu(\Omega)$. □

Chapter 13

Integral functionals of the calculus of variations

This chapter is devoted to the study of the sequential lower semicontinuity of certain types of functionals which occur in many variational problems. As noted in chapter 11, lower semicontinuity is the key tool to apply the direct methods of the calculus of variations, and we deal in the sections below with some different cases, depending on the spaces the functionals are defined on. We will see that, due to the integral form of the functionals under consideration, the convexity or quasi-convexity conditions play a central role in all the results. We first complement Section 11.2 by establishing necessary and sufficient conditions for more general integral functionals to be lower semicontinuous. Then to complement Section 11.3, we deal with lower semicontinuity of functionals defined on the space of measures, on BV and SBV. We do not pretend to be exhaustive in this very widely studied field; we intend only to give here some principal results. For other cases and details see [22], [97], [102], [116], [143], [185].

13.1 Lower semicontinuity in the scalar case

In this section we consider integral functionals of the form

$$F(u) = \int_\Omega f(x, u, Du)\, dx, \tag{13.1}$$

where u varies on a Sobolev space $W^{1,p}(\Omega)$. We stress the fact that in this section we restrict our attention to the case of functions u which take their values in $\mathbf{R}$; some differences with the case of functionals defined on vector valued functions will be discussed in the next section.

To study the sufficient conditions for the lower semicontinuity of functionals of the form (13.1), it is convenient to consider first the case of functionals of the form

$$F(u, v) = \int_\Omega f\big(x, u(x), v(x)\big)\, d\mu(x), \tag{13.2}$$

where $(\Omega, \mathcal{A}, \mu)$ is a measure space with the measure μ nonnegative and finite (i.e., $\mu \in \mathbf{M}^+(\Omega)$), $f : \Omega \times \mathbf{R}^m \times \mathbf{R}^n \to [0, +\infty]$ is an $\mathcal{A} \otimes \mathbf{B}_m \otimes \mathbf{B}_n$-measurable function ($\mathbf{B}_m$ and $\mathbf{B}_n$,

respectively, denote the σ-algebras of Borel subsets of $\mathbf{R}^m$ and $\mathbf{R}^n$), and u, v, respectively, vary in the spaces $L^1_\mu(\Omega;\mathbf{R}^m)$, $L^1_\mu(\Omega;\mathbf{R}^n)$ of μ integrable $\mathbf{R}^m$, $\mathbf{R}^n$ valued functions.

The theorem below is a lower semicontinuity result for functionals of the form (13.2); the link with the case (13.1) will be discussed later.

Theorem 13.1.1. *Assume the function f satisfies the following conditions:*

(i) *for μ-a.e. $x \in \Omega$ the function $f(x,\cdot,\cdot)$ is lower semicontinuous on $\mathbf{R}^m \times \mathbf{R}^n$;*

(ii) *for μ-a.e. $x \in \Omega$ and for every $s \in \mathbf{R}^m$ the function $f(x,s,\cdot)$ is convex on $\mathbf{R}^n$.*

Then the functional F defined in (13.2) *is sequentially lower semicontinuous on the space $L^1_\mu(\Omega;\mathbf{R}^m) \times L^1_\mu(\Omega;\mathbf{R}^n)$ endowed with the strong topology on $L^1_\mu(\Omega;\mathbf{R}^m)$ and the weak topology on $L^1_\mu(\Omega;\mathbf{R}^n)$.*

Proof. Let $u_h \to u$ strongly in $L^1_\mu(\Omega;\mathbf{R}^m)$ and $v_h \to v$ weakly in $L^1_\mu(\Omega;\mathbf{R}^n)$; we have to prove that

$$F(u,v) \le \liminf_{h\to+\infty} F(u_h,v_h). \tag{13.3}$$

Possibly passing to subsequences, we may assume without loss of generality that the lim inf in the right-hand side of (13.3) is a finite limit, that is,

$$\lim_{h\to+\infty} F(u_h,v_h) = c \in \mathbf{R}. \tag{13.4}$$

Since the sequence (v_h) is weakly compact in $L^1_\mu(\Omega;\mathbf{R}^n)$, we may use the Dunford–Pettis theorem, Theorem 2.4.5, and the De La Vallée–Poussin criterion, Theorem 2.4.4, to conclude that there exists a function $\vartheta : [0,+\infty[\to [0,+\infty[$, which can be taken convex and strictly increasing, with a superlinear growth, that is,

$$\lim_{t\to+\infty} \frac{\vartheta(t)}{t} = +\infty,$$

such that

$$\sup_{h\in\mathbf{N}} \int_\Omega \vartheta(|v_h|)\,d\mu \le 1. \tag{13.5}$$

Setting

$$\begin{cases} H(t) = \sqrt{t\vartheta(t)}, \\ \Phi(t) = \vartheta\big(H^{-1}(t)\big), \\ \xi_h(x) = H\big(|v_h(x)|\big), \end{cases}$$

it is easy to see that

(i) H is strictly increasing and $H(t)/t \to +\infty$ as $t \to +\infty$,

(ii) Φ is strictly increasing and $\Phi(t)/t \to +\infty$ as $t \to +\infty$,

(iii) $\vartheta(t)/H(t) \to +\infty$ as $t \to +\infty$,

(iv) $\Phi\big(\xi_h(x)\big) = \vartheta\big(|v_h(x)|\big)$.

Therefore, by (13.5) we have

$$\sup_{h\in\mathbf{N}}\int_\Omega \Phi(\xi_h)\,d\mu \le 1.$$

We can use now the Dunford–Pettis theorem again to deduce that the sequence (ξ_h) is weakly compact in $L^1_\mu(\Omega)$, hence (up to extracting a subsequence) we may assume that $\xi_h \to \eta$ weakly in $L^1_\mu(\Omega)$ for a suitable η. By the Mazur theorem, a suitable sequence of convex combinations of (ξ_h, v_h) is strongly convergent in $L^1_\mu(\Omega)\times L^1_\mu(\Omega;\mathbf{R}^n)$ to (η, v). More precisely, there exist $N_h \to +\infty$ and $\alpha_{i,h} \ge 0$ with

$$\sum_{i=N_h+1}^{N_{h+1}} \alpha_{i,h} = 1$$

such that the sequences

$$\eta_h = \sum_{i=N_h+1}^{N_{h+1}} \alpha_{i,h}\xi_i, \qquad v_h = \sum_{i=N_h+1}^{N_{h+1}} \alpha_{i,h} v_i$$

strongly converge to η in $L^1_\mu(\Omega)$ and to v $L^1_\mu(\Omega;\mathbf{R}^n)$, respectively. Possibly passing to subsequences we may also assume that $\eta_h \to \eta$, $v_h \to v$, and $u_h \to u$ pointwise μ-a.e. on Ω.

Consider now a point $x \in \Omega$ where all the convergences above occur, and set

$$\begin{cases} \varepsilon_h = \max\big\{|u(x) - u_i(x)| \;:\; N_h < i \le N_{h+1}\big\},\\ \lambda_h = \displaystyle\sum_{i=N_h+1}^{N_{h+1}} \alpha_{i,h} f\big(x, u_i(x), v_i(x)\big),\\ \mathcal{A}_h = \big\{(v,\eta,\lambda)\in\mathbf{R}^{n+2} \;:\; \eta = H(|v|),\ \exists s\in\mathbf{R}^m,\ |s-u(x)|\le\varepsilon_h,\ \lambda \ge f(x,s,v)\big\}. \end{cases}$$

We have $\varepsilon_h \to 0$ and by definition of v_h, η_h, λ_h we obtain that $\big(v_h(x), \eta_h(x), \lambda_h(x)\big)$ belongs to the convex hull $co\mathcal{A}_h$ of $\mathcal{A}_h$. Since $\mathcal{A}_h \subset \mathbf{R}^{n+2}$, by the Carathéodory theorem on convex hulls in Euclidean spaces the vector $\big(v_h(x), \eta_h(x), \lambda_h(x)\big)$ can be written as a convex combination of $n+3$ elements of $\mathcal{A}_h$, that is, there exist

$$\beta_{i,h} \ge 0, \quad v_{i,h}\in\mathbf{R}^n, \quad \eta_{i,h}\ge 0, \quad \lambda_{i,h}\ge 0 \qquad (i = 1,\dots,n+3)$$

such that $(v_{i,h}, \eta_{i,h}, \lambda_{i,h}) \in \mathcal{A}_h$ for every index i, and

$$\begin{cases} \displaystyle\sum_{i=1}^{n+3}\beta_{i,h} = 1, \qquad \sum_{i=1}^{n+3}\beta_{i,h}v_{i,h} = v_h(x),\\ \displaystyle\sum_{i=1}^{n+3}\beta_{i,h}\eta_{i,h} = \eta_h(x), \qquad \sum_{i=1}^{n+3}\beta_{i,h}\lambda_{i,h} = \lambda_h(x). \end{cases}$$

Therefore, for suitable $s_{i,h}\in\mathbf{R}^m$ with $|s_{i,h}-u(x)|\le\varepsilon_h$ we have

$$\lambda_{i,h} \ge f(x, s_{i,h}, v_{i,h}).$$

By extracting subsequences, without loss of generality, we may assume that for every index i the sequence $|v_{i,h}|$ tends to a limit and, denoting by I the set of indices i such that this limit is finite, again by passing to subsequences, we may also assume that

$$\begin{cases} v_{i,h} \to v_i & \forall i \in I, \\ |v_{i,h}| \to +\infty & \forall i \notin I, \\ \beta_{i,h} \to \beta_i & \forall i = 1, \dots, n+3. \end{cases}$$

Since

$$\sum_{i=1}^{n+3} \beta_{i,h} H(|v_{i,h}|) = \eta_h(x) \to \eta(x),$$

the set I cannot be empty. From the relation

$$\sum_{i=1}^{n+3} \beta_{i,h} v_{i,h} = v_h(x) \to v(x),$$

we obtain that $\beta_i = 0$ for every $i \notin I$. Moreover, from

$$\eta_h(x) = \sum_{i=1}^{n+3} \beta_{i,h} \eta_{i,h} \geq \sum_{i \notin I} \beta_{i,h} \eta_{i,h} = \sum_{i \notin I} \beta_{i,h} |v_{i,h}| \frac{H(|v_{i,h}|)}{|v_{i,h}|}$$

we get

$$\beta_{i,h} |v_{i,h}| \to 0 \qquad \forall i \notin I$$

so that

$$\sum_{i \in I} \beta_i = 1, \qquad \sum_{i \in I} \beta_i v_i = v(x).$$

We now use the assumptions on the function f to obtain

$$\begin{aligned} f\big(x, u(x), v(x)\big) &\leq \sum_{i \in I} \beta_i f\big(x, u(x), v_i\big) \\ &\leq \liminf_{h \to +\infty} \sum_{i \in I} \beta_{i,h} f(x, s_{i,h}, v_{i,h}) \\ &\leq \liminf_{h \to +\infty} \sum_{i=1}^{n+3} \beta_{i,h} f(x, s_{i,h}, v_{i,h}) \\ &\leq \liminf_{h \to +\infty} \lambda_h(x) \end{aligned}$$

so that by Fatou's lemma,

$$\int_\Omega f(x, u, v)\, d\mu \leq \liminf_{h \to +\infty} \int_\Omega \lambda_h(x)\, d\mu \tag{13.6}$$

$$= \liminf_{h \to +\infty} \sum_{i=N_h+1}^{N_{h+1}} \alpha_{i,h} \int_\Omega f(x, u_i, v_i)\, d\mu. \tag{13.7}$$

Fix now $\varepsilon > 0$; by using (13.4) we obtain, for h large enough,

$$\int_\Omega f(x, u_i, v_i)\, d\mu \le c + \varepsilon \qquad \forall i \in [N_h + 1, N_{h+1}]$$

and so by (13.7) $F(u, v) \le c + \varepsilon$. The proof is then achieved by taking $\varepsilon \to 0^+$. □

Remark 13.1.1. It is easy to see that the result of Theorem 13.1.1 remains true if the measure μ is only assumed to be σ-finite.

The result above, under the slightly stronger assumption that $f(x, \cdot, \cdot)$ is continuous for μ-a.e. $x \in \Omega$, was first obtained by De Giorgi in 1968 in an unpublished paper. The original proof by De Giorgi is obtained by approximating from below the convex function $f(x, s, \cdot)$ by finite suprema of affine functions $f_k(x, s, \cdot)$ for which the proof is easier, and then passing to the limit as $k \to +\infty$; the interested reader may find further details about this type of proof in the book by Buttazzo [97]. The proof reported above follows on the contrary the scheme of the proof which was given in 1977 by Ioffe [157].

By using the result of Theorem 13.1.1 we can easily give some sufficient conditions for the sequential lower semicontinuity of functionals of the form (13.1) on the Sobolev space $W^{1,1}(\Omega)$. More precisely, the following result holds.

Theorem 13.1.2. *Let Ω be a Lipschitz domain of $\mathbf{R}^n$ and let $f : \Omega \times \mathbf{R}^m \times \mathbf{R}^{mn} \to [0, +\infty]$ be a function verifying the assumptions of Theorem* 13.1.1. *Then the functional F defined in* (13.1) *is sequentially weakly lower semicontinuous on the Sobolev space $W^{1,1}(\Omega; \mathbf{R}^m)$.*

Proof. Let (u_h) be a sequence in $W^{1,1}(\Omega; \mathbf{R}^m)$ converging weakly to some function u. Setting $v_h = Du_h$ we have that v_h converges weakly to $v = Du$ in $L^1(\Omega; \mathbf{R}^{mn})$ and, by the Rellich theorem, u_h converge strongly to u in $L^1(\Omega; \mathbf{R}^m)$. By Theorem 13.1.1 we have

$$F(u) = \int_\Omega f(x, u, v)\, dx \le \liminf_{h \to +\infty} \int_\Omega f(x, u_h, v_h)\, dx = \liminf_{h \to +\infty} F(u_h),$$

which then proves the assertion. □

In the so-called scalar case (i.e., when $m = 1$), and in the case of ordinary integrals as well (i.e., when $n = 1$), we will prove below that the convexity of the integrand f with respect to the gradient is also a necessary condition for the semicontinuity. This is no longer true in the vector valued case $m > 1$ for multiple integrals, as we will discuss in the next section. For the sake of simplicity we here limit ourselves to consider only the case of functionals of the form

$$F(u) = \int_\Omega f(Du)\, dx; \tag{13.8}$$

the more general case (13.1) presents only technical differences in the proof, and we refer to one of the books mentioned at the beginning of this chapter for the details.

Theorem 13.1.3. *Assume that either $m = 1$ or $n = 1$ and that the functional F in* (13.8) *is sequentially weakly* lower semicontinuous in the Sobolev space $W^{1,\infty}$, in the sense that*

$$F(u) \le \liminf_{h \to +\infty} F(u_h) \tag{13.9}$$

for every sequence u_h converging to u uniformly in Ω and with Du_h uniformly bounded in Ω. Then the function f is convex and lower semicontinuous.

PROOF. We give the proof only in the case $m = 1$, the other one being similar. Let $z_1, z_2 \in \mathbf{R}^n$, let $t \in]0,1[$, and let $z = tz_1 + (1-t)z_2$. Denote the linear function $u_z(x) = z \cdot x$ by u_z and define

$$\begin{cases} z_0 = \dfrac{z_2 - z_1}{|z_2 - z_1|}, \\ \Omega^1_{hj} = \left\{ x \in \Omega \ : \ \frac{j-1}{h} < z_0 \cdot x < \frac{j-1+t}{h} \right\}, \qquad j \in \mathbf{Z},\ h \in \mathbf{N}, \\ \Omega^2_{hj} = \left\{ x \in \Omega \ : \ \frac{j-1+t}{h} < z_0 \cdot x < \frac{j}{h} \right\}, \qquad j \in \mathbf{Z},\ h \in \mathbf{N}, \\ \Omega^1_h = \bigcup \left\{ \Omega^1_{hj} \ : \ j \in \mathbf{Z} \right\}, \\ \Omega^2_h = \bigcup \left\{ \Omega^2_{hj} \ : \ j \in \mathbf{Z} \right\}, \\ u_h(x) = \begin{cases} c^1_{hj} + z_1 \cdot x \text{ if } x \in \Omega^1_{hj}, \\ c^2_{hj} + z_2 \cdot x \text{ if } x \in \Omega^2_{hj}, \end{cases} \end{cases}$$

where

$$c^1_{hj} = \frac{(j-1)(1-t)}{h}|z_2 - z_1|, \qquad c^2_{hj} = -\frac{jt}{h}|z_2 - z_1|.$$

It is easy to verify that, as $h \to +\infty$,

$$\frac{\operatorname{meas}(\Omega^1_h)}{\operatorname{meas}(\Omega)} \to t, \qquad \frac{\operatorname{meas}(\Omega^2_h)}{\operatorname{meas}(\Omega)} \to 1 - t. \tag{13.10}$$

Moreover, the functions u_h are Lipschitz continuous and for every $x \in \Omega^1_{hj}$ we have

$$\begin{aligned} \left|u_h(x) - u_z(x)\right| &= \left|c^1_{hj} + (z_1 - z) \cdot x\right| = (1-t)\left|\frac{j-1}{h}|z_2 - z_1| + (z_1 - z_2) \cdot x\right| \\ &= (1-t)|z_2 - z_1|\left|\frac{j-1}{h} - z_0 \cdot x\right| \le \frac{t(1-t)}{h}|z_2 - z_1|. \end{aligned}$$

Analogously, a similar computation gives for every $x \in \Omega^2_{hj}$

$$\left|u_h(x) - u_z(x)\right| \le \frac{t(1-t)}{h}|z_2 - z_1|.$$

Therefore $u_h \to u_z$ uniformly on Ω. Morever, it is immediate to see that the gradients Du_h are uniformly bounded on Ω, so that the sequence (u_h) converges to u_z weakly* in $W^{1,\infty}(\Omega)$. By the assumption (13.9), using also (13.10), we then obtain

$$\begin{aligned} f(z)\operatorname{meas}(\Omega) = F(u_z) &\le \liminf_{h \to +\infty} F(u_h) \\ &= \liminf_{h \to +\infty} \left(f(z_1)\operatorname{meas}(\Omega^1_h) + f(z_2)\operatorname{meas}(\Omega^2_h) \right) \\ &= t f(z_1)\operatorname{meas}(\Omega) + (1-t) f(z_2)\operatorname{meas}(\Omega), \end{aligned}$$

which proves the convexity of f. The lower semicontinuity of f on $\mathbf{R}^n$ is a straightforward consequence of the sequential lower semicontinuity assumption (13.9) on F. □

13.2 Lower semicontinuity in the vectorial case

We have seen in the previous section that the convexity assumption on the integrand $f(x, s, \cdot)$ is necessary and sufficient for the sequential weak lower semicontinuity of the functional

$$F(u) = \int_\Omega f(x, u, Du)\, dx \tag{13.11}$$

in the case of scalar functions u. On the contrary, in the case of functions u with vector values, the convexity of the integrand with respect to the gradient variable describes only a small class of weakly lower semicontinuous functionals. For this reason, according to Morrey [185] we introduce the notion of quasi-convexity defined in chapter 11 in the context of relaxation theory.

Definition 13.2.1. *A Borel function $f : \mathbf{R}^{mn} \to [0, +\infty]$ is said to be quasi-convex if*

$$f(z)\,\mathrm{meas}(A) \le \int_A f\big(z + D\phi(x)\big)\, dx \tag{13.12}$$

for a suitable (hence for all) bounded open subset A of $\mathbf{R}^n$, every $m \times n$ matrix z, and every $\phi \in \mathbf{C}_0^1(A; \mathbf{R}^m)$.

Remark 13.2.1. It is possible to prove that when either $m = 1$ or $n = 1$ quasi-convexity reduces to the usual convexity; this will actually follow from Theorem 13.1.3 once the equivalence between lower semicontinuity and quasi-convexity will be proved. On the other hand, there are many examples of quasi-convex functions which are not convex, as, for instance, the function

$$z \mapsto f(z) = |\det z|.$$

If $f : \mathbf{R}^{mn} \to [0, +\infty]$ is quasi-convex, then for every $z \in \mathbf{R}^{mn}$ and $s \in \mathbf{R}^m$ the function $\phi_{s,z} : \mathbf{R}^n \to [0, +\infty]$ defined by

$$\phi_{s,z}(\xi) = f(z + s \otimes \xi) \tag{13.13}$$

is convex. This fact follows again from the results of the previous section on the scalar case, once we remark that by (13.12) we obtain

$$\phi_{s,z}(\xi)\,\mathrm{meas}(A) \le \int_A \phi_{s,z}\big(\xi + D\phi(x)\big)\, dx$$

for every vector $\xi \in \mathbf{R}^n$ and every scalar function $\phi \in \mathbf{C}_0^1(A)$. The property given by (13.13) is called rank-one convexity. From the convexity of the functions $\phi_{s,z}$ defined in (13.13), we obtain that every rank-one convex function f of class $\mathbf{C}^2(\mathbf{R}^{mn})$ satisfies the so-called Legendre–Hadamard condition:

$$\frac{\partial^2 f}{\partial z_{ij}\partial z_{hk}}(z_0)\alpha_i\alpha_h\beta_j\beta_k \ge 0$$

for all $z_0 \in \mathbf{R}^{mn}$ and for all $\alpha \in \mathbf{R}^m$, $\beta \in \mathbf{R}^n$. (The summation convention over repeated indices is adopted.) Moreover, we get that every rank-one convex finite-valued function is locally Lipschitz on $\mathbf{R}^{mn}$. This is made precise in Lemma 13.2.1.

Remark 13.2.2. A wide class of quasi-convex functions is the class of polyconvex functions, introduced by Ball in [54]. A function f is called polyconvex if it can be written in the form

$$f(z) = g\big(X(z)\big) \qquad \forall z \in \mathbf{R}^{m\times n},$$

where $X(z)$ denotes the vector of all subdeterminants of the matrix z and g is a convex function. For instance, if $m = n = 2$, every polyconvex function is of the form $g(z, \det z)$ with g convex on $\mathbf{R}^4 \times \mathbf{R}$; analogously, if $m = n = 3$, every polyconvex function is of the form $g(z, \operatorname{adj} z, \det z)$ with g convex on $\mathbf{R}^9 \times \mathbf{R}^9 \times \mathbf{R}$ and where $\operatorname{adj} z$ denotes the adjugate matrix of z, that is the transpose of the matrix of cofactors of z.

Lemma 13.2.1. *Let $f : \mathbf{R}^{mn} \to \mathbf{R}$ be a rank-one convex function such that*

$$0 \le f(z) \le c(1 + |z|^p) \quad \forall z \in \mathbf{R}^{mn}.$$

Then, for a suitable constant $k > 0$ we have

$$\big|f(z) - f(w)\big| \le k|z - w|\big(1 + |z|^{p-1} + |w|^{p-1}\big) \quad \forall z, w \in \mathbf{R}^{mn}.$$

Proof. Being f convex with respect to each column vector, it is enough to prove the inequality in the case f convex. Again, arguing component by component, we may assume that f is a convex function of only one variable t. These functions are differentiable almost everywhere, and we have

$$f'(t) \le \frac{f(t+h) - f(t)}{h} \quad \forall h > 0, \tag{13.14}$$

$$f'(t) \ge \frac{f(t+h) - f(t)}{h} \quad \forall h < 0. \tag{13.15}$$

Taking $h = 1 + |t|$ in (13.14) and $h = -1 - |t|$ in (13.15) we obtain for a.e. $t \in \mathbf{R}$

$$|f'(t)| \le \frac{f(t+h)}{|h|} \le \frac{c}{1+|t|}\big(1 + |t|^p + |h|^p\big) \le c\big(1 + |t|^{p-1}\big),$$

from which the conclusion follows easily. □

For a systematic study of the properties of quasi-convex functions, see [186], [2], [177], [185], and [116]. The main interest of the notion of quasi-convexity consists in its relation with the lower semicontinuity of integral functionals, in the sense specified by the following theorem.

Theorem 13.2.1. *Let $p \ge 1$ and let $f : \Omega \times \mathbf{R}^m \times \mathbf{R}^{mn} \to \mathbf{R}$ be a Carathéodory integrand such that*

$$0 \le f(x, s, z) \le c\big(a(x) + |s|^p + |z|^p\big) \tag{13.16}$$

for all $(x, s, z) \in \Omega \times \mathbf{R}^m \times \mathbf{R}^{mn}$, where $c \geq 0$ is a constant, and $a \in L^1(\Omega)$. Then the following conditions are equivalent:

(i) *for a.e. $x \in \Omega$ and every $s \in \mathbf{R}^m$ the function $f(x, s, \cdot)$ is quasi-convex;*

(ii) *the functional F defined by* (13.11) *is sequentially lower semicontinuous on $W^{1,p}(\Omega; \mathbf{R}^m)$ with respect to its weak topology.*

PROOF. We give here the proof only in the basic case $f = f(z)$, where

$$0 \leq f(z) \leq c\big(1 + |z|^p\big);$$

the proof of the general case can be found in the references mentioned above.

We start by proving that the quasi-convexity condition (i) implies the lower semicontinuity condition (ii). We have to prove that

$$F(u) \leq \liminf_{h \to +\infty} F(u_h) \tag{13.17}$$

whenever $u_h \to u$ weakly in $W^{1,p}(\Omega; \mathbf{R}^m)$. This will be done in three steps.

First step. Inequality (13.17) *holds when u is affine and $u_h = u$ on $\partial\Omega$.* In fact, in this case Du is a constant matrix z and $u_h - u \in W_0^{1,p}(\Omega; \mathbf{R}^m)$ so that, by definition of quasi-convexity, we get

$$F(u) = |\Omega| f(z) \leq \int_\Omega f\big(z + D(u_h - u)\big)\, dx = F(u_h),$$

hence (13.17).

Second step. Inequality (13.17) *holds when u is affine.* We use a slicing method near the boundary of Ω (see also the proof of Proposition 11.2.3). Let Ω_0 be a compact subset of Ω, let $R = \frac{1}{2}\mathrm{dist}(\Omega_0, \partial\Omega)$, and let $N \in \mathbf{N}$. For every integer $i = 1, \ldots, N$ define

$$\Omega_i = \left\{ x \in \Omega \,:\, \mathrm{dist}(x, \Omega_0) < \frac{iR}{N} \right\}$$

and let φ_i be a smooth function such that

$$0 \leq \varphi_i \leq 1, \quad \varphi_i = 0 \text{ on } \Omega \setminus \Omega_i, \quad \varphi_i = 1 \text{ on } \Omega_{i-1}, \quad |D\varphi_i| \leq \frac{2N}{R}.$$

Finally take

$$v_{i,h} = u + \varphi_i \cdot (u_h - u).$$

For every $i = 1, \ldots, N$ we have $v_{i,h} \to u$ weakly in $W^{1,p}(\Omega; \mathbf{R}^m)$ as $h \to +\infty$, and $v_{i,h} = u$ on $\partial\Omega$ so that, by step 1,

$$\begin{aligned} F(u) &\leq F(v_{i,h}) \\ &= \int_{\Omega_{i-1}} f(Du_h)\, dx + \int_{\Omega_i \setminus \Omega_{i-1}} f(Dv_{i,h})\, dx + \int_{\Omega \setminus \Omega_i} f(Du)\, dx \end{aligned}$$

$$
\begin{aligned}
&\le F(u_h) + c\int_{\Omega_i\setminus\Omega_{i-1}} \bigl(1+|Dv_{i,h}|^p\bigr)\,dx + c\int_{\Omega\setminus\Omega_i} \bigl(1+|Du|^p\bigr)\,dx\\
&\le F(u_h) + c\int_{\Omega_i\setminus\Omega_{i-1}} \Bigl(1+|Du_h|^p+|Du|^p+|u_h-u|^p\frac{N^p}{R^p}\Bigr)dx\\
&\qquad + c\int_{\Omega\setminus\Omega_i} \bigl(1+|Du|^p\bigr)\,dx\\
&\le F(u_h) + c\int_{\Omega_i\setminus\Omega_{i-1}} \left(|Du_h|^p+|u_h-u|^p\frac{N^p}{R^p}\right)dx + c\int_{\Omega\setminus\Omega_0} \bigl(1+|Du|^p\bigr)\,dx.
\end{aligned}
$$

Summing for $i=1,\dots,N$ and dividing by N gives

$$
F(u) \le F(u_h) + \frac{c}{N}\int_{\Omega\setminus\Omega_0} \left(|Du_h|^p+|u_h-u|^p\frac{N^p}{R^p}\right)dx + c\int_{\Omega\setminus\Omega_0} \bigl(1+|Du|^p\bigr)\,dx.
$$

Passing to the limit as $h\to+\infty$ and taking into account that (u_h) is bounded in $W^{1,p}(\Omega;\mathbf{R}^m)$ yields

$$
F(u) \le \liminf_{h\to+\infty} F(u_h) + \frac{c}{N} + c\int_{\Omega\setminus\Omega_0} \bigl(1+|Du|^p\bigr)\,dx.
$$

Now, to achieve the proof of step 2, it is enough to pass to the limit as $N\to+\infty$ and as $\Omega_0\uparrow\Omega$.

Third step. Inequality (13.17) *holds in the general case* $u\in W^{1,p}(\Omega;\mathbf{R}^m)$. Let us fix $\varepsilon>0$ and let w be a piecewise affine function such that $\|u-w\|_{W^{1,p}}<\varepsilon$. In particular, there exist open sets Ω_i and constant matrices z_i such that $Dw=z_i$ on Ω_i. Setting

$$
w_{i,h}(x) = u_h(x) - u(x) + z_i x \quad \text{on } \Omega_i,
$$

we have $w_{i,h}\to z_i x$ weakly in $W^{1,p}(\Omega_i;\mathbf{R}^m)$ so that, by step 2,

$$
\int_{\Omega_i} f(z_i)\,dx \le \liminf_{h\to+\infty}\int_{\Omega_i} f(Dw_{i,h})\,dx.
$$

By Lemma 13.2.1 we get

$$
\begin{aligned}
&\left|F(u_h) - \sum_i \int_{\Omega_i} f(Dw_{i,h})\,dx\right| \le \sum_i \int_{\Omega_i} \bigl|f(Du_h)-f(Dw_{i,h})\bigr|\,dx\\
&\le c\sum_i \int_{\Omega_i} |Du_h-Dw_{i,h}|\bigl(1+|Du_h|^{p-1}+|Dw_{i,h}|^{p-1}\bigr)\,dx\\
&\le c\sum_i \int_{\Omega_i} |Du-Dw|\bigl(1+|Du_h|^{p-1}+|Du-Dw|^{p-1}\bigr)\,dx\\
&\le c\left(\int_\Omega |Du-Dw|^p\,dx\right)^{1/p}\left(\int_\Omega \bigl[1+|Du_h|^p+|Du-Dw|^p\bigr]\,dx\right)^{1-1/p}\\
&\le c\varepsilon.
\end{aligned}
$$

Analogously,

$$\left| F(u) - \sum_i \int_{\Omega_i} f(z_i)\,dx \right| \le c\varepsilon.$$

Therefore,

$$\begin{aligned} F(u) &\le c\varepsilon + \sum_i \int_{\Omega_i} f(z_i)\,dx \\ &\le c\varepsilon + \sum_i \liminf_{h\to+\infty} \int_{\Omega_i} f(Dw_{i,h})\,dx \\ &\le c\varepsilon + \liminf_{h\to+\infty} \sum_i \int_{\Omega_i} f(Dw_{i,h})\,dx \\ &\le 2c\varepsilon + \liminf_{h\to+\infty} F(u_h) \end{aligned}$$

and the conclusion follows by taking the limit as $\varepsilon \to 0^+$.

We prove now that lower semicontinuity condition (ii) implies quasiconvexity of f. The proof is based on the following result, which is, as said in Example 2.4.2 about oscillation phenomena, a straightforward consequence of a general ergodic theorem. (See also Lemma 12.3.1.)

Lemma 13.2.2. *Let Q be any open cube in $\mathbf{R}^n$ of size $L > 0$, v a function in $L^p(Q, \mathbf{R}^m)$ and $\tilde{v}$ its Q-periodic extension, i.e., the function of $L^p_{loc}(\mathbf{R}^n, \mathbf{R}^m)$ defined by*

$$\tilde{v}(x) = v(x - z) \text{ if } x \in Q + z,\ z \in L\mathbf{Z}^n.$$

Then the sequence $(v_h)_{h\in\mathbf{N}}$ defined by $v_h(x) = \tilde{v}(hx)$, weakly converges in $L^p_{loc}(\mathbf{R}^n, \mathbf{R}^m)$ (weak if $p = +\infty$) to its mean value $\frac{1}{\text{meas}(Q)} \int_Q v(x)\,dx$.*

The proof is a straightforward consequence of Proposition 13.2.1 that we establish below because of its own interest.

Let now z be an $m \times n$ matrix, l_z the linear function defined by $l_z(x) = zx$, $u \in \mathbf{C}^1_0(Q, \mathbf{R}^m)$, $\tilde{u}$ its Q-periodic extension, and set for every $x \in \mathbf{R}^n$, $u_h(x) = \tilde{u}(hx)/h$. It is easily seen that u_h strongly converges to 0 in $L^p_{loc}(\mathbf{R}^n, \mathbf{R}^m)$. On the other hand, according to Lemma 13.2.2, Du_h weakly converges to

$$\frac{1}{\text{meas}(Q)} \int_Q Du(x)\,dx = 0$$

in $L^p_{loc}(\mathbf{R}^m, \mathbf{R}^{mn})$. Therefore $u_h + l_z$ weakly converges to l_z in $W^{1,p}(Q, \mathbf{R}^m)$ and, by hypothesis (ii),

$$\liminf_{h\to+\infty} \int_Q f(Du_h + z)\,dx \ge \int_Q f(z)\,dz = \text{meas}(Q) f(z). \tag{13.18}$$

A change of scale and the periodicity assumption on $\tilde{u}$ gives

$$\frac{1}{\text{meas}(Q)}\int_Q f(Du_h+z)\,dx = \frac{1}{\text{meas}(hQ)}\int_{hQ} f(D\tilde{u}+z)\,dx$$
$$= \frac{1}{\text{meas}(Q)}\int_Q f(Du+z)\,dx$$

so that (13.18) yields

$$\frac{1}{\text{meas}(Q)}\int_Q f(Du+z)\,dx \geq f(z),$$

which completes the proof. □

Remark 13.2.3. When $p=+\infty$ the result of Theorem 13.2.1 still holds if we substitute the weak topology with the weak* topology of $W^{1,\infty}(\Omega;\mathbf{R}^m)$, and condition (13.16) with

$$0 \leq f(x,s,z) \leq \alpha(x,|s|,|z|) \tag{13.19}$$

for all $(x,s,z)\in\Omega\times\mathbf{R}^m\times\mathbf{R}^{mn}$, where $\alpha(x,t,\tau)$ is a function which is summable in x and increasing in t and τ.

Remark 13.2.4. From what was presented above, it follows that the implications

$$f \text{ convex} \quad\Rightarrow\quad f \text{ polyconvex} \quad\Rightarrow\quad f \text{ quasi-convex} \quad\Rightarrow\quad f \text{ rank-one convex}$$

hold true. We stress the fact that, as shown in Theorem 13.2.1, quasi-convexity is the right property to use when dealing with the lower semicontinuity of integral functionals of the calculus of variations. However, due to its intrinsic definition, it is not easy to work with quasi-convex functions, while polyconvexity and rank-one convexity conditions are much more explicit: in many cases, the sufficiency of the first and the necessity of the second are of great help.

None of the implications above can be reversed; this can be seen very easily for the first one (indeed $|\det z|$ is a polyconvex function which is not convex), whereas more delicate counterexamples are needed for the remaining ones. See the book by Dacorogna [116] and the paper by Sverak [213] for details concerning these topics. The only problem that remains open in the study of the implications above is the equivalence between quasi-convexity and rank-one convexity in the case $m=n=2$: neither counterexamples nor proofs of the equivalence are known.

We state now the ergodic theorem stated in Lemma 13.2.2, which is a generalization of the convergence result described in Example 2.4.2 about oscillations phenomena, and a particular case of Lemma 12.3.1.

Proposition 13.2.1. *With the notation of Lemma* 12.3.1, *let Q be a cube in $\mathbf{R}^n$ of the form $\Pi_{i=1}^n(a_i,a_i+L)$ for some $L>0$, and $\mathcal{A}:\mathcal{B}_b(\mathbf{R}^n)\to\mathbf{R}$ satisfying*

(i) *$\mathcal{A}_{A\cup B}=\mathcal{A}_A+\mathcal{A}_B$ for every disjoint set of $\mathcal{B}_b(\mathbf{R}^n)$,*

(ii) *$\mathcal{A}_{A+z}=\mathcal{A}_A$ for every set of $\mathcal{B}_b(\mathbf{R}^n)$ and every z in $L\mathbf{Z}^n$.*

Then, for any bounded Borel convex subset B of $\mathbf{R}^n$,

$$\lim_{h\to+\infty} \frac{\mathcal{A}_{hB}}{\operatorname{meas}(hB)} = \frac{1}{\operatorname{meas}(Q)}\mathcal{A}_Q.$$

Consequently, if v belongs to $L^p(Q,\mathbf{R}^m)$ and $\tilde{v}$ denotes its Q-periodic extension, then v_h defined by $v_h(x) = \tilde{v}(hx)$, $h \in \mathbf{N}$, weakly converges in $L^p_{loc}(\mathbf{R}^n,\mathbf{R}^m)$ (weakly if $p = +\infty$) to its mean value $\frac{1}{\operatorname{meas}(Q)}\int_Q v(x)\,dx$.*

Proof. When $B = Q$, the first assertion is a straightforward consequence of the following decomposition:

$$hQ = \bigcup_{z\in L\mathbf{Z}^n\cap hQ} (z+Q).$$

For the proof of the general case, see [106]. We are going to the proof of the second assertion. Let U be an open ball of $\mathbf{R}^n$; then v_h is bounded in $L^p(U,\mathbf{R}^m)$. Indeed

$$\frac{1}{\operatorname{meas}(U)}\int_U |v_h|^p\,dx = \frac{1}{\operatorname{meas}(hU)}\int_{hU} |\tilde{v}|^p\,dx,$$

which converges thanks to the first assertion. Therefore, in the case when $p > 1$, there exists a subsequence of $(v_h)_{h\in\mathbf{N}}$ (not relabeled) and $u \in L^p(U,\mathbf{R}^m)$ such that $v_h \rightharpoonup u$ weakly in $L^p(U,\mathbf{R}^m)$. To identify the weak limit u, we work in the space of measures $\mathbf{M}(U,\mathbf{R}^m)$. Obviously $v_h \rightharpoonup u\,\mathcal{L}^n\lfloor U$ weakly in $\mathbf{M}(U,\mathbf{R}^m)$. For a.e. x_0 in U, ρ in $(0,1)\setminus N$ where N is a countable subset of $\mathbf{R}$ (see Lemma 4.2.2), according to the first assertion, we have

$$\begin{aligned}
u(x_0) &= \lim_{\rho\to 0} \frac{1}{\operatorname{meas}(B_\rho(x_0))}\int_{B_\rho(x_0)} u(x)\,dx \\
&= \lim_{\rho\to 0}\lim_{h\to+\infty} \frac{1}{\operatorname{meas}(B_\rho(x_0))}\int_{B_\rho(x_0)} v_h(x)\,dx \\
&= \lim_{\rho\to 0}\lim_{h\to+\infty} \frac{1}{\operatorname{meas}(hB_\rho(x_0))}\int_{hB_\rho(x_0)} \tilde{v}(x)\,dx \\
&= \frac{1}{\operatorname{meas}(Q)}\int_Q v(x)\,dx.
\end{aligned}$$

It remains to treat the case $p = 1$. We establish the uniform integrability of the sequence $(v_h)_{h\in\mathbf{N}}$ in $L^1(U,\mathbf{R}^m)$; the conclusion will follow by the Dunford–Pettis theorem, Theorem 2.4.5, and by the above procedure for identifying the weak limit. We now use a truncation argument. Let $\delta > 0$ intended to go to $+\infty$ and set

$$\begin{cases}
v_{h,\delta} = v_h \wedge \delta \vee (-\delta), \\
v_\delta = v \wedge \delta \vee (-\delta), \\
w_{h,\delta} = v_h - v_{h,\delta}, \\
w_\delta = v - v_\delta.
\end{cases}$$

Note that $v_{h,\delta} = (v_\delta)_h$. For every Borel subset A of U, we have

$$\int_A |v_h|\,dx \le \int_U |w_{h,\delta}|\,dx + \int_A |v_{h,\delta}|\,dx. \tag{13.20}$$

On the other hand, according to the first assertion,

$$\lim_{h\to+\infty}\int_U |w_{h,\delta}|\,dx = \frac{1}{\text{meas}(Q)}\int_Q |w_\delta|\,dx.$$

Then, given $\varepsilon > 0$, since there exists δ large enough such that

$$\frac{1}{\text{meas}(Q)}\int_Q |w_\delta|\,dx < \frac{\varepsilon}{4},$$

there exists $h(\varepsilon) \in \mathbf{N}$ such that

$$\sup_{h\geq h(\varepsilon)}\int_U |w_{h,\delta}|\,dx < \frac{\varepsilon}{2}. \tag{13.21}$$

Collecting (13.20) and (13.21), we deduce

$$\sup_{h\geq h(\varepsilon)}\int_A |v_h|\,dx < \frac{\varepsilon}{2} + \delta\ \text{meas}(A)$$

and the conclusion follows by taking $\text{meas}(A) < \frac{\varepsilon}{2\delta}$. □

Remark 13.2.5. Proposition 13.2.1 is a particular case, in a deterministic setting, of the ergodic theorem for spatial processes established by Nguyen and Zessin which generalizes the famous Birkhoff ergodic theorem. For a proof, see [194]. For an application to homogenization, see [106]. For a systematic study of ergodic theorems, see [165], and for applications to homogenization, see [119], [81], [179], [1], and references therein. For an application to one-dimensional models in fracture mechanics, see Theorem 14.2.2.

13.3 Lower semicontinuity for functionals defined on the space of measures

In this subsection, we establish a semicontinuity result for integral functionals defined on the space $\mathbf{M}(\Omega, \mathbf{R}^m)$. The set Ω is a separable locally compact, metrizable, σ-compact topological space, for example, $\mathbf{R}^N$ or an open subset of $\mathbf{R}^N$. As a consequence, we recover, in a convex setting, some lower semicontinuity results concerning integral functionals defined on $BV(\Omega)$ obtained in Section 11.3.

Let $f : \mathbf{R}^m \to [0, +\infty]$ be a lower semicontinuous convex function. The recession function of a quasi-convex function was defined in Theorem 11.3.1. For the convex function f, it is easily seen that the limit

$$\lim_{t\to+\infty}\frac{f(w+ta)}{t}$$

exists and does not depend on the choice of w. This limit then coincides with the recession function defined in Theorem 11.3.1. More precisely, we define

$$f^\infty(a) := \lim_{t\to+\infty}\frac{f(w+ta)}{t}$$

(when $f \equiv +\infty$ we set $f^\infty \equiv +\infty$). It is straightforward to show that f^∞ is convex, lower semicontinuous, and is positively homogeneous of degree one, that is, satisfies $f^\infty(ta) = tf^\infty(a)$ for all $t \in \mathbf{R}^+$ and all a in $\mathbf{R}^m$.

Theorem 13.3.1. *Let $f : \mathbf{R}^m \to [0, +\infty]$ be a lower semicontinuous convex function and μ be a Borel measure in $\mathbf{M}^+(\Omega)$. Then the integral functional defined on $\mathbf{M}(\Omega, \mathbf{R}^m)$ by*

$$F(\lambda) = \int_\Omega f\left(\frac{d\lambda}{d\mu}\right) d\mu + \int_\Omega f^\infty\left(\frac{d\lambda^s}{d|\lambda^s|}\right) d|\lambda^s|,$$

where $\lambda = d\lambda/d\mu \cdot \mu + d\lambda^s$ is the Radon–Nikodym decomposition of λ with respect to μ, is lower semicontinuous for the weak convergence in $\mathbf{M}(\Omega, \mathbf{R}^m)$.

Proof. Our strategy is to apply localization Lemma 4.2.2. Since f is lower semicontinuous and convex, there exist $a_h \in \mathbf{R}$ and $b_h \in \mathbf{R}^m$ such that (see Theorem 9.3.1)

$$f(s) = \sup\{a_h + b_h \cdot s \ : \ h \in \mathbf{N}\} \qquad \forall s \in \mathbf{R}^m.$$

Let us set, for all $h \in \mathbf{N}$ and all Borel subsets B of Ω,

$$\begin{cases} f_h(s) = [a_h + b_h \cdot s]^+, \\ F_h(\lambda, B) = \displaystyle\int_B f_h\left(\frac{d\lambda}{d\mu}\right) d\mu + \int_B f_h^\infty\left(\frac{d\lambda^s}{d|\lambda^s|}\right) d|\lambda^s|. \end{cases}$$

Clearly $f = \sup\{f_h \ : \ h \in \mathbf{N}\}$. Let $\lambda \in \mathbf{M}(\Omega, \mathbf{R}^m)$ and let $\mathcal{N}$ be a μ-measurable subset of Ω where λ^s is concentrated. We moreover set $\nu = \mu + |\lambda^s|$ and

$$\tilde f(x) = \begin{cases} f\left(\dfrac{d\lambda}{d\mu}(x)\right) & \text{if } x \in \Omega \setminus \mathcal{N}, \\ f^\infty\left(\dfrac{d\lambda^s}{d|\lambda^s|}(x)\right) & \text{if } x \in \mathcal{N}, \end{cases}$$

$$\tilde f_h(x) = \begin{cases} f_h\left(\dfrac{d\lambda}{d\mu}(x)\right) & \text{if } x \in \Omega \setminus \mathcal{N}, \\ f_h^\infty\left(\dfrac{d\lambda^s}{d|\lambda^s|}(x)\right) & \text{if } x \in \mathcal{N}. \end{cases}$$

We have $F(\lambda) = \int_\Omega \tilde f \, d\nu$, $F_h(\lambda) = \int_\Omega \tilde f_h \, d\nu$ and, according to Lemma 4.2.2,

$$F(\lambda) = \sup\left\{\sum_{i\in I} F_i(\lambda, A_i) \ : \ A_i \text{ disjoint open subsets of } \Omega\right\}.$$

To conclude, it suffices to notice that for all $h \in \mathbf{N}$ and all open subsets A of Ω, $\lambda \mapsto F_h(\lambda, A)$ is (sequentially) lower semicontinuous on $\mathbf{M}(\Omega, \mathbf{R}^m)$. □

Remark 13.3.1. A more general semicontinuity result has been established in [74] for functionals of the form

$$F(\lambda) = \int_\Omega f\left(\frac{d\lambda}{d\mu}\right) d\mu + \int_{\Omega\setminus A_\lambda} f^\infty(\lambda^s) + \int_\Omega g\,(\lambda(x))\; d\mathcal{H}^0, \tag{13.22}$$

where $f : \mathbf{R}^m \to [0, +\infty]$ is assumed to be convex and lower semicontinuous, $g : \mathbf{R}^m \to [0, +\infty]$ is subadditive and lower semicontinuous, $f(0) = g(0) = 0$, and $f^\infty = g^0$, where

$$g^0(s) = \lim_{t\to 0^+} \frac{g(ts)}{t}.$$

The set A_λ is the subset of the atoms of λ and $\lambda(x)$ is the value $\lambda(\{x\})$.

The result of Remark 13.3.1 is complete in the following sense: every functional lower semicontinuous on $\mathbf{M}(\Omega, \mathbf{R}^m)$ for the weak convergence has an integral representation of the form (13.22) provided that it satisfies a local property: a functional F defined on $\mathbf{M}(\Omega, \mathbf{R}^m)$ is said to be local iff

$$\lambda_1 \perp \lambda_2 \quad \Rightarrow \quad F(\lambda_1 + \lambda_2) = F(\lambda_1) + F(\lambda_2).$$

The following theorem is established in Bouchitté and Buttazzo [75].

Theorem 13.3.2. *Let $F : \mathbf{M}(\Omega, \mathbf{R}^m) \to [0, +\infty]$ be a local lower semicontinuous functional for the weak convergence of $\mathbf{M}(\Omega, \mathbf{R}^m)$. Then F is of the form*

$$F(\lambda) = \int_\Omega f\left(x, \frac{d\lambda}{d\mu}\right) d\mu + \int_{\Omega\setminus A_\lambda} f^\infty\left(x, \frac{d\lambda^s}{d|\lambda^s|}\right) d|\lambda^s| + \int_{A_\lambda} g(x, \lambda(x))\, d\mathcal{H}^0,$$

where $f, g : \Omega \times \mathbf{R}^m \to [0, +\infty]$ are two Borel functions satisfying

(a) *$f(x, \cdot)$ is convex, lower semicontinuous on $\mathbf{R}^m$ and $f(x, 0) = 0$ for μ-a.e. $x \in \Omega$;*

(b) *g is lower semicontinuous on $\Omega \times \mathbf{R}^m$, $g(x, \cdot)$ is subadditive on $\mathbf{R}^m$, and $g(x, 0) = 0$ for all $x \in \Omega$;*

(c) *$f^\infty = g^0$ on $(\Omega \setminus N) \times \mathbf{R}^m$ where N is a countable subset of Ω.*

Note that the lower semicontinuity hypothesis does not generally imply the convexity of F as it is the case, however, for functionals defined on $L^p(\Omega)$. Indeed, consider the functional

$$F(\lambda) = \int_\Omega \left|\frac{d\lambda}{d\mu}\right|^2 d\mu + \mathcal{H}^0(A_\lambda) + \chi_{\{\lambda^s=0 \text{ on } \Omega\setminus A_\lambda\}}(\lambda).$$

It is easily seen that F is local and lower semicontinuous but nonconvex because of the term $\mathcal{H}^0(A_\lambda)$ and that

$$\chi_{\{\lambda^s=0 \text{ on } \Omega\setminus A_\lambda\}}(\lambda) = \int_{\Omega\setminus A_\lambda} f^\infty\left(x, \frac{d\lambda^s}{d|\lambda^s|}\right) d|\lambda^s|,$$

where $f(x, s) = |s|^2$ (see Bouchitté and Buttazzo [76], Ambrosio and De-Giorgi [122]).

When F is convex, its integral representation was earlier established in Ambrosio and Buttazzo [19].

Theorem 13.3.3. *Let $F : \mathbf{M}(\Omega, \mathbf{R}^m) \to [0, +\infty]$ be a convex, local, and lower semicontinuous functional for the weak convergence of $\mathbf{M}(\Omega, \mathbf{R}^m)$. Then F is of the form*

$$F(\lambda) = \int_\Omega f\left(x, \frac{d\lambda}{d\mu}\right) d\mu + \int_\Omega f^\infty\left(x, \frac{d\lambda^s}{d|\lambda^s|}\right) d|\lambda^s|,$$

where $f : \Omega \times \mathbf{R}^n \to \mathbf{R}$ is a Borel function such that $f(x, \cdot)$ is convex in $\mathbf{R}^m$ for μ-a.e. $x \in \Omega$ and $f^\infty(x, \cdot)$ is the recession function of $f(x, \cdot)$.

13.4 Functionals with linear growth: Lower semicontinuity in BV and SBV

13.4.1 Lower semicontinuity and relaxation in BV

The main objective of this section is, in a convex situation, to provide an alternative proof of the general relaxation result established in Theorem 11.3.1 for integral functionals defined in $L^1(\Omega)$ by

$$F(u) = \begin{cases} \displaystyle\int_\Omega f(\nabla u)\, dx \text{ if } u \in W^{1,1}(\Omega), \\ +\infty \text{ otherwise,} \end{cases} \quad G(u) = \begin{cases} \displaystyle\int_\Omega f(\nabla u)\, dx \text{ if } u \in W_0^{1,1}(\Omega), \\ +\infty \text{ otherwise,} \end{cases}$$

when Ω is an open bounded subset of $\mathbf{R}^N$ with a Lipschitz boundary Γ, and f is convex and satisfies suitable linear growth conditions. Let us first state the following corollary of Theorem 13.3.1 where we do not assume growth conditions on f.

Proposition 13.4.1. *Let $Du = \nabla u\, \mathcal{L}^N \lfloor \Omega + D^s u$ be the Lebesgue–Nykodym decomposition of the measure Du and $f : \mathbf{R}^N \to [0, +\infty]$ a lower semicontinuous convex function. Then, the integral functional defined on $BV(\Omega)$ by*

$$F(u) = \int_\Omega f(\nabla u)\, dx + \int_\Omega f^\infty\left(\frac{D^s u}{|D^s u|}\right) |D^s u|$$

is lower semicontinuous for the weak convergence of $BV(\Omega)$.

Proof. Take $\mu = \mathcal{L}^N \lfloor \Omega$, $\lambda = Du$, and apply Theorem 13.3.1 to the functional defined by

$$\tilde{F}(Du) := \int_\Omega f(\nabla u)\, dx + \int_\Omega f^\infty\left(\frac{D^s u}{|D^s u|}\right) |D^s u|.$$

Indeed, $u_n \rightharpoonup u$ weakly in $BV(\Omega)$ implies $Du_n \rightharpoonup Du$ weakly in $\mathbf{M}(\Omega, \mathbf{R}^N)$. □

In what follows, we denote the integral

$$\int_\Omega f^\infty\left(\frac{D^s u}{|D^s u|}\right) |D^s u|$$

by $\int_\Omega f^\infty(D^s u)$ and the functional F by

$$F(u) := \int_\Omega f(Du).$$

We consider a convex function $f : \mathbf{R}^N \to \mathbf{R}^+$ satisfying for all a in $\mathbf{R}^N$ the growth condition

$$0 \leq f(a) \leq \beta(1 + |a|) \tag{13.23}$$

for some positive constant β. We recall that the Fenchel transform of f is the function $f^* : \mathbf{R}^N \to [-\infty, +\infty]$ defined for all b in $\mathbf{R}^N$ by

$$f^*(b) = \sup_{a \in \mathbf{R}^N} \{b.a - f(a)\}.$$

It is well known (see Section 9.3) that f^* is a $\mathbf{R} \cup \{+\infty\}$-value convex function, proper and lower semicontinuous. According to the growth condition (13.23), a straightforward calculation shows that its domain K is contained in the closed ball $\overline{\mathbf{B}(0, \beta)}$ of $\mathbf{R}^N$, with radius β and centered at 0 and that for all $b \in K$, $f^*(b) \geq -\beta$. We assume moreover that f^* is bounded from above on its domain: there exists $C > 0$ such that

$$\forall b \in K,\ f^*(b) \leq C. \tag{13.24}$$

Approximating Theorem 10.1.2 may be generalized as follows.

Theorem 13.4.1. *Let f be a nonnegative convex function satisfying* (13.23) *and* (13.24), *then the space $\mathbf{C}^\infty(\Omega) \cap BV(\Omega)$ is dense in $BV(\Omega)$ equipped with the intermediate convergence associated with f. Namely, for all u in $BV(\Omega)$, there exists u_n in $\mathbf{C}^\infty(\Omega) \cap BV(\Omega)$ such that*

$$u_n \to u \text{ strongly in } L^1(\Omega);$$
$$\int_\Omega |Du_n|\, dx \to \int_\Omega |Du|;$$
$$\int_\Omega f(Du_n)\, dx \to \int_\Omega f(Du).$$

Proof. Proceed exactly as in the proof of Theorem 10.1.2. For a proof explained in detail, consult Temam [219]. □

With the intention of obtaining a complete description of the two lower semicontinuous envelopes of the functionals F and G, we assume a coerciveness condition on f: there exists a positive constant α such that

$$\forall a \in \mathbf{R}^N,\ \alpha(|a| - 1) \leq f(a). \tag{13.25}$$

Theorem 13.4.2. *Let Ω be a Lipschitz bounded open subset of $\mathbf{R}^N$ with boundary Γ and f a nonnegative and convex function satisfying* (13.23), (13.24), *and* (13.25). *Then the lsc*

envelopes $\overline{F}$ and $\overline{G}$ of the functionals F and G in $L^1(\Omega)$ equipped with its strong topology are defined by

$$\overline{F}(u) = \begin{cases} \displaystyle\int_\Omega f(\nabla u)\,dx + \int_\Omega f^\infty(D^s u) \text{ if } u \in BV(\Omega), \\ +\infty \text{ otherwise;} \end{cases}$$

$$\overline{G}(u) = \begin{cases} \displaystyle\int_\Omega f(\nabla u)\,dx + \int_\Omega f^\infty(D^s u) + \int_\Gamma f^\infty(\gamma_0(u)\nu)\,d\mathcal{H}^{N-1} \text{ if } u \in BV(\Omega), \\ +\infty \text{ otherwise.} \end{cases}$$

PROOF. We argue as in the proof of Proposition 11.3.2. For the function F, we must establish that for all u in $L^1(\Omega)$,

$$\text{if } u_n \to u \text{ in } L^1(\Omega), \text{ then } \overline{F}(u) \le \liminf_{n\to+\infty} F(u_n),$$
$$\text{there exists } u_n \text{ converging to } u \text{ in } L^1(\Omega), \text{ such that } \overline{F}(u) \ge \limsup_{n\to+\infty} F(u_n).$$

These two assertions are straightforward consequences of Proposition 13.4.1 and Theorem 13.4.1. Indeed, let $u_n \to u$ strongly in $L^1(\Omega)$, such that $\liminf_{n\to+\infty} F(u_n) < +\infty$. From (13.25), for a nonrelabeled subsequence, $u_n \rightharpoonup u$ in $BV(\Omega)$ and, according to Proposition 13.4.1,

$$\overline{F}(u) \le \liminf_{n\to+\infty} F(u_n).$$

On the other hand, for $u \in BV(\Omega)$, Theorem 13.4.1 provides a sequence of functions u_n in $\mathbf{C}^\infty(\Omega) \cap BV(\Omega)$ such that $\overline{F}(u) \ge \limsup_{n\to+\infty} F(u_n)$.

For the function G, the proof is exactly that of Proposition 11.3.2. Let us recall that it suffices to apply the previous result related to F after enlarging the set Ω to obtain the first assertion and to approach Ω from below to derive the second. □

13.4.2 Compactness and lower semicontinuity in SBV

The following result, due to Ambrosio, is a key tool in the so-called direct method in the calculus of variations when working with integral functionals defined in $SBV(\Omega)$ equipped with the strong topology of $L^1(\Omega)$ (see chapter 14). In this section, Ω denotes an open bounded subset of $\mathbf{R}^N$.

Theorem 13.4.3. *Let $(u_n)_{n\in\mathbf{N}}$ be a sequence in $SBV(\Omega)$ satisfying for $p > 1$,*

$$\sup_{n\in\mathbf{N}} \left\{ \|u_n\|_\infty + \int_\Omega |\nabla u_n|^p\,dx + \mathcal{H}^{N-1}(S_{u_n}) \right\} < +\infty.$$

Then there exists a subsequence $(u_{n_k})_{k\in\mathbf{N}}$ and a function u in $SBV(\Omega)$, such that

$$u_{n_k} \to u \text{ strongly in } L^1_{loc}(\Omega);$$

$$\nabla u_{n_k} \rightharpoonup \nabla u \text{ weakly in } L^p(\Omega, \mathbf{R}^N);$$

$$\mathcal{H}^{N-1}(S_u) \le \liminf_{k\longrightarrow+\infty} \mathcal{H}^{N-1}(S_{u_{n_k}}).$$

Moreover, the Lebesgue part and the jump part of the derivatives converge separately. More precisely, $\nabla u_{n_k} \rightharpoonup \nabla u$ *weakly in* $L^1(\Omega)$ *and* $Ju_{n_k} \rightharpoonup Ju$ *weakly in* $\mathbf{M}(\Omega, \mathbf{R}^N)$.

PROOF. In what follows C denotes various constants which do not depend on n. Thanks to inequalities

$$\int_\Omega |\nabla u_n|^p \, dx \leq C \text{ and } |u_n|_\infty \leq C,$$

we have $\|u_n\|_{BV(\Omega)} \leq C$. Indeed, according to Remark 10.3.4, for $\mathcal{H}^{N-1}$ a.e. x in S_{u_n}, $|u_n^+(x)| \leq \|u_n\|_{L^\infty(\Omega)} \leq C$ and $|u_n^-(x)| \leq \|u_n\|_{L^\infty(\Omega)} \leq C$. From the compactness of the embedding of $BV(U)$ into $L^1(U)$, for each regular open subset of Ω, there exists a subsequence $(u_{n_k})_{k\in\mathbf{N}}$ and u in $BV(\Omega)$ such that

$$\begin{aligned} &u_{n_k} \rightharpoonup u \text{ weakly in } BV(\Omega), \\ &u_{n_k} \to u \text{ strongly in } L^1_{loc}(\Omega). \end{aligned}$$

Let us show that $u \in SBV(\Omega)$. Since $u_{n_k} \in SBV(\Omega)$, according to Theorem 10.5.1, there exists a Borel measure μ_{n_k} in $\mathbf{M}(\Omega \times \mathbf{R}, \mathbf{R}^N)$ such that for all $\Phi \in \mathbf{C}_c^1(\Omega, \mathbf{R}^N)$ and all $\varphi \in \mathbf{C}_0^1(\mathbf{R})$,

$$\begin{aligned} &\int_{\Omega\times\mathbf{R}} \varphi(s)\Phi(x) \, \mu_{n_k}(dx, ds) = -\int_\Omega \left(\varphi'(u_{n_k}) \, \nabla u_{n_k} \, . \, \Phi(x) + \varphi(u_{n_k}) \operatorname{div} \Phi(x)\right) dx, \\ &|\mu_{n_k}|(\Omega \times \mathbf{R}) = 2 \, H^{N-1}(S_{u_{n_k}}). \end{aligned}$$

The second equality and the hypothesis $\mathcal{H}^{N-1}(S_{u_{n_k}}) \leq C$ yield the existence of a Borel measure μ in $\mathbf{M}(\Omega \times \mathbf{R}, \mathbf{R}^N)$ and a subsequence (not relabeled) of $(\mu_{n_k})_{k\in\mathbf{N}}$, weakly converging to μ in $\mathbf{M}(\Omega \times \mathbf{R}, \mathbf{R}^N)$. On the other hand, for a further nonrelabeled subsequence, there exists $a \in L^1(\Omega, \mathbf{R}^N)$ such that ∇u_{n_k} weakly converges to a in $L^1(\Omega, \mathbf{R}^N)$. Going to the limit in the first equality, we obtain

$$\int_{\Omega\times\mathbf{R}} \varphi(s)\Phi(x) \, \mu(dx, ds) = -\int_\Omega \left(\varphi'(u) \, a \, . \, \Phi(x) + \varphi(u) \operatorname{div} \Phi(x)\right) dx,$$

which, from Theorem 10.5.1, yields $u \in SBV(\Omega)$ and $a = \nabla u$.

On the other hand, according to the lower semicontinuity of the total variation, one has

$$\begin{aligned} 2\mathcal{H}^{N-1}(S_u) &= |\mu|(\Omega \times \mathbf{R}) \\ &\leq \liminf_{k\longrightarrow+\infty} |\mu_{n_k}|(\Omega \times \mathbf{R}) \\ &= \liminf_{k\longrightarrow+\infty} 2\mathcal{H}^{N-1}(S_{u_{n_k}}). \end{aligned}$$

Finally, since Du_{n_k} weakly converges to Du in $\mathbf{M}(\Omega, \mathbf{R}^N)$ and ∇u_{n_k} weakly converges to ∇u in $L^1(\Omega, \mathbf{R}^N)$, one has $Ju_{n_k} = Du_{n_k} - \nabla u_{n_k}\mathcal{L}^N$ weakly converges to $Ju = Du - \nabla u\mathcal{L}^N$ in $\mathbf{M}(\Omega, \mathbf{R}^N)$. □

When $(u_n)_{n\in\mathbf{N}}$ is not bounded in $L^\infty(\Omega)$, we obtain the same result provided that $(u_n)_{n\in\mathbf{N}}$ be bounded in $BV(\Omega)$. Note also that the boundedness of $(\nabla u_n)_{n\in\mathbf{N}}$ in $L^p(\Omega)$

with $p > 1$, implies the equi-integrability of $(\nabla u_n)_{n\in\mathbf{N}}$. In this sense, the next theorem generalizes Theorem 13.4.3.

Theorem 13.4.4. *Let $(u_n)_n$ be a sequence in $SBV(\Omega)$ satisfying*

(i) $\sup_{n\in\mathbf{N}}\{|u_n|_{BV(\Omega)}\} < +\infty$;

(ii) *the approximate gradients ∇u_n are equi-integrable (i.e., $(\nabla u_n)_{n\in\mathbf{N}}$ is relatively compact with respect to the weak topology of $L^1(\Omega, \mathbf{R}^N)$);*

(iii) *the sequence $(\mathcal{H}^{N-1}(S_{u_n}))_{n\in\mathbf{N}}$ is bounded.*

Then there exists a subsequence $(u_{n_k})_{k\in\mathbf{N}}$ weakly converging to some $u \in SBV(\Omega)$, such that

$$u_{n_k} \to u \text{ strongly in } L^1_{loc}(\Omega);$$

$$\nabla u_{n_k} \rightharpoonup \nabla u \text{ weakly in } L^1(\Omega, \mathbf{R}^N);$$

$$Ju_{n_k} \rightharpoonup Ju \text{ weakly in } \mathbf{M}(\Omega, \mathbf{R}^N);$$

$$\mathcal{H}^{N-1}(S_u) \le \liminf_{k\to+\infty} \mathcal{H}^{N-1}(S_{u_{n_k}}).$$

Proof. According to Theorem 2.4.5, equi-integrability condition (ii) yields the existence of a subsequence of $(\nabla u_n)_{n\in\mathbf{N}}$ and a in $L^1(\Omega, \mathbf{R}^N)$ such that $\nabla u_n \rightharpoonup a$ in $L^1(\Omega, \mathbf{R}^N)$. Then we argue as in the proof of Theorem 13.4.3 and adopt the same notation. We only have to justify the convergence of

$$\int_\Omega \varphi'(u_{n_k})\, \nabla u_{n_k} \,.\, \Phi(x)\, dx$$

to

$$\int_\Omega \varphi'(u)\, \nabla u \,.\, \Phi(x)\, dx.$$

According to Egorov's theorem, since $\varphi'(u_{n_k})$ converges a.e. to $\varphi'(u)$, for all $\varepsilon > 0$ there exists a Borel subset Ω_ε of Ω such that $\mathcal{L}^N(\Omega \setminus \Omega_\varepsilon) < \varepsilon$ and $\lim_{k\to+\infty} \sup_{x\in\Omega_\varepsilon} |\varphi(u_{n_k}) - \varphi(u)| = 0$. Let us write

$$\int_\Omega \varphi'(u_{n_k})\, \nabla u_{n_k} \,.\, \Phi(x)\, dx = \int_{\Omega_\varepsilon} \varphi'(u_{n_k})\, \nabla u_{n_k} \,.\, \Phi(x)\, dx + \int_{\Omega\setminus\Omega_\varepsilon} \varphi'(u_{n_k})\, \nabla u_{n_k} \,.\, \Phi(x)\, dx.$$

Since moreover $\sup_{k\in\mathbf{N}} \int_\Omega |\nabla u_{n_k}|\, dx < +\infty$, one easily obtains that the first term in the right-hand side tends to

$$\int_{\Omega_\varepsilon} \varphi'(u)\, \nabla u \,.\, \Phi(x)\, dx.$$

Letting $\varepsilon \to 0$, the conclusion then follows, provided that we establish

$$\lim_{\varepsilon\to 0} \lim_{k\to+\infty} \int_{\Omega\setminus\Omega_\varepsilon} \varphi'(u_{n_k})\, \nabla u_{n_k} \,.\, \Phi(x)\, dx = 0,$$

which is a straightforward consequence of

$$\left| \int_{\Omega\setminus\Omega_\varepsilon} \varphi'(u_{n_k})\ \nabla u_{n_k}\ .\ \Phi(x)\ dx \right| \leq C \int_{\Omega\setminus\Omega_\varepsilon} |\nabla u_{n_k}|\ dx$$

and equi-integrability of $(\nabla u_n)_{n\in\mathbf{N}}$. □

Remark 13.4.1. If one of the two conditions (ii) and (iii) is not satisfied, the conclusion may fail. Indeed, in the Cantor–Vitali example of Section 10.4, u_n belongs to $SBV(0,1)$, weakly converges to the Cantor–Vitali function u in $BV(0,1)$ and $\mathcal{H}^0(S_{u_n}) = 0$. Nevertheless $(\nabla u_n)_{n\in\mathbf{N}}$ is not equi-integrable since $\int_{(0,1)} \nabla u_n\ dx = 1$ does not converge to $\int_{(0,1)} \nabla u\ dx = 0$. With the notation of this example, consider now v_n in $SBV(0,1)$ defined by $v_n = u_n 1_{(0,1)\setminus C_n}$. It is easily seen that v_n weakly converges to u in $BV(0,1)$ and that $\nabla v_n = 0$ is obviously equi-integrable. But $\mathcal{H}^0(S_{v_n}) = 2(2^n - 1)$ is not uniformly bounded.

Condition (iii) of Theorem 13.4.4 can be weakened by the following condition (iii)′: there exists a function $\psi : [0,+\infty) \to [0,+\infty]$ such that $\psi(t)/t \to +\infty$ as $t \to 0$ and

$$\sup_{n\in\mathbf{N}} \int_{S_{u_n}} \psi\left(|u_n^+ - u_n^-|\right)\ d\mathcal{H}^{N-1} < +\infty.$$

For a proof, consult Braides [82].

Remark 13.4.2. Theorem 13.4.4 obviously holds in the vectorial case, i.e., when the considered sequences belong to $SBV(\Omega, \mathbf{R}^m)$.

We now deal with the lower semicontinuity property of functionals of the form

$$\int_\Omega f(\nabla u)\ dx + \int_\Omega g(u^+, u^-) h(\nu_u)\ d\mathcal{H}^{N-1}\lfloor S_u,$$

where f, g, and h verify suitable conditions.

Theorem 13.4.5. *Let us consider a function $f : \mathbf{R}^N \to \mathbf{R}^+$ satisfying the De La Vallée–Poussin criterion: f is convex and*

$$\lim_{a\to+\infty} \frac{f(a)}{|a|} = +\infty.$$

Let moreover $g : \mathbf{R}\times\mathbf{R} \to \mathbf{R}^+$ be a lower semicontinuous symmetric and subadditive function, i.e.,

$$g(a,b) = g(b,a) \leq g(b,c) + g(c,a) \quad \forall\, a,\ b,\ c \in \mathbf{R},$$

and assume that $g(a,b) \geq \max(\psi(|a-b|, \delta|a-b|)$ for all $a,\ b \in \mathbf{R}$ where the function $\psi : [0,+\infty) \to [0,+\infty]$ satisfies the condition $\psi(t)/t \to +\infty$ as $t \to 0$, and δ is some positive constant.

Let finally $h : \mathbf{R}^N \to [0,+\infty)$ be a convex, even function, positively homogeneous of degree 1 and satisfying, for all $\nu \in \mathbf{R}^N$, $h(\nu) \geq c|\nu|$ for some positive constant c. Then

the functional defined in $L^1(\Omega)$ *by*

$$F(u) = \begin{cases} \displaystyle\int_\Omega f(\nabla u)\,dx + \int_\Omega g(u^+, u^-)h(\nu_u)\,d\mathcal{H}^{N-1}\lfloor S_u & \text{if } u \in SBV(\Omega), \\ +\infty & \text{otherwise} \end{cases}$$

is lower semicontinuous for the strong topology of $L^1(\Omega)$.

If the lower semicontinuous, symmetric, and subadditive function g *only satisfies the condition* $g(a, b) \geq \psi(|a - b|)$, *then, given a nonempty compact subset* K *of* $\mathbf{R}$, *the functional* $\tilde{F}$ *defined in* $L^1(\Omega)$ *by*

$$\tilde{F}(u) = \begin{cases} \displaystyle\int_\Omega f(\nabla u)\,dx + \int_\Omega g(u^+, u^-)h(\nu_u)\,d\mathcal{H}^{N-1}\lfloor S_u & \text{if } u \in SBV(\Omega) \text{ and } u(x) \in K, \\ +\infty & \text{otherwise} \end{cases}$$

is lower semicontinuous for the strong topology of $L^1(\Omega)$.

SKETCH OF THE PROOF. The proof is based on the following lemma.

Lemma 13.4.1. *Let* $g : \mathbf{R} \times \mathbf{R} \to \mathbf{R}^+$ *be a lower semicontinuous, symmetric, and subadditive function and* $h : \mathbf{R}^N \to [0, +\infty)$ *be a convex, even function, positively homogeneous of degree 1. Then*

$$\int_\Omega g(u^+, u^-)h(\nu)\,d\mathcal{H}^{N-1}\lfloor S_u \leq \liminf_{n\to+\infty} \int_\Omega g(u_n^+, u_n^-)h(\nu_{u_n}))\,d\mathcal{H}^{N-1}\lfloor S_{u_n}$$

whenever u_n, u *satisfy the thesis of Theorem* 13.4.4 *(with condition* (iii) *or* (iii)′*).*

For the proof of Lemma 13.4.1, consult Braides [82, Theorem 2.12]. Let $(u_n)_{n\in\mathbf{N}}$ be a sequence strongly converging to some u in $L^1(\Omega)$ and such that $\liminf_{n\to+\infty} F(u_n) < +\infty$. From the De La Vallée–Poussin criterion, there exists a subsequence of $(u_n)_{n\in\mathbf{N}}$ (not relabeled) such that

$$\nabla u_{n_k} \text{ weakly converges in } L^1(\Omega, \mathbf{R}^N).$$

On the other hand, according to the coercivity assumption on g,

$$\sup_{n\in\mathbf{N}} \int_\Omega |u_n^+ - u_n^-|\,d\mathcal{H}^{N-1}\lfloor S_{u_n} < +\infty.$$

The sequence $(u_n)_{n\in\mathbf{N}}$ is then bounded in $BV(\Omega)$. Moreover, from the assumption made on ψ, condition (iii′) of Remark 13.4.1 is satisfied. Thus conditions (i), (ii), and (iii)′ of Theorem 13.4.4 are fulfilled. The conclusion then follows from the convexity of f and Lemma 13.4.1. The proof of the lower semicontinuity property of the functional $\tilde{F}$ follows the same scheme. The boundedness

$$\sup_{n\in\mathbf{N}} \int_\Omega |u_n^+ - u_n^-|\,d\mathcal{H}^{N-1}\lfloor S_{u_n} < +\infty$$

is now satisfied thanks to $|u_n^+(x)| \leq \|u_n\|_{L^\infty(\Omega)} \leq C$ and $|u_n^-(x)| \leq \|u_n\|_{L^\infty(\Omega)} \leq C$ for $\mathcal{H}^{N-1}$ a.e. x in S_{u_n} (cf. Remark 10.3.4). □

Remark 13.4.3. In the vectorial case, Theorem 13.4.5 holds with the same conditions on g and h (now $g : \mathbf{R}^m \times \mathbf{R}^m \to \mathbf{R}^+$), and when $f : \mathbf{R} \times \mathbf{M}^{m\times N} \to \mathbf{R}$ is quasi-convex and satisfies the growth conditions of order $p > 1$: for all $A \in \mathbf{M}^{m\times N}$,

$$|A|^p \leq f(x, A) \leq \beta(1 + |A|^p)$$

for some positive constants α and β. For a proof, consult Ambrosio [15]. This result will be essential in Section 14.2 to establish the existence of a solution for the weak Griffith model in the framework of fracture mechanics.

Chapter 14

Application in mechanics and computer vision

14.1 Problems in pseudoplasticity

14.1.1 Introduction

This section is devoted to the study of the equilibrium of a three-dimensional elastoplastic material occupying a bounded domain $\Omega \subset \mathbf{R}^3$ as reference configuration and subjected to body and surface forces. The unknown displacement vector field u solves a minimization problem of the form

$$\inf \left\{ \int_\Omega W(\varepsilon(v))\, dx - L(v) : v \in \mathcal{A} \right\},$$

where $\varepsilon(v)$ denotes the linearized strain tensor $\varepsilon_{i,j}(v) = \frac{1}{2}(\frac{\partial u_i}{\partial x_j} + \frac{\partial u_j}{\partial x_i})$. The linear mapping $v \mapsto L(v)$ accounts for the exterior loading and is of the form

$$L(v) = \int_\Omega f \cdot v\, dx + \int_{\Gamma_1} g \cdot v\, d\mathcal{H}^2,$$

where f denotes the body forces and g the surface forces on a part Γ_1 of the boundary. The body is assumed to be clamped on $\Gamma_0 = \Gamma \setminus \overline{\Gamma}_1$, with $\mathcal{H}^2(\Gamma_0) > 0$. We will denote the space of 3×3 symmetric matrices by $\mathbf{M}_S$ and its subspace of matrices with null trace by $\mathbf{M}_S^D$, so that $\mathbf{M}_S = \mathbf{M}_S^D \oplus \mathbf{R}I$. The constitutive equation of the material is such that the restriction W^D of the stored energy density W to $\mathbf{M}_S^D$ satisfies a linear growth at infinity. The admissible set $\mathcal{A}$ of displacement fields is a subset of the space $\{v \in LD(\Omega) \ : \ v = 0 \text{ on } \Gamma_0\}$, where

$$LD(\Omega) = \{v \in L^1(\Omega, \mathbf{R}^3) : \varepsilon(v) \in L^1(\Omega, \mathbf{M}_S)\}.$$

From a mathematical point of view, due to the linear growth of W^D at infinity, two difficulties may appear:

• The value of the infimum may be infinite. This problem leads to the theory of *yield design* (or *limit load*), which consists in analyzing the set of $\lambda \in \mathbf{R}^+$ such that

$$\inf \left\{ \int_\Omega W(\varepsilon(v))\, dx - \lambda L(v) \ : \ v \in \mathcal{A} \right\} > -\infty.$$

From the mechanical point of view, this analysis predicts the load capacity of the structure. For more details about limit analysis, see Section 15.4.

• Even if the infimum is finite, the problem has generally no solution. This is not surprising from a mechanical point of view, since the observed displacements are sometimes discontinuous on surfaces. A well-adapted space must contain admissible displacements with discontinuity on two-dimensional surfaces. This space, derived from $BV(\Omega, \mathbf{R}^3)$ where the measure $\varepsilon(u)$ plays the role of the measure Du, is precisely

$$BD(\Omega) := \big\{v \in L^1(\Omega, \mathbf{R}^3) \ : \ \varepsilon(u) \in \mathbf{M}(\Omega, \mathbf{M}_S)\big\}.$$

It will be described in detail in Subsection 14.1.3. The integral functional $\int_\Omega W(\varepsilon(u))$ of the measure $\varepsilon(u)$ will be defined in Subsection 14.1.2. Unfortunately, the new problem

$$\inf\left\{\int_\Omega W(\varepsilon(v)) - L(v) \ : \ v \in \tilde{\mathcal{A}}\right\},$$

where now the set $\tilde{\mathcal{A}}$ of admissible displacements is a subset of the space

$$\{v \in BD(\Omega) \ : \ v = 0 \text{ on } \Gamma_0\},$$

has generally no solution. Indeed, plastification phenomena may appear on the boundary, and discontinuities are sometimes observed on the part Γ_0 of the boundary. The boundary condition $v = 0$ in the trace sense must be replaced by a surface energy of the form

$$\int_{\Gamma_0} W^\infty(\gamma_0(v)_\tau \otimes_s \nu)\, d\mathcal{H}^2.$$

We have denoted by γ_0 the trace operator and by ν the exterior unit normal to Γ_0. The symmetric matrix field $\gamma_0(v)_\tau \otimes_s \nu$ will be defined further. Roughly, the function W^∞ describes the behavior of W at infinity on straight lines generated by $\gamma_0(v)_\tau \otimes_s \nu$. The set of admissible displacement fields, still denoted by $\tilde{\mathcal{A}}$, is now a subset of $BD(\Omega)$ whose elements satisfy a weaker boundary condition and will be described further.

The mathematical theory of relaxation introduced in chapter 11 allows us to sum up this discussion as follows:

$$\inf\left\{\int_\Omega W(\varepsilon(v)) + \int_{\Gamma_0} W^\infty(\gamma_0(v)_\tau \otimes_s \nu)\, d\mathcal{H}^2 - L(v) \ : \ v \in \tilde{\mathcal{A}}\right\} \qquad (\overline{\mathcal{P}})$$

is the relaxed problem of

$$\inf\left\{\int_\Omega W(\varepsilon(v))\, dx - L(v) \ : \ v \in \mathcal{A}\right\}, \qquad (\mathcal{P})$$

that is,

$$v \mapsto \int_\Omega W(\varepsilon(v)) + \int_{\Gamma_0} W^\infty(\gamma_0(v)_\tau \otimes_s \nu)\, d\mathcal{H}^2 + \mathrm{Ind}_{\tilde{\mathcal{A}}}$$

is the lower semicontinuous envelope of $v \mapsto \int_\Omega W(\varepsilon(v))\, dx + \mathrm{Ind}_{\mathcal{A}}$ when $BD(\Omega)$ is equipped with its weak convergence. Moreover $\inf(\mathcal{P}) = \min(\overline{\mathcal{P}})$. The next sections are devoted to a precise description and to a proof of this relaxation scheme.

14.1.2 The Hencky model

To illustrate the previous general considerations, we deal with the description of the Hencky model. The reference configuration Ω is assumed to have a boundary of class $\mathbf{C}^1$. Let λ and μ be two given positive constants, namely, the Lamé coefficients of the material, and set $k = \lambda + 2\mu/3$, the compression stiffness. The constitutive equation of the material is such that there exists a potential W of the form

$$W(E) = W^D(E^D) + \frac{k}{2}\left(\mathrm{tr}(E)\right)^2$$

for all $E = E^D + (1/3)\ \mathrm{tr}(E)I$ in $\mathbf{M}_S = \mathbf{M}_S^D \otimes \mathbf{R}I$. The density W^D is more precisely defined by $W^D(E^D) = \phi(|E^D|)$, where $\Phi : \mathbf{R}^+ \longrightarrow \mathbf{R}$ has a quadratic growth up to $\frac{k}{\mu\sqrt{2}}$, and a linear growth beyond this threshold. More precisely,

$$\phi(s) = \begin{cases} \mu s^2 \text{ if } s \leq \dfrac{k}{\mu\sqrt{2}}, \\ s\, k\sqrt{2} - \dfrac{k^2}{2\mu} \text{ if } s \geq \dfrac{k}{\mu\sqrt{2}}. \end{cases}$$

The proof of Lemma 14.1.1 below may be easily established and is left to the reader.

Lemma 14.1.1. *The function W^D is convex and fulfills the three conditions* (13.23), (13.24), *and* (13.25).

The set of admissible displacement fields is the set of finite energy, that is,

$$\mathcal{A} = \{v \in LD(\Omega) : \mathrm{div}(v) \in L^2(\Omega),\ v = 0 \text{ on } \Gamma_0\},$$

and problem $(\mathcal{P})$ is precisely

$$\inf\left\{\int_\Omega W^D(\varepsilon^D(v))\,dx + \frac{k}{2}\int_\Omega (\mathrm{div}(v))^2\,dx - L(v)\ :\ v \in \mathcal{A}\right\}.$$

Let us recall that for a function $v : \Omega \mapsto \mathbf{R}^N$, its divergence in the distributional sense is given by $\mathrm{div}\ v := \sum_{i=1}^N \frac{\partial u_i}{\partial x_i}$. The boundary condition $v = 0$ on Γ_0 must be taken in the trace sense. The trace operator is indeed well defined from $LD(\Omega)$ into $L^1(\Gamma_0, \mathbf{R}^3)$. For a proof, it suffices to adapt the trace Theorem 10.2.1 (see Remark 10.2.2 or see, for instance, Temam [219]).

We define now the relaxed problem. For all a in $\mathbf{R}^3$ and every unit vector ν in $\mathbf{R}^3$, we denote the tangential and normal components of a relative to ν, by a_τ and a_ν, respectively. In other words $a_\tau = a - (a\cdot\nu)\,\nu$, where $a\cdot\nu$ denotes the scalar product of a and ν in $\mathbf{R}^3$. For all a and b in $\mathbf{R}^3$, we define their symmetric tensor product by $a \otimes_s b := 1/2(a_i b_j + a_j b_i)_{i,j}$. The relaxed problem is precisely

$$\begin{aligned}(\overline{\mathcal{P}}) \qquad \inf\Bigg\{ &\int_\Omega W^D(\varepsilon^D(v)) + \frac{k}{2}\int_\Omega (\mathrm{div}(v))^2\,dx \\ &+ \int_{\Gamma_0} W^\infty(\gamma_0(v)_\tau \otimes_s \nu)\,d\mathcal{H}^2 - L(v)\ :\ v \in \tilde{\mathcal{A}}\Bigg\},\end{aligned} \tag{14.1}$$

where γ_0 is the trace operator from $BD(\Omega)$ into $L^1(\Gamma)$ and the set of admissible displacements is

$$\tilde{\mathcal{A}} = \{v \in BD(\Omega) \ : \ \text{div } v \in L^2(\Omega), \ v_\nu = 0 \text{ on } \Gamma_0\}.$$

Note that the boundary condition $v = 0$ (taken in the trace sense) on Γ_0 in problem $(\mathcal{P})$ has been relaxed, in problem $(\overline{\mathcal{P}})$, by the surface energy

$$\int_{\Gamma_0} W^\infty(\gamma_0(v)_\tau \otimes_s \nu) \, d\mathcal{H}^2$$

and the weaker boundary condition $v_\nu = 0$ on Γ_0. This last condition must be taken in the trace sense $\gamma_\nu(v) = 0$ where γ_ν is a linear continuous operator from the space

$$\{v \in L^1(\Omega, \mathbf{R}^3) : \text{div } v \in L^2(\Omega)\}$$

into the dual $\mathbf{C}^1(\Gamma)'$ of $\mathbf{C}^1(\Gamma)$. The existence of this trace operator γ_ν will be established in Theorem 14.1.3.

The integral $\int_\Omega W^D(\varepsilon^D(v))$ must be taken in the sense of measures defined in Sections 13.3 and 13.4.1. Let us recall that the measure $W^D(\varepsilon^D(v))$ denotes the Borel measure $W^D(e^D) \ \mathcal{L}^3 \lfloor \Omega + (W^D)^\infty(\varepsilon^D(v)^S)$ where $e^D \ \mathcal{L}^3 \lfloor \Omega + \varepsilon^D(v)^S$ is the Lebesgue–Nikodym decomposition of the measure $\varepsilon^D(v)$, and that the recession function $(W^D)^\infty$ of W^D is defined for all E in $\mathbf{M}^D_S$ by

$$(W^D)^\infty(E) = \lim_{t \to +\infty} \frac{W^D(tE)}{t}.$$

Consequently, by definition one has

$$\int_\Omega W^D(\varepsilon^D(v)) := \int_\Omega W^D(e^D) \, dx + \int_\Omega (W^D)^\infty(\varepsilon^D(v)^S).$$

When the singular part $\varepsilon^D(v)^S$ of $\varepsilon^D(v)$ vanishes, we also denote the measure $e^D(v)\mathcal{L}^3 \lfloor \Omega$ by $\varepsilon^D(v)\mathcal{L}^3 \lfloor \Omega$. In the same spirit, if $e \ \mathcal{L}^3 \lfloor \Omega + \varepsilon(v)^S$ is the Lebesgue–Nikodym decomposition of the measure $\varepsilon(v)$, we denote the measure $e(v)\mathcal{L}^3 \lfloor \Omega$ by $\varepsilon(v)\mathcal{L}^3 \lfloor \Omega$ when $\varepsilon(v)^S = 0$.

14.1.3 The spaces $BD(\Omega)$, $M(\text{div})$, and $U(\Omega)$

Unless differently specified, the set Ω is, for the moment, a bounded open subset of $\mathbf{R}^3$. As said before, a well-adapted space for relaxing the above model is the space defined below.

Definition 14.1.1. *The subspace*

$$BD(\Omega) := \{v \in L^1(\Omega, \mathbf{R}^3) : \varepsilon(v) \in \mathbf{M}(\Omega, \mathbf{M}_S)\}$$

of $L^1(\Omega, \mathbf{R}^3)$ is called the space of bounded deformations. The measure $\varepsilon(v) \in \mathbf{C}_0'(\Omega, \mathbf{M}_S)$ is defined by its action on all φ in $\mathbf{C}_0(\Omega, \mathbf{M}_S)$:

$$\langle \varepsilon(u), \varphi \rangle = \sum_{i,j} \langle \varepsilon(v)_{i,j}, \varphi_{i,j} \rangle,$$

where the brackets in the right-hand side, denote the action of the signed measure $\varepsilon(v)_{i,j}$ *on the scalar function* $\varphi_{i,j}$ *for the duality* $(\mathbf{C}_0'(\Omega), \mathbf{C}_0(\Omega))$.

Remark 14.1.1. The action of $\varepsilon(v)$ on φ will also be written as $\int_\Omega \varphi\, \varepsilon(v)$ which is also well defined on bounded Du-integrable functions φ. Note that, when v is regular (for instance, belongs to $W^{1,1}(\Omega, \mathbf{R}^3)$),

$$\begin{aligned}\int_\Omega \varphi\ \varepsilon(v) \ &= \ \int_\Omega \varphi(x)\ \varepsilon(v)(x)\ dx \\ &:= \int_\Omega \varphi(x)\ :\ \varepsilon(v)(x)\ dx,\end{aligned}$$

where for two 3×3 matrices A and B, $A : B$ denotes their Hilbert–Schmidt scalar product defined by $A : B := \text{trace}(A^T B)$. This is why we also write $\int_\Omega \varphi : \varepsilon(v)$ for the integral $\int_\Omega \varphi\, \varepsilon(v)$ with respect to the measure $\varepsilon(v)$.

Under these considerations, we leave the reader to adapt the definitions of the weak and intermediate convergences and the proof of the approximating Theorem 10.1.2 or its generalization, Theorem 13.4.1. It suffices to argue with the components of u and to replace everywhere Du by $\varepsilon(u)$ (see also Remark 10.2.2). Because of its importance in Subsection 14.1.4, we state the approximating theorem.

Theorem 14.1.1. *Let* $f : \mathbf{M}_S \to \mathbf{R}^+$ *be a convex function satisfying* (13.23) *and* (13.24). *The space* $\mathbf{C}^\infty(\Omega, \mathbf{M}_S) \cap BD(\Omega)$ *is dense in* $BD(\Omega)$ *equipped with the intermediate convergence associated with* f. *More precisely, for all* u *in* $BD(\Omega)$, *there exists* u_n *in* $\mathbf{C}^\infty(\Omega, \mathbf{M}_S) \cap BD(\Omega)$ *such that*

$$\begin{cases} u_n \to u \text{ strongly in } L^1(\Omega, \mathbf{R}^3); \\ \displaystyle\int_\Omega |\varepsilon(u_n)|\, dx \to \int_\Omega |\varepsilon(u)|; \\ \displaystyle\int_\Omega f(\varepsilon(u_n))\, dx \to \int_\Omega f(\varepsilon(u)); \\ \displaystyle\int_\Omega f(\varepsilon^D(u_n))\, dx \to \int_\Omega f(\varepsilon^D(u)). \end{cases}$$

In the same spirit, we state without proof the trace theorem.

Theorem 14.1.2. *Let* Ω *be a Lipschitz open bounded subset of* $\mathbf{R}^3$. *There exists a linear continuous map* γ_0 *from* $BD(\Omega)$ *onto* $L^1_{\mathcal{H}^2}(\Gamma, \mathbf{R}^3)$ *satisfying*

(i) $\forall u \in \mathbf{C}(\overline{\Omega}, \mathbf{R}^3) \cap BD(\Omega)$, $\gamma_0(u) = u\lfloor_\Gamma$;

(ii) $\forall \varphi \in \mathbf{C}(\overline{\Omega}, \mathbf{M}_S)$

$$\int_\Omega \varphi : \varepsilon(u) = -\int_\Omega u \cdot \text{div}\ \varphi\, dx + \int_\Gamma \gamma_0(u) \otimes_S \nu : \varphi\, d\mathcal{H}^2,$$

where ν is the outer unit normal at $\mathcal{H}^2$ almost all x on Γ and div φ *denotes the vector valued function defined by* $(\text{div } \varphi)_i := \sum_{j=1}^{3} \frac{\partial \varphi_{i,j}}{\partial x_j}$, $i = 1, \ldots, 3$.

As a consequence of the Green's formula (ii) above, one can adapt the two first examples of Section 10.2.

Example 14.1.1. Consider two disjoint Lipschitz open bounded subsets Ω_1 and Ω_2 of an open bounded subset Ω of $\mathbf{R}^3$ such that $\overline{\Omega} = \overline{\Omega}_1 \cup \overline{\Omega}_2$ and set $\Gamma_{1,2} := \partial\Omega_1 \cap \partial\Omega_2$ which is assumed to satisfy $\mathcal{H}^2(\Gamma_{1,2}) > 0$. We denote the trace operators from $BD(\Omega_1)$ onto $L^1_{\mathcal{H}^2}(\partial\Omega_1, \mathbf{R}^3)$ and $BD(\Omega_2)$ onto $L^1_{\mathcal{H}^2}(\partial\Omega_2, \mathbf{R}^3)$ by γ_1 and γ_2, respectively. Let u_1 and u_2 be, respectively, two elements of $BD(\Omega_1)$ and $BD(\Omega_2)$ and define

$$u = \begin{cases} u_1 \text{ in } \Omega_1, \\ u_2 \text{ in } \Omega_2. \end{cases}$$

Then u belongs to $BD(\Omega)$ and

$$\varepsilon(u) = \varepsilon(u_1)\lfloor_{\Omega_1} + \varepsilon(u_2)\lfloor_{\Omega_2} + [u] \otimes_s \nu \mathcal{H}^{N-1}\lfloor_{\Gamma_{1,2}},$$

where $[u] = \gamma_1(u_1) - \gamma_2(u_2)$ and $\nu(x)$ is the unit inner normal at x to $\Gamma_{1,2}$, considered as a part of the boundary of Ω_1.

Example 14.1.2. By slightly modifying the previous example, if we set

$$v = \begin{cases} u \text{ in } \Omega, \\ 0 \text{ in } \mathbf{R}^3 \setminus \overline{\Omega}, \end{cases}$$

where Ω is a Lipschitz bounded open subset of $\mathbf{R}^3$ and $u \in BD(\Omega)$, we see that v belongs to $BD(\mathbf{R}^3)$ and that

$$\varepsilon(v) = \varepsilon(u)\lfloor_{\Omega} + u^+ \otimes_s \nu \, \mathcal{H}^{N-1}\lfloor_{\Gamma},$$

where Γ is the boundary of Ω, ν denotes the inner unit vector normal to Γ, and u^+ denotes the trace of u on Γ.

Remark 14.1.2. Remark 10.2.1 also holds in this situation. More precisely, let Ω be a Lipschitz open bounded subset of $\mathbf{R}^3$. The approximating Theorem 14.1.1 may easily be improved as follows: the regular approximating functions of $u \in BD(\Omega)$ have all their traces equal to that of u on the boundary of Ω. For a proof, see, for instance, [219].

We define now the space $\mathbf{M}(\text{div})$ which is involved in the definition of the admissible set of displacement fields in the Hencky model.

Definition 14.1.2. *We denote the space of all the functions in $L^1(\Omega, \mathbf{R}^3)$ such that their divergence belongs to $L^2(\Omega)$, by* $\mathbf{M}(\text{div})$*:*

$$\mathbf{M}(\text{div}) := \{v \in L^1(\Omega, \mathbf{R}^3) \ : \ \text{div } v \in L^2(\Omega)\}.$$

On $\mathbf{M}(\text{div})$, we can define a trace notion.

Theorem 14.1.3. *Let Ω be a $\mathbf{C}^2$ open bounded subset of $\mathbf{R}^3$. There exists a linear continuous map γ_v from $\mathbf{M}(\mathrm{div})$ into $\mathbf{C}^1(\Gamma)'$ satisfying*

(i) $\forall u \in \mathbf{C}(\overline{\Omega}, \mathbf{R}^3) \cap BD(\Omega)$, $\gamma_v(u) = u \cdot v\lfloor_\Gamma$;

(ii) *for all u in $\mathbf{M}(\mathrm{div})$, all $\varphi \in \mathbf{C}^1(\Gamma)$, and all $\Phi \in \mathbf{C}^1(\overline{\Omega})$ such that $\Phi\lfloor_\Gamma = \varphi$,*

$$\langle \gamma_v(u), \varphi \rangle = \int_\Omega u \cdot D\Phi \, dx + \int_\Omega \Phi \ \mathrm{div}\ u \, dx.$$

For a proof, consult Temam [219, Proposition 7.2]. Note that this theorem also makes sense when div u is a Borel measure. In this case, the space $\mathbf{M}(\mathrm{div})$ is defined by

$$\mathbf{M}(\mathrm{div}) := \{v \in L^1(\Omega, \mathbf{R}^3) \ : \ \mathrm{div}\ v \in \mathbf{M}(\Omega)\}.$$

For our application we consider only the case when div $v \in L^2(\Omega)$.

We consider now the space $U(\Omega) := BD(\Omega) \cap \mathbf{M}(\mathrm{div})$ equipped with two convergences. Precisely, for all sequence $(u_n)_{n\in\mathbf{N}}$ in $U(\Omega)$, one defines

• the weak convergence, defined by the weak convergence of u_n to u in $BD(\Omega)$ and the strong convergence of div u_n to div u in $L^2(\Omega)$;

• the intermediate convergence associated with a convex function f, defined by the intermediate convergence of u_n to u in $BD(\Omega)$ associated with f together with the strong convergence of div u_n to div u in $L^2(\Omega)$.

The following result completes the approximating Theorem 14.1.1 in the spirit of Remark 14.1.2. For a proof, consult Theorems 3.4 and 5.3 in [219].

Theorem 14.1.4. *Let $f : \mathbf{M}_S \to \mathbf{R}$ be a convex function satisfying*

$$\alpha(|A| - 1) \le f(A) \le \beta(1 + |A|)$$

for all A in $\mathbf{M}_S$. Then for all u in $U(\Omega)$, there exists u_n in $\mathbf{C}^\infty(\Omega, \mathbf{M}_S) \cap BD(\Omega)$ such that

$$\begin{cases} u_n \to u \text{ strongly in } L^1(\Omega, \mathbf{R}^3); \\ \displaystyle\int_\Omega |\varepsilon(u_n)| \, dx \to \int_\Omega |\varepsilon(u)|; \\ \displaystyle\int_\Omega f(\varepsilon(u_n)) \, dx \to \int_\Omega f(\varepsilon(u)); \\ \mathrm{div}\ u_n \to \mathrm{div}\ u \text{ strongly in } L^2(\Omega); \\ \displaystyle\int_\Omega f(\varepsilon^D(u_n)) \, dx \to \int_\Omega f(\varepsilon^D(u)); \\ \gamma_0(u_n) = \gamma_0(u). \end{cases}$$

14.1.4 Relaxation of the Hencky model

In this subsection we show that the lower semicontinuous envelope of the Hencky functional energy for the weak topology of $U(\Omega)$ is the functional of $(\overline{\mathcal{P}})$ in (14.1). We moreover establish the convergence of the corresponding energy to the energy of the relaxed functional. Here, Ω is a bounded open subset of $\mathbf{R}^3$ of class $\mathbf{C}^2$.

Theorem 14.1.5. *The lower semicontinuous envelope for the weak convergence of $U(\Omega)$, of the integral functional defined on $U(\Omega)$ by*

$$F(v) = \begin{cases} \displaystyle\int_\Omega W^D(\varepsilon^D(v))\,dx + \frac{k}{2}\int_\Omega (\mathrm{div}(v))^2\,dx \text{ if } v \in LD(\Omega) \cap \mathbf{M}(\mathrm{div}),\ v = 0 \text{ on } \Gamma_0, \\ +\infty \text{ otherwise,} \end{cases}$$

is the functional defined on $U(\Omega)$ by

$$\overline{F}(v) = \begin{cases} \displaystyle\int_\Omega W^D(\varepsilon^D(v)) + \frac{k}{2}\int_\Omega (\mathrm{div}(v))^2\,dx + \int_{\Gamma_0} (W^D)^\infty(\gamma_0(v)_\tau \otimes_s \nu)\,d\mathcal{H}^2, \\ \qquad\qquad \text{if } v \in U,\ \gamma_\nu(v) = 0 \text{ on } \Gamma_0, \\ +\infty \text{ otherwise.} \end{cases}$$

Sketch of the proof. Arguing exactly as in the proof of Theorem 13.4.2, where Theorem 13.4.1 is replaced by Theorem 14.1.4, we obtain

$$\overline{F}(v) = \begin{cases} \displaystyle\int_\Omega W^D(\varepsilon^D(v)) + \frac{k}{2}\int_\Omega (\mathrm{div}(v))^2\,dx + \int_{\Gamma_0} (W^D)^\infty(\gamma_0(v) \otimes_s \nu)\,d\mathcal{H}^2, \\ \qquad\qquad \text{if } v \in U,\ v_\nu = 0 \text{ on } \Gamma_0, \\ +\infty \text{ otherwise,} \end{cases}$$

where condition $v_\nu = 0$ on Γ_0 is intended in the trace sense: for H^2 a.e. x on Γ_0, $\gamma_\nu(v)(x) = 0$. Indeed, the map $v \mapsto \int_\Omega (\mathrm{div}(v))^2\,dx$ is continuous for the weak convergence of $U(\Omega)$. Note also that according to Theorem 14.1.3, the trace operator defined from $\mathbf{M}(\mathrm{div})$ into $\mathbf{C}^1(\Gamma)'$ is continuous when the two spaces are equipped with their weak convergences. Finally, for $\mathcal{H}^2$ a.e. x on Γ_0, since $v_\nu = 0$ in the trace sense on Γ_0, we have $\gamma_0(v)_\tau \otimes_s \nu(x) = \gamma_0(v) \otimes_s \nu(x)$. Indeed, it is easily seen that $\gamma_0(v) \cdot \nu = \gamma_\nu(v)\ \mathcal{H}^2_{\lfloor \Gamma_0}$ a.e. □

Corollary 14.1.1. *The energy of the Hencky model*

$$\inf\left\{ \int_\Omega W^D(\varepsilon^D(v))\,dx + \frac{k}{2}\int_\Omega (\mathrm{div}(v))^2\,dx - L(v)\ :\ v \in \mathcal{A} \right\},$$

where $\mathcal{A} = \{v \in LD(\Omega) : \mathrm{div}(v) \in L^2(\Omega),\ v = 0 \text{ on } \Gamma_0\}$, relaxes to

$$\min\left\{ \int_\Omega W^D(\varepsilon^D(v)) + \frac{k}{2}\int_\Omega (\mathrm{div}(v))^2\,dx + \int_{\Gamma_0} W^\infty(\gamma_0(v)_\tau \otimes_s \nu)\,dH^2 - L(v)\ :\ v \in \tilde{\mathcal{A}} \right\},$$

where $\tilde{\mathcal{A}} = \{v \in BD(\Omega)\ :\ \mathrm{div}(v) \in L^2(\Omega),\ v_\nu = 0 \text{ on } \Gamma_0\}$.

Sketch of the proof. According to the general theory of relaxation (see Theorem 11.1.2), it suffices to prove that any minimizing sequence related to the Hencky energy possesses a subsequence weakly converging in the space U. This assertion may be easily established thanks to coercivity condition (13.25). □

14.2 Some variational models in fracture mechanics

14.2.1 A few considerations in fracture mechanics

Let us consider an elastic brittle medium whose reference configuration is a bounded domain Ω of $\mathbf{R}^3$. Griffith's theory of fracture mechanics asserts that the energy necessary to produce a crack K included in Ω is proportional to the crack area $\mathcal{H}^2(K)$. Consequently, the elastic deformation energy outside the crack must be completed by an additional energy whose simplest form is $\lambda\mathcal{H}^2(K)$. The constant λ is the Griffith coefficient, introduced for fracture initiation (see [149], [200]). The elastic energy of the deformable body under consideration then takes the form

$$E(u, K) = \int_{\Omega\setminus K} f(\nabla u)\, dx + \lambda\mathcal{H}^2(K),$$

where u denotes the deformation vector field and ∇u the deformation gradient. Under suitable conditions on f, the functional E makes sense in a classical way if, for instance, K is a closed set, and u belongs to $\mathbf{C}^1(\Omega\setminus K, \mathbf{R}^3)$. From the inequality $\mathcal{H}^2(K) \le \lambda^{-1}E(u, K)$, we see that when $E(u, K) < +\infty$, the crack surface K is a two-Hausdorff-dimensional closed set of Ω and its Lebesgue measure is zero. Thus, the crack surface K can be seen as the set of discontinuity points for the measurable function u, more precisely, the measurable representative of u defined on Ω, satisfying the convention of Remark 10.3.2. It is worth noticing the analogy between this model and the strong model introduced in image segmentation by Mumford–Shah and discussed in Section 12.4 (see also Section 14.3 for complements). Following the idea developed in Section 12.4, one may define a weak formulation in the setting of SBV functions introduced in Section 10.5 (completed by Remark 10.5.1) by considering the functional

$$E(u) = \int_\Omega f(\nabla u)\, dx + \lambda\mathcal{H}^2(S_u),\ u \in SBV(\Omega, \mathbf{R}^3),$$

where ∇u is the density of the regular part of the measure Du and S_u the jump set of u. Actually, one can deal with functionals of the more general form

$$E(u) = \int_\Omega f(x, \nabla u)\, dx + \int_\Omega g(x, u^+(x), u^-(x), \nu_u(x))\, d\mathcal{H}^2\lfloor S_u.$$

The bulk energy density f accounts for the elastic deformation outside the crack and g for the density energy necessary to produce a crack of surface S_u. The meaning of the presence of the terms $u^+(x)$, $u^-(x)$, and $\nu_u(x)$ is the following: the fracture energy may depend on the crack opening and, for nonisotropic materials, on the crack surface orientation.

According to Theorem 13.4.5 and Remark 13.4.3, suitable conditions on f and g ensure semicontinuity of E so that direct methods of the calculus of variations provide the existence of solutions for optimization problems related to the energy functionals of the type E. More precisely, we consider a lower semicontinuous symmetric and subadditive function $g : \mathbf{R}^3 \times \mathbf{R}^3 \to \mathbf{R}^+$, i.e.,

$$g(a, b) = g(b, a) \le g(b, c) + g(c, a)\ \forall\, a,\ b,\ c \in \mathbf{R}^3.$$

We assume moreover that $g(a,b) \geq \psi(|a-b|)$ for all $a, b \in \mathbf{R}^3$, where $\psi : [0,+\infty) \to [0,+\infty]$, satisfies $\psi(t)/t \to +\infty$ as $t \to 0$. Subadditivity assumption on the function g forces the crack material to possess a minimal number of connected components. We finally consider a convex, even function $h : \mathbf{R}^3 \to [0,+\infty)$, positively homogeneous of degree 1 and satisfying for all $\nu \in \mathbf{R}^N$, $h(\nu) \geq c|\nu|$, where c is a given positive constant.

Theorem 14.2.1. *Let $f : \mathbf{R} \times M^{3\times 3} \to \mathbf{R}$ be a quasi-convex function satisfying the following growth conditions of order $p > 1$: there exist two positive constants α and β such that for all $A \in M^{3\times 3}$,*

$$\alpha|A|^p \leq f(x,A) \leq \beta(1+|A|^p).$$

Let g and h be two functions satisfying the conditions introduced above, and let $L \in L^\infty(\Omega, \mathbf{R}^3)$ (the exterior loading). Then, given a nonempty compact subset K of $\mathbf{R}^3$, there exists a solution of the minimum problem

$$\min\left\{ \int_\Omega f(\nabla u)\,dx + \int_\Omega g(u^+,u^-)h(\nu_u)\,d\mathcal{H}^2\lfloor S_u + \int_\Omega L\cdot u\,dx \;:\; u \in SBV(\Omega,\mathbf{R}^3),\right.$$
$$\left. u(x) \in K \;\text{ for a.e. } x\right\}.$$

Proof. Following the classical direct methods of the calculus of variation, the conclusion is a straightforward consequence of the lower semicontinuity of the functional energy, which has been established in Theorem 13.4.5 and Remark 13.4.3. The coercivity easily follows from the confinement condition $u(x) \in K$ for a.e. x. □

A similar type of result is described in detail in Section 14.3 for the Mumford–Shah model in image segmentation. In the second assertion of Theorem 14.2.1, the confinement condition $u \in K$ does not seem to be natural for all problems in fracture mechanics. To remove such a condition, we have in general to state the problems in the space $GSBV$ of functions whose truncations are in SBV (see [17], [82], and references therein). A similar statement can be given for boundary value problems. In this case we know (see the previous section or Theorem 11.3.1) that the Dirichlet boundary conditions $u = u_0$ on a subset Γ_0 of the boundary Γ of Ω, assumed to be Lipschitz, are relaxed into a surface energy at the boundary. Taking, for instance, $g = 1$, we have to consider minimization problems of the form

$$\min\left\{ \int_\Omega f(\nabla u)\,dx + \int_\Omega h(\nu_u)\,d\mathcal{H}^2\lfloor S_u + \int h(\nu)\,d\mathcal{H}^2\lfloor\Gamma_0 + \int_\Omega L\cdot u\,dx \;:\; u \in SBV(\Omega,\mathbf{R}^3),\right.$$
$$\left. u(x) \in K \text{ for a.e. } x\right\},$$

where ν is the inner unit normal to Γ_0. Existence of a solution is obtained in a similar way. The argument above also works for more general minimum problems of the form (see, for instance, [118])

$$\min\left\{ \int_\Omega f(\nabla u)\,dx + \int_\Omega g(u^+,u^-)h(\nu_u)\,d\mathcal{H}^2\lfloor S_u + \int_\Omega V(x,u)\,dx : u \in SBV(\Omega,\mathbf{R}^3)\right\},$$

where f, g, h are as above, and the potential V is integrable in x and satisfies

$$V(x,s) \geq \theta(|s|) - a(x)$$

with $a \in L^1(\Omega)$ and $\theta(t)/t \to +\infty$ as $t \to +\infty$. The case of Theorem 14.2.1 corresponds to $V(x, s) = L(x) \cdot s + \delta_K(s)$.

Functionals of the type E can be approximated by elliptic functionals via a variational procedure due to Ambrosio and Tortorelli and described in Section 12.4.3 for the Mumford–Shah energy. These approximating functionals are more adapted to a numerical treatment (see [61], [80]). In the one-dimensional case, the two models introduced in Subsections 14.2.2 and 14.2.3 provide an alternative and more direct way for defining a discrete variational approximation of E.

Other models, proposed by Barenblatt [58], can be similarly weakened in the space $SBV(\Omega, \mathbf{R}^3)$ by considering energies of the type

$$E(u) = \int_\Omega f(\nabla u)\, dx + \int_\Omega g(|u^+ - u^-|)\, d\mathcal{H}^2 \lfloor S_u,$$

where $g(t) \to 0$ as $t \to 0$. Optimization problems related to such energies functionals do not provide minimizing sequences satisfying Theorem 14.2.1. Indeed, we lose the control of the Hausdorff measure of the jump set. Therefore, minimizing sequences may converge to a function whose jump part is not concentrated on an $(N-1)$-Hausdorff dimensional set and, consequently, it does not belong to $SBV(\Omega, \mathbf{R}^3)$. To select minimizers, one can follow a singular perturbation approach, involving the notion of viscosity solutions (consult Attouch [29]). This procedure has already been described in details for the phase transition model in Section 12.4.2. Precisely, the method consists in perturbing the functional E by the fracture initiation energy $\varepsilon\mathcal{H}^2(S_u)$ and introducing functionals of Griffith type

$$E_\varepsilon(u) = E(u) + \varepsilon\mathcal{H}^2(S_u)$$

($\varepsilon > 0$ intended to tend to zero) for which Theorem 14.2.1 applies. The cluster points of ε-minimizers related to E_ε then belong to $SBV(\Omega, \mathbf{R}^3)$ and minimize E among minimizers with jump set of minimal $\mathcal{H}^{N-1}$ measure (see [17] and [85]). For a specific study of problems involving fracture mechanics in the modern framework of SBV functions, we refer the reader to [17], [22], [138], [139], and references therein.

In sections below, we would like to supply a justification of weak Griffith's models, in the one-dimensional case, by taking into account the microscopic scale and the statistical local energy distribution. More precisely, we deal with two discrete systems of material points, which, in the reference configuration, occupy the points of the lattice $\varepsilon\mathbf{Z}$, $\varepsilon = 1/n$ included in the interval $[0, 1]$. For each material point placed at $x \in \varepsilon\mathbf{Z}$ in the reference configuration, $u(x) \in \mathbf{R}$ denotes its new position in the deformed state. Each point interacts only with its nearest neighbors. We aim at describing the continuous limit of the discrete energy in the sense of Γ-convergence. We show that the continuous variational limit of the second model takes the form of Griffith's model discussed above. In the first model, we show that convexity conditions satisfied by local density energies entail the presence of a Cantor-part energy: the Griffith's initiation energy is completed by an additional term $C\,u'_c(0, 1)$, where u'_c is the Cantor part of u' and C a suitable constant. It is worth noticing that the discrete energies considered may also be viewed as discrete variational approximating functionals of Griffith's type in the one-dimensional case. This approach is very close to that of Chambolle consisting in approximating the Mumford–Shah functionals in the two-dimensional case (see Chambolle [107]) and to some recent works by Braides [84] and Braides and Gelli [89].

14.2.2 A first model in one dimension

The interaction between each pair $\{z\varepsilon, (z+1)\varepsilon\}$ of contiguous points is described by a random energy

$$\varepsilon W_{\omega_z}\left(\frac{u(\varepsilon(z+1)) - u(\varepsilon z)}{\varepsilon}\right),$$

where W_{ω_z} belongs to a finite set $\{W_j,\ j \in J\}$. Each density $W_j,\ j \in J$, mapping $\mathbf{R}$ into $[0, +\infty]$, is assumed to be of Hencky pseudoplastic type (see Section 14.1) and must take the value $+\infty$ if two neighboring points occupy the same location. The total energy of the interaction between points located in $[0, 1] \cap \varepsilon\mathbf{Z}$ is

$$E_\varepsilon(\omega, u) = \sum_{z=0}^{n-1} \varepsilon\ W_{\omega_z}\left(\frac{u(\varepsilon(z+1)) - u(\varepsilon z)}{\varepsilon}\right),$$

where $\omega = (\omega_z)_{z\in\mathbf{Z}} \in J^{\mathbf{Z}}$. One may think that two neighboring material points are connected by randomly chosen nonlinear springs, and this model also describes a system of n springs which are randomly distributed. We still denote the continuous extension of u, which is affine on $(\varepsilon z, \varepsilon(z+1))$ for all $z \in \mathbf{Z}$, by u. Thus the discrete energy may be considered as defined on $L^1(0, 1)$. This first model is described by the almost sure Γ-convergence of $u \mapsto E_\varepsilon(\omega, u)$ toward a deterministic energy functional living in $BV(0, 1)$.

Let us give some specific notations about BV-functions in the one-dimensional case. For each function u in $BV(0, 1)$, one writes $u' = u'_a\ dt + u'_s$ the Lebesgue–Nikodym decomposition of its distributional derivative u' with respect to the Lebesgue measure dt on $(0, 1)$. The singular part with respect to dt has the following decomposition: $u'_s = \sum_{t\in S_u}(u^+ - u^-)\delta_t + u'_c$, where u^+ and u^- are, respectively, the approximate upper and lower limits of u, S_u is the jump set $\{t \in (0, 1) : u^+(t) \neq u^-(t)\}$ of u, and u'_c is the singular diffuse part, also called the Cantor part of u'. Let $BV^+(0, 1)$ be the subset of all the functions u in $BV(0, 1)$ such that $u'_a > 0$ a.e. in $(0, 1)$ and $u'_s \geq 0$. We will establish that the (deterministic) limit energy functional is defined on $BV^+(0, 1)$ by

$$E(u) := \int_0^1 W^{hom}(u'_a)\ dt + C\ u'_s\big((0, 1)\big),$$

where C is a positive constant. The density $e \mapsto W^{hom}(e)$ is obtained as the almost sure limit of a suitable subadditive ergodic process. This mathematical result expresses that the mechanical macroscopic behavior of a string can be interpreted as the variational limit of a discrete system at a microscopic scale. Moreover, we will express the duality principle $(W^{hom})^* = \sum_{j\in J} p_j W_j^*$, where p_j is the probability presence of each density W_j. For other models in a deterministic setting and some discussions about the possible existence of a fracture site related to these models, consult [87] and [17].

Let us give some more details on the random discrete model. As said above, we consider a finite set Λ of functions $W_j,\ j \in J$, of probability presence p_j, satisfying the three following conditions:

(i) $W_j : \mathbf{R} \to \mathbf{R}^+ \cup \{+\infty\}$ is convex, finite for $e > 0$, $W_j(1) = 0$, and there exists $\alpha > 0$ such that $\alpha(e - 1) \leq W_j(e)$ for all $e \geq 0$;

(ii) there exists $\beta > 0$ such that $W_j(e) \leq \beta(1+e)$ for all $e > 1$;

(iii) $\lim_{e\to 0^+} W_j(e) = +\infty$ and $W_j(e) = +\infty$ when $e \leq 0$.

The assumption $W_j(1) = 0$ means that no energy is needed when no deformation occurs. Assumption (iii) means that an infinite amount of energy is needed to squeeze a pair of material points down to a single one and that there is no interpenetrability of the matter. These density functions are of Hencky pseudoplastic type owing to the convexity and the linear growth conditions.

The fundamental stochastic setting that we will need for describing the discrete energy and its asymptotic behavior is the discrete dynamical system $(\Omega, \mathcal{T}, \mathbf{P}, (T_z)_{z\in\mathbf{Z}})$: $\Omega = \Lambda^{\mathbf{Z}}$, $(\Omega, \mathcal{T}, \mathbf{P})$ is the product probability space of the Bernoulli probability space on Λ constructed from p_j, $j \in J$, the transformation T_z is the shift defined for all $z \in \mathbf{Z}$ by $T_z\big((\omega_s)_{s\in\mathbf{Z}}\big) = (\omega_{s+z})_{s\in\mathbf{Z}}$. The expectation operator will be denoted by $\mathbf{E}$. To write in a continuous form the random energy functional, we consider the random function defined for all $(\omega, t, e) \in \Omega \times \mathbf{R} \times \mathbf{R}$ by

$$W(\omega, t, e) = W_{\omega_z}(e) \text{ when } t \in [z, z+1).$$

Let $\mathcal{A}_\varepsilon(0,1)$ denote the space of continuous functions on $(0,1)$ which are affine on each interval $(\varepsilon z, \varepsilon(z+1))$ of $(0,1)$. More generally when $s = (b-a)/n$, $\mathcal{A}_s(a,b)$ will denote the space of continuous functions on (a,b) which are affine on each interval $(a+sz, a+s(z+1))$ of (a,b). The total energy due to the interactions between the points of $[0,1] \cap \varepsilon\mathbf{Z}$ is the functional defined in $L^1(0,1)$ by

$$E_\varepsilon(\omega, u) = \begin{cases} \displaystyle\sum_{z=0}^{n-1} \varepsilon\, W\Big(\omega, z, \frac{u(\varepsilon(z+1)) - u(\varepsilon z)}{\varepsilon}\Big) & \text{if } u \in \mathcal{A}_\varepsilon(0,1), \\ +\infty \quad \text{otherwise,} \end{cases}$$

or, in a continuous form,

$$E_\varepsilon(\omega, u) = \begin{cases} \displaystyle\int_0^1 W\Big(\omega, \frac{t}{\varepsilon}, u'(t)\Big)\, dt & \text{if } u \in \mathcal{A}_\varepsilon(0,1), \\ +\infty \quad \text{otherwise,} \end{cases}$$

where $t \mapsto u(t)$ also denotes the piecewise affine extension of u. Note that the domain of $E_\varepsilon(\omega, .)$ is the subset of all functions u in $\mathcal{A}_\varepsilon(0,1)$ whose distributional derivative u' is positive.

To solve the problem, we make use of the following concept of subadditive random process which extends, in a probabilistic setting, the notion of subadditive process developed in Lemma 12.3.1. Consider a probability space $(\Omega, \mathcal{T}, \mathbf{P})$ and a group $(\tau_s)_{s\in\mathbf{Z}^d}$ of $\mathbf{P}$-preserving transformations on $(\Omega, \mathcal{T})$, i.e.,

(i) τ_s is $\mathcal{T}$-measurable;

(ii) $\mathbf{P}(\tau_s E) = \mathbf{P}(E)$ for all E in $\mathcal{T}$ and all s in $\mathbf{Z}^d$;

(iii) $\tau_s \circ \tau_t = \tau_{s+t}$, $\tau_{-s} = \tau_s^{-1}$ for all s and t in $\mathbf{Z}^d$.

The group $(\tau_s)_{s\in Z^d}$ is said to be ergodic if every set E in $\mathcal{T}$, such that for all $s \in \mathbf{Z}^d$ $\tau_s E = E$, has a probability measure equal to 0 or 1. A sufficient condition to ensure ergodicity is the following mixing condition: for all E and F in $\mathcal{T}$

$$\lim_{|s|\to+\infty} \mathbf{P}(\tau_s E \cap F) = \mathbf{P}(E)\mathbf{P}(F),$$

which expresses an asymptotic independence. We summarize these conditions by saying that $(\Omega, \mathcal{T}, \mathbf{P}, (\tau_z)_{z\in\mathbf{Z}})$ is a discrete ergodic dynamical system. By using classical probabilistic arguments, it is easily seen that $(\Omega, \mathcal{T}, \mathbf{P}, (T_z)_{z\in\mathbf{Z}})$, previously defined, is a discrete ergodic dynamical system.

We are now going to recall the important notion of random subadditive process in a general probabilistic setting. The notation $[a, b),\ a \in Z^d,\ b \in Z^d$, stands for $\prod_{i=1}^d [a_i, b_i)$ and $\mathcal{J}$ is the set of all subsets $[a, b)$ with a, b in $\mathbf{Z}^d$. For every I in $\mathcal{J}$, $|I|$ denotes its Lebesgue measure. Finally, the positive number $\rho(I)$ denotes the supremum of the set $\{r \geq 0 \ : \ \exists \bar{B}_r(x) \subset I\}$ where $\bar{B}_r(x)$ is the closed ball of $\mathbf{Z}^d$ with radius r, centered at x. Given a discrete ergodic dynamical system $(\Omega, \mathcal{T}, \mathbf{P}, (\tau_z)_{z\in\mathbf{Z}^d})$, a random subadditive process with respect to the group $(\tau_s)_{s\in\mathbf{Z}^d}$ is a set function $\mathcal{S} : \mathcal{J} \longrightarrow L^1(\Omega, \mathcal{T}, \mathbf{P})$ satisfying (i) and (ii):

(i) for all $I \in \mathcal{J}$ such that there exists a finite family $(I_j)_{j\in J}$ of disjoint sets in $\mathcal{J}$ with $I = \bigcup_{j\in J} I_j$,

$$\mathcal{S}_I(\cdot) \leq \sum_{j\in J} \mathcal{S}_{I_j}(\cdot);$$

(ii) $\forall I \in \mathcal{J},\ \forall s \in \mathbf{Z}^d, \mathcal{S}_{s+I} = \mathcal{S}_I \circ \tau_s$.

A sequence $(A_n)_{n\in\mathbf{N}}$ of sets of $\mathcal{J}$ is said to be regular if there exists an increasing sequence $(I_n)_{n\in\mathbf{N}}$ of $\mathcal{J}$ and a positive constant C independent of n such that $A_n \subset I_n$ and $|I_n| \leq C|A_n|$ for all $n \in \mathbf{N}$. The following subadditive ergodic theorem is due to Ackoglu and Krengel. It will be applied in the next section, with $d = 1$, to the discrete ergodic dynamical system $(\Omega, \mathcal{T}, \mathbf{P}, (T_z)_{z\in\mathbf{Z}})$.

Theorem 14.2.2. *Let $\mathcal{S}$ be a subadditive process with respect to the group $(\tau_s)_{z\in\mathbf{Z}^d}$ which is assumed to be ergodic. Suppose that*

$$\inf\left\{\int_\Omega \frac{\mathcal{S}_I(\omega)}{|I|}\,\mathbf{P}(d\omega) : |I| \neq 0\right\} > -\infty$$

and let $(A_n)_{n\in\mathbf{N}}$ be a regular sequence of $\mathcal{J}$ satisfying $\lim_{n\to+\infty} \rho(A_n) = +\infty$. Then almost surely

$$\lim_{n\to+\infty} \frac{\mathcal{S}_{A_n}(\omega)}{|A_n|} = \inf_{m\in\mathbf{N}^*}\left\{E\frac{\mathcal{S}_{[0,m[^d}}{m^d}\right\}.$$

For a proof see [3] and, for some extensions, see [165] and [170].

The space $L^1(0, 1)$, equipped with its norm, plays the role of the metric space (X, d) in the Γ-convergence definition of chapter 12.

Theorem 14.2.3. *The energy functional $E_\varepsilon(\omega, .)$ Γ-converges almost surely to the functional E defined in $L^1(0,1)$ by*

$$E(u) = \begin{cases} \displaystyle\int_0^1 W^{hom}(u_a')\,dt + W^{hom,\infty}(1)\,u_s'\big((0,1)\big) \text{ if } u \in BV^+(0,1), \\ +\infty \text{ otherwise.} \end{cases}$$

The density W^{hom} is defined as follows: $W^{hom}(e) = +\infty$ if $e \le 0$ and, if $e > 0$, one has ω-a.s.,

$$\begin{aligned} W^{hom}(e) &= \lim_{n\to+\infty} \inf\left\{ \frac{1}{n}\int_0^n W(\omega, t, e+v')\,dt \ :\ v \in W_0^{1,1}(0,n)\right\} \\ &= \inf_{n\in\mathbf{N}^*} E\left(\inf\left\{ \frac{1}{n}\int_0^n W(.,t,e+v')\,dt \ :\ v \in W_0^{1,1}(0,n)\right\}\right). \end{aligned}$$

Moreover, W^{hom} verifies properties (i), (ii), *and* (iii) *of the functions W_j and its Legendre–Fenchel transform is given by $(W^{hom})^* = \sum_{j\in J} p_j W_j^*$.*

The proof is established by means of Propositions 14.2.1 and 14.2.2, each giving, respectively, the lower bound and the upper bound for a subsequence in the definition of Γ-convergence. Before stating Proposition 14.2.1, we introduce a parametrized subadditive process, i.e., a family of subadditive processes which will be used to define the limit problem. For this purpose, for $\delta \in (0, \delta_0]$, $j \in J$, we consider the truncated functions $T_\delta W_j = W_j \wedge L_\delta$, where L_δ is the affine function defined by $L_\delta(t) = W_j(\delta) + \tau_i(t-\delta)$, $\tau_j \in \partial W_j(\delta)$ (the subdifferential of W_j at δ) and set

$$\begin{cases} W_\delta(\omega, t, e) = T_\delta \omega_z(e) \text{ when } t \in [z, z+1), \\ W_0 = W. \end{cases}$$

The nonincreasing family $(W_\delta)_{\delta\in(0,\delta_0]}$ satisfies for $0 < \delta \le \delta_0$, $\omega \in \Omega$, $t \in \mathbf{R}$, and $e \in \mathbf{R}$:

$$\begin{cases} W_\delta \le W,\ W_\delta(e) = W(e) \text{ for } e \ge \delta, \\ \lim_{\delta\to 0} W_\delta(\omega, t, e) = W(\omega, t, e), \\ e \mapsto W_\delta(\omega, t, e) \text{ is convex}, \\ \alpha(|e| - 1) \le W_\delta(\omega, t, e) \le \beta_\delta(1 + |e|), \end{cases}$$

where β_δ is a positive constant depending only on δ.

Let $\mathcal{F}(\Omega, \mathbf{R}^+ \cup \{+\infty\})$ be the set of all the measurable functions from Ω into $\mathbf{R}^+ \cup \{+\infty\}$ and consider the parametrized subadditive process $\mathcal{S}$ defined by

$$\mathcal{S} :\ \mathcal{J} \times [0, \delta_0] \times \mathbf{R} \longrightarrow \mathcal{F}(\Omega, \mathbf{R}^+ \cup \{+\infty\}),\ (A, \delta, e) \mapsto \mathcal{S}_A(\delta, e, .),$$

where

$$\mathcal{S}_A(\delta, e, \omega) = \inf\left\{ \int_A W_\delta(\omega, t, e+v')\,dt \ :\ v \in W_0^{1,1}(\overset{\circ}{A})\right\}.$$

It is worth noticing that the domain of $e \mapsto \mathcal{S}_A(\delta, e, \omega)$ is $\mathbf{R}$ when $\delta \in (0, \delta_0]$ while the one of $e \mapsto \mathcal{S}_A(0, e, \omega)$ is $]0, +\infty[$. Indeed, if $\mathcal{S}_A(0, e, \omega) < +\infty$, for $e \le 0$, there exists $v \in W_0^{1,1}(\overset{\circ}{A})$ such that $W(\omega, t, v'(t)) < +\infty$ t a.e. in $\overset{\circ}{A}$. Therefore $v'(t) > -e$ for t a.e. in $\overset{\circ}{A}$ and $\int_A v'\,dt = 0 > -e$, a contradiction.

It is easily seen that for all fixed (δ, e) in $\big((0, \delta_0] \times \mathbf{R}\big) \cup \big(\{0\} \times (0, +\infty)\big)$, the map $A \mapsto \mathcal{S}_A(\delta, e, .)$ is a subadditive process satisfying $\mathcal{S}_A(\delta, e, \omega) \leq |A| \max\{T_\delta W_j(e) : j \in J\}$ and that all conditions of Theorem 14.2.2 are fulfilled. Consequently, there exists a set Ω' of full probability such that for all (δ, e) in $\big((0, \delta_0] \cap \mathbf{Q}\big) \times \mathbf{R}$ and all $\omega \in \Omega'$,

$$W_\delta^{hom}(e) := \lim_{n \to +\infty} \frac{\mathcal{S}_{[0,n)}(\delta, e, \omega)}{n} = \inf_{m \in \mathbf{N}^*} \mathbf{E} \frac{\mathcal{S}_{[0,m)}}{m}(\delta, e, .). \tag{14.2}$$

On the other hand, for all $e > 0$, there exists a set Ω_e of full probability such that

$$W^{hom}(e) := \lim_{n \to +\infty} \frac{\mathcal{S}_{[0,n)}(0, e, \omega)}{n} = \inf_{m \in \mathbf{N}^*} \left\{ \mathbf{E} \frac{\mathcal{S}_{[0,m)}}{m}(0, e, .) \right\}$$

and $W^{hom}(e) = +\infty$ if $e \leq 0$. The independence of the first set Ω' with respect to e comes from the equi-Lipschitz property of $e \mapsto \mathcal{S}_{[0,n)}(\delta, e, \omega)$ (see [119] or [179]). The following lemma states the continuity at $\delta = 0^+$ of the function $\delta \mapsto W_\delta^{hom}(e)$ when $e > 0$.

Lemma 14.2.1. *For all fixed $e > 0$, there exists $\overline{\delta}(e) > 0$ such that for all $\delta \in (0, \overline{\delta}(e)] \cap \mathbf{Q}$, $W_\delta^{hom}(e) = W^{hom}(e)$.*

Proof. *First step*. This step is also valid for all $e \in \mathbf{R}$ and all $\omega \in \Omega$. Let $\delta > 0$, we show that there exists $u_{\delta,n}(\omega) \in \mathcal{A}_1(0, n) \cap W_0^{1,1}(0, n)$ such that

$$\frac{\mathcal{S}_{[0,n)}(\delta, e, \omega)}{n} = \frac{1}{n} \int_0^n W_\delta(\omega, t, e + u'_{\delta,n}(\omega))\, dt.$$

To shorten notation, we set $W_n(e) = \dfrac{\mathcal{S}_{[0,n)}(\delta, e, \omega)}{n}$. From a classical calculation, the Fenchel conjugate of $e \mapsto W_n(e)$ is given for all σ in $\mathbf{R}$ by

$$W_n^*(\sigma) = \frac{1}{n} \int_0^n W_\delta^*(\omega, t, \sigma)\, dt.$$

Set $A_j(\omega, n) = \{t \in [0, n] \ : \ W_\delta(\omega, t, .) = T_\delta W_j\}$ and $\lambda_j(\omega, n) = \text{meas}(A_j(\omega, n))$ for $j \in J$. Thus

$$W_n^*(\sigma) = \sum_{j \in J} \frac{\lambda_j(\omega, n)}{n} (T_\delta W_j)^*(\sigma) \tag{14.3}$$

and, by classical subdifferential rules (see chapter 9),

$$\partial W_n^*(\sigma) = \sum_{j \in J} \frac{\lambda_j(\omega, n)}{n} \partial (T_\delta W_j)^*(\sigma). \tag{14.4}$$

Let now $\sigma \in \partial W_n(e)$. Thus $e \in \partial W_n^*(\sigma)$ and from (14.4), there exists $U_{j,\delta,n}(\omega) \in \partial (T_\delta W_j)^*(\sigma)$ such that

$$e = \sum_{j \in J} \frac{\lambda_j(\omega, n)}{n} U_{j,\delta,n}(\omega).$$

The function

$$u_{\delta,n}(\omega)(x) = \int_0^x \sum_{j\in J} \big(U_{j,\delta,n}(\omega) - e\big) 1_{A_j(\omega,n)}(s)\, ds$$

answers the question. Indeed, $u_{\delta,n}$ belongs to $\mathcal{A}_1(0,n) \cap W_0^{1,1}(0,n)$. On the other hand, according to $U_{j,\delta,n}(\omega) \in \partial(T_\delta W_j)^*(\sigma)$, (14.3), and the fact that $e \in \partial W_n^*(\sigma)$,

$$\begin{aligned}
\frac{1}{n}\int_0^n W_\delta(\omega, t, e + u'_{\delta,n}(\omega))\, dt &= \sum_{j\in J} \frac{\lambda_j(\omega,n)}{n} T_\delta W_j(U_{j,\delta,n}(\omega)) \\
&= \sum_{j\in J} \frac{\lambda_j(\omega,n)}{n} \big(\sigma U_{j,\delta,n}(\omega) - T_\delta W_j^*(\sigma)\big) \\
&= \sigma e - W_n^*(\sigma) \\
&= W_n(e).
\end{aligned}$$

Second step. From classical probabilistic arguments, there exists a set Ω'' of full probability such that for all $\omega \in \Omega''$,

$$\lim_{n\to+\infty} \frac{\lambda_j(\omega,n)}{n} = p_j. \tag{14.5}$$

We pick up ω_0 in $\Omega_e \cap \Omega' \cap \Omega''$ and claim that $\inf_n U_{j,\delta,n}(\omega_0) \geq \delta$ for all $\delta \in \mathbf{Q}^{*+}$ small enough, say, $\delta \in (0, \overline{\delta}(e)] \cap \mathbf{Q}$. Otherwise there exist two sequences $(\delta_k)_k$ and $(n_k)_k$ converging to 0 and $+\infty$ such that $U_{j,\delta_k,n_k}(\omega_0) \leq \delta_k$. But, according to the first step

$$\begin{aligned}
\frac{\mathcal{S}_{[0,n_k)}(0,e,\omega_0)}{n_k} &\geq \frac{\mathcal{S}_{[0,n_k)}(\delta_k,e,\omega_0)}{n_k} \\
&= \sum_{j\in J} \frac{\lambda_j(\omega_0,n_k)}{n_k} T_{\delta_k} W_j(U_{j,\delta_k,n_k}(\omega_0)) \\
&\geq \sum_{j\in J} \frac{\lambda_j(\omega_0,n_k)}{n_k} W_j(\delta_k)
\end{aligned}$$

and letting $k \to +\infty$, we obtain $W^{hom}(e) = +\infty$, a contradiction.

Last step. According to the two previous steps, for $\delta \in (0, \overline{\delta}(e)] \cap \mathbf{Q}$ one has

$$\begin{aligned}
\frac{\mathcal{S}_{[0,n)}(0,e,\omega_0)}{n} &\geq \frac{\mathcal{S}_{[0,n)}(\delta,e,\omega_0)}{n} \\
&= \sum_{j\in J} \frac{\lambda_j(\omega_0,n)}{n} T_\delta W_j(U_{j,\delta,n}(\omega_0)) \\
&= \sum_{j\in J} \frac{\lambda_j(\omega_{\check{y}},n)}{n} W_j(U_{j,\delta,n}(\omega_0)) \\
&= \frac{1}{n}\int_0^n W(\omega_0, t, u'_{\delta,n}(\omega_0))\, dt \\
&\geq \frac{\mathcal{S}_{[0,n[}(0,e,\omega_0)}{n}.
\end{aligned}$$

Therefore $\dfrac{\mathcal{S}_{[0,n)}(0,e,\omega_0)}{n} = \dfrac{\mathcal{S}_{[0,n)}(\delta,e,\omega_0)}{n}$. Letting $n \to +\infty$, we obtain that $W^{hom}(e) = W^{hom}_{\delta}(e)$ as soon as $\delta \in (0, \overline{\delta}(e)] \cap \mathbf{Q}$. □

Proposition 14.2.1. *Let Ω', Ω'' be the subsets of full probability defined in* (14.2) *and* (14.5) *and let u, u_ε in $L^1(0,1)$ be such that $u_\varepsilon \to u$ strongly in $L^1(0,1)$. Then for all ω in $\Omega' \cap \Omega''$,*

$$E(u) \le \liminf_{\varepsilon\to 0} E_\varepsilon(\omega, u_\varepsilon).$$

Moreover the domain of $\Gamma - \liminf_{\varepsilon\to 0} E_\varepsilon(\omega, .)$ is included in $BV^+(0,1)$.

Proof. We fix ω in $\Omega' \cap \Omega''$.

First step. If $\liminf_{\varepsilon\to 0} E_\varepsilon(\omega, u_\varepsilon) < +\infty$, by the coercivity condition on the functions W_j (property (i)), u belongs to $BV(0,1)$ and obviously $u' \ge 0$. Let us now consider, for $\delta \in]0, \delta_0] \cap \mathbf{Q}$, the truncated energy

$$E_{\varepsilon,\delta}(\omega, u) = \begin{cases} \displaystyle\int_0^1 W_\delta(\omega, \frac{t}{\varepsilon}, u'(t))\, dt \text{ if } u \in \mathcal{A}_\varepsilon(0,1), \\ +\infty \text{ otherwise,} \end{cases}$$

and the corresponding energy with domain $W^{1,1}(0,1)$

$$\tilde{E}_{\varepsilon,\delta}(\omega, u) = \begin{cases} \displaystyle\int_0^1 W_\delta(\omega, \frac{t}{\varepsilon}, u'(t))\, dt \text{ if } u \in W^{1,1}(0,1), \\ +\infty \text{ otherwise,} \end{cases}$$

which satisfies all the properties of random integral functionals considered in [1]. From the inequalities $E_\varepsilon(\omega, .) \ge E_{\varepsilon,\delta}(\omega, .) \ge \tilde{E}_{\varepsilon,\delta}(\omega, .)$ and according to [1] we then deduce

$$\liminf_{\varepsilon\to 0} E_\varepsilon(\omega, u_\varepsilon) \ge \liminf_{\varepsilon\to 0} E_{\varepsilon,\delta}(\omega, u_\varepsilon) \ge E_\delta(u), \tag{14.6}$$

where

$$E_\delta(u) = \begin{cases} \displaystyle\int_0^1 W^{hom}_\delta(u'_a)\, dt + W^{hom,\infty}_\delta(1) u'_s\Big((0,1)\Big) \text{ if } u \in BV(0,1), \\ +\infty \text{ otherwise.} \end{cases}$$

Second step. It is not restrictive to assume $\liminf_{\varepsilon\to 0} E_\varepsilon(\omega, u_\varepsilon) < +\infty$. Obviously $u'_a \ge 0$ and $u'_s \ge 0$. We would like to let δ going to 0 in (14.6) and apply Lemma 14.2.1. It remains to prove that $u'_a > 0$ a.e. in $(0,1)$. Otherwise, there exists a Borel set N of $(0,1)$ with $\mathrm{meas}(N) \ne 0$, such that $u'_a = 0$ on N. We have

$$+\infty > \liminf_{\varepsilon\to 0} E_\varepsilon(\omega, u_\varepsilon) \ge \mathrm{meas}(N) W^{hom}_\delta(0), \tag{14.7}$$

where, from the first step in the proof of Lemma 14.2.1,

$$\begin{cases} W^{hom}_\delta(0) = \displaystyle\lim_{n\to+\infty} \sum_{j\in J} \frac{\lambda_j(\omega, n)}{n} T_\delta W_j(U_{j,\delta,n}(\omega)), \\ \displaystyle\sum_{j\in J} \frac{\lambda_j(\omega, n)}{n} U_{j,\delta,n}(\omega) = 0. \end{cases}$$

By the coercivity assumption (i), an easy calculation leads to $\sup_n |U_{j,\delta,n}(\omega)| \le C$. Thus there exists $U_{j,\delta}(\omega)$ in $\mathbf{R}$ satisfying, up to a subsequence, $\lim_{n\to+\infty} U_{j,\delta,n}(\omega) = U_{j,\delta}(\omega)$ and

$$\begin{cases} W_\delta^{hom}(0) = \displaystyle\sum_{j\in J} p_j T_\delta W_j(U_{j,\delta}(\omega)), \\ \displaystyle\sum_{j\in J} p_j U_{j,\delta}(\omega) = 0. \end{cases}$$

The second equality yields the existence of an index j_δ such that $U_{j_\delta,\delta} \le 0$ and the first equality gives

$$\begin{aligned} W_\delta^{hom}(0) &\ge p_{j_\delta} T_\delta W_{j_\delta}(U_{j_\delta,\delta}(\omega)) \\ &\ge p_{j_\delta} T_\delta W j_\delta(0) \\ &\ge \min_j p_j \min_j T_\delta W_j(0) \end{aligned}$$

so that $\lim_{\delta\to 0} W_\delta^{hom}(0) = +\infty$. Letting $\delta \to 0$ in (14.7) leads to a contradiction.

Last step. Letting $\delta \to 0$ in (14.6), according to the monotone convergence theorem and Lemma 14.2.1, we finally obtain

$$\liminf_{\varepsilon\to 0} E_\varepsilon(\omega, u_\varepsilon) \ge \int_0^1 W^{hom}(u_a')\,dt + W^{hom,\infty}(1)\,u_s'\big((0,1)\big).$$

The proof is then achieved. □

To establish the upper bound, we will apply the following lemma.

Lemma 14.2.2. (i) *Let $e > 0$ and let $i \in \mathbf{N}$ be a fixed integer. There exists $u_{i,n}(\omega)$ in $\mathcal{A}_1(in,(i+1)n) \cap W_0^{1,1}(in,(i+1)n)$ such that*

$$\mathcal{S}_{[in,(i+1)n)}(0,e,\omega) = \int_{in}^{(i+1)n} W(\omega,t,e+u_{i,n}'(\omega))\,dt.$$

(ii) *The map $W^{hom} : \mathbf{R} \to [0,+\infty]$ is convex, continuous, and $W^{hom}(1) = 0$.*

Proof. For establishing (i), reproduce the first step of the proof of Lemma 14.2.1 with $(0,n)$ replaced by $(in,(i+1)n)$ and W_δ by W. Note also that there exist $U_{j,\delta} > 0$, $j = 1,2,3$ (depending on e), satisfying $\sup_n U_{j,n} < +\infty$, such that

$$\begin{cases} \dfrac{\mathcal{S}_{[0,n)}(0,e,\omega)}{n} = \displaystyle\sum_{j\in J} \frac{\lambda_j(\omega,n)}{n} W_j(U_{j,n}(\omega)), \\ \displaystyle\sum_{j\in J} \frac{\lambda_j(\omega,n)}{n} U_{j,n} = e,\ U_{j,n} > 0. \end{cases}$$

The convexity of the map $e \mapsto W^{hom}(e)$ is a consequence of Jensen's inequality fulfilled for $e > 0$ and established by a straightforward calculation. Consequently, this map

is continuous on its domain $\mathbf{R}^{*+}$ and, to prove (ii), we must show that $\lim_{e\to 0^+} W^{hom}(e) = +\infty$. Let $e_k > 0$ tend to 0 and fix ω in $\Omega'' \cap_{k\in\mathbf{N}} \Omega_{e_k}$. Letting $n \to +\infty$ (up to a subsequence) in

$$\begin{cases} \dfrac{\mathcal{S}_{[0,n[}(0, e_k, \omega)}{n} = \displaystyle\sum_{j\in J} \frac{\lambda_j(\omega, n)}{n} W_j(U_{j,n,k}(\omega)), \\ \displaystyle\sum_{j\in J} \frac{\lambda_j(\omega, n)}{n} U_{j,n,k} = e_k, \end{cases}$$

we obtain the existence of numbers $U_{j,k} \geq 0$ such that

$$\begin{cases} W^{hom}(e_k) = \displaystyle\sum_{j\in J} p_j W_j(U_{j,k}(\omega)), \\ \displaystyle\sum_{j\in J} p_j U_{j,k} = e_k. \end{cases}$$

Note that this shows that $W^{hom}(1) = 0$. Obviously $U_{j,k} \to 0$ when $k \to +\infty$. Thus

$$\lim_{k\to+\infty} W^{hom}(e_k) = \sum_{j\in J} p_j \lim_{k\to+\infty} W_j(U_{j,k}) = +\infty$$

and the proof is complete. □

We establish now the upper bound.

Proposition 14.2.2. *There exists a subset Ω''' of full probability such that for all $\omega \in \Omega'''$, there exists a subsequence $(n_k)_{k\in\mathbf{N}}$ of positive integers satisfying the following: for all u in $L^1(0,1)$, there exists u_{n_k} in $L^1(0,1)$ possibly depending on ω, converging to u, and such that* $\limsup_{k\to+\infty} E_{1/n_k}(\omega, u_{n_k}) \leq E(u)$.

Proof. *First step*. We prove Proposition 14.2.2 for all u belonging to $SBV_{\#}(0,1) \cap BV^+(0,1)$ where $SBV_{\#}(0,1)$ denotes the subspace of all functions of $SBV(0,1)$ having a finite jump set.

It is not restrictive to assume that $S_u = \{t_0\}$. For $m \in \mathbf{N}^*$, consider the decomposition $(0,1) = \cup_{i=0}^{m-1}(i/m, (i+1)/m)$. Let us set i_0 to denote the integer such that $t_0 \in [i_0/m, (i_0+1)/m)$ and by u_m the interpolate function of u with respect to this decomposition. We set $e_{i,m} := u'_m\lfloor(i/m, (i+1)/m)$ if $i \neq i_0$ and $e_{i_0,m} = m\big(u((i_0+1)/m) - u(i_0/m)\big)$. For each $i \in \{1, \ldots, m-1\}$ we consider a sequence $e_{i,m,\eta}$ in $\mathbf{Q}^{*+}$ such that $\lim_{\eta\to 0} e_{i,m,\eta} = e_{i,m}$. By continuity of W^{hom} and Jensen's inequality, we have

$$\begin{aligned} \lim_{\eta\to 0} \sum_{i=0,\ i\neq i_0}^{m-1} \frac{1}{m} W^{hom}(e_{i,m,\eta}) &\leq \sum_{i=0,\ i\neq i_0}^{m-1} \int_{\frac{i}{m}}^{\frac{i+1}{m}} W^{hom}(u'_a)\, dt \\ &\leq \int_0^1 W^{hom}(u'_a)\, dt \end{aligned}$$

and

$$\lim_{m\to+\infty}\lim_{\eta\to 0}\frac{1}{m}W^{hom}(e_{i_0,m,\eta}) = \lim_{m\to+\infty}\frac{1}{m}W^{hom}\left(m\left(u\left(\frac{i_0+1}{m}\right)-u\left(\frac{i_0}{m}\right)\right)\right)$$
$$= W^{hom,\infty}([u](t_0))$$

so that

$$\lim_{m\to+\infty}\lim_{\eta\to 0}\sum_{i=0}^{m-1}\frac{1}{m}W^{hom}(e_{i,m,\eta}) \le E(u). \tag{14.8}$$

But Lemma 14.2.2 implies the existence of $u_{i,\eta,n}$ in $\mathcal{A}_1(in,(i+1)n)\cap W_0^{1,1}(in,(i+1)n)$ (we have dropped the dependence on ω) such that

$$\frac{\mathcal{S}_{[in,(i+1)n)}(0,e_{i,m,\eta},\omega)}{n} = \frac{1}{n}\int_{in}^{(i+1)n} W(\omega,t,e_{i,m,\eta}+u'_{i,\eta,n}\,dt$$
$$= m\int_{i/m}^{(i+1)/m} W(\omega,mnt,e_{i,m}+u'_{i,\eta,n}(mnt))\,dt.$$

Therefore, as the sequence of intervals $\big((in,(i+1)n)\big)_{n\in\mathbf{N}^*}$ is regular, according to Theorem 14.2.2, there exists $\Omega''' = \cap_{e\in\mathbf{Q}^{*+}}\Omega_e$ of full probability such that for all $\omega\in\Omega'''$

$$W^{hom}(e_{i,m,\eta}) = \lim_{n\to+\infty} m\int_{i/m}^{(i+1)/m} W(\omega,mnt,e_{i,m,\eta}+u'_{i,\eta,n}(mnt))\,dt. \tag{14.9}$$

Combining (14.8) and (14.9), we obtain

$$\lim_{m\to+\infty}\lim_{\eta\to 0}\lim_{n\to+\infty}\int_0^1 W(\omega,mnt,v'_{m,\eta,n})\,dt \le E(u),$$

where

$$v_{m,\eta,n}(t) = u_m(t)+\sum_{i=0}^{m-1}1_{(i/m,(i+1)/m)}(t)\frac{1}{mn}u_{i,\eta,n}(nmt).$$

An easy calculation shows that $\lim_{m\to+\infty}\lim_{\eta\to 0}\lim_{n\to+\infty}v_{m,\eta,n}=u$ strongly in $L^1(0,1)$ and that $v_{m,\eta,n}$ belongs to $\mathcal{A}_{1/nm}(0,1)$. Therefore, by using a diagonalization argument, there exists a map $n\mapsto(m(n),\eta(n))$ such that

$$\begin{cases}\lim_{n\to+\infty}E_{\frac{1}{m(n)n}}(\omega,v_{m(n),\eta(n),n})\le E(u),\\ \lim_{n\to+\infty}v_{m(n),\eta(n),n}=u.\end{cases}$$

We complete the proof by denoting $k\mapsto n_k$, the subsequence $n\mapsto m(n)n$, and setting $u_{nm(n)} := v_{m(n),\eta(n),n}$.

Second step. We prove Proposition 14.2.2 for u belonging to $SBV^+(0,1)$. Let $S_u=\{t_0,\dots,t_i,\dots\}$ be the jump set of u and let u_l the function of $SBV_\#(0,1)$ with jump set $S_{u_l}=\{t_0,\dots,t_l\}$ defined by $u_l(0^+)=u(0^+)$ and $u'_l=u'_a+\sum_{i=0}^l[u](.)\delta_{t_i}$. According to the first step, there exists u_{l,n_k} strongly converging to u_l in $L^1(0,1)$ such that

$$\limsup_{k\to+\infty}E_{1/n_k}(\omega,u_{l,n_k})\le E(u_l).$$

Letting l tend to $+\infty$ and by using a diagonalization argument, there exists a map $k \mapsto l(k)$ such that the sequence $u_{l(k),n_k}$, still denoted by u_{n_k}, strongly converges to u in $L^1(0,1)$ and satisfies

$$\limsup_{k\to+\infty} E_{1/n_k}(\omega, u_{n_k}) \le E(u).$$

Last step. According to the previous step, we have $\Gamma - \limsup_{k\to+\infty} E_{1/n_k}(\omega, .) \le \tilde{E}$, where

$$\tilde{E}(u) = \begin{cases} \displaystyle\int_0^1 W^{hom}(u'_a)\,dt \; + W^{hom,\infty}(1) \sum_{t\in S_u} [u](t) \text{ if } u \in SBV^+(0,1), \\ +\infty \text{ otherwise.} \end{cases}$$

Its lower semicontinuous envelope for the strong topology of $L^1(0,1)$ is (see [76])

$$E(u) = \begin{cases} \displaystyle\int_0^1 W^{hom}(u'_a)\,dt \; + W^{hom,\infty}(1) u'_s\Big((0,1)\Big) \text{ if } u \in BV^+(0,1), \\ +\infty \text{ otherwise,} \end{cases}$$

which ends the proof. □

Proof of Theorem 14.2.3. Propositions 14.2.1 and 14.2.2 and the last assertion of Theorem 12.1.1 imply that for all ω in the set $\Omega \cap \Omega'' \cap \Omega'''$ of full probability, $E_\varepsilon(\omega, .)$ Γ-converges to E. It remains to show that W^{hom} satisfies the three properties of the functions W_j. Convexity, property (iii), and $W^{hom}(1) = 0$ have been proved in Lemma 14.2.2(ii) and the growth conditions are trivially satisfied. Finally, we establish $(W^{hom})^* = \sum_{j\in J} p_j W_j^*$. We make precise the notations introduced in the proof of Lemma 14.2.1 by pointing out the dependence of W_n with respect to the parameter δ. We then set

$$W_n^\delta(e) = \frac{\mathcal{S}_{[0,n)}(\delta, e, \omega)}{n}.$$

Its Fenchel conjugate is given for all σ in $\mathbf{R}$ by

$$(W_n^\delta)^*(\sigma) = \frac{1}{n}\int_0^n W_\delta^*(\omega, t, \sigma)\,dt.$$

Equation (14.3) becomes now

$$(W_n^\delta)^*(\sigma) = \sum_{j\in J} \frac{\lambda_j(\omega, n)}{n} (T_\delta W_j)^*(\sigma). \tag{14.10}$$

According to the strong law of large numbers, the right-hand side of (14.10) tends almost surely to

$$\sum_{j\in J} p_j (T_\delta W_j)^*(\sigma)$$

when n goes to infinity. Note that this pointwise limit is also the Γ-limit of

$$\sigma \mapsto \sum_{j\in J} \frac{\lambda_j(\omega, n)}{n} (T_\delta W_j)^*(\sigma)$$

defined on $\mathbf{R}$.

On the other hand the pointwise limit of W_n^δ toward W_δ^{hom} obtained by the subadditive ergodic theorem is also the Γ-limit of $\sigma \mapsto W_n^\delta(\sigma)$ defined on $\mathbf{R}$. Indeed, the sequence of functions $(W_\delta^{hom})_{n\in\mathbf{N}^*}$ is equi-Lipschitz, so that pointwise and Γ-limit agree (see Corollary 2.59 in [28]). According to the continuity of the Fenchel conjugate with respect to the Mosco-convergence, hence here, to the Γ-convergence, we deduce from (14.10)

$$(W_\delta^{hom})^*(\sigma) = \sum_{j\in J} p_j (T_\delta W_j)^*(\sigma) \ \forall \sigma \in \mathbf{R}. \tag{14.11}$$

We would now like to go to the limit on δ in (14.11). Since the sequence $(W_\delta^{hom})_\delta$ of lower semicontinuous functions defined on $\mathbf{R}$ increases to the lower semicontinuous function W^{hom} when δ tends to 0, we have $W_\delta^{hom} \to W^{hom}$ in the sense of Γ-convergence for functionals defined on $\mathbf{R}$. The same argument implies that $T_\delta W_j$ Γ-converges to W_j. According to the continuity of the Fenchel conjugate with respect to the Γ-convergence (note that Mosco and Γ-convergence agree), we finally deduce our result by going to the limit on δ in (14.11). □

Remark 14.2.1. Concerning the upper bound, for all u in $SBV_\#(0,1) \cap BV^+(0,1)$, we have proved the existence of u_ε strongly converging to u in $L^1(0,1)$ and having the traces of u at 0^+ and 1^-.

It is possible to generalize the previous study when the density functions W_j satisfy a growth condition of order $p > 1$. In this case $W^{hom,\infty}(1) = +\infty$ and the limit functional given by Theorem 14.2.3 becomes

$$E(u) = \begin{cases} \displaystyle\int_0^1 W^{hom}(u')\,dt \text{ if } u' \in L^p(0,1),\ u' > 0, \\ +\infty \text{ otherwise.} \end{cases}$$

Let us consider the functional

$$\tilde{E}_\varepsilon(\omega, u) = \begin{cases} \displaystyle\int_0^1 W(\omega, \frac{t}{\varepsilon}, u')\,dt \text{ if } u \in W^{1,1}(0,1), \\ +\infty \text{ otherwise.} \end{cases}$$

Then, according to Proposition 14.2.2, we have $\Gamma - \limsup_{\varepsilon\to 0} \tilde{E}_\varepsilon \le E$ almost surely. On the other hand, the truncation argument of the first step in the proof of Proposition 14.2.2, implies that for $u_\varepsilon \to u$ in $L^1(0,1)$,

$$\liminf_{\varepsilon\to 0} \tilde{E}_\varepsilon(\omega, u_\varepsilon) \ge \liminf_{\varepsilon\to 0} \tilde{E}_{\varepsilon,\delta}(\omega, u_\varepsilon) \ge E_\delta(u).$$

Arguing as in the second step of this proof, we also obtain $\liminf_{\varepsilon\to 0} \tilde{E}_\varepsilon(\omega, u_\varepsilon) \ge E(u)$. Thus the functional $\tilde{E}(\omega, .)$ Γ-converges to the functional E.

14.2.3 A second model in one dimension

Keeping the same probabilistic setting, we study a new discrete model for which interaction between each pair of contiguous points is described by a random energy density which is

no longer assumed to be convex but which satisfies the same conditions in a neighborhood of 0^+. This energy functional is precisely assumed to be subadditive beyond a random threshold e_ε satisfying $\lim_{\varepsilon\to 0} \varepsilon e_\varepsilon = 0$. In this case, the total energy is of the form

$$F_\varepsilon(\omega, u) = \sum_{z=0}^{n-1} \varepsilon\ W_\varepsilon\left(\omega, z, \frac{u_{z+1} - u_z}{\varepsilon}\right)$$

and almost surely Γ-converges to a deterministic energy functional defined on the subset $SBV^+(0,1) := SBV(0,1) \cap BV^+(0,1)$ by

$$F(u) = \int_0^1 f(u'_a)\, dt + \sum_{t\in S_u} g([u]).$$

More precisely, we consider a finite number W_j, $j \in J$ of density functions satisfying

(i) $W_j : \mathbf{R} \to \mathbf{R}^+ \cup \{+\infty\}$ is convex, finite for $e > 0$, $W_j(1) = 0$ and there exists $\alpha > 0$ such that $\alpha(e-1)^2 \le W_j(e)$ for all e in $[0, +\infty)$;

(ii) there exists $\beta > 0$ such that $W_j(e) \le \beta(1 + e^2)$ for all $e > 1$;

(iii) $\lim_{e\to 0^+} W_j(e) = +\infty$ and $W_j(e) = +\infty$ when $e \le 0$.

Note that (i), (ii), and (iii) are the conditions of the first model but with growth conditions of order 2. On the other hand, let g be a subadditive, continuous function with at most linear growth, mapping $[0, +\infty)$ into $(0, +\infty)$ and satisfying $\inf_{[0,+\infty)} g > 0$. We then consider the density functions $W_{j,\varepsilon}$, $j \in J$, from $\mathbf{R}$ into $[0, +\infty)$ defined by

$$W_{j,\varepsilon}(e) = \begin{cases} W_j(e) \text{ if } e \le 1, \\ W_j(e) \wedge \frac{1}{\varepsilon} g(\varepsilon(e-1)) \text{ if } e > 1. \end{cases}$$

According to growth conditions, it is easily seen that there exists $e_{j,\varepsilon} > 1$ satisfying $\lim_{\varepsilon\to 0} e_{j,\varepsilon} = +\infty$, $\lim_{\varepsilon\to 0} \varepsilon e_{j,\varepsilon} = 0$, and such that

$$W_{j,\varepsilon}(e) = \begin{cases} W_j(e) \text{ if } e \le e_{j,\varepsilon}, \\ \dfrac{1}{\varepsilon} g(\varepsilon(e-1)) \text{ if } e > e_{j,\varepsilon}. \end{cases}$$

An example of such functions is given by $W_j = (e-1)^2$ when $e \ge 1$, $W_j(e) = -\ln e$ when $0 \le e \le 1$, and $g(e) = 1 + e$ or $g(e) = 1 + \sqrt{e}$ or $g(e) = \sqrt{1+e}$.

As said above, the subadditivity assumption on g forces the crack to possess a minimal number of connected components. Denoting by $e_{W_{j,\varepsilon}}$ every threshold $e_{j,\varepsilon}$, we now can define the random threshold $\omega \mapsto e_\varepsilon(\omega)$ by $e_\varepsilon(\omega) = (e_{\omega_z,\varepsilon})_{z\in\mathbf{Z}}$ and the random function W_ε for all $t \in [z, z+1)$ by

$$W_\varepsilon(\omega, t, e) = \begin{cases} \omega_z(e) \text{ if } e \le e_{\omega_z,\varepsilon}, \\ \dfrac{1}{\varepsilon} g(\varepsilon(e-1)) \text{ if } e > e_{\omega_z,\varepsilon}. \end{cases}$$

The total energy modeling interactions between contiguous points of $[0,1] \cap \varepsilon\mathbf{Z}$ is the functional defined on $L^1(0,1)$ by

$$F_\varepsilon(\omega,u) = \begin{cases} \displaystyle\sum_{z=0}^{n-1} \varepsilon\ W_\varepsilon\left(\omega, z, \frac{u(\varepsilon(z+1)) - u(\varepsilon z)}{\varepsilon}\right) \text{ if } u \in \mathcal{A}_\varepsilon(0,1), \\ +\infty \text{ otherwise,} \end{cases}$$

or, in a continuous form, by

$$F_\varepsilon(\omega,u) = \begin{cases} \displaystyle\int_0^1 W_\varepsilon\left(\omega, \frac{t}{\varepsilon}, u'\right)\ dt \text{ if } u \in \mathcal{A}_\varepsilon(0,1), \\ +\infty \text{ otherwise.} \end{cases}$$

We equip $L^1(0,1)$ with its strong convergence. The main result is given below.

Theorem 14.2.4. *The functional F_ε Γ-converges almost surely to the functional F defined in $L^1(0,1)$ by*

$$F(u) = \begin{cases} \displaystyle\int_0^1 W^{hom}(u'_a)\ dt + \sum_{t\in S_u} g([u](t)) \text{if } u \in SBV^+(0,1), \\ +\infty \text{ otherwise,} \end{cases}$$

where W^{hom} is the limit density defined in the first model.

The proof follows the lines of the first model proof: first we establish the lower bound, then the upper one for a subsequence. For more general models, but in a deterministic setting, see [88].

We would like to write the functional F_ε so that the contribution of W_j and g are separated. To this end, we consider the space $SBV_\varepsilon(0,1)$ of all the functions of $SBV(0,1)$ whose restriction to each interval $(\varepsilon z, \varepsilon(z+1))$ included in $(0,1)$ is affine, and we associate to each function u of $\mathcal{A}_\varepsilon(0,1)$ the function $\tilde{u}$ in $SBV_\varepsilon(0,1)$ defined for all $t \in \varepsilon[z, z+1)$ by

$$\tilde{u}(t) = \begin{cases} u(t) \text{ if } \dfrac{u(\varepsilon(z+1)) - u(\varepsilon z)}{\varepsilon} \le e_{\omega_z,\varepsilon}, \\ t - \varepsilon z + u(\varepsilon z) \text{ otherwise.} \end{cases}$$

Note that actually $\tilde{u}$ is a random function, but we have dropped the dependence on ω to shorten notations. Then, for all $u \in \mathcal{A}_\varepsilon(0,1)$, one has

$$\begin{aligned} F_\varepsilon(\omega,u) = & \sum_{\{z\,:\,u(\varepsilon(z+1))-u(\varepsilon z)\le \varepsilon e_{\omega_z,\varepsilon}\}} \varepsilon W\left(\omega, z, \frac{u(\varepsilon(z+1)) - u(\varepsilon z)}{\varepsilon}\right) \\ & + \sum_{\{z\,:\,u(\varepsilon(z+1))-u(\varepsilon z)> \varepsilon e_{\omega_z,\varepsilon}\}} g\left(\varepsilon\left(\frac{u(\varepsilon(z+1)) - u(\varepsilon z)}{\varepsilon} - 1\right)\right) \\ \ge & \int_0^1 W(\omega, \frac{t}{\varepsilon}, \tilde{u}'_a)\ dt + \sum_{t\in S_{\tilde{u}}} g([\tilde{u}](t)) := \tilde{F}_\varepsilon(\omega,\tilde{u}), \end{aligned}$$

where $\omega \mapsto W(\omega, t, e)$ is the random function defined in the first model. The proof of the lower bound in the definition of Γ-convergence is based on the following lemma.

Lemma 14.2.3. *Assume that* $\sup_\varepsilon F_\varepsilon(\omega, u_\varepsilon) < +\infty$ *and that* u_ε *strongly converges to some* u *in* $L^1(0,1)$. *Then* $\tilde{u}_\varepsilon$ *strongly converges to* u *in* $L^1_{loc}(0,1)$. *Moreover* u *belongs to* $SBV^+(0,1)$ *and*

$$\sum_{t\in S_u} g([u]) \le \liminf_{\varepsilon\to 0} \sum_{t\in S_{\tilde{u}_\varepsilon}} g([\tilde{u}_\varepsilon]).$$

Proof. In the proof of the first statement, the difficulty stems from the lack of coercivity of W_ε with respect to $(u'_\varepsilon - 1)^+$. Nevertheless, note that for all $i,\ j$ in $\{1,\dots,n-1\}$,

$$\int_{i\varepsilon}^{j\varepsilon} |u_\varepsilon - \tilde{u}_\varepsilon|\, dt \le \frac{\varepsilon}{2} \int_{i\varepsilon}^{j\varepsilon} (u'_\varepsilon - 1)^+ dt,$$

and that

$$\sup_\varepsilon \int_0^1 (u'_\varepsilon - 1)^- dt < +\infty.$$

Thus for $(a,b) \subset\subset (0,1)$, where $a,\ b$ satisfy $\lim_{\varepsilon\to 0} u_\varepsilon(a) = u(a)$ and $\lim_{\varepsilon\to 0} u_\varepsilon(b) = u(b)$, one has

$$\sup_\varepsilon \int_a^b (u'_\varepsilon - 1)^+ dt < +\infty.$$

Let now $(a,b) \subset\subset (0,1)$ as above. From $\sup_\varepsilon \tilde{F}_\varepsilon(\omega, \tilde{u}_\varepsilon) < +\infty$ and using the previous estimate, the coercivity assumption on W and the fact that $\inf_{[0,+\infty)} g > 0$, we have

$$\begin{cases} \sup_\varepsilon \|\tilde{u}_\varepsilon\|_{BV(a,b)} < +\infty; \\ (\tilde{u}'_{\varepsilon,a})_\varepsilon \text{ equi-integrable on } (a,b); \\ \sup_\varepsilon H^0(S_{\tilde{u}_\varepsilon \lfloor (a,b)}) < +\infty. \end{cases}$$

Thus, according to compactness Theorem 13.4.4, u belongs to $SBV(a,b)$ and there exists a subsequence (not relabeled) satisfying

$$\begin{cases} \tilde{u}_\varepsilon \rightharpoonup u \text{ weakly in } BV(a,b), \\ \tilde{u}'_{\varepsilon,a} \rightharpoonup u'_a \text{ weakly in } L^1(a,b), \\ \displaystyle\sum_{t\in S_{\tilde{u}_\varepsilon \lfloor (a,b)}} [\tilde{u}_\varepsilon \lfloor (a,b)]\delta_t \rightharpoonup \sum_{t\in S_{u\lfloor (a,b)}} [u \lfloor (a,b)]\delta_t \text{ weakly in } M(a,b), \\ H^0(S_{u\lfloor (a,b)}) \le \liminf_{\varepsilon\to 0} H^0(S_{\tilde{u}_\varepsilon \lfloor (a,b)}). \end{cases}$$

Note that thanks to the coercivity assumption on W_j, $(\tilde{u}'_{\varepsilon,a})_\varepsilon$ is equi-integrable on $(0,1)$ and thus weakly converges to u'_a in $L^1(0,1)$. Moreover, u' is a Borel measure on $(0,1)$ as a nonnegative distribution, so that u belongs to $BV(0,1)$. Therefore u belongs to $SBV(0,1)$ owing to the characterization of the space $SBV(0,1)$ (see Theorem 10.5.1). To prove the last assertion, it is enough to establish

$$\sum_{t\in S_{u\lfloor (a,b)}} g([u]) \le \liminf_{\varepsilon\to 0} \sum_{t\in S_{\tilde{u}_\varepsilon}} g([\tilde{u}_\varepsilon])$$

and to let a go to 0 and b to 1. Set $\mu_\varepsilon = g([\tilde{u}_\varepsilon \lfloor (a,b)])\, H^0 \lfloor S_{\tilde{u}_\varepsilon \lfloor (a,b)}$. From $\sup_\varepsilon \tilde{F}_\varepsilon(\omega, \tilde{u}_\varepsilon) < +\infty$, up to a subsequence, μ_ε weakly converges to a Borel measure $\mu \in M(a,b)$. Let $\mu = \theta\, H^0 \lfloor S_{\tilde{u} \lfloor (a,b)} + \mu_s$ its Lebesgue–Nikodym decomposition with respect to the Borel measure $H^0 \lfloor S_{\tilde{u} \lfloor (a,b)}$. It suffices now to establish $\theta(t_0) \geq g([u](t_0))\ t_0$ for $H^0 \lfloor S_{\tilde{u} \lfloor (a,b)}$ a.e. t_0 in $(0,1)$. For $H^0 \lfloor S_{\tilde{u} \lfloor (a,b)}$ a.e. t_0 in $(0,1)$ and for a.e. $\rho > 0$, we have

$$\begin{aligned}
\theta(t_0) &= \lim_{\rho \to 0} \lim_{\varepsilon \to 0} \frac{\mu_\varepsilon(B_\rho(x_0))}{H^0 \lfloor S_{\tilde{u} \lfloor (a,b)}(B_\rho(x_0))} \\
&= \lim_{\rho \to 0} \lim_{\varepsilon \to 0} \mu_\varepsilon(B_\rho(x_0)) \\
&= \lim_{\rho \to 0} \lim_{\varepsilon \to 0} \sum_{t \in B_\rho(x_0) \cap S_{\tilde{u}_\varepsilon \lfloor (a,b)}} g([\tilde{u}_\varepsilon](t)) \\
&\geq \liminf_{\rho \to 0} \liminf_{\varepsilon \to 0} g\Big(\sum_{t \in B_\rho(x_0) \cap S_{\tilde{u}_\varepsilon \lfloor (a,b)}} [\tilde{u}_\varepsilon](t) \Big) \\
&\geq \liminf_{\rho \to 0} g\Big(\sum_{t \in B_\rho(x_0) \cap S_{u \lfloor (a,b)}} [\tilde{u}](t) \Big) \\
&= g([u](t_0)),
\end{aligned}$$

where we have used the subadditivity and continuity assumptions on g. □

We now establish the lower bound.

Proposition 14.2.3. *There exists a set Ω' of full probability such that for all ω in Ω' and u, u_ε in $L^1(0,1)$ satisfying $u_\varepsilon \to u$ strongly in $L^1(0,1)$,*

$$F(u) \leq \liminf_{\varepsilon \to 0} F_\varepsilon(\omega, u_\varepsilon).$$

Proof. We assume $\liminf_{\varepsilon \to 0} F_\varepsilon(\omega, u_\varepsilon) < +\infty$. Define v_ε in $W^{1,1}(0,1)$ by $v_\varepsilon(t) = \int_0^t \tilde{u}'_{a,\varepsilon}\, ds$. Since $\tilde{u}'_{a,\varepsilon}$ weakly converges to u'_a in $L^2(0,1)$, v_ε strongly converges in $L^2(0,1)$ to the function v of $L^2(0,1)$ defined by $v(t) = \int_0^t u'_a\, ds$. Therefore, according to Remark 14.2.1 seen for the first model, there exists a set Ω' of full probability such that, for all $\omega \in \Omega'$,

$$\liminf_{\varepsilon \to 0} \int_0^1 W\left(\omega, \frac{t}{\varepsilon}, v'_\varepsilon\right) dt \geq \int_0^1 W^{hom}(v')\, dt.$$

Therefore

$$\liminf_{\varepsilon \to 0} \int_0^1 W\left(\omega, \frac{t}{\varepsilon}, \tilde{u}'_{a,\varepsilon}\right) dt \geq \int_0^1 W^{hom}(u'_a)\, dt,$$

and $u'_a > 0$. We end the proof by applying Lemma 14.2.3. □

To conclude the proof of Theorem 14.2.4, we now establish the upper bound.

Proposition 14.2.4. *There exists a subset Ω'' of full probability such that for all $\omega \in \Omega''$, there exists a subsequence $(n_k)_{k \in \mathbf{N}}$ of positive integers satisfying: for all u in $L^1(0,1)$, there exists u_{n_k} in $L^1(0,1)$ possibly depending on ω, converging to u and such that* $\limsup_{k \to +\infty} F_{1/n_k}(\omega, u_{n_k}) \leq F(u)$.

PROOF. Without loss of generality, one may assume $F(u) < +\infty$ and $S_u = \{t_0\}$. Let $m \in \mathbf{N}^*$ and $i_0 \in \mathbf{N}$ be such that $t_0 \in [i_0/m, (i_0+1)/m)$ and set $I_m = (0, i_0/m) \cup ((i_0+1)/m, 1)$. According to Remark 14.2.1, where $(0,1)$ is replaced by $(0, i_0/m)$ or $((i_0+1)/m, 1)$, there exists a subset Ω'' of full probability which can be chosen independent of m, such that for all $\omega \in \Omega''$, there exists $w_n \in \mathcal{A}_{1/mn}(0, i_0/m) \cap \mathcal{A}_{1/mn}(i_0+1)/m, 1)$ satisfying

$$\begin{cases} \lim\limits_{n\to+\infty} w_n = u \text{ strongly in } L^1(I_m), \\ w_n\left(\dfrac{i_0}{m}\right) = u\left(\dfrac{i_0}{m}\right) \text{ and } w_n\left(\dfrac{i_0+1}{m}\right) = u\left(\dfrac{i_0+1}{m}\right), \\ \limsup\limits_{n\to+\infty} \displaystyle\int_{I_m} W(\omega, nmt, w_n')\, dt \le \int_{I_m} W^{hom}(u_a')\, dt \le \int_0^1 W^{hom}(u_a')\, dt. \end{cases}$$

Noticing that $W_{1/mn} \le W$, we obtain

$$\limsup_{n\to+\infty} \int_{I_m} W_{1/mn}(\omega, nmt, w_n')\, dt \le \int_0^1 W^{hom}(u_a')\, dt. \tag{14.12}$$

Consider now the function $w_{n,m}$ defined as follows:

$$\begin{cases} w_{n,m} = w_n \text{ on } I_m; \\ w_{n,m}\left(\dfrac{i_0}{m}\right) = u\left(\dfrac{i_0}{m}\right),\ w_{n,m}\left(\dfrac{i_0+1}{m}\right) = u\left(\dfrac{i_0+1}{m}\right); \\ w_{n,m} \text{ is affine on } \left(\dfrac{i_0}{m},\ \dfrac{i_0+1}{m} - \dfrac{1}{nm}\right) \text{ with } w_{n,m}' = 1; \\ w_{n,m} \text{ is affine on } \left(\dfrac{i_0+1}{m} - \dfrac{1}{nm}, \dfrac{i_0+1}{m}\right). \end{cases}$$

Clearly $w_{n,m}$ belongs to $\mathcal{A}_{1/mn}(0,1)$ and the slope e of its restriction to $(\frac{i_0+1}{m} - \frac{1}{nm}, \frac{i_0+1}{m})$ satisfies

$$\frac{1}{mn} e = u\left(\frac{i_0+1}{m}\right) - u\left(\frac{i_0}{m}\right) - \frac{n-1}{mn} > u\left(\frac{i_0+1}{m}\right) - u\left(\frac{i_0}{m}\right) - \frac{1}{m},$$

where the last term tends to $[u](t_0) > 0$. Since $\frac{1}{mn} e_{\omega_{i_0+1}, 1/mn}$ tends to 0, for n large enough we have $e > e_{\omega_{i_0+1}, 1/mn}$. A straightforward calculation then yields

$$\int_{(0,1)\setminus \overline{I}_m} W_{1/mn}(\omega, nmt, w_{n,m}')\, dt = g\left(u\left(\frac{i_0+1}{m}\right) - u\left(\frac{i_0}{m}\right) - \frac{1}{m}\right)$$

so that

$$\limsup_{m\to+\infty} \limsup_{n\to+\infty} \int_{(0,1)\setminus \overline{I}_m} W_{1/mn}(\omega, nmt, w_{n,m}')\, dt = g([u](t_0)). \tag{14.13}$$

Finally, combining (14.12) and (14.13), we obtain

$$\limsup_{m\to+\infty} \limsup_{n\to+\infty} \int_0^1 W_{1/mn}(\omega, nmt, w_{n,m}')\, dt \le F(u).$$

On the other hand, clearly

$$\limsup_{m\to+\infty}\limsup_{n\to+\infty} w_{n,m} = u \text{ strongly in } L^1(0,1)$$

and we complete the proof by a diagonalization argument as in Proposition 14.2.2. □

Remark 14.2.2. One can slightly generalize this model by assuming the function g of random type (see [159], [160]). More precisely, let $\{g_j,\ j \in J\}$ be a finite set of Lipschitz functions mapping $[0,+\infty)$ into $(0,+\infty)$, satisfying $\inf_{[0,+\infty)} g > 0$, and not necessarily subadditive. We set $\Omega = \{(w_j, g_j),\ j \in J\}^{\mathbf{Z}}$ and

$$W_{j,\varepsilon}(e) = \begin{cases} W_j(e) \text{ if } e \le 1, \\ W_j(e) \wedge \frac{1}{\varepsilon} g_j(\varepsilon(e-1)) \text{ if } e > 1, \end{cases}$$

which, as previously, leads to the random function defined by

$$W_\varepsilon(\omega, t, e) = \begin{cases} \omega_z^1(e) \text{ if } e \le e_{\omega_z,\varepsilon}, \\ \dfrac{1}{\varepsilon}\omega_z^2(\varepsilon(e-1)) \text{ if } e > e_{\omega_z,\varepsilon} \end{cases}$$

for all $\omega = ((\omega_z^1, \omega_z^2))_{z\in\mathbf{Z}}$ in Ω, t in $[z, z+1)$, and e in $\mathbf{R}$. Then, the corresponding total energy, almost surely Γ-converges to the functional

$$F(u) = \begin{cases} \displaystyle\int_0^1 W^{hom}(u_a')\,dt + \sum_{t\in S_u} g^{hom}([u](t)) \text{ if } u \in SBV^+(0,1), \\ +\infty \text{ otherwise.} \end{cases}$$

Setting $g(\omega, t, .) = \omega_z^2$ for all t in $[z, z+1)$, the density $g^{hom}(a)$ at $a \in \mathbf{R}$ is the almost sure deterministic limit of the process

$$\begin{cases} \mathcal{G}_{(-T,T)}(a) = \inf\Big\{ \displaystyle\sum_{t\in(-T,T)\cap S_v} g(\omega, t, [v](t)) \ :\ v \in SBV^+_{0,a}(-T,T)\Big\}, \\ SBV^+_{0,a}(-T,T) = \{v \in SBV^+(-T,T) \ :\ v_a' = 0,\ v(-T) = 0,\ v(T) = a\} \end{cases}$$

when T goes to $+\infty$. It is easily seen that $a \mapsto g^{hom}(a)$ is subadditive.

14.3 The Mumford–Shah model

Let us recall the Mumford–Shah model discussed in Section 12.4. Let Ω be a bounded open subset of $\mathbf{R}^N$ and g a given function in $L^\infty(\Omega)$. Denoting by $\mathcal{F}$ the class of the closed sets of Ω, for all K in $\mathcal{F}$ and all u in $\mathbf{C}^1(\Omega \setminus K)$ we defined the functional

$$E(u, K) := \int_\Omega |u-g|^2\,dx + \int_{\Omega\setminus K} |\nabla u|^2\,dx + \mathcal{H}^{N-1}(K)$$

and the associated optimization problem which is the strong formulation of the Mumford–Shah model in image segmentation:

$$\inf\{E(u, K) \ :\ (u, K) \in \mathbf{C}^1(\Omega \setminus K) \times \mathcal{F}\}. \tag{14.14}$$

When Ω is a rectangle in $\mathbf{R}^2$ and $g(x)$ is the light signal striking Ω at a point x, (14.14) is the Mumford–Shah model of image segmentation: K may be considered as the outline of the given light image in computer vision. In Section 12.4 we introduced the corresponding weak formulation,

$$m_w := \inf \left\{ \int_\Omega |u-g|^2\,dx + \int_\Omega |\nabla u|^2\,dx + \mathcal{H}^{N-1}(S_u) \ : \ u \in SBV(\Omega) \right\}, \qquad (14.15)$$

where ∇u denotes the density of the Lebesgue part of Du. One may now establish the weak existence result.

Theorem 14.3.1. *There exists at least a solution of the weak problem* (14.15).

Proof. We adopt the strategy of the so-called direct methods in the calculus of variations. From the hypothesis $g \in L^\infty(\Omega)$, we may assume that all the admissible functions u in (14.15) are uniformly bounded in $L^\infty(\Omega)$ by $|g|_{L^\infty(\Omega)}$. Indeed, let $c = |g|_{L^\infty(\Omega)}$ and consider the truncated function $u_c = c \wedge u \vee (-c)$. It is easily seen that u_c belongs to $SBV(\Omega)$ and satisfies

$$\begin{cases} S_{u_c} \subset S_u, \\ \displaystyle\int_\Omega |\nabla u_c|^2\,dx \le \int_\Omega |\nabla u|^2\,dx, \\ \displaystyle\int_\Omega |u_c - g|^2\,dx \le \int_\Omega |u-g|^2\,dx \end{cases}$$

(note that only the last inequality requires the explicit value of c). Let $(u_n)_n$ be a minimizing sequence of (14.15). It obviously satisfies

$$|u_n|_\infty + \int_\Omega |\nabla u_n|^2\,dx + H^{N-1}(S_{u_n}) \le C,$$

where C is a constant which does not depend on n. According to Theorem 13.4.3, there exists a subsequence $(u_{n_k})_k$ and a function u^* in $SBV(\Omega)$ such that

$$\begin{cases} u_{n_k} \to u^* \text{ in } L^1_{loc}(\Omega), \\ \nabla u_{n_k} \rightharpoonup \nabla u^* \text{ in } L^2(\Omega, \mathbf{R}^N), \\ \mathcal{H}^{N-1}(S_{u^*}) \le \liminf\limits_{k \longrightarrow +\infty} H^{N-1}(S_{u_{n_k}}). \end{cases}$$

According to Fatou's lemma and to the weak lower semicontinuity of the norm of the space $L^2(\Omega, \mathbf{R}^N)$, we deduce

$$\begin{aligned} &\int_\Omega |u^*-g|^2\,dx + \int_\Omega |\nabla u^*|^2\,dx + \mathcal{H}^{N-1}(S_{u^*}) \\ &\le \liminf_{k\to+\infty} \int_\Omega |u_{n_k}-g|^2\,dx + \liminf_{k\to+\infty} \int_\Omega |\nabla u_{n_k}|^2\,dx + \liminf_{k\to+\infty} \mathcal{H}^{N-1}(S_{u_{n_k}}) \\ &\le \liminf_{k\to+\infty} \left(\int_\Omega |u_{n_k}-g|^2\,dx + \int_\Omega |\nabla u_{n_k}|^2\,dx + \mathcal{H}^{N-1}(S_{u_{n_k}}) \right) = m_w, \end{aligned}$$

which proves that u^* is a solution of (14.15). □

We establish now the existence of a solution for the strong problem (14.14).

Theorem 14.3.2. *There exists at least a solution of the strong problem* (14.14).

PROOF. The proof proceeds in three steps.

First step. Let m_s be the value of the infimum of problem (14.14). We begin by showing that $m_w \leq m_s$. Indeed, arguing as in the previous proof, one may assume that all the admissible functions of (14.14) are uniformly bounded in $L^\infty(\Omega)$. Moreover, according to Example 10.5.1, for all K in $\mathcal{F}$ such that $\mathcal{H}^{N-1}(K) < +\infty$, the space $W^{1,1}(\Omega \setminus K) \cap L^\infty(\Omega)$ is included in $SBV(\Omega)$ and all its elements satisfy $\mathcal{H}^{N-1}(S_u \setminus K) = 0$.

Second step. We establish that any solution u^* of (14.15) satisfies

$$\begin{cases} u^* \in \mathbf{C}^1(\Omega \setminus \overline{S}_{u^*}), \\ \mathcal{H}^{N-1}(\overline{S}_{u^*} \cap \Omega \setminus S_{u^*}) = 0. \end{cases}$$

We only prove the first assertion. The second is more involved and we refer the reader to the paper of De Giorgi, Carriero, and Leaci [123]. Let $B_\rho(x)$ be the open ball centered at x, with radius ρ small enough so that $B_\rho(x) \subset \Omega \setminus \overline{S}_{u^*}$. Then u^* belongs to $W^{1,2}(B_\rho(x))$ and minimizes the problem

$$\inf \left\{ \int_{B_\rho(x)} |\nabla v|^2 \, dx + \int_{B_\rho(x)} |v - g|^2 \, dx \ : \ v \in u^* + W_0^{1,2}(B_\rho(x)) \right\}.$$

Thus u^* is a solution of the Dirichlet problem

$$\begin{cases} -\Delta v + v = g \text{ in } B_\rho(x), \\ v = u^* \text{ on } \partial B_\rho(x). \end{cases}$$

According to classical results on regularity properties of the solutions of Dirichlet problems (see, for instance, [90], [193]) we have $u^* \in \mathbf{C}^1(B_\rho(x))$.

Last step. Collecting the two previous steps we straightforwardly deduce that $m_w = m_s$ and that $(u^*, \overline{S}_{u^*} \cap \Omega)$ is a solution of (14.14). □

For other variational models in computer vision and image processing, see [183], [41], [42], [43], and references therein.

Chapter 15

Variational problems with a lack of coercivity

As we have seen in Section 3.2, every minimization problem of a coercive lower semicontinuous function admits a solution. On the other hand, without the coercivity assumption, in general we cannot apply the direct method and the existence of a minimizer may fail. This may occur even if the cost function is convex, as it happens, for instance, in the case

$$\min\{e^x \ : \ x \in \mathbf{R}\}.$$

However, some minimum problems, even if not coercive, still admit a solution, as, for instance, the case

$$\min\{x^2 \ : \ (x, y) \in \mathbf{R}^2\}$$

trivially shows.

In this chapter we present some methods which allow us to identify the noncoercive minimum problems which admit a solution. The history of these tools goes back to Stampacchia [210] and Fichera [135], who developed them to treat noncoercive cases in the framework of variational inequalities and of unilateral contact problems in elasticity, respectively. On the other hand, at least in the finite dimensional convex situations, the geometrical tool of *recession function* was introduced by Rockafellar in 1964, and this has been shown to be very useful in a large number of cases. The theory we present in Section 15.1 for the convex cases and in Section 15.2 for the general ones appeared first in the paper by Baiocchi et al. in 1988 [50] and makes it possible to treat in a unified way problems of geometrical type as well as problems coming from continuum mechanics.

15.1 Convex minimization problems and recession functions

In this section we will treat convex minimum problems, not necessarily coercive, and we will prove the existence of minimizers provided some compatibility conditions are satisfied.

The simplest example of a variational noncoercive minimization problem is the classical Neumann problem presented in Section 6.2,

$$\min\left\{\frac{1}{2}\int_\Omega |Du|^2\,dx - \langle L, u\rangle \ : \ u \in H^1(\Omega)\right\},$$

where Ω is a connected bounded Lipschitz domain of $\mathbf{R}^n$ and L belongs to the dual space $\big(H^1(\Omega)\big)'$. The Euler–Lagrange equation of the minimization problem above can be written in the weak form

$$\int_\Omega Du\,D\phi\,dx = \langle L, \phi\rangle \qquad \forall \phi \in H^1(\Omega). \tag{15.1}$$

Now, if the term L on the right-hand side of (15.1) is of the form

$$\langle L, \phi\rangle = \int_\Omega f\phi\,dx + \int_{\partial\Omega} g\phi\,d\mathcal{H}^{n-1}$$

with $f \in L^2(\Omega)$ and $g \in L^2(\partial\Omega)$, which we write as $L = f + g$, integrating by parts the left-hand side of (15.1) we obtain the PDE problem

$$\begin{cases} -\Delta u = f & \text{in } \Omega, \\ \dfrac{\partial u}{\partial \nu} = g & \text{on } \partial\Omega. \end{cases}$$

It is well known that a solution of the problem above exists iff the compatibility condition

$$\langle L, 1\rangle = 0 \tag{15.2}$$

is fulfilled. This can be seen in a simple way by remarking that if a solution of (15.1) exists, condition (15.2) follows straightforwardly by taking as a test function $\phi = 1$; on the other hand, if the compatibility condition (15.2) is fulfilled, then by the Poincaré inequality (Theorem 5.3.1), the minimization problem becomes coercive as soon as it is restricted to the class of functions in $H^1(\Omega)$ with zero average. Therefore we obtain the existence of a solution u for the problem written in weak form as

$$\int_\Omega Du\,D\psi\,dx = \langle L, \psi\rangle \qquad \forall \psi \in H^1(\Omega),\ \int_\Omega \psi\,dx = 0.$$

Now, this implies the existence of a solution for (15.1) by noticing that every function ϕ in $H^1(\Omega)$ can be written as $\phi = \psi + c$, where ψ is a function with zero average and $c \in \mathbf{R}$, so that condition (15.2) yields

$$\int_\Omega Du\,D\phi\,dx = \int_\Omega Du\,D\psi\,dx = \langle L, \psi\rangle = \langle L, \phi\rangle.$$

In this section, (V, σ) will denote a real locally convex Hausdorff topological vector space, and $F : V \to]-\infty, +\infty]$ will be a proper convex and sequentially σ-lower semicontinuous mapping. The minimization problem we are interested in is

$$\min\big\{F(v) \ :\ v \in V\big\}.$$

From the discussions above we know that the existence or nonexistence of a solution depends on some compatibility conditions that we want to identify. To do this we recall the classical definition of recession function introduced by Rockafellar [201] in the finite dimensional case (see also Sections 11.3 and 13.3).

Definition 15.1.1. *Given a proper convex and sequentially σ-lower semicontinuous functional $F : V \to]-\infty, +\infty]$, the recession functional F^∞ of F is defined, for every $v \in V$, by*

$$F^\infty(v) = \lim_{t \to +\infty} \frac{F(v_0 + tv)}{t}, \tag{15.3}$$

where v_0 is any element of $\operatorname{dom} F = \{v \in V \ : \ F(v) < +\infty\}$.

The main properties of the recession functional F^∞ are listed in the following proposition.

Proposition 15.1.1. *We have the following:*

(i) *The limit in* (15.3) *exists and is independent of v_0.*

(ii) *The functional F^∞ can equivalently be expressed by*

$$\begin{aligned} F^\infty(v) &= \sup\big\{F(u+v) - F(u) \ : \ u \in \operatorname{dom} F\big\} \\ &= \sup\left\{\frac{F(v_0 + tv) - F(v_0)}{t} \ : \ t > 0\right\} \end{aligned}$$

for any $v_0 \in \operatorname{dom} F$.

(iii) *F^∞ is proper, convex, sequentially σ-lower semicontinuous, and positively 1-homogeneous, that is,*

$$F^\infty(tv) = tF^\infty(v) \qquad \forall t \geq 0, \quad \forall v \in V.$$

(iv) *For every $F_1, \ldots, F_n$ proper, convex, sequentially σ-lower semicontinuous mappings, with* $(\operatorname{dom} F_1) \cap \ldots \cap (\operatorname{dom} F_n) \neq \emptyset$, *it is*

$$\left(\sum_{i=1}^n F_i\right)^\infty = \sum_{i=1}^n F_i^\infty.$$

(v) *$F^\infty(v) + F^\infty(-v) \geq 0$ for every $v \in V$.*

Proof. Let us prove that the limit in (15.3) exists. This follows from the fact that for every $v_0 \in \operatorname{dom} F$ and $v \in V$ the function $\phi(t) = F(v_0 + tv)$ is convex on $\mathbf{R}$; hence the mapping $t \mapsto \big(\phi(t) - \phi(0)\big)/t$ is nondecreasing and so it admits a limit as $t \to +\infty$. Moreover, the fact that the definition of F^∞ does not depend on $v_0 \in \operatorname{dom} F$ follows from property (ii).

Let us prove now that for every $v_0 \in \operatorname{dom} F$ and $v \in V$ it is

$$\lim_{t \to +\infty} \frac{F(v_0 + tv)}{t} = \sup\left\{\frac{F(v_0 + tv) - F(v_0)}{t} \ : \ t > 0\right\}.$$

The inequality $\leq$ is trivial because

$$\begin{aligned} \lim_{t \to +\infty} \frac{F(v_0 + tv)}{t} &= \lim_{t \to +\infty} \frac{F(v_0 + tv) - F(v_0)}{t} \\ &\leq \sup\left\{\frac{F(v_0 + tv) - F(v_0)}{t} \ : \ t > 0\right\}. \end{aligned}$$

To prove the opposite inequality, fix $s > 0$ and $t > s$; by the convexity of F we have

$$\begin{aligned} F(v_0 + sv) &= F\left(\left(1 - \frac{s}{t}\right)v_0 + \frac{s}{t}(v_0 + tv)\right) \\ &\leq \left(1 - \frac{s}{t}\right)F(v_0) + \frac{s}{t}F(v_0 + tv), \end{aligned}$$

so that

$$\frac{F(v_0 + sv) - F(v_0)}{s} \leq \frac{F(v_0 + tv) - F(v_0)}{t}.$$

By letting first $t \to +\infty$, and then taking the supremum for $s > 0$, we get

$$\sup\left\{\frac{F(v_0 + sv) - F(v_0)}{s} \ : \ s > 0\right\} \leq \lim_{t\to+\infty} \frac{F(v_0 + tv)}{t}.$$

To conclude the proof of (ii) it remains to prove that for every $v_0 \in \operatorname{dom} F$ and $v \in V$ the equality

$$\sup_{u\in\operatorname{dom} F} \big(F(u + v) - F(u)\big) = \sup_{t>0} \frac{F(v_0 + tv) - F(v_0)}{t} \tag{15.4}$$

holds. Let $u, v_0 \in \operatorname{dom} F$ and $v \in V$; by using the convexity and lower semicontinuity of F we obtain

$$\begin{aligned} F(u + v) &\leq \liminf_{t\to+\infty} F\left(\left(1 - \frac{1}{t}\right)u + \frac{1}{t}(v_0 + tv)\right) \\ &\leq \liminf_{t\to+\infty} \left[\left(1 - \frac{1}{t}\right)F(u) + \frac{1}{t}F(v_0 + tv)\right] \\ &= F(u) + \lim_{t\to+\infty} \frac{F(v_0 + tv) - F(v_0)}{t}; \end{aligned}$$

hence, inequality $\leq$ in (15.4) is proved. To prove the opposite inequality, we denote by S the left-hand side of (15.4); it is clear that without loss of generality we may assume $S < +\infty$. Then $u + v \in \operatorname{dom} F$ for every $u \in \operatorname{dom} F$ and so, from $F(u + v) \leq S + F(u)$, we deduce for every integer $k \geq 0$

$$F(u + kv) = F(u) + \sum_{i=1}^{k} \big[F(u + iv) - F(u + (i - 1)v)\big] \leq F(u) + kS.$$

Take now two nonnegative integers h, k; by using the convexity of F and the inequality above we obtain

$$\begin{aligned} F\left(u + \frac{h}{k}v\right) &= F\left(\left(1 - \frac{1}{k}\right)u + \frac{u + hv}{k}\right) \\ &\leq \left(1 - \frac{1}{k}\right)F(u) + \frac{1}{k}F(u + hv) \\ &\leq \left(1 - \frac{1}{k}\right)F(u) + \frac{1}{k}\left(F(u) + hS\right) = F(u) + \frac{h}{k}S. \end{aligned}$$

Finally, by using the lower semicontinuity of F, we have

$$F(u + tv) \leq F(u) + tS \qquad \forall t \geq 0.$$

Thus, taking $u = v_0$, we obtain

$$\frac{F(v_0 + tv) - F(v_0)}{t} \leq S \qquad \forall t > 0,$$

which concludes the proof of (15.4).

The fact that F^∞ is proper, convex, and sequentially σ-lower semicontinuous follows from assertion (ii). Indeed, for every $u \in \operatorname{dom} F$ the mapping $v \mapsto F(u + v) - F(u)$ is clearly convex and sequentially σ-lower semicontinuous, and so is F^∞, thanks to the well-known properties of supremum of convex functions. The fact that F^∞ is positively 1-homogeneous follows easily from the definition. Indeed, given $v \in V$ and $s > 0$, we have

$$F^\infty(sv) = \lim_{t \to +\infty} \frac{F(v_0 + tsv)}{t}$$

and, setting $\tau = ts$,

$$F^\infty(sv) = s \lim_{\tau \to +\infty} \frac{F(v_0 + \tau v)}{\tau} = sF^\infty(v).$$

Assertion (iv) follows straightforwardly by the definition of recession function. Finally, to prove (v), we may reduce ourselves to the case when both $F^\infty(v)$ and $F^\infty(-v)$ are finite; otherwise the statement is trivial. Therefore, by using (ii), we deduce that

$$u + v \in \operatorname{dom} F, \quad u - v \in \operatorname{dom} F \qquad \forall u \in \operatorname{dom} F;$$

hence, by using (ii) again, taking $u - v$ instead of u in the supremum associated to $F^\infty(v)$, we have

$$\begin{aligned} F^\infty(v) + F^\infty(-v) \geq & \sup_{u \in \operatorname{dom} F} \big(F(u) - F(u - v)\big) \\ & + \sup_{u \in \operatorname{dom} F} \big(F(u - v) - F(u)\big) \geq 0, \end{aligned}$$

which concludes the proof of Proposition 15.1.1. □

Here are some simple examples in which the recession functional can be explicitly computed.

Example 15.1.1. Let $V = \mathbf{R}$ and let $F(x) = e^x$ for every $x \in \mathbf{R}$. Then an easy calculation shows that in this case

$$F^\infty(x) = \begin{cases} 0 & \text{if } x \leq 0, \\ +\infty & \text{if } x > 0. \end{cases}$$

Analogously, if $V = \mathbf{R}^2$ and $F(x) = x_1^2$, then

$$F^\infty(x) = \begin{cases} 0 & \text{if } x_1 = 0, \\ +\infty & \text{if } x_1 \neq 0. \end{cases}$$

As we will see more precisely later, this example shows how the recession function indicates the *directions of coercivity*.

Example 15.1.2. Let $F : V \to [0, +\infty]$ be a nonnegative convex σ-lower semicontinuous functional which is positively homogeneous of degree $p > 1$, that is,

$$F(tv) = t^p F(v) \qquad \forall t > 0,\ \forall v \in V.$$

Then we have

$$F^\infty(v) = \begin{cases} 0 & \text{if } F(v) = 0, \\ +\infty & \text{otherwise.} \end{cases}$$

On the other hand, if F is positively homogeneous of degree 1, then it is clear that $F^\infty = F$.

Example 15.1.3. Let Ω be an open subset of $\mathbf{R}^n$ with a Lipschitz boundary, and let $V = W^{1,p}(\Omega; \mathbf{R}^m)$ $(p \geq 1)$ be the Sobolev space of all $\mathbf{R}^m$-valued functions which are in $L^p(\Omega)$ along with their first derivatives. Consider the functional

$$F(u) = \int_\Omega f(x, Du)\, dx \qquad \forall u \in W^{1,p}(\Omega, \mathbf{R}^m),$$

where

(i) $f : \Omega \times \mathbf{R}^{mn} \to [0, +\infty]$ is a Borel function,

(ii) for a.e. $x \in \Omega$ the function $f(x, \cdot)$ is convex and lower semicontinuous on $\mathbf{R}^{mn}$,

(iii) there exists $u_0 \in W^{1,p}(\Omega, \mathbf{R}^m)$ such that $F(u_0) < +\infty$.

It is well known (see, for instance, Section 13.1) that under the assumptions above, the functional F turns out to be proper, convex, and sequentially lower semicontinuous with respect to the weak topology of $W^{1,p}(\Omega, \mathbf{R}^m)$. Moreover, we have

$$F^\infty(u) = \int_\Omega f^\infty(x, Du)\, dx \qquad \forall u \in W^{1,p}(\Omega, \mathbf{R}^m),$$

where $f^\infty(x, \cdot)$ is the recession function of $f(x, \cdot)$. In fact, because of the convexity of f, for all $u \in W^{1,p}(\Omega, \mathbf{R}^m)$ the function

$$g(x, t) = \frac{f\big(x, Du_0(x) + t\,Du(x)\big) - f\big(x, Du_0(x)\big)}{t}$$

is nondecreasing with respect to t for a.e. $x \in \Omega$. Therefore the monotone convergence theorem gives

$$\begin{aligned} F^\infty(u) &= \lim_{t \to +\infty} \frac{F(u_0 + tu) - F(u_0)}{t} \\ &= \lim_{t \to +\infty} \int_\Omega g(x, t)\, dx = \int_\Omega f^\infty\big(x, Du(x)\big)\, dx. \end{aligned}$$

As a consequence, if

$$F(u) = \int_\Omega |Du|^p\, dx \qquad \forall u \in W^{1,p}(\Omega, \mathbf{R}^m)$$

with $p > 1$, we get

$$F^\infty(u) = \begin{cases} 0 & \text{if } u \text{ is locally constant in } \Omega, \\ +\infty & \text{otherwise.} \end{cases}$$

We are now in a position to give a first result on necessary conditions in convex minimization.

Proposition 15.1.2. *Assume that*

$$\inf \left\{ F(v) \ : \ v \in V \right\} > -\infty$$

(which always occurs, for instance, if F admits a minimum point on V). Then

$$F^\infty(v) \geq 0 \qquad \forall v \in V. \tag{15.5}$$

PROOF. Let m be the infimum of F on V, let v_0 be any point in dom F, and let $v \in V$. Since F is proper the infimum m is finite and so, by the definition of F^∞ we obtain

$$F^\infty(v) = \lim_{t \to +\infty} \frac{F(v_0 + tv)}{t} \geq \lim_{t \to +\infty} \frac{m}{t} = 0. \quad \square$$

Example 15.1.4. Let $P : V \to [0, +\infty]$ be a convex sequentially σ-lower semicontinuous functional which is positively homogeneous of degree $p > 1$ (for instance, the pth power of a seminorm), let $L \in V'$, and let F be the functional defined by

$$F(v) = P(v) - \langle L, v \rangle \qquad \forall v \in V.$$

By Example 15.1.2 we have

$$F^\infty(v) = \begin{cases} -\langle L, v \rangle & \text{if } P(v) = 0, \\ +\infty & \text{otherwise;} \end{cases}$$

hence, by the necessary condition of Proposition 15.1.2 we deduce that if the minimum problem

$$\min \left\{ F(v) \ : \ v \in V \right\}$$

admits a solution, then the linear functional L has to satisfy the compatibility condition

$$\langle L, v \rangle \leq 0 \qquad \forall v \in V \text{ with } P(v) = 0.$$

Example 15.1.5. The necessary condition of Proposition 15.1.2 is clearly not sufficient to obtain the existence of a minimizer, as the example of the function $F(x) = e^x$ with $V = \mathbf{R}$ shows. In fact, in this case it is $F^\infty \geq 0$ (see Example 15.1.1) but the function F has no minimum points on $\mathbf{R}$.

We give now an existence result for convex minimum problems without coercivity; the existence of a minimizer for the functional F will be obtained by adding to the necessary condition (15.5) some more requirements, namely, semicontinuity, compactness, and

compatibility conditions, in the sense specified below. We will use the notation $\ker F^\infty$ for the set $\{v \in V \ : \ F^\infty(v) = 0\}$, which, if condition (15.5) is satisfied, is a sequentially σ-closed convex cone.

Theorem 15.1.1. *Let V be a reflexive and separable Banach space with norm $\|\cdot\|$, and let $F : V \to]-\infty, +\infty]$ be a proper convex sequentially weakly lower semicontinuous functional. Assume that the following conditions are satisfied:*

(i) compactness: *if $t_h \to +\infty$, $v_h \to v$ weakly, and $F(t_h v_h)$ is bounded from above, then $\|v_h - v\| \to 0$;*

(ii) necessary condition: $F^\infty(v) \geq 0$ *for every $v \in V$;*

(iii) compatibility: $\ker F^\infty$ *is a linear subspace of V.*

Then the minimum problem

$$\min\{F(v) \ : \ v \in V\} \tag{15.6}$$

admits at least a solution.

Proof. For convenience, we divide the proof into several steps.

Step 1. For every $h \in \mathbf{N}$ consider the minimum problem

$$\min\{F(v) \ : \ v \in B_h\}, \tag{$\wp_h$}$$

where $B_h = \{v \in V \ : \ \|v\| \leq h\}$. Since F is sequentially weakly lower semicontinuous and B_h is sequentially weakly compact, by the direct method of the calculus of variations (see Section 3.2) we obtain that for every $h \in \mathbf{N}$ there exists a solution v_h of problem ($\wp_h$).

Step 2. If for some $h \in \mathbf{N}$ it is $\|v_h\| < h$, we claim that the proof of the theorem is achieved because such v_h is a solution of problem (15.6). Indeed, due to the convexity of F, for every $v \in V$ and every $\theta \in]0, 1[$

$$F\big(v_h + \theta(v - v_h)\big) \leq \theta F(v) + (1-\theta)F(v_h);$$

hence

$$F(v) - F(v_h) \geq \frac{1}{\theta}\Big[F\big(v_h + \theta(v - v_h)\big) - F(v_h)\Big].$$

Due to the definition of v_h, the right-hand side is nonnegative whenever $v_h + \theta(v - v_h) \in B_h$, which always occurs when θ is choosen small enough to have $\theta\big(\|v\| - \|v_h\|\big) \leq h - \|v_h\|$. Therefore, to conclude the proof of the theorem, it remains to show that the case

$$\|v_h\| = h \qquad \forall h \in \mathbf{N} \tag{15.7}$$

leads to existence of a solution of problem (15.6) too. Thus, we assume for the rest of the proof that (15.7) holds.

Step 3. For every $h \in \mathbf{N}$ set $w_h = v_h/h$; the sequence (w_h) is weakly compact in V. Then we may extract a subsequence $\big($which we still denote by $(w_h)\big)$ weakly converging to some $w \in V$. We have

$$F(hw_h) = F(v_h) \le F(v_0) < +\infty \tag{15.8}$$

for every h large enough, where v_0 is any point in dom F. Therefore, by the compactness assumption (i), we obtain $\|w_h - w\| \to 0$. Moreover, by using the lower semicontinuity and convexity of F, we have for every $t > 0$

$$\begin{aligned} F(tw + v_0) &\le \liminf_{h\to+\infty} F\left(tw_h + \left(1 - \frac{t}{h}\right)v_0\right) \\ &\le \liminf_{h\to+\infty} \left(\frac{t}{h}F(hw_h) + \left(1 - \frac{t}{h}\right)F(v_0)\right) \le F(v_0), \end{aligned}$$

where the last inequality follows from (15.8). Hence

$$F^\infty(w) = \lim_{t\to+\infty} \frac{F(tw + v_0)}{t} \le \lim_{t\to+\infty} \frac{F(v_0)}{t} = 0,$$

which, together with the necessary condition (ii), implies

$$w \in \ker F^\infty.$$

Step 4. By the compatibility condition (iii) we obtain $F^\infty(-w) = 0$, and this implies, by Proposition 15.1.1(ii), that

$$F(v - tw) \le F(v) \qquad \forall v \in V,\ \forall t \ge 0. \tag{15.9}$$

In particular, by taking $v = v_h$ and $t = h$, from the fact that $\|w_h - w\| \to 0$ we have $\|v_h - hw\| < h$ for h large enough, and then from (15.9) it follows that $v_h - hw$ is a solution of problem $(\wp_h)$ for h large enough, with $\|v_h - hw\| < h$. Hence we found a solution of problem $(\wp_h)$, for h large enough, with norm strictly less than h, and by repeating the argument used in Step 2, this provides a solution of the minimum problem (15.6)

Remark 15.1.1. In the case of nonreflexive Banach spaces V the result above still holds; it is enough to assume that V is a Banach space with norm $\|\cdot\|$ and to consider a topology σ on V coarser than the norm topology and such that (V, σ) is a Hausdorff vector space with the closed unit ball of $(V, \|\cdot\|)$ sequentially σ-compact. This happens, for instance, when V is the dual W' of a separable Banach space W and σ is the weak* topology on V. Then the proof above can be repeated if $F : V \to]-\infty, +\infty]$ is assumed to be proper convex and sequentially σ-lower semicontinuous and such that (ii), (iii) hold together with the compactness condition:

> if $t_h \to +\infty$, $v_h \to v$ with respect to σ, and $F(t_h v_h)$ is bounded from above, then $\|v_h - v\| \to 0$.

Remark 15.1.2. Notice that if dom F is bounded in $(V, \|\cdot\|)$, then conditions (i), (ii), and (iii) are automatically fulfilled. In fact, in this case the set $\ker F^\infty$ reduces to $\{0\}$, as can be seen immediately.

Remark 15.1.3. Consider the particular case when the so-called condition of Lions–Stampacchia type (see [173])

$$F^\infty(v) > 0 \qquad \forall v \neq 0 \tag{15.10}$$

holds. Then we obtain immediately that assumptions (ii) and (iii) are fulfilled. By Theorem 15.1.1 we obtain that, if the compactness assumption (i) also holds, then the set of solutions of problem (15.6) is nonempty. In this case the set of solutions is also bounded. In fact, by contradiction, assume there exist v_h solutions of (15.6) with $\|v_h\| \to +\infty$; arguing as in the proof of Theorem 15.1.1, it is possible to prove that the sequence $w_h = v_h/\|v_h\|$ converges in norm to some $w \in \ker F^\infty$. From assumption (15.10) it follows that $w = 0$, and this contradicts the fact that $\|w_h\| = 1$ for every h.

We consider now the particular case of quadratic forms on a Hilbert space. More precisely, let V be a separable Hilbert space, let $a : V \times V \to \mathbf{R}$ be a symmetric bilinear continuous form, and let $L \in V'$. Define

$$F(v) = \frac{1}{2}a(v, v) - \langle L, v\rangle \qquad \forall v \in V$$

and consider the minimum problem

$$\min\big\{F(v) \ : \ v \in V\big\}, \tag{15.11}$$

which is equivalent to the equation

$$a(v, \cdot) = L \qquad \text{in } V'.$$

On the bilinear form a we assume that the following conditions are satisfied:

$$a(v, v) \geq 0 \qquad \forall v \in V, \tag{15.12}$$

$$v_h \to 0 \text{ weakly, and } a(v_h, v_h) \to 0 \quad \Rightarrow \quad v_h \to 0 \text{ strongly.} \tag{15.13}$$

In this case, since $\ker(F^\infty - L) = (\ker F^\infty) \cap (\ker L)$, as a corollary of Theorem 15.1.1 we obtain the following result.

Proposition 15.1.3. *The minimum problem* (15.11) *admits a solution iff the compatibility condition*

$$L \textit{ is orthogonal to } \ker a \textit{ (i.e., } \langle L, v\rangle = 0 \textit{ whenever } a(v, v) = 0) \tag{15.14}$$

is fulfilled. Note that the condition above reads simply $\ker a \subset \ker L$.

PROOF. An easy computation shows that

$$F^\infty(v) = \begin{cases} -\langle L, v\rangle & \text{if } a(v, v) = 0, \\ +\infty & \text{if } a(v, v) \neq 0; \end{cases}$$

then, by Proposition 15.2, if problem (15.11) admits a solution, we must necessarily have $F^\infty \geq 0$, that is, (15.14). Conversely, let us assume (15.14) holds. The weak lower semicontinuity of the functional F is then a consequence of the continuity of the bilinear form a, the compactness assumption (i) follows from property (15.13), and finally the necessary condition (ii) and the compatibility condition (iii) follow from property (15.14). Therefore, by Theorem 15.1.1 we obtain that the minimum problem (15.11) admits a solution. □

Example 15.1.6. Consider the variational formulation of the classical Neumann problem

$$\min\left\{\frac{1}{2}\int_\Omega |Du|^2\,dx - \langle L, u\rangle \;:\; u \in H^1(\Omega)\right\},$$

where Ω is a bounded regular open subset of $\mathbf{R}^n$ and $L \in \left(H^1(\Omega)\right)'$. If we consider the bilinear form

$$a(u,v) = \int_\Omega Du\,Dv\,dx \qquad \forall u, v \in H^1(\Omega)$$

and apply Proposition 15.1.3, we obtain that a solution exists iff the compatibility condition

$$\langle L, 1\rangle = 0$$

is fulfilled. In terms of partial differential equations, when $L = f + g$ with $f \in L^2(\Omega)$ and $g \in L^2(\partial\Omega)$, this means that the problem

$$\begin{cases} -\Delta u = f & \text{in } \Omega, \\ \dfrac{\partial u}{\partial \nu} = g & \text{on } \partial\Omega \end{cases}$$

admits a solution iff

$$\int_\Omega f(x)\,dx + \int_{\partial\Omega} g(x)\,d\mathcal{H}^{n-1}(x) = 0.$$

For every subset K of V we denote by χ_K the indicator function of K defined by

$$\chi_K(v) = \begin{cases} 0 & \text{if } v \in K, \\ +\infty & \text{otherwise.} \end{cases}$$

Notice that if K is a nonempty sequentially σ-closed convex subset of V, then the function χ_K turns out to be a proper convex sequentially σ-lower semicontinuous mapping.

Definition 15.1.2. *For every nonempty sequentially σ-closed convex subset K of V we define the recession cone K^∞ of K by setting*

$$K^\infty = \mathrm{dom}(\chi_K)^\infty. \tag{15.15}$$

For every function $F : V \to]-\infty, +\infty]$ we denote by epi F the epigraph of F

$$\mathrm{epi}\, F = \left\{(v,t) \in V \times \mathbf{R} \;:\; t \geq F(v)\right\}. \tag{15.16}$$

We remark that epi F is a nonempty sequentially closed convex subset of $V \times \mathbf{R}$ whenever F is a proper sequentially σ-lower semicontinuous convex function.

Proposition 15.1.4. *If $F : V \to]-\infty, +\infty]$ is a proper convex sequentially σ-lower semicontinuous function, we have*

$$\text{epi } F^\infty = (\text{epi } F)^\infty.$$

PROOF. By Proposition 15.1.1(ii) it is

$$(v,t) \in \text{epi } F^\infty \iff t \geq F(u+v) - F(u) \quad \forall u \in \text{dom } F.$$

In other words, if $(s,u) \in \text{epi } F$ we have

$$F(u+v) \leq t + F(u) \leq t + s,$$

that is,

$$(v,t) + \text{epi } F \subset \text{epi } F,$$

or equivalently,

$$\chi_{\text{epi } F}\big((v,t) + (w,\alpha)\big) = 0 \qquad \forall (w,\alpha) \in \text{epi } F.$$

By Proposition 15.1.1 again, we obtain

$$\big(\chi_{\text{epi } F}\big)^\infty (v,t) = 0,$$

that is, $(v,t) \in (\text{epi } F)^\infty$. □

The definition given in (15.15) of recession cone is a very useful tool in convex analysis (see [201], [51], [52], and [79], where it is called "cône asymptote"); let us now state its main properties.

Proposition 15.1.5. *For any nonempty sequentially σ-closed convex subset K of V the set K^∞ is a convex sequentially σ-closed cone and the following properties hold true:*

$$K^\infty = \bigcap_{t>0} t^{-1}(K - v_0) \ \forall v_0 \in K, \tag{15.17}$$

$$K \text{ is a cone} \iff K^\infty = K, \tag{15.18}$$

$$0 \in K \iff K^\infty \subset K, \tag{15.19}$$

$$K \text{ is bounded} \Rightarrow K^\infty = \{0\}, \tag{15.20}$$

$$\text{dom } F^\infty \subset (\text{dom } F)^\infty \text{ whenever } F : V \to]-\infty, +\infty] \text{ is convex and lsc.} \tag{15.21}$$

Moreover, a point $w \in V$ belongs to K^∞ iff one of the following conditions is fulfilled:

$$k + w \in K \ \forall k \in K, \tag{15.22}$$

$$k + tw \in K \ \forall k \in K \ \forall t \geq 0, \tag{15.23}$$

$$\exists k \in K \ : \ k + tw \in K \ \forall t \geq 0. \tag{15.24}$$

PROOF. By Proposition 15.1.1(ii) and the definition (15.15) of K^∞ we have that $v \in K^\infty$ iff

$$\chi_K(v_0 + tv) - \chi_K(v_0) \leq 0 \qquad \forall t > 0,\ \forall v_0 \in K,$$

which is equivalent to

$$v + tv_0 \in K \qquad \forall t > 0,\ \forall v_0 \in K.$$

Therefore (15.17) follows.

The proof of (15.18) follows from Proposition 15.1.1(iii). Assertion (15.19) follows easily from (15.17) by taking $v_0 = 0$.

Always from (15.17) we obtain that K^∞ reduces to $\{0\}$ whenever K is bounded, that is, (15.20).

The proof of (15.21) simply follows by remarking that when $F^\infty(v) < +\infty$, then $v_0 + tv \in \operatorname{dom} F$ for every $t > 0$ and every $v_0 \in \operatorname{dom} F$.

Finally, (15.22), (15.23), (15.24) follow from Proposition 15.1.1(ii). □

Remark 15.1.4. We remark that the inclusion in (15.21) may be strict: this can be seen by taking, for instance, $F(v) = \|v\|^2$. In this case we have

$$\operatorname{dom} F = (\operatorname{dom} F)^\infty = V \qquad \text{whereas} \qquad \operatorname{dom} F^\infty = \{0\}.$$

In the finite dimensional case, implication (15.20) can be reversed, as the following proposition shows.

Proposition 15.1.6. *Let K be a convex closed subset of $\mathbf{R}^n$ such that $K^\infty = \{0\}$. Then K is bounded.*

PROOF. Assume by contradiction that K is unbounded; then for every $h \in \mathbf{N}$ there exists $x_h \in K$ with $|x_h| \geq h$. The sequence $y_h = x_h/|x_h|$ is bounded, so that we may extract a subsequence (which we still denote for simplicity by y_h) converging to some y with $|y| = 1$. We claim that $y \in K^\infty$. In fact, fix $t > 0$ and let x_0 be any point in K; since K is convex, for h large enough we have

$$\frac{t}{|x_h|} x_h + \left(1 - \frac{t}{|x_h|}\right) x_0 \in K,$$

and, since K is closed, passing to the limit as $h \to +\infty$,

$$ty + x_0 \in K,$$

which implies that y belongs to K^∞. This gives a contradiction because by assumption $K^\infty = \{0\}$ and $|y| = 1$. □

In general infinite dimensional topological vector spaces, Proposition 15.1.6 is false, as shown by the following example.

Example 15.1.7. Let V be a separable (infinite dimensional) Hilbert space, and let $(e_n)_{n \in \mathbf{N}}$ be a complete orthonormal system in V. Consider the set

$$K = \left\{ v \in V \ :\ |(v, e_n)| \leq n \text{ for every } n \in \mathbf{N} \right\},$$

where $(\cdot,\cdot)$ denotes the scalar product in V. It is easy to see that K is a nonempty convex weakly closed subset of V; moreover, we have $K^\infty = \{0\}$. In fact, if $v \in K^\infty$, by using the fact that $0 \in K$, we obtain from (15.17)

$$tv \in K \qquad \text{for every } t > 0,$$

that is,

$$|(v, e_n)| \le \frac{n}{t} \text{ for every } t > 0 \text{ and } n \in \mathbf{N}.$$

Therefore, as $t \to +\infty$, we get

$$(v, e_n) = 0 \ \forall n \in \mathbf{N},$$

and so $v = 0$. Nevertheless, K is unbounded, as it contains the points $v_n = ne_n$ for every $n \in \mathbf{N}$.

We will specialize now our existence results on noncoercive minimum problems to the case when the functional F can be written in the form

$$F(v) = J(v) - \langle L, v\rangle + \chi_K(v).$$

This situation arises, for instance, in many problems of mathematical physics, where J represents the stored energy functional depending on the nature of the body, L describes the action of the applied forces, and K is the set of admissible configurations which takes into account the physical constraints of the problems. In this case, the minimization problem we are dealing with takes the form

$$\min\big\{J(v) - \langle L, v\rangle \ : \ v \in K\big\}.$$

Theorem 15.1.2. *Assume that*

$$J : V \to [0, +\infty] \text{ is a proper convex sequentially } \sigma\text{-lsc functional}, \tag{15.25}$$

$$L : V \to \mathbf{R} \text{ is a linear } \sigma\text{-continuous functional}, \tag{15.26}$$

$$K \subset V \text{ is a nonempty convex sequentially } \sigma\text{-closed set}, \tag{15.27}$$

and consider the minimum problem

$$\min\big\{J(v) - \langle L, v\rangle \ : \ v \in K\big\}. \tag{15.28}$$

Then a necessary condition for the existence of at least a minimizer is

$$J^\infty(v) \ge \langle L, v\rangle \qquad \forall v \in K^\infty. \tag{15.29}$$

On the other hand, the minimum problem (15.28) *admits at least a solution provided the necessary condition* (15.29) *holds and the following compactness and compatibility conditions are fulfilled: if $t_h \to +\infty$, $v_h \in K$, $v_h \to v$ weakly, and $J(t_h v_h) - t_h\langle L, v_h\rangle$ is bounded from above, then*

$$\|v_h - v\| \to 0, \tag{15.30}$$

$$K^\infty \cap \ker(J^\infty - L) \text{ is a linear subspace of } V. \tag{15.31}$$

PROOF. Noticing that by Proposition 15.1.1(iv) the equality

$$(J - L + \chi_K)^\infty = J^\infty - L + \chi_{K^\infty}$$

holds, and taking Proposition 15.1.2 into account, we obtain that condition (15.29) is necessary for the existence of at least a solution of the minimum problem (15.28).

Analogously, compactness and compatibility conditions (i) and (iii) become in this case (15.30) and (15.31), so that the conclusion follows by Theorem 15.1.1. □

Remark 15.1.5. When J is a quadratic form, or more generally when J^∞ takes only the values 0 and $+\infty$ (which, for instance, occurs if J is positively p-homogeneous with $p > 1$; see Example 15.1.2, condition (15.31)), it can be written in the simpler form

$$K^\infty \cap \ker J^\infty \cap \ker L \qquad \text{is a linear subspace of } V, \tag{15.32}$$

which will be used in the following.

Let us discuss the structural assumptions of the existence result of Theorem 15.1.1 above.

Example 15.1.8. The compactness assumption (i) in the existence theorem, Theorem 15.1.1, cannot be dropped, as the following example shows. Let V be an infinite dimensional separable Hilbert space, and let $(e_n)_{n\in\mathbf{N}}$ be a complete orthonormal system in V. We denote by $(\cdot,\cdot)$ the scalar product in V, and we define a functional $F : V \to \mathbf{R}$ by setting

$$F(v) = \sum_{n\in\mathbf{N}} 2^{-n}|(v, e_n) - 1|^2 \qquad \forall v \in V.$$

It is easy to see that the functional F is finite-valued, convex, and weakly lower semicontinuous. Moreover, for every $v \in V$

$$\begin{aligned} F^\infty(v) &= \lim_{t\to+\infty} \frac{F(tv)}{t} \\ &= \lim_{t\to+\infty} \sum_{n\in\mathbf{N}} 2^{-n}\left[t|(v, e_n)|^2 - 2(v, e_n) + \frac{1}{t}\right]. \end{aligned}$$

Hence,

$$F^\infty(v) = \begin{cases} 0 & \text{if } v = 0, \\ +\infty & \text{if } v \neq 0, \end{cases}$$

so that the necessary condition (ii) and the compatibility condition (iii) of Theorem 15.1 are fulfilled. Nevertheless, the functional F does not admit any minimum point in V. In fact, taking for every $k \in \mathbf{N}$

$$v_k = \sum_{i=1}^{k} e_i$$

we get

$$\inf_{v\in V} F(v) \le F(v_k) = \sum_{n=k+1}^{\infty} 2^{-n}$$

so that

$$\inf_{v\in V} F(v) = 0.$$

But there are no points $v \in V$ such that $F(v) = 0$. Indeed, $F(v) = 0$ would imply $(v, e_n) = 1$ for every $n \in \mathbf{N}$, which contradicts the equality

$$\|v\|^2 = \sum_{n\in\mathbf{N}} |(v, e_n)|^2.$$

Example 15.1.9. The compatibility condition (iii) in Theorem 15.1.1 cannot be dropped, as the following example shows. Take $V = \mathbf{R}$ and define $F : \mathbf{R} \to]-\infty, +\infty]$ by

$$F(x) = \begin{cases} -\log x & \text{if } x > 0, \\ +\infty & \text{if } x \le 0. \end{cases}$$

The function F is proper, convex, and lower semicontinuous, and the compactness condition (i) is fulfilled since the dimension of V is finite. Moreover, a simple calculation yields

$$F^\infty(x) = \begin{cases} 0 & \text{if } x \ge 0, \\ +\infty & \text{if } x < 0, \end{cases}$$

and so the necessary condition (ii) is fulfilled too, but

$$\inf\big\{F(x) \ : \ x \in \mathbf{R}\big\} = -\infty.$$

We conclude this section by showing how Theorem 15.1.1 can be used to determine whether the algebraic difference of two closed convex sets in a Banach space is closed. Here we consider a reflexive Banach space V and two nonempty closed convex subsets A, B of V; the algebraic difference $A - B$ is defined by

$$A - B = \{a - b \ : \ a \in A, \ b \in B\}.$$

Example 15.1.10. We emphasize that even when dealing with cones in a finite dimensional space, the convex set $A - B$ may be not closed: take, for instance, $V = \mathbf{R}^3$ and

$$A = \big\{x \in \mathbf{R}^3 \ : \ x_1 \ge 0, \ x_2 \ge 0, \ x_3 \ge 0, \ x_1x_3 \ge x_2^2\big\},$$
$$B = \big\{x \in \mathbf{R}^3 \ : \ x_2 = x_3 = 0\big\}.$$

The sets A and B are closed convex cones, but a simple calculation gives

$$A - B = \big\{x \in \mathbf{R}^3 \ : \ x_2 > 0, \ x_3 > 0\big\} \cup \big\{x \in \mathbf{R}^3 \ : \ x_2 = 0, \ x_3 \ge 0\big\},$$

which is not closed. Indeed, for every $n \in \mathbf{N}$ we have that the point $(0, 1, 1/n)$ belongs to $A - B$, whereas their limit $(0, 1, 0)$ is not in $A - B$.

In the previous example the convex sets A and B are both unbounded. On the other hand, if A and B are two closed convex subsets of a reflexive Banach space and at least one of them is bounded, then $A - B$ is closed (weak and strong closedness coincide,

due to convexity). Indeed, if A is bounded and $x_h = a_h - b_h$ tends weakly to x with $a_h \in A$ and $b_h \in B$, then up to subsequences we have $a_h \to a \in A$ weakly, hence $b_h = x_h - a_h \to x - a \in B$ weakly, so that $x \in A - B$.

The following lemma characterizes the closed convex subsets of V.

Lemma 15.1.1. *Let K be a nonempty convex subset of a reflexive Banach space V. Then the following conditions are equivalent:*

(i) *K is closed.*

(ii) *For every $u \in V$ the function $v \mapsto \|u - v\|$ has a minimum on K.*

PROOF. Assume K is closed and let $u \in V$; set

$$M = \inf \big\{ \|u - v\| \ : \ v \in K \big\},$$

and for every $h \in \mathbf{N}$ let $v_h \in K$ be such that

$$\|u - v_h\| \le M + \frac{1}{h}.$$

The sequence (v_h) is bounded; since V is reflexive, possibly passing to subsequences, we may assume that v_h converges weakly to some v which belongs to K, because K is weakly closed (being strongly closed and convex). By the weak lower semicontinuity of the norm, we obtain

$$\|u - v\| \le \liminf_{h \to +\infty} \|u - v_h\| \le M,$$

which proves (ii).

Conversely, assume (ii) holds, and let (v_h) be a sequence in K strongly convergent to some $v \in V$. By (ii) there exists $\overline{w} \in K$ such that

$$\|\overline{w} - v\| \le \|w - v\| \qquad \forall w \in K.$$

In particular,

$$\|\overline{w} - v\| \le \|v_h - v\| \qquad \forall h \in \mathbf{N};$$

hence, as $h \to +\infty$, we obtain $\overline{w} = v$, and this proves that $v \in K$. □

We are now in a position to prove a closure result for the difference of two closed sets.

Theorem 15.1.3. *Let V be a reflexive Banach space, and let A and B be two nonempty closed convex subsets of V. Assume that A is locally compact for the strong topology and that*

$$A^\infty \cap B^\infty \text{ is a linear subspace.} \tag{15.33}$$

Then the convex set $A - B$ is closed.

PROOF. By Lemma 15.1.1 it is enough to prove that for every $u \in V$ the function $v \mapsto \|u - v\|$ has a minimum on $A - B$, or equivalently that the problem

$$\min \big\{ F(a, b) \ : \ (a, b) \in V \times V \big\} \tag{15.34}$$

has at least a solution, where

$$F(a,b) = \|u - a + b\| + \chi_A(a) + \chi_B(b).$$

We apply to the minimum problem (15.34) the existence theorem, Theorem 15.1.1. Since F is a nonnegative functional, the necessary condition $F^\infty \geq 0$ is immediately fulfilled. To prove the compactness condition (i) take $a_h \to a$ weakly in V, $b_h \to b$ weakly in V, $t_h \to +\infty$ such that $F(t_h a_h, t_h b_h) \leq C$; then by the definition of F we have

$$t_h a_h \in A, \qquad t_h b_h \in B, \qquad \|u - t_h a_h + t_h b_h\| \leq C.$$

Since the convex set A is assumed to be locally compact for the strong topology of V and $t_h a_h \in A$, the convergence $a_h \to a$ is actually strong; moreover, by $\|u - t_h a_h + t_h b_h\| \leq C$ we obtain that $\|b_h - a_h\| \to 0$ and so also $b_h \to a$ strongly.

Let us finally prove the compatibility condition (iii). By the definition of F we find

$$F^\infty(a,b) = \lim_{t\to+\infty} \frac{F(a_0 + ta, b_0 + tb)}{t},$$

where $a_0 \in A$ and $b_0 \in B$. Therefore

$$\begin{aligned} F^\infty(a,b) &= \lim_{t\to+\infty} \|a - b\| + \chi_A(a_0 + ta) + \chi_B(b_0 + tb) \\ &= \|a - b\| + \chi_{A^\infty}(a) + \chi_{B^\infty}(b) \end{aligned}$$

so that

$$\begin{aligned} \ker F^\infty &= \big\{(a,b) \ : \ a \in A^\infty,\ b \in B^\infty,\ a = b\big\} \\ &= \big\{(a,a) \ : \ a \in A^\infty \cap B^\infty\big\}. \end{aligned}$$

Therefore the compatibility condition (iii) follows from assumption (15.33). □

Remark 15.1.6. The requirement that A is locally compact for the strong topology is clearly fulfilled when A is finite dimensional. On the other hand, there are closed convex subsets A which are locally compact for the strong topology but not finite dimensional. For instance, it is enough to take in a separable Hilbert space V the set

$$A = \big\{v \in V \ : \ |(v, e_n)| \leq 1/n \text{ for all } n \in \mathbf{N}\big\},$$

where $(e_n)_{n\in\mathbf{N}}$ is a complete orthonormal system in V.

The results of this section allow us to study the problem of the lower semicontinuity for the inf-convolution of two convex functions. We recover some results of Section 9.2. If V is a reflexive Banach space and $f, g : V \to]-\infty, +\infty]$ are two proper convex functions, we recall that the inf-convolution $f \#_e g$ is defined by

$$(f \#_e g)(w) = \inf\{f(u) + g(v) \ : \ u, v \in V,\ u + v = w\}. \tag{15.35}$$

For instance, if $f = \chi_A$ and $g = \chi_B$, with A, B convex subsets of V, we have $f \#_e g = \chi_{A+B}$. Moreover, it is easy to see that in terms of epigraphs the inf-convolution operation turns out to simply reduce to the algebraic sum, that is,

$$\mathrm{epi}_{f\#_e g} = \mathrm{epi}_f + \mathrm{epi}_g.$$

As we saw in Example 15.1.10, it may happen that f and g are both lower semicontinuous but the lower semicontinuity does not occur for the inf-convolution $f\#_e g$. To see when $f\#_e g$ is lower semicontinuous, we investigate equivalently on the closedness of $\mathrm{epi}_{f\#_e g}$ and we obtain the following result.

Proposition 15.1.7. *Let $f, g : V \to]-\infty, +\infty]$ be two proper convex lower semicontinuous functions. Assume the following:*

(i) compactness: *if $t_h \to +\infty$, u_h and v_h converge weakly, $u_h + v_h \to 0$ strongly, and $f(t_h u_h) + g(t_h v_h)$ is bounded from above, then u_h and v_h converge strongly;*

(ii) compatibility: *if $f^\infty(v) + g^\infty(-v) \le 0$, then $f^\infty(-v) + g^\infty(v) \le 0$.*

Then the inf-convolution $f\#_e g$ defined in (15.35) *is lower semicontinuous.*

Proof. By Lemma 15.1.1 it is enough to show that for every $M \in \mathbf{R}$ and every $u \in V$ the function $(t, v) \mapsto |t - M| + \|v - u\|$ admits a minimum on $\mathrm{epi}_{f\#_e g}$. In other words, we have to show the existence of a solution for the minimum problem

$$\min\big\{|t - M| + \|v - u\| \ : \ t \ge (f\#_e g)(v)\big\}.$$

It is easy to see that for a fixed v, the optimal t in the minimum problem above is given by $t = M \vee (f\#_e g)(v)$, so that we have to show the existence of a solution for the problem

$$\min\big\{M \vee (f\#_e g)(v) + \|v - u\| \ : \ v \in V\big\}.$$

By the definition of inf-convolution this fact turns out to be equivalent to the existence of a solution for

$$\min\big\{M \vee \big(f(x) + g(y)\big) + \|x + y - u\| \ : \ x, y \in V\big\}.$$

Setting $\Phi(x, y) = M \vee \big(f(x) + g(y)\big) + \|x + y - u\|$ we apply to Φ the existence theorem, Theorem 15.1.1. The convexity and the weak lower semicontinuity of Φ follow straightforwardly. To prove the compactness assumption (i) of Theorem 15.1.1, take $x_h \to x$ and $y_h \to y$ weakly in V, and $t_h \to +\infty$ such that $\Phi(t_h x_h, t_h y_h)$ is bounded from above. Dividing by t_h we obtain

$$\limsup_{h\to+\infty} \left(\frac{f(t_h x_h)}{t_h} + \frac{g(t_h y_h)}{t_h} \right)^+ + \|x_h + y_h\| \le 0,$$

which implies that $x_h + y_h \to 0$ strongly in V. By the compactness assumption (i), the convergence of x_h and of y_h is actually strong in V.

Since $\Phi^\infty(x, y) = \big(f^\infty(x) + g^\infty(y)\big)^+ \|x + y\|$ the necessary condition (ii) of Theorem 15.1.1 is fulfilled. It remains to prove the compatibility condition (iii). Since

$$\ker \Phi^\infty = \big\{(x, y) \in V \times V \ : \ x + y = 0, \ f^\infty(x) + g^\infty(y) \le 0\big\},$$

the fact that $\ker \Phi^\infty$ is a subspace follows immediately from assumption (ii). □

Example 15.1.11. Let Ω be a bounded connected open subset of $\mathbf{R}^n$ and let X be the Sobolev space $H^1(\Omega)$. We denote by (x, y) the points of Ω, where x represents some k coordinates and y the remaining $n-k$. Consider the functionals

$$F(u) = \int_\Omega |D_x u|^2 \, dx \, dy - \langle f, u \rangle,$$

$$G(u) = \int_\Omega |D_y u|^2 \, dx \, dy - \langle g, u \rangle,$$

where f and g are in the dual space of X. The functionals F and G are convex and lower semicontinuous on $H^1(\Omega)$; we want to see if their inf-convolution $F\#_e G$ is still lower semicontinuous. By Proposition 15.1.7 above it is enough to verify the compactness assumption (i) and the compatibility assumption (ii).

If $t_h \to +\infty$, u_h and v_h converge weakly, $u_h + v_h \to 0$ strongly, and $F(t_h u_h) + G(t_h v_h)$ is bounded from above, then dividing by t_h^2 we have that

$$D_x u_h \to 0 \text{ in } L^2(\Omega), \qquad D_y v_h \to 0 \text{ in } L^2(\Omega),$$

which, together with the fact that $u_h + v_h \to 0$ strongly in $H^1(\Omega)$, implies that u_h and v_h converge actually strongly in $H^1(\Omega)$.

Finally, an easy calculation gives the expressions of the recession functions of F and G:

$$F^\infty(u) = \begin{cases} +\infty & \text{if } D_x u \neq 0, \\ -\langle f, u \rangle & \text{if } D_x u \equiv 0, \end{cases} \qquad G^\infty(v) = \begin{cases} +\infty & \text{if } D_y v \neq 0, \\ -\langle g, v \rangle & \text{if } D_y v \equiv 0. \end{cases}$$

Therefore, to prove the compatibility assumption, and hence the lower semicontiuity of $F\#_e G$, it is enough to assume that $\langle f - g, 1 \rangle = 0$.

15.2 Nonconvex minimization problems and topological recession

In this section we will consider general minimum problems of the form

$$\min \big\{ F(v) \ : \ v \in V \big\}, \tag{15.36}$$

where F is a possibly nonconvex functional noncoercive as well. Problems of this kind arise, for instance, in nonlinear elasticity (see Section 11.2), and due to the lack of convexity the results of previous sections cannot be applied. We will introduce a new kind of recession functional for general nonconvex functions and we will prove an abstract existence result under lower semicontinuity, compactness, and compatibility conditions, which will be expressed by means of this new tool.

As in Section 15.1, (V, σ) will denote a real locally convex Hausdorff topological vector space, and $F : V \to]-\infty, +\infty]$ will be a proper (not necessarily convex) mapping.

Definition 15.2.1. *The topological recession functional F_∞ of F is defined for every $v \in V$ by*

$$F_\infty(v) = \liminf_{\substack{t \to +\infty \\ w \to v}} \frac{F(tw)}{t}. \tag{15.37}$$

The main properties of the functional F_∞ are listed in the following proposition.

Proposition 15.2.1. *We have the following:*

(i) F_∞ *is* σ*-lower semicontinuous and positively homogeneous of degree* 1.

(ii) $F_\infty = F$ *whenever* F *is* σ*-lower semicontinuous and positively homogeneous of degree* 1.

(iii) $F_\infty = F^\infty$ *whenever* F *is proper, convex, and* σ*-lower semicontinuous.*

(iv) $(F+G)_\infty(v) \geq F_\infty(v) + G_\infty(v)$ *for every mapping* $G : V \to]-\infty, +\infty]$ *and for every* $v \in V$ *such that the sum at the right-hand side is defined.*

(v) *The equality* $(F+G)_\infty = F_\infty + G_\infty$ *holds in the following cases:*

(v_a) F *and* G *are proper, convex,* σ*-lower semicontinuous, and* $\operatorname{dom} F \cap \operatorname{dom} G \neq \emptyset$*;*

(v_b) G *is positively homogeneous of degree* 1*, finite, and* σ*-continuous;*

(v_c) F *is convex and* σ*-lower semicontinuous,* $F(0) < +\infty$*,* G *is* σ*-lower semicontinuous and positively homogeneous of degree* 1.

Proof. The proof of properties (i), (ii), (iv), (v_b) can be obtained immediately from Definition 15.2.1; property (v_a) follows from property (iii) and from Proposition 15.1.1(iv); property (v_c) follows from property (iv) and from the fact that using properties (ii), (iii), and Proposition 15.1.1(i), we get for every $v \in V$

$$\begin{aligned}(F+G)_\infty(v) &\leq \liminf_{t\to+\infty} \frac{F(tv)+G(tv)}{t} \\ &= \left(\liminf_{t\to+\infty} \frac{F(tv)}{t}\right) + G(v) \\ &= F^\infty(v) + G(v) = F_\infty(v) + G_\infty(v).\end{aligned}$$

It remains to prove property (iii). Let $v \in V$ and let $v_0 \in \operatorname{dom} F$ be fixed; taking $w_t = v + v_0/t$ we get

$$F_\infty(v) \leq \liminf_{t\to+\infty} \frac{F(tw_t)}{t} = \liminf_{t\to+\infty} \frac{F(v_0+tv)}{t} = F^\infty(v).$$

On the other hand, by using the convexity and the lower semicontinuity of F, for every $s > 0$ we have

$$\begin{aligned}F(v_0+sv) &\leq \liminf_{\substack{t\to+\infty \\ w\to v}} F\left(\left(1-\frac{s}{t}\right)v_0 + \frac{s}{t}tw\right) \\ &\leq \liminf_{\substack{t\to+\infty \\ w\to v}} \left[\left(1-\frac{s}{t}\right)F(v_0) + \frac{s}{t}F(tw)\right] \\ &= F(v_0) + s\liminf_{\substack{t\to+\infty \\ w\to v}} \frac{F(tw)}{t} = F(v_0) + sF_\infty(v).\end{aligned}$$

Therefore,

$$\frac{F(v_0 + sv) - F(v_0)}{s} \leq F_\infty(v)$$

for every $s > 0$, and passing to the limit as $s \to +\infty$, we obtain $F^\infty(v) \leq F_\infty(v)$. □

Analogously to what we made in Section 15.2 in the convex case, for every nonempty subset K of V we may define the topological recession cone K_∞ of K.

Definition 15.2.2. *Let K be a nonempty subset of V. The topological recession cone K_∞ of K is defined by*

$$K_\infty = \mathrm{dom}(\chi_K)_\infty. \tag{15.38}$$

Remark 15.2.1. By Definition 15.2.1 it follows that a point u belongs to K_∞ iff

$$\forall s > 0 \quad \forall U \in \Im(u) \quad \exists t > s \ : \ U \cap \frac{1}{t} K \neq \emptyset,$$

where $\Im(u)$ denotes the family of all σ-neighborhoods of u. In other words,

$$K_\infty = \bigcap_{s>0} \mathrm{cl}_\sigma \left(\bigcup_{t>s} \frac{1}{t} K \right), \tag{15.39}$$

where cl_σ denotes the closure with respect to σ.

The following proposition contains a list of properties of the topological recession cones.

Proposition 15.2.2. *We have the following:*

(i) *K_∞ is a σ-closed cone (possibly nonconvex);*

(ii) *$K_\infty = K$ whenever K is a σ-closed cone;*

(iii) *$K_\infty = K^\infty$ whenever K is a nonempty convex σ-closed set;*

(iv) *$K_\infty = \{0\}$ whenever K is bounded (the converse is false even in the convex case, as shown in Example* 15.1.7*);*

(v) *if V is finite dimensional, then $K_\infty = \{0\}$ implies that K is bounded;*

(vi) *$K_\infty = (K \cup H)_\infty = (K + H)_\infty$ for every bounded subset H of V;*

(vii) *$(\mathrm{epi}\, F)_\infty = \mathrm{epi}\, F_\infty$ for every proper function $F : V \to]-\infty, +\infty]$.*

Proof. Properties (i), (ii), (iii) follow from the definition (15.38) of K_∞ and from Proposition 15.2.1(i), (ii), (iii), respectively.

To prove property (iv), let U be a closed neighborhood of 0; since K is bounded, there exists $s > 0$ such that $K \subset tU$ for every $t > s$. By the characterization of K_∞ given by (15.39) this implies that $K_\infty \subset U$, and since U is arbitrary, we get $K_\infty = \{0\}$.

Let us prove property (v). Assume by contradiction that K is unbounded; then for every $h \in \mathbf{N}$ there exists $x_h \in K$ with $|x_h| \geq h$. Since the sequence $y_h = x_h/|x_h|$ is bounded, we may extract a subsequence (still denoted by y_h) converging to some $y \in V$ with $|y| = 1$. Therefore

$$\begin{aligned}(\chi_K)_\infty(y) &\leq \liminf_{h\to+\infty} \chi_K(|x_h|y_h) \\ &= \liminf_{h\to+\infty} \chi_K(x_h) = 0,\end{aligned}$$

so that $y \in K_\infty$, and this is impossible because $y \neq 0$.

Let us prove property (vi). Since the inclusion $K_\infty \subset (K \cup H)_\infty$ is obvious, it is enough to prove the opposite inclusion. Let $x \in (K \cup H)_\infty$; if $x = 0$ we have $x \in K_\infty$ because, by (i), K_∞ is a closed cone. If $x \neq 0$ let U_0, U be two disjoint neighborhoods of 0 and x, respectively. Since H is bounded, there exists $s_0 > 0$ such that $H \subset tU_0$ for every $t > s_0$; moreover, since $x \in (K \cup H)_\infty$, by (15.39) we have

$$\forall s > 0 \quad \forall W_x \quad \exists t > s \ : \ W_x \cap \frac{1}{t}(K \cup H) \neq \emptyset, \tag{15.40}$$

where we denoted by W_x a generic neighborhood of x. When $s \geq s_0$ and $W_x \subset U$ we have

$$W_x \cap \frac{1}{t}H \subset U \cap U_0 = \emptyset \qquad \forall t > s,$$

so that, by (15.40),

$$W_x \cap \frac{1}{t}K \neq \emptyset.$$

By (15.39) this proves that $x \in K_\infty$. The equality $K_\infty = (K \cup H)_\infty$ can be proved in a similar way.

Let us now prove property (vii). If $(u, \xi) \in (\text{epi } F)_\infty$, denoting by U and I generic neighborhoods of u and ξ, respectively, by (15.39) we have

$$\forall s > 0 \quad \forall U \quad \forall I \quad \exists t > s \quad \exists v \in U \quad \exists \eta \in I \ : \ F(tv) \leq t\eta.$$

Therefore $F_\infty(u) \leq \xi$, so that $(u, \xi) \in \text{epi } F_\infty$. On the other hand, if $(u, \xi) \in \text{epi } F_\infty$, we have $F_\infty(u) \leq \xi$; hence by the definition of F_∞ we obtain

$$\forall s > 0 \quad \forall U \quad \forall \eta > \xi \quad \exists t > s \quad \exists v \in U \ : \ \frac{F(tv)}{t} < \eta.$$

Therefore $(u, \xi) \in \text{epi } F_\infty$. □

Remark 15.2.2. It can be useful to give a characterization of the topological recession functional F_∞ in terms of converging nets: actually, it is easy to show that for every $v \in V$

$$F_\infty(v) = \inf \left\{ \liminf_{\lambda\in\Lambda} \frac{F(t_\lambda v\lambda)}{t_\lambda} \ : \ t_\lambda \to +\infty, \ v_\lambda \to v \right\}, \tag{15.41}$$

where Λ is an arbitrary directed set and (t_λ), (v_λ) are nets indexed by Λ. Recall that a directed set is a set together with a relation $\geq$ which is both transitive and reflexive such that for any two elements a, b there exists another element c with $c \geq a$ and $c \geq b$.

Definition (15.37) or the equivalent one (15.41) is very general and provides a good extension of the notion of convex recession functional introduced in the last section. However, in many situations, it is not easy to deal with neighborhoods or nets; for this reason, we introduce now a new definition of recession functional, in which only the behavior of converging sequences is involved. More precisely, for every $v \in V$ we set

$$F_\infty^{seq}(v) = \inf\left\{\liminf_{h\to+\infty} \frac{F(t_h v_h)}{t_h} \ : \ t_h \to +\infty,\ v_h \to v\right\}, \tag{15.42}$$

where (t_h) and (v_h) are sequences. It is clear from (15.41) and (15.42) that $F_\infty \le F_\infty^{seq}$, and that $F_\infty = F_\infty^{seq}$ whenever the space (V,σ) is metrizable. Moreover, by a proof similar to the one of Proposition 15.2.1, it is possible to show that the following properties for F_∞^{seq} hold.

Proposition 15.2.3. *We have the following:*

(i) F_∞^{seq} *is positively homogeneous of degree* 1.

(ii) $F_\infty^{seq} = F$ *whenever* F *is sequentially* σ*-lower semicontinuous and positively homogeneous of degree* 1.

(iii) $F_\infty^{seq} = F$ *whenever* F *is proper, convex, and sequentially* σ*-lower semicontinuous.*

(iv) $(F+G)_\infty^{seq}(v) \ge F_\infty^{seq}(v) + G_\infty^{seq}(v)$ *for every mapping* $G : V \to]-\infty,+\infty]$ *and for every* $v \in V$ *such that the sum at the right-hand side is defined.*

(v) *The equality* $(F+G)_\infty^{seq}(v) = F_\infty^{seq}(v) + G_\infty^{seq}(v)$ *holds in the following cases:*

(v_a) F *and* G *are proper, convex, sequentially* σ*-lower semicontinuous, and* $\mathrm{dom}\, F \cap \mathrm{dom}\, G \neq \emptyset$*;*

(v_b) G *is positively homogeneous of degree* 1 *and sequentially* σ*-continuous;*

(v_c) F *is convex and sequentially* σ*-lower semicontinuous,* $F(0) < +\infty$*,* G *is sequentially* σ*-lower semicontinuous and positively homogeneous of degree* 1.

Remark 15.2.3. We point out that in general the functional F_∞^{seq} is neither σ-lower semicontinuous nor sequentially σ-lower semicontinuous. However, this simpler definition will be sufficient to obtain an existence theorem for minimizers (Theorem 15.2.1) which will be used in many applications. We stress the fact that the explicit computation of F_∞^{seq} may be difficult; nevertheless, to apply the existence result of Theorem 15.2.1, in many cases it will be enough to show qualitative properties of F_∞^{seq} which are easy to obtain thanks to Proposition 15.2.3.

In an analogous way, for every nonempty subset K of V we may introduce the sequential recession cone K_∞^{seq} by setting

$$K_\infty^{seq} = \mathrm{dom}(\chi_K)_\infty^{seq}. \tag{15.43}$$

In other words, it is

$$x \in K_\infty^{seq} \quad \Longleftrightarrow \quad \exists t_h \to +\infty \quad \exists x_h \to x \quad \forall h \in \mathbf{N} \quad t_h x_h \in K.$$

The set K_∞^{seq} turns out to be a cone (possibly not sequentially σ-closed) with vertex 0 and, by a proof similar to the one of Proposition 15.2.2, we obtain the following properties.

Proposition 15.2.4. *We have the following:*

(i) $K_\infty^{seq} = K$ *if* K *is a sequentially* σ*-closed cone.*

(ii) $K_\infty^{seq} = K^\infty$ *if* K *is a nonempty sequentially* σ*-closed set.*

(iii) $K_\infty^{seq} = \{0\}$ *if* K *is bounded.*

(iv) $K_\infty^{seq} = \{0\}$ *implies* K *bounded if* V *is finite dimensional.*

(v) $K_\infty^{seq} = (K \cup H)_\infty^{seq} = (K + H)_\infty^{seq}$ *if* $H \subset V$ *is bounded.*

The following result provides a general necessary condition for the existence of minimizers.

Proposition 15.2.5. *Assume that*

$$\inf \left\{ F(v) \ : \ v \in V \right\} > -\infty.$$

Then we have

$$F_\infty(v) \geq 0 \ \ \forall v \in V \ \ \textit{(hence } F_\infty^{seq} \geq 0\textit{, because } F_\infty \leq F_\infty^{seq}\textit{).}$$

PROOF. Let m be the infimum of F on V, and let $v \in V$. By the definition of F_∞ we get

$$F_\infty(v) = \liminf_{\substack{t \to +\infty \\ w \to v}} \frac{F(tw)}{t} \geq \liminf_{t \to +\infty} \frac{m}{t} = 0. \qquad \square$$

We recall that even in the convex case, the condition $F^\infty \geq 0$ is not sufficient for the existence of a solution of problem

$$\min \left\{ F(v) \ : \ v \in V \right\} \tag{15.44}$$

(see, for instance, Example 15.1.9). To show an existence result for problem (15.44), analogously to what was done in Section 15.1 we set

$$\ker F_\infty^{seq} = \left\{ v \in V \ : \ F_\infty^{seq}(v) = 0 \right\}.$$

Theorem 15.2.1. *Assume* V *is a Banach space with norm* $\| \cdot \|$*, let* σ *be a topology on* V *coarser than the norm topology and such that* (V, σ) *is a Hausdorff vector space with the closed unit ball of* $(V, \| \cdot \|)$ *sequentially* σ*-compact, and let* $F : V \to]-\infty, +\infty]$ *be a sequentially* σ*-lower semicontinuous functional. Assume also that the following conditions are satisfied:*

(i) compactness: *if* $t_h \to +\infty$, $v_h \to v$ *with respect to* σ, *and* $F(t_h v_h)$ *is bounded from above, then* $\|v_h - v\| \to 0$;

(ii) necessary condition: $F_\infty^{seq}(v) \geq 0$ *for every* $v \in V$;

(iii) compatibility: *for every* $u \in \ker F_\infty^{seq}$ *there exists* $t > 0$ *such that* $F(v - tu) \leq F(v)$ *for all* $v \in V$.

Then the minimum problem (15.44) *has at least a solution.*

PROOF. The proof is similar to the one of Theorem 15.1, and we will follow it step by step.

Step 1. For every $h \in \mathbf{N}$ let v_h be a solution of the minimum problem

$$\min\big\{F(v) \ : \ \|v\| \leq h\big\}. \tag{$\wp_h$}$$

Since F is sequentially σ-lower semicontinuous and $\{v \in V \ : \ \|v\| \leq h\}$ is sequentially σ-compact, by the direct method of the calculus of variations (see Corollary 3.2.3) we obtain that there exists a solution v_h of problem $(\wp_h)$. Moreover, by the lower semicontinuity of F and by the assumptions on σ, we may choose v_h such that

$$\|v_h\| = \min\big\{\|w\| \ : \ w \text{ solves } (\wp_h)\big\}. \tag{15.45}$$

Step 2. If the sequence (v_h) is bounded in norm, then by the lower semicontinuity of F and by the σ-compactness of bounded sets, the existence of a solution of problem (15.44) follows from the direct method of the calculus of variations. Indeed, if (v_{h_k}) is a subsequence of (v_h) which is σ-convergent to some $\overline{v} \in V$, we have

$$F(\overline{v}) \leq \liminf_{k \to +\infty} F(v_{h_k}) = \inf\big\{F(v) \ : \ v \in V\big\}.$$

Step 3. It remains to show that the case (v_h) unbounded cannot occur. By contradiction, assume that a subsequence of $\|v_h\|$ (which we still index by h) tends to $+\infty$. Since the normalized vectors $w_h = v_h/\|v_h\|$ are bounded, there exists a subsequence of (w_h) (which we still index by h) which σ-converges to some $w \in V$. We have

$$F(v_{h+1}) \leq F(v_h) \ \ \forall h \in \mathbf{N},$$

so that $F(v_h)$ is bounded from above. Hence

$$F_\infty^{seq}(w) \leq \liminf_{h \to +\infty} \frac{F(\|v_h\| w_h)}{\|v_h\|} = \liminf_{h \to +\infty} \frac{F(v_h)}{\|v_h\|} \leq 0,$$

which, togheter with the necessary condition (ii), gives

$$w \in \ker F_\infty^{seq}. \tag{15.46}$$

Step 4. We have $\|v_h\| \to +\infty$, $w_h \to w$ with respect to σ, and $F(\|v_h\| w_h) = F(v_h)$ bounded from above. By the compactness assumption (i) we obtain $\|w_h - w\| \to 0$, and this prevents w from being zero, because $\|w_h\| = 1$ for all $h \in \mathbf{N}$. From (15.46) and from the compatibility condition (iii) we get that there exists $t > 0$ such that

$$F(v_h - tw) \leq F(v_h) \ \ \forall h \in \mathbf{N}. \tag{15.47}$$

Finally,

$$\begin{aligned} \|v_h - tw\| &= \left\| \left(1 - \frac{t}{\|v_h\|}\right) v_h + t(w_h - w) \right\| \\ &\leq \left(1 - \frac{t}{\|v_h\|}\right) \|v_h\| + t\|w_h - w\| \\ &= \|v_h\| + t\big(\|w_h - w\| - 1\big). \end{aligned}$$

The right-hand side of the last equality is strictly less than $\|v_h\|$ for h large enough, and this is in contradiction with (15.47) and (15.45). □

Remark 15.2.4. Looking at the proof of Theorem 15.2.1 we see that the compactness condition (i) can be imposed only on sequences (v_h) which σ-converge to an element $v \in \ker F_\infty^{seq}$.

Remark 15.2.5. In Theorem 15.2.1 a weaker form of the compatibility condition (iii) can be used, namely,

> for every $u \in \ker F_\infty^{seq}$ there exists $R > 0$ such that for every $v \in V$ with $\|v\| \geq R$ there exists $t > 0$ such that $F(v - tu) \leq F(v)$.

Moreover, an inspection of the proof of Theorem 15.2.1 shows that in the case of a product space $V_1 \times \cdots \times V_n$ the compatibility condition (iii) can be replaced by the following weaker one:

> for every $(u_1, \ldots, u_n) \in \ker F_\infty^{seq}$ there exist positive numbers $t_1, \ldots, t_n$ such that
>
> $$F(v_1 - t_1 u_1, \ldots, v_n - t_n u_n) \leq F(v_1, \ldots, v_n) \qquad \forall\, (v_1, \ldots, v_n) \in V_1 \times \cdots \times V_n.$$

Remark 15.2.6. Theorem 15.2.1 includes in a certain sense the classical direct method of the calculus of variations, which gives the existence of a solution of problem (15.44) under the following assumptions:

(i) F is sequentially σ-lower semicontinuous.

(ii) There exist $\alpha > 0$ and $b \in \mathbf{R}$ such that

$$F(v) \geq \alpha\|v\| + b \ \ \forall v \in V.$$

Indeed, by (ii) we obtain

$$t_h \to +\infty,\ v_h \to v,\ F(t_h v_h) \le c \quad \Rightarrow \quad \|v_h\| \to 0 \text{ and } v = 0,$$

so that the compactness hypothesis (i) is satisfied. Moreover, by (iii) again we get

$$F_\infty^{seq}(v) \ge \alpha \|v\| \ \ \forall v \in V,$$

so that $\ker F_\infty^{seq}$ reduces to $\{0\}$. Hence, hypotheses (ii) and (iii) of Theorem 15.2.1 are also satisfied, and problem (15.44) admits at least a solution. In particular (ii) holds if $\operatorname{dom} F$ is bounded.

As in Section 15.1, we now specialize our results to the case when the functional F is of the form

$$F(v) = J(v) - \langle L, v\rangle + \chi_K(v),$$

where

$J : V \to [0, +\infty]$ is a sequentially σ-lower semicontinuous functional;
$L : V \to \mathbf{R}$ is a linear σ-continuous functional;
$K \subset V$ is a sequentially σ-closed set.

Taking into account Propositions 15.39 and 15.40, we have that a necessary condition for the existence of a solution of the minimum problem

$$\min\left\{J(v) - \langle L, v\rangle \ :\ v \in K\right\}$$

is given by

$$J_\infty^{seq}(v) \ge \langle L, v\rangle \ \ \forall v \in K, \tag{15.48}$$

whereas the compatibility condition (iii) can be written as

(iii′) for every $u \in K_\infty^{seq}$ with $J_\infty^{seq}(u) = \langle L, u\rangle$, there exists $t > 0$ such that for every $v \in K$

$$\begin{cases} v - tu \in K, \\ J(v - tu) + t\langle L, u\rangle \le J(v). \end{cases}$$

In many situations the functional J has a "superlinear growth" so that the functional J_∞^{seq} reduces to

$$J_\infty^{seq}(v) = \begin{cases} 0 & \text{if } v \in \ker J_\infty^{seq}, \\ +\infty & \text{otherwise;} \end{cases}$$

in this case the necessary condition (ii) becomes

(ii′) $\langle L, v\rangle \le 0 \ \ \forall v \in K_\infty^{seq} \cap \ker J_\infty^{seq}$,

whereas the compatibility condition (iii) is given by (iii′) with $u \in K_\infty^{seq} \cap \ker J_\infty^{seq} \cap \ker L$.

15.3 Some examples

In this section we show some examples that can be treated with the theory of noncoercive minimum problems developed in the previous sections of the chapter. Some other cases with applications to problems from mechanics were presented in chapter 14.

Consider the following minimization problem:

$$\min\left\{\int_\Omega\left(\frac{1}{2}|Dv|^2+B(v)\right)dx-\langle L,v\rangle:\quad u\in H^1(\Omega)\right\},\tag{15.49}$$

where Ω is a bounded connected open subset of $\mathbf{R}^n$, L is in the dual space of $H^1(\Omega)$, and $B:\mathbf{R}\to]-\infty,+\infty]$ is a convex lower semicontinuous function.

Proposition 15.3.1. *For the existence of a solution of problem* (15.49) *the assumption*

$$-B^\infty(-1)\le\frac{\langle L,1\rangle}{\operatorname{meas}(\Omega)}\le B^\infty(1)\tag{15.50}$$

is necessary, while

$$-B^\infty(-1)<\frac{\langle L,1\rangle}{\operatorname{meas}(\Omega)}<B^\infty(1)\tag{15.51}$$

is sufficient. The case $-B^\infty(-1)=B^\infty(1)$ corresponds to an affine function B for which the condition $B^\infty(1)\operatorname{meas}(\Omega)=\langle L,1\rangle$ is necessary and sufficient for the existence.

Proof. It is immediate to see that the functional

$$F(u)=\int_\Omega\left(\frac{1}{2}|Dv|^2+B(v)\right)dx-\langle L,v\rangle$$

is sequentially weakly lower semicontinuous on $H^1(\Omega)$. To verify the compactness property (i) of Theorem 15.1.1 we notice that, being the function B convex and lower semicontinuous, we have for suitable $a,b\in\mathbf{R}$

$$B(s)\ge as+b\qquad\forall s\in\mathbf{R}.\tag{15.52}$$

Therefore, if $t_h\to+\infty$, $v_h\to v$ weakly, and $F(t_hv_h)$ is bounded, we obtain from (15.52)

$$\int_\Omega t_h^2|Dv_h|^2\,dx-ct_h\|v_h\|_{H^1}\le c$$

for a suitable positive constant c. Therefore, dividing by t_h^2, we obtain that Dv_h strongly converges to zero in $L^2(\Omega)$ and hence v_h strongly converges in $H^1(\Omega)$ to a constant function.

The computation of the recession functional F^∞ gives for every $v\in H^1(\Omega)$

$$F^\infty(v)=\begin{cases}\int_\Omega B^\infty(v)-\langle L,v\rangle & \text{if } v \text{ is constant,}\\ +\infty & \text{otherwise.}\end{cases}$$

Therefore, the necessity of condition (15.50) follows from Proposition 15.1.2.

To apply the existence Theorem 15.1.1 it remains to prove the compatibility condition (iii). By above expression of F^∞, we have that $v \in \ker F^\infty$ iff v is constant and

$$\operatorname{meas}(\Omega) B^\infty(v) = v\langle L, 1\rangle.$$

Thus (iii) is a consequence of assumption (15.51). □

Remark 15.3.1. Consider the partial differential equation

$$\begin{cases} -\Delta u + b(u) = h(x) & \text{in } \Omega, \\ \partial u/\partial \nu = g & \text{on } \partial\Omega, \end{cases}$$

where b is the maximal monotone graph $b = \partial B$. It is well known that this problem is equivalent to the minimization problem (15.49) via the Euler–Lagrange equation, where $L = h + g$. Then, it is easy to see that condition (15.50) can be expressed by saying that $\frac{1}{\operatorname{meas}(\Omega)}\langle L, 1\rangle$ belongs to the closure of the range of b, while condition (15.51) can be expressed by saying that $\frac{1}{\operatorname{meas}(\Omega)}\langle L, 1\rangle$ belongs to the algebraic interior of the range of b.

Example 15.3.1. We particularize the discussion above to the case of the differential equation

$$\begin{cases} -\Delta u + u^+ = h & \text{in } \Omega, \\ \partial u/\partial \nu = g & \text{on } \partial\Omega, \end{cases} \tag{15.53}$$

where $h \in H^{-1}(\Omega)$ and $g \in H^{-1/2}(\partial\Omega)$. A simple calculation shows that the functional whose Euler–Lagrange equation is (15.53) is

$$F(v) = \int_\Omega \left(\frac{1}{2}|Dv|^2 + |v^+|^2\right) dx - \langle L, v\rangle, \qquad v \in H^1(\Omega), \tag{15.54}$$

where $L = h + g$, and, due to convexity of the functional F, solving (15.53) is equivalent to minimizing (15.54). As a consequence of Proposition 15.3.1 we obtain immediately that

(i) if $\langle L, 1\rangle < 0$, then no solution of (15.53) exists;

(ii) if $\langle L, 1\rangle > 0$, then (15.53) has a solution.

It remains to consider the case

$$\langle L, 1\rangle = 0. \tag{15.55}$$

If h is uniformly continuous and $g = 0$, a proof of existence can be found in [93]. Note that in this case such regularity assumptions imply that the solution w of the associated linearized problem

$$\begin{cases} -\Delta w = h & \text{in } \Omega, \\ \partial w/\partial \nu = g & \text{on } \partial\Omega \end{cases} \tag{15.56}$$

is essentially bounded. More generally, the following proposition holds.

Proposition 15.3.2. *Let $h \in H^{-1}(\Omega)$ and $g \in H^{-1/2}(\partial\Omega)$ satisfy* (15.55)*. Then the following conditions are equivalent:*

(i) *there exists a solution of problem* (15.53)*;*

(ii) *the linearized problem* (15.56) *admits a negative solution.*

Moreover, any solution u of (15.53) *solves* (15.56) *and* $u^+ = 0$.

PROOF. Assume (i) holds. If u solves (15.53) and $u^+ \neq 0$, for every constant $c > 0$ we have

$$F(u - c) < F(u),$$

which contradicts the fact that u is a minimum point of the functional F. Hence $u^+ = 0$ and u solves (15.56), that is, u is a negative solution of (15.56). Assume now (ii). Let w be a negative solution of (15.56); then, w obviously solves (15.53). □

Remark 15.3.2. For instance, by well-known regularity results for solutions of elliptic partial differential equations, condition (ii) of Proposition 15.3.2 is satisfied if $\partial\Omega$ is smooth, $h \in L^p(\Omega)$, and $g \in W^{1-1/p,p}(\partial\Omega)$ with $p > n/2$.

Remark 15.3.3. The argument used in the proof of Proposition 15.3.2 applies also to problems of the form

$$\min\left\{\int_\Omega \left(\frac{1}{2}|Dv|^2 + B(v^+)\right) dx - \langle L, v\rangle : \quad v \in H^1(\Omega)\right\}, \tag{15.57}$$

for any convex and strictly increasing $B : \mathbf{R}^+ \to \mathbf{R}^+$. In this case the associated Euler–Lagrange equation reads as

$$\begin{cases} -\Delta u + \partial B(u^+) \ni h & \text{in } \Omega, \\ \partial u/\partial \nu = g & \text{on } \partial\Omega \end{cases} \tag{15.58}$$

being $L = h + g$ and ∂B the subdifferential of the convex function B, whereas the associated linearized equation coincides with (15.56).

Remark 15.3.4. Consider the obstacle (from below) problem (see also [141])

$$\min\left\{\int_\Omega \frac{1}{2}|Dv|^2\, dx - \langle L, v\rangle : \quad v \in H^1(\Omega),\ v \leq 0\right\}. \tag{15.59}$$

By introducing the function

$$B(s) = \begin{cases} 0 & \text{if } t \leq 0, \\ +\infty & \text{if } t > 0, \end{cases}$$

problem (15.59) can be written in the form (15.58) and the argument above works in the same way. The Euler–Lagrange equation is usually written in this case as a *variational inequality:*

$$u \in H^1(\Omega),\ u \leq 0 \qquad \int_\Omega DuD(v-u)\, dx - \langle L, v - u\rangle \geq 0 \quad \forall v \in H^1(\Omega),\ v \leq 0.$$

Consider now, in a Hilbert space V, a linear continuous functional $L : V \to \mathbf{R}$, a closed convex subset $K \subset V$, and a symmetric continuous bilinear form $a : V \times V \to \mathbf{R}$ which we assume to be nonnegative, that is,

$$a(v, v) \geq 0 \qquad \forall v \in V.$$

If F denotes the functional

$$F(v) = \frac{1}{2}a(v, v) - (L, v) \qquad \forall v \in V,$$

we may consider the minimum problem

$$\min\{F(v) \ : \ v \in K\}. \tag{15.60}$$

Proposition 15.3.3. *The following conditions are equivalent:*

(i) *There exists $u \in K$ such that $F(u) \leq F(v)$ for every $v \in K$.*

(ii) *There exists $u \in K$ such that $a(u, u - v) \leq (L, u - v)$ for every $v \in K$.*

Proof. Assume (i). Then, since K is convex, for every $v \in V$ we have $tv + (1 - t)u \in K$ and the map

$$g_v(t) = F\big(tv + (1 - t)u\big), \qquad t \in [0, 1],$$

achieves its minimum at $t = 0$, hence

$$0 \leq g_v'(0) = -a(u, u) + a(u, v) - (L, v) + (L, u),$$

that is (ii).

Assume now (ii). We have to prove that the map g_v above achieves its minimum at $t = 0$. Since

$$g_v''(t) = a(v, v) + a(u, u) - 2a(u, v) \geq 0,$$

the function $g_v(t)$ is convex on $[0, 1]$. Therefore it is enough to show that $g_v'(0) \geq 0$, which holds true because $g_v'(0) = -a(u, u) + a(u, v) - (L, v) + (L, u)$. □

Remark 15.3.5. Since the map $v \mapsto a(v, v)$ is quadratic, the recession functional associated to F is given by

$$F^\infty(v) = \begin{cases} -(L, v) & \text{if } a(v, v) = 0, \\ +\infty & \text{otherwise;} \end{cases}$$

hence (see Theorem 15.1.2), so that a solution of (15.60) exists, a necessary condition is

$$(L, v) \leq 0 \qquad \forall v \in K^\infty \cap \ker a,$$

whereas, if a satisfies the compactness condition

$$v_h \to 0 \text{ weakly}, \ a(v_h, v_h) \to 0 \quad \to \quad v_h \to 0 \text{ strongly},$$

a sufficient one is

$$K^\infty \cap \ker a \cap \ker L \quad \text{is a subspace.}$$

Problem (15.60) written in the form (ii) above is called a variational inequality. For instance, the obstacle problems are of this type, where the Hilbert space is $H^1(\Omega)$ and the convex set K is

$$K = \left\{ v \in H^1(\Omega) \ : \ v \geq \psi \text{ q.e. in } \overline{\Omega} \right\}$$

where q.e. is intended in the sense of capacity (see Section 5.8). Let us consider more particularly this situation in the case the bilinear form a is given by means of an elliptic operator,

$$a(u, v) = \int_\Omega \sum_{i,j=1}^n a_{ij}(x) D_i u D_j u \, dx,$$

whose coefficient a_{ij} are symmetric, measurable, and satisfy the ellipticity condition

$$c_1|\xi|^2 \leq \sum_{i,j=1}^n a_{ij}(x)\xi_i\xi_j \leq c_2|\xi|^2 \qquad \forall \xi \in \mathbf{R}^n.$$

We denote by A the corresponding elliptic operator

$$Au = -\sum_{i,j=1}^n D_i\big(a_{ij}(x) D_j u\big).$$

Proposition 15.3.4. *The obstacle problem* (15.60)*, equivalent to the variational inequality* (ii) *of Proposition* 15.3.3*, is also equivalent to the complementary problem*

$$\begin{cases} u - \psi \geq 0, \\ Au - L \geq 0, \\ \langle Au - L, u - \psi \rangle = 0. \end{cases} \tag{15.61}$$

Proof. Let u be a solution of the obstacle problem (15.60); by formulation (ii) of Proposition 15.3.3, for every nonnegative function $\phi \in H^1(\Omega)$, taking $v = u + \phi$, we obtain

$$\langle Au, \phi \rangle \geq \langle L, \phi \rangle,$$

that is, $Au - L \geq 0$. On the other hand, taking $v = \psi$ we have

$$0 \leq \langle Au - L, u - \psi \rangle = \langle Au, u - \psi \rangle - \langle L, u - \psi \rangle \leq 0,$$

that is, $\langle Au - L, u - \psi \rangle = 0$.

On the other hand, if u verifies (15.61), writing a generic $v \geq \psi$ in the form $\psi + \phi$ with $\phi \geq 0$, we have

$$\langle Au - L, u - v \rangle = \langle Au - L, u - \psi \rangle - \langle Au - L, \phi \rangle leq 0,$$

so that by Proposition 15.3.3, u solves the obstacle problem. □

Remark 15.3.6. Since $Au - L$ is a nonnegative distribution, by the Riesz–Schwartz theorem there exists a nonnegative Borel measure μ on Ω such that $Au - L = \mu$. Moreover, by (15.61) the measure μ is concentrated on the *coincidence set* $\{u = \psi\}$. For further details on the obstacle problems see, for instance, the book by Kinderlehrer and Stampacchia [163]. Here we want only to remark that the obstacle problem is a mathematical model for determining the shape of a thin elastic membrane subject to a vertical load L and where the unknown u represents the vertical displacement of the membrane. The measure μ in this framework has a natural interpretation as the upward force due to the constraint reaction of the rigid obstacle.

As a further example we consider now the case of minimum problems of the form

$$\min\left\{\int_0^T \left[\frac{1}{2}|u'|^2 + V(t,u)\right] dt \ : \ u \in H^1(0,T;\mathbf{R}^n),\ u(0) = u(T)\right\}, \tag{15.62}$$

where V is a Borel function, with $V(t,\cdot)$ lower semicontinuous on $\mathbf{R}^n$, and such that

$$V(t,s) \geq -a(t) - b(t)|s|^q \qquad \text{for a.e. } t \in (0,T) \text{ and for every } s \in \mathbf{R}^n \tag{15.63}$$

for suitable $q < 2$ and $a(t)$, $b(t)$ in $L^1(0,T)$. If $V(t,\cdot)$ is smooth the minimum problem above has the Euler–Lagrange equation

$$\begin{cases} -u'' + \nabla V(t,u) = 0, \\ u(0) = u(T),\ u'(0) = u'(T). \end{cases} \tag{15.64}$$

We are therefore looking for solutions of the differential equations $-u'' + \nabla V(t,u) = 0$ which are periodic on the interval $[0,T]$. It is convenient to denote by H^1_T the space $\left\{u \in H^1(0,T;\mathbf{R}^n) \ : \ u(0) = u(T)\right\}$ and by F the functional

$$F(u) = \int_0^T \left[\frac{1}{2}|u'|^2 + V(t,u)\right] dt.$$

Notice that in general the functional F is not convex. Nevertheless the sequential weak lower semicontinuity of F is straightforward. Let us prove now the compactness property (i) of Theorem 15.2.1. If $t_h \to +\infty$ and $u_h \to u$ weakly in H^1_T, with $F(t_h u_h)$ bounded from above, then we deduce by (15.63) that $\int_0^T |u_h'|^2\,dt \to 0$, and this implies immediately the strong convergence of the sequence $\{u_h\}$ to a constant.

To prove the necessary condition $F_\infty(u) \geq 0$ for every $u \in H^1_T$, we notice that it is trivially fulfilled when u is a nonconstant function, because in this case we have $F_\infty(u) = +\infty$. Indeed, if $t_h \to +\infty$ and $u_h \to u$ weakly in H^1_T with u nonconstant, we have

$$\liminf_{h\to+\infty} \int_0^T |u_h'|^2\,dt \geq \int_0^T |u'|^2\,dt > 0.$$

Therefore, by (15.63) and using the fact that $q < 2$, we obtain

$$\liminf_{h\to+\infty} \frac{F(t_h u_h)}{t_h} \geq \liminf_{h\to+\infty} \int_0^T \left[\frac{t_h}{2}|u_h'|^2 - b(t) t_h^{q-1}|u_h|^q\right] dt = +\infty.$$

The necessary condition $F_\infty(u) \geq 0$ is then reduced to

$$F_\infty(c) \geq 0 \qquad \text{for every } c \in \mathbf{R}^n. \tag{15.65}$$

Here below we assume (15.65) is satisfied and we will particularize some special cases when additional assumptions on the potential V are made.

When $V(t, \cdot)$ is convex on $\mathbf{R}^n$, then the functional F turns out to be convex, and we are in the framework of Section 13.1. Assume there exists a function $u_0 \in H^1_T$ such that $V\big(t, u_0(t)\big)$ is integrable and introduce for every $c \in \mathbf{R}^n$ the convex function $\Phi_c : \mathbf{R} \to]-\infty, +\infty]$ given by

$$\Phi_c(r) = \int_0^T V\big(t, u_0(t) + cr\big)\, dt.$$

Then by applying Theorem 15.1.1 it is easy to see that the existence of at least a solution to problem (15.62) occurs provided for every $c \in \mathbf{R}^n$ the function Φ_c is either constant or such that

$$\lim_{|r| \to +\infty} \Phi_c(r) = +\infty. \tag{15.66}$$

For instance, if

$$V(t, s) = V(s) - f(t)s$$

we have that condition (15.66) is fulfilled for every $f \in L^1(0, T; \mathbf{R}^n)$ whenever the function $V(s)$ has a superlinear growth, that is,

$$\lim_{|s| \to +\infty} \frac{V(s)}{|s|} = +\infty.$$

Another case in which the existence of minimizers for problem (15.62) can be easily obtained is when the potential V satisfies a Lipschitz condition of the form

$$\big|V(t, s_1) - V(t, s_2)\big| \leq k(t)|s_1 - s_2| \qquad \text{for a.e. } t \in (0, T) \text{ and for every } s_1.s_2 \in \mathbf{R}^n. \tag{15.67}$$

Here we assume that $k \in L^1(0, T)$, and that the potential V satisfies the coercivity condition

$$\lim_{|s| \to +\infty} \int_0^T V(t, s)\, dt = +\infty. \tag{15.68}$$

In this case problem (15.62) is actually coercive, and the existence result then follows immediately from the direct methods of the calculus of variations of Section 3.2. Indeed, by using the Lipschitz condition (15.67) we obtain for every $u \in H^1_T$

$$F(u) \geq \int_0^T \left[\frac{1}{2}|u'|^2 + V\big(t, u(0)\big)\right] dt - \int_0^T k(t)\big|u(t) - u(0)\big|\, dt.$$

Moreover, since

$$\big|u(t) - u(0)\big| = \left|\int_0^t u'(s)\, ds\right| \leq \left(T \int_0^T |u'|^2\, dt\right)^{1/2},$$

we deduce that

$$F(u) \geq \int_0^T \left[\frac{1}{2}|u'|^2 + V\big(t, u(0)\big)\right] dt - C\bigg(\int_0^T |u'|^2\, dt\bigg)^{1/2}$$

for a suitable positive constant C. Now, by using the coercivity assumption (15.68), and the fact that on H^1_T the norm is equivalent to

$$\bigg(\int_0^T |u'|^2\, dt + |u(0)|^2\bigg)^{1/2},$$

we obtain that $\|u\|_{H^1_T} \to +\infty$ yields $F(u) \to +\infty$, hence the coercivity of F. The lower semicontinuity of F is a straightforward consequence of the results of Section 13.1.

Consider now the case when the function $V(t,\cdot)$ is periodic. More precisely, we assume that the function V is nonnegative, or more generally bounded from below by an L^1 function, and that for suitable independent vectors $\tau_i \in \mathbf{R}^n$ we have for a.e. $t \in (0,T)$ and for every $s \in \mathbf{R}^n$

$$V(t, s+\tau_i) = V(t,s), \qquad i = 1, \dots, n.$$

Note that in this case the functional F is not convex in general and that, by the nonnegativity of V, the necessary condition (15.65) is clearly fulfilled. Thus, to apply Theorem 15.2.1 and obtain the existence of a minimizer for F, it is enough to show that the compatibility condition (iii′) of Remark 15.2.5 holds. In other words we have to find for every $c \in \mathbf{R}^n$ a vector $\mu \in \mathbf{R}^n$ whose components μ_i are positive, such that

$$F(u_1 - \mu_1 c_1, \dots, u_n - \mu_n c_n) \leq F(u_1, \dots, u_n) \qquad \forall u \in H^1_T.$$

This can be achieved if we choose for instance $\mu_i = |\tau_i|/|c_i|$ for all $i = 1, \dots, n$, with $\mu_i = 1$ if $c_i = 0$.

15.4 Limit analysis problems

We consider the so-called limit analysis problems which consist in minimizing functionals of the form

$$F(x) - \gamma L(x), \quad x \in X,$$

where X is a normed space, $F : X \to]-\infty, +\infty]$ is a (possibly nonconvex) functional, and $L \in X'$. We are interested in characterizing the values of γ for which the minimum is attained. As an application we consider nonconvex minimum problems defined on the space of measures and on the BV space.

Let X be a normed space; we consider on X the norm topology τ and another (linear Hausdorff) topology σ weaker than τ and such that the unit ball $\{x \in X : \|x\| \leq 1\}$ is σ-compact. Let $F : X \to]-\infty, +\infty]$ be a functional proper and σ-lsc (on all τ-bounded sets), and let $L : X \to \mathbf{R}$ be a linear σ-continuous functional. The limit analysis problem associated to F and L consists in finding the values $\gamma \in \mathbf{R}$ for which the minimum problem

$$\min\{F(x) - \gamma L(x) : x \in X\} \tag{15.69}$$

admits at least a solution. When F is convex, the problem was studied by Bouchitté and Suquet in [78], where the following result is proved.

Theorem 15.4.1. *Assume that F is proper, convex, σ-lsc, and σ-coercive in the sense that*

$$\lim_{\|x\|\to+\infty} F(x) = +\infty.$$

Then, setting

$$\gamma^* = \min\{F^\infty(x) \ : \ x \in X, \ L(x) = 1\},$$
$$\gamma_* = -\min\{F^\infty(x) \ : \ x \in X, \ L(x) = -1\},$$

the following statements are equivalent:

(i) *$F - \gamma L$ is σ-coercive.*

(ii) *$\gamma_* < \gamma < \gamma^*$.*

Proof. Since F is proper, there exists $x_0 \in X$ such that $F(x_0) < +\infty$; then, by considering the functional $F(x + x_0) - F(x_0)$ it is easy to see that we may reduce ourselves to assume without any loss of generality that $F(0) = 0$.

Assume now (i). By the properties of recession functions we have

$$(F - \gamma L)^\infty(x) \geq \frac{F(tx) - \gamma L(tx)}{t} \qquad \forall t > 0, \ \forall x \in X;$$

if for some $x \neq 0$ we had $(F - \gamma L)^\infty(x) \leq 0$, this would be in contradiction to the coerciveness of $F - \gamma L$. Then, taking as x^* the solution of

$$\gamma^* = \min\left\{F^\infty(x) \ : \ L(x) = 1\right\}$$

we obtain

$$0 < (F - \gamma L)^\infty(x^*) = F^\infty(x^*) - \gamma L(x^*) = \gamma^* - \gamma.$$

Analogously, if x_* is the solution of

$$-\gamma_* = \min\left\{F^\infty(x) \ : \ L(x) = -1\right\},$$

we get

$$0 < (F - \gamma L)^\infty(x_*) = F^\infty(x_*) - \gamma L(x_*) = -\gamma^* + \gamma,$$

that is, (ii).

Assume now (ii) and let $0 < \gamma < \gamma^*$. By contradiction, if $F - \gamma L$ were not coercive, we could find a sequence (x_h) such that $\|x_h\| \to +\infty$ and

$$(F - \gamma L)(x_h) \leq M \tag{15.70}$$

for a suitable $M \in \mathbf{R}$. Note that $L(x_h) \to +\infty$ because otherwise (15.70) and the coerciveness of F would prove that (x_h) is bounded. Setting $y_h = x_h/\|x_h\|$ and $t_h = \|x_h\|$ we

get that (y_h) is σ-relatively compact, hence (up to subsequences) converging to a suitable $y \in X$. By using the lower semicontinuity and convexity of F, and the fact that $F(0) = 0$, we have for every $t > 0$

$$\begin{aligned}\frac{F(ty)}{t} &\le \liminf_{h\to+\infty} \frac{F(ty_h)}{t} \le \liminf_{h\to+\infty} \frac{F(t_h y_h)}{t_h} \\ &\le \liminf_{h\to+\infty} \gamma \frac{L(x_h)}{\|x_h\|} = \gamma L(y)\end{aligned}$$

so that, as $t \to +\infty$,

$$F^\infty(y) \le \gamma L(y). \tag{15.71}$$

Being F lower semicontinuous and σ-coercive, by Proposition 15.1.3 we have $F^\infty \ge 0$, hence $L(y) \ge 0$ by (15.71). The case $L(y) > 0$ must be excluded because, taking $z = y/L(y)$ we get $L(z) = 1$ and

$$\gamma^* \le F^\infty(z) = \frac{F^\infty(y)}{L(y)} \le \gamma,$$

which contradicts (ii). Therefore $L(y) = 0$ and, by (15.71), $F^\infty(y) = 0$. Arguing as in the first part of the proof we get that by the coerciveness of F, this implies $y = 0$, so that using (15.70) and recalling that $L(x_h) \to +\infty$,

$$F\left(\frac{x_h}{L(x_h)}\right) \le \frac{F(x_h)}{L(x_h)} \le \frac{M}{L(x_h)} + \gamma.$$

Hence $x_h/L(x_h)$ is bounded in X (by the coerciveness of F), and so $y_h/L(y_h)$ is bounded too. But this contradicts the fact that $\|y_h\| = 1$ whereas $L(y_h) \to 0$. An analogous argument can be used in the case $\gamma_* < \gamma < 0$. □

When F is not necessarily convex we have the following result.

Theorem 15.4.2. *Assume that*

(i) *F is proper and sequentially σ-lsc;*

(ii) *for all $z \in \ker F_\infty$ there exists $\eta = \eta(z) > 0$ such that*

$$F(x - \eta z) \le F(x) \qquad \forall x \in X;$$

(iii) *there exist a seminorm $P : X \to [0, +\infty[$ satisfying the compactness condition* (2.3.11) *and a number $C \ge 0$ such that*

$$F(x) \ge P(x) - C \qquad \forall x \in X.$$

Then, setting

$$\begin{aligned}\gamma^* &= \inf\{F_\infty(x) \ : \ L(x) = 1\}, \\ \gamma_* &= -\inf\{F_\infty(x) \ : \ L(x) = -1\},\end{aligned}$$

we have

(a) *if the minimum problem* (15.69) *admits a solution, then* $\gamma_* \le \gamma \le \gamma^*$*;*

(b) *if* $\gamma_* < \gamma < \gamma^*$*, then the minimum problem* (15.69) *admits a solution.*

PROOF. To prove statement (a) we apply Proposition 15.2.5. The necessary condition for the existence of a solution to problem (15.69) is

$$(F - \gamma L)_\infty \ge 0 \qquad \text{on } X,$$

which, thanks to Proposition 15.2.1, becomes

$$F_\infty(x) \ge \gamma L(x) \qquad \forall x \in X. \tag{15.72}$$

Taking $L(x) = -1$ in (15.72) gives

$$\gamma \ge \sup\big\{ - F_\infty(x) \ : \ L(x) = -1\big\} = \gamma_*;$$

similarly, taking $L(x) = 1$ in (15.72) gives

$$\gamma \le \inf\big\{F_\infty(x) \ : \ L(x) = 1\big\} = \gamma^*.$$

To prove statement (b) we are going to verify all the hypotheses of Theorem 15.2.1 for the functional $G = F - \gamma L$ which is clearly sequentially σ-lsc. As seen in the proof of statement (a), the necessary condition (ii) of Theorem 15.2.1 is equivalent to inequalities

$$\gamma_* \le \gamma \le \gamma^*.$$

To verify the compactness hypothesis (i) of Theorem 15.2.1, let $t_h \to +\infty$ and let $x_h \to x$ be a sequence σ-converging in X such that

$$F(t_h x_h) - \gamma L(t_h x_h) \le C. \tag{15.73}$$

Dividing by t_h we get

$$\frac{F(t_h x_h)}{t_h} - \gamma L(x_h) \le \frac{C}{t_h}$$

so that, by definition of the topological recession functional,

$$F_\infty(x) \le \gamma L(x).$$

If $L(x) > 0$ we would obtain

$$F_\infty\left(\frac{x}{L(x)}\right) \le \gamma < \gamma^* = \inf\big\{F_\infty(y) \ : \ L(y) = 1\big\},$$

which gives a contradiction. In an analogous way we can exclude the case $L(x) < 0$. Therefore we have

$$L(x) = 0$$

and so also $F_\infty(x) = 0$. Now, if $L(t_h x_h)$ is bounded from above, by (15.73) we would get

$$P(t_h x_h) \le C,$$

and so, since P satisfies the compactness assumption (ii) of Theorem 15.2.1, we would obtain $x_h \to x$ strongly in X. Otherwise, if $L(t_h x_h)$ (or a subsequence of it) tends to $+\infty$, dividing by $L(t_h x_h)$ in (15.73) we get

$$P\left(\frac{x_h}{L(x_h)}\right) \le C.$$

Since $L(x_h) \to L(x) = 0$, by the compactness assumption of P we obtain again $x_h \to x$. Therefore the compactness hypothesis (2.3.11) is satisfied.

Finally we verify the compatibility conditions (iii) of Theorem 15.2.1. Let $z \in \ker(F_\infty - \gamma L)$, i.e.,

$$F_\infty(z) = \gamma L(z).$$

As before, using the strict inequalities $\gamma_* < \gamma < \gamma^*$ we obtain $F_\infty(z) = L(z) = 0$, and so, by assumption (ii)

$$F(x - \eta z) - \gamma L(x - \eta z) = F(x - \eta z) - \gamma L(x) \le F(x) - \gamma L(x)$$

for all $x \in X$, that is, the compatibility condition (iii) is satisfied. □

Remark 15.4.1. If the infimum

$$\inf\left\{F_\infty(x) \ : \ L(x) = 1\right\} \qquad \left(\text{respectively, } \inf\left\{F_\infty(x) \ : \ L(x) = -1\right\}\right)$$

is not attained, then we can also accept in Theorem 15.4.2 (b) $\gamma = \gamma^*$ (respectively, $\gamma = \gamma_*$).

As an application of the limit analysis theorems above we consider the case of functionals defined on measures. We refer to Section 13.3 for the theory of convex functionals on measures and to Bouchitté and Buttazzo [74], [75], [76] for further details on nonconvex functionals defined on measures.

Consider a measure space $(\Omega, \mathcal{B}, \mu)$, where Ω is a separable locally compact metric space, $\mathcal{B}$ is the σ-algebra of all Borel subsets of Ω, and $\mu : \mathcal{B} \to [0, +\infty[$ is a positive, finite, nonatomic measure. Consider a functional defined on $\mathbf{M}(\Omega; \mathbf{R}^n)$ of the form first considered by Bouchitté and Buttazzo [74],

$$F(\lambda) = \int_\Omega f\left(\frac{d\lambda}{d\mu}\right) d\mu + \int_{\Omega \setminus A_\lambda} f^\infty(\lambda^s) + \int_{A_\lambda} g(\lambda(x))\, d\#. \tag{15.74}$$

Here

$f : \mathbf{R}^n \to [0, +\infty]$ is a proper, convex, lsc function with $f(0) = 0$;

f^∞ is its recession function;

$g : \mathbf{R}^n \to [0, +\infty]$ is an lsc function with $g(0) = 0$ satisfying the subadditivity condition

$$g(s_1 + s_2) \le g(s_1) + g(s_2) \qquad \forall s_1, s_2 \in \mathbf{R}^n;$$

$\lambda = \frac{d\lambda}{d\mu}\mu + \lambda^s$ is the Lebesgue–Nikodym decomposition of λ into absolutely continuous and singular parts with respect to μ;

A_λ is the set of all atoms of λ;

$\lambda(x)$ is the value $\lambda(\{x\})$;

is the counting measure.

As already recalled, Bouchitté and Buttazzo [74] proved that if the condition

$$f^\infty(s) = \lim_{t\to 0^+} \frac{g(ts)}{t} \qquad \forall s \in \mathbf{R}^n$$

is fulfilled, then the functional (15.74) is sequentially weakly*-lsc on $\mathbf{M}(\Omega; \mathbf{R}^n)$.

For our purposes, it is convenient to introduce for every function $g : \mathbf{R}^n \to [0, +\infty]$ the functions

$$g^\infty(s) = \liminf_{t\to+\infty} \frac{g(ts)}{t},$$
$$g^0(s) = \limsup_{t\to 0^+} \frac{g(ts)}{t}.$$

The following proposition holds (see [74]).

Proposition 15.4.1. *Let $g : \mathbf{R}^n \to [0, +\infty[$ be an lsc and subadditive function, with $g(0) = 0$. Then, we have the following:*

(i) *the functions g^0 and g^∞ are convex, lsc, and positively 1-homogeneous;*

(ii) $g^0(s) = \sup_{t>0} \frac{g(ts)}{t} = \lim_{t\to 0^+} \frac{g(ts)}{t}$ *for every $s \in \mathbf{R}^n$;*

(iii) $g^\infty(s) = \inf_{t>0} \frac{g(ts)}{t} = \lim_{t\to+\infty} \frac{g(ts)}{t}$ *for every $s \in \mathbf{R}^n$.*

Remark 15.4.2. From Proposition 15.72 it follows easily that

$$g^\infty(s) \le g(s) \le g^0(s) \qquad \text{for every } s \in \mathbf{R}^n.$$

The following theorem gives sufficient conditions on f and g to apply Theorem 15.4.2 to functionals of the form (15.74).

Theorem 15.4.3. *Let $f : \mathbf{R}^n \to [0, +\infty]$, and $g : \mathbf{R}^n \to [0, +\infty[$ be given functions. Assume that*

(i) *f is convex, lsc, and proper on $\mathbf{R}^n$, and $f(0) = 0$;*

(ii) *there exist $C_1 > 0$ and $D \in \mathbf{R}$ such that*

$$f(x) \ge C_1|x| - D \qquad \forall x \in \mathbf{R}^n;$$

(iii) *g is lsc and subadditive on $\mathbf{R}^n$, and $g(0) = 0$;*

(iv) *there exists* $C_2 > 0$ *such that*

$$g(x) \geq C_2|x| \qquad \forall x \in \mathbf{R}^n;$$

(v) $g^0 = f^\infty$ *in* $\mathbf{R}^n$*;*

(vi) $H \in \mathbf{C}_0(\Omega; \mathbf{R}^n)$.

Let $F : \mathbf{M}(\Omega; \mathbf{R}^n) \to [0, +\infty]$ *be the functional defined in* (15.74)*. Then, setting*

$$\gamma^* = \inf\{F_\infty(\lambda) \ : \ \langle H, \lambda\rangle = 1\},$$
$$\gamma_* = -\inf\{F_\infty(\lambda) \ : \ \langle H, \lambda\rangle = -1\},$$

we have

(a) *if the functional* $F - \gamma\langle H, \cdot\rangle$ *admits a minimum on* $\mathbf{M}(\Omega; \mathbf{R}^n)$*, then* $\gamma_* \leq \gamma \leq \gamma^*$*;*

(b) *the functional* $F - \gamma\langle H, \cdot\rangle$ *admits a minimum on* $\mathbf{M}(\Omega; \mathbf{R}^n)$*, for every* γ *such that* $\gamma_* < \gamma < \gamma^*$.

Proof. By the assumptions made on f and g the functional F is sequentially weakly*-lsc on $\mathbf{M}(\Omega; \mathbf{R}^n)$. Moreover, by assumptions (ii) and (iv), we have

$$F(\lambda) \geq C_1 \int_\Omega \left|\frac{d\lambda}{d\mu}\right| d\mu + C_1 \int_{\Omega\setminus A_\lambda} |\lambda^s| + C_2 \int_{A_\lambda} |\lambda^s(x)|\, d\# - D\mu(\Omega)$$

so that

$$F(\lambda) \geq C\|\lambda\| - b \tag{15.75}$$

for suitable $C > 0$ and $b \in \mathbf{R}$. Finally, from (15.75) we get

$$\ker F_\infty = \{0\}$$

so that hypothesis (ii) of Theorem 15.4.2 is satisfied too, and hence the conclusions follow from Theorem 15.4.2. □

We give now an explicit formula for the bounds γ^* and γ_*. To obtain this result we need first an explicit representation for the topological recession function F_∞. We use a representation theorem for the relaxed functional associated to integrals of the form (15.74). More precisely, given a functional $F : \mathbf{M}(\Omega; \mathbf{R}^n) \to [0, +\infty]$ of the form

$$F(\lambda) = \begin{cases} \displaystyle\int_\Omega f\Big(\frac{d\lambda}{d\mu}\Big)\, d\mu + \int_{A_\lambda} g\big(\lambda(x)\big)\dot{\#} & \text{if } \lambda^s = 0 \text{ on } \Omega \setminus A_\lambda, \\ +\infty & \text{otherwise,} \end{cases}$$

we consider its relaxed functional $\overline{F}$ defined by

$$\overline{F} = \sup\big\{G \ : \ G \leq F, \ G \text{ sequentially weakly* lsc on } \mathbf{M}(\Omega; \mathbf{R}^n)\big\}.$$

Bouchitté and Buttazzo [75] proved that if $f, g : \mathbf{R}^n \to [0, +\infty]$ satisfy the assumptions

- f is convex and lsc on $\mathbf{R}^n$, and $f(0) = 0$,
- there exist $\alpha > 0$ and $\beta \geq 0$ such that

$$f(s) \geq \alpha|s| - \beta \qquad \forall s \in \mathbf{R}^n,$$

- g is subadditive and lsc on $\mathbf{R}^n$, and $g(0) = 0$,
- $g^0(s) \geq \alpha|s|$ for every $s \in \mathbf{R}^n$,

then the following integral representation holds for $\overline{F}$:

$$\overline{F}(\lambda) = \int_\Omega \overline{f}\left(\frac{d\lambda}{d\mu}\right) d\mu + \int_{\Omega\setminus A_\lambda} (\overline{f})^\infty(\lambda^s) + \int_{A_\lambda} \overline{g}(\lambda(x))\, d\#, \tag{15.76}$$

where

$$\overline{f} = f\#_e g^0, \qquad \overline{g} = f^\infty \#_e g.$$

To characterize the topological recession function for functionals of the form (15.74) we introduce the functional

$$G^\infty(\lambda) = \int_\Omega g^\infty(\lambda) = \int_\Omega g^\infty\left(\frac{d\lambda}{d\mu}\right) d\mu + \int_\Omega g^\infty(\lambda^s).$$

Theorem 15.4.4. *Under the assumptions of Theorem* 15.4.3, *we have*

$$F_\infty(\lambda) = G^\infty(\lambda) \qquad \forall \lambda \in \mathbf{M}(\Omega; \mathbf{R}^n).$$

Proof. We prove first that $F_\infty(\lambda) \leq G^\infty(\lambda)$ for every $\lambda \in \mathbf{M}(\Omega; \mathbf{R}^n)$. Let $\lambda \in \mathbf{M}(\Omega; \mathbf{R}^n)$ be a measure with a finite number of atoms. Then,

$$\begin{aligned} F_\infty(\lambda) &\leq \liminf_{t\to+\infty} \left[\int_\Omega \frac{1}{t} f\left(t\frac{d\lambda}{d\mu}\right) d\mu + \int_{\Omega\setminus A_\lambda} f^\infty(\lambda^s) + \int_{A_\lambda} \frac{g(t\lambda(x))}{t}\, d\# \right] \\ &\leq \int_\Omega f^\infty\left(\frac{d\lambda}{d\mu}\right) d\mu + \int_{\Omega\setminus A_\lambda} f^\infty(\lambda^s) + \int_{A_\lambda} g^\infty\big(\lambda(x)\big)\, d\#. \end{aligned} \tag{15.77}$$

Now, let λ be any measure in $\mathbf{M}(\Omega; \mathbf{R}^n)$. Setting $A_\lambda^h = \big\{x \in A_\lambda \,:\, |\lambda|(x) < 1/h\big\}$ and $\lambda_h = \lambda \cdot 1_{\Omega\setminus A_\lambda^h}$, we get $\lambda_h \to \lambda$; moreover, since $A_{\lambda_h} = \big\{x \in A_\lambda \,:\, |\lambda|(x) \geq 1/h\big\}$, we have that λ_h has a finite number of atoms. From the weak*-lower semicontinuity of F_∞, taking into account (15.77), we have

$$\begin{aligned} &F_\infty(\lambda) \leq \liminf_{h\to+\infty} F_\infty(\lambda_h) \leq \liminf_{h\to+\infty} \left[\int_\Omega f^\infty\Big(\frac{d\lambda}{d\mu}\Big)\, d\mu + \int_{\Omega\setminus A_\lambda} f^\infty(\lambda^s) + \int_{A_\lambda\setminus A_\lambda^h} g^\infty\big(\lambda^s(x)\big)\, d\# \right] \\ &\leq \int_\Omega f^\infty\Big(\frac{d\lambda}{d\mu}\Big)\, d\mu + \int_{\Omega\setminus A_\lambda} \chi_{\{0\}}(\lambda^s) + \int_{A_\lambda} g^\infty\big(\lambda^s(x)\big)\, d\#. \end{aligned} \tag{15.78}$$

By computing the relaxation of the first and the last terms of (15.78), we get by (15.76)

$$F_\infty(\lambda) \leq \int_\Omega (f^\infty \#_e g^\infty)\left(\frac{d\lambda}{d\mu}\right) d\mu + \int_{\Omega\setminus A_\lambda} (f^\infty \#_e g^\infty)(\lambda^s) + \int_{A_\lambda} (f^\infty \#_e g^\infty)\big(\lambda^s(x)\big)\, d\#$$

for every $\lambda \in \mathbf{M}(\Omega;\mathbf{R}^n)$. From (v) of Theorem 15.4.3 we deduce that $f^\infty \#_e g^\infty = g^\infty$ so that

$$F_\infty(\lambda) \le G^\infty(\lambda) \qquad \forall \lambda \in \mathbf{M}(\Omega;\mathbf{R}^n).$$

We prove now the opposite inequality. We claim that for every $\varepsilon > 0$ there exists a $k_\varepsilon > 0$ such that

$$g^\infty(s) \le f(s) + \varepsilon|s| + k_\varepsilon \qquad \forall s \in \mathbf{R}^n. \tag{15.79}$$

By contradiction, assume there exists an $\varepsilon_0 > 0$ such that for every $k \in \mathbf{N}$ there exists a $s_k \in \mathbf{R}^n$ with

$$g^\infty(s_k) > f(s_k) + \varepsilon_0|s_k| + k.$$

Setting $v_k = s_k/|s_k|$, and $t_k = |s_k|$ we have

$$g^\infty(v_k) > \frac{f(t_k v_k)}{t_k} + \varepsilon_0 + \frac{k}{t_k}. \tag{15.80}$$

Since $|v_k| = 1$, it is not restrictive to assume $v_k \to v$ for some $v \in \mathbf{R}^n$. If (t_k) is bounded we get, by using Proposition 15.4.1(i),

$$g(v) \ge g^\infty(v) = +\infty,$$

which is impossible since g is finite. Therefore, we can assume $t_k \to +\infty$. Passing to the limit in (15.80) yields, taking into account Proposition 15.4.1(i) again,

$$g^\infty(v) \ge f^\infty(v) + \varepsilon_0 = g^0(v) + \varepsilon_0 > g^\infty(v),$$

which is a contradiction. Then (15.79) holds, and from the weak*-lower semicontinuity and 1-homogeneity of G^∞ we get

$$\begin{aligned} G^\infty(\lambda) &= \inf\left\{\liminf_{h\to+\infty} \frac{G^\infty(t_h\lambda_h)}{t_h}\right\} \\ &\le \inf\left\{\liminf_{h\to+\infty} \frac{1}{t_h}\left[\int_\Omega \left(f\left(t_h\frac{d\lambda_h}{d\mu}\right) + \varepsilon\left|t_h\frac{d\lambda_h}{d\mu}\right|\right) d\mu\right.\right. \\ &\qquad \left.\left. + \int_{\Omega\setminus A_{\lambda_h}} f^\infty(t_h\lambda_h^s) + \int_{A_{\lambda_h}} g\big(t_h\lambda_h^s(x)\big)\, d\#\right]\right\} \\ &\le \inf\left\{\liminf_{h\to+\infty}\left[\frac{F(t_h\lambda_h)}{t_h} + \varepsilon\|\lambda_h\|\right]\right\}, \end{aligned}$$

where the infimum is taken over all $t_h \to +\infty$ and all $\lambda_h \to \lambda$. Taking (t_h) and (λ_h) such that

$$F_\infty(\lambda) = \liminf_{h\to+\infty} \frac{F(t_h\lambda_h)}{t_h},$$

we have

$$G^\infty(\lambda) \le F_\infty(\lambda) + \varepsilon \limsup_{h\to+\infty} \|\lambda_h\| \le F_\infty(\lambda) + \varepsilon C.$$

Letting $\varepsilon \to 0$, we get

$$G^\infty(\lambda) \le F_\infty(\lambda),$$

and the proof is achieved. □

By virtue of Theorem 15.4.4 we can write

$$\gamma^* = \inf\{G^\infty(\lambda) \ : \ \langle H, \lambda\rangle = 1\},$$
$$\gamma_* = -\inf\{G^\infty(\lambda) \ : \ \langle H, \lambda\rangle = -1\}$$

or equivalently

$$\gamma^* = \frac{1}{\sup\{\langle H, \lambda\rangle \ : \ G^\infty(\lambda) = 1\}},$$

$$\gamma_* = \frac{1}{\inf\{\langle H, \lambda\rangle \ : \ G^\infty(\lambda) = 1\}}.$$

The last expressions allow us to compute explicitly γ^* and γ_* in terms of g^∞ and H only. Indeed, by using the definition of the Fenchel transform, it is easy to see that

$$\frac{1}{\sup\{\langle H, \lambda\rangle \ : \ G^\infty(\lambda) = 1\}} = \sup\left\{t \ : \ (G^\infty)^*(tH) = 0\right\}.$$

Therefore, since

$$(G^\infty)^*(w) = \begin{cases} 0 & \text{if } (g^\infty)^*(w) \equiv 0, \\ +\infty & \text{otherwise,} \end{cases}$$

we obtain

$$\gamma^* = \sup\left\{t \ : \ (g^\infty)^*(tH) \equiv 0\right\} = \left[\sup_{x,s} \frac{H(x)s}{g^\infty(s)}\right]^{-1}.$$

Analogously, it is

$$\gamma_* = \inf\left\{t \ : \ (g^\infty)^*(tH) \equiv 0\right\} = \left[\inf_{x,s} \frac{H(x)s}{g^\infty(s)}\right]^{-1}.$$

For instance, if $g^\infty(s) = c|s|$, we get

$$\gamma^* = \frac{c}{\|H\|_{C_0(\Omega;\mathbf{R}^n)}}, \qquad \gamma_* = -\frac{c}{\|H\|_{C_0(\Omega;\mathbf{R}^n)}}.$$

The result we obtained allows us to study the limit analysis problem for a class of nonconvex functionals defined on BV. More precisely, let $\Omega =]a, b[$ be an open interval of $\mathbf{R}$, and assume that f and g satisfy all the hypotheses of Theorem 15.4.3 Consider the nonconvex functional $F : BV(\Omega; \mathbf{R}^n) \to [0, +\infty]$ defined by

$$F(u) = \int_\Omega f(\nabla u)\, dx + \int_{\Omega\setminus S_u} f^\infty(D^s u) + \int_{S_u} g(D^s u(x))\, d\#(x), \tag{15.81}$$

where ∇u and $D^s u$, respectively, denote the absolutely continuous and the singular parts of Du with respect to the Lebesgue measure, and S_u is the set of jumps of u, that is, the set of all points $x \in \Omega$ such that the left and right traces $u^+(x)$ and $u^-(x)$ do not coincide. Setting $\lambda = Du$, the functionals of type (15.81) can be interpreted in terms of functionals of type (15.74) on $\mathbf{M}(\Omega; \mathbf{R}^n)$.

The Neumann problem. We deal with functionals G defined on $BV(\Omega;\mathbf{R}^n)$ by

$$G(u)=\int_\Omega f(\nabla u)\,dx+\int_{\Omega\setminus S_u} f^\infty(D^s u)+\int_{S_u} g(D^s u(x))\,d\#(x)-\gamma\langle L,u\rangle, \qquad (15.82)$$

where

$$\langle L,u\rangle=\int_\Omega hu\,dx+\int_\Omega \phi Du$$

with $h\in L^1(\Omega;\mathbf{R}^n)$ and $\phi\in C_0(\Omega;\mathbf{R}^n)$. It is easily verified that $\langle L,1\rangle=0$ is a necessary condition in order to get a minimum for the functional (15.82). Therefore, setting

$$H(x)=\int_a^x h(y)\,dy,$$

we have that $H\in C_0(\Omega;\mathbf{R}^n)$, and integrating by parts,

$$\langle L,u\rangle=\langle \phi-H, Du\rangle, \qquad \text{with } \phi-H\in C_0(\Omega;\mathbf{R}^n).$$

Hence, by the limit analysis result above, a necessary condition for the existence of a minimizer of the functional G in (15.82) is $\gamma_*\le\gamma\le\gamma^*$, whereas a sufficient condition is $\gamma_*<\gamma<\gamma^*$, with

$$\gamma_*=\left[\inf_{x,s}\frac{(\phi(x)-H(x))\,s}{g^\infty(s)}\right]^{-1},$$
$$\gamma^*=\left[\sup_{x,s}\frac{(\phi(x)-H(x))\,s}{g^\infty(s)}\right]^{-1}.$$

The Dirichlet problem. To deal with the Dirichlet problem associated to functionals of the form (15.82), it is convenient to consider an open interval Ω_0 containing Ω and the space

$$BV_0=\big\{u\in BV(\Omega_0;\mathbf{R}^n)\ :\ u=0 \text{ on } \Omega_0\setminus\Omega\big\}.$$

Therefore, given $h\in L^1(\Omega;\mathbf{R}^n)$ and $\phi\in\mathbf{C}(\overline{\Omega};\mathbf{R}^n)$, and denoting by $\tilde h\in L^1(\Omega_0;\mathbf{R}^n)$ and $\tilde\phi\in C_0(\Omega_0;\mathbf{R}^n)$ some extensions of h and ϕ to Ω_0, we may set for every $u\in BV_0$

$$\langle\tilde L,u\rangle=\int_{\Omega_0}\tilde h u\,dx+\int_{\Omega_0}\tilde\phi Du=\int_\Omega hu\,dx+\int_{\overline{\Omega}}\phi Du$$

and consider the problem

$$\min\left\{\int_{\Omega_0} f(\nabla u)\,dx+\int_{\Omega_0\setminus S_u} f^\infty(D^s u)+\int_{S_u} g(D^s u(x))\,d\#(x)-\gamma\langle\tilde L,u\rangle\ :\ u\in BV_0\right\}, \qquad (15.83)$$

where S_u denotes now the set of jumps of u on Ω_0. If $H\in\mathbf{C}_0(\Omega_0;\mathbf{R}^n)$ is such that $H'=h$ a.e. on Ω, we have

$$\langle\tilde L,u\rangle=\int_\Omega H'u\,dx+\int_{\Omega_0}\tilde\phi Du=\int_{\overline{\Omega}}(\phi-H)Du$$

and problem (15.83) can be written as

$$\min\left\{\int_{\overline{\Omega}} f\left(\frac{d\lambda}{dx}\right)dx + \int_{\overline{\Omega}\setminus A_\lambda} f^\infty(\lambda^s) + \int_{A_\lambda} g(\lambda(x))\,d\# - \gamma\langle\phi - H, \lambda\rangle : \lambda \in \mathbf{M}(\overline{\Omega}; \mathbf{R}^n),\right.$$
$$\left.\lambda(\overline{\Omega}) = 0\right\}.$$

Therefore, arguing as in the previous case, we obtain that a necessary (respectively, sufficient) condition for existence in the Dirichlet problem (15.83) is $\gamma_* \le \gamma \le \gamma^*$ (respectively, $\gamma_* < \gamma < \gamma^*$), with

$$\gamma_* = \left[\inf\left\{\tfrac{\langle\phi-H,\lambda\rangle}{\int_{\overline{\Omega}} g^\infty(\lambda)} \;:\; \lambda(\overline{\Omega}) = 0\right\}\right]^{-1},$$
$$\gamma^* = \left[\sup\left\{\tfrac{\langle\phi-H,\lambda\rangle}{\int_{\overline{\Omega}} g^\infty(\lambda)} \;:\; \lambda(\overline{\Omega}) = 0\right\}\right]^{-1}.$$

Remark 15.4.3. The Neumann and Dirichlet problems can be considered in the more general and interesting case of functions u defined on a subset Ω of $\mathbf{R}^n$. The functional $F : BV(\Omega; \mathbf{R}^m) \to [0, +\infty]$ is then of the form (see Sections 10.3 and 10.4)

$$F(u) = \int_\Omega f(\nabla u)\,dx + \int_{\Omega\setminus S_u} f^\infty(D^s u) + \int_{S_u} g([u], \nu_u)\,d\mathcal{H}^{n-1},$$

where $\mathcal{H}^{n-1}$ is the Hausdorff $(n-1)$-dimensional measure, $[u]$ is the jump of u along S_u, and ν_u is the normal versor to S_u. In this case the associated limit analysis problems have not been studied, and even the study of general conditions on f and g which imply the lower semicontinuity of F leaves some open questions.

Chapter 16

An introduction to shape optimization problems

In this section we give a quick introduction to shape optimization problems in a rather general framework, and we discuss some of their features, especially in relation with the existence of an optimal solution. Our goal is not to give here a detailed presentation of the many problems and results in this very wide field, but only to show that several optimization problems, often very important for applications in mechanics and engineering, cannot be formulated by taking a Banach function space of the competing admissible choices: a more appropriate framework consists in taking as admissible controls the elements of a class of domains. We refer the reader interested in a deeper knowledge and analysis of this fascinating field to one of the several books on the subject [153], [9], [198], [209], to the notes by Tartar [218], or to the recent collection of lecture notes by Bucur and Buttazzo [95].

A shape optimization problem is a minimization problem where the unknown variable runs over a class of domains; then every shape optimization problem can be written in the form

$$\min\big\{F(A)\ :\ A\in\mathcal{A}\big\}, \tag{16.1}$$

where $\mathcal{A}$ is the class of admissible domains and F is the cost function that one has to minimize over $\mathcal{A}$.

It must be noticed that the class $\mathcal{A}$ of admissible domains does not have any linear or convex structure, so in shape optimization problems it is meaningless to speak of convex functionals and similar notions. Moreover, even if several topologies on families of domains are available, in general there is not an a priori choice of a topology which allows us to apply the direct methods of the calculus of variations for obtaining the existence of at least an optimal domain.

We want to stress that, as also happens in other kinds of optimal control problems, in several situations an optimal domain does not exist; this is mainly because in these cases the minimizing sequences are highly oscillating and converge to a limit object only in a suitable "relaxed" sense. Then we may have, in these cases, only the existence of a relaxed solution that in general is not a domain, and whose characterization may change from problem to problem.

A general procedure to relax optimal control problems can be successfully developed by using the Γ-convergence scheme which provides the right topology that has to be used for sequences of admissible controls. In particular, for shape optimization problems, this provides the right notion of convergence for sequences of domains. Presenting in a detailed way the abstract framework of relaxed optimal control problems through the Γ-convergence would require us to develop several preliminary tools as background. This goes beyond our purposes, so we simply refer the interested reader to [64], where this framework was first introduced, or to [95], [101].

Coming back to the minimization problem (16.1), in general, unless some geometrical constraints on the admissible sets are assumed, or some very special cases of cost functionals are considered, the existence of an optimal domain may fail. In these situations the discussion will then be focused on the relaxed solutions that always exist.

As usually happens in all optimization problems, to give a qualitative description of the optimal solutions of a shape optimization problem, it is important to derive the so-called necessary conditions of optimality. These conditions have to be derived from the comparison of the cost of an optimal solution A_{opt} to the cost of other suitable admissible choices, close enough to A_{opt}. This procedure is what is usually called a *variation* near the solution. The difficulty in obtaining necessary conditions of optimality for shape optimization problems consists in the fact that being the unknown variables domains, the notion of neighborhood is not a priori clear; the possibility of choosing a domain variation could then be rather wide. The same method can be applied, when no classical solution exists, to relaxed solutions, and this will provide some qualitative information about the behavior of the minimizing sequences of the original problem.

Finally, for some particular problems presenting special behaviors or symmetries, one would like to exhibit explicit solutions (balls, ellipsoids, …). This could be very difficult, even for simple problems, and often, instead of having established results, one can give only conjectures.

In general, since the explicit computations are difficult, one should develop efficient numerical schemes to produce approximated solutions. This is a challenging field we will not enter; we refer the interested reader to the books and papers available on the subject [9], [198], [209].

In the following examples we show that several classical optimization problems can be written in the form (16.1).

16.1 The isoperimetric problem

The isoperimetric problem is certainly the oldest shape optimization problem; it seems to go back to the Greek golden period of mathematics (Archimedes, Zenodorus, …), and a legend about Queen Dido shows that the question was clearly formulated long ago. The problem with constraint Q (see, for instance, [59]) consists in finding among all Borel subsets A of a given closed set $Q \subset \mathbf{R}^N$ the one which minimizes the perimeter, once its Lebesgue measure, or more generally the quantity $\int_A f(x)\,dx$ for a given function $f \in L^1_{loc}(\mathbf{R}^N)$, is prescribed. With this notation the isoperimetric problem can be then formulated in the form (16.1) if we take

$$F(A) = \mathrm{Per}(A),$$

$$\mathcal{A} = \left\{ A \subset Q \ : \ \int_A f(x)\,dx = c \right\}.$$

Here the perimeter of a Borel set is the one defined in chapter 10 as

$$\mathrm{Per}(A) = \int |D1_A| = \mathcal{H}^{N-1}(\partial^* A),$$

where $D1_A$ is the distributional derivative of the characteristic function of A, $\mathcal{H}^{N-1}$ is the $(N-1)$-dimensional Hausdorff measure introduced in Section 4.1, and $\partial^* A$ is the reduced boundary defined in Section 10.3. By using the properties of the BV spaces seen in chapter 10, when Q is bounded we obtain the lower semicontinuity and the coercivity of the perimeter for the L^1 convergence, which enables us to apply the direct methods of the calculus of variations of Section 3.2 and to obtain straightforwardly the existence of an optimal solution for the problem

$$\min \left\{ \mathrm{Per}(A) \ : \ A \subset Q, \int_A f\,dx = c \right\}. \tag{16.2}$$

It is also very simple to show that in general the problem above may have no solution if we drop the assumption that Q is bounded (see, for instance, [95]). Take indeed $f \equiv 1$, $c = \pi$, and Q the countable union of all closed disks in $\mathbf{R}^2$ of the form $\overline{B(x_n, r_n)}$, where $x_n = (2n, 0)$ and $r_n = 1 - 1/n$ (see Figure 16.1). It is then easy to see that the infimum of problem (16.2) is 2π, whereas no admissible domain in $\mathcal{A}$ provides the value 2π to the cost functional.

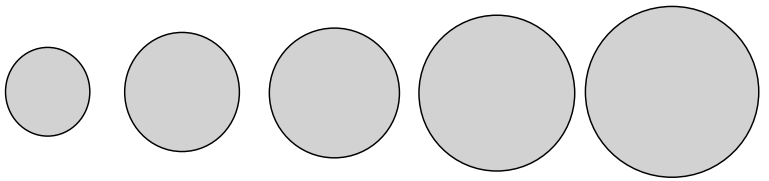


Figure 16.1. *An unbounded set Q.*

On the other hand, it is very well known that the classical isoperimetric problem, with $Q = \mathbf{R}^N$ and $f \equiv 1$, admits a solution which is any ball of measure c, even if the complete proof of this fact requires very delicate tools, especially when the dimension N is larger than 2. A complete characterization of pairs (Q, f) which provide the existence of a solution for the problem (16.2) seems to be difficult.

A variant of the isoperimetric problem consists in not counting some parts of the boundary ∂A in the cost functional. More precisely, if Q is the closure of an open set Ω with a Lipschitz boundary, we may consider problem (16.2) with $\mathrm{Per}(A)$ replaced by the cost functional

$$\mathrm{Per}_\Omega(A) = \int_\Omega |D1_A| = \mathcal{H}^{N-1}(\Omega \cap \partial^* A)$$

which does not count the part of ∂A which is included in $\partial\Omega$. The existence of a solution when Ω is bounded still holds, as above, together with nonexistence examples when this

boundedness condition is dropped. Indeed, it is enough to take f, c, Q as above and to observe that the infimum of problem (16.2) is in this case zero, whereas no admissible domain provides the value zero to the cost functional.

16.2 The Newton problem

Another classical question which can be considered as a shape optimization problem is the determination of the best aerodynamical profile for a body in a fluid stream under some constraints on its size. This problem, at least within the class of radially symmetric bodies, which makes the problem one-dimensional, was first considered by Newton, who gave a rather simple variational expression for the aerodynamical resistance of a convex body in a fluid stream. Here are his words (from *Principia Mathematica*):

If in a rare medium, consisting of equal particles freely disposed at equal distances from each other, a globe and a cylinder described on equal diameter move with equal velocities in the direction of the axis of the cylinder, (then) the resistance of the globe will be half as great as that of the cylinder.... I reckon that this proposition will be not without application in the building of ships.

Indeed, if we make the assumption that the resistance is due to the impact of fluid particles against the body surface, if all the particles are supposed independent (which is quite reasonable if the fluid is rarefied), and if the tangential friction is neglected, simple geometric considerations lead us to obtain for the resistance along the direction of the fluid stream the expression

$$R(u) = \int_\Omega \frac{1}{1+|Du|^2}\,dx, \tag{16.3}$$

where we normalized to one all the physical multiplicative constants involving the density and the velocity of the fluid. Here Ω represents the cross section of the body at the basis level, and $u(x)$ is a function whose graph is the body upper boundary. Since the validity of the model requires that all particles hit the body at most once, we consider only convex bodies, which turns out to require Ω convex and $u : \Omega \to [0, +\infty[$ concave.

Note that the integral functional F above is neither convex nor coercive. Therefore, obtaining an existence theorem for minimizers via the usual direct method in the calculus of variations may fail. Indeed, if we do not impose any further constraint on the competing functions u, the infimum of the functional in (16.3) turns out to be zero, as immediately seen by taking, for instance,

$$u_n(x) = n\,\mathrm{dist}(x, \partial\Omega)$$

for every $n \in \mathbf{N}$ and by letting $n \to +\infty$. Therefore, no function u can minimize the functional F, because $F(u) > 0$ for every function u.

A complete discussion of the problem can be found in [95], where all the relevant references are quoted. Here we simply recall that the problem

$$\min\left\{ \int_\Omega \frac{1}{1+|Du|^2}\,dx \;:\; u \text{ concave},\ 0 \le u \le M \right\}$$

admits a solution u_{opt}. Some interesting necessary conditions of optimality can be deduced: for instance (see [166]), it can be proved that on every open set ω where u_{opt} is of class C^2

we obtain

$$\det D^2 u(x) = 0 \qquad \forall x \in \omega.$$

In particular, this excludes that in the case $\Omega = B(0, R)$ the solution u_{opt} is radially symmetric. The optimal radially symmetric profile and a nonsymmetric profile which is better that all the radial ones are shown, respectively, in Figures 16.2 and 16.3.

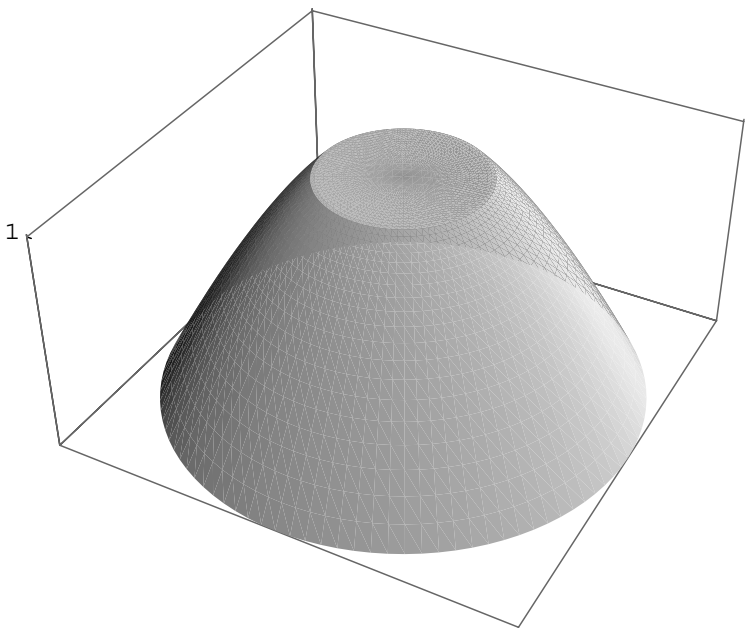


Figure 16.2. *The optimal radial shape for $M = R$.*

It is interesting to notice that with simple calculations, one can write the optimization problem above in the form (16.1) by taking the cost functional as a boundary integral,

$$F(A) = \int_{\partial A} j\big(x, \nu(x)\big)\, d\mathcal{H}^{N-1},$$

for a suitable integrand $j(x, s)$, being $\nu(x)$ the exterior normal unit vector to ∂A at x and $\mathcal{H}^{N-1}$ the Hausdorff $(N-1)$-dimensional measure (see [103]).

16.3 Optimal Dirichlet free boundary problems

We consider now the model example of a Dirichlet problem over an unknown domain, which has to be optimized according to a given cost functional. More precisely, we consider a given bounded open subset Ω of $\mathbf{R}^N$, an admissible class $\mathcal{A}$ of subsets of Ω, a given function

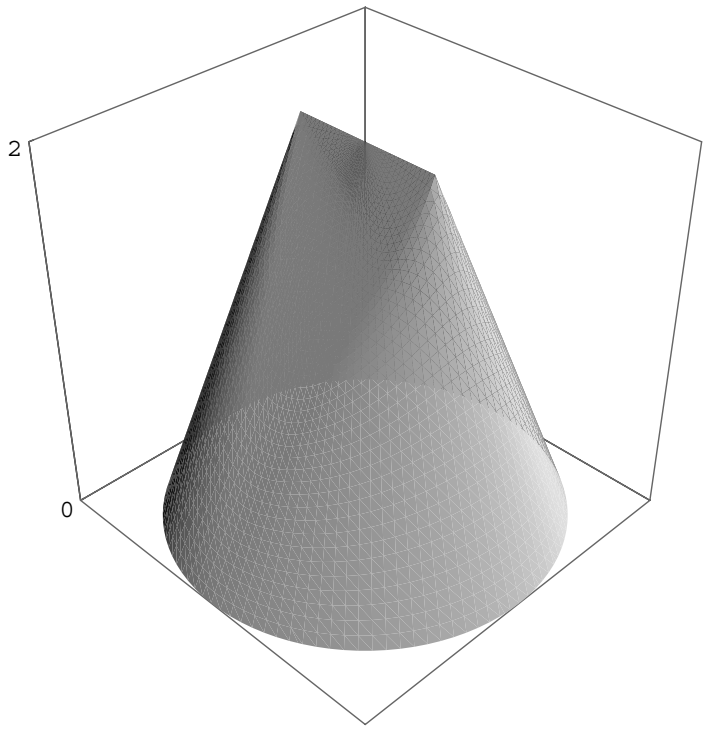


Figure 16.3. *A nonradial profile better than all radial ones for $M = 2R$.*

$f \in L^2(\Omega)$, and a cost functional of the form

$$F(A) = \int_\Omega j(x, u_A)\, dx. \tag{16.4}$$

Here the integrand $j : \Omega \times \mathbf{R} \to \mathbf{R}$ is given, and we denote by u_A the unique solution of the elliptic problem

$$\begin{cases} -\Delta u = f & \text{in } A, \\ u \in H_0^1(A), \end{cases} \tag{16.5}$$

extended by zero to $\Omega \setminus A$.

It is well known that in general one should not expect the existence of an optimal solution; below we show an example where the existence of an optimal domain does not occur. The problem we consider is of the form (16.1), where the admissible class $\mathcal{A}$ consists of all subdomains of a given bounded open subset Ω of $\mathbf{R}^N$ and the cost functional F is of the form (16.4) with

$$j(x, s) = |s - \overline{u}(x)|^2$$

for a prescribed desired state $\overline{u}$. In the thermostatic model the shape optimization problem (16.1) with the choices above consists in finding an optimal distribution, inside Ω, of the

Dirichlet region $\Omega \setminus A$ to achieve a temperature which is as close as possible to the desired temperature $\overline{u}$, once the heat sources f are prescribed.

For simplicity, we consider a uniformly distributed heat source, that is, we take $f \equiv 1$, and we take the desired temperature $\overline{u}$ constantly equal to $c > 0$. Therefore, problem (16.1) becomes

$$\min \left\{ \int_\Omega |u_A - c|^2 \, dx \ : \ -\Delta u_A = 1 \text{ in } A, \ u_A \in H_0^1(A) \right\}. \tag{16.6}$$

We will actually show that for small values of the constant c no regular domain A can solve problem (16.6) above; the proof of nonexistence of any domain is slightly more delicate and requires additional tools like the capacitary form of necessary conditions of optimality (see, for instance, [95], [98], [99], [109]).

Proposition 16.3.1. *If $c > 0$ is small enough, then problem* (16.6) *has no smooth solutions.*

Proof. The nonexistence proof can be obtained by contradiction. Assume indeed that a regular domain A solves the optimization problem (16.6) and that A does not coincide with the whole set Ω. Take a point $x_0 \in \Omega \setminus \overline{A}$ and a small ball B_ε of radius ε sufficiently small, centered at x_0. If u_A denotes the solution of (16.5) corresponding to A, and if ε is small enough so that B_ε does not intersect A, then the solution $u_{A\cup B_\varepsilon}$, corresponding to the admissible choice $A \cup B_\varepsilon$, is given by

$$u_{A\cup B_\varepsilon}(x) = \begin{cases} u_A(x) & \text{if } x \in A, \\ (\varepsilon^2 - |x - x_0|^2)/4 & \text{if } x \in B_\varepsilon, \\ 0 & \text{otherwise.} \end{cases}$$

Therefore, we obtain

$$F(A) = \int_A |u_A - c|^2 \, dx + \int_{B_\varepsilon} c^2 \, dx + \int_{\Omega\setminus(A\cup B_\varepsilon)} c^2 \, dx,$$

$$F(A \cup B_\varepsilon) = \int_A |u_A - c|^2 \, dx + \int_{B_\varepsilon} \left| \frac{\varepsilon^2 - |x - x_0|^2}{4} - c \right|^2 dx + \int_{\Omega\setminus(A\cup B_\varepsilon)} c^2 \, dx.$$

By using the minimality of A this then yields

$$\begin{aligned} c^2 \operatorname{meas}(B_\varepsilon) &\le \int_{B_\varepsilon} \left| \frac{\varepsilon^2 - |x - x_0|^2}{4} - c \right|^2 dx \\ &= N\varepsilon^{-N} \operatorname{meas}(B_\varepsilon) \int_0^\varepsilon \left| \frac{\varepsilon^2 - r^2}{4} - c \right|^2 r^{N-1} \, dr \\ &= c^2 \operatorname{meas}(B_\varepsilon) + \frac{1}{16} \int_0^\varepsilon (\varepsilon^2 - r^2)(\varepsilon^2 - r^2 - 8c) r^{N-1} \, dr, \end{aligned}$$

which, for a fixed $c > 0$, turns out to be false if ε is small enough.

Thus all regular domains $A \neq \Omega$ are ruled out by the argument above. We can now exclude also the case $A = \Omega$ if c is small enough, by comparing the full domain Ω to the empty set $\emptyset$. This gives, taking into account that $u_\emptyset \equiv 0$,

$$F(\Omega) = \int_\Omega |u_\Omega - c|^2 \, dx,$$

$$F(\emptyset) = \int_\Omega c^2\,dx,$$

so that we have $F(\emptyset) < F(\Omega)$ if c is small enough. Hence all regular subdomains of Ω are excluded, and the nonexistence proof is achieved. □

Nevertheless, the existence of an optimal domain occurs for problem (16.1) in some particular cases:

(i) when severe geometrical constraints on the class of admissible domains are imposed;

(ii) when the cost functional fulfills some particular qualitative assumptions;

(iii) when the problem is of a very special type, involving, for instance, only the eigenvalues of the Laplace operator, and where neither geometrical constraints nor monotonicity of the cost is required.

See the lecture notes of Bucur and Buttazzo [95] for a more complete discussion on this topic; here, for the sake of simplicity, we list only some situations when conditions (i), (ii), and (iii) occur.

Concerning case (i), a rather simple situation when the existence of an optimal solution occurs is when the class of admissible domains fulfills the geometrical constraint which is called the *exterior cone condition*. It consists in requiring that there exists a fixed height h and opening ω such that for every domain A of the admissible class $\mathcal{A}$ and every point $x_0 \in \partial A$ a cone with height h, opening ω, and vertex at x_0 is contained in $\Omega \setminus A$. This condition is weaker than an equi-Lipschitz conditions on the class of admissible domains; conditions weaker than the exterior cone but which still imply the existence of an optimal domain can be given in terms of capacity (see [95]).

Concerning case (ii), we consider the problem

$$\min\big\{F(A)\ :\ A \in \mathcal{A},\ \text{meas}(A) \le m\big\}, \tag{16.7}$$

where the volume constraint $\text{meas}(A) \le m$ has been added. The cost functional F is assumed to be monotone nonincreasing with respect to the set inclusion, that is,

$$A_1 \subset A_2 \quad \Rightarrow \quad F(A_2) \le F(A_1).$$

Moreover, F is assumed to be lower semicontinuous with respect to the γ-convergence on the class of domains, defined by

$$A_n \to A \text{ in the } \gamma\text{-convergence} \quad \Longleftrightarrow \quad u_{A_n} \to u_A \text{ weakly in } H_0^1(\Omega),$$

where u_{A_n} and u_A are the solutions of (16.5) in A_n and A, respectively, with the right-hand side $f \equiv 1$. Under the assumptions above, Buttazzo and Dal Maso in [100] showed that problem (16.5) admits an optimal solution. It is important to stress that several optimal shape problems can be written in the form (16.7) with a cost functional F which is nonincreasing for the set inclusion and γ-lower semicontinuous. For instance, if L denotes a second-order elliptic operator of the form

$$Lu = -\,\text{div}\big(a(x)Du\big)$$

with the $N \times N$ matrix $a(x)$ symmetric, uniformly elliptic, and with bounded and measurable coefficients, we may consider the spectrum $\Lambda(A)$ of L associated to the Dirichlet boundary conditions on A:

$$Lu = \lambda u, \qquad u \in H_0^1(A).$$

It is known that $\Lambda(A)$ is given by a nonnegative sequence $\lambda_k(A)$ which tends to $+\infty$, that every $\lambda_k(A)$ is a nonincreasing set function with respect to A, and that the mappings $\lambda_k(A)$ are γ-continuous. Therefore the existence result above applies and we obtain that the problem

$$\min\big\{\Phi\big(\Lambda(A)\big) \ : \ A \in \mathcal{A},\ \mathrm{meas}(A) \le m\big\}$$

admits an optimal solution whenever the mapping $\Phi : \mathbf{R}^{\mathbf{N}} \to [0, +\infty]$ is

- nondecreasing in the sense that

$$\Lambda_k^1 \le \Lambda_k^2 \ \forall\, k \quad \Rightarrow \quad \Phi(\Lambda^1) \le \Phi(\Lambda^2);$$

- lower semicontinuous in the sense that

$$\Lambda_k^n \to \Lambda_k \ \forall\, k \quad \Rightarrow \quad \Phi(\Lambda) \le \liminf_{n \to +\infty} \Phi(\Lambda^n).$$

In particular, for a fixed integer k, all cost functionals of the form

$$F(A) = \phi\big(\lambda_1(A), \ldots \lambda_k(A)\big)$$

with a mapping $\phi : \mathbf{R}^k \to \mathbf{R}$ continuous and nondecreasing in each variable verify the assumptions above.

Concerning case (iii), we may still prove the existence of an optimal domain for problem (16.7) for some special form of the cost functional. The case which has been considered in [96] (see also [95]) is when

$$F(A) = \phi\big(\lambda_1(A), \lambda_2(A)\big),$$

where $\lambda_1(A)$ and $\lambda_2(A)$ are the first two eigenvalues of the Laplace operator $-\Delta$ on $H_0^1(A)$. Due to the special form of the cost functional and to the fact that the operator is $-\Delta$, it is possible to show that an optimal solution exists without any monotonicity assumption on ϕ by requiring only that ϕ is a lower semicontinuous function on $\mathbf{R}^2$.

16.4 Optimal distribution of two conductors

Another interesting case of a shape optimization problem is the optimal distribution of two given conductors into a given set. If Ω denotes a bounded open subset of $\mathbf{R}^N$ (the prescribed container), denoting by α and β the conductivities of the two materials, the problem consists in filling Ω with the two materials in the best performing way according to some given cost functional. The volume of each material can also be prescribed. It is convenient to denote by A the domain where the conductivity is α and by $a_A(x)$ the conductivity coefficient

$$a_A(x) = \alpha 1_A(x) + \beta 1_{\Omega \setminus A}(x).$$

In this way the state equation becomes

$$\begin{cases} -\operatorname{div}\big(a_A(x)Du\big) = f \text{ in } \Omega, \\ u = 0 \text{ on } \partial\Omega, \end{cases} \tag{16.8}$$

where f is the (given) source density, and we denote by u_A its unique solution.

It is well known (see, for instance, [164], [192]) that if we take as a cost functional an integral of the form

$$\int_\Omega j(x, 1_A, u_A, Du_A)\,dx,$$

in general an optimal configuration does not exist. However, the addition of a perimeter penalization is enough to imply the existence of classical optimizers. In other words, if we take as a cost the functional

$$J(u, A) = \int_\Omega j(x, 1_A, u, Du)\,dx + \sigma \operatorname{Per}_\Omega(A),$$

where $\sigma > 0$, the problem can be written as an optimal control problem in the form

$$\min\big\{J(u, A) \ : \ A \subset \Omega,\ u \text{ solves (16.8)}\big\}. \tag{16.9}$$

A volume constraint of the form $\operatorname{meas}(A) = m$ could also be present. The main ingredient for the proof of the existence of an optimal classical solution is the following result.

Theorem 16.4.1. *Let $a_n(x)$ be a sequence of $N \times N$ symmetric matrices with measurable coefficients, such that the uniform ellipticity condition*

$$c_0|z|^2 \le a_n(x)z \cdot z \le c_1|z|^2 \qquad \forall x \in \Omega,\ \forall z \in \mathbf{R}^N \tag{16.10}$$

holds with $0 < c_0 \le c_1$. Given $f \in H^{-1}(\Omega)$ we denote by u_n the unique solution of the problem

$$-\operatorname{div}\big(a_n(x)Du\big) = f, \qquad u_n \in H^1_0(\Omega). \tag{16.11}$$

If $a_n(x) \to a(x)$ a.e. in Ω, then $u_n \to u$ weakly in $H^1_0(\Omega)$, where u is the solution of (16.11) *with a_n replaced by a.*

Proof. By the uniform ellipticity condition (16.10) we have

$$c_0 \int_\Omega |Du_n|^2\,dx \le \int_\Omega f u_n\,dx$$

and, by the Poincaré inequality we have that u_n are bounded in $H^1_0(\Omega)$ so that a subsequence (still denoted by the same indices) converges weakly in $H^1_0(\Omega)$ to some v. All we have to show is that $v = u$, or equivalently that

$$-\operatorname{div}\big(a(x)Dv\big) = f. \tag{16.12}$$

This means that for every smooth test function ϕ we have

$$\int_\Omega a(x)Dv D\phi\,dx = \langle f, \phi\rangle.$$

Then it is enough to show that for every smooth test function ϕ we have

$$\lim_{n\to+\infty}\int_\Omega a_n(x)Du_nD\phi\,dx=\int_\Omega a(x)Dv D\phi\,dx.$$

This is an immediate consequence of the fact that ϕ is smooth, $Du_n\to Dv$ weakly in $L^2(\Omega)$, and $a_n\to a$ a.e. in Ω remaining bounded.

Another way to show that (16.12) holds is to verify that v minimizes the functional

$$F(w)=\int_\Omega a(x)Dw Dw\,dx-2\langle f,w\rangle,\qquad w\in H_0^1(\Omega).\tag{16.13}$$

Since the function $\alpha(s,z)=sz\cdot z$, which is defined for all $z\in\mathbf{R}^N$ and for symmetric positive definite $N\times N$ matrices s, is convex in z and lower semicontinuous in s, the functional

$$\Phi(a,\xi)=\int_\Omega a(x)\xi\cdot\xi\,dx$$

is sequentially lower semicontinuous with respect to the strong L^1 convergence on a and the weak L^1 convergence on ξ (see, for instance, Theorem 13.1.1 and [97]). Therefore we have

$$F(v)=\Phi(a,Dv)-2\langle f,v\rangle\le\liminf_{n\to+\infty}\Phi(a_n,Du_n)-2\langle f,u_n\rangle=\liminf_{n\to+\infty}F(u_n).$$

Since u_n minimizes the functional F_n defined as in (16.13) with a replaced by a_n, we also have for every $w\in H_0^1(\Omega)$

$$F_n(u_n)\le F_n(w)=\int_\Omega a_n(x)Dw Dw\,dx-2\langle f,w\rangle,$$

so that taking the limit as $n\to+\infty$ and using the convergence $a_n\to a$ we obtain

$$\liminf_{n\to+\infty}F_n(u_n)\le\int_\Omega a(x)Dw Dw\,dx-2\langle f,w\rangle=F(w).$$

Thus $F(v)\le F(w)$, which shows what is required. □

Remark 16.4.1. The result above can be rephrased in terms of G-convergence by saying that for uniformly elliptic operators of the form $-\operatorname{div}\big(a(x)Du\big)$ the G-convergence is weaker than the L^1-convergence of coefficients. Analogously, we can say that the functionals

$$G_n(w)=\int_\Omega a_n(x)Dw Dw\,dx$$

Γ-converge, with respect to the $L^2(\Omega)$ convergence, to the functional G defined in the same way with a in the place of a_n.

Corollary 16.4.1. *If $A_n\to A$ in $L^1(\Omega)$, then $u_{A_n}\to u_A$ weakly in $H_0^1(\Omega)$.*

A more careful inspection of the proof of Theorem 16.4.1 shows that the following stronger result holds.

Theorem 16.4.2. *Under the same assumptions of Theorem* 16.4.1, *the convergence of u_n is actually strong in $H^1_0(\Omega)$.*

PROOF. We have already seen that $u_n \to u$ weakly in $H^1_0(\Omega)$, which gives $Du_n \to Du$ weakly in $L^2(\Omega)$. Denoting by $c_n(x)$ and $c(x)$ the square root matrices of $a_n(x)$ and $a(x)$, respectively, we have that $c_n \to c$ a.e. in Ω remaining equibounded. Then $c_n(x)Du_n$ converges to $c(x)Du$ weakly in $L^2(\Omega)$. Multiplying (16.4) by u_n and integrating by parts we obtain

$$\begin{aligned}\int_\Omega a(x)Du\,Du\,dx &= \langle f, u\rangle = \lim_{n\to+\infty}\langle f, u_n\rangle \\ &= \lim_{n\to+\infty}\int_\Omega a_n(x)Du_n\,Du_n\,dx.\end{aligned}$$

This implies that

$$c_n(x)Du_n \to c(x)Du \qquad \text{strongly in } L^2(\Omega).$$

Multiplying now by $\big(c_n(x)\big)^{-1}$ we finally obtain the strong convergence of Du_n to Du in $L^2(\Omega)$. □

We are now in a position to obtain an existence result for the optimization problem (16.2). On the function j we only assume that it is nonnegative, Borel measurable, and such that $j(x, s, z, w)$ is lower semicontinuous in (s, z, w) for a.e. $x \in \Omega$.

Theorem 16.4.3. *Under the assumptions above, the minimum problem* (16.2) *admits at least a solution.*

PROOF. Let (A_n) be a minimizing sequence; then $\mathrm{Per}_\Omega(A_n)$ are bounded so that, up to extracting subsequences, we may assume (A_n) is strongly convergent in the L^1_{loc} sense to some set $A \subset \Omega$. We claim that A is a solution of problem (16.2). Let us denote by u_n a solution of problem (16.1) associated to A_n; by Theorem 16.4.2, (u_n) converges strongly in $H^1_0(\Omega)$ to some $u \in H^1_0(\Omega)$. Then by the lower semicontinuity of the perimeter (see Proposition 10.1.1) and by Fatou's lemma we have

$$J(u, A) \le \liminf_{n\to+\infty} J(u_n, A_n),$$

which proves the optimality of A. □

Remark 16.4.2. The same proof works when volume constraints of the form $\mathrm{meas}(A) = m$ are present. Indeed this constraint passes to the limit when $A_n \to A$ strongly in $L^1(\Omega)$.

The existence result above shows the existence of a classical solution for the optimization problem (16.2). This solution is simply a set with finite perimeter and additional assumptions have to be made to prove further regularity. For instance, in [18] Ambrosio and Buttazzo considered the similar problem

$$\min\Big\{E(u, A) + c\,\mathrm{Per}_\Omega(A) \;:\; u \in H^1_0(\Omega),\; A \subset \Omega\Big\},$$

where $c > 0$ and

$$E(u, A) = \int_\Omega \left[a_A(x)|Du|^2 + 1_A(x)g_1(x, u) + 1_{\Omega\setminus A}g_2(x, u)\right] dx.$$

They showed that every solution A is actually an open set provided g_1 and g_2 are Borel measurable and satisfy inequalities

$$g_i(x, s) \geq \gamma(x) - k|s|^2, \qquad i = 1, 2,$$

where $\gamma \in L^1(\Omega)$ and $k < \alpha\lambda_1$, λ_1 being the first eigenvalue of $-\Delta$ on Ω.

Bibliography

[1] Y. Abddaimi, C. Licht, G. Michaille. *Stochastic homogenization of an integral functional of quasiconvex function with linear growth.* Asymptotic Analysis **15** (1997), 183–202.

[2] N. Acerbi, N. Fusco. *Semicontinuity problems in the calculus of variations.* Mech. Anal. **86** (1984), 125–145.

[3] M. A. Ackoglu, U. Krengel. *Ergodic theorem for superadditive processes.* J. Reine Angew. Math. **323** (1981), 53–67.

[4] S. Adly, M. Thera, E. Ernst. *Stability of the solution set of non-coerecive variational inequalities.* Comm. Contemp. Math. **4** (2002), no. 1, 145–160.

[5] S. Adly, E. Ernst, M. Thera. *A characterization of convex and semicoercive functionals.* J. Convex Anal. **8** (2001), no. 1, 127–148.

[6] S. Agmon, A. Douglis, L. Nirenberg. *Estimates near the boundary for solutions of elliptic partial differential equations satisfying general boundary conditions* I. Comm. Pure Appl. Math. **12** (1959), 623–727.

[7] S. Agmon, A. Douglis, L. Nirenberg. *Estimates near the boundary for solutions of elliptic partial differential equations satisfying general boundary conditions* II. Comm. Pure Appl. Math. **17** (1964), 35–92.

[8] G. Alberti. *Rank one properties for derivatives of functions with bounded variation.* Proc. Roy. Soc. Edinburgh **A-123** (1993), 239–274.

[9] G. Allaire. *Shape optimization by the homogeneization method.* Appl. Math. Sci. **146**, Springer-Verlag, New York, 2002.

[10] F. Alvarez, J.-P. Mandallena. *Multiparameter homogenization by localization and blow-up.* Proc. Roy. Soc. Edinburgh **134A** (2004), 801–814.

[11] F. Alvarez, J.-P. Mandallena. *Homogenization of multiparameter integrals.* Nonlinear Anal. **50** (2002), 839–870.

[12] A. Ambrosetti, P. Rabinowitz. *Dual variational methods in critical point theory and applications.* J. Funct. Anal. **14** (1973), 349–381.

[13] L. Ambrosio. *A compactness theorem for a special class of functions of bounded variation.* Boll. Un. Mat. It. **3-B** (1989), 857–881.

[14] L. Ambrosio. *A new proof of the SBV compactness theorem.* Università di Pisa, 2.148 **754**, Sept. 1993.

[15] L. Ambrosio. *On the lower semicontinuity of quasi-convex integrals in $SBV(\Omega;\mathbf{R}^k)$.* Nonlinear Anal. **23** (1994), 405–425.

[16] L. Ambrosio. *Corso introduttivo alla Theoria Geometrica della misura ed alle Superfici Minime.* Appunti dei corsi tenuti da documenti della Scuola, Scuola Normale Superiore, Pisa, 1997.

[17] L. Ambrosio, A. Braides. *Energies in SBV and variational models in fracture mechanics.* Homogenization and Applications to Material Sciences **9**. D. Cioranescu, A. Damlamian, P. Donato eds, Gakuto, Gakkotosho, Tokyo, Japan (1997), 1–22.

[18] L. Ambrosio, G. Buttazzo. *An optimal design problem with perimeter penalization.* Calc. Var. **1** (1993), 55–69.

[19] L. Ambrosio, G. Buttazzo. *Weak lower semicontinuous envelope of functionals defined on a space of measures.* Ann. Mat. Pura Appl. **150** (1988), 311–340.

[20] L. Ambrosio, G. Dal Maso. *A general chain rule for distributional derivatives.* Proc. Amer. Math. Soc. **108** (1990), 691–702.

[21] L. Ambrosio, G. Dal Maso. *On the relaxation in $BV(\Omega;\mathbf{R}^m)$ of quasi-convex integrals.* J. Funct. Anal. **109** (1992), 76–97.

[22] L. Ambrosio, N. Fusco, D. Pallara. *Functions of bounded variation and free discontinuity problems.* Oxford Mathematical Monographs, Oxford University Press, New York, 2000.

[23] L. Ambrosio, P. Tilli. *Topics on analysis in metric spaces.* Oxford Lecture Series in Math. Appl. **25**, Oxford University Press, Oxford, UK, 2004.

[24] L. Ambrosio, V. M. Tortorelli. *Approximation of functionals depending on jumps by elliptic functionals via Γ-convergence.* Comm. Pure Appl. Math. **18** (1990), 999–1036.

[25] O. Anza Hafsa. *Variational formulations on thin elastic plates with constraints.* J. Convex Anal. (to appear).

[26] O. Anza Hafsa, J.-P. Mandallena. *Interchange of infimum and integral.* Calc. Var. **18** (2003), 433–449.

[27] O. Anza Hafsa, J.-P. Mandallena. *Relaxation of second order geometric integrals and non-local effects.* J. Nonlinear Convex Anal. **5** (2004), 295–306.

[28] H. Attouch. *Variational convergence for functions and operators*. Applicable Mathematics Series, Pitman Advanced Publishing Program, Boston, 1985.

[29] H. Attouch. *Viscosity solutions of minimization problems.* SIAM J. Optim. **6** (1996), 769–806.

[30] H. Attouch, G. Bouchitté, M. Mabrouk. *Variational formulations for semilinear elliptic equations involving measures.* Nonlinear Variational Problems, vol II, A. Marino, A. Murthy, eds., Pitman Res. Notes in Math. **193** (1989), 1–56.

[31] H. Attouch, H. Brezis. *Duality for the sum of convex functions in general Banach spaces.* Aspects of Mathematics and Its Applications, J. Barroso, ed., North-Holland, Amsterdam (1986), 125–133.

[32] H. Attouch, T. Champion. *L_p regularization of the non-parametric minimal surface problem.* Ill-Posed Variational Problems and Regularization Techniques, Lecture Notes in Econom. and Math. Systems **477** (1999), 25–34.

[33] H. Attouch, R. Cominetti. *L_p approximation of variational problems in L_1 and L_∞.* Nonlinear Anal. **36** (1999), no. 3, 373–399.

[34] H. Attouch, A. Damlamian. *Applications des méthodes de convexité et monotonie à l'étude de certaines équations quasi-linéaires.* Proc. Roy. Soc. Edinburgh **79A** (1977), 107–129.

[35] H. Attouch, C. Picard. *Variational inequalities with varying obstacles: The general form of the limit problem.* J. Funct. Anal. **50** (1983), no. 3, 329–386.

[36] H. Attouch, H. Riahi. *Stability results for Ekeland's ε-variational principle and cone extremal solutions.* Math. Oper. Res. **18** (1993), no. 1, 173–201.

[37] H. Attouch, A. Soubeyran. *Towards stable routines: Improving and satisficing enough by exploration-exploitation on an unknown landscape.* Working paper GREQAM, Montpellier (2004).

[38] H. Attouch, M. Thera. *A general duality principle for the sum of two operators.* J. Convex Anal. **3** (1996), no. 1, 1–24.

[39] H. Attouch, R. J.-B. Wets. *Epigraphical analysis.* In Analyse non linéaire, Perpignan 1987, and Ann. Inst. H. Poincaré, Anal. Non Linéaire **6** (1989), 73–100.

[40] H. Attouch, R. J.-B. Wets. *Quantitative stability of variational systems:* I. *The epigraphical distance.* Trans. Amer. Math. Soc. **328** (1991), no. 2, 695–729.

[41] G. Aubert, R. Deriche, P. Kornprobst. *Computing optical flow via variational techniques.* SIAM J. Appl. Math. **60** (1999), 156–182.

[42] G. Aubert, P. Kornprobst. *A mathematical study of the relaxed optical flow problem in the space $BV(\Omega)$.* SIAM J. Math. Anal. **30** (1999), 1282–1308.

[43] G. Aubert, P. Kornprobst. *Mathematical problems in image processing: Partial differential equations and the calculus of variations.* Appl. Math. Sci. **147**, Springer-Verlag, New York, 2002.

[44] Th. Aubin. *Problèmes isopérimétriques et espaces de Sobolev.* J. Diff. Geom., **11** (1976), 573–598.

[45] J.P. Aubin. *Mathematical methods of game and economic theory.* North-Holland, Amsterdam, 1979.

[46] J.P. Aubin, I. Ekeland. *Applied nonlinear analysis.* John Wiley, New York, 1984.

[47] J.P. Aubin, H. Frankowska. *Set-valued analysis*. Systems Control Found. Appl. **2**, Birkhäuser-Verlag, Boston, 1990.

[48] A. Auslender. *Noncoercive optimization problems.* Math. Oper. Res. **21** (1996), 769–782.

[49] D. Azé. *Eléments d'Analyse Convexe et Variationnelle.* Ellipse, Paris, 1997.

[50] C. Baiocchi, G. Buttazzo, F. Gastaldi, F. Tomarelli. *General existence theorems for unilateral problems in continuum mechanics.* Arch. Ration. Mech. Anal. **100** (1988), no. 2, 149–189.

[51] C. Baiocchi, F. Gastaldi. *Inéquations variationnelles non coercives.* C.R. Acad. Sci. Paris **299** (1984), 647–650.

[52] C. Baiocchi, F. Gastaldi. *Some existence results on noncoercive variational inequalities.* Ann. Scuola Norm. Sup. Pisa (4) **13** (1986), 617–659.

[53] E.J. Balder. *Lectures on Young measures theory and its applications in economics.* Workshop di Teoria della Misura e Analisi Reale, Grado, 1997, Rend. Instit. Univ. Trieste **31** (2000), suppl. 1, 1–69.

[54] J. M. Ball *Convexity conditions and existence theorems in nonlinear elasticity.* Arch. Ration. Mech. Anal. **63** (1977), 13–23.

[55] J.M. Ball. *A version of the fundamental theorem for Young measures.* PDE's and Continuum Models of Phase Transitions, M. Rascle, D. Serre, M. Slemrod, eds., Lecture Notes in Physics 344, Springer-Verlag, New York, 1989, 207–215.

[56] J.M. Ball, R.D. James. *Fine phase mixtures as minimizers of energy.* Arch. Ration. Mech. Anal. **100** (1987), 13–52.

[57] J.M. Ball, F. Murat. *$W^{1,p}$-quasiconvexity and variational problems for multiple integrals.* J. Funct. Anal. **58** (1984), 225–253.

[58] G. I. Barenblatt. *The mathematical theory of equilibrium cracks in brittle fracture.* Adv. Appl. Mech. **7** (1962), 55.

[59] E. Barozzi, E. H. A. Gonzalez. *Least area problems with a volume constraint.* Variational methods for equilibrium problems of fluids, Astérisque **118** (1984), 33–53.

[60] G. Beer. *Lipschitz regularization and the convergence of convex functions.* Numer. Funct. Anal. Optim. **15** (1994), no. 1–2, 31–46.

[61] G. Bellettini, A. Coscia. *Discrete approximation of a free discontinuity problem.* Numer. Funct. Anal. Optim. **15** (1994), no. 3-4 , 201–224.

[62] M. Bellieud, G. Bouchitté. *Homogenization of elliptic problems in a fiber reinforced structure. Nonlocal effect.* Ann. Scuola Norm. Sup. Pisa Cl. Sci. (4) **26** (1998), no. 3, 407–436.

[63] M. Bellieud, G. Bouchitté. *Homogenization of a soft elastic material reinforced by fibers.* Asymptot. Anal. **32** (2002), no. 2, 153–183.

[64] M. Belloni, G. Buttazzo, L. Freddi. *Completion by Gamma-convergence for optimal control problems.* Ann. Fac. Sci. Toulouse Math. **2** (1993), 149–162.

[65] A. Beurling, J. Deny. *Dirichlet spaces.* Proc. Nat. Acad. Sci. USA **45** (1959), 208–215.

[66] K. Bhattacharya. *Wedge-like microstructure in martensite.* Acta Metal. **39** (1991), 2431–2444.

[67] K. Bhattacharya. *Self accommodation in martensite.* Arch. Ration. Mech. Anal. **120** (1992), 201–244.

[68] K. Bhattacharya, N. Firoozye, R. D. James, R. V. Kohn. *Restrictions on microstructures.* Proc. Roy. Soc. Edinburgh **124A** (1994), 843–878.

[69] K. Bhattacharya, R. V. Kohn. *Elastic energy minimization and the recoverable strains of polycrystalline shape-memory materials.* Arch. Ration. Mech. Anal. **139** (1997), 99–180.

[70] E. Bishop, R. Phelps. *The support functional of a convex set.* Proc. Symp. Pure Math. Amer. Math. Soc. **7** (1962), 27–35.

[71] J.M. Borwein, A. Lewis. *Convex analysis and nonlinear optimization.* Canadian Mathematical Society Books in Mathematics, Springer-Verlag, New York, 2000.

[72] J.M. Borwein, A. Lewis, D. Noll. *Maximum entropy spectral analysis using first order information. Part I: Fisher information and convex duality.* Math. Oper. Res. **21** (1996), 442–468.

[73] J.M. Borwein, A. Lewis, R. Nussbaum. *Entropy minimization, DAD problems and doubly stochastic kernels.* J. Funct. Anal. **123** (1994), 264–307.

[74] G. Bouchitté, G. Buttazzo. *New lower semicontinuity results for non convex functionals defined on measures.* Nonlinear Anal. **15** (1990), 679–692.

[75] G. Bouchitté, G. Buttazzo. *Integral representation of nonconvex functionals defined on measures.* Ann. Inst. H. Poincaré Anal. Non Linéaire **9** (1992), 101–117.

[76] G. Bouchitté, G. Buttazzo. *Relaxation for a class of nonconvex functionals defined on measures.* Ann. Inst. H. Poincaré Anal. Non Linéaire **10** (1993), 345–361.

[77] G. Bouchitté, G. Buttazzo, P. Seppecher. *Energies with respect to a measure and applications to low dimensional structures.* Calc. Var. Partial Differential Equations **5** (1997), no. 1, 37–54.

[78] G. Bouchitté, P. Suquet. *Equi-coercivity of variational problems: The role of recession functions.* Proceedings of the Séminaire du Collège de France, Vol XII, (Paris, 1991–1993), 31–54, Pitman Res. Notes Math. Ser., 302.

[79] Bourbaki. *Elements de mathématiques–Espaces vectoriels topologiques,* Act. Sci. Ind., 1189, Hermann, Paris 1966.

[80] B. Bourdin, G.A. Francfort, J.-J. Marigo. *Numerical experiments in revisited brittle fracture.* J. Mech. Phys. Solids **48** (2000), no. 4, 797–826.

[81] A. Bourgeat, A. Mikelic, S. Wright. *Stochastic two-scale convergence in the mean and applications.* J. Reine Angew. Math. **456** (1994), 19–51.

[82] A. Braides. *Approximation of free-discontinuity problems.* Lecture Notes in Math. **1694**, Springer-Verlag, New York, 1998.

[83] A. Braides. Γ*-convergence for beginners.* Oxford Lecture Ser. Math. Appl. **22**, Oxford University Press, Oxford, UK, 2002.

[84] A. Braides. *Discrete approximation of functionals with jumps and creases.* Homogenization, 2001 (Naples), Gakuto Internat. Ser. Math. Sci. Appl. **18**, Gakkōtosho, Tokyo, 2003, 147–153.

[85] A. Braides, A. Coscia. *A singular perturbation approach to problems in fracture mechanics.* Math. Mod. Meth. Appl. Sci. **3** (1993), 302–340.

[86] A. Braides, G. Dal Maso. *Non-local approximation of the Mumford-Shah functional.* Calc. Var. Partial Differential Equations **5** (1997), 293–322.

[87] A. Braides, G. Dal Maso, A. Garroni. *Variational formulation of softening phenomena in fracture mechanics: The one dimensional case.* Arch. Ration. Mech. Anal. **146** (1999), 23–58.

[88] A. Braides, M.S. Gelli. *Continuum limits of discrete systems without convexity hypotheses.* Math. Mech. Solids **7** (2002), 41–66.

[89] A. Braides, M. S. Gelli. *Limits of discrete systems with long-range interactions.* J. Convex Anal. **9** (2002), no. 2, 363–399.

[90] H. Brezis. *Analyse Fonctionnelle, Théorie et Applications.* Masson, Paris, 1983.

[91] H. Brezis. *Intégrales convexes dans les espaces de Sobolev.* Israel J. Math. **13** (1972), 9–23.

[92] H. Brezis, F.E. Browder. *A general principle on ordered sets in nonlinear functional analysis.* Adv. Math., **21** (1976), no. 3, 353–364.

[93] H. Brezis, L. Nirenberg. *Characterizations of the ranges of some nonlinear operators and applications to boundary value problems.* Ann. Scuola Norm. Sup. Pisa Cl. Sci. (4) **5** (1978), 225–326.

[94] H. Buchwalter. *Variations sur l'Analyse.* Ellipse, Paris, 1992.

[95] D. Bucur, G. Buttazzo. *Variational Methods in Some Shape Optimization Problems.* Lecture notes, Dipartimento di Matematica Università di Pisa and Scuola Normale Superiore di Pisa, Series Appunti di Corsi della Scuola Normale Superiore (2002).

[96] D. Bucur, G. Buttazzo, I. Figueiredo. *On the attainable eigenvalues of the Laplace operator.* SIAM J. Math. Anal. **30** (1999), 527–536.

[97] G. Buttazzo. *Semicontinuity, relaxation and integral representation in the calculus of variations.* Pitman Res. Notes Math. Ser. **207**, Longman, Harlow, UK, 1989.

[98] G. Buttazzo, G. Dal Maso. *Shape optimization for Dirichlet problems: Relaxed solutions and optimality conditions.* Bull. Amer. Math. Soc. **23** (1990), 531–535.

[99] G. Buttazzo, G. Dal Maso. *Shape optimization for Dirichlet problems: Relaxed formulation and optimality conditions.* Appl. Math. Optim. **23** (1991), 17–49.

[100] G. Buttazzo, G. Dal Maso. *An existence result for a class of shape optimization problems.* Arch. Ration. Mech. Anal. **122** (1993), 183–195.

[101] G. Buttazzo, L. Freddi. *Relaxed optimal control problems and applications to shape optimization.* Lecture Notes, NATO-ASI Summer School Nonlinear Analysis, Differential Equations and Control, Montreal, Kluwer, Dordrecht, The Netherlands 1999, 159–206.

[102] G. Buttazzo, M. Giaquinta, S. Hildebrandt. *One-dimensional variational problems. An introduction.* Oxford Lecture Ser. Math. Appl. 15, Oxford University Press, New York, 1998.

[103] G. Buttazzo, P. Guasoni. *Shape optimization problems over classes of convex domains.* J. Convex Anal. **4** (1997), 343–351.

[104] C. Castaing, P. Raynaud de Fitte, M. Valadier. *Young measure on topological spaces. With applications in control theory and probability theory.* Math. Appl. 571, Kluwer Academic Publishers, Dordrecht, Netherlands, 2004.

[105] C. Castaing, M. Valadier. *Convex analysis and measurable multifunctions.* Lecture Notes in Math. **590**, Springer-Verlag, Berlin, 1977.

[106] E. Chabi, G. Michaille. *Ergodic theory and application to nonconvex homogenization.* Set Valued Anal. **2** (1994), 117–134.

[107] A. Chambolle. *Image segmentation by variational methods: Mumford and Shah functional and the discrete approximations.* SIAM J. Appl. Math. **55** (1995), 827–863.

[108] M. Chipot, C. Collins, D. Kinderlehrer. *Numerical analysis of oscillations in multiple well problems.* Numer. Math. **70** (1995), 259–282.

[109] M. Chipot, G. Dal Maso. *Relaxed shape optimization: The case of nonnegative data for the Dirichlet problem.* Adv. Math. Sci. Appl. **1** (1992), 47–81.

[110] M. Chipot, D. Kinderlehrer. *Equilibrium configurations of crystals.* Arch. Ration. Mech. Anal. **103** (1988), 237–277.

[111] Ph. Ciarlet. *The finite element method for elliptic problems.* North-Holland, Amsterdam, 1978.

[112] F. H. Clarke. *Optimization and nonsmooth analysis.* John Wiley, New York, 1983.

[113] C. Combari, L. Thibault. *On the graph convergence of subdifferentials of convex functions.* Proc. Amer. Math. Soc. **126** (1998), 2231–2240.

[114] G. Cortesani, R. Toader. *Nonlocal approximation of nonisotropic free-discontinuity problems.* SIAM J. Appl. Math. **59** (1999), 1507–1519.

[115] R. Courant, D. Hilbert. *Methoden der Mathematischen Physik.* Berlin, Vol I (1931), Vol II (1937). English ed. Interscience, New York, Vol I (1953), Vol II (1962).

[116] B. Dacorogna. *Direct methods in the calculus of variations.* Appl. Math. Sci. **78**, Springer-Verlag, Berlin, 1989.

[117] G. Dal Maso. *An introduction to* Γ*-convergence.* Birkhäuser, Boston, 1993.

[118] G. Dal Maso, G. Franfort, R. Toader. *Quasi-static evolution in brittle fracture: the case of bounded solutions*. Calculus of Variations: Topics from the Mathematical Heritage of E. De Giorgi, Quad. Mat., 14, Dept. Math., Seconda Univ. Napoli, Caserts, 2004, 245–266. Also available on the Web at `http://cvgmt.sns.it/cgi/get.cgi/papers/dalfratoa04a/`.

[119] G. Dal Maso, L. Modica. *Nonlinear stochastic homogenization and ergodic theory.* J. Reine Angew. Math. **363** (1986), 27–43.

[120] R. Dautray, J. L. Lions. *Analyse mathématique et calcul numérique pour les sciences et les techniques.* Masson, Paris, 1984.

[121] D. G. De Figueiredo. *Lectures at the Tata Institute of Fundamental Research.* Springer-Verlag, New York, 1989.

[122] E. De Giorgi, L. Ambrosio. *Un nuovo tipo di funzionale del calcolo delle variazioni.* Atti Accad. Naz. Lincei Rend. Cl. Sci. Fis. Mat. Natur. **82** (1988), 199–210.

[123] E. De Giorgi, M. Carriero, A. Leaci. *Existence theorem for a minimum problem with free discontinuity set.* Arch. Ration. Mech. Anal. **108** (1989), 195–218.

[124] E. De Giorgi, T. Franzoni. *Su un tipo di convergenza variazionale.* Atti Accad. Naz. Lincei Rend. Cl. Sci. Fis. Mat. Natur. (8) **58** (1975), 842–850.

[125] C. Dellacherie, P.A. Meyer. *Probabilité et Potentiel.* Hermann, Paris, 1975.

[126] F. Demengel, R. Temam. *Convex functions of measures and applications.* Indiana Univ. Math. J. **33** (1984), 673–709.

[127] N. Dunford, J.T. Schwartz. *Linear operators.* Interscience, New York, 1958.

[128] I. Ekeland. *Remarques sur les problèmes variationnels*. CRAS, t. 275, 1057–1059, 1972.

[129] I. Ekeland. *Remarques sur les problèmes variationnels II*. CRAS, t. 276, 1347–1349, 1973.

[130] I. Ekeland. *On the variational principle.* J. Math. Anal. Appl. **47** (1974), 324–353.

[131] I. Ekeland, R.Temam. *Convex analysis and variational problems.* North-Holland, Amsterdam, 1978.

[132] L.C. Evans, R.F. Gariepy. *Measure theory and fine properties of functions.* Stud. Adv. Math., CRC Press, Boca Raton, FL, 1992.

[133] K.F. Falconer. *The geometry of fractal sets.* Cambridge University Press, Cambridge, UK, 1986.

[134] H. Federer. *Geometric measure theory.* Springer-Verlag, Berlin, 1969.

[135] G. Fichera. *Problemi elastostatici con vincoli unilaterali: il problema di signorini con ambigue condizioni al contorno*. Atti Accad. Naz. Lincei Mem. Cl. Sci. Fis. Mat. Natur. Sez. I (8) **7**, 1963/1964, 91–140.

[136] W. Fleming, R. Rishel. *An integral formula for total gradient variation.* Arch. Math. **11** (1960), 218–222.

[137] I. Fonseca, S. Müller, P. Pedregal. *Analysis of concentration and oscillation effects generated by gradients.* SIAM J. Math. Anal. **29** (1998), no. 3, 736–756.

[138] G. A. Francfort, J.-J. Marigo. *Une approche variationnelle de la mécanique du défaut.* Actes du 30ème Congrès d'Analyse Numérique: CANum'98 (Arles 1998), 57–74, ESAIM Proc. **6**, Soc. Math. Appl. Indust., Paris, 1999.

[139] G. A. Francfort, J.-J. Marigo. *Cracks in fracture mechanics: A time indexed family of energy minimizers.* Variations of domain and free-boundary problems in solid mechanics (Paris, 1997), 197–202, Solid Mech. Appl. **66**, Kluwer Academic Publishers, Dordrecht, The Netherlands, 1999.

[140] K. Friedrichs. *Spektraltheorie halbbeschränkter Operatorem*, I, II. Math. Ann. **109** (1934), 465–487; 685–713.

[141] F. Gastaldi, F. Tomarelli. *Some remarks on nonlinear and non-coercive variational inequalities.* Boll. U.M.I. (7), **1**-B (1987) 143–165.

[142] G. Geymonat. *Introduction à la localisation.* Cours LMGC, Université Montpellier II, Mars-Avril 2003.

[143] M. Giaquinta, S. Hildebrandt. *Calculus of variations, I, II.* Springer-Verlag, Berlin, 1996.

[144] E. Giusti. *Minimal surfaces and functions of bounded variation.* Birkhäuser, Boston, 1984.

[145] D. Gilbarg, N. Trudinger. *Elliptic partial differential equations of second order.* Springer-Verlag, Berlin, 1977.

[146] R. Glowinski. *Numerical methods for nonlinear variational problems.* Springer-Verlag, New York, 1984.

[147] R. Glowinski, J. L. Lions, R. Trémolière. *Numerical analysis of variational inequalities.* North-Holland, Amsterdam, 1981.

[148] M. Gobbino. *Finite difference approximation of the Mumford-Shah functional.* Comm. Pure Appl. Math. **51** (1998), 197–228.

[149] A. A. Griffith. *The phenomenon of rupture and flow in solids.* Phil Trans. Royal Soc. London A **221** (1920), 163–198.

[150] P. Grisvard. *Problèmes aux limites dans des domaines avec points de rebroussement.* Ann. Fac. Sci. Toulouse Math. **6** (1995), no. 3, 561–578.

[151] P. Grisvard. *Singularities in boundary value problems.* Rech. Math. Appl. **22**, Masson, Paris, Springer-Verlag, Berlin, 1992.

[152] M. E. Gurtin. *On phase transitions with bulk, interfacial, and boundary energy.* Arch. Ration. Mech. Anal. **96** (1986), 243–264.

[153] A. Henrot, M. Pierre. *Variation et optimisation de formes. Une analyse géométrique.* Mathematiques et Applications **48**, Springer, New York, 2005.

[154] D. Hilbert. *Mathematical Problems*. Lecture delivered before the First International Congress of Mathematicians at Paris 1900. Bull. Amer. Math. Soc. **8**, 437–479. (Reprinted in Browder, F., ed. 1976, Vol. 1, 1–34.)

[155] J.B. Hiriart-Urruty, C. Lemaréchal. *Convex analysis and mininimization algorithms* I, II. Springer-Verlag, New York, 1993.

[156] S.I. Huhjaev, A.I. Vol'pert. *Analysis in classes of discontinuous functions and equations of mathematical physics.* Martinus Nijhoff, Dordrecht, The Netherlands, 1985.

[157] A. D. Ioffe. *On lower semicontinuity of integral functionals I. II.* SIAM J. Control Optim. **15** (1977), 521–538 and 991–1000.

[158] A. D. Ioffe, V. M. Tihomirov. *Theory of extremal problems.* North-Holland, Amsterdam, 1979.

[159] O. Iosifescu, C. Licht, G. Michaille. *Variational limit of a one dimensional discrete and statistically homogeneous system of material points.* Asymptot. Anal. **28** (2001), 309–329.

[160] O. Iosifescu, C. Licht, G. Michaille. *Variational limit of a one dimensional discrete and statistically homogeneous system of material points.* CRAS, t. 322, Série I, 575–580, 2001.

[161] J. L. Joly. *Une famille de topologies sur l'ensemble des fonctions convexes pour lesquelles la polarité est bicontinue.* J. Math. Pures Appl. **52** (1973), 421–441.

[162] D. Kinderlehrer, P. Pedregal. *Characterization of Young measures generated by gradients.* Arch. Ration. Mech. Anal. **119** (1991), 329–365.

[163] D. Kinderlherer, G. Stampacchia. *An introduction to variational inequalities and their applications.* Classics Appl. Math. **31**, SIAM, Philadelphia, 2000.

[164] R. V. Kohn, G. Strang. *Optimal design and relaxation of variational problems*, I,II,III. Comm. Pure Appl. Math. **39** (1986), 113–137, 139–182, 353–377.

[165] U. Krengel. *Ergodic theorems.* Studies in Mathematics, Walter de Gruyter, Berlin, New York, 1985.

[166] T. Lachand-Robert, M.A. Peletier. *An example of non-convex minimization and an application to Newton's problem of the body of least resistance.* Ann. Inst. H. Poincaré Anal. Non Linéaire **18** (2001), no. 2, 179–198.

[167] B. Larrouturou, P. L. Lions. *Méthodes Mathématiques pour les Sciences de l'ingénieur, Optimisation et Analyse Numérique.* Ecole Polytechnique, Palaiseau, 1996.

[168] H. Le Dret, A. Raoult. *The nonlinear membrane model as variational limit in nonlinear three-dimensional elasticity.* J. Math. Pures Appl., (9) **74** (1995), no. 6, 549–578.

[169] L. Leghmizi, C. Licht, G. Michaille. *The nonlinear membrane model: A Young measure and varifold formulation.* ESAIM: COCV **11** (2005), 449–472.

[170] C. Licht, G. Michaille. *Global-local subadditive ergodic theorems and application to homogenization in elasticity.* Ann. Math. Blaise Pascal **9** (2002), 21–62.

[171] E.H. Lieb. *Sharp constants in the Hardy-Littelwood-Sobolev and related inequalities.* Ann. Math. **118** (1983), no. 2, 349–374.

[172] J.L. Lions. *Quelques méthode de résolution des problèmes aux limites non linéaires.* Dunod, Paris, 1969.

[173] J.L. Lions, G. Stampacchia. *Variational inequalities.* Comm. Pure Appl. Math. **20** (1967), 493–519.

[174] S. Luckhaus, L. Modica. *The Gibbs-Thompson relation within the gradient theory of phase transitions.* Arch. Ration. Mech. Anal. **107** (1989), 71–83.

[175] J.-P. Mandallena. *On the relaxation of nonconvex superficial integral functionals.* J. Math. Pures Appl. **79** (2000), 1011–1028.

[176] J.-P. Mandallena. *Quasiconvexification of geometric integrals.* Ann. Mat. Pura Appl. (to appear).

[177] P. Marcellini. *Approximation of quasiconvex functions, and lower semicontinuity of multiple integrals.* Manuscripta Math. **51** (1985), 1–28.

[178] C.M. Marle. *Mesure et probabilités.* Hermann, Paris, 1974.

[179] K. Messaoudi, G. Michaille. *Stochastic homogenization of nonconvex integral functionals.* Math. Modelling Numer. Anal. **28** (1994), no. 3, 329–356.

[180] L. Modica. *The gradient theory of phase transitions and the minimal interface criterion.* Arch. Ration. Mech. Anal. **98** (1987), 123–142.

[181] B. Mohammadi, J.-H. Saïac. *Pratique de la Simulation Numérique.* Dunod, Paris, 2003.

[182] J.J. Moreau. *Fonctionnelles convexes.* Cours Collège de France 1967, new edition CNR, Facoltà di Ingegneria di Roma, Roma, 2003.

[183] J.M. Morel, S. Solimini. *Variational models in image segmentation.* Birkhäuser Boston, Boston, 1995.

[184] F. Morgan. *Geometric measure theory: A beginners guide.* Academic Press, New York, 1988.

[185] C. B. Morrey. *Multiple integrals in the calculus of variations.* Springer-Verlag, Berlin, 1966.

[186] C.B. Morrey. *Quasiconvexity and the semicontinuity of multiple integrals.* Pacific J. Math. **2** (1952), 25–53.

[187] U. Mosco. *Convergence of convex sets and of solutions of variational inequalities.* Adv. Math. **3** (1969), 510–585.

[188] U. Mosco. *On the continuity of the Young-Fenchel transformation.* J. Math. Anal. Appl. **35** (1971), 518–535.

[189] J. Moser. *A Harnack inequality for paraboloc differential equations.* Comm. Pure Appl. Math. **17** (1964), 101–134.

[190] S. Müller, V. Sverak. *Attainment results for the two-well problem by convex integration.* Geometric Analalysis and the Calculus of Variations, International Press, Cambridge, MA (1996), 239–251.

[191] D. Mumford, J. Shah. *Optimal approximation by piecewise smooth functions and associated variational problem.* Comm. Pure Appl. Math. **17** (1989), 577–685.

[192] F. Murat, L. Tartar. *Optimality conditions and homogenization.* Proceedings of Nonlinear Variational Problems, Isola d'Elba 1983, Res. Notes Math. **127**, Pitman, London (1985), 1–8.

[193] J. Nečas. *Les méthodes directes en théorie des équations elliptiques.* Masson, Paris, 1967.

[194] X.X. Nguyen, H. Zessin. *Ergodic theorems for spatial processes.* Z. Wah. Vew. Gebiette **48** (1979), 133–158.

[195] P. Pedregal. *Parametrized measures and variational principle.* Birkhäuser Boston, Boston, 1997.

[196] P. Pedregal. *Laminates and microstructure.* European J. Appl. Math. **4** (1993), 121–149.

[197] R.R. Phelps. *Convex functions, monotone operators and differentiability.* Lecture Notes in Math. **1364**, Springer-Verlag, New York, 1989.

[198] O. Pironneau. *Optimal shape design for elliptic systems.* Springer Ser. Comput. Phys., Springer-Verlag, New York, 1984.

[199] P.A. Raviart, J.M. Thomas. *Introduction à l'analyse numérique des équations aux dérivées partielles.* Masson, Paris, 1983.

[200] J. R. Rice. *Mathematical analysis in the mechanics of fracture.* Fracture: An Advanced Treatise, H. Liebowitz ed., **2** (1969), 191–311.

[201] R.T. Rockafellar. *Convex analysis.* Princeton University Press, Princeton, NJ, 1970.

[202] R.T. Rockafellar. *Integrals which are convex functionals.* Pacific J. Math. **24** (1968), no. 3, 525–539.

[203] R.T. Rockafellar. *Convex integral functionals, II.* Pacific J. Math. **39** (1971), 439–469.

[204] R.T. Rockafellar. *Conjugate convex functions in optimal control and the calculus of variations.* J. Math. Appl. **32** (1970), 174–222.

[205] R.T. Rockafellar, R.J.B. Wets. *Variational analysis.* Springer-Verlag, Berlin, Heidelberg, New York, 1998.

[206] W. Rudin. *Analyse réelle et complexe.* Masson, Paris, 1975.

[207] H. A. Simon. *Bounded rationality in social science: Today and tomorrow.* Mind Society, **1** (2000), 25–39.

[208] S. Sobolev. *Methode nouvelle a résoudre le probléme de Cauchy pour les équations linéaires hyperboliques normales.* Mat. Sbornik **1** (1936), 39–72.

[209] J. Sokolowski, J.-P. Zolesio. *Introduction to shape optimization. Shape sensitivity analysis.* Springer Series in Comput. Math. **16**, Springer-Verlag, Berlin (1992).

[210] G. Stampacchia. *Formes bilinéaires coercitives sur les ensembles convexes.* C. R. Acad. Sci. Paris, **258** (1964), 4413–4416.

[211] V. Sverak. *New examples of quasiconvex functions.* Arch. Ration. Mech. Anal. **119** (1992), 293–300.

[212] V. Sverak. *Quasiconvex functions with subquadratic growth.* Proc. Roy. Soc. Lond. **A433** (1991), 723–725.

[213] V. Sverak. *Rank-one convexity does not imply quasiconvexity.* Proc. Roy. Soc. Edinburgh **120A** (1992), 185–189.

[214] V. Sverak. *On the problem of two well.* Microstructure and Phase Transition, D. Kinderlehrer et al., eds., IMA Vol. Math. Appl. **54**, Springer-Verlag, New York (1993), 183–190.

[215] M.A. Sychev. *A new approach to Young measure theory, relaxation and convergence in energy.* Ann. Inst. Henri Poincaré **16** (1999), no. 6, 773–812.

[216] G. Talenti. *Best constants in Sobolev inequality.* Ann. Mat. Pura Appl. **110** (1976), 353–372.

[217] L. Tartar. *H-measures, a new approach for studying homogenization, oscillations and concentration effects in partial differential equations.* Proc. Roy. Soc. Edinburgh **115A** (1990), 193–230.

[218] L. Tartar. *An introduction to the homogenization method in optimal design.* Optimal Shape Design. Lecture Notes in Math. **1740**, Springer-Verlag, Berlin (2000), 47–156.

[219] R. Temam. *Problèmes Mathématiques en Plasticité.* Gauthier-Villars, Paris, 1983.

[220] L. Thibault. *Sequential convex subdifferential calculus and sequential Lagrange multipliers.* SIAM J. Control Optim. **35** (1997), 1434–1444.

[221] D. Torralba. *Applications aux transitions de phases et à la méthode barrière logarithmique.* Thèse de l'Université Montpellier II, 1996.

[222] N.S. Trudinger. *On Harnack type inequalities and their applications to quasilinear elliptic equations.* Comm. Pure Appl. Math. **20** (1967), 721–747.

[223] M. Valadier. *Young measures.* Methods of Nonconvex Analysis, A. Cellina ed. Lecture Notes in Math. **1446**, Springer-Verlag, Berlin (1990), 152–188.

[224] M. Valadier. *A course on Young measures.* Rend. Instit. Mat. Univ. Trieste **26** suppl. (1994), 349–394.

[225] K. Yosida. *Functional analysis.* Springer-Verlag, New York, 1971.

[226] L. C. Young. *Lectures on calculus of variations and optimal control theory.* W. B. Saunders, Philadelphia, 1969.

[227] E. Zeidler. *Nonlinear functional analysis and its applications* III. Variational Methods and Optimization, Springer-Verlag, Berlin, 1984.

[228] K. Zhang. *Rank-one connections at infinity and quasiconvex hulls.* J. Convex Anal. **7** (2000), no. 1, 19–45.

[229] W. P. Ziemer. *Weakly differentiable functions.* Springer-Verlag, Berlin, 1989.

Index