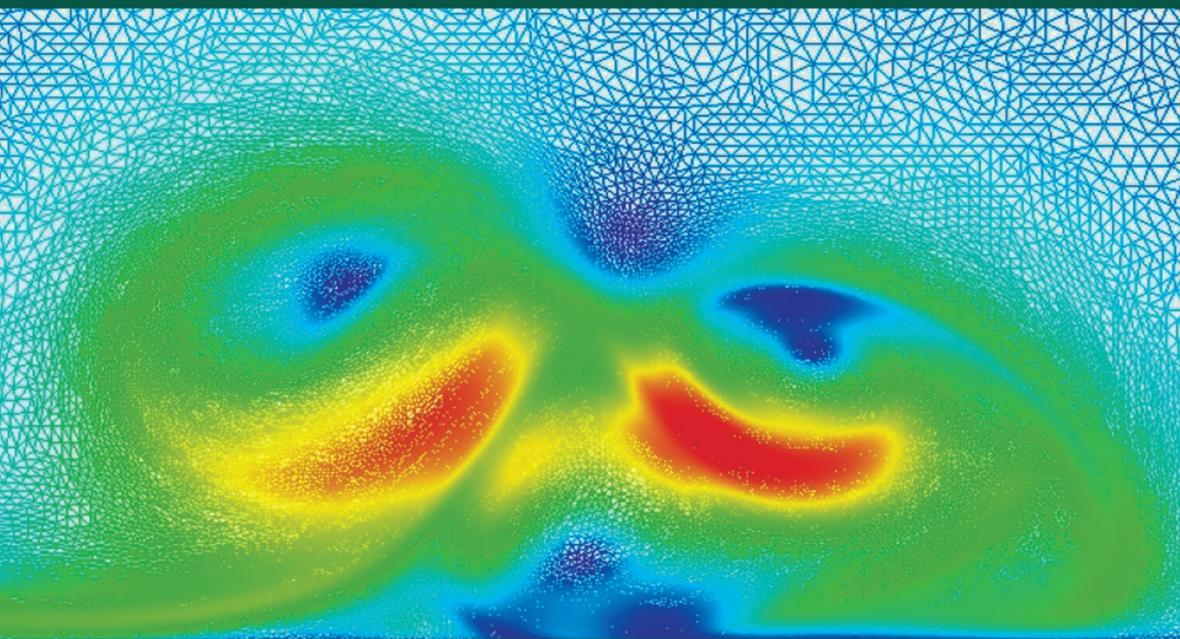


**MATHEMATICS AND STATISTICS**



**Mathematics  
for Modeling and  
Scientific Computing**

**Thierry Goudon**

**ISTE**

**WILEY**



**Mathematics for Modeling and Scientific Computing**



*Series Editor*  
*Jacques Blum*

---

# **Mathematics for Modeling and Scientific Computing**

---

Thierry Goudon

**ISTE**

**WILEY**

First published 2016 in Great Britain and the United States by ISTE Ltd and John Wiley & Sons, Inc.

Apart from any fair dealing for the purposes of research or private study, or criticism or review, as permitted under the Copyright, Designs and Patents Act 1988, this publication may only be reproduced, stored or transmitted, in any form or by any means, with the prior permission in writing of the publishers, or in the case of reprographic reproduction in accordance with the terms and licenses issued by the CLA. Enquiries concerning reproduction outside these terms should be sent to the publishers at the undermentioned address:

ISTE Ltd  
27-37 St George's Road  
London SW19 4EU  
UK

[www.iste.co.uk](http://www.iste.co.uk)

John Wiley & Sons, Inc.  
111 River Street  
Hoboken, NJ 07030  
USA

[www.wiley.com](http://www.wiley.com)

© ISTE Ltd 2016

The rights of Thierry Goudon to be identified as the author of this work have been asserted by him in accordance with the Copyright, Designs and Patents Act 1988.

Library of Congress Control Number: 2016949851

---

British Library Cataloguing-in-Publication Data  
A CIP record for this book is available from the British Library  
ISBN 978-1-84821-988-5

---

---

# Contents

---

<b>Preface</b>	ix
<b>Chapter 1. Ordinary Differential Equations</b>	1
1.1. Introduction to the theory of ordinary differential equations	1
1.1.1. Existence–uniqueness of first-order ordinary differential equations	1
1.1.2. The concept of maximal solution	11
1.1.3. Linear systems with constant coefficients	16
1.1.4. Higher-order differential equations	20
1.1.5. Inverse function theorem and implicit function theorem	21
1.2. Numerical simulation of ordinary differential equations, Euler schemes, notions of convergence, consistence and stability	27
1.2.1. Introduction	27
1.2.2. Fundamental notions for the analysis of numerical ODE methods	29
1.2.3. Analysis of explicit and implicit Euler schemes	33
1.2.4. Higher-order schemes	50
1.2.5. Leslie’s equation (Perron–Frobenius theorem, power method)	51
1.2.6. Modeling red blood cell agglomeration	78
1.2.7. SEI model	87
1.2.8. A chemotaxis problem	93
1.3. Hamiltonian problems	102
1.3.1. The pendulum problem	106
1.3.2. Symplectic matrices; symplectic schemes	112
1.3.3. Kepler problem	125
1.3.4. Numerical results	129

<b>Chapter 2. Numerical Simulation of Stationary Partial Differential Equations: Elliptic Problems</b>	141
2.1. Introduction	141
2.1.1. The 1D model problem; elements of modeling and analysis	144
2.1.2. A radiative transfer problem	155
2.1.3. Analysis elements for multidimensional problems	163
2.2. Finite difference approximations to elliptic equations	166
2.2.1. Finite difference discretization principles	166
2.2.2. Analysis of the discrete problem	173
2.3. Finite volume approximation of elliptic equations	180
2.3.1. Discretization principles for finite volumes	180
2.3.2. Discontinuous coefficients	187
2.3.3. Multidimensional problems	189
2.4. Finite element approximations of elliptic equations	191
2.4.1. $\mathbb{P}_1$ approximation in one dimension	191
2.4.2. $\mathbb{P}_2$ approximations in one dimension	197
2.4.3. Finite element methods, extension to higher dimensions	200
2.5. Numerical comparison of FD, FV and FE methods	204
2.6. Spectral methods	205
2.7. Poisson–Boltzmann equation; minimization of a convex function, gradient descent algorithm	217
2.8. Neumann conditions: the optimization perspective	224
2.9. Charge distribution on a cord	228
2.10. Stokes problem	235
<b>Chapter 3. Numerical Simulations of Partial Differential Equations: Time-dependent Problems</b>	267
3.1. Diffusion equations	267
3.1.1. $L^2$ stability (von Neumann analysis) and $L^\infty$ stability: convergence	269
3.1.2. Implicit schemes	276
3.1.3. Finite element discretization	281
3.1.4. Numerical illustrations	283
3.2. From transport equations towards conservation laws	291
3.2.1. Introduction	291
3.2.2. Transport equation: method of characteristics	295
3.2.3. Upwinding principles: upwind scheme	299
3.2.4. Linear transport at constant speed; analysis of FD and FV schemes	301

3.2.5. Two-dimensional simulations . . . . .	326
3.2.6. The dynamics of prion proliferation . . . . .	329
3.3. Wave equation . . . . .	345
3.4. Nonlinear problems: conservation laws . . . . .	354
3.4.1. Scalar conservation laws . . . . .	354
3.4.2. Systems of conservation laws . . . . .	387
3.4.3. Kinetic schemes . . . . .	393
<b>Appendices</b> . . . . .	407
<b>Appendix 1</b> . . . . .	409
<b>Appendix 2</b> . . . . .	417
<b>Appendix 3</b> . . . . .	427
<b>Appendix 4</b> . . . . .	433
<b>Appendix 5</b> . . . . .	443
<b>Bibliography</b> . . . . .	447
<b>Index</b> . . . . .	455



---

## Preface

---

*Early he rose, far into the night he would wait, To count,  
to cast up, and to calculate, Casting up, counting,  
calculating still, For new mistakes for ever met his view.*

*Jean de La Fontaine*

(The Money-Hoarder and Monkey, Book XII, Fable 3).

This book was inspired by a collection of courses of varied natures and at different levels, all of which focused on different aspects of scientific computing. Therefore, it owes much to the motivation of students from the universities of Nice and Lille, and the Ecole Normale Supérieure. The writing style adopted in this book is based on my experience as a longtime member of the jury for the *agrégation* evaluations, particularly in the modeling examination. In fact, a substantial part of the examples on the implementation of numerical techniques was drawn directly from the texts made public by the evaluation's jury (see <http://agreg.org>), and a part of this course was the foundation for a series of lectures given to students preparing for the Moroccan *agrégation* evaluations. However, some themes explored in this book go well beyond the scope of the evaluations. They include, for example, the rather advanced development of Hamiltonian problems, the fine-grained distinction between the finite-difference method and the finite-volume method, and the discussion of nonlinear hyperbolic problems. The latter topic partially follows a course given at the IFCAM (Indo-French Centre for Applied Mathematics) in Bangalore. A relatively sophisticated set of tools is developed on this topic: this heightened level can be explained by the fact that the questions it explores are of great practical importance. It provides a relevant introduction for those who might want to learn more, and it prepares them for reading more advanced and specialized works.

Numerical analysis and scientific calculus courses are often considered a little scary. This fear is often due to the fact that the subject can be difficult on several counts:

– problems of interest are very strongly motivated by their potential applications (for example, in physics, biology, engineering, finance). Therefore, it is impossible to restrict the discussion strictly to the field of mathematics, and the intuitions motivating the math are strongly guided by the specificities of its applications. As a result, the subject requires a certain degree of scientific familiarity that goes beyond mere technical dexterity.

– Numerical analysis uses a very ample technical background. This is a subject that we cannot address by utilizing a small and previously delimited set of tools. Rather, we must draw from different areas of mathematics<sup>1</sup>, sometimes in rather unexpected ways, for example by using linear algebra to analyze the behavior of numerical approximations to differential equations. However, this somewhat roundabout way of finding answers is what makes the subject so exciting.

– Finally, it is often difficult to produce categorical statements and conclusions. For example, although it can be shown that several numerical schemes produce an approximate solution that “converges” towards the solution of the problem of interest (when the numerical parameters are sufficiently small), in practice, some methods are more suitable than others, according to qualitative criteria that are not always easy to formalize. Similarly, the choice of method may depend on the criteria that are considered most important for the target application context. Many questions do not have definite, clear-cut answers. The answer to the question “how should we approach this problem?” is often “it depends”: numerically simulating a physical phenomenon through calculations performed by a computer is a real, delicate and nuanced art. This art must be based on strong technical mastery of mathematical tools and deep understanding of the underlying physical phenomena they study.

The aim of this book is to fully address these challenges and, by design, to “mix everything up”. Therefore, the book will include many classical results from analysis and algebra, details for certain equation resolution algorithms, examples from science and technology, and numerical illustrations. Some “theoretical” tools will be introduced by studying an application example, even if it means repurposing it for an entirely different field. Nevertheless, the book does follow a certain structure, which is organized into three main sections, focused on numerical solutions to (ordinary and partial) differential equations. The first chapter addresses the solution of ordinary differential equations, with a very broad overview of its essential theoretical basis

---

<sup>1</sup> The following quote is quite telling: *[...] in France, there was even some snobbishness surrounding pure mathematics: when a gifted student was identified, he was told: “Do your PhD in pure mathematics”. On the other hand, average students were advised to focus on applied mathematics, under the rationale that that was “all they were capable of doing”! But the opposite is true: it is impossible to do applied mathematics without first knowing how to do pure mathematics properly.* J.A. Dieudonné, [SCH 90, p. 104].

(for example, Cauchy–Lipschitz, qualitative analysis, linear problems). This chapter details the analysis of classical schemes (explicit and implicit Euler methods), and distinguishes various concepts of stability, which are more or less relevant depending on the context. This set of concepts is illustrated by a series of examples, which are motivated mostly by the description of biological systems. A large section, with fairly substantial technical content, is devoted to the particular case of Hamiltonian systems. The second chapter deals with numerical solutions to elliptic boundary value problems, once again with a detailed exploration of basic functional analysis tools. Although the purpose is mostly restricted to the one-dimensional framework and to the model problem  $\lambda u(x) - \frac{\partial}{\partial x}(k(x)\frac{\partial}{\partial x}u(x)) = f(x)$  on  $]0, 1[$  with homogeneous Dirichlet conditions, different discretization families are distinguished: finite differences, finite volumes, finite element and spectral methods. Techniques related to optimization are also presented through the simulation of complex problems such as Boltzmann–Poisson equations, load optimization and Stokes’ problem. The last chapter deals with evolutionary partial differential equations, again addressing only the one-dimensional case. Questions of stability and consistency are addressed in this chapter, first for the heat equation and then for hyperbolic problems. The transport and wave equations can be considered “classics”. In contrast, discussion of nonlinear equations, scalars or systems with the simulations of Euler equations for gas dynamics as a final target, leads to more advanced topics. The book does not contain exercise problems. However, readers are invited to carry out the simulations illustrated in the book on their own. This work of numerical experimentation will allow readers to develop an intuition of the mathematical phenomena and notions it presents by playing with numerical and modeling parameters, which will lead to a complete understanding of the subject.

My colleagues and collaborators have had an important influence on building my personal mathematical outlook; they helped me discover points of view that I was not familiar with, and they have made me appreciate notions that I admit had remained beyond my understanding during my initial schooling and even during the early stages of my career. This is the place to thank them for their patience with me and for everything they have taught me. In particular, I am deeply indebted to Frédéric Poupaud and Michel Rascle, as well as Stella Krell and Magali Ribot in Nice, Caterina Calgaro and Emmanuel Creusé in Lille, Virginie Bonnaillie - Noël, Frédéric Coquel, Benoît Desjardins, Frédéric Lagoutière, and Pauline Lafitte in Paris. I also thank my colleagues on the *agrégation* jury, especially Florence Bachman, Guillaume Dujardin, Denis Favennec, Hervé Le Dret, Pascal Noble and Gregory Vial. A large number of developments were directly inspired by our passionate conversations. Finally, Claire Scheid, Franck Boyer and Sébastien Minjeaud were kind and patient enough to proofread some of the passages of the manuscript; their advice and suggestions have led to many improvements.

Thierry GOUDON  
August 2016



---

# Ordinary Differential Equations

---

## 1.1. Introduction to the theory of ordinary differential equations

### 1.1.1. Existence–uniqueness of first-order ordinary differential equations

The most important result from the theory of ordinary differential equations ensures the existence and uniqueness of solutions to equations of the form

$$y'(t) = f(t, y(t)), \quad y(t_{\text{Init}}) = y_{\text{Init}} \quad [1.1]$$

where

$$y_{\text{Init}} \in \Omega \subset \mathbb{R}^D, \quad f : I \times \Omega \longrightarrow \mathbb{R}^D$$

Here  $I$  is an open interval of  $\mathbb{R}$  containing  $t_{\text{Init}}$ , and  $\Omega$  is an open set of  $\mathbb{R}^D$  containing  $y_{\text{Init}}$ . The variable  $t \in I$  is called the *time variable*, and the variable  $y$  is referred to as the *state variable*. When the function  $f$  depends only on the state variable, equation [1.1] is said to be *autonomous*. We say that a function  $t \mapsto y(t)$  is a solution of [1.1] if

- $y$  is defined on an interval  $J$  that contains  $t_{\text{Init}}$  and is included in  $I$ ;
- $y(t_{\text{Init}}) = y_{\text{Init}}$  and for all  $t \in J$ ,  $y(t) \in \Omega$ ;
- $y$  is differentiable on  $J$  and for all  $t \in J$ ,  $y'(t) = f(t, y(t))$ .

**THEOREM 1.1** (Picard–Lindelöf<sup>1</sup>).— Assume that  $f$  is a *continuous* function on  $I \times \Omega$  and that for every  $(t_*, y_*) \in I \times \Omega$ , there exist  $\rho_* > 0$  and  $L_* > 0$ , such that

---

1 Also known as the Cauchy–Lipschitz Theorem.

$B(y_*, \rho_*) \subset \Omega$ ,  $[t_* - \rho_*, t_* + \rho_*] \subset I$ , and if  $y, z \in B(y_*, \rho_*)$  and  $|t - t_*| \leq \rho_*$ , then we have

$$|f(t, y) - f(t, z)| \leq L_* |y - z|.$$

Then for every  $(t_*, y_*) \in I \times \Omega$ , there exist  $r_* > 0$  and  $h_* > 0$ , such that if  $|y_{\text{Init}} - y_*| \leq r_*$  and  $|t_{\text{Init}} - t_*| \leq r_*$ , the problem [1.1] has a solution  $y : ]t_{\text{Init}} - h_*, t_{\text{Init}} + h_*[ \rightarrow \Omega$ , which is a class  $C^1$  function.

This solution is unique in the sense that if  $z$  is a function of class  $C^1$  defined on  $]t_{\text{Init}} - h_*, t_{\text{Init}} + h_*[$  satisfying [1.1], then  $y(t) = z(t)$  for all  $t \in ]t_{\text{Init}} - h_*, t_{\text{Init}} + h_*[$ .

Finally, if  $f$  is a function of class  $C^k$  on  $I \times \Omega$ , then  $y$  is a function of class  $C^{k+1}$ .

This statement calls for a number of comments, which we present in detail here.

1) Theorem 1.1 assumes that the function  $f$  satisfies a certain regularity property with respect to the state variable, this property is stronger than mere continuity:  $f$  must be Lipschitz continuous in the state variable, at least locally.<sup>2</sup> In particular, note that *if  $f$  is a function of class  $C^1$  on  $I \times \Omega$ , then it satisfies the assumptions of theorem 1.1*. This regularity hypothesis cannot be completely ruled out. However, we will see later that it can be relaxed slightly.

2) Theorem 1.1 only defines the solution in a neighborhood of the initial time  $t_{\text{Init}}$ . Once the question of existence–uniqueness is settled, it can be worthwhile to take interest in the solution’s *lifespan*: is the solution only defined on a bounded interval, or does it exist for all times? We will see that the answer depends on estimates that can be established for the solution of [1.1].

The “classic proof” for the Picard–Lindelöf theorem is based on a fixed point argument that requires the following statement.

**THEOREM 1.2** (Banach theorem).— Let  $E$  be a vector space with norm  $\|\cdot\|$ , for which it is assumed that  $E$  is complete. Let  $\mathcal{T} : E \rightarrow E$  be a strict contraction mapping, that is to say, such that there exists  $0 < k < 1$ , which satisfies the following inequality for all  $x, y \in E$ :

$$\|\mathcal{T}(x) - \mathcal{T}(y)\| \leq k \|x - y\|.$$

Then,  $\mathcal{T}$  has a unique fixed point in  $E$ .

---

<sup>2</sup> The terminology “Lipschitz continuous with respect to the second variable”, found in many books, is likely to lead to unfortunate confusion and misinterpretations, for example when addressing an autonomous system in dimension 2.

PROOF.– Let us begin by establishing uniqueness, assuming existence: if  $x$  and  $y$  satisfy  $\mathcal{T}(x) = x$  and  $\mathcal{T}(y) = y$ , then  $\|\mathcal{T}(x) - \mathcal{T}(y)\| = \|x - y\| \leq k\|x - y\|$ , which implies that  $x = y$  because  $0 < k < 1$ . In order to show existence, let us examine the sequence defined iteratively by  $x_{n+1} = \mathcal{T}(x_n)$ , starting at any  $x_0 \in E$ . We have

$$\begin{aligned}\|x_{n+p} - x_n\| &\leq \|x_{n+p} - x_{n+p-1}\| + \dots + \|x_{n+1} - x_n\| \\ &\leq \|\mathcal{T}(x_{n+p-1}) - \mathcal{T}(x_{n+p-2})\| + \dots + \|\mathcal{T}(x_n) - \mathcal{T}(x_{n-1})\| \\ &\leq k \left( \|x_{n+p-1} - x_{n+p-2}\| + \dots + \|x_n - x_{n-1}\| \right) \leq \|x_1 - x_0\| \sum_{j=n}^{n+p-1} k^j.\end{aligned}$$

Since  $0 < k < 1$ , the series  $\sum_{j=0}^{\infty} k^j$  converges. It follows that the sequence  $(x_n)_{n \in \mathbb{N}}$  is Cauchy in the complete space  $E$ . So, it has a limit  $x$  and by continuity of the mapping  $\mathcal{T}$ , we obtain  $\mathcal{T}(x) = x$ .  $\square$

PROOF OF THEOREM 1.1.– The proof of theorem 1.1 is based on a functional analysis argument: the subtle trick works with a vector space whose “points” are functions. In this case, it is important to distinguish between:

- the function  $t \mapsto y(t)$ , which is a point in the functional space (here  $C^0([t_{\text{Init}}, T[; \mathbb{R}^D]$ , for example);
- and its value  $y(t)$  for a fixed  $t$ , which is a point in the state space  $\mathbb{R}^D$ .

We will justify theorem 1.1 in the case where  $f$  is *globally* Lipschitz with respect to the state variable: we assume that  $f$  is defined on  $\mathbb{R} \times \mathbb{R}^D$  and that there exists an  $L > 0$ , such that for all  $x, y \in \mathbb{R}^D$  and any  $t \in \mathbb{R}$ , we have

$$|f(t, x) - f(t, y)| \leq L|x - y|. \quad [1.2]$$

This technical limitation is important, but enables us to only focus on the key elements of the proof. A proof of the general case can be found in [BEN 10] or [ARN 88, Chapter 4], and later we present a somewhat different approach, which starts from the perspective of numerical approximations. The starting point of the proof is to integrate [1.1] to obtain

$$y(t) = y_{\text{Init}} + \int_{t_{\text{Init}}}^t f(s, y(s)) \, ds, \quad [1.3]$$

such that the solution  $y$  of [1.1] is interpreted as a fixed point of the mapping

$$\begin{aligned}\mathcal{T} : C^0(\mathbb{R}; \mathbb{R}^D) &\longrightarrow C^0(\mathbb{R}; \mathbb{R}^D) \\ [t \mapsto z(t)] &\longmapsto \left[ t \mapsto y_{\text{Init}} + \int_{t_{\text{Init}}}^t f(s, z(s)) \, ds \right].\end{aligned}$$

We will see that this point of view, which transforms from a differential equation to an integral equation (the formulation of [1.3]), is also useful for finding numerical approximations to the solutions of [1.1]. It is now necessary to construct a functional space and a norm in order for  $\mathcal{T}$  to be a contraction. Thus, the sequence defined by  $y_0$  given in  $C^0(\mathbb{R}; \mathbb{R}^D)$  and  $y_{n+1} = \mathcal{T}(y_n)$  will converge to a fixed point of  $\mathcal{T}$ , which will be the solution to [1.1] (Picard method). We only focus on time  $t \geq t_{\text{Init}}$ . We introduce the auxiliary function

$$\mathcal{A}(t) = 1 + |y_{\text{Init}}| + \int_{t_{\text{Init}}}^t |f(s, 0)| \, ds > 0$$

and we set

$$\|y\| = \sup_{t \geq t_{\text{Init}}} \left( \frac{e^{-Mt}}{\mathcal{A}(t)} |y(t)| \right)$$

with  $M > 0$  that remains to be defined. We denote the subspace of functions  $z \in C^0([t_{\text{Init}}, \infty[; \mathbb{R}^D)$ , such that  $\|z\| < \infty$  as  $\mathcal{E}$ . With norm  $\|\cdot\|$ , this space is complete. If  $y = \mathcal{T}(z)$ , with  $z \in \mathcal{E}$ , we can write

$$y(t) = \left( y_{\text{Init}} + \int_{t_{\text{Init}}}^t f(s, 0) \, ds \right) + \int_{t_{\text{Init}}}^t (f(s, z(s)) - f(s, 0)) \, ds$$

and deduce that  $y \in \mathcal{E}$ . Indeed, this function  $t \mapsto y(t)$  is continuous and satisfies

$$\begin{aligned}\frac{e^{-Mt}}{\mathcal{A}(t)} |y(t)| &\leq 1 + \frac{e^{-Mt}}{\mathcal{A}(t)} \int_{t_{\text{Init}}}^t L e^{Ms} \mathcal{A}(s) \times \frac{e^{-Ms} |z(s)|}{\mathcal{A}(s)} \, ds \\ &\leq 1 + \frac{e^{-Mt}}{\mathcal{A}(t)} \int_{t_{\text{Init}}}^t L e^{Ms} \mathcal{A}(s) \, ds \|z\|,\end{aligned}$$

for every  $t \geq t_{\text{Init}}$ . Finally, using integration by parts, we calculate

$$\begin{aligned}\frac{e^{-Mt}}{\mathcal{A}(t)} \int_{t_{\text{Init}}}^t L e^{Ms} \mathcal{A}(s) \, ds &= \frac{e^{-Mt}}{\mathcal{A}(t)} \frac{L}{M} \int_{t_{\text{Init}}}^t \frac{d}{ds} (e^{Ms}) \times \mathcal{A}(s) \, ds \\ &= \frac{L}{M} \times \frac{e^{-Mt}}{\mathcal{A}(t)} \left( e^{Mt} \mathcal{A}(t) - e^{Mt_{\text{Init}}} (1 + |y_{\text{Init}}|) - \int_{t_{\text{Init}}}^t e^{Ms} |f(s, 0)| \, ds \right) \leq \frac{L}{M}.\end{aligned}$$

Thus, we have  $\|y\| < \infty$ . Similarly, we have

$$\begin{aligned}\|\mathcal{T}(y) - \mathcal{T}(z)\| &\leq \frac{e^{-Mt}}{\mathcal{A}(t)} L \int_{t_{\text{Init}}}^t e^{Ms} \mathcal{A}(s) \times \frac{e^{-Ms}}{\mathcal{A}(s)} |y(s) - z(s)| ds \\ &\leq \frac{L}{M} \|y - z\|.\end{aligned}$$

By choosing  $M > L$ , the mapping  $\mathcal{T}$  appears as a contraction in the complete space  $(\mathcal{E}, \|\cdot\|)$ . The Banach theorem ensures the existence and uniqueness of a fixed point, since the relation  $y = \mathcal{T}(y)$  proves that  $t \mapsto y(t)$  is a continuous and even  $C^1$  function because  $f$  is continuous. We can easily adapt the proof in order to expand the resulting solution to time  $t \leq t_{\text{Init}}$ . Interestingly, by assuming  $f$  is *globally Lipschitz* continuous with respect to the state variable, see [1.2], it has been possible to directly show that the solution is defined for any time. This fact is important and should be justified on its own.  $\square$

**THEOREM 1.3** (Picard–Lindelöf theorem, assuming *global Lipschitz continuity*).— Let  $f$  be a continuous function defined on  $\mathbb{R} \times \mathbb{R}^D$ , which satisfies [1.2]. Then, for every  $y_{\text{Init}} \in \mathbb{R}^D$ , the equation [1.1] has a unique solution  $y$  of class  $C^1$  defined on  $\mathbb{R}$ .

Let us now return to the comments for theorem 1.1 on the regularity of the function  $f$ . First, in problem [1.1], if we interpret the equation as [1.3], the assumption that  $f$  is continuous in the time variable might be weakened; it suffices to assume integrability. For example, the proof for theorem 1.3 can be slightly modified in order to justify the existence and uniqueness of a fixed point of the mapping  $\mathcal{T}$  in a space of continuous functions on  $[t_{\text{Init}}, \infty[$  by assuming that there exists  $t \mapsto L(t)$ , a function locally integrable on  $[t_{\text{Init}}, \infty[$ , such that for every  $t \geq t_{\text{Init}}$ , and all  $x, y \in \mathbb{R}^D$ , we have

$$|f(t, x) - f(t, y)| \leq L(t) |x - y|.$$

This hypothesis generalizes [1.2] (which amounts to the case of  $L(t) = L$ ). We therefore obtain a function  $y$  as a solution to [1.3], which is continuous but not necessarily  $C^1$ , and the equation [1.1] is only satisfied in a generalized sense. Next, assume that the function  $f$  defined on  $\mathbb{R} \times \mathbb{R}^D$  is only continuous and bounded on  $\mathbb{R} \times \mathbb{R}^D$  (there exists  $C > 0$ , such that for every  $t, x$ , we have  $|f(t, x)| \leq C$ ). We introduce a regularizing sequence by defining, for  $\ell \in \mathbb{N} \setminus \{0\}$ ,

$$\rho_\ell(x) = \ell^D \rho(\ell x),$$

where  $\rho \in C_c^\infty(\mathbb{R}^D)$  is a smooth function with compact support in  $\mathbb{R}^D$ , such that

$$0 \leq \rho(x) \leq 1, \quad \int_{\mathbb{R}^D} \rho(x) dx = 1.$$

We define

$$f_\ell(t, x) = \int_{\mathbb{R}^D} f(t, x - y) \rho_\ell(y) dy = f(t, \cdot) \star \rho_\ell(x).$$

(For more details about these convolution regularization techniques, the reader may refer to [GOU 11, Section 4.4]). Thus, for every fixed  $\ell$ ,  $f_\ell$  is continuous and globally Lipschitz in the state variable, because by writing

$$\begin{aligned} \rho_\ell(z_2) - \rho_\ell(z_1) &= \int_0^1 \frac{d}{d\theta} (\rho_\ell(z_1 + \theta(z_2 - z_1))) d\theta \\ &= \ell^D \int_0^1 (\nabla \rho)(\ell(z_1 + \theta(z_2 - z_1))) \cdot \ell(z_2 - z_1) d\theta \end{aligned}$$

we obtain

$$\begin{aligned} |f_\ell(t, x) - f_\ell(t, y)| &\leq \int_{\mathbb{R}^D} |f(t, z)| \left| \int_0^1 (\nabla \rho)(\ell(x - z + \theta(y - x)) \right. \\ &\quad \left. \cdot \ell(x - y) d\theta \right| \ell^D dz \\ &\leq C\ell|x - y| \times \int_0^1 \int_{\mathbb{R}^D} |\nabla \rho(z')| dz' d\theta = C\ell \|\nabla \rho\|_{L^1} |x - y|, \end{aligned}$$

where we use Fubini's theorem (for functions with positive values) and then a variable change  $z' = \ell(x - z - \theta(x - y))$ ,  $dz' = \ell^D dz$ . By theorem 1.1 for every  $\ell \in \mathbb{N} \setminus \{0\}$ , there exists a function  $t \mapsto y_\ell(t)$  of class  $C^1$ , which is defined on  $\mathbb{R}$  and is a solution to

$$y'_\ell(t) = f_\ell(t, y_\ell(t)), \quad y_\ell(t_{\text{Init}}) = y_{\text{Init}}.$$

However,  $f_\ell$  is uniformly bounded with respect to  $\ell$ ; in particular, we have  $|f_\ell(t, x)| \leq C$ . From this, we deduce that for every  $0 < T < \infty$ , the set  $\{t \in [-T, T] \mapsto y_\ell(t), \ell \in \mathbb{N} \setminus \{0\}\}$  is equibounded and equicontinuous in  $C^0([-T, T])$ . The Arzela–Ascoli theorem (see [GOU 11], theorem 7.49 and example 7.50) allows us to extract a subsequence  $(y_{\ell_k})_{k \in \mathbb{N}}$  that converges uniformly on  $[-T, T]$ , whereas  $\lim_{k \rightarrow \infty} \ell_k = +\infty$ . We can therefore see that

$$|f_{\ell_k}(s, y_{\ell_k}(s)) - f(s, y(s))| \leq \int_{\mathbb{R}^D} |f(s, y_{\ell_k}(s) - z/\ell_k) - f(s, y(s))| \rho(z) dz.$$

Since  $y_{\ell_k}(s)$  tends to  $y(s)$  and  $f$  is continuous, the integrand tends to 0 when  $k \rightarrow \infty$ , and, moreover, it is still dominated by  $2C\rho(z) \in L^1(\mathbb{R}^D)$ . Lebesgue's

theorem implies that  $\lim_{k \rightarrow \infty} f_{\ell_k}(s, y_{\ell_k}(s)) = f(s, y(s))$ . Letting  $k$  tend to  $+\infty$  in the integral relation

$$y_{\ell_k}(t) = y_{\text{Init}} + \int_{t_{\text{Init}}}^t f_{\ell_k}(s, y_{\ell_k}(s)) \, ds,$$

for every  $t \in [-T, T]$ , with  $t_{\text{Init}} < T < \infty$ , by the Lebesgue theorem, we obtain

$$y(t) = y_{\text{Init}} + \int_{t_{\text{Init}}}^t f(s, y(s)) \, ds$$

and finally we can conclude that  $y$  is a solution of [1.1]. The continuity of  $f$  is therefore sufficient to show the existence of a solution to the equation [1.1]. This is the *Cauchy–Peano theorem*. However, the solution is not in general unique. A simple counter example, in dimension  $D = 1$ , is given by the equation

$$y'(t) = \sqrt{y(t)}, \quad y(0) = 0.$$

It is clear that  $t \mapsto y(t) = 0$  is a solution on  $\mathbb{R}$ . But it is not the only one because  $t \mapsto z(t) = t^2/4$  is also a solution. The assumptions of theorem 1.1 are not satisfied because  $y \mapsto \sqrt{y}$  is continuous but not Lipschitz around  $y = 0$  (the derivative is  $\frac{1}{2\sqrt{y}}$ , which tends to  $+\infty$  when  $y \rightarrow 0$ ).

Thus, we note that on its own, the continuity of  $f$  in the state variable is not enough to ensure the truth of an existence–uniqueness statement as strong as theorem 1.1, as shown by the example  $f(t, y) = \sqrt{y}$ . However, we can slightly relax the assumption of regularity with respect to the state variable stated in theorem 1.1 and justify the existence–uniqueness of solutions to the differential problem. Sometimes, more sophisticated statements like these are necessary. We will study one such example in detail. The following statement is crucial in the analysis of the equations for incompressible fluid mechanics [CHE 95].

**DEFINITION 1.1.–** Let  $\mu : [0, \infty[ \rightarrow [0, \infty[$  be a continuous, strictly increasing function, such that  $\mu(0) = 0$ . We denote by  $C_\mu(\Omega)$  the set of functions  $u$  that are continuous on  $\Omega$ , and for which there exists a  $C > 0$ , such that

$$|u(x) - u(y)| \leq C\mu(|x - y|)$$

for all  $x, y \in \Omega$ .  $C_\mu(\Omega)$  is a Banach space for the norm

$$\|u\|_\mu = \|u\|_{L^\infty} + \sup_{x \neq y} \frac{|u(x) - u(y)|}{\mu(|x - y|)}.$$

Therefore, it is convenient to consider the right hand side of equation [1.1] as “a function in the state variable, with the time variable as a parameter”: for (almost) every  $t \in I$ , we assume that  $y \mapsto f(t, y)$  is a function in  $C_\mu(\Omega; \mathbb{R}^D)$ . We say that  $f \in L^1(I; C_\mu(\Omega))$  when there exists an  $L \in L^1(I)$  with strictly positive values, such that for all  $x, y \in \Omega$

$$|f(t, x)| \leq L(t), \quad |f(t, x) - f(t, y)| \leq L(t) \mu(|x - y|). \quad [1.4]$$

**THEOREM 1.4** (Osgood theorem).— Let  $\mu : [0, \infty[ \rightarrow [0, \infty[$  be a continuous, strictly increasing function, such that  $\mu(0) = 0$ . Assume further that

$$\int_0^\infty \frac{ds}{\mu(s)} = +\infty. \quad [1.5]$$

Let  $I$  be an interval of  $\mathbb{R}$  that contains  $t_{\text{Init}}$  and  $f \in L^1(I; C_\mu(\Omega))$ . Then, for every  $y_{\text{Init}} \in \Omega$ , there is an interval  $\mathcal{I} \subset I$  containing  $t_{\text{Init}}$ , such that the equation [1.1] has a unique solution  $t \mapsto y(t)$  defined on  $\mathcal{I}$ . Equation [1.1] is understood in the sense of [1.3] being satisfied. (We say that  $y$  is a *mild solution* of [1.1]).

**PROOF.**— In fact, this statement raises two questions of a somewhat different nature. We can easily adapt the proof of the Cauchy–Peano theorem to demonstrate the existence of a continuous solution to [1.3] with the simple assumption that  $f \in L^1(I; C(\Omega))$ . However, this does not justify uniqueness, which has been found to be wrong, with the mere continuity of  $f$ . Extending uniqueness to functions satisfying [1.4] is a result due to [OSG 98]. Another problem is to justify the convergence of the sequence of Picard iterations. Given  $y_0 \in C^0[t_{\text{Init}}, \infty[$ , then

$$y_{k+1}(t) = y_{\text{Init}} + \int_{t_{\text{Init}}}^t f(s, y_k(s)) ds, \quad [1.6]$$

converges to one (and therefore *the*) solution of [1.3]. This issue has been studied for its part in [WIN 46].

Let us first prove uniqueness. Suppose that there are two solutions,  $y_1$  and  $y_2$ , to [1.3], and define  $z = |y_2 - y_1|$ . Therefore, for  $t \geq t_{\text{Init}}$ , we obtain

$$0 \leq z(t) \leq \int_{t_{\text{Init}}}^t L(s) \mu(z(s)) ds, \quad z(t_{\text{Init}}) = 0.$$

We then focus on functions  $\alpha : [t_{\text{Init}}, \infty[ \rightarrow [0, \infty]$ , such that

$$0 \leq \alpha(t) \leq \alpha_0 + \int_{t_{\text{Init}}}^t L(s) \mu(\alpha(s)) ds.$$

Let us first assume that  $\alpha_0 > 0$ . We define

$$\Omega : t \in [t_{\text{Init}}, \infty[ \mapsto \int_{t_{\text{Init}}}^t L(s) \mu(\alpha(s)) \, ds.$$

This function is continuous, non-negative, and satisfies  $\Omega(t_{\text{Init}}) = 0$ . In addition, since  $\Omega$  is defined as the integral of a positive function, it is non-decreasing and therefore differentiable almost everywhere, with

$$0 \leq \Omega'(t) = L(t) \mu(\alpha(t)) \leq L(t) \mu(\alpha_0 + \Omega(t))$$

since  $L(t) \geq 0$  and  $\mu$  are non-decreasing. Because  $\alpha_0 + \Omega(t) \geq \alpha_0 > 0$  and  $\mu$  is strictly increasing, we have  $\mu(\alpha_0 + \Omega(t)) > \mu(0) = 0$ , making the following relation true:

$$\int_{t_{\text{Init}}}^t \frac{\Omega'(s)}{\mu(\alpha_0 + \Omega(s))} \, ds \leq \int_{t_{\text{Init}}}^t L(s) \, ds.$$

Recalling that  $\Omega(t_{\text{Init}}) = 0$  and with the change of variables  $\sigma = \alpha_0 + \Omega(s)$ ,  $d\sigma = \Omega'(s) \, ds$ , it becomes

$$\int_{\alpha_0}^{\alpha_0 + \Omega(t)} \frac{d\sigma}{\mu(\sigma)} = \int_{t_{\text{Init}}}^t \frac{\Omega'(s)}{\mu(\alpha_0 + \Omega(s))} \, ds \leq \int_{t_{\text{Init}}}^t L(s) \, ds.$$

In particular, the function  $t \mapsto z(t) = |y_2(t) - y_1(t)|$  satisfies this relation for any  $\alpha_0 > 0$ . Suppose that  $z$  is not equal to zero. In this case, we can find  $t_* > t_{\text{Init}}$ ,  $\eta > 0$ , such that  $z(t) > 0$  on an interval of the form  $[t_* - \eta, t_*] \subset [t_{\text{Init}}, t_*]$ . Accordingly, the function  $\Omega$ , which is associated with  $\alpha = z$ , is strictly positive at  $t_*$ . We obtain

$$\int_{\alpha_0}^{\Omega(t_*)} \frac{d\sigma}{\mu(\sigma)} \leq \int_{\alpha_0}^{\alpha_0 + \Omega(t_*)} \frac{d\sigma}{\mu(\sigma)} \leq \int_{t_{\text{Init}}}^{t_*} L(s) \, ds$$

for every  $0 < \alpha_0 < \Omega(t_*)$ , recalling that  $L$  is integrable. Letting  $\alpha_0$  tend to 0, we are led to a contradiction with [1.5].

We then examine the behavior of the sequence  $(y_k)_{k \in \mathbb{N}}$  defined by [1.6]. We first check that, when  $T$  is small enough, this sequence is well defined and remains bounded in  $C([t_{\text{Init}}, T])$ : we set  $M > 0$  and show that for every  $k \in \mathbb{N}$  and  $t \in [t_{\text{Init}}, T]$ ,

$$|y_k(t) - y_{\text{Init}}| \leq M.$$

Indeed, we have

$$\begin{aligned} |y_{k+1}(t) - y_{\text{Init}}| &= \left| \int_{t_{\text{Init}}}^t \left\{ (f(s, y_k(s)) - f(s, y_{\text{Init}})) + f(s, y_{\text{Init}}) \right\} ds \right| \\ &\leq \int_{t_{\text{Init}}}^t L(s) (\mu(|y_k(s) - y_{\text{Init}}|) + 1) ds. \end{aligned}$$

If  $|y_k(t) - y_{\text{Init}}| \leq M$ , for a given  $M > 0$ , since  $\mu$  is increasing, it follows that

$$|y_{k+1}(t) - y_{\text{Init}}| \leq (\mu(M) + 1) \int_{t_{\text{Init}}}^T L(s) ds$$

is satisfied for all  $t_{\text{Init}} \leq t \leq T$ . Let us now introduce  $T > 0$ , such that

$$(\mu(M) + 1) \int_{t_{\text{Init}}}^T L(s) ds < M,$$

which is permissible because  $\lim_{T \rightarrow t_{\text{Init}}} \int_{t_{\text{Init}}}^T L(s) ds = 0$ . Given this preliminary result, we note that for all integers  $k, p$  and  $t \in [t_{\text{Init}}, T]$ ,

$$|y_{k+p}(t) - y_k(t)| \leq \int_{t_{\text{Init}}}^t L(s) \mu(|y_{k+p-1}(s) - y_{k-1}(s)|) ds.$$

Since  $\mu$  is monotonous, we can say that  $z_k(t) = \sup_p |y_{k+p}(t) - y_k(t)|$  satisfies

$$0 \leq z_k(t) \leq \int_{t_{\text{Init}}}^t L(s) \mu(z_{k-1}(s)) ds.$$

Applying the Fatou lemma [GOU 11, lemma 3.27] to the sequence of positive functions  $\mu(2M) - \mu(z_k(t))$ , we can deduce that  $Z(t) = \limsup_{k \rightarrow \infty} z_k(t)$  satisfies

$$0 \leq Z(t) \leq \int_{t_{\text{Init}}}^t L(s) \mu(Z(s)) ds.$$

The same reasoning that made it possible to demonstrate the uniqueness of the solution to [1.3] applies here and allows us to conclude that  $Z(t) = 0$  on  $[t_{\text{Init}}, T]$ . Therefore,  $(y_k)_{k \in \mathbb{N}}$  is a Cauchy sequence in the complete space  $C([t_{\text{Init}}, T])$ . This sequence has a limit  $y$  for uniform convergence on  $[t_{\text{Init}}, T]$ . By making  $k \rightarrow \infty$  in [1.6], we effectively show that  $y$  satisfies [1.3].  $\square$

### 1.1.2. The concept of maximal solution

It is important to recall that theorem 1.1 only ensures the existence of a solution to [1.1] defined in a *neighborhood* of the initial time  $t_{\text{Init}}$ . This is an inherently *local* result. Additional information is required to prove that the solution is globally defined. The case of a function  $f$  that is globally Lipschitz continuous with respect to the state variable is one of those special situations in which the solution of [1.1] is defined for all times. However, in dimension  $D = 1$ , the example of

$$y'(t) = y^2(t), \quad y(t_{\text{Init}}) = y_{\text{Init}}$$

shows that the solution can explode in finite time, even if the second member  $f$  is a function of class  $C^1$  on  $\mathbb{R} \times \mathbb{R}^D$ . In this case, we, in fact, have

$$y(t) = \frac{1}{1/y_{\text{Init}} - (t - t_{\text{Init}})}$$

which is therefore not defined beyond<sup>3</sup>  $T^* = t_{\text{Init}} + 1/y_{\text{Init}}$ .

Defining a solution for [1.1], therefore, requires determining both an interval  $\mathcal{I}$  containing  $t_{\text{Init}}$  and a function  $y$  defined on  $\mathcal{I}$  (with values in  $\Omega$ ). Given two solutions  $(\tilde{\mathcal{I}}, \tilde{y})$  and  $(\mathcal{I}, y)$  for [1.1], we can say that  $(\tilde{\mathcal{I}}, \tilde{y})$  extends  $(\mathcal{I}, y)$  if

$$\mathcal{I} \subset \tilde{\mathcal{I}} \quad \text{and} \quad \tilde{y}|_{\mathcal{I}} = y.$$

We say that  $(\tilde{\mathcal{I}}, \tilde{y})$  is a *maximal solution* if the solution  $(\tilde{\mathcal{I}}, \tilde{y})$  can only be extended by itself.

**PROPOSITION 1.1.-** Let  $f$  satisfy the assumptions of theorem 1.1. For every  $t_{\text{Init}} \in I$  and  $y_{\text{Init}} \in \Omega$ , there is a unique maximal solution  $(\tilde{\mathcal{I}}, \tilde{y})$  for [1.1] and  $\tilde{\mathcal{I}} \subset I$  is an open interval of  $\mathbb{R}$ .

**PROOF.-** The set of solutions  $(\mathcal{I}, y)$  of [1.1] contains at least  $(\{t_{\text{Init}}\}, y_{\text{Init}})$ . Let us consider two solutions  $(\tilde{\mathcal{I}}, \tilde{y})$  and  $(\mathcal{I}, y)$ . In particular,  $\tilde{\mathcal{I}}$  and  $\mathcal{I}$  are two intervals containing  $t_{\text{Init}}$ . Let  $\mathcal{J} = \tilde{\mathcal{I}} \cap \mathcal{I}$ . Then,  $\mathcal{J} \subset I \subset \mathbb{R}$  is also an interval that at least contains  $t_{\text{Init}}$ , and we have  $\tilde{y}(t_{\text{Init}}) = y(t_{\text{Init}})$ . Let  $\mathcal{U} = \{t \in \mathcal{J}, \tilde{y}(t) = y(t)\}$ . By theorem 1.1,  $\mathcal{U}$  is an open set of  $\mathbb{R}$  containing  $t_{\text{Init}}$  (there is a solution of the differential equation with initial values  $(t, y(t)) = (\tilde{y}(t), y(t))$ , which is uniquely defined on an open interval  $]t - \eta, t + \eta[$ , with  $\eta$  depending on  $(t, y(t))$ ). Since  $y$  and  $\tilde{y}$  are themselves solutions to this problem, we have  $y(s) = \tilde{y}(s)$  for every  $s \in ]t - \eta, t + \eta[ \dots$ ).

---

<sup>3</sup> For  $y_{\text{Init}} > 0$  (respectively  $y_{\text{Init}} < 0$ ), the formula is only meaningful for  $t < T^*$  (respectively  $t > T^*$ ).

However,  $\mathcal{U}$  is also closed in  $\mathcal{J}$  by the continuity of functions  $\tilde{y}$  and  $y$ : if  $(t_n)_{n \in \mathbb{N}}$  is a sequence of elements of  $\mathcal{U}$ , which tends to  $t \in \mathcal{J}$ , then  $\tilde{y}(t) = y(t)$  and  $t \in \mathcal{U}$  (we can also note that  $\mathcal{U}$  is the inverse image of the closed set  $\{0\}$  by the continuous function  $y - \tilde{y}$ ). It follows that  $\mathcal{U} = \mathcal{J}$ , by characterization of connected sets of  $\mathbb{R}$ . We define  $\mathcal{V}$  as the union of all the intervals in which we can define a solution of [1.1]. For  $t \in \mathcal{V}$ , there exists a unique solution  $s \mapsto y(s)$  of [1.1] defined in a neighborhood of  $s = t$ ; we thus define  $\tilde{y}(t) = y(t)$ . By construction,  $(\mathcal{V}, \tilde{y})$  is the maximal solution of [1.1].

We introduce the set

$$\mathcal{A} = \{T > 0, \text{ such that there exists a solution } y_T \text{ of [1.1]}$$

$$\text{which satisfies } y_T(t_{\text{Init}}) = y_{\text{Init}} \text{ and is defined on } [t_{\text{Init}}, t_{\text{Init}} + T] \subset I\}.$$

As a result of theorem 1.1,  $\mathcal{A}$  is non-empty. We define

$$T^* = \sup \mathcal{A} = \sup \mathcal{V}.$$

Let  $T, T' \in \mathcal{A}$ ,  $T < T'$ . By uniqueness of the solution of the differential equation associated with the point  $(t_{\text{Init}}, y_{\text{Init}})$ , it follows that  $([t_{\text{Init}}, t_{\text{Init}} + T'], y_{T'})$  extends  $([t_{\text{Init}}, t_{\text{Init}} + T], y_T)$ ; in particular,  $\mathcal{A}$  is an interval. Let  $t_{\text{Init}} < t < t_{\text{Init}} + T^*$ . Then, there exists a  $T \in \mathcal{A}$ , such that  $t < T + t_{\text{Init}} < T^* + t_{\text{Init}}$  and  $y_T$  is the unique solution of [1.1] defined on  $[t_{\text{Init}}, t_{\text{Init}} + T]$ . Therefore, we conclude that there exists a unique solution to [1.1] associated with  $(t_{\text{Init}}, y_{\text{Init}})$  and defined on the interval  $[t_{\text{Init}}, t_{\text{Init}} + T^*]$ , which is an open set of  $[t_{\text{Init}}, +\infty] \cap I$ . Finally, if  $T^*$  is a finite element of  $\mathcal{A}$ , then  $(T^*, \tilde{y}(T^*)) \in I \times \Omega$  and theorem 1.1 extends the solution beyond  $T^*$ . We conclude that  $\mathcal{J}$  is an open interval.  $\square$

NOTE.– An important special case that allows us to estimate solutions is when there exists a “stationary point”. Indeed, if  $f(t, x_0) = 0$  for every  $t \in I$ , then the constant function  $t \mapsto x_0$  is a solution to [1.1], with  $y_{\text{Init}} = x_0$  being the initial data (for any initial time  $t_{\text{Init}}$ ). Uniqueness implies that no other solution of [1.1] can pass through  $x_0$ .

Let  $(\mathcal{J}, y)$  be the maximal solution of [1.1]. We can write  $\mathcal{J} = ]T_*, T^*[$ . If  $T^*$  is finite, we call *right end* the associated set

$$\begin{aligned} \omega^* &= \{b \in \Omega \text{ such that there exists a sequence } (t_n)_{n \in \mathbb{N}} \text{ satisfying} \\ &t_n \in I, \lim_{n \rightarrow \infty} t_n = T^* \text{ and } \lim_{n \rightarrow \infty} y(t_n) = b\}. \end{aligned}$$

We then compare  $\mathcal{J}$  with the definition set  $I$  of the data for the problem [1.1].

**THEOREM 1.5.**— We have

$$(T^*, \omega^*) \subset \overline{I \times \Omega} \setminus I \times \Omega = \partial(I \times \Omega),$$

as well as analogous definitions and results for the “left end”.

**PROOF.**— Suppose that  $T^* < \infty$  and  $\omega^* \neq \emptyset$ , that is,  $b \in \omega^*$ . By definition, we always have  $(T^*, b) \in \overline{I \times \Omega}$ . More specifically, if we denote the maximal solution of [1.1] as  $([T_*, T^*[, y)$ , there exists a sequence  $(t_n)_{n \in \mathbb{N}}$ , such that

$$(t_n, y(t_n)) \in I \times \Omega, \quad \lim_{n \rightarrow \infty} t_n = T^*, \quad \lim_{n \rightarrow \infty} y(t_n) = b.$$

Suppose that  $T^* \in I$  and  $\omega^* \in \Omega$ . We can therefore find  $\tau, r > 0$ , such that

- $\mathcal{C} = [T^* - \tau, T^* + \tau] \times \overline{B(b, r)} \subset I \times \Omega$ ;
- for every  $(t, y) \in \mathcal{C}$ ,  $|f(t, y)| \leq M$  and  $r > 2M\tau$ ;
- $f$  is locally Lipschitz continuous in the state variable on  $\mathcal{C}$ .

We define  $\mathcal{C}' = [T^* - \tau/3, T^* + \tau/3] \times \overline{B(b, r/3)}$ . Then, given that  $(s, z) \in \mathcal{C}'$ , we can see that  $\mathcal{C}'' = [s - 2\tau/3, s + 2\tau/3] \times \overline{B(z, 2r/3)}$ . By construction, we have  $\mathcal{C}' \subset \mathcal{C}'' \subset \mathcal{C}$ . The local existence theorem ensures the existence of a function  $t \mapsto \psi(t)$ , defined in a neighborhood  $\mathcal{J}$  of  $s$ , which is a solution of  $\psi'(t) = f(t, \psi(t))$ , with  $\psi(s) = z$ . Moreover, the neighborhood can be chosen in such a way that  $(t, \psi(t)) \in \mathcal{C}''$  for every  $t \in \mathcal{J}$ .

Furthermore, there exists an integer  $N$ , such that  $|t_N - T^*| \leq \tau/3$  and  $|y(t_N) - b| \leq r/3 : (t_N, X(t_N)) \in \mathcal{C}' \subset \mathcal{C}''$ . We can find a function  $t \mapsto \psi(t)$  that satisfies  $\psi'(t) = f(t, \psi(t))$  and  $\psi(t_N) = y(t_N)$ . By writing

$$\psi(t) = y(t_N) + \int_{t_N}^t f(s, \psi(s)) \, ds,$$

we can see that if  $|t - t_N| \leq 2\tau/3$ ,

$$\begin{aligned} |\psi(t) - b| &\leq \left| y(t_N) - b + \int_{t_N}^t f(s, \psi(s)) \, ds \right| \\ &\leq |y(t_N) - b| + M|t - t_N| \leq r/3 + 2M\tau/3 < 2r/3. \end{aligned}$$

That is,  $\psi(t) \in \overline{B(b, 2r/3)}$ ,  $(t, \psi(t)) \in \mathcal{C}''$ . It follows that  $\psi$  is defined on an interval  $[t_N - 2\tau/3, t_N + 2\tau/3]$ .

We have two solutions to the differential equation  $x'(t) = f(t, x(t))$ , which pass through  $y(t_N)$  at the instant  $t_N$ : the maximal solution  $y$  and the solution  $\psi$  that was just presented. By uniqueness, they must be one and the same. However,  $T^* = T^* - t_N + t_N < t_N + 2\tau/3$ , so we could have extended the maximal solution, which is a contradiction.  $\square$

The statement of theorem 1.5 can be a bit abstruse, but it is possible to deduce some more practical formulations.

**COROLLARY 1.1.**— If  $T^* < \infty$ , then

- either  $y'(t) = f(t, y(t))$  is unbounded in a neighborhood of  $T^*$ ,
- or else  $t \mapsto y(t)$  has a limit  $b \in \mathbb{R}^D$  when  $t \rightarrow T^*$ , and  $(T^*, b) \notin I \times \Omega$ .

**PROOF.**— Suppose that  $T^*$  is finite and that  $t \mapsto y'(t) = f(t, y(t))$  is bounded by  $M > 0$ . It follows that  $|y(t) - y(s)| \leq M|t - s|$ . We can therefore infer that for every sequence  $(t_n)_{n \in \mathbb{N}}$  that converges to  $T^*$ ,  $(y(t_n))_{n \in \mathbb{N}}$  is a Cauchy sequence and thus has a limit. Moreover, this limit does not depend on the sequence considered. In other words, there exists a  $b \in \mathbb{R}^D$ , such that  $\lim_{t \rightarrow T^*} y(t) = b$ . Theorem 1.5 ensures that  $(T^*, b) \notin I \times \Omega$ . This result is subtle: there is no guarantee that  $y$  is not extensible by continuity in  $T^*$  and that, by that same token,  $f(t, y)$  can be extended by continuity in  $(T^*, b)$ .  $\square$

**COROLLARY 1.2 (blow up criterion).**— If  $f$  is defined on  $\mathbb{R} \times \mathbb{R}^D$  and one of the ends  $T_*$  or  $T^*$ , denoted as  $\bar{T}$ , is finite, we have  $\lim_{t \rightarrow \bar{T}} |y(t)| = +\infty$ .

**PROOF.**— Let  $(t_n)_{n \in \mathbb{N}}$  be a sequence that converges to  $T^* < \infty$ , such that  $(y(t_n))_{n \in \mathbb{N}}$  remains bounded. We can thus extract a subsequence, such that  $\lim_{k \rightarrow \infty} y(t_{n_k}) = b \in \mathbb{R}^N$ . By theorem 1.5,  $(T^*, b)$  is not in the domain of  $f$ . However, since the domain is the entire space  $\mathbb{R} \times \mathbb{R}^D$ , we have a contradiction.  $\square$

It is useful to have simple and representative examples in mind for the variety of situations that can occur (in one dimension):

- The function  $f(t, y) = y^2$  is defined on all of  $\mathbb{R}$ ; it is locally but not globally Lipschitz continuous. The solutions to [1.1] are not defined for all times: there exists a finite lifetime  $T^*(y_{\text{Init}})$  and  $\lim_{t \rightarrow T^*(y_{\text{Init}})} y(t) = +\infty$ .
- With  $f(t, y) = -1/y$ , we have  $I \times \Omega = \mathbb{R} \times (\mathbb{R} \setminus \{0\})$ . For  $t_{\text{Init}} = 0$  and  $y_{\text{Init}} > 0$ , we get  $y(t) = \sqrt{y_{\text{Init}}^2 - 2t}$ : lifetime in the future  $T^*(y_{\text{Init}})$  is finite and  $\lim_{t \rightarrow T^*(y_{\text{Init}})} y(t) = 0$ , at the edge of  $f$ 's domain. Moreover, we have  $\lim_{t \rightarrow T^*(y_{\text{Init}})} y'(t) = -\infty$ .
- The function  $f(t, y) = \sqrt{y}$  is locally Lipschitz continuous on  $\mathbb{R} \times \Omega$ , with  $\Omega = ]0, \infty[$ , and continuous at  $(t, 0)$  for every  $t \in \mathbb{R}$ . For  $t_{\text{Init}} = 0$  and  $y_{\text{Init}} > 0$ ,

we obtain  $y(t) = (\sqrt{y_{\text{Init}}} + t/2)^2$ . The lifespan in the future  $T^*(y_{\text{Init}})$  is infinite, but the lifespan in the past  $T_*(y_{\text{Init}})$  is finite with  $\lim_{t \rightarrow T_*(y_{\text{Init}})} y(t) = 0$ , at the edge of  $f$ 's domain. Here we have  $\lim_{t \rightarrow T^*(y_{\text{Init}})} y'(t) = 0$ .

– The function  $f(t, y) = \sin(1/y)$  is locally Lipschitz continuous on  $\mathbb{R} \times \Omega$ , with  $\Omega = \mathbb{R} \setminus \{0\}$ , and cannot be extended by continuity at 0. Since this function is bounded, if the lifespan  $T^*$  of the maximal solution for [1.1] is finite, then  $y(t)$  should tend to 0 when  $t \rightarrow T^*$ . However, every initial value  $y_{\text{Init}} > 0$  can be bounded from above and below by equilibrium points of the form  $\frac{1}{n\pi}$ . The maximal solution is therefore globally defined:  $T^*(y_{\text{Init}}) = +\infty$ .

By virtue of corollary 1.2, in order to establish that the maximal solution of [1.1] is defined for all times, it suffices to provide estimates that rule out the possibility that the solution blows up. A very useful tool for obtaining this kind of estimates is as follows.

**LEMMA 1.1 (Grönwall's Lemma).**– Let  $\alpha_0 \in \mathbb{R}$  and  $a, b$  and  $\alpha$  be continuous, real-valued functions on  $[0, T]$ ,  $0 < T < \infty$ , with  $a(t) \geq 0$  for all  $t \in [0, T]$ . Suppose that

$$\alpha(t) \leq \alpha_0 + \int_0^t a(s)\alpha(s) \, ds + \int_0^t b(s) \, ds$$

for  $0 \leq t \leq T$ . Then, we have

$$\alpha(t) \leq \alpha_0 \exp \left( \int_0^t a(s) \, ds \right) + \int_0^t b(s) \exp \left( \int_s^t a(\sigma) \, d\sigma \right) \, ds. \quad [1.7]$$

**PROOF.**– We denote the majorant in [1.7] as  $\beta(t)$ , which is a differentiable function that satisfies

$$\beta'(t) = a(t)\beta(t) + b(t), \quad \beta(0) = \alpha_0.$$

We set  $\delta(t) = \beta(t) - \alpha(t)$ , and we wish to show that it is positive-valued. Now, we have

$$\begin{aligned} \delta(t) &= \beta(0) + \int_0^t \beta'(s) \, ds - \alpha(t) \\ &= \alpha_0 + \int_0^t (a(s)\beta(s) + b(s)) \, ds - \alpha(t) \geq \int_0^t a(s)\delta(s) \, ds. \end{aligned}$$

We denote the right-hand term as  $\Delta(t)$ , which is a differentiable function and satisfies

$$\Delta'(t) = a(t)\delta(t) \geq a(t)\Delta(t)$$

since  $a(t) \geq 0$ , with  $\Delta(0) = 0$ . It follows that

$$\frac{d}{dt} \left( \exp \left( - \int_0^t a(s) ds \right) \Delta(t) \right) = \exp \left( - \int_0^t a(s) ds \right) [\Delta'(t) - a(t)\Delta(t)] \geq 0.$$

This implies that

$$\exp \left( - \int_0^t a(s) ds \right) \Delta(t) \geq \Delta(0) = 0$$

and that  $\Delta(t) \geq 0$ . Finally, we conclude that  $\delta(t) \geq \Delta(t) \geq 0$ .  $\square$

**COROLLARY 1.3.**— Let  $f$  be a continuous, locally Lipschitz function defined on  $\mathbb{R} \times \mathbb{R}^D$ . Suppose that there exists an  $M > 0$ , such that for all  $t, y$ , we have  $|f(t, y)| \leq M(1 + |y|)$ . Then, the maximal solution of [1.1] is defined on all of  $\mathbb{R}$ .

**PROOF.**— The maximal solution satisfies the following estimate, for  $t > t_{\text{init}}$ ,

$$|y(t)| \leq |y_{\text{Init}}| + \int_{t_{\text{init}}}^t |f(s, y(s))| ds \leq |y_{\text{Init}}| + M \int_{t_{\text{init}}}^t (1 + |y(s)|) ds.$$

Grönwall's lemma allows us to infer that

$$|y(t)| \leq e^{M(t-t_{\text{Init}})} (|y_{\text{Init}}| + M(t - t_{\text{Init}})).$$

This estimate rules out the possibility that the solution explodes in finite time.  $\square$

### 1.1.3. Linear systems with constant coefficients

For linear systems with constant coefficients, the solution can be expressed with the matrix exponential: the solution of

$$y'(t) = Ay(t) + b(t), \quad y(0) = y_{\text{Init}} \quad [1.8]$$

is given by

$$y(t) = e^{At} y_{\text{Init}} + \int_0^t e^{A(t-s)} b(s) ds$$

where

$$e^{At} = \sum_{k=0}^{\infty} \frac{(At)^k}{k!}.$$

This formula allows us to determine the qualitative properties of the solution. In some cases, especially for small dimensions, it allows us to find explicit expressions. The process relies on the Jordan–Chevalley decomposition (see, for example, [SER 01, Prop. 2.7.2]).

**THEOREM 1.6** (Jordan–Chevalley decomposition).— Every matrix  $A \in \mathcal{M}_N(\mathbb{C})$  can be written in the form  $A = D + N$ , where  $D$  is diagonalizable on  $\mathbb{C}$ ,  $N$  is nilpotent and  $DN = ND$ .

We write the eigenvalues  $\lambda_1, \dots, \lambda_N$ , of  $D$  repeated according to their multiplicity, with  $p+1$  as the nilpotence index of  $N$ . Since  $D$  and  $N$  commute, we can write

$$\begin{aligned} e^{At} &= e^{(D+N)t} = e^{Dt}e^{Nt} \\ &= P\text{diag}(e^{\lambda_1 t}, \dots, e^{\lambda_N t})P^{-1}\left(\mathbb{I} + tN + \frac{t^2 N^2}{2!} + \dots + \frac{t^p N^p}{p!}\right). \end{aligned}$$

This method can be used for “small” systems: once the eigenvalues, eigenvectors and generalized eigenvectors of the matrix  $A$  are known, we have an explicit formulation of  $e^{tA}$ . More generally, the spectral properties of the matrix are reflected in the qualitative behavior of the solutions to [1.8]. In order to show this relationship, we recall the following notions and notation:

–  $\lambda \in \mathbb{C}$  is an *eigenvalue of  $A$*  whenever  $E_\lambda = \text{Ker}(A - \lambda\mathbb{I}) \neq \{0\}$ , a vector subspace called *eigenspace associated with the eigenvalue  $\lambda$* .

– We also define the *characteristic space associated with an eigenvalue  $\lambda$* :  $\tilde{E}_\lambda = \text{Ker}((A - \lambda\mathbb{I})^k)$ , where  $k$  is the largest integer, such that

$$\text{Ker}((A - \lambda\mathbb{I})^{k-1}) \subset \text{Ker}((A - \lambda\mathbb{I})^k), \quad \text{Ker}((A - \lambda\mathbb{I})^{k-1}) \neq \text{Ker}((A - \lambda\mathbb{I})^k).$$

– The dimension of  $E_\lambda$  is the *geometric multiplicity of eigenvalue  $\lambda$* , and the dimension of  $\tilde{E}_\lambda$  is the *algebraic multiplicity of eigenvalue  $\lambda$* . When the two values are the same and  $k > 1$ ,  $\lambda$  is said to be a semisimple eigenvalue; if  $\dim(\tilde{E}_\lambda) = \dim(E_\lambda) = 1$ , then  $\lambda$  is said to be a simple eigenvalue. The algebraic multiplicity of  $\lambda$  is, in fact, given by the order of  $\lambda$ , as the root of the characteristic polynomial  $\mu \mapsto \det(A - \mu\mathbb{I})$  of the matrix  $A$ .

– It is possible to decompose  $\mathbb{C}^N = \bigoplus_{\lambda \in \sigma(A)} \tilde{E}_\lambda$ , where  $\sigma(A)$  is the set of eigenvalues (the spectrum) of  $A$ . In particular, if all the eigenvalues are semisimple, then the matrix  $A$  is diagonalizable.

**NOTE.**— It would be well to avoid hasty conclusions when determining the Jordan–Chevalley decomposition of a matrix. For example, it could be tempting to decompose

$$A = \begin{pmatrix} 2 & 3 \\ 0 & 4 \end{pmatrix} = \underbrace{\begin{pmatrix} 2 & 0 \\ 0 & 4 \end{pmatrix}}_D + \underbrace{\begin{pmatrix} 0 & 3 \\ 0 & 0 \end{pmatrix}}_{=N}.$$

It is indeed a “diagonal(-izable)+nilpotent” decomposition, but those matrices do not commute:

$$DN = \begin{pmatrix} 0 & 6 \\ 0 & 0 \end{pmatrix}, \quad ND = \begin{pmatrix} 0 & 12 \\ 0 & 0 \end{pmatrix}.$$

In fact, the matrix  $A$  has two different eigenvalues, 2 and 4, and therefore it is diagonalizable. It is similar to

$$A' = \begin{pmatrix} 2 & 0 \\ 0 & 4 \end{pmatrix}$$

with the change-of-basis matrix

$$P = \begin{pmatrix} 1 & 1 \\ 0 & 2/3 \end{pmatrix}.$$

**THEOREM 1.7.–** The mapping  $t \mapsto e^{tA}$  is bounded on  $[0, \infty[$  if and only if all the eigenvalues of  $A$  have a negative or zero real component and the strictly imaginary eigenvalues are semisimple. Furthermore,  $e^{tA}$  tends to 0 when  $t \rightarrow \infty$  if the eigenvalues of  $A$  have a strictly negative real component.

**PROOF.–** We denote the distinct eigenvalues of  $A$  as  $\lambda_1, \dots, \lambda_p$  and introduce the projection operators  $P_j$  for the projection onto  $\tilde{E}_{\lambda_j}$  parallel to  $\oplus_{k \neq j} \tilde{E}_{\lambda_k}$ . Thus, for each  $j \in \{1, \dots, p\}$ , we can find an integer  $k_j$ , such that

$$(A - \lambda_j \mathbb{I})^{k_j} P_j = 0, \tag{1.9}$$

and the identity matrix can be written as the sum of the projectors

$$\mathbb{I} = \sum_{j=1}^p P_j.$$

We can then write

$$\begin{aligned} e^{tA} &= e^{tA} \sum_{j=1}^p P_j = \sum_{j=1}^p e^{\lambda_j t} e^{t(A - \lambda_j \mathbb{I})} P_j \\ &= \sum_{j=1}^p e^{\lambda_j t} \left( \sum_{\ell=0}^{\infty} \frac{t^\ell}{\ell!} (A - \lambda_j \mathbb{I})^\ell \right) P_j. \end{aligned}$$

If  $\operatorname{Re}(\lambda_j) < 0$  for every  $j \in \{1, \dots, p\}$ , this quantity therefore tends to 0 when  $t \rightarrow \infty$ , by [1.9]. It remains to study the case where  $A$  has eigenvalues with a real

component equal to zero. If those eigenvalues are semisimple, then the index  $k_j$  associated with them is  $k_j = 1$ , and we can infer that  $e^{tA}$  is bounded for every  $t \geq 0$ .

Reciprocally, suppose that there exists an eigenvalue  $\lambda$ , such that  $\operatorname{Re}(\lambda) > 0$ . Let  $v \neq 0$  be an associated eigenvalue. We have  $Av = \lambda v$ , so  $e^{tA}v = e^{\lambda t}v$  is not bounded on  $[0, \infty[$ . Finally, suppose that there exists a  $\lambda \in \sigma(A) \cap i\mathbb{R}$  that is not semisimple. We can find two vectors  $v, w \neq 0$ , such that

$$Av = \lambda v, \quad Aw = \lambda w + v.$$

It follows that  $A^\ell w = \lambda^\ell w + \ell \lambda^{\ell-1} v$ , and therefore

$$e^{tA}w = \sum_{\ell=0}^{\infty} \frac{(t\lambda)^\ell}{\ell!} w + \sum_{\ell=1}^{\infty} \frac{\ell t^\ell \lambda^{\ell-1}}{\ell!} v = e^{\lambda t}(w + tv).$$

We can infer that, for a large enough  $t$ ,  $|e^{tA}w| \geq t|v| - |w|$ , which tends to  $+\infty$  when  $t \rightarrow \infty$ , and therefore  $e^{tA}$  is not bounded on  $[0, \infty[$ . Finally, if there exists  $\lambda \in \sigma(A) \cap i\mathbb{R}$ , whether or not it is simple,  $e^{tA}$  does not tend to 0 when  $t \rightarrow \infty$  because for every eigenvector  $v \neq 0$  associated with  $\lambda$ , we have  $e^{tA}v = e^{t\lambda}v$ , which has a constant norm.  $\square$

This rationale is also important in understanding the behavior of certain nonlinear equations. Indeed, we are interested in problems of the type

$$\frac{d}{dt}y(t) = f(y(t)), \tag{1.10}$$

where the function  $f : \mathbb{R}^D \rightarrow \mathbb{R}^D$  is regular enough to apply theorem 1.1. Suppose that there exists an equilibrium point  $\bar{y}$ ; that is, a solution to

$$f(\bar{y}) = 0.$$

Of course, the constant function  $t \mapsto y(t) = \bar{y}$  is a solution to [1.10] with the initial condition  $y(t_{\text{Init}}) = \bar{y}$ . We can now study the *stability* of this particular solution: if we choose an initial value for [1.10] that is close to  $\bar{y}$ , does the corresponding solution remain close to  $\bar{y}$ ? Does it tend to  $\bar{y}$  when  $t \rightarrow \infty$ ?

The answer to this question lies in linearizing equation [1.10]: we can write  $y(t) = \bar{y} + \tilde{y}(t)$ , where the perturbation is assumed to be “small”. We thus have

$$\frac{d}{dt}\tilde{y}(t) = f(\bar{y} + \tilde{y}(t)) = 0 + \nabla f(\bar{y})\tilde{y}(t) + R(t),$$

where  $\nabla f(\bar{y})$  is the Jacobian matrix of  $f$  at  $\bar{y}$ , whose coefficients are  $\partial_{y_j} f_i(\bar{y})$  and the remainder term  $R(t)$  satisfies  $\frac{|R(t)|}{|\tilde{y}(t)|} \rightarrow 0$  when  $|\tilde{y}(t)| \rightarrow 0$ . Therefore, locally, within

a neighborhood of the equilibrium point  $\bar{y}$ , the dynamic is described by the behavior of solutions of the *linear* system

$$\frac{d}{dt}z(t) = Az(t), \quad A = \nabla f(\bar{y}).$$

If all the eigenvalues of  $A$  have a strictly negative real component, we have  $\lim_{t \rightarrow +\infty} |z(t)| = 0$  and the equilibrium point  $\bar{y}$  is said to be *locally stable*. If there are eigenvalues with a strictly positive real component, there are unstable trajectories that depart from the point of equilibrium.

#### 1.1.4. Higher-order differential equations

Differential equations of order  $n > 1$  can be converted into first-order equations of the form [1.1]. Indeed, if we have

$$u^{(n)}(t) = \Phi(t, u(t), u'(t), \dots, u^{(n-1)}(t)),$$

and given  $(u, u', \dots, u^{(n-1)})(t = t_{\text{Init}})$ , then we set  $y(t) = (u(t), u'(t), \dots, u^{(n-1)}(t)) = (y_0, \dots, y_{n-1}) \in (\mathbb{R}^D)^n$ , which satisfies the equation

$$\begin{aligned} \frac{d}{dt}y(t) &= \frac{d}{dt} \begin{pmatrix} y_0 \\ y_1 \\ \vdots \\ y_{n-2} \\ y_{n-1} \end{pmatrix}(t) = \begin{pmatrix} u'(t) \\ u''(t) \\ \vdots \\ u^{(n-1)}(t) \\ u^{(n)}(t) \end{pmatrix} = \begin{pmatrix} y_1(t) \\ y_2(t) \\ \vdots \\ y_{n-1}(t) \\ \Phi(t, y(t)) \end{pmatrix} \\ &= \begin{pmatrix} 0 & \mathbb{I} & 0 & \cdots & 0 \\ \vdots & & & & \vdots \\ 0 & & 0 & \cdots & 0 \\ & & & \ddots & 0 \\ & & & & \mathbb{I} \end{pmatrix} y(t) + \begin{pmatrix} 0 \\ \vdots \\ 0 \\ \Phi(t, y(t)) \end{pmatrix} = f(t, y(t)) \end{aligned}$$

on which we can apply the Picard–Lindelöf theorem, with the function  $f$  defined by

$$f : (t, y) \in \mathbb{R} \times (\mathbb{R}^D)^n \mapsto \begin{pmatrix} 0 & \mathbb{I} & 0 & \cdots & 0 \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ 0 & \cdots & 0 & \mathbb{I} & 0 \\ 0 & \cdots & 0 & 0 & \Phi(t, y) \end{pmatrix} y + \begin{pmatrix} 0 \\ \vdots \\ 0 \\ \Phi(t, y) \end{pmatrix}.$$

### 1.1.5. Inverse function theorem and implicit function theorem

Let us finish this review of differential calculus with some analysis tools that will prove to be useful in studying the behavior of numerical schemes that allows us to find approximations for the solutions of [1.1].

**THEOREM 1.8** (Inverse Function Theorem).— Let  $\Omega \subset \mathbb{R}^D$  be an open set and  $f : \Omega \rightarrow \mathbb{R}^D$  a function of class  $C^1$ . Let  $\bar{x} \in \Omega$ , such that the Jacobian Matrix  $\nabla f(\bar{x}) \in \mathcal{M}_D(\mathbb{R})$  is invertible. There exists an open neighborhood  $\mathcal{V} \subset \mathbb{R}^D$  of  $\bar{x}$  and an open neighborhood  $\mathcal{W} \subset \mathbb{R}^D$  of  $f(\bar{x})$ , such that

- $f|_{\mathcal{V}}$  is a bijection of  $\mathcal{V}$  in  $\mathcal{W}$ ;
- the inverse mapping  $x \mapsto f^{-1}(x)$  is continuous on  $\mathcal{W}$ ;
- $f^{-1}$  is, in fact, a function of class  $C^1$  on  $\mathcal{W}$  and  $\nabla(f^{-1})(f(x)) = [\nabla f(x)]^{-1}$ .

**PROOF.**— Recall that, for  $f : x = (x_1, \dots, x_D) \in \mathbb{R}^D \mapsto f(x) = (f_1(x), \dots, f_D(x))$ , the Jacobian matrix of  $f$  evaluated at point  $x$ , denoted as  $\nabla f(x) \in \mathcal{M}_D(\mathbb{R})$ , is defined as

$$\nabla f(x) = \begin{pmatrix} \partial_{x_1} f_1(x) & \partial_{x_2} f_1(x) & \cdots & \partial_{x_D} f_1(x) \\ \partial_{x_1} f_2(x) & \partial_{x_2} f_2(x) & \cdots & \partial_{x_D} f_2(x) \\ \vdots & \vdots & & \vdots \\ \partial_{x_1} f_D(x) & \partial_{x_2} f_D(x) & \cdots & \partial_{x_D} f_D(x) \end{pmatrix}.$$

This matrix identifies the derivative of  $f$  at point  $x$  as a linear mapping on  $\mathbb{R}^D$  because

$$f(x+h) = f(x) + \nabla f(x)h + |h|\epsilon(h), \quad \text{where } \lim_{|h| \rightarrow 0} \epsilon(h) = 0.$$

Even if it means replacing  $f$  by the mapping  $x \mapsto [\nabla f(\bar{x})]^{-1}(f(\bar{x} + x) - f(\bar{x}))$ , we may assume that  $\bar{x} = 0$ ,  $f(0) = 0$  and  $\nabla f(0) = \mathbb{I}$ . Since  $f$  is a function of class  $C^1$  on the open set  $\Omega$ , there exists an  $r > 0$ , such that the ball  $B(0, r)$  is included in  $\Omega$  and for every  $x \in B(0, r)$ , we have  $\|\nabla f(x) - \nabla f(0)\| = \|\nabla f(x) - \mathbb{I}\| \leq 1/2$ , where  $\|A\| = \sup \{|Ax|, x \in \mathbb{R}^D, |x| = 1\}$  denotes the norm of a matrix  $A$ . For  $x \in B(0, r)$ , we can therefore write  $\nabla f(x) = \mathbb{I} - u(x)$ , where  $\|u(x)\| \leq 1/2$ , so that  $\nabla f(x)$  is invertible, with  $[\nabla f(x)]^{-1} = \sum_{n=0}^{\infty} u(x)^n$ , whose series is normally convergent. We also note that for every  $x \in B(0, r)$ ,

$$\|[\nabla f(x)]^{-1}\| \leq \sum_{n=0}^{\infty} \|u(x)\|^n \leq \sum_{n=0}^{\infty} \frac{1}{2^n} = 2.$$

The classic proof of the inverse function theorem relies on Banach's fixed-point theorem: we will demonstrate the locally bijective nature of  $f$  through a fixed point argument. We choose a point  $y$  in  $B(0, r/2)$  and introduce the mapping

$$\Phi : x \mapsto y + x - f(x).$$

This mapping is  $C^1$  on  $B(0, r)$  and satisfies  $\|\nabla \Phi(x)\| = \|\mathbb{I} - \nabla f(x)\| \leq 1/2$  for every  $x \in B(0, r)$ . On  $B(0, r)$ , it follows that, on the one hand,

$$|\Phi(x_1) - \Phi(x_2)| \leq \frac{1}{2}|x_1 - x_2|,$$

and on the other hand,

$$\begin{aligned} |\Phi(x)| &= |y + x - f(x)| \leq |y| + |x - f(x)| \\ &= |y| + |\Phi(0) - \Phi(x)| \leq |y| + \frac{|x|}{2} \leq r/2 + r/2 = r. \end{aligned}$$

Therefore,  $\Phi$  is a contraction of  $B(0, r)$  in  $B(0, r)$ . It has a single fixed point, denoted as  $x_*$ . We have  $\Phi(x_*) = x_* = y + x_* - f(x_*)$ , so  $f(x_*) = y$ . We denote the intersection of the reciprocal image of  $B(0, r/2)$  through  $f$  – that is to say, the set  $\{x \in \Omega, f(x) \in B(0, r/2)\}$  – with  $B(0, r)$  as  $\mathcal{V}$ . Since  $f(0) = 0$  and  $f$  is continuous,  $\mathcal{V}$  is an open set containing 0, which shows that the mapping  $f$  is a bijection of  $\mathcal{V}$  in  $\mathcal{W} = B(0, r/2)$ . We define  $\Psi : x \mapsto x - f(x)$ . For all  $x_1, x_2 \in B(0, r)$ , we have

$$\begin{aligned} |x_1 - x_2| &= |\Psi(x_1) + f(x_1) - \Psi(x_2) - f(x_2)| \\ &\leq |\Psi(x_1) - \Psi(x_2)| + |f(x_1) - f(x_2)| \leq \frac{|x_1 - x_2|}{2} + |f(x_1) - f(x_2)|, \end{aligned}$$

because  $\Psi$  corresponds to the function  $\Phi$  that was previously studied with  $y = 0$ . It follows that  $|x_1 - x_2| \leq 2|f(x_1) - f(x_2)|$  and therefore  $|f^{-1}(x_1) - f^{-1}(x_2)| \leq$

$2|x_1 - x_2|$ , which proves that  $f^{-1}$  is continuous. We now identify the derivative of  $f^{-1}$ . Let  $x \in \mathcal{V}$  and  $y = f(x) \in \mathcal{W}$ . Let  $z \in \mathbb{R}^D$ , such that  $y + z \in \mathcal{W}$ . We define  $h = f^{-1}(y+z) - f^{-1}(y) = f^{-1}(f(x)+z) - x$ , which means that  $f(x+h) = f(x)+z$ . Note that  $|h| \leq 2|z|$ . We have

$$\begin{aligned} |f^{-1}(y+z) - f^{-1}(y) - [\nabla f(x)]^{-1}z| &= |h - [\nabla f(x)]^{-1}(f(x+h) - f(x))| \\ &= | - [\nabla f(x)]^{-1}(f(x+h) - f(x) - \nabla f(x)h)| \\ &\leq 2|f(x+h) - f(x) - \nabla f(x)h| \\ &\leq 2|h| |\epsilon(h)| \quad \text{where } \lim_{|h| \rightarrow 0} \epsilon(h) = 0. \end{aligned}$$

We rewrite the remainder term as a function of the increment of  $z$  by writing  $\eta(z) = \epsilon(f^{-1}(y+z) - f^{-1}(y)) = \epsilon(h)$ . By the continuity of  $f^{-1}$ , we still have  $\lim_{|z| \rightarrow 0} \eta(z) = 0$ . This inequality thus becomes

$$|f^{-1}(y+z) - f^{-1}(y) - [\nabla f(x)]^{-1}z| \leq 4|z| |\eta(z)|,$$

which proves that  $\nabla f^{-1}(y) = [\nabla f(x)]^{-1} = [\nabla f(f^{-1}(y))]^{-1}$ . (In dimension  $D = 1$ , the formula  $(f^{-1})'(y) = \frac{1}{f'(f^{-1}(y))}$  is easily recognizable.)  $\square$

NOTE.– It is possible to relax the regularity assumption by assuming that  $f$  is merely differentiable, with  $\nabla f(x)$  invertible for all  $x \in \Omega$ , irrespective of the continuity of its derivative. In the  $C^1$  case, Banach's fixed-point theorem directly guaranteed the existence and uniqueness of the solution  $z$  to the equation  $f(z) = y$ , for a fixed  $y$ . The continuity of  $x \mapsto \nabla f(x)$  stands as the key argument in constructing a suitable contraction. In fact, the surjective nature of  $f$  is not problematic. Indeed, suppose that  $f(0) = 0$  and  $\nabla f(0) = \mathbb{I}$ . Then, by the definition of differentiability,  $f(x) - x = f(x) - f(0) - \nabla f(0)x = |x|\epsilon(x)$ , with  $\lim_{|x| \rightarrow 0} \epsilon(x) = 0$ . So, there exists an  $r > 0$ , such that for every  $|x| \leq r$ , we have  $|\epsilon(x)| \leq 1/2$  and therefore  $|f(x) - x| \leq r/2$ . It follows that for every  $|y| \leq r/2$ , the function  $\Phi : x \mapsto y + x - f(x)$  satisfies  $|\Phi(x)| \leq |y| + |x - f(x)| \leq r$  for every  $x \in B(0, r)$ . Therefore,  $\Phi$  is a continuous function that maps the ball  $B(0, r)$  onto itself. By Brouwer's theorem (see Appendix 4, theorem A4.1),  $\Phi$  has a fixed point  $x_*$  in  $B(0, r)$ , which is the solution of  $f(x_*) = y$ . Demonstrating that  $f$  is also injective is the more subtle point. In dimension  $D = 1$ , the conclusion follows from a simple application of Rolle's theorem. We introduce the differentiable function  $\psi : x \mapsto \psi(x) = |f(x) - y|^2$ . If  $\psi(x_1) = \psi(x_2) = 0$ , then there exists a  $u \in ]x_1, x_2[$ , such that  $\psi$  is maximal in  $u$  and  $\psi'(u) = 0 = 2(f(u) - y)f'(u)$ , which contradicts the assumption that  $f'$  does not vanish. This approach can be generalized to every dimension [SAI 02].

**COROLLARY 1.4 (Implicit Function Theorem).**— Let  $\Omega$  be an open set of  $\mathbb{R}^P \times \mathbb{R}^Q$ . Let  $f : (x, y) \in \Omega \mapsto f(x, y) \in \mathbb{R}^Q$  be a  $C^1$  function. Let  $J(x, y)$  denote the matrix of partial derivatives of  $f$  with respect to the variable  $y \in \mathbb{R}^Q$  evaluated at  $(x, y) \in \Omega$  (with components  $\partial_{y_j} f_i(x, y)$ , for  $i, j \in \{1, \dots, Q\}$ ). If  $J(\bar{x}, \bar{y})$  is invertible, then there exists

- an open set  $\mathcal{U} \subset \mathbb{R}^P$  containing  $\bar{x}$ , an open set  $\mathcal{V} \subset \mathbb{R}^Q$  containing  $f(\bar{x}, \bar{y})$ , an open set  $\mathcal{W} \subset \Omega \subset \mathbb{R}^P \times \mathbb{R}^Q$  containing  $(\bar{x}, \bar{y})$ ,
- and a mapping  $\varphi : \mathcal{U} \times \mathcal{V} \rightarrow \mathbb{R}^Q$  of class  $C^1$ ,

such that for every  $x \in \mathcal{U}$ ,  $z \in \mathcal{V}$ ,  $\varphi(x, z)$  is the only solution  $y$  of the equation  $f(x, y) = z$ , with  $(x, y) \in \mathcal{W}$ . In particular, for all  $x \in \mathcal{U}$ ,  $z \in \mathcal{V}$ , we have  $f(x, \varphi(x, z)) = z$ .

**PROOF.**— We apply the inverse function theorem to the mapping defined by

$$\begin{aligned}\Phi : \mathcal{U} \times \mathcal{V} &\longrightarrow \mathbb{R}^P \times \mathbb{R}^Q \\ (x, y) &\longmapsto (x, f(x, y)).\end{aligned}$$

This mapping is indeed of class  $C^1$  and its derivative, which is a continuous linear mapping of  $\mathbb{R}^P \times \mathbb{R}^Q$  onto itself, is associated with the matrix

$$\nabla \Phi(x, y) = \begin{pmatrix} \mathbb{I} & 0 \\ \tilde{J}(x, y) & J(x, y) \end{pmatrix}$$

where  $\mathbb{I}$  is the identity matrix in  $\mathbb{R}^P$  and  $\tilde{J}(x, y)$  designates the matrix of  $f$ 's partial derivatives with respect to the variable  $x \in \mathbb{R}^P$  evaluated at  $(x, y) \in \Omega$ , which has  $Q$  rows and  $P$  columns. This matrix is invertible in  $(\bar{x}, \bar{y})$ , and its inverse can be expressed in the following form:

$$\begin{pmatrix} h \\ k \end{pmatrix} \longmapsto \begin{pmatrix} h \\ J(\bar{x}, \bar{y})^{-1}(h - \tilde{J}(\bar{x}, \bar{y})k) \end{pmatrix}.$$

The inverse function theorem guarantees the existence of an open set  $\mathcal{W}$  containing  $(\bar{x}, \bar{y})$  as well as an open set  $\tilde{\mathcal{W}}$  containing  $(\bar{x}, f(\bar{x}, \bar{y}))$ , such that  $\Phi$  is a class  $C^1$  diffeomorphism of  $\mathcal{W}$  on  $\tilde{\mathcal{W}}$ . Since the first component of  $\Phi(x, y)$  is simply  $x$ , the inverse mapping is of the form  $\Phi^{-1}(x, z) = (x, \varphi(x, z))$ .  $\square$

**COROLLARY 1.5 (Constrained optimization, Lagrange Multipliers).**— Let  $\Omega$  be an open set of  $\mathbb{R}^D$ . Let  $f$  and  $g_1, \dots, g_Q$  be  $C^1$  functions defined on  $\Omega$  with values in  $\mathbb{R}$ . We define the set

$$\mathcal{C} = \{x \in \Omega, g_1(x) = \dots = g_Q(x) = 0\}.$$

Suppose that  $f$  has a local extremum on  $\mathcal{C}$  in  $\bar{x} \in \mathcal{C}$  and that the family  $(\nabla g_1(\bar{x}), \dots, \nabla g_Q(\bar{x}))$  is linearly independent in  $\mathbb{R}^D$  (in particular,  $Q \leq D$ ). Then, there exists real numbers  $\lambda_1, \dots, \lambda_Q$ , called *Lagrange multipliers*, such that

$$\nabla f(\bar{x}) - \sum_{k=1}^Q \lambda_k \nabla g_k(\bar{x}) = 0.$$

PROOF.– The assumption about the  $g_k$ 's can be reformulated by saying that the matrix with  $D$  columns and  $Q$  rows

$$\begin{pmatrix} \partial_{x_1} g_1(\bar{x}) & \partial_{x_2} g_1(\bar{x}) & \dots & \partial_{x_D} g_1(\bar{x}) \\ \partial_{x_1} g_2(\bar{x}) & \partial_{x_2} g_2(\bar{x}) & \dots & \partial_{x_D} g_2(\bar{x}) \\ \vdots & \vdots & & \vdots \\ \partial_{x_1} g_Q(\bar{x}) & \partial_{x_2} g_Q(\bar{x}) & \dots & \partial_{x_D} g_Q(\bar{x}) \end{pmatrix}$$

is of rank  $Q$ . We can therefore extract an invertible  $Q \times Q$  submatrix. We let  $P = D - Q \geq 0$ . Even though it may involve renumbering the variables, we may suppose that

$$\Gamma = \begin{pmatrix} \partial_{x_{P+1}} g_1(\bar{x}) & \partial_{x_{P+2}} g_1(\bar{x}) & \dots & \partial_{x_{P+Q}} g_1(\bar{x}) \\ \partial_{x_{P+1}} g_2(\bar{x}) & \partial_{x_{P+2}} g_2(\bar{x}) & \dots & \partial_{x_{P+Q}} g_2(\bar{x}) \\ \vdots & \vdots & & \vdots \\ \partial_{x_{P+1}} g_Q(\bar{x}) & \partial_{x_{P+2}} g_Q(\bar{x}) & \dots & \partial_{x_{P+Q}} g_Q(\bar{x}) \end{pmatrix}$$

is invertible and we decompose  $x \in \mathbb{R}^D$  in the form  $(\hat{x}, y) \in \mathbb{R}^P \times \mathbb{R}^Q$ . By theorem 1.4 applied to the function

$$G : (\hat{x}, y) \in \mathbb{R}^P \times \mathbb{R}^Q \mapsto (g_1(\hat{x}, y), \dots, g_Q(\hat{x}, y)) \in \mathbb{R}^Q,$$

there exists an open set  $\mathcal{U}$  of  $\mathbb{R}^P$  containing  $\hat{x}$  and a mapping

$$\varphi : \hat{x} \in \mathcal{U} \mapsto \varphi(\hat{x}) = (\varphi_1(\hat{x}), \dots, \varphi_Q(\hat{x})) \in \mathbb{R}^Q$$

such that  $g_1(\hat{x}, y) = \dots = g_P(\hat{x}, y) = 0$ , with  $\hat{x} \in \mathcal{U}$ , if and only if  $y = \varphi(\hat{x})$ . Moreover, we have  $\bar{y} = \varphi(\hat{x})$ . Theorem 1.4 applies because the matrix  $\Gamma$  corresponds to the matrix of partial derivatives of  $(\hat{x}, y) \mapsto (g_1(\hat{x}, y), \dots, g_Q(\hat{x}, y))$  with respect to

the variables  $y \in \mathbb{R}^Q$ . We define the following function:  $H : \hat{x} \in \mathcal{U} \mapsto f(\hat{x}, \varphi(\hat{x})) \in \mathbb{R}$ . Since  $(\hat{x}, \varphi(\hat{x})) \in \mathcal{C}$ , by assumption, the function  $H$ , which is  $C^1$  on the open set  $\mathcal{U}$ , has an extremum in  $\hat{x}$ . It follows that

$$\nabla_{\hat{x}} H(\hat{x}) = 0 = \nabla_{\hat{x}} f(\hat{x}, \bar{y}) + \Phi(\hat{x})^T \nabla_y f(\hat{x}, \bar{y}),$$

where  $\Phi$  is the matrix-valued function<sup>4</sup> defined by

$$\hat{x} \longmapsto \Phi(\hat{x}) = \begin{pmatrix} \partial_{\hat{x}_1} \varphi_1(\hat{x}) & \partial_{\hat{x}_2} \varphi_1(\hat{x}) & \dots & \partial_{\hat{x}_P} \varphi_1(\hat{x}) \\ \partial_{\hat{x}_1} \varphi_2(\hat{x}) & \partial_{\hat{x}_2} \varphi_2(\hat{x}) & \dots & \partial_{\hat{x}_P} \varphi_2(\hat{x}) \\ \vdots & \vdots & & \vdots \\ \partial_{\hat{x}_1} \varphi_Q(\hat{x}) & \partial_{\hat{x}_2} \varphi_Q(\hat{x}) & \dots & \partial_{\hat{x}_P} \varphi_Q(\hat{x}) \end{pmatrix}.$$

Now, for all  $k \in \{1, \dots, Q\}$  and  $\hat{x} \in \mathcal{U}$ , we have  $g_k(\hat{x}, \varphi(\hat{x})) = 0$ , so the derivatives with respect to  $\hat{x}$  satisfy

$$\nabla_{\hat{x}} [g_k(\hat{x}, \varphi(\hat{x}))] = 0 = (\nabla_{\hat{x}} g_k)(\hat{x}, \varphi(\hat{x})) + \Phi(\hat{x})^T (\nabla_y g_k)(\hat{x}, \varphi(\hat{x})).$$

We can deduce that the matrix with  $Q + 1$  rows and  $D$  columns,

$$\begin{pmatrix} \partial_{\hat{x}_1} f(\bar{x}) & \partial_{\hat{x}_2} f(\bar{x}) & \dots & \partial_{\hat{x}_P} f(\bar{x}) & \partial_{y_1} f(\bar{x}) & \dots & \partial_{y_Q} f(\bar{x}) \\ \partial_{\hat{x}_1} g_1(\bar{x}) & \partial_{\hat{x}_2} g_1(\bar{x}) & \dots & \partial_{\hat{x}_P} g_1(\bar{x}) & \partial_{y_1} g_1(\bar{x}) & \dots & \partial_{y_Q} g_1(\bar{x}) \\ \vdots & \vdots & & \vdots & & & \vdots \\ \partial_{\hat{x}_1} g_Q(\bar{x}) & \partial_{\hat{x}_2} g_Q(\bar{x}) & \dots & \partial_{\hat{x}_P} g_Q(\bar{x}) & \partial_{y_1} g_Q(\bar{x}) & \dots & \partial_{y_Q} g_Q(\bar{x}) \end{pmatrix}$$

has a rank of at most  $Q$  because these relations show that the first  $P$  columns are expressed as linear combinations of the last  $Q$ . We conclude<sup>5</sup> that the  $Q + 1$  lines of this matrix constitute a linearly dependent family in  $\mathbb{R}^D$ . Therefore, there exists  $(\mu_0, \mu_1, \dots, \mu_Q) \in \mathbb{R}^{Q+1} \setminus \{0\}$ , such that  $\mu_0 \nabla f(\bar{x}) + \sum_{k=1}^Q \mu_k \nabla g_k(\bar{x}) = 0$ . Since  $(\nabla g_1(\bar{x}), \dots, \nabla g_Q(\bar{x}))$  is free in  $\mathbb{R}^D$ , we necessarily have  $\mu_0 \neq 0$ . Finally, we let  $\lambda_k = -\frac{\mu_k}{\mu_0}$  to obtain the desired result.  $\square$

<sup>4</sup> It is important to properly understand this formula and to be able to pass from the matrix formulation to its formulation in coordinates  $\partial_{\hat{x}_j} H(\hat{x}) = 0 = (\partial_{\hat{x}_j} f)(\hat{x}, \varphi(\hat{x})) + \sum_{\ell=1}^Q (\partial_{y_\ell} f)(\hat{x}, \varphi(\hat{x})) \partial_{\hat{x}_j} \varphi_\ell(\hat{x})$  and vice versa.

<sup>5</sup> For example, by emphasizing the fact that a matrix and its transpose have the same rank.

## 1.2. Numerical simulation of ordinary differential equations, Euler schemes, notions of convergence, consistency and stability

### 1.2.1. Introduction

Before beginning, it is worthwhile to recall the stakes of numerical simulation of [1.1]. Even though theorem 1.1 guarantees the existence and uniqueness of solutions, and we have some qualitative information about the solution, as well as some estimates, mostly, formulas relating the time  $t$  and the solution  $y$  of [1.1] are not known at this moment. The main problem lies in finding a sequence of operations that can be efficiently performed by a computer, which can produce an *approximation* of the solution. This approximation procedure is of an essentially *discrete* nature, and numerical analysis precisely involves the study of the relation between discrete and continuous descriptions of the solution. More specifically, for the sake of simplicity, supposing that  $t_{\text{Init}} = 0$ , we seek to approach the solution  $t \mapsto y(t)$  of [1.1] on a fixed time interval  $[0, T]$ . We introduce a subdivision in that interval  $[0, T]$ :

$$0 = t_0 < t_1 < \dots < t_N = T$$

for a given integer  $N \neq 0$ . For  $n \in \{1, \dots, N\}$ , we can write

$$\Delta t_n = t_n - t_{n-1},$$

and

$$\Delta t = \max \{\Delta t_n, \quad n \in \{1, \dots, N\}\} > 0.$$

Next, we limit the discussion to the case of a uniform subdivision, where

$$\Delta t = t_n - t_{n-1} > 0 \quad \text{for all } n \in \{1, \dots, N\}.$$

This seemingly obvious remark is fundamental, the final time  $T$ , the interval  $\Delta t$  and the number  $N$  of intervals that form the subdivision are related by

$$N \times \Delta t = T.$$

In what follows we write  $N = N_{\Delta t}$  to clearly denote this dependence. Given the initial value  $y_0 = y_{\text{Init}} \in \mathbb{R}^D$  in the state space, a *numerical scheme* defines states  $y_1, \dots, y_{N_{\Delta t}}$  that are explicitly calculable and that are supposed to be approximations to the solution at the discrete points  $y(t_1), \dots, y(t_{N_{\Delta t}})$ . It is therefore necessary to distinguish  $y(t_n)$ , evaluation of the function  $t \mapsto y(t)$ , which is the solution of [1.1] at a time  $t = t_n$  from  $y_n$ , the  $n^{\text{th}}$  iterated term of a sequence that is constructed to approximate the solution.

The construction of the approximation is based on the integrated form of the equation

$$y(t_{n+1}) = y(t_n) + \int_{t_n}^{t_{n+1}} f(s, y(s)) \, ds,$$

which is satisfied for  $n \in \{0, \dots, N - 1\}$  and, in fact, used to prove theorem 1.1. The idea is to use formulas from numerical integration, some aspects of which are reviewed in section Appendix 2. The simplest approach consists of the “left rectangle rule”. We thus obtain the recurrence formula

$$y_{n+1} = y_n + \Delta t \, f(t_n, y_n). \quad [1.11]$$

This is known as the *explicit Euler method*. If the initial value  $y_0 = y(t_0) = y_{\text{Init}}$  is known, this formula effectively defines the elements  $y_1, \dots, y_{N_{\Delta t}}$ . It now becomes necessary to determine its relation to the solution of [1.1] and to know if it indeed provides an approximation of  $y(t_1), \dots, y(T)$ . It is also possible to assume a construction based on the right rectangle rule, which gives

$$y_{n+1} = y_n + \Delta t \, f(t_{n+1}, y_{n+1}). \quad [1.12]$$

This is known as the *implicit Euler method*. However, this method has one difficulty: the formula does not allow us to directly calculate  $y_{n+1}$  simply by knowing the previous iteration  $y_n$ . In order to determine  $y_{n+1}$ , it is necessary to solve an equation. More specifically,  $y_{n+1}$  is a zero for the mapping

$$z \mapsto \Phi(z) = z - y_n - \Delta t \, f(t_{n+1}, z).$$

Therefore, in order to obtain a numerical value for  $y_{n+1}$ , it would be necessary to operate a root-finding method for the mapping  $z \mapsto \Phi(z)$  (and, in fact, in that case we would only use an approximation of  $y_{n+1}$ , the solution of [1.12]). Nevertheless, we will see that, despite its undeniably greater complexity, the method [1.12] also has advantages with respect to method [1.11] in some cases.

One variant of these methods is the *predictor–corrector method*, which is defined by the following relations: given  $y_n$ , we can write

$$y^* = y_n + \frac{\Delta t}{2} f(t_n, y_n),$$

$$y_{n+1} = y_n + \Delta t f(t_n + \Delta t/2, y^*) = y_n + \Delta t f\left(t_n + \Delta t/2, y_n + \frac{\Delta t}{2} f(t_n, y_n)\right).$$

This corresponds to an approximation of  $\int_{t_n}^{t_n + \Delta t} f(s, y(s)) ds$  using a “midpoint rule”. However, since we do not know the value of  $y(t_n + \Delta t/2)$ , we replace it with its approximation  $y^*$  obtained by using the explicit Euler method.

Before going into the analysis of these methods, we highlight two important points:

- As mentioned above, numerical approximations of solutions to differential equations and the development of mathematical methods adapted to these questions are justified by the fact that explicit expressions of these solutions in terms of the time variable are not generally known. Nevertheless, it is always useful to evaluate the performance of numerical methods by testing them on problems whose explicit solutions are known. This experimental confirmation ensures the robustness of a given numerical analysis.

- In the following sections, we will attempt to estimate the errors  $y_n - y(t_n)$ . These estimates involve the solution and possibly some of its derivatives. These calculations assume that the solution is “fairly regular”, possibly even beyond the simple  $C^1$  regularity provided by theorem 1.1.

### 1.2.2. Fundamental notions for the analysis of numerical ODE methods

In order to analyze the behavior of these numerical methods and justify that they indeed define an appropriate approximation of the solution to [1.1], we will now introduce a certain number of definitions that will allow us to formalize what we may refer to as the “properties” of a numerical approximation. Recall that the working interval  $[0, T]$  is subdivided into intervals separated by

$$\Delta t = \max \{t_n - t_{n-1}, n \in \{1, \dots, N_{\Delta t} = T/\Delta t\}\}.$$

Given initial data  $y_0 = y_{\text{Init}}$ , a numerical scheme defines the points  $y_1, \dots, y_{N_{\Delta t}}$  with a relation of the form

$$y_{n+1} = y_n + (t_{n+1} - t_n) \Phi(t_n, \Delta t, y_n) = y_n + \Delta t \Phi(t_n, \Delta t, y_n) \quad [1.13]$$

for  $n \in \{0, \dots, N_{\Delta t} - 1\}$ . We assume throughout the discussion that the function  $\Phi$  depends on those arguments in a regular manner. With this notation, for the explicit Euler scheme [1.11], we simply have  $\Phi(t, h, y) = f(t, y)$ . For the implicit Euler scheme [1.12], we have  $\Phi(t, h, y) = f(t + h, \phi(t, h, y))$ , where  $\phi(t, h, y)$  is the solution of the equation  $z - y_n - \Delta t f(t + h, z) = 0$ , which, for the moment, we assume that it exists and is unique. It will be easy to adapt the discussion to schemes with variable time steps. The principal notion is as follows: a scheme is said to be *convergent* whenever the elements of the sequence  $y_0, y_1, \dots, y_N$  “resemble” the evaluations of the solution  $y(0), y(t_1), \dots, y(t_N) = y(T)$ . More formally, we can state the following.

**DEFINITION 1.2.**— We say that the scheme [1.13] is *convergent* if, given the solution  $t \mapsto y(t)$ , of [1.1] associated with the initial value  $y_{\text{Init}} = y_0$ , we have

$$\lim_{\Delta t \rightarrow 0} \left( \sup_{n \in \{0, \dots, N_{\Delta t}\}} |y_n - y(t_n)| \right) = 0.$$

In order to show that a scheme is convergent, it is necessary to define new notions that characterize the “properties” of a numerical scheme. The notion of consistence evaluates the coherence of a scheme with the equation: if we replace the approximation  $y_n$  with the solution evaluated at a point  $t_n$  in [1.13], an error emerges, but it must tend to 0 when  $\Delta t \rightarrow 0$ , and sometimes it can be calculated precisely as a function of the step  $\Delta t$ . The notion of stability controls the accumulation of errors at each step in time, since, more specifically, the different  $y(t_n)$ ’s do not quite satisfy [1.13].

**DEFINITION 1.3.**— We say that the scheme [1.13] is *consistent* if, given the solution of [1.1] associated with an initial value  $y_{\text{Init}}$ ,  $t \mapsto y(t)$ , and

$$\epsilon_n = \frac{y(t_{n+1}) - y(t_n)}{\Delta t} - \Phi(t_n, \Delta t, y(t_n)),$$

we have

$$\lim_{\Delta t \rightarrow 0} \left( \Delta t \sum_{n=0}^{N_{\Delta t}-1} |\epsilon_n| \right) = 0.$$

We say the scheme [1.13] is (*consistent*) *of order p* if there exists a  $C > 0$ , independent of  $\Delta t$ , such that we have

$$\Delta t \sum_{n=0}^{N_{\Delta t}-1} |\epsilon_n| \leq C \Delta t^p.$$

**NOTE.**— If we introduce the piecewise constant function

$$\epsilon_{\Delta t} : t \in [0, T] \longmapsto \sum_{n=0}^{N_{\Delta t}-1} \epsilon_n \mathbf{1}_{n\Delta t \leq t < (n+1)\Delta t}$$

then the quantity to be studied may be interpreted as the  $L^1$  norm of this function

$$\Delta t \sum_{n=0}^{N_{\Delta t}-1} |\epsilon_n| = \int_0^T |\epsilon_{\Delta t}(t)| dt.$$

The quantity  $\epsilon_n$  that was introduced in definition 1.3 is called the *local consistency error*. In practice, this quantity can be controlled in a uniform manner; that is to say, to find a  $C > 0$ , such that for all  $0 < \Delta t \leq \Delta t_*$ , and every  $n \in \mathbb{N}$ , we have

$$|\epsilon_n| \leq C \Delta t^p.$$

An estimate of this kind implies that the scheme is of order  $p$ : the *global consistency error*, on the time interval  $[0, T]$  with  $T = N_{\Delta t} \Delta t$ , is controlled by

$$\Delta t \sum_{n=0}^{N_{\Delta t}-1} |\epsilon_n| \leq C N_{\Delta t} \Delta t (\Delta t)^p = CT(\Delta t)^p.$$

Importantly, in these estimates, the constant  $C$ , in general, depends on the  $L^\infty$  norms of the derivatives of the solution  $y$  on  $[0, T]$ ; therefore, it is also a function of the final time  $T$ . Determination of this kind of estimates also requires that the solution be regular beyond the mere class  $C^1$  required in the Picard–Lindelöf theorem. In fact, there exists a practical criterion for determining if a method is consistent.

**PROPOSITION 1.2.–** If  $\Phi(t, 0, y) = f(t, y)$  for all  $t$  and  $y$ , then [1.13] is consistent.

**PROOF.–** Recall that the step  $\Delta t > 0$  and the number of intervals required to reach the final time  $T$  are related by the relation  $T = \Delta t N_{\Delta t}$ . First, note that

$$\begin{aligned} \Delta t \sum_{n=0}^{N_{\Delta t}-1} |\epsilon_n| &= \Delta t \sum_{n=0}^{N_{\Delta t}-1} \left| \frac{y(t_n + \Delta t) - y(t_n)}{\Delta t} - \Phi(t_n, \Delta t, y(t_n)) \right| \\ &= \Delta t \sum_{n=0}^{N_{\Delta t}-1} \left| \int_0^1 (f(t_n + \tau \Delta t, y(t_n + \tau \Delta t)) - \Phi(t_n, \Delta t, y(t_n))) d\tau \right|. \end{aligned}$$

For  $t \in [0, T]$ , the solution  $t \mapsto y(t)$  remains within a compact domain  $K \subset \mathbb{R}^D$ . The functions

$$t \mapsto y(t), \quad (t, y) \mapsto f(t, y), \text{ and} \quad (t, h, y) \mapsto \Phi(t, h, y)$$

are continuous and therefore uniformly continuous on the compact domains  $[0, T]$ ,  $[0, T] \times K$  and  $[0, T] \times [0, 1] \times K$ , respectively. Let  $\alpha > 0$ . Then, there exists a  $\delta(\alpha) > 0$ , such that for all  $0 < h < \delta(\alpha)$ ,  $t \in [0, T]$  and  $\tau \in [0, 1]$ , we have

$$\begin{aligned} |f(t + h\tau, y(t + h\tau)) - \Phi(t, h, y(t))| &\leq |f(t + h\tau, y(t + h\tau)) - f(t, y(t))| \\ &\quad + |\Phi(t, 0, y(t)) - \Phi(t, h, y(t))| \leq \alpha. \end{aligned}$$

It follows that

$$\Delta t \sum_{n=0}^{N_{\Delta t}-1} |\epsilon_n| \leq \Delta t \sum_{n=0}^{N_{\Delta t}-1} \alpha = \alpha \Delta t N_{\Delta t} \leq T \alpha.$$

for all  $0 < \Delta t < \delta(\alpha)$  because  $N_{\Delta t} = T/\Delta t$ .  $\square$

**DEFINITION 1.4.**— The scheme [1.13] is said to be *stable* if, when considering  $y_1, \dots, y_{N_{\Delta t}}$  defined by [1.13] and the elements  $z_1, \dots, z_{N_{\Delta t}}$  defined by

$$z_{n+1} = z_n + \Delta t \Phi(t_n, \Delta t, z_n) + \eta_n$$

with given  $z_0$  and  $\eta_1, \dots, \eta_{N_{\Delta t}}$ , there exists an  $M > 0$ , independent of  $\Delta t$ , such that

$$|y_n - z_n| \leq M \left( |y_{\text{Init}} - z_0| + \sum_{j=0}^{n-1} |\eta_j| \right).$$

The combination of consistency and stability properties allows us to determine that a scheme is convergent.

**THEOREM 1.9.**— A scheme of the form [1.13] that is stable and consistent is convergent.

**PROOF.**— We will use the stability property with  $z_n = y(t_n)$ ,  $\eta_n = \Delta t \epsilon_n$ , such that

$$\max_{n \in \{0, \dots, N_{\Delta t}\}} |y_n - y(t_n)| \leq M \left( |y_0 - y(0)| + \Delta t \sum_{n=0}^{N_{\Delta t}-1} |\epsilon_n| \right)$$

is satisfied for some  $M > 0$ . Of course, we begin the scheme with the initial value prescribed for the ODE:  $y_0 = y(0)$ . The consistency of the scheme ensures that

$$\max_{n \in \{0, \dots, N_{\Delta t}\}} |y_n - y(t_n)| \xrightarrow{\Delta t \rightarrow 0} 0.$$

If the scheme has the order  $p$ , then

$$\max_{n \in \{0, \dots, N_{\Delta t}\}} |y_n - y(t_n)| \leq MCT \Delta t^p$$

(Here, it is important to recall that the estimate involves the final time  $T$ , even by means of the constant  $C$ ; in particular, note that for a given accuracy, the greater the final time  $T$ , the smaller the step  $\Delta t$  must be).  $\square$

### 1.2.3. Analysis of explicit and implicit Euler schemes

Consistency analysis involves successive derivatives of the solution  $t \mapsto y(t)$  of [1.1] and uses Taylor developments taken up to convenient order. Recall that given a second member  $f$  that is Lipschitz continuous in the state variable, the Picard–Lindelöf theorem ensures that the maximal solution has only  $C^1$  regularity. Therefore, consistency analysis should be accompanied by additional regularity assumptions.

**THEOREM 1.10** (Taylor's Formula With Integral Remainder).— Let  $t \mapsto y(t)$  be a  $C^n$  function, with  $n \geq 1$ . Then, for every  $k \in \{1, \dots, n\}$ , we have

$$y(t+h) = \sum_{\ell=0}^{k-1} y^{(\ell)}(t) \frac{h^\ell}{\ell!} + \int_0^1 y^{(k)}(t+sh) h^k \frac{(1-s)^{k-1}}{(k-1)!} ds.$$

**PROOF.**— This theorem can be proved very easily using induction and integration by parts. Indeed, for  $k = 1$ , it is true that

$$y(t+h) - y(t) = \int_0^1 \frac{d}{ds}(y(t+sh)) ds = \int_0^1 y'(t+sh) h ds.$$

Suppose that the relation holds for all  $k \in \{1, \dots, n-1\}$ . We integrate the remainder term using integration by parts

$$\begin{aligned} & \int_0^1 \underbrace{y^{(k)}(t+sh) h^k}_{:=F(s)} \underbrace{\frac{(1-s)^{k-1}}{(k-1)!}}_{:=g(s)=G'(s)} ds = \left[ \underbrace{y^{(k)}(t+sh) h^k}_{=F(s)} \underbrace{\frac{-(1-s)^k}{k!}}_{=G(s)} \right]_0^1 \\ & + \int_0^1 \underbrace{y^{(k+1)}(t+sh) h^{k+1}}_{=f(s)=F'(s)} \underbrace{\frac{(1-s)^k}{k!}}_{=-G(s)} ds \\ & = y^{(k)}(t) \frac{h^k}{k!} + \int_0^1 y^{(k+1)}(t+sh) h^{k+1} \frac{(1-s)^k}{k!} ds, \end{aligned}$$

This enables us to extend the formula to  $k+1$ . □

**LEMMA 1.2.**— Suppose that the solution  $y$  of [1.1] is a function of class  $C^2$  on  $[0, T]$ . The explicit and implicit Euler schemes are consistent of order 1.

**PROOF.**— For the explicit Euler scheme, by manipulating Taylor's formula, we obtain

$$\begin{aligned} & |y(t_n + \Delta t) - y(t_n) - \Delta t f(t_n, y(t_n))| \\ & = \left| y'(t_n) \Delta t + \int_0^1 y''(t_n + s\Delta t) \Delta t^2 (1-s) ds - \Delta t f(t_n, y(t_n)) \right| \leq \frac{\|y''\|_\infty}{2} \Delta t^2. \end{aligned}$$

An identical manipulation applies for the implicit scheme.  $\square$

**NOTE.–** Consider the case where the differential system is autonomous with  $f(y) = -\nabla\Phi(y)$  for some function  $\Phi : \mathbb{R}^D \rightarrow \mathbb{R}$ . We assume that

–  $\Phi$  is  $C^1$ ;

–  $\Phi$  is  $\alpha$ -convex: for every  $x, y \in \mathbb{R}^D$ , we have  $(\nabla\Phi(x) - \nabla\Phi(y)) \cdot (x - y) \geq \alpha|x - y|^2$ , with  $\alpha > 0$ ;

–  $\Phi$  is coercive:  $\lim_{|x| \rightarrow \infty} \Phi(x) = +\infty$ .

Under these assumptions, the function  $\Phi$  has a unique minimum on  $\mathbb{R}^D$ , denoted as  $\bar{y}$ . Indeed, the minimum is unique as a consequence of strict convexity (which itself is a product of the fact that  $\Phi$  is  $\alpha$ -convex): if  $\bar{y}$  and  $y_*$  satisfy  $\Phi(\bar{y}) = \Phi(y_*) = m = \min\{\Phi(y), y \in \mathbb{R}^D\}$ , then, for every  $0 \leq \lambda \leq 1$ , we have  $m \leq \Phi(\lambda\bar{y} + (1 - \lambda)y_*) < \lambda\Phi(\bar{y}) + (1 - \lambda)\Phi(y_*)$ , which is absurd because the right-hand term is equal to  $\lambda m + (1 - \lambda)m = m$ . To justify the existence of a minimum, consider a minimizing sequence  $(y_n)_{n \in \mathbb{N}}$ :  $\lim_{n \rightarrow \infty} \Phi(y_n) = \inf\{\Phi(y), y \in \mathbb{R}^D\}$ . Because  $\Phi$  is coercive, we can be sure that this sequence is bounded. Even if it is necessary to extract a subsequence, we may assume that  $\lim_{n \rightarrow \infty} y_n = \bar{y} \in \mathbb{R}^D$ . For every  $0 < \lambda \leq 1$  and all  $x, y \in \mathbb{R}^D$ , we have

$$\frac{\Phi(x + \lambda(y - x)) - \Phi(x)}{\lambda} = \frac{\Phi((1 - \lambda)x + \lambda y) - \Phi(x)}{\lambda} \leq \Phi(y) - \Phi(x).$$

Using this inequality with  $x = \bar{y}$ ,  $y = y_n$  and by making  $\lambda$  tend to 0, we obtain

$$\Phi(\bar{y}) + \nabla\Phi(\bar{y}) \cdot (y_n - \bar{y}) \leq \Phi(y_n)$$

since  $\lim_{\lambda \rightarrow 0} \frac{\Phi(x + \lambda(y - x)) - \Phi(x)}{\lambda} = \nabla\Phi(x) \cdot (y - x)$ . When  $n \rightarrow \infty$ , it follows that  $\Phi(\bar{y}) + 0 \leq m$ , and therefore the limit  $\bar{y}$  minimizes  $\Phi$ . Since the minimum of  $\Phi$  occurs at  $\bar{y}$ , we have  $\nabla\Phi(\bar{y}) = 0$ .

The asymptotic behavior of the differential system  $\frac{d}{dt}y = -\nabla\Phi(y)$  is given by

$$\lim_{t \rightarrow \infty} y(t) = \bar{y}.$$

In order to demonstrate this remarkable property, we calculate

$$\begin{aligned} \frac{d}{dt} \frac{|y(t) - \bar{y}|^2}{2} &= (y(t) - \bar{y}) \cdot (-\nabla\Phi(y(t))) \\ &= -(\nabla\Phi(y(t)) - \nabla\Phi(\bar{y})) \cdot (y(t) - \bar{y}) \leq -\alpha|y(t) - \bar{y}|^2, \end{aligned}$$

which implies the convergence of  $y(t)$  towards the equilibrium point  $\bar{y}$  at an exponential rate as time increases. The explicit Euler scheme for this problem corresponds exactly to the gradient descent algorithm with a fixed step  $\Delta t$ , which allows us to minimize the functional  $\Phi$  (see section 2.7).

**LEMMA 1.3.**— Suppose that the function  $f$  is  $L$ -Lipschitz continuous in the state variable. The explicit Euler scheme is stable, and the stability constant is  $e^{LT}$ .

**PROOF.**— We can write

$$L = \sup_{0 \leq t \leq T, y \neq z} \frac{|f(t, y) - f(t, z)|}{|y - z|}.$$

By reproducing the notation used in definition 1.3, we have

$$\begin{aligned} |z_{n+1} - y_{n+1}| &= |z_n - y_n + \Delta t(f(t_n, z_n) - f(t_n, y_n)) + \eta_n| \\ &\leq (1 + L\Delta t)|z_n - y_n| + |\eta_n|. \end{aligned}$$

The basic inequality  $1 + x \leq e^x$ , which is satisfied by every  $x \geq 0$ , gives

$$\begin{aligned} |z_{n+1} - y_{n+1}| &\leq e^{L\Delta t}|z_n - y_n| + |\eta_n| \\ &\leq e^{L\Delta t}(e^{L\Delta t}|z_{n-1} - y_{n-1}| + |\eta_{n-1}|) + |\eta_n| \\ &\leq \dots \leq e^{(n+1)L\Delta t}|z_0 - y_0| + \sum_{k=0}^n e^{L(n-k)\Delta t}|\eta_k| \\ &\leq e^{LT}\left(|z_0 - y_0| + \sum_{k=0}^n |\eta_k|\right). \end{aligned} \quad \square$$

**THEOREM 1.11.**— The explicit Euler scheme is stable and consistent of order 1. Therefore, it is convergent and the error between the approximate solution and the actual solution tends to 0 with  $\Delta t$ .

This statement ensures that the explicit Euler scheme actually does the work expected of it: it provides an approximation of the real solution, which gets better as time steps become smaller. Nevertheless, if we study the approximations in more detail, we realize that there are practical limitations that can be quite significant, so that they may even make it unrealistic to simulate certain problems using this method. Indeed, the stability constant is  $e^{LT}$ . When  $L \gg 1$ , in order to obtain a given degree of precision for a given final time  $T$ , it becomes necessary to choose a very small time step and therefore to perform a great deal of operations, which means a prohibitive computation time. (For example, suppose that  $L = 10^{43}$  and that we want a solution on the interval  $[0, 1]$  with a degree of precision of  $10^{-2}$ . Since the error is bounded by  $e^L \Delta t$ , it is necessary to take  $\Delta t = 10^{-2} e^{-10^{43}}$ !).

It is important to be wary of abuse of language assimilating these difficulties to a lack of stability: the explicit Euler method *is* stable. However, the stability estimate is too poor for the method to provide pertinent results, within realistic computational conditions. This discussion shows just how subtle the analysis of numerical schemes can be. We will see later that, in some cases, the implicit Euler scheme [1.12] can be more effective in terms of this stability criterion.

We will now see that the “numerical scheme” approach allows us to construct an alternative proof of the Picard–Lindelöf theorem and some of its generalizations. Suppose that

$$f \text{ is a continuous function on an open set } \mathcal{U} \subset \mathbb{R} \times \mathbb{R}^D. \quad [1.14]$$

Let  $(t_{\text{Init}}, y_{\text{Init}}) \in \mathcal{U}$ . We take  $T_0 > 0$  and  $r_0 > 0$  small enough for the cylinder  $\mathcal{C}_0 = [t_{\text{Init}}, t_{\text{Init}} + T_0] \times \overline{B}(y_{\text{Init}}, r_0) \subset \mathcal{U}$ . Since the function  $f$  is continuous and  $\mathcal{C}_0$  is a compact set, we can set

$$M_0 = \sup \{ |f(t, y)|, (t, y) \in \mathcal{C}_0 \} < \infty.$$

We let<sup>6</sup>  $T = \min(T_0, r_0/M_0)$ . For a fixed  $N \in \mathbb{N} \setminus \{0\}$ , we introduce a subdivision of  $[t_{\text{Init}}, t_{\text{Init}} + T]$ , with step  $h = T/N > 0$ , defined by

$$t_0 = t_{\text{Init}} < t_1 < \dots < t_N = t_{\text{Init}} + T, \quad t_{n+1} = t_n + h, \quad h = \frac{T}{N}.$$

Consider the vectors  $y_0, y_1, \dots, y_N$  defined by  $y_0 = y_{\text{Init}}$  and the recurrent relation defined by the explicit Euler scheme

$$y_{n+1} = y_n + hf(t_n, y_n).$$

We associate these vectors with the sequence of functions defined on  $[t_{\text{Init}}, t_{\text{Init}} + T]$  by

$$\begin{aligned} z^{(N)}(t) &= \sum_{n=0}^{N-1} (y_n + (t - t_n)f(t_n, y_n)) \mathbf{1}_{t_n \leq t \leq t_{n+1}} \\ &= \sum_{n=0}^{N-1} \left( y_n + \frac{(t - t_n)}{h} (y_{n+1} - y_n) \right) \mathbf{1}_{t_n \leq t \leq t_{n+1}}. \end{aligned}$$

---

<sup>6</sup> Note that if  $t$  has the homogeneity of time and  $y$  that of length, then  $f$  and  $M_0$  have the homogeneity of speed, whereas  $r_0$  is homogeneous to speed; therefore,  $r_0/M_0$  is homogeneous to time.

LEMMA 1.4.– The functions  $z^{(N)}$  are continuous on  $[t_{\text{Init}}, t_{\text{Init}} + T]$  and satisfy

$$z^{(N)}(t) \in \overline{B}(y_{\text{Init}}, r_0)$$

for every  $t \in [t_{\text{Init}}, t_{\text{Init}} + T]$  and  $N \in \mathbb{N} \setminus \{0\}$ .

PROOF.– The functions  $z^{(N)}$  are piecewise  $\mathbb{P}_1$ , with continuous connections. In particular, these functions are continuous and piecewise  $C^1$ . An immediate induction shows that for every  $n \in \{0, \dots, N\}$ , we have  $|y_n - y_0| \leq nhM_0$ . In light of the definition of  $h, T, r_0$  and  $M_0$ , this shows, in passing, that  $(t_n, y_n) \in \mathcal{C}_0$ . Therefore, if  $t_n \leq t \leq t_{n+1}$ ,  $n \in \{0, \dots, N-1\}$ , then

$$\begin{aligned} |z^{(N)}(t) - y_{\text{Init}}| &\leq |y_n - y_{\text{Init}} + (t - t_n)f(t_n, y_n)| \\ &\leq nhM_0 + (t - t_n)M_0 = (t - t_{\text{Init}})M_0 \leq TM_0 \leq r_0. \end{aligned}$$

It follows that  $(t, z^{(N)}(t)) \in \mathcal{C}_0 \subset \mathcal{U}$ .  $\square$

Moreover, the functions  $z^{(N)}$  are piecewise  $C^1$  and for everywhere other than points  $t_0, t_1, \dots, t_N$ , we have  $\frac{d}{dt}z^{(N)}(t) = \sum_{n=0}^{N-1} f(t_n, y_n) \mathbf{1}_{t_n < t < t_{n+1}}$ , which implies  $|\frac{d}{dt}z^{(N)}(t)| \leq M_0$  for almost every  $t \in [t_{\text{Init}}, t_{\text{Init}} + T]$ . We can therefore apply the Arzela–Ascoli theorem [GOU 11, theorem 7.49]: the sequence of functions  $\{t \mapsto z^{(N)}(t), N \in \mathbb{N} \setminus \{0\}\}$  forms a relatively compact set in  $C^0([t_{\text{Init}}, t_{\text{Init}} + T]; \mathbb{R}^D)$ , and there exists an extracted sequence that uniformly converges on  $[t_{\text{Init}}, t_{\text{Init}} + T]$  towards a continuous function  $z$ . We introduce the functions

$$f^{(N)}(t) = \sum_{n=0}^{N-1} f(t_n, y_n) \mathbf{1}_{t_n \leq t \leq t_{n+1}}.$$

Note that for every  $n \in \{0, \dots, N-1\}$  and  $t_n \leq t < t_{n+1}$ , we have

$$\begin{aligned} y_{\text{Init}} + \int_{t_{\text{Init}}}^t f^{(N)}(s) ds &= y_{\text{Init}} + \sum_{k=0}^{n-1} \int_{t_k}^{t_{k+1}} f^{(N)}(s) ds + \int_{t_n}^t f^{(N)}(s) ds \\ &= y_0 + \sum_{k=0}^{n-1} h f(t_k, y_k) + (t - t_n) f(t_n, y_n) \\ &= y_n + (t - t_n) f(t_n, y_n) = z^{(N)}(t), \end{aligned}$$

by using  $y_1 = y_0 + h f(t_0, y_0)$ ,  $y_2 = y_1 + h f(t_1, y_1) = y_0 + h f(t_0, y_0) + h f(t_1, y_1)$ . It follows that

$$z^{(N)}(t) = y_{\text{Init}} + \int_{t_{\text{Init}}}^t f(s, z^{(N)}(s)) ds + \int_{t_{\text{Init}}}^t (f^{(N)}(s) - f(s, z^{(N)}(s))) ds.$$

The function  $f$  is continuous and therefore uniformly continuous on the compact set  $\mathcal{C}_0$ : for every  $\epsilon > 0$ , there exists  $\eta > 0$ , such that for all  $(s_1, y_1) \in \mathcal{C}_0$  and  $(s_2, y_2) \in \mathcal{C}_0$ , if  $|s_1 - s_2| \leq \eta$  and  $|y_1 - y_2| \leq \eta$ , then  $|f(s_1, y_1) - f(s_2, y_2)| \leq \epsilon$ . Thus, since  $\epsilon > 0$  is fixed, whenever  $h < \eta$ , we obtain

$$\begin{aligned} & \left| \int_{t_{\text{Init}}}^t (f^{(N)}(s) - f(s, z^{(N)}(s))) \, ds \right| \\ & \leq \sum_{n=0}^{N-1} \int_{t_n}^{t_{n+1}} |f(t_n, z^{(N)}(s)) - f(s, z^{(N)}(s))| \, ds \end{aligned}$$

which is bounded from above by  $Nh\epsilon = T\epsilon$ . We apply this result by considering the extracted sequence  $(z^{(N_\ell)})_{\ell \in \mathbb{N}}$ , which converges uniformly towards  $z$  on  $[t_{\text{Init}}, t_{\text{Init}} + T]$ : by making  $\ell$  tend to  $+\infty$ , and by using the continuity of  $f$ , we obtain

$$z(t) = y_{\text{Init}} + \int_{t_{\text{Init}}}^t f(s, z(s)) \, ds. \quad [1.15]$$

This proves the Cauchy–Peano existence theorem. We now strengthen the regularity assumptions on  $f$  with respect to the state variable by assuming that the function is locally Lipschitz continuous.

**DEFINITION 1.5.–** We say that  $f$  is locally Lipschitz continuous in the state variable on the open set  $\mathcal{U} \subset \mathbb{R} \times \mathbb{R}^D$  if for every compact set  $K \subset \mathcal{U}$ , there exists a  $L_K > 0$ , such that for all  $s, y_1, y_2$  satisfying  $(s, y_1) \in K$  and  $(s, y_2) \in K$ , the following relation holds:

$$|f(s, y_1) - f(s, y_2)| \leq L_K |y_1 - y_2|.$$

Under this assumption, and retaining the preceding paragraph's notation, we obtain

$$z^{(N)}(t) - z^{(M)}(t) = \int_{t_{\text{Init}}}^t (f^{(N)}(s) - f^{(M)}(s)) \, ds$$

and therefore

$$\begin{aligned} |z^{(N)}(t) - z^{(M)}(t)| & \leq \int_{t_{\text{Init}}}^t |f^{(N)}(s) - f(s, z^{(N)}(s))| \, ds \\ & + \int_{t_{\text{Init}}}^t |f(s, z^{(N)}(s)) - f(s, z^{(M)}(s))| \, ds \\ & + \int_{t_{\text{Init}}}^t |f(s, z^{(M)}(s)) - f^{(M)}(s)| \, ds. \end{aligned}$$

By following the reasoning similar to that made above, thanks to the uniform continuity of  $f$  on  $\mathcal{C}_0$ , for every  $\epsilon > 0$ , we have

$$\int_{t_{\text{Init}}}^t |f^{(N)}(s) - f(s, z^{(N)}(s))| ds \leq T\epsilon, \quad \int_{t_{\text{Init}}}^t |f^{(M)}(s) - f(s, z^{(M)}(s))| ds \leq T\epsilon$$

given that  $N$  and  $M$  are large enough. We denote the Lipschitz constant of  $f$  on  $\mathcal{C}_0$  as  $L_0$ , such that

$$\int_{t_{\text{Init}}}^t |f(s, z^{(N)}(s)) - f(s, z^{(M)}(s))| ds \leq L_0 \int_{t_{\text{Init}}}^t |z^{(N)}(s) - z^{(M)}(s)| ds.$$

It follows that

$$|z^{(N)}(t) - z^{(M)}(t)| \leq 2T\epsilon + L_0 \int_{t_{\text{Init}}}^t |z^{(N)}(s) - z^{(M)}(s)| ds$$

for large enough  $N$  and  $M$ . Grönwall's inequality implies that

$$|z^{(N)}(t) - z^{(M)}(t)| \leq 2Te^{L_0 T} \epsilon.$$

Therefore,  $(z^{(N)})_{N \in \mathbb{N}}$  is a Cauchy sequence in  $C^0([t_{\text{Init}}, t_{\text{Init}} + T])$  for the uniform norm, and it converges uniformly towards a continuous function  $z$  on  $[t_{\text{Init}}, t_{\text{Init}} + T]$ . Using the continuity of  $f$ , we also show that  $z$  satisfies the integral formulation in [1.15].

Finally, note that the locally Lipschitz nature of  $f$  also allows us to justify uniqueness, by again applying Grönwall's inequality. Let  $z_1, z_2$  be continuous functions satisfying

$$z_j(t) = y_{\text{Init},j} + \int_{t_{\text{Init}}}^t f(s, z_j(s)) ds$$

on  $[t_{\text{Init}}, t_{\text{Init}} + T]$ . Let  $K \subset \mathcal{U}$  denote a compact domain, such that  $(t, z_j(t)) \in K$  for every  $t \in [t_{\text{Init}}, t_{\text{Init}} + T]$  and  $j \in \{1, 2\}$ . We thus have

$$|z_1(t) - z_2(t)| \leq |y_{\text{Init},1} - y_{\text{Init},2}| + L_K \int_{t_{\text{Init}}}^t |z_1(s) - z_2(s)| ds$$

which implies that

$$|z_1(t) - z_2(t)| \leq |y_{\text{Init},1} - y_{\text{Init},2}| e^{L_K(t-t_{\text{Init}})}.$$

We will now focus on a specific class of differential equations, by assuming certain properties for the second member of equation [1.1].

**DEFINITION 1.6.**— We say the problem [1.1] is *dissipative* if the function  $f$  satisfies

$$(f(t, y_1) - f(t, y_2), y_1 - y_2) \leq 0$$

for all  $t, y_1, y_2$ , such that  $(t, y_1) \in \mathcal{U}$ ,  $(t, y_2) \in \mathcal{U}$ .

Whenever the function  $f$  is differentiable, we have a practical criterion that uses the Jacobian matrix (with respect to the state variable) of  $f$ .

**LEMMA 1.5.**— Assume that  $f$  is differentiable. Then, the problem is dissipative if for every  $(t, y) \in \mathcal{U}$  and every vector  $\xi \in \mathbb{R}^D$ , we have

$$\nabla_y f(t, y) \xi \cdot \xi \leq 0.$$

As mentioned before, the construction of the implicit Euler scheme is based on the possibility of solving the equation

$$z = y_n + \Delta t f(t_n + \Delta t, z)$$

whose unknown is  $z$  and where  $y_n \in \mathbb{R}^D$ ,  $t_n$  and  $\Delta t > 0$  are fixed. If this equation does indeed have a solution, then it defines the next iteration of the scheme, which will become an approximation to the solution of [1.1] at the instant  $t_{n+1}$ . One such approach involves introducing the function  $\Phi : y \mapsto y_n + \Delta t f(t_n + \Delta t, y)$  and showing the conditions that justify the existence of a fixed point. Therefore, by using Banach's theorem, we show that

- if the function  $f$  is globally Lipschitz continuous in the state variable, there exists an  $L > 0$ , such that for all  $t \in [t_{\text{Init}}, t_{\text{Init}} + T]$  and  $y_1, y_2 \in \mathcal{U}$ , we have  $|f(t, y_1) - f(t, y_2)| \leq L|y_1 - y_2|$ ,

- and if  $\Delta t < \frac{1}{L}$ ,

then the function  $\Phi$  is a contraction on  $\mathbb{R}^D$  and has a unique fixed point  $z = y_{n+1}$ . This statement is not very satisfying because it introduces a constraint on the time step  $\Delta t$ , in order to simply allow for the effective construction of the scheme. This restriction can be relaxed somewhat for dissipative problems.

**PROPOSITION 1.3.**— Suppose that the problem [1.1] is dissipative. In this case, *for every time step  $\Delta t > 0$* , the implicit Euler scheme [1.12] is well defined; the scheme is of order 1, stable and therefore convergent.

**PROOF.**— Let  $n \in \{0, \dots, N - 1\}$ . For fixed  $y_n$  and  $\Delta t > 0$ , to determine the next iterated term  $y_{n+1}$  using the implicit Euler scheme, it becomes necessary to find the root of the function

$$y \mapsto y - y_n - h f(t_n + h, y).$$

We introduce the mapping

$$F_n : (h, y) \in \mathbb{R} \times \mathbb{R}^D \mapsto y - y_n - h f(t_n + h, y) \in \mathbb{R}^D.$$

Note that  $F_n(0, y_n) = y_n - y_n = 0$ . Let  $(\bar{h}, \bar{y})$  be a root of  $F_n : F_n(\bar{h}, \bar{y}) = 0$ . The Jacobian matrix of  $F_n$  at point  $(\bar{h}, \bar{y})$  is expressed as

$$\nabla_y F_n(\bar{h}, \bar{y}) = \mathbb{I} - \bar{h} \nabla_y f(t_n + \bar{h}, \bar{y}).$$

Due to the dissipative nature of the problem, for every  $\xi \in \mathbb{R}^D$ , we have

$$\nabla_y F_n(\bar{h}, \bar{y}) \xi \cdot \xi \geq |\xi|^2$$

and  $\nabla_y F_n(\bar{h}, \bar{y})$  is an invertible matrix (because the linear mapping associated with it is one-to-one). By the implicit function theorem 1.4, starting from  $(\bar{h}, \bar{y})$ , we can define a curve  $h \mapsto y_h \in \mathbb{R}^D$  in a neighborhood of  $\bar{h}$ , such that

$$F_n(h, y_h) = 0, \quad y_{\bar{h}} = \bar{y}.$$

We introduce the set

$$\mathcal{E} = \{\bar{h} > 0, \text{such that for all } 0 < h < \bar{h}, \text{there exists}$$

$$y_h \in \mathbb{R}^D, \text{a solution of } F_n(h, y_h) = 0\}.$$

We just saw that this set is non-empty. Note first that with a fixed  $\bar{h}$ , there exists at most one vector  $\bar{y}$  that cancels  $y \mapsto F_n(\bar{h}, y)$ . Indeed, if  $\bar{y}_1$  and  $\bar{y}_2$  satisfy  $\bar{y}_1 - y_n - \bar{h} f(t + \bar{h}, \bar{y}_1) = \bar{y}_2 - y_n - \bar{h} f(t + \bar{h}, \bar{y}_2)$ , then

$$\begin{aligned} |\bar{y}_1 - \bar{y}_2|^2 &= \bar{h} (f(t + \bar{h}, \bar{y}_1) - f(t + \bar{h}, \bar{y}_2)) \cdot (\bar{y}_1 - \bar{y}_2) \\ &= \bar{h} \int_0^1 \nabla_y f(t + \bar{h}, \bar{y}_2 + s(\bar{y}_1 - \bar{y}_2)) (\bar{y}_1 - \bar{y}_2) \cdot (\bar{y}_1 - \bar{y}_2) \, ds \leq 0 \end{aligned}$$

which indeed implies that  $\bar{y}_1 = \bar{y}_2$ . Suppose that  $h_* = \sup \mathcal{E} < \infty$ . Consider a sequence  $(h_k)_{k \in \mathbb{N}}$  of elements from  $\mathcal{E}$ , such that  $\lim_{k \rightarrow \infty} h_k = h_*$ , and a sequence  $(\bar{y}_k)_{k \in \mathbb{N}}$  of roots of  $F_n$ , which is associated with it:

$$\begin{aligned} \bar{y}_k &= y_n + h_k f(t + h_k, \bar{y}_k) = y_n + h_k (f(t + h_k, \bar{y}_k) \\ &\quad - f(t + h_k, y_n)) + h_k f(t + h_k, y_n). \end{aligned}$$

From this relation, we infer the following estimate:

$$\begin{aligned} |\bar{y}_k - y_n|^2 &\leq h_k |f(t + h_k, y_n)| |\bar{y}_k - y_n| \\ &+ h_k \underbrace{\left( f(t + h_k, \bar{y}_k) - f(t + h_k, y_n) \right) \cdot (\bar{y}_k - y_n)}_{\leq 0} \end{aligned}$$

which implies

$$|\bar{y}_k| \leq |\bar{y}_k - y_n| + |y_n| \leq |y_n| + h_k |f(t + h_k, y_n)|.$$

Therefore, the sequence  $(\bar{y}_k)_{k \in \mathbb{N}}$  is bounded. We now find a subsequence, such that  $\lim_{\ell \rightarrow \infty} h_{k_\ell} = h_*$  and  $\lim_{\ell \rightarrow \infty} \bar{y}_{k_\ell} = \bar{y}_* \in \mathbb{R}^D$ . By the continuity of  $f$ , we obtain

$$\bar{y}_* - y_n - h_* f(t_n + h_*, \bar{y}_*) = F_n(h_*, \bar{y}_*) = 0.$$

Furthermore, if  $0 < h < h_*$ , then, for a large enough  $k$ , we have  $0 < h < h_k < h_*$ , and therefore there exists a  $y_h \in \mathbb{R}^D$ , such that  $F(h, y_h) = 0$ . It follows that  $h_* \in \mathcal{E}$ . This contradicts the definition of  $h_*$  because the implicit function theorem would make it possible to construct roots  $(h, y_h)$  of  $F_n$  with  $h > h_*$ . We conclude that  $h_* = +\infty$  and the implicit Euler scheme is well defined, with no restriction on the step  $h$ .

To study stability, we consider the sequences defined by  $y_0, z_0$  and the relations

$$y_{n+1} = y_n + h f(t + h, y_{n+1}), \quad z_{n+1} = z_n + h f(t + h, z_{n+1}) + \eta_n$$

where  $(\eta_n)_{n \in \mathbb{N}}$  is a given sequence in  $\mathbb{R}^D$ . We have

$$\begin{aligned} |y_{n+1} - z_{n+1}|^2 &= |y_{n+1} - z_{n+1}|(|y_n - z_n| + |\eta_n|) \\ &+ h \underbrace{\left( f(t + h, y_{n+1}) - f(t + h, z_{n+1}) \right) \cdot (y_{n+1} - z_{n+1})}_{\leq 0} \\ &\leq |y_{n+1} - z_{n+1}|(|y_n - z_n| + |\eta_n|). \end{aligned}$$

It follows that

$$|y_{n+1} - z_{n+1}| \leq |y_n - z_n| + |\eta_n| \leq |y_0 - z_0| + \sum_{k=0}^n |\eta_k|.$$

The stability constant is 1: here, a crucial difference emerges with respect to the stability analysis for the explicit Euler scheme [1.11] whose stability constant depends on the Lipschitz constant of  $f$  and the final time.

Finally, in order to describe consistency, we can write

$$\begin{aligned}\epsilon_n &= y(t_{n+1}) - y(t_n) - hf(t_{n+1}, y(t_{n+1})) \\ &= \int_{t_n}^{t_{n+1}} y'(s) \, ds - \int_{t_n}^{t_{n+1}} \, ds \times y'(t_{n+1}) = - \int_{t_n}^{t_{n+1}} \left( \int_s^{t_{n+1}} y''(\sigma) \, d\sigma \right) \, ds.\end{aligned}$$

This makes it possible to bound it from above by, for example,

$$|\epsilon_n| \leq \|y''\|_{L^\infty(0,T)} h^2.$$

We can therefore conclude our discussion of the implicit Euler scheme's convergence.  $\square$

Explicit and implicit Euler schemes are both first-order consistent, stable and therefore convergent. Nevertheless, for certain differential equations, there might be a good reason to prefer the implicit method, which provides more favorable estimations. These phenomena are very subtle and their analysis is delicate.

To this end, we will attempt to quantify the propensity that a numerical solution might have to distance itself from the real solution. Let  $T > 0$  be a simulation time that is strictly less than the lifetime of the solution  $t \mapsto y(t)$ , which we assume to be defined from the initial time  $t_{\text{Init}} = 0$ . We can write

$$M = \max \{|y(t)|, 0 \leq t \leq T\}.$$

There is no apparent reason for the approximate solution  $y_n$  to remain in the ball  $B(0, M)$ . Also, for a given  $C > 0$ , we seek to know under what conditions the approximate solution remains in  $B(0, M + C)$ . In order to perform this study, we assume that  $B(0, M + C) \subset \mathcal{U}$  and denote the Lipschitz constant of  $f$  on  $[0, T] \times \mathcal{U}$  as  $L$  (this is to say that  $|f(t, x) - f(t, y)| \leq L|x - y|$  for all  $0 \leq t \leq T$  and  $x, y \in \mathcal{U}$ ). Recall that

$$y_n = e_n + y(t_n), \quad e_n = y_n - y(t_n)$$

such that  $|y_n| \leq M + |e_n|$  and the question comes down to estimating the local error  $e_n$ . For the implicit Euler scheme, we have

$$y_{n+1} = y_n + hf(t_{n+1}, y_{n+1}),$$

$$y(t_{n+1}) = y(t_n) + hf(t_{n+1}, y(t_{n+1})) + \epsilon_n$$

such that

$$e_{n+1} = e_n + h(f(t_{n+1}, y_{n+1}) - f(t_{n+1}, y(t_{n+1}))) - \epsilon_n.$$

Using the fact that the problem is dissipative, it follows that

$$\begin{aligned} |e_{n+1}|^2 &= (e_n - \epsilon_n) \cdot e_{n+1} \\ &\quad + \underbrace{h(f(t_{n+1}, y_{n+1}) - f(t_{n+1}, y(t_{n+1}))) \cdot (y_{n+1} - y(t_{n+1}))}_{\leq 0} \\ &\leq (|e_n| + |\epsilon_n|) |e_{n+1}| \end{aligned}$$

and therefore,  $|e_{n+1}| \leq |e_n| + |\epsilon_n|$ . However, we have seen that

$$\begin{aligned} |\epsilon_n| &= |y(t_{n+1}) - y(t_n) - hf(t_{n+1}, y(t_{n+1}))| \\ &= \left| \int_{t_n}^{t_{n+1}} y'(s) \, ds - \int_{t_n}^{t_{n+1}} \, ds \times y'(t_{n+1}) \right| \\ &\leq \int_{t_n}^{t_{n+1}} |y'(s) - y'(t_{n+1})| \, ds = \int_{t_n}^{t_{n+1}} \left| \int_s^{t_{n+1}} y''(\sigma) \, d\sigma \right| \, ds \\ &\leq h \int_{t_n}^{t_{n+1}} |y''(\sigma)| \, d\sigma. \end{aligned}$$

We can therefore establish by direct induction that

$$|e_n| \leq |e_0| + h \int_0^{t_n} |y''(\sigma)| \, d\sigma.$$

(Again, this relation shows that the implicit Euler scheme converges.) We can thus obtain the estimate

$$|y_n| \leq M + h \|y''\|_{L^1(0,T)}.$$

It follows that the numerical solution remains in  $B(0, M + C)$ , given that  $h \leq \frac{C}{\|y''\|_{L^1(0,T)}}$ .

With the explicit Euler scheme, we have

$$y_{n+1} = y_n + hf(t_n, y_n),$$

$$y(t_{n+1}) = y(t_n) + hf(t_n, y(t_n)) + \epsilon_n$$

so

$$|e_{n+1}| \leq |e_n| + h |f(t_n, y_n) - f(t_n, y(t_n))| + |\epsilon_n|.$$

We assume that  $y_n \in B(0, M + C)$ . Using the fact that  $f$  is  $L$ -Lipschitz in the state variable on  $[0, T] \times B(0, M + C)$ , and the estimate on the local consistency error  $|\epsilon_n| \leq \frac{\|y''\|_\infty}{2} h^2$ , we obtain

$$|e_{n+1}| \leq (1 + Lh)|e_n| + \frac{\|y''\|_\infty}{2} h^2$$

which leads to

$$|e_{n+1}| \leq e^{LT} \left( |e_0| + \frac{\|y''\|_\infty}{2L} h \right).$$

Therefore, the desired estimate ( $y_{n+1} \in B(0, M + C)$ ) is satisfied under the restrictive condition

$$h e^{LT} \frac{\|y''\|_\infty}{2L} \leq C.$$

With a fixed  $T$  and  $C$ , the greater the  $L$  is, the smaller the time interval  $h$  must be, and the condition is more restrictive than that shown for the implicit scheme by using the dissipative nature of the equation.

This discussion reveals that, in fact, several notions of stability for a numerical scheme coexist, which are more or less pertinent according to the problem being considered, and can be related to a scheme's capacity to

- a) not amplify perturbations introduced by the numerical approximation in an uncontrolled manner (stability with respect to errors)<sup>7</sup>;
- b) produce an approximate solution that remains bounded within a domain as close as possible to that of the solution to the continuous problem;
- c) provide a solution that preserves the known properties (positiveness<sup>8</sup>, conservation of energy, etc.) of the continuous solution;

---

<sup>7</sup> The relevant errors are those *produced by the scheme itself*: the continuous problem's solution does not satisfy the recurrence formula that defines the scheme and we can wonder how this *consistency* error is propagated by the scheme. This consistency error is the leading phenomena; in particular, it largely dominates the so-called "machine round-off" that results from the way real numbers are represented on a computer!

<sup>8</sup> On this point, we can compare the implicit and explicit Euler schemes for solving  $y'(t) = -ay(t)$ , with  $a > 0$ . The continuous solution is known: if  $y(0) = y_{\text{init}}$ , then  $y(t) = e^{-at} y_{\text{init}}$ , which is non-negative if  $y_{\text{init}} \geq 0$ . The explicit Euler scheme can be written as  $y_{n+1} = (1 - a\Delta t)y_n$ ; it preserves only non-negative values under the constraint  $\Delta t < 1/a$ . The implicit Euler scheme can be written as  $y_{n+1} = (1 + a\Delta t)^{-1}y_n$ , and guarantees that all iterations are non-negative regardless of  $\Delta t > 0$ . For the implicit scheme, if  $a < 0$ , there arise constraints on the time step.

d) dissipate certain quantities.

The last two aspects will be discussed in more detail when we study Hamiltonian problems.

To conclude the presentation of Euler schemes, we will now return to the practical implementation of the implicit method. We have seen that when  $y_n$  is known, each new iteration is constructed as the solution of the nonlinear problem

$$z = y_n + \Delta t f(t_n + \Delta t, z).$$

We introduce the following function, which is parameterized by  $t_n$ ,  $y_n$  and  $\Delta t$ :

$$\Psi : z \in \mathbb{R}^D \longrightarrow z - y_n - \Delta t f(t_n + \Delta t, z).$$

We therefore seek a process that would allow us to calculate (an approximation to) the root of the function  $\Psi$ . The construction relies on an iterative process – *Newton's method* – whose main idea is as follows. If  $z_*$  satisfies  $\Psi(z_*) = 0$ , Taylor's formula leads to

$$\Psi(z_*) = 0 = \Psi(z) + \nabla \Psi(z)(z_* - z) + \|z_* - z\| \epsilon(z_* - z) \text{ with } \lim_{|h| \rightarrow 0} \epsilon(h) = 0.$$

The idea involves considering only the leading part in the right-hand term, which is therefore a linear function of  $z_*$ , in order to define a sequence, and we will attempt to show that it indeed approaches  $z_*$ . More specifically, we take a vector  $z^{(0)}$  and then define the following iterative scheme:

a) Constructing the matrix

$$M^{(k)} = \begin{pmatrix} \partial_{z_1} \Psi_1(z^{(k)}) & \partial_{z_2} \Psi_1(z^{(k)}) & \dots & \partial_{z_D} \Psi_1(z^{(k)}) \\ \partial_{z_1} \Psi_2(z^{(k)}) & \partial_{z_2} \Psi_2(z^{(k)}) & \dots & \partial_{z_D} \Psi_2(z^{(k)}) \\ \vdots & \vdots & & \vdots \\ \partial_{z_1} \Psi_D(z^{(k)}) & \partial_{z_2} \Psi_D(z^{(k)}) & \dots & \partial_{z_D} \Psi_D(z^{(k)}) \end{pmatrix},$$

(It is the Jacobian matrix of  $\Psi$  evaluated at  $z^{(k)}$ ).

b) Solving the linear system  $M^{(k)} \zeta^{(k)} = \Psi(z^{(k)})$ ,

c) Letting  $z^{(k+1)} = z^{(k)} - \zeta^{(k)}$ .

In other words, since  $M^{(k)} = \nabla\Psi(z^{(k)})$ ,  $z^{(k+1)}$  satisfies

$$\Psi(z^{(k)}) + \nabla\Psi(z^{(k)}) \cdot (z^{(k+1)} - z^{(k)}) = 0$$

(or  $z^{(k+1)} = z^{(k)} - [\nabla\Psi(z^{(k)})]^{-1}\Psi(z^{(k)})$ ). However, in practice, it is preferable to solve *one* linear system rather than calculating the inverse of a matrix. The efficiency of this algorithm is confirmed by the following statement.

**THEOREM 1.12** (Convergence of Newton's Method).— Let  $\Psi : \mathbb{R}^D \rightarrow \mathbb{R}^D$  be a  $C^2$  function and  $z_* \in \mathbb{R}^D$ , such that  $\Psi(z_*) = 0$ . Suppose that the matrix  $\nabla\Psi(z_*) \in \mathcal{M}_D(\mathbb{R})$  is invertible. There exists a  $\epsilon > 0$ , such that if  $|z^{(0)} - z_*| \leq \epsilon$ , then for every  $k \in \mathbb{N}$ , we have  $|z^{(k)} - z_*| \leq \epsilon$  and the sequence  $(z^{(k)})_{k \in \mathbb{N}}$  converges towards  $z_*$ .

We will see that convergence of  $z^{(k)}$  towards  $z_*$  is fast: we will effectively establish that there exists a  $\kappa > 0$ , such that

$$|z^{(k+1)} - z_*| \leq \kappa |z^{(k)} - z_*|^2. \quad [1.16]$$

We say the convergence is *quadratic*. Indeed, we can write  $r^{(k)} = \kappa |z^{(k)} - z_*|$ , which satisfies  $r^{(k+1)} \leq |r^{(k)}|^2$ . An immediate recurrence leads to the estimate

$$r^{(k)} \leq (r^{(0)})^{2^k}$$

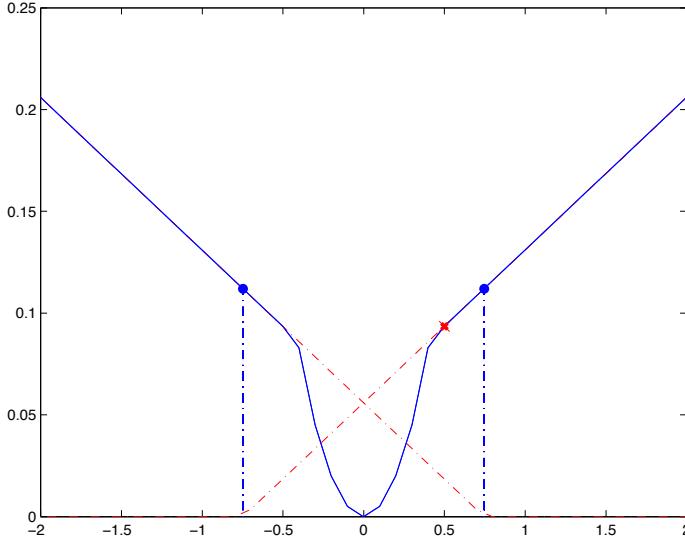
which proves that  $\lim_{k \rightarrow \infty} |z^{(k)} - z_*| = \lim_{k \rightarrow \infty} \frac{r^{(k)}}{\kappa} = 0$ , given that  $r^{(0)} \leq \kappa\epsilon < 1$ .

However, the success of the method depends on the fact that the initial iteration is “close enough” to the desired point  $z_*$ . In practice, we do not know  $z_*$  or  $\kappa$  (and therefore  $\epsilon$ ). It may be that the algorithm does not converge and that it does not allow for a correct determination of  $z_*$ . In that case, we must find another way to approach  $z_*$ , even if somewhat approximately, but sufficiently well to return to Newton's method. Finally, as we will see in the proof, the assumption that  $\nabla\Psi(z_*)$  is invertible is crucial. Figure 1.1 shows a convergence defect for the method. The function  $\Psi$  that we studied is equal to  $x^2/2$  in the neighborhood of the origin: it cancels itself out at that point, which is also a minimum of the function. Other than in a “small” neighborhood of 0, the function  $\Psi$  is linear, with a rather small slope. Moreover, the function is constructed in such a way as to gather that behavior regularly. As can be seen in the figure, when Newton's method starts from a point that is not close enough to the origin, it oscillates between two values.

**PROOF OF THEOREM 1.12.** The first stage consists of ensuring that the matrix  $\nabla\Psi(z)$  remains invertible when  $z$  is in a sufficiently small neighborhood of  $z_*$ . We can write

$$\nabla\Psi(z) = \nabla\Psi(z_*) - (\nabla\Psi(z_*) - \nabla\Psi(z)) = \nabla\Psi(z_*)(\mathbb{I} - M(z))$$

$$\text{with } M(z) = [\nabla\Psi(z_*)]^{-1}(\nabla\Psi(z_*) - \nabla\Psi(z)).$$



**Figure 1.1.** Illustration of non-convergence for Newton's method.  
The algorithm starts from a point marked as  $x$  and oscillates between two values indicated by a •

However, recall that if  $H \in \mathcal{M}_D(\mathbb{R})$  satisfies  $\|H\| < 1$ , then  $(\mathbb{I} - H)$  is invertible. The series  $\sum_{k=0}^{\infty} H^k$  is normally convergent and  $(\mathbb{I} - H)^{-1} = \sum_{k=0}^{\infty} H^k$ . Therefore,  $\nabla\Psi(z)$  is invertible whenever the matrix  $M(z)$  has a norm strictly smaller than 1, as a product of invertible matrices. For  $z = z_*$ , we have  $M(z_*) = 0$  and  $z \mapsto \nabla\Psi(z)$  is continuous. It follows that there exists an  $\epsilon_1 > 0$ , such that for every  $|z - z_*| < \epsilon_1$ , we have  $\|M(z)\| < 1$  and therefore  $\nabla\Psi(z)$  is invertible. Moreover, for  $|z - z_*| \leq \epsilon_1$ , we have

$$\begin{aligned}\|\nabla\Psi(z)\|^{-1} &= \|(\mathbb{I} - M(z))^{-1}\nabla\Psi(z_*)\|^{-1} \leq \|(\mathbb{I} - M(z))^{-1}\| \|\nabla\Psi(z_*)\|^{-1} \\ &\leq \|\nabla\Psi(z_*)\|^{-1} \sum_{k=0}^{\infty} \|M(z)\|^k = \frac{\|\nabla\Psi(z_*)\|^{-1}}{1 - \|M(z)\|}.\end{aligned}$$

Next, we write  $\kappa_1 > 0$ , such that for all  $|z - z_*| \leq \epsilon_1$ , we have

$$\|\nabla\Psi(z)\|^{-1} \leq \kappa_1.$$

Suppose that  $|z^{(k)} - z_*| \leq \epsilon < \epsilon_1$ , for a certain  $\epsilon > 0$ . We define  $z^{(k+1)}$  as a solution to the linear problem

$$\nabla \Psi(z^{(k)})(z^{(k+1)} - z^{(k)}) + \Psi(z^{(k)}) = 0.$$

Since  $\Psi(z_*) = 0$ , we can rewrite the relation in the form of

$$\nabla \Psi(z^{(k)})(z^{(k+1)} - z_*) - \nabla \Psi(z^{(k)})(z^{(k)} - z_*) + \Psi(z^{(k)}) - \Psi(z_*) = 0.$$

We therefore have

$$\begin{aligned} \nabla \Psi(z^{(k)})(z^{(k+1)} - z_*) &= \nabla \Psi(z^{(k)})(z^{(k)} - z_*) \\ &+ \int_0^1 \frac{d}{dt} \left( \Psi(z^{(k)} + t(z_* - z^{(k)})) \right) dt \\ &= \int_0^1 \left( \nabla \Psi(z^{(k)}) - \nabla \Psi(z^{(k)} + t(z_* - z^{(k)})) \right) (z^{(k)} - z_*) dt \\ &= - \int_0^1 \int_0^1 t D^2 \Psi(z^{(k)} + \theta t(z_* - z^{(k)})) (z^{(k)} - z_*, z^{(k)} - z_*) d\theta dt. \end{aligned}$$

However,  $|z^{(k)} + \theta t(z_* - z^{(k)}) - z_*| = (1 - \theta t)|z^{(k)} - z_*| \leq \epsilon \leq \epsilon_1$  for every  $0 \leq \theta, t \leq 1$ , and since  $\Psi$  is a function of class  $C^2$ , there exists a  $\kappa_2 > 0$ , such that for all  $|u - z_*| \leq \epsilon_1$ , and every  $\xi \in \mathbb{R}^D$ , we have  $|D^2 \Psi(u)(\xi, \xi)| \leq \kappa_2 |\xi|^2$ . It follows that

$$\left| \int_0^1 \int_0^1 t D^2 \Psi(z^{(k)} + \theta t(z_* - z^{(k)})) (z^{(k)} - z_*, z^{(k)} - z_*) d\theta dt \right| \leq \kappa_2 |z^{(k)} - z_*|^2.$$

By writing  $\kappa = \kappa_1 \kappa_2$ , we deduce that

$$|z^{(k+1)} - z_*| \leq \kappa |z^{(k)} - z_*|^2.$$

However, we assumed that  $|z^{(k)} - z_*| \leq \epsilon < \epsilon_1$ , implying that  $|z^{(k+1)} - z_*| \leq \kappa \epsilon^2$ . In order to repeat the argument, we will impose  $0 < \epsilon < \min(\frac{1}{\kappa}, \epsilon_1)$  in such a way that all elements of the sequence thus defined satisfy  $|z^{(k)} - z_*| \leq \epsilon$ . This discussion allows us to show [1.16], with  $\kappa \epsilon < 1$ , by construction and therefore theorem 1.12 is proved.  $\square$

Returning to the implicit Euler method, we define  $y_{n+1}$  by applying Newton's algorithm to the function  $\Psi(z) = z - y_n - \Delta t f(t_n + \Delta t, z)$ , with  $z^{(0)} = y_n$ . In order to make  $\lim_{k \rightarrow \infty} z^{(k)}$  numerically accessible, we set a stopping criterion and stop the iterative process when the relative error  $\frac{|z^{(k+1)} - z^{(k)}|}{|z^{(k)}|}$  exceeds a previously fixed tolerance limit.

### 1.2.4. Higher-order schemes

The idea for obtaining higher-order methods involves taking inspiration from higher-order numerical integration methods. The trapezoidal rule (see section Appendix 2) allows us to obtain the scheme

$$y_{n+1} = y_n + \frac{\Delta t}{2} (f(t_n, y_n) + f(t_{n+1}, y_{n+1})),$$

which is of an implicit nature. Therefore, it would be necessary to analyze the conditions that effectively make it possible to define  $y_{n+1}$ . It is *Crank–Nicolson’s method*. Assuming that  $f$  is regular enough and denoting a function that tends to 0 when  $h \rightarrow 0$  as  $h \mapsto \epsilon(h)$ , the local consistency error can be expressed as

$$\begin{aligned} \epsilon_n &= y(t_{n+1}) - y(t_n) - \frac{\Delta t}{2} (f(t_n, y(t_n)) + f(t_{n+1}, y(t_{n+1}))) \\ &= y'(t_n) \Delta t + y''(t_n) \frac{\Delta t^2}{2} - \frac{\Delta t}{2} \left( 2 \underbrace{f(t_n, y(t_n))}_{=y'(t_n)} + \partial_t f(t_n, y(t_n)) \Delta t \right. \\ &\quad \left. + \nabla_y f(t_n, y(t_n)) \underbrace{(y(t_{n+1}) - y(t_n))}_{=y'(t_n) \Delta t + \Delta t \epsilon(\Delta t)} + \Delta t^2 \epsilon(\Delta t) \right). \end{aligned}$$

However,  $y'(t) = f(t, y(t))$  and by differentiation, we have

$$\begin{aligned} y''(t) &= \frac{d}{dt} [f(t, y(t))] = \partial_t f(t, y(t)) + \nabla_y f(t, y(t)) \cdot y'(t) \\ &= \partial_t f(t, y(t)) + \nabla_y f(t, y(t)) \cdot f(t, y(t)). \end{aligned}$$

It follows that the local consistency error is  $\epsilon_n = \Delta t^2 \epsilon(\Delta t)$ . The Crank–Nicolson scheme is therefore second-order consistent.

In order to obtain higher orders, it is necessary to approach the integral

$$\int_{t_n}^{t_{n+1}} f(s, y(s)) ds$$

in the form of a linear combination  $\sum_{j=1}^J \beta_j f(t_n^j, y_n^j)$ , where the intermediate values  $y_n^j$  are themselves obtained as combinations of the previous points

$$y_n^j = y_n + \sum_{k=1}^{j-1} \alpha_{j,k} f(t_n^k, y_n^k)$$

with

$$t_n^j = t_n + \gamma^j \Delta t, \quad 0 < \gamma^j < 1, \quad \sum_{j=1}^J \alpha_{j,k} = \gamma^k.$$

We thus obtain the family of *Runge–Kutta schemes*. The values of the coefficients are tabulated. These methods remain explicit, but the step  $t_{n+1}$  is achieved in several stages.

### 1.2.5. Leslie's equation (Perron–Frobenius theorem, power method)

We consider here a population dynamics problem. We divide the population into  $n$  age classes. At instant  $t \geq 0$ ,  $x_j(t)$  is the number of individuals in the  $j^{\text{th}}$  class. Its dynamics depends

- on the fertility rates of the age classes:  $f_j > 0$  is the average rate of newborns—that is to say, individuals within class 1 – produced by individuals in the class  $j$ ;
- on the transition rate  $0 \leq t_{j+1,j} < 1$ , which gives the proportion of individuals in the age category  $j$  that can reach the age category  $j + 1$ .

The population's behavior is completely described by the Leslie matrix

$$L = \begin{pmatrix} f_1 & f_2 & \cdots & f_n \\ t_{2,1} & 0 & \ddots & 0 \\ 0 & t_{3,2} & \ddots & \vdots \\ \vdots & & \ddots & 0 \\ 0 & \cdots & 0 & t_{n,n-1} \\ \vdots & & & 0 \end{pmatrix}$$

We will analyze how the asymptotic behavior of the population for large times is related to the remarkable spectral properties of the matrix  $L$ . For more details about Leslie's model, we refer the reader to [CUS 98] and the original articles [LES 45, LES 48].

### 1.2.5.1. Discrete time model

We begin by considering discrete times. The population's evolution between two instants,  $k$  and  $k + 1$ , is defined by the linear relation

$$X^{(k+1)} = LX^{(k)}$$

where  $X^{(k)} = (x_1^{(k)}, \dots, x_n^{(k)})$  is the vector describing the population distribution in age classes at instant  $k \in \mathbb{N}$ . Even if it is not quite accurate from a biological point of view, we can begin by studying the scalar case involving only a single class. In this case, the sequence  $x^{(k)}$  is a simple geometric sequence

$$x^{(k)} = (t + f)x^{(k-1)} = \dots = (t + f)^k x^{(0)}$$

and the asymptotic behavior depends on the ratio

$$\lim_{k \rightarrow \infty} x^{(k)} = \begin{cases} x^{(0)} & \text{if } t + f = 1, \\ 0 & \text{if } t + f < 1, \\ +\infty & \text{if } t + f > 1. \end{cases}$$

In other words, the dynamic is governed by the number  $R = \frac{f}{1-t}$ , with  $f > 0$  and  $0 \leq t < 1$ , which we describe in the form of a series

$$R = f \sum_{k=0}^{\infty} t^k = f + ft + ft^2 + ft^3 + \dots$$

where  $t$  is the probability of passing from age  $k$  to age  $k + 1$ , in particular from age 1 to age 2,  $t^2$  is the probability of surviving until age 2,  $t^3$  is the probability of surviving up to age 3, etc. We multiply these numbers by  $f$ , the probable number of babies per individuals of a given age. Therefore,  $R$  is interpreted as the number of babies per individual over an entire lifetime. In the vector case, the population's asymptotic behavior, on a long time scale, can be deduced from the spectral properties of the matrix  $L$ . More specifically, we will use the Perron–Frobenius theorem [FRO 12, PER 07], which, as we will see, has several versions with somewhat different conclusions.

**THEOREM 1.13 (Perron–Frobenius Theorem).**— Let  $A \in M_n(\mathbb{R})$  be a matrix with strictly positive coefficients  $a_{k,j}$ . Then,

- i) the spectral radius  $\rho := \sup \{|\lambda|, \text{ where } \lambda \in \mathbb{C} \text{ is an eigenvalue of } A\}$  is a simple eigenvalue of  $A$ ;
- ii) there exists an eigenvector  $v$  of  $A$  associated with the eigenvalue  $\rho$  whose components are strictly positive;

- iii) every other eigenvalue  $\lambda$  of  $A$  satisfies  $|\lambda| < \rho$ ;
- iv) there exists an eigenvector  $\phi$  of  $A^T$  associated with the eigenvalue  $\rho$  and whose components are strictly positive.

COROLLARY 1.6.– Let  $A \in M_n(\mathbb{R})$  be a matrix with non-negative coefficients  $a_{k,j}$ . Its spectral radius  $\rho(A)$  is an eigenvalue and there exists an eigenvector for that eigenvalue whose coordinates are non-negative.

The statement in theorem 1.13 (which we will prove further below) does not apply to the Leslie matrix  $L$ , since that matrix's coefficients are not all positive. However, we will be able to use the following variant.

COROLLARY 1.7.– Let  $B \in M_n(\mathbb{R})$ . We assume there exists a  $q \in \mathbb{N}$ , such that the coefficients of  $B^q$  are strictly positive<sup>9</sup>. Then, the conclusions of the Perron–Frobenius theorem apply to the matrix  $B$ .

This corollary applies to the Leslie matrix since the fertility rates  $f_j$  and the transition rates  $t_{j+1,j}$  are all strictly positive. This assumption excludes, in particular, the case of populations whose younger age group  $j \in \{1, \dots, j_0\}$  is infertile, as well as those whose older age groups  $j \in \{j_1, \dots, n\}$  have become sterile. To justify the fact that the Leslie matrix  $L$  satisfies the assumptions of corollary 1.7, we provide an induction argument. Let  $k \in \{1, \dots, n-1\}$ . We assume that  $L^k$  is such that  $(L^k)_{m,i} > 0$  for every  $m \in \{1, \dots, k\}$ ,  $i \in \{1, \dots, n\}$  and  $(L^k)_{k+1,1} > 0$ ,  $(L^k)_{k+2,2} > 0, \dots, (L^k)_{k+(n-k),n-k} > 0$ . This property is satisfied by  $k = 1$ . We use the relation

$$(L^{k+1})_{i,j} = \sum_{\ell=1}^n (L^k)_{i,\ell} L_{\ell,j},$$

knowing that those matrices' coefficients are positive or zero. On the one hand, we have

$$(L^{k+1})_{m,j} \geq (L^k)_{m,1} f_j > 0$$

for all  $m \in \{1, \dots, k\}$  and  $j \in \{1, \dots, n\}$  (the rows of strictly positive coefficients remain strictly positive) and

$$(L^{k+1})_{k+1,j} \geq (L^k)_{k+1,1} f_j > 0$$

---

<sup>9</sup> A matrix of this kind is called *primitive*.

for every  $j \in \{1, \dots, n\}$  (the row  $k+1$  is comprised of strictly positive coefficients). On the other hand, we obtain

$$(L^{k+1})_{k+1+j,j} \geq (L^k)_{k+(j+1),j+1} t_{j+1,j} > 0$$

for every  $j \in \{1, \dots, n - (k+1)\}$ . The induction hypothesis is passed on to the rank  $k+1$ . It follows that the coefficients of  $L^n$  are all strictly positive and therefore  $L$  is primitive.

As in the scalar case, we can always write

$$X^{(k)} = LX^{(k-1)} = L^k X^{(0)}.$$

We let  $\rho_L > 0$  designate the spectral radius of  $L$  and use the Jordan–Chevalley decomposition, which allows us to write

$$L = P^{-1}(D + N)P, \quad D = \begin{pmatrix} \rho_L & 0 & \cdots & 0 \\ 0 & \mu_1 & & \\ \vdots & & \ddots & \\ 0 & & & \mu_{n-1} \end{pmatrix}, \quad |\mu_j| < \rho_L,$$

$N^{k_0} = 0$  for a certain  $k_0 \in \mathbb{N}$ ,  $DN = ND$ .

It follows that, for  $k \gg k_0$ ,

$$L^k = P^{-1}(D + N)^k P = P^{-1} \left( \sum_{\ell=0}^k C_k^\ell D^{k-\ell} N^\ell \right) P = P^{-1} \left( \sum_{\ell=0}^{k_0} C_k^\ell D^{k-\ell} N^\ell \right) P.$$

In that sum, we have

– on the one hand,

$$D^{k-\ell} = \rho_L^k \begin{pmatrix} \rho_L^{-\ell} & 0 & \cdots & 0 \\ 0 & \kappa_1^{k-\ell} \rho_L^{-\ell} & & \\ \vdots & & \ddots & \\ 0 & & & \kappa_{n-1}^{k-\ell} \rho_L^{-\ell} \end{pmatrix}$$

with  $\kappa_j = \frac{\lambda_j}{\rho_L}$  having a modulus strictly less than 1, so that

$$\lim_{k \rightarrow \infty} \kappa_j^{k-\ell} \rho_L^{-\ell} = 0,$$

– on the other hand, we can estimate the binomial coefficients from above

$$C_k^\ell = \frac{k!}{\ell!(k-\ell)!} \leq \frac{k!}{1 \times (k-k_0)!} \leq k(k-1)\dots(k-k_0+1) \leq k^{k_0}.$$

We can therefore reorganize the sum to write

$$X^{(k)} = \rho_L^k \left( Y^{(k)} + \sum_{j=1}^{n-1} \kappa_j^k Z^{(j,k)} \right)$$

where  $Y^{(k)}$  and  $Z^{(j,k)}$  are vectors whose growth in  $k$  is polynomial, at most in  $k^{k_0}$ . Moreover, the system has a conservation property: by the Perron–Frobenius theorem, we can find two vectors,  $\phi$  and  $e$ , with strictly positive coordinates, such that

$$Le = \rho_L e, \quad L^\top \phi = \rho_L \phi, \quad |e| = 1, \quad \phi \cdot e = 1,$$

and we thus obtain

$$X^{(k)} \cdot \phi = L^k X^{(0)} \cdot \phi = X^{(0)} \cdot (L^\top)^k \phi = \rho_L^k X^{(0)} \cdot \phi.$$

It follows that

$$\frac{X^{(k)}}{\rho_L^k} \cdot \phi = Y^{(k)} \cdot \phi + r^{(k)} = X^{(0)} \cdot \phi, \quad \text{with } \lim_{k \rightarrow \infty} r^{(k)} = 0.$$

Since the coordinates of  $\phi$  are strictly positive, it follows that the coordinates  $(\frac{X_j^{(k)}}{\rho_L^k})_{k \in \mathbb{N}}$  form a bounded sequence. In light of the estimates of  $Y^{(k)}$  and  $Z^{(j,k)}$ , this implies that  $Y^{(k)} = \bar{Y}$  is constant, with  $\bar{Y} \cdot \phi = X^{(0)} \cdot \phi$ . Finally, we obtain

$$\lim_{k \rightarrow \infty} \frac{X^{(k+1)}}{\rho_L^k} = \rho_L \times \lim_{k \rightarrow \infty} Y^{(k+1)} = \rho_L \bar{Y} = \lim_{k \rightarrow \infty} \frac{LX^{(k)}}{\rho_L^k} = L\bar{Y}.$$

This means that  $\bar{Y}$  is an eigenvector of  $L$  associated with the eigenvalue  $\rho_L$ , so  $\bar{Y} = \alpha e$  for some  $\alpha \in \mathbb{C}$ . We can express this constant with

$$\bar{Y} \cdot \phi = \alpha = X^{(0)} \cdot \phi.$$

We therefore conclude that

$$X^{(k)} \sim_{k \rightarrow \infty} \rho_L^k X^{(0)} \cdot \phi e.$$

Therefore, the total population

$$\sum_{j=1}^n X^{(k)} \begin{cases} \text{tends to } X^{(0)} \cdot \phi \sum_{j=1}^n e_j > 0 \text{ if } \rho_L = 1, \\ \text{tends to } +\infty \text{ if } \rho_L > 1, \\ \text{tends to } 0 \text{ if } \rho_L < 1. \end{cases}$$

In order to continue the discussion, we consider the following statement:

LEMMA 1.6.– Let  $A$  be a matrix whose coefficients are positive or zero, and which we assume to be primitive.

i) If  $z \neq 0$  has positive or zero coordinates and satisfies  $Az \geq \rho(A)z$ , then the coordinates of  $z$  are strictly positive and  $Az = \rho(A)z$ .

ii) Let  $B \neq A$  be a matrix whose coefficients satisfy  $B_{ij} \geq A_{ij} \geq 0$ . Then,  $\rho(B) > \rho(A)$ .

PROOF.– According to the Perron–Frobenius theorem applied to  $A^\top$ , we can find a vector  $\phi$  whose coordinates are strictly positive, such that  $A^\top \phi = \rho(A)\phi$ . We write  $y = Az - \rho(A)z$  whose coordinates are non-negative. If  $y \neq 0$ , we have  $y \cdot \phi > 0$ , which contradicts the fact that

$$y \cdot \phi = (A - \rho(A)\mathbb{I})z \cdot \phi = z \cdot (A^\top - \rho(A)\mathbb{I})\phi$$

is zero because  $\phi$  is an eigenvector of  $A^\top$  associated with  $\rho(A)$ .

We thus have  $Az = \rho(A)z$ . This implies that  $A^k z = \rho(A)^k z$ , since the vector has strictly positive values for a large enough  $k$  because  $A$  is primitive. Item i) is thus demonstrated. Next, the inequalities  $B_{ij} \geq A_{ij} \geq 0$  imply  $[B^k]_{ij} \geq [A^k]_{ij} \geq 0$ . Because  $A$  is primitive,  $B$  is also primitive. By the Perron–Frobenius theorem,  $\rho(A)$  is an eigenvector of  $A$  and has an eigenvector  $x$  with strictly positive values. We assume  $|x| = 1$ . For each coordinate, it follows that

$$0 \leq \rho(A)^k x_j \leq (A^k x)_j \leq (B^k x)_j$$

which implies

$$\rho(A)^k \leq |A^k x| \leq |B^k x| \leq \|B^k\|,$$

so  $\rho(A) \leq \|B^k\|^{1/k}$  and, finally, by lemma A1.2,

$$\rho(A) \leq \lim_{k \rightarrow \infty} \|B^k\|^{1/k} = \rho(B).$$

Suppose that  $\rho(A) = \rho(B)$ . In this case, we have

$$0 \leq \rho(A)x_j = \rho(B)x_j \leq (Ax)_j \leq (Bx)_j$$

and, since  $B$  is primitive, item i) implies that  $x$  is also an eigenvector of  $B$ , for the eigenvalue  $\rho(B)$  and its coordinates are strictly positive. We therefore have  $(B - A)x = 0$  and for all  $i, j \in \{1, \dots, N\}$ ,

$$\sum_{k=1}^N (B_{ik} - A_{ik})x_k = 0 \geq (B_{ij} - A_{ij})x_j$$

with  $B_{ij} \geq A_{ij} \geq 0$  and  $x_j > 0$ , which is only possible if  $A_{ij} = B_{ij}$ . (This property reinforces lemma A1.2 in the case of primitive matrices).  $\square$

We seek to interpret the analysis of asymptotic behavior by way of an analogy with the scalar case, by showing a simple criterion that makes it possible to distinguish between cases of population extinction and explosion. To this end, we decompose

$$L = F + T,$$

$$F = \begin{pmatrix} f_1 & f_2 & \cdots & f_n \\ 0 & \ddots & & 0 \\ \vdots & & & \vdots \\ 0 & \ddots & & 0 \end{pmatrix}, T = \begin{pmatrix} 0 & & & & 0 \\ t_{2,1} & \ddots & & & \vdots \\ 0 & t_{3,2} & \ddots & & 0 \\ \vdots & & \ddots & & \vdots \\ 0 & & & t_{n,n-1} & 0 \end{pmatrix}$$

and we assume (for the moment) that  $R = \rho(F(\mathbb{I} - T)^{-1}) > 0$ . We will see further below that this quantity is easy to calculate. Note that

$$(\mathbb{I} - T)^{-1} = \sum_{k=0}^{\infty} T^k$$

has non-negative coefficients, since those of  $T$  are also non-negative. It is valid to express things in this way because the coefficients of the matrix  $T$  are in  $[0, 1[$  and, in

particular,  $\|T\|_\infty < 1$ . Likewise,  $A = F(\mathbb{I} - T)^{-1} \neq 0$  has non-negative coefficients and its spectral radius  $R$  is strictly positive. Therefore,  $A$  has an eigenvector  $z \neq 0$ , whose coordinates are non-negative, and is associated with the eigenvalue  $R$ , its spectral radius, according to corollary 1.6. Thus, we obtain

$$\begin{aligned} \frac{F}{R}(\mathbb{I} - T)^{-1}z &= z = (\mathbb{I} - T) \underbrace{(\mathbb{I} - T)^{-1}z}_{=\phi} \\ &= \sum_{k=0}^{\infty} T^k z \end{aligned}$$

where the coordinates of  $\phi \neq 0$  are non-negative. We can rewrite this equation as

$$\left(T + \frac{F}{R}\right)\phi = \phi.$$

However, just as  $T + F$  is primitive, the matrix  $T + F/R$  is primitive, and we have found an eigenvector  $\phi$  with non-negative coordinates for that matrix, associated with the eigenvalue 1. It follows from the Perron–Frobenius theorem that  $\rho(T + F/R) = 1$ . In the same way, we show that  $\phi$  is the eigenvector of  $(RT + F)$  for the eigenvalue  $R$ , and thus  $\rho(RT + F) = R$ . The number  $R$  allows us to describe the asymptotic behavior of the discrete dynamic system. We distinguish between three cases:

– if  $R = 1$ , then  $\rho_L = 1 = R$ . In this case, the total population converges towards a positive finite value;

– if  $R > 1$ , we have  $F_{ij}/R < F_{ij}$  and  $RT_{ij} > T_{ij}$ . It follows that  $T_{ij} + F_{ij}/R \leq T_{ij} + F_{ij} \leq RT_{ij} + F_{ij}$  and lemma 1.6 ensures that  $\rho(T + F/R) = 1 < \rho(T + F) = \rho_L < \rho(RT + F) = R$ . In this case, the total population tends towards  $+\infty$ ;

– if  $R < 1$ , we also obtain  $\rho(T + F/R) = 1 > \rho(T + F) = \rho_L > \rho(RT + F) = R$ . In this case, the total population tends to 0.

It remains to find the number  $R$  for the Leslie matrix. We determine the inverse of  $\mathbb{I} - T$ : if  $y = (\mathbb{I} - T)x$ , the relations

$$\begin{aligned} x_1 &= y_1, \\ x_2 - \tau_{21}x_1 &= y_2, \\ \dots \\ x_m - \tau_{mm-1}x_{m-1} &= y_m, \end{aligned}$$

lead to

$$\begin{aligned} x_1 &= y_1, \\ x_2 &= y_2 + \tau_{21}y_1, \\ \dots \\ x_m &= y_m + \tau_{mm-1}(y_{m-1} + \tau_{m-1m-2}(y_{m-2} + \dots(y_2 + \tau_{21}y_1)\dots)), \end{aligned}$$

so

$$(\mathbb{I} - T)^{-1} = \begin{pmatrix} 1 & & & & & \\ & 0 & \cdots & & & 0 \\ t_{2,1} & & \ddots & & & \\ \vdots & & & \ddots & & \\ & & & & \ddots & 0 \\ t_{21}\tau_{32}\dots t_{mm-1} & t_{32}\dots t_{mm-1} & \cdots & t_{mm-1} & 1 \end{pmatrix}.$$

The matrix product therefore takes a simpler form

$$F(\mathbb{I} - T)^{-1} = \begin{pmatrix} R & * & \cdots & * \\ 0 & \cdots & & 0 \\ \vdots & & & \vdots \\ 0 & \cdots & & 0 \end{pmatrix}$$

from which we can deduce the spectral radius

$$R = f_1 + f_2 t_{21} + \dots + f_m t_{21} t_{32} \dots t_{mm-1}.$$

The total population's behavior – extinction or explosion – is uniquely determined by this quantity.

### 1.2.5.2. Model in continuous time

We now pass from the description in discrete time to that in continuous time in terms of ordinary differential equations:  $X^{(k+1)} - X^{(k)} = (L - \mathbb{I})X^{(k)}$  defines the population's variation by unit of time  $t^{k+1} - t^k$ . Assuming that this unit of time is small with respect to the observation scale, we are driven to consider  $\frac{X^{(k+1)} - X^{(k)}}{\delta t} = (L - \mathbb{I})X^{(k)}$ , with  $0 < \delta t \ll 1$ , which can be interpreted as a discrete approximation (which is just the Euler scheme) for the differential system

$$\frac{d}{dt}X = (L - \mathbb{I})X, \quad X|_{t=0} = X_{\text{Init}}. \quad [1.17]$$

It is a simple homogeneous linear equation with constant coefficients. The Picard–Lindelöf theorem (linear version) applies and the problem has a unique solution defined on  $\mathbb{R}$ . We know that this solution is defined for all  $t \in \mathbb{R}$  by the relation

$$X(t) = \exp(t(L - \mathbb{I}))X_{\text{Init}}.$$

Because of their physical interpretation, the components  $x_j(t)$  are non-negative numbers. The differential system must preserve this property. We write  $Y(t) = e^t X(t)$ , which satisfies  $\frac{d}{dt}Y = LY$  and  $Y(t) = e^{Lt}Y(0) = e^{Lt}X_{\text{Init}}$ . However, the matrix  $L$  has non-negative coefficients, so for every  $t \geq 0$ ,  $e^{Lt}$  also has non-negative coefficients. Therefore, if the components of  $X_{\text{init}}$  are non-negative, those of  $Y(t)$  are also non-negative, as well as those of  $X(t)$ . This remark, which is necessary to make the model coherent, will be crucial for analyzing the problem (for more details and further development, see [RAP 12]).

Note that the spectrum of the differential system's matrix  $\tilde{L} = L - \mathbb{I}$  can be obtained from the spectrum of  $L$  from a simple shift, with the same eigenvectors

$$\sigma(\tilde{L}) = \{\lambda - 1, \lambda \in \sigma(L)\}.$$

Let  $\rho_L$  be the spectral radius of  $L$ . Then, by corollary 1.7,  $\rho = \rho_L - 1$  is an eigenvalue of  $\tilde{L}$  and has an eigenvector  $V$  whose components are strictly positive. Furthermore,  $\tilde{L}^\top$  has an eigenvector  $\phi$  associated with  $\rho$ , whose components are strictly positive. We can assume that

$$|V| = 1, \quad V \cdot \phi = 1.$$

Interestingly, the system [1.17] leaves the quantity  $e^{-\rho t} X(t) \cdot \phi$  unchanged.

LEMMA 1.7.– For all  $t \geq 0$ , we have  $e^{-\rho t} X(t) \cdot \phi = X_{\text{Init}} \cdot \phi$ .

PROOF.– We simply calculate

$$\frac{d}{dt}[X(t) \cdot \phi] = \frac{d}{dt}X(t) \cdot \phi = \tilde{L}X(t) \cdot \phi = X(t) \cdot \tilde{L}^\top \phi = \rho X(t) \cdot \phi.$$

Because  $X(t) \cdot \phi|_{t=0} = X_{\text{Init}} \cdot \phi$ , it follows that for all  $t \in \mathbb{R}$ ,  $X(t) \cdot \phi = e^{\rho t} X_{\text{Init}} \cdot \phi$ .  $\square$

We now return to the Jordan–Chevalley decomposition of the matrix  $\tilde{L}$ : there exist  $D$  and  $N$ , such that  $\tilde{L} = D + N$  and

$$-DN = ND,$$

- there exists an invertible  $P \in \mathcal{M}(\mathbb{C}^n)$ , such that

$$PDP^{-1} = \begin{pmatrix} \rho & 0 & \cdots & 0 \\ 0 & \lambda_1 & & \\ \vdots & & \ddots & \\ 0 & 0 & \cdots & \lambda_{n-1} \end{pmatrix}$$

where  $\lambda_j$  are the eigenvalues of  $\tilde{L}$ , which are distinct from  $\rho$ .

- there exists an  $r \in \mathbb{N}$  that satisfies  $N^r = 0$ .

We can also use the Jordan form of the matrix  $\tilde{L}$ , which is of the form

$$\begin{pmatrix} \rho & 0 & \cdots & 0 \\ 0 & \mathcal{J}_1 & & \\ \vdots & & \ddots & \\ 0 & 0 & \cdots & \mathcal{J}_\ell \end{pmatrix}$$

where the upper left-hand block has a size of  $1 \times 1$  and the blocks  $\mathcal{J}_k$  are of the form

$$\mathcal{J}_k = \begin{pmatrix} \lambda & 1 & 0 & \cdots & 0 \\ 0 & & & & \\ \vdots & & & & \\ 0 & 0 & \cdots & 0 & \lambda \end{pmatrix}$$

with  $\lambda$ , an eigenvalue of  $\tilde{L}$ .

We thus obtain

$$X(t) = e^{\tilde{L}t} X_{\text{init}} = P^{-1} \begin{pmatrix} e^{\rho t} & 0 & \cdots & 0 \\ 0 & e^{\lambda_1 t} & & \\ \vdots & & \ddots & 0 \\ 0 & \cdots & 0 & e^{\lambda_{n-1} t} \end{pmatrix} P \sum_{k=0}^{r-1} \frac{t^k \tilde{N}^k}{k!} X_{\text{Init.}}$$

Thus, for each element of vector  $X(t)$ , we have an expression of the form

$$x_j(t) = P_{\rho,j}(t)e^{\rho t} + \sum_{k=1}^{n-1} P_{\lambda_k,j}(t)e^{\lambda_k t}$$

where the functions  $t \mapsto P_{\lambda_k,j}(t)$  are polynomials. It follows that

$$e^{-\rho t} x_j(t) = P_{\rho,j}(t) + \sum_{k=1}^{n-1} P_{\lambda_k,j}(t)e^{(\lambda_k - \rho)t}.$$

However, each eigenvalue  $\lambda_k \neq \rho$  of  $\tilde{L}$  can be written as  $\lambda_k = \mu_k - 1$ , where  $\mu_k$  is an eigenvalue of  $L$  that satisfies  $\operatorname{Re}(\mu_k) \leq |\operatorname{Re}(\mu_k)| \leq |\mu_k| < \rho_L$ . It follows that  $\operatorname{Re}(\lambda_k - \rho) = \operatorname{Re}(\mu_k - \rho_L) < 0$ , and therefore

$$\lim_{t \rightarrow \infty} \left( \sum_{k=1}^{n-1} P_{\lambda_k,j}(t)e^{(\lambda_k - \rho)t} \right) = 0.$$

Moreover,  $\sum_{j=1}^n e^{\rho t} x_j(t) \phi_j = X_{\text{Init}} \cdot \phi$  is a sum of strictly positive terms, so for each  $j \in \{1, \dots, n\}$ , the function  $t \mapsto e^{\rho t} x_j(t)$  is bounded. We infer that the polynomial functions  $t \mapsto P_{\rho,j}(t)$  are bounded. They are therefore constant functions, denoted as  $p_j \in \mathbb{C}$ , and we have

$$\lim_{t \rightarrow \infty} e^{-\rho t} x_j(t) = p_j.$$

However, note that

$$\frac{d}{dt} (e^{-\rho t} x_j(t)) = -\rho e^{-\rho t} x_j(t) + [\tilde{L} e^{-\rho t} x(t)]_j \xrightarrow[t \rightarrow \infty]{} -\rho p_j + [\tilde{L} p]_j = [(\tilde{L} - \rho \mathbb{I}) p]_j,$$

by denoting the vector of coordinates  $p_1, \dots, p_n$  as  $p$ . This quantity can also be calculated in the following form:

$$\begin{aligned}\frac{d}{dt} (e^{-\rho t} x_j(t)) &= \frac{d}{dt} \left( p_j + \sum_{k=1}^{n-1} P_{\lambda_k, j}(t) e^{(\lambda_k - \rho)t} \right) \\ &= \sum_{k=1}^{n-1} e^{(\lambda_k - \rho)t} ((\lambda_k - \rho) P_{\lambda_k, j}(t) + P'_{\lambda_k, j}(t)) \xrightarrow[t \rightarrow \infty]{} 0\end{aligned}$$

again using the fact that  $\text{Re}(\lambda_k) < \rho$ . The vector  $p = (p_1, \dots, p_n)$  therefore satisfies  $(\tilde{L} - \rho \mathbb{I})p = 0$ , which implies that  $p = \alpha V \in \text{Ker}(\tilde{L} - \rho \mathbb{I})$  for some  $\alpha \in \mathbb{C}$ . Finally, we identify the constant  $\alpha$  by using the conservation law

$$e^{\rho t} X(t) \cdot \phi = X_{\text{Init}} \cdot \phi \xrightarrow[t \rightarrow \infty]{} p \cdot \phi = \alpha V \cdot \phi = \alpha = X_{\text{Init}} \cdot \phi$$

(which is a strictly positive real number). In summary, we have demonstrated the following relation:

**PROPOSITION 1.4.**— Let  $t \mapsto X(t)$  be the solution of [1.17]. Then, when  $t \rightarrow \infty$ ,  $X(t)$  behaves like  $X_{\text{Init}} \cdot \phi e^{\rho t} V$ .

In terms of population, this means that there will be extinction whenever  $\rho < 0$ , that is to say,  $\rho_L < 1$ . There will also be exponential growth whenever  $\rho > 0$ , that is to say,  $\rho_L > 1$ . This kind of behavior has been most famously studied by Malthus [MAL 98], and identifying a leading eigenvalue for a transition matrix therefore plays a very important role in several biological applications. Another illustration of this type of method, with important practical applications, involves the analysis of criticality in neutron transport [DAU 84, Chapter 1, Section 5, Paragraph 3] and its applications to nuclear security problems.

**NOTE.**— It is useful to describe in detail a somewhat different approach to systems of differential equations

$$\frac{d}{dt} y = A y$$

where  $A$  is a matrix for which all coefficients are strictly positive (and therefore theorem 1.13 applies directly). In this case, we can present dissipation properties for the differential system that translates the convergence to the asymptotic state. The Perron–Frobenius theorem allows us to effectively find  $V = (v_1, \dots, v_n)$  and  $\phi = (\phi_1, \dots, \phi_n)$ , the positive-coordinate eigenvectors of  $A$  and  $A^\top$ , respectively,

which are associated with the spectral radius and normalized by  $|V| = 1$ ,  $V \cdot \phi = 1$ . We calculate

$$\begin{aligned} & \frac{1}{2} \frac{d}{dt} \sum_{j=1}^n \left( e^{-\rho t} \frac{y_j(t)}{v_j} \right)^2 v_j \phi_j \\ &= \sum_{j=1}^n e^{-2\rho t} \frac{y_j(t)}{v_j} \left( \frac{-\rho y_j(t)}{v_j} + \frac{(Ay(t))_j}{v_j} \right) v_j \phi_j \\ &= \sum_{j=1}^n e^{-2\rho t} \left( -\rho v_j \phi_j \left( \frac{y_j(t)}{v_j} \right)^2 + \sum_{k=1}^n a_{j,k} y_k(t) \frac{y_j(t)}{v_j} \phi_j \right) \end{aligned}$$

However, we have

$$\rho v_j \phi_j = \frac{1}{2} \sum_{k=1}^n a_{j,k} v_k \phi_j + \frac{1}{2} \sum_{k=1}^n a_{k,j} \phi_k v_j$$

which allows us to write

$$\begin{aligned} & \frac{1}{2} \frac{d}{dt} \sum_{j=1}^n \left( e^{-\rho t} \frac{y_j(t)}{v_j} \right)^2 v_j \phi_j \\ &= -\frac{1}{2} \sum_{j=1}^n \sum_{k=1}^n e^{-2\rho t} a_{j,k} \left( v_k \phi_j \left( \frac{y_j(t)}{v_j} \right)^2 + v_k \phi_j \left( \frac{y_k(t)}{v_k} \right)^2 - 2v_k \phi_j \frac{y_j(t)}{v_j} \frac{y_k(t)}{v_k} \right) \\ &= -\frac{1}{2} \sum_{j=1}^n \sum_{k=1}^n a_{j,k} v_k \phi_j \left( e^{-\rho t} \frac{y_k(t)}{v_k} - e^{-\rho t} \frac{y_j(t)}{v_j} \right)^2. \end{aligned}$$

We denote the right-hand term as  $-D(t)$ . This calculation shows that the positive-valued function  $H : t \mapsto \sum_{j=1}^n \left( e^{-\rho t} \frac{y_j(t)}{v_j} \right)^2 v_j \phi_j$  is decreasing on  $\mathbb{R}^+$  and thus has a limit  $H_\infty \geq 0$  when  $t \rightarrow \infty$ , and also  $t \mapsto D(t) \in L^2(\mathbb{R}^+)$ . Let  $(t^{(\nu)})_{\nu \in \mathbb{N}}$  be an increasing sequence that tends towards  $+\infty$ . We can apply the Arzela–Ascoli theorem to the sequence of functions  $z_j^{(\nu)}(t) = e^{-\rho(t+t^{(\nu)})} \frac{y_j(t+t^{(\nu)})}{v_j}$  because these functions are uniformly bounded. Even if it means extracting a subsequence denoted as  $(t^{\nu_m})_{m \in \mathbb{N}}$ , we may assume that  $z_j^{(\nu_m)}$  converges uniformly on  $[0, T]$ ,  $0 < T < \infty$ , towards a limit  $\ell_j^\infty(t)$ . However, we have

$$H(t + t^{(\nu_m)}) + \int_{t^{(\nu_m)}}^{t+t^{(\nu_m)}} D(s) ds = H(t + t^{(\nu_m)}) + \int_0^t D(s + t^{(\nu_m)}) ds = H(t^{(\nu_m)}).$$

So, by letting  $m \rightarrow \infty$ , we get

$$\lim_{m \rightarrow \infty} \int_0^t D(s + t^{(\nu_m)}) \, ds = 0.$$

Since  $D(t) \geq 0$ , by Fatou's lemma, it follows that

$$D(s + t^{(\nu_m)}) \xrightarrow[m \rightarrow \infty]{} 0 = \sum_{k,j=1}^n a_{j,k} v_k \phi_j \left| \ell_j^\infty(s) - \ell_k^\infty(s) \right|^2.$$

We assume that the coefficients  $a_{j,k}$  are strictly positive and the Perron–Frobenius theorem ensures that the coordinates  $v_k$  and  $\phi_j$  are also strictly positive. It follows that  $\ell_j^\infty(t) = \ell^\infty(t)$  does not depend on  $j$ . The conservation relation allows us to identify the function  $\ell^\infty$ :

$$\begin{aligned} \sum_{j=1}^n e^{-\rho(t+t^{(\nu_m)})} y_j(t + t^{(\nu_m)}) \phi_j &= \sum_{j=1}^n e^{-\rho(t+t^{(\nu_m)})} \frac{y_j(t + t^{(\nu_m)})}{v_j} v_j \phi_j \\ &= y_{\text{Init}} \cdot \phi \xrightarrow[m \rightarrow \infty]{} \ell^\infty(t) v \cdot \phi \end{aligned}$$

gives

$$\ell^\infty(t) = \frac{y_{\text{Init}} \cdot \phi}{v \cdot \phi} = y_{\text{Init}} \cdot \phi,$$

which does not depend on  $j$  or  $t$ . Since the limit is defined uniquely, we have identified the asymptotic behavior of the system when  $t \rightarrow \infty$ :  $y(t)$  behaves like  $e^{\rho t} \ell^\infty v$  when  $t \rightarrow \infty$ .

This kind of approach can be related to several different contexts: Lyapunov functional for dynamic systems or “entropy functional” for systems with partial derivatives. For extensions and applications motivated by biology and population dynamics, the reader may refer to P. Michel’s dissertation [MIC 05].

### 1.2.5.3. Proof of the Perron–Frobenius theorem

Before advancing to the proof of the Perron–Frobenius theorem, we note that corollary 1.7 can be deduced from the following general result.

LEMMA 1.8.– Let  $A \in \mathcal{M}_n(\mathbb{C})$ . For every polynomial  $P$ , we have

$$\sigma(P(A)) = \{P(\lambda), \lambda \in \sigma(A)\}.$$

PROOF.– For every  $\lambda \in \mathbb{C}$ , the mapping  $z \in \mathbb{C} \mapsto \lambda - P(z)$  is a polynomial, which can be written as

$$\lambda - P(z) = a \prod_{j=1}^m (z - z_j)$$

where  $m$  is the degree of the polynomial and the roots  $z_j$  depend on  $\lambda$ .

Suppose that  $\lambda \neq P(\mu)$  for every  $\mu \in \sigma(A)$ . Then, for every  $j \in \{1, \dots, m\}$ , the matrix  $A - z_j \mathbb{I}$  is invertible (otherwise, there would exist a  $j_0$ , such that  $z_{j_0} \in \sigma(A)$  and  $\lambda - P(z_{j_0}) = 0$ , which is a contradiction) and  $\lambda \mathbb{I} - P(A)$  is written as the product of invertible operators, so it is invertible. In other words, we have  $\lambda \notin \sigma(P(A))$ .

For the reciprocal result, we provide a general analysis that treats the case of operators in infinite dimensions, given that the matrix solution case provides a faster argument. Let  $\lambda \notin \sigma(P(A))$ :  $\lambda \mathbb{I} - P(A)$  is invertible.

– Let  $x \in \text{Ker}(A - z_j \mathbb{I})$ :

$$(\lambda \mathbb{I} - P(A))x = a \prod_{k \neq j} (A - z_k \mathbb{I})(A - z_j \mathbb{I})x = a \prod_{k \neq j} (A - z_k \mathbb{I})0 = 0.$$

Since  $\lambda \mathbb{I} - P(A)$  is invertible, it follows that  $x = 0$ , so for every  $j$ , the operator  $A - z_j \mathbb{I}$  is injective.

– Let  $y$  be a given vector. There exists an  $x$ , such that  $(\lambda \mathbb{I} - P(A))x = y = (A - z_j \mathbb{I})\tilde{x}$ , with  $\tilde{x} = a \prod_{k \neq j} (A - z_k \mathbb{I})x$ , which proves that  $(A - z_j \mathbb{I})$  is surjective.

– Finally, we can write

$$\mathbb{I} = (\lambda \mathbb{I} - P(A))^{-1}(\lambda \mathbb{I} - P(A)) = (\lambda \mathbb{I} - P(A))^{-1}a \prod_{k \neq j} (A - z_k \mathbb{I}) (A - z_j \mathbb{I})$$

in such a way that  $(A - z_j \mathbb{I})^{-1} = (\lambda \mathbb{I} - P(A))^{-1}a \prod_{k \neq j} (A - z_k \mathbb{I})$  is bounded as the product of bounded operators.

These three points show that if  $\lambda \notin \sigma(P(A))$ , then for all  $j \in \{1, \dots, m\}$ ,  $z_j \notin \sigma(A)$ . Suppose that  $\lambda \notin \sigma(P(A))$  is written as  $\lambda = P(\mu)$ , with  $\mu \in \sigma(A)$ . We obtain  $\lambda - P(\mu) = 0 = a \prod_{j=1}^m (\mu - z_j)$ , so there is an index  $k$ , such that  $\mu = z_k$ , which would contradict the fact that the  $z_j$ s are not in the spectrum of  $A$ . We conclude that  $\lambda \neq P(\mu)$  for all  $\mu \in \sigma(A)$ .

We will, however, remain attentive to the fact that we cannot identify the eigenvectors in general because we can have  $\lambda_1 \neq \lambda_2$  eigenvalues of  $A$ , with  $\lambda_1^p = \lambda_2^p \in \sigma(A^p)$ . Nevertheless, if the Perron–Frobenius theorem applies to  $A^p$  in

such a way that the dominant eigenvalue  $\rho$  of  $A^p$  is simple, then there exists an eigenvalue  $\lambda$  of  $A$ , such that  $\lambda^p = \rho$  and the associated eigenspace has dimension 1.  $\square$

**PROOF OF THEOREM 1.13.** We begin by noting that if all the coefficients of matrix  $A$  are strictly positive, then for every  $x \neq 0$  whose components are non-negative, the components of  $Ax$  are strictly positive. (The reciprocal statement is also true: if  $A$  has this property, then by applying  $A$  to the vectors of the canonical basis, we show that the coefficients  $A_{k,j}$  of  $A$  are strictly positive). We introduce the sets

$$\mathcal{C} = \{x = (x_j)_{1 \leq j \leq n} \in \mathbb{R}^n, \text{ such that } x_j \geq 0 \text{ for all } j \in \{1, \dots, n\}\},$$

$$\mathcal{E} = \{t \geq 0, \text{ such that there exists } x \in \mathcal{C} \setminus \{0\} \text{ satisfying } Ax - tx \in \mathcal{C}\}.$$

We will employ the following observations:

– The set  $\mathcal{E}$  contains at least 0 by assumption on  $A$ . Then, by denoting  $e$  the vector for which all components are equal to 1,  $Ae$  has strictly positive coordinates, and for a sufficiently small  $\varepsilon > 0$ , we have  $Ae - \varepsilon e \in \mathcal{C}$ , so  $\varepsilon \in \mathcal{E}$ . For example,

$$\varepsilon = \frac{1}{2} \min_{k \in \{1, \dots, n\}} \sum_{j=1}^n a_{k,j} > 0$$

works. Therefore,  $\mathcal{E}$  is not reduced to  $\{0\}$ .

– Let  $t \in \mathcal{E}$ ,  $x \in \mathcal{C} \setminus \{0\}$ , such that  $Ax - tx \in \mathcal{C}$ , and let  $t' \in [0, t]$ . Therefore,  $Ax - t'x = Ax - tx + (t - t')x \in \mathcal{C}$  because  $t - t' \geq 0$ , and the coordinates of  $x$  are non-negative. This ensures that  $t' \in \mathcal{E}$ , and  $\mathcal{E}$  is therefore an interval.

– Let  $x \in \mathcal{C} \setminus \{0\}$ . Since  $a_{k,j}$  and  $x_j$  are non-negative, we get

$$\begin{aligned} 0 \leq Ax \cdot e &= \sum_{k=1}^n \sum_{j=1}^n a_{k,j} x_j = \sum_{j=1}^n \left( \sum_{k=1}^n a_{k,j} \right) x_j \\ &\leq \left( \max_{m \in \{1, \dots, n\}} \sum_{k=1}^n a_{k,m} \right) \sum_{j=1}^n x_j = \left( \max_{m \in \{1, \dots, n\}} \sum_{k=1}^n a_{k,m} \right) x \cdot e, \end{aligned}$$

where  $x \cdot e > 0$ . Let  $t \in \mathcal{E}$  and  $x \in \mathcal{C} \setminus \{0\}$ , such that  $Ax - tx \in \mathcal{C} \setminus \{0\}$ . Therefore,  $(Ax - tx) \cdot e \geq 0$ , that is to say,  $tx \cdot e \leq Ax \cdot e$ . It follows that

$$0 \leq t \leq \max_{j \in \{1, \dots, n\}} \sum_{k=1}^n a_{k,j}$$

and  $\mathcal{E}$  is therefore bounded.

– Finally, let  $(t_p)_{p \in \mathbb{N}}$  be a sequence of elements of  $\mathcal{E}$  that converges to  $t \in \mathbb{R}$ . For every  $p \in \mathbb{N}$ , there exists a  $x_p \in \mathcal{C} \setminus \{0\}$ , such that  $Ax_p - t_p x_p \in \mathcal{C}$ . Because the  $x_p$ s are non-zero, we can consider the sequence  $(y_p)_{p \in \mathbb{N}} = \left(\frac{x_p}{|x_p|}\right)_{p \in \mathbb{N}}$ , which is bounded in a finite dimension space: by the Bolzano–Weierstrass theorem, we can extract a subsequence  $(y_{p_k})_{k \in \mathbb{N}}$  that converges to  $y$ , with  $|y| = 1$ . Since the coordinates of  $y_p$  are non-negative, those of  $y$  are also non-negative, and  $y \in \mathcal{C} \setminus \{0\}$ . Because for every  $k \in \mathbb{N}$ , we have  $Ay_{p_k} - t_{p_k} y_{p_k} \in \mathcal{C}$ , by passing to the limit, we deduce that  $Ay - ty \in \mathcal{C}$ . Therefore,  $t \in \mathcal{E}$ , and  $\mathcal{E}$  is closed.

Since the set  $\mathcal{E}$  is a closed and bounded interval that is not reduced to  $\{0\}$ ,

$$0 < \rho = \max \mathcal{E} < \infty$$

is well defined and belongs to  $\mathcal{E}$ . In particular, there exists an  $x \neq 0$  with non-negative coordinates, such that  $Ax \geq \rho x$ . If  $y = Ax - \rho x \neq 0$ , then  $y \in \mathcal{C} \setminus \{0\}$  and therefore  $Ay$  has strictly positive coordinates. For a small enough  $\epsilon > 0$ , we have  $Ay - \epsilon Ax = A(Ax) - (\rho + \epsilon)Ax \in \mathcal{C}$ . For example,

$$\epsilon = \frac{1}{2} \frac{\min_{j \in \{1, \dots, n\}} (Ay)_j}{\max_{j \in \{1, \dots, n\}} (Ax)_j}$$

works. Because  $Ax \in \mathcal{C} \setminus \{0\}$ , it contradicts the definition of  $\rho$ , and we conclude that necessarily  $Ax = \rho x$ . However,  $x \neq 0$  has non-negative coordinates, so  $Ax$  has strictly positive coordinates. Moreover, since  $\rho > 0$ , we have  $x_j > 0$  for every  $j \in \{1, \dots, n\}$ . The vector  $x$  obtained is indeed an eigenvector of  $A$  associated with the eigenvalue  $\rho$  and its components are all strictly positive.

In order to show that  $\text{Ker}(A - \rho \mathbb{I})$  has dimension 1, we will use the following result.

**LEMMA 1.9.–** Let  $(w_1, \dots, w_n) \in \mathbb{C}^n$ , such that  $|w_1 + \dots + w_n| = |w_1| + \dots + |w_n|$ . Therefore, there exists  $\theta \in [0, 2\pi[$ , such that for every  $j \in \{1, \dots, n\}$ , we have  $w_j = e^{i\theta} |w_j|$ .

**PROOF.–** We begin by noting that the assumption on the vector  $w$  implies that

$$\begin{aligned} |w_1 + \dots + w_n|^2 &= (w_1 + \dots + w_n)(\overline{w_1} + \dots + \overline{w_n}) \\ &= \sum_{k=1}^n |w_k|^2 + 2 \sum_{1 \leq j < \ell \leq n} \operatorname{Re}(\overline{w_j} w_\ell) \\ &= (|w_1| + \dots + |w_n|)^2 = \sum_{k=1}^n |w_k|^2 + 2 \sum_{1 \leq j < \ell \leq n} |w_j| |w_\ell|. \end{aligned}$$

However, for every  $z \in \mathbb{C}$ , we have  $\operatorname{Re}(z) \leq |z|$ , with equality occurring if and only if  $z \in \mathbb{R}^+$ . It follows that for  $j \neq \ell \in \{1, \dots, n\}^2$ , we have  $\operatorname{Re}(\overline{w_j}w_\ell) = |w_j||w_\ell|$  and  $\overline{w_j}w_\ell \in \mathbb{R}^+$ . We distinguish between the following two cases:

– if  $w_1 = \dots = w_n = 0$ , then  $\theta = 0$  works;

– otherwise, let  $\ell \in \{1, \dots, n\}$ , such that  $w_\ell \neq 0$ , and  $\theta \in [0, 2\pi[$ , such that  $w_\ell = e^{i\theta}|w_\ell|$ . Let  $j \neq \ell$ . If  $w_j = 0$ , we can always write  $w_j = e^{i\theta_j}|w_j|$ . Otherwise, we can write  $w_j = e^{i\theta_j}|w_j|$ , with  $0 \leq \theta_j < 2\pi$ , such that, by this discussion, we have

$$\operatorname{Re}(\overline{w_j}w_\ell) = |w_j||w_\ell| \cos(\theta - \theta_j) = |w_j||w_\ell|$$

which implies that  $\theta - \theta_j$  is of the form  $2m\pi$ ,  $m \in \mathbb{Z}$ . In fact, it follows that  $\theta_j = \theta$ .

In both cases, we have found a real number  $\theta \in [0, 2\pi[$ , such that  $\forall j \in \{1, \dots, n\}$ , we have  $w_j = e^{i\theta}|w_j|$ .  $\square$

For  $z = (z_1, \dots, z_n) \in \mathbb{C}^n$ , we let  $|z|$  denote the vector of coordinates  $|z_1|, \dots, |z_n|$ . The coordinates of  $A|z|$  are  $(\sum_{j=1}^n a_{k,j}|z_j|)_{k \in \{1, \dots, n\}}$  and the coordinates of  $|Az|$  are  $(|\sum_{j=1}^n a_{k,j}z_j|)_{k \in \{1, \dots, n\}}$ . If  $|Az| = A|z|$ , we have  $|\sum_{j=1}^n a_{1,j}z_j| = \sum_{j=1}^n a_{1,j}|z_j|$ . We can therefore apply lemma 1.9 to the vector of coordinates  $w_j = a_{1,j}z_j$ : there exists  $\theta \in [0, 2\pi[$ , such that  $\forall j \in \{1, \dots, n\}$ ,  $a_{1,j}z_j = e^{i\theta}|a_{1,j}z_j|$ . However, for all  $j \in \{1, \dots, n\}$ ,  $a_{1,j} > 0$ , so  $z_j = e^{i\theta}|z_j|$ . Reciprocally, if there exists  $\theta \in [0, 2\pi[$ , such that  $\forall j \in \{1, \dots, n\}$ ,  $z_j = e^{i\theta}|z_j|$ , so

$$\forall k \in \{1, \dots, n\}, \quad \left| \sum_{j=1}^n a_{k,j}z_j \right| = \sum_{j=1}^n a_{k,j}|z_j|$$

which proves that  $A|z| = |Az|$  if and only if there exists a  $\theta \in [0, 2\pi[$ , such that  $z_j = e^{i\theta}|z_j|$  for every  $j \in \{1, \dots, n\}$ .

We will use the fact that for every vector  $z \in \mathbb{C}^n$ , we have

$$\left| \sum_{j=1}^n a_{k,j}z_j \right| \leq \sum_{j=1}^n a_{k,j}|z_j| \quad \text{for all } k \in \{1, \dots, n\} \quad [1.18]$$

this is to say,  $A|z| - |Az| \in \mathcal{C}$ . We know that  $\operatorname{Ker}(A - \rho\mathbb{I})$  contains a vector  $x$  with strictly positive coordinates, which we already assume is normalized so that  $|x| = 1$ . We decompose the argument as follows:

– Let  $z$  be an eigenvector of  $A$  for the eigenvalue  $\lambda \in \mathbb{C}$ . By equation [1.18], we have  $|Az| = |\lambda| |z| \leq A|z|$ , which proves that  $|\lambda| \leq \rho$ . The spectrum of  $A$  is contained in the disk of radius  $\rho$ , which is therefore the spectral radius.

– Let us first suppose that  $\lambda = \rho$  and  $z \neq 0$  satisfy  $(x|z) = 0$ . Then,  $|z| \in \mathcal{C} \setminus \{0\}$  and  $|Az| = \rho|z|$ . By equation [1.18], we have  $A|z| - |Az| = A|z| - \rho|z| \in \mathcal{C}$ . However, we see that this, in fact, implies  $A|z| = \rho|z| = |Az|$ . We can thus deduce the existence of  $\theta \in [0, 2\pi[$ , such that  $z = e^{i\theta}|z|$ . Returning to the orthogonality condition, it follows that  $(z|x) = e^{-i\theta} \sum_{k=1}^n |z_k| x_k = 0$ , and since the  $x_k$ s are strictly positive, this is impossible. Therefore, we necessarily have  $z = 0$ .

– For every  $z \in \text{Ker}(A - \rho\mathbb{I})$ , we may decompose it as  $z = \alpha x + w$ , where  $\alpha \in \mathbb{C}$  and  $(w|x) = 0$ . Therefore,  $w = z - \alpha x \in \text{Ker}(A - \rho\mathbb{I})$ , and this discussion shows that  $w = 0$ . In this way, we deduce

$$\text{Ker}(A - \rho\mathbb{I}) = \text{Span}\{x\}.$$

– Finally, let  $\lambda \neq \rho$  be an eigenvalue, such that  $|\lambda| = \rho$ , and let  $z$  be an associated eigenvector. Thus,  $A|z| - |Az| = A|z| - \rho|z| \in \mathcal{C}$ , by equation [1.18]. This implies that  $|z| \in \text{Ker}(A - \rho\mathbb{I}) = \text{Span}\{x\}$ ; moreover, there exists a  $\theta \in [0, 2\pi[$ , such that  $z = e^{i\theta}|z|$ . It follows that  $z \in \text{Span}\{x\}$ , which contradicts  $\lambda \neq \rho$ . Finally, for every eigenvalue of  $A$ , we have

$$|\lambda| < \rho.$$

Let  $v$  be an eigenvector of  $A$  belonging to  $\mathcal{C}$  and associated with the eigenvalue  $\lambda$  of  $A$ , and let  $\phi$  be an eigenvector of  $A^\top$  with strictly positive coordinates associated with the eigenvalue  $\rho$ . ( $\phi$  exists because  $A$  and  $A^\top$  have the same eigenvalues, and the coefficients of  $A^\top$  are strictly positive: the results obtained for  $A$  therefore apply to  $A^\top$ ). Thus,  $\phi^\top Av = \lambda\phi^\top v = \lambda\phi \cdot v = (A^\top\phi)^\top v = \rho\phi \cdot v$ . However, since  $v$  and  $\phi$  belong to  $\mathcal{C}$  and  $v \neq 0$ ,  $\phi$  has strictly positive coordinates, and we have  $\phi \cdot v > 0$ . It follows that  $\lambda = \rho$ , and

$$v \in \text{Ker}(A - \rho\mathbb{I}) = \text{Span}\{x\},$$

which finishes the proof of theorem 1.13.  $\square$

**PROOF OF COROLLARY 1.6.** – For  $n \in \mathbb{N} \setminus \{0\}$ , we let

$$A^{(n)} = A + \frac{1}{n}E,$$

where  $E$  is the matrix for which all coefficients are equal to 1. Theorem 1.13 applies to  $A_n$  because its coefficients are strictly positive. Therefore, there exists a  $\rho^{(n)} = \rho(A^{(n)}) > 0$  and an eigenvector  $x^{(n)}$  with strictly positive coordinates, such that

$$A^{(n)}x^{(n)} = \rho^{(n)}x^{(n)}, \quad |x^{(n)}| = 1. \tag{1.19}$$

Using the monotony property of the spectral radius stated in lemma A1.2, we have

$$\rho(A) \leq \rho(A^{(m)}) = \rho^{(m)} \leq \rho(A^{(n)}) = \rho^{(n)} \quad \text{for all } m \geq n.$$

The sequence  $(\rho^{(n)})_{n \in \mathbb{N} \setminus \{0\}}$  is non-increasing and has a limit when  $n \rightarrow \infty$ , denoted as  $\bar{\rho}$ . We note that  $\bar{\rho} \geq \rho(A)$ . Moreover, since the sequence  $(x^{(n)})_{n \in \mathbb{N} \setminus \{0\}}$  is bounded, the Bolzano–Weierstrass theorem allows us to extract a convergent subsequence; we denote its limit as  $\lim_{\ell \rightarrow \infty} x^{(n_\ell)} = \bar{x}$ . The coordinates of  $\bar{x}$  are non-negative. Finally, by passing to the limit  $\ell \rightarrow \infty$  in equation [1.19], we obtain

$$A\bar{x} = \bar{\rho}\bar{x}, \quad |\bar{x}| = 1.$$

Therefore,  $\bar{x} \neq 0$  is an eigenvector of  $A$  that is associated with the eigenvalue  $\bar{\rho} \geq \rho(A)$ , which implies that  $\bar{\rho} = \rho(A)$ .  $\square$

In fact, in those statements, the key notion is that of *irreducibility*, which is related to the connectedness of the graph of the matrix  $A$ : the matrix  $A \in \mathcal{M}_n$  is associated with a graph composed of  $n$  vertices where an edge connects vertices  $i$  and  $j$  when  $A_{i,j} \neq 0$ . Moreover, we say that  $A$  is irreducible if for every ordered pair  $(i, j)$ , there exists a path (that is to say, a sequence of such edges) connecting  $i$  to  $j$ . Another way to state this property is to say that for every  $i, j$ , there exists an integer  $k(i, j)$ , such that  $[A^{k(i,j)}]_{i,j} > 0$ . The smallest integer  $k(i, j)$  satisfying that criterion is called the *path length* for the path connecting  $i$  to  $j$ . For an irreducible matrix  $A$  with non-negative coefficients, by developing

$$(\mathbb{I} + A)^k = \sum_{m=0}^k C_k^m A^m, \quad C_k^m = \frac{k!}{m!(k-m)!}$$

we realize that, for a large enough  $k$ , all the coefficients of  $(\mathbb{I} + A)^k$  are strictly positive:  $\mathbb{I} + A$  is primitive (and, clearly, the reciprocal statement is also true). Therefore, corollary 1.7 applies to  $\mathbb{I} + A$ , whose spectral radius is denoted as  $\rho$ . In particular,  $\rho - 1$  is the eigenvector of  $A$  and the associated eigensubspace is spanned by an eigenvector  $x$  with strictly positive coordinates. We recall from the proof of the Perron–Frobenius theorem that  $\rho$  can be characterized by

$$\begin{aligned} \rho^k &= \max \{t \geq 0, \text{such that there exists an } x \in \mathcal{C} \setminus \{0\} \\ &\quad \text{satisfying } (\mathbb{I} + A)^k x - tx \in \mathcal{C}\}. \end{aligned} \tag{1.20}$$

We note also that  $\rho > 1$ . Indeed, if  $x$  designates the eigenvector of  $(\mathbb{I} + A)$  with strictly positive coordinates and normalized by  $|x|_\infty = 1$ , assuming that  $x_{i_0} = 1$ , we obtain, for a large enough  $k$ ,

$$\rho^k x_{i_0} = \rho^k = [(\mathbb{I} + A)^k x]_{i_0} = x_{i_0} + \sum_{m=1}^k \sum_{\ell=1}^n C_k^m [A^m]_{i_0, \ell} x_\ell > x_{i_0} = 1.$$

It follows that the spectral radius of  $A$  is strictly positive. Let  $\lambda \in \mathbb{C}$  be an eigenvalue of  $A$  and  $z$  an associated eigenvector. Since the coefficients of  $A$  are positive or equal to zero, we can write

$$|\lambda z_j| = |\lambda| |z_j| = \left| \sum_{\ell=1}^n A_{j,\ell} z_\ell \right| \leq \sum_{\ell=1}^n A_{j,\ell} |z_\ell|,$$

which implies

$$(1 + |\lambda|) |z_j| \leq \sum_{\ell=1}^n (\delta_{j,\ell} + A_{j,\ell}) |z_\ell|.$$

Therefore, the vector  $y$  with (non-negative) coordinates  $|z_1|, \dots, |z_n|$  satisfies  $(1 + |\lambda|)y \leq (\mathbb{I} + A)y$  as well as  $(1 + |\lambda|)^k y \leq (\mathbb{I} + A)^k y$ . It follows from the definition in equation [1.20] of  $\rho$  that  $\rho - 1 > 0$  is the spectral radius of  $A$ . We have therefore demonstrated a version of theorem 1.13 for irreducible matrices with non-negative coefficients.

**COROLLARY 1.8.–** Let  $A \in M_n(\mathbb{R})$  be an *irreducible* matrix with non-negative coefficients. Then, the spectral radius  $\rho(A) > 0$  is a simple eigenvector of  $A$ , and there exists an eigenvector whose coordinates are strictly positive for that eigenvalue.

It is important to distinguish the assumptions and conclusions of the different versions of theorem 1.13, corollaries 1.6, 1.7 and 1.8. For a matrix with non-negative coefficients, as in the case of corollary 1.6, we have no information on the dimension of the eigensubspace associated with the spectral radius. There can exist non-collinear eigenvectors with non-negative coordinates. Furthermore, there can be several eigenvalues whose modulus is equal to the spectral radius. The identity matrix provides a simple example of this situation. For an irreducible matrix, as in corollary 1.8, the eigenspace associated with the spectral radius does indeed have dimension 1, and there is only one normalized eigenvector with strictly positive coordinates. However, there can be other eigenvalues whose modulus is equal to the spectral radius, as is shown by the example of the matrix

$$\begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}.$$

This matrix is irreducible, but is not primitive (its square is the identity matrix and its eigenvalues are  $+1$  and  $-1$ ). For more details on this theory and its interpretations in terms of graphs, the reader may refer to, for example, [HOR 90].

#### 1.2.5.4. Computation of the leading eigenvalue

Sections 1.2.5.1 and 1.2.5.2 show that all the asymptotic behavior of the Leslie system, which ultimately corresponds to the situations observed in practice, is directed by the dominant eigenvalue of the matrix  $L$ . The question therefore comes down to finding the largest eigenvalue of a matrix satisfying the assumptions of the Perron–Frobenius theorem: the *power method*, which we will now describe, provides a practical way to calculate  $\rho$  and  $V$ .

Suppose that:

- a)  $A \in \mathcal{M}_N$  is a diagonalizable matrix;
- b) there exists a unique eigenvalue  $\lambda_1$ , such that  $\rho = |\lambda_1|$ , the spectral radius of  $A$ : if we denote the eigenvalues of  $A$  as  $\lambda_1, \dots, \lambda_N$ , we have  $|\lambda_1| > |\lambda_{p+1}| \geq |\lambda_N|$ , where  $p = \dim(\text{Ker}(A - \lambda_1 \mathbb{I}))$ . In this case, we say that  $\lambda_1$  is the dominant eigenvalue of  $A$ .

We (randomly) choose two vectors  $x^{(0)} \neq 0$  and  $\phi \neq 0$ . We construct the sequence defined by

$$x^{(n+1)} = \frac{Ax^{(n)}}{\phi \cdot Ax^{(n)}},$$

where  $x^{(0)}$  is taken as the first iteration, while  $\phi$  defines a normalization factor. The role of this factor can be motivated as follows: we seek to construct a sequence that converges to an eigenvector. If we start specifically from  $y \in \text{Ker}(A - \lambda_1 \mathbb{I})$ , then  $A^k y = \lambda_1^k$  can oscillate, “explode” or vanish according to the value of  $\lambda_1$ . The normalization factor prevents this kind of behavior.

Thanks to assumption a), we can develop  $x^0$  on a basis  $\{\psi_1, \dots, \psi_N\}$  composed of eigenvectors, with the first  $p$  elements corresponding to the eigenvalue  $\lambda_1$ :

$$x^0 = \sum_{k=1}^N \alpha_k \psi_k.$$

We note that

$$\begin{aligned} A^n x^0 &= \lambda_1^n (u + r^{(n)}), \\ u &= \sum_{k=1}^p \alpha_k \psi_k \in \text{Ker}(A - \lambda_1 \mathbb{I}), \\ r^{(n)} &= \sum_{k=p+1}^N \alpha_k \left( \frac{\lambda_k}{\lambda_1} \right)^n \psi_k. \end{aligned} \tag{1.21}$$

As a result of assumption b), we obtain

$$\lim_{n \rightarrow \infty} r^{(n)} = 0.$$

We will now show that

$$x^{(n)} = \frac{A^n x^{(0)}}{\phi \cdot A^n x^{(0)}}. \quad [1.22]$$

We argue by induction. The property is clearly satisfied for  $n = 1$ . We suppose that it is satisfied for an integer  $n$ . Therefore, it follows that

$$x^{(n+1)} = \frac{A \frac{A^n x^{(0)}}{\phi \cdot A^n x^{(0)}}}{\phi \cdot A \frac{A^n x^{(0)}}{\phi \cdot A^n x^{(0)}}} = \frac{A^{n+1} x^{(0)}}{\phi \cdot A^n x^{(0)}} \frac{\phi \cdot A^n x^{(0)}}{\phi \cdot A^{n+1} x^{(0)}}$$

and the property is also satisfied for  $n + 1$ . We combine equations [1.21] and [1.22] to obtain

$$x^{(n)} = \frac{\lambda_1^n (u + r^{(n)})}{\phi \cdot \lambda_1^n (u + r^{(n)})} = \frac{u + r^{(n)}}{\phi \cdot (u + r^{(n)})} \xrightarrow{n \rightarrow \infty} \frac{u}{\phi \cdot u} \in \text{Ker}(A - \lambda_1 \mathbb{I}).$$

Finally, we conclude that

$$Ax^{(n)} = \frac{A(u + r^{(n)})}{\phi \cdot (u + r^{(n)})} = \frac{\lambda_1 u + Ar^{(n)}}{\phi \cdot (u + r^{(n)})} \xrightarrow{n \rightarrow \infty} \lambda_1 \frac{u}{\phi \cdot u}.$$

The method allows us to find the eigenvalue  $\lambda_1 = \lim_{n \rightarrow \infty} Ax^{(n)} \cdot \phi$  and the associated eigenvector  $v = \frac{u}{\phi \cdot u}$  normalized by the condition  $v \cdot \phi = 1$ .

The method works because  $u \neq 0$ . However, in practice, we do not know  $\text{Ker}(A - \lambda_1 \mathbb{I})$  and cannot exclude the possibility that the vector  $x^{(0)}$  is chosen orthogonal to  $\text{Ker}(A - \lambda_1 \mathbb{I})$ . In practice, we run the algorithm with several different choices of data in order to ensure that the “correct” eigenvalue is found. In fact, by choosing the initial vector at random, following a uniform law, the probability of choosing a vector orthogonal to the eigenspace is null. Finally, we note that the rate of convergence of the method is governed by the ratio  $|\lambda_2|/|\lambda_1|$ : the larger the “spectral gap”, the faster the convergence occurs.

The above detailed proof assumes that  $A$  is diagonalizable: this is assumption a). It is an important restriction that can be relaxed, and the proof for the general case

follows a similar argument. We use the Jordan decomposition  $A = PJP^{-1}$ , with the Jordan matrix  $J \in \mathcal{M}_N$  composed of blocks of the form:

$$J(\lambda, q) = \begin{pmatrix} \lambda & 1 & 0 & \cdots & 0 \\ 0 & \ddots & & & \\ \vdots & & \ddots & & 0 \\ 0 & & & \ddots & 1 \\ 0 & \cdots & 0 & \lambda & \end{pmatrix} \in \mathcal{M}_q,$$

where  $\lambda$  describes the set of eigenvalues of  $A$ . The matrix  $J$  contains precisely  $r$  such blocks; the number of blocks with the same eigenvalue  $\lambda$  corresponds to the geometric multiplicity  $\dim(\text{Ker}(A - \lambda\mathbb{I}))$ , while the sum of those blocks of size  $q$  gives the algebraic multiplicity of  $\lambda$ , that is to say, its multiplicity as a root of the characteristic polynomial (if  $\lambda$  is semisimple, the corresponding Jordan blocks have no nilpotent part). We therefore develop the first iteration on the basis defined for the Jordan decomposition (and whose coordinates in the canonical basis can be read on the columns of the change-of-basis matrix  $P$ ):

$$x^{(0)} = \sum_{k=1}^N \alpha_k \psi_k.$$

We continue to assume that the first  $p$  vectors of this basis are associated with the dominant eigenvalue  $\lambda_1$ . We note that  $P^{-1}\psi_k = e_k$ , the column made of 0, except for the  $k$ th row, which is 1. We thus write

$$A^n x^{(0)} = PJ^n P^{-1} x^{(0)} = \lambda_1^n (u^{(n)} + r^{(n)}),$$

$$u^{(n)} = \sum_{k=1}^p \alpha_k P \left( \frac{J}{\lambda_1} \right)^n e_k,$$

$$r^{(n)} = \sum_{k=p+1}^N \alpha_k P \left( \frac{J}{\lambda_1} \right)^n e_k.$$

For a large enough  $n$ ,  $J(\lambda, q)^n$  is given by

$$\left( \begin{array}{cccc} \lambda^n & C_n^1 \lambda^{n-1} & C_n^2 \lambda^{n-2} & \dots & C_n^{q-1} \lambda^{n-q+1} \\ 0 & \ddots & \ddots & & \vdots \\ \vdots & & \ddots & & C_n^2 \lambda^{n-2} \\ 0 & & & C_n^1 \lambda^{n-1} & \\ 0 & \dots & 0 & \lambda^n & \end{array} \right) \quad [1.23]$$

with  $C_n^k = \frac{n!}{k!(n-k)!}$ . We can verify this formula by writing  $J(\lambda, q) = \lambda \mathbb{I} + N$ , with  $N^q = 0$ , and by using the relation  $(\lambda \mathbb{I} + N)^n = \sum_{k=0}^{q-1} C_n^k \lambda^{n-k} N^k$ . Since  $r^{(n)}$  is expressed by only using blocks associated with eigenvalues such that  $|\lambda|/|\lambda_1| < 1$ , we have  $\lim_{n \rightarrow \infty} r^{(n)} = 0$  as a result of the following estimate, which is valid for all  $k \in \{1, \dots, q\}$ ,  $n \gg q$ ,  $\lambda \neq 0$ ,

$$\begin{aligned} C_n^k \left| \frac{\lambda^{n-k}}{\lambda_1^n} \right| &\leq \frac{n(n-1)\dots(n-k+1)}{k!} \left( \frac{|\lambda|}{|\lambda_1|} \right)^n \frac{1}{|\lambda|^k} \\ &\leq \frac{n^k}{k! |\lambda|^k} \left( \frac{|\lambda|}{|\lambda_1|} \right)^n \\ &\leq \frac{1}{k! |\lambda|^k} \exp \left( n \ln \left( \frac{|\lambda|}{|\lambda_1|} \right) + k \ln(n) \right) \xrightarrow[n \rightarrow \infty]{} 0. \end{aligned}$$

When  $\dim(\text{Ker}(A - \lambda_1 \mathbb{I})) = p$ , there are  $p$  Jordan blocks (of size  $1 \times 1!$ ) reduced to  $\lambda_1$ , the  $\psi_k = P e_k$  are eigenvectors of  $A$  associated with  $\lambda_1$  and we can directly reproduce the arguments of the diagonalizable case. It remains to discuss the case when the eigenvalue  $\lambda_1$  has  $L$  non-trivial Jordan blocks of size  $1 \leq q_\ell \leq p$ ,  $\sum_{\ell=1}^L q_\ell = p$ . We have (with the convention  $q_0 = 0$ )

$$u^{(n)} = \sum_{\ell=1}^L \sum_{k=q_{\ell-1}+1}^{q_\ell} \alpha_k P \left( \frac{J(\lambda_1, q_\ell)}{\lambda_1} \right)^n e_k.$$

We thus need understand the asymptotic behavior of the matrices  $J(\lambda_1, q)^n$ , which are of size  $q \times q$ . To this end, we work with the canonical basis  $(\kappa_1, \dots, \kappa_q)$  of  $\mathbb{R}^q$  and note that

$$\left( \frac{J(\lambda_1, q)}{\lambda_1} \right)^n \kappa_k = \sum_{m=1}^k C_n^{k-m} \lambda_1^{m-k} \kappa_m$$

(This expression can be read on the  $k$ th column of the matrix in equation [1.23]). The leading term in these sums corresponds to  $k = q$ ,  $m = 1$  (element located top and on the right of the matrix)<sup>10</sup>. The behavior of  $u^{(n)}$  is therefore directly related to the size of the largest Jordan block. By assuming that it corresponds to the elements  $\{\psi_1, \dots, \psi_{q_1}\}$  of the Jordan basis, when  $n \rightarrow \infty$ ,  $u^{(n)}$  behaves like

$$C_n^{q-1} \lambda_1^{1-q} \alpha_1 Pe_1,$$

where we recall that  $APe_1 = A\psi_1 = \lambda_1\psi_1 = PJe_1 = \lambda_1 Pe_1$  is the eigenvector associated with  $\lambda_1$ . Finally, we arrive at

$$x^{(n)} \xrightarrow[n \rightarrow \infty]{ } \frac{Pe_1}{\phi \cdot Pe_1}$$

and

$$Ax^{(n)} \xrightarrow[n \rightarrow \infty]{ } \frac{APe_1}{\phi \cdot Pe_1} = \lambda_1 \frac{Pe_1}{\phi \cdot Pe_1}$$

We will find several commentaries, variants and extensions of this method in Chapter 10 of [LAS 04] and section 10.4 of [SER 01]. In particular, the case where several eigenvectors have a modulus equal to the spectral radius – which case is excluded by the Perron–Frobenius theorem – leads to difficulties. We conclude this analysis with the following statement.

**THEOREM 1.14** (Convergence of the Power Method).— We assume there exists a unique  $\lambda \in \sigma(A)$ , such that  $|\lambda| = \rho(A)$ . We decompose  $\mathbb{R}^N = E \oplus F$ , where  $E$  and  $F$  are stable subspaces by  $A$ , such that  $\sigma(A|_E) = \{\lambda\}$  and  $\lambda \notin \sigma(A|_F)$ . If  $x^{(0)}$  is not orthogonal to  $E$ , then the sequence defined by the power method is such that  $\lim_{n \rightarrow \infty} Ax^{(n)} \cdot \phi = \lambda$  and  $\lim_{n \rightarrow \infty} x^{(n)} = v$ , such that  $Av = \lambda v$  and  $v \cdot \phi = 1$ .

### 1.2.5.5. Numerical simulations

We will perform several simulations to show the phenomena described above. We consider a system of size  $10 \times 10$  with initial values  $X_{\text{Init}} = (22, 18, 15, 15, 10, 11, 9, 8, 6, 3)$ ,  $f_1 = 0.05$ ,  $f_2 = 0.1$ ,  $f_3 = 0.8$ ,  $f_4 = 0.9$ ,  $f_5 = 0.6$ ,  $f_6 = 0.4$ ,  $f_7 = 0.2$ ,  $f_8 = 0.08$ ,  $f_9 = 0.06$  and  $f_{10} = 0.002$  as the fertility rate, and  $t_{2,1} = 0.85$ ,  $t_{3,2} = 0.94$ ,  $t_{4,3} = 0.9$ ,  $t_{5,4} = 0.83$ ,  $t_{6,5} = 0.76$ ,  $t_{7,6} = 0.7$ ,  $t_{8,7} = 0.6$ ,  $t_{9,8} = 0.1$

---

<sup>10</sup>Indeed, for  $k' > k$  and  $n$  being “large” compared to  $k$  and  $k'$ , we have  $\frac{C_n^{k'}}{C_n^k} = \frac{n(n-1)\dots(n-k)\dots(n-k'+1)}{k'(k'-1)\dots k+1} \frac{k(k-1)\dots 1}{n(n-1)\dots(n-k+1)} = \frac{(n-k)\dots(n-k'+1)}{k'(k'-1)\dots(k+1)} > 1$ .

and  $t_{10,9} = 0.02$  as its transition rate. The power method gives  $\rho_L = 1.1896445$  with associated eigenvectors, respectively, for  $L$  and  $L^\top$

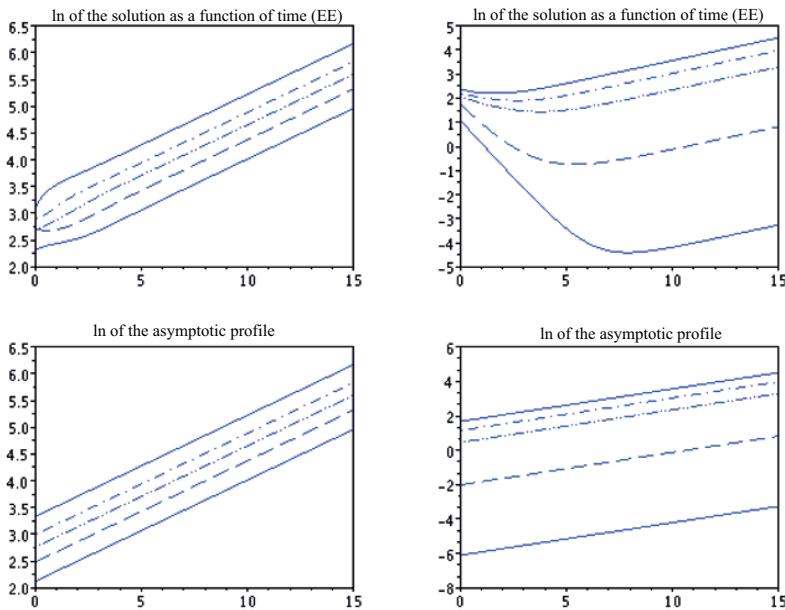
$$V = \begin{pmatrix} 1. \\ 0.7144990 \\ 0.5645625 \\ 0.4271077 \\ 0.2979879 \\ 0.1903681 \\ 0.1120146 \\ 0.0564951 \\ 0.0047489 \\ 0.0000798 \end{pmatrix} \quad \phi = \begin{pmatrix} 0.2596440 \\ 0.3633933 \\ 0.4322813 \\ 0.3406064 \\ 0.2066518 \\ 0.1184944 \\ 0.0530123 \\ 0.0185617 \\ 0.0131025 \\ 0.0004365 \end{pmatrix}.$$

Simulating this linear problem [1.17] is not particularly difficult. Figure 1.2 shows the graphs of the approximations of  $\ln(x_j)$  as a function of time, compared with the asymptotic profiles (simulations carried out with the explicit Euler scheme), until the final time  $T = 15$ . In qualitative terms, we confirm the behavior predicted by the theory. However, a direct evaluation of the difference in the asymptotic profile can leave something to be desired: see Figure 1.3 for a simulation performed by the Crank–Nicolson scheme and time step  $\Delta t = 1/1000$ . In particular, the quantity  $e^{-\rho t} X(t) \cdot \phi$  is not preserved by the schemes employed, and the errors produce significant discrepancies on large temporal scales.

### 1.2.6. Modeling red blood cell agglomeration

In blood, red blood cells (or erythrocytes) are subject to attractive interaction forces that drive them to form aggregates. We here provide a simplified model of this kind of phenomenon. Red blood cells move along the segment  $[0, 1]$  (which represents a blood vessel!). We consider  $N + 2$  blood cells, whose center points are located at  $x_i(t)$ , with  $i \in \{0, \dots, N + 1\}$ , as a function of time. The outermost cells must remain fixed at the ends of the interval studied:  $x_0(t) = 0$ ,  $x_{N+1}(t) = 1$ . In this description, we represent the blood cells as point-based particles, although we take account of their size to describe the interaction efforts. The particles are effectively subject to forces exerted by their neighboring particles. At large distances, those forces are attractive and tend to 0 when the distance between the particles tends to  $+\infty$ . When the distance between particles is less than a certain threshold  $a > 0$ , the interaction force becomes repulsive. These effects are described by a function  $\varphi : ]0, \infty[ \rightarrow \mathbb{R}$ , such that

$$\varphi(d) > 0 \text{ if } d > a, \quad \varphi(d) < 0 \text{ if } d < a, \quad \lim_{d \rightarrow \infty} \varphi(d) = 0, \quad \lim_{d \rightarrow 0} \varphi(d) = -\infty.$$



**Figure 1.2.** Logarithm of the numerical solution (explicit Euler) versus logarithm of the asymptotic profile as a function of time (components 1 to 5 on the left, 6 to 10 on the right)

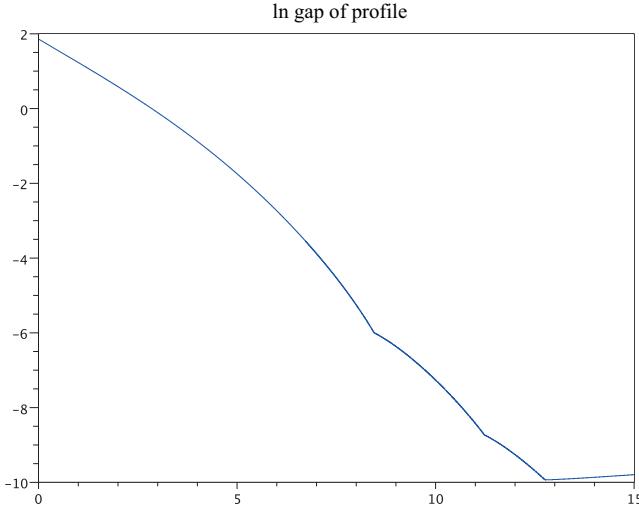
For example, we can take

$$\varphi(d) = \frac{\gamma}{d} \ln\left(\frac{d}{a}\right), \quad [1.24]$$

with a given  $\gamma > 0$ . Therefore, the particle located in  $x_{i+1}$  exerts the force  $\varphi(x_{i+1} - x_i)$  on the particle  $i$ . Finally, the particles undergo a friction force, exerted by the carrying fluid. This force is opposite to the movement and proportional to the velocity of the particle considered. Applying the fundamental principle of dynamics, the movement of the particles is described by the following system of differential equations:

$$\frac{d^2}{dt^2}x_i = \varphi(x_{i+1} - x_i) - \varphi(x_i - x_{i-1}) - \lambda \frac{dx_i}{dt} \quad [1.25]$$

for  $i \in \{1, \dots, N\}$ , with  $\lambda > 0$ .



**Figure 1.3.** Logarithm of the difference between the numerical approximation (using Crank–Nicolson) and the profile as a function of time

We write  $\xi_i = \frac{d}{dt}x_i$  and  $t \mapsto U(t) = (x_1(t), \dots, x_N(t), \xi_1(t), \dots, \xi_N(t))$ . We rewrite the problem in the form of a (autonomous) first-order differential equation satisfied by  $U$ :

$$\frac{d}{dt}U(t) = F(U(t)). \quad [1.26]$$

To this end, we introduce the open set

$$\Omega = \{(y, \theta) \in \mathbb{R}^{2N}, 0 < y_1 < y_2 < \dots < y_N < 1\}$$

and the function  $F : Y = (y_1, \dots, y_N, \theta_1, \dots, \theta_N) \in \Omega \mapsto F(Y) \in \mathbb{R}^{2N}$  is defined by

$$\begin{aligned} F_i(Y) &= \theta_i, \\ F_{N+i}(Y) &= \varphi(y_{i+1} - y_i) - \varphi(y_i - y_{i-1}) - \lambda\theta_i \end{aligned}$$

for  $i \in \{1, \dots, N\}$  and using the convention  $y_0 = 0, y_{N+1} = 1$ .  $F$  is a function of class  $C^1$  on the open set  $\Omega$ . For each initial value that sets the positions and velocities of the particles in  $\Omega$ , the Picard–Lindelöf theorem guarantees the existence and uniqueness of a class  $C^1$  solution of [1.25], which is defined on an interval  $[0, T[, 0 < T \leq \infty$ .

We will see that the system [1.25] has remarkable “energetic” properties. Indeed, let  $\Psi$  be an antiderivative of  $\varphi$ . We can write

$$E(t) = \frac{1}{2} \sum_{i=1}^N \left| \frac{d}{dt} x_i \right|^2 + \sum_{i=0}^N \Psi(x_{i+1} - x_i).$$

This quantity sums the kinetic energy of the particles and the potential energy due to their interaction forces. We can compute

$$\begin{aligned} \frac{d}{dt} E &= \sum_{i=1}^N \frac{d}{dt} x_i \frac{d^2}{dt^2} x_i + \sum_{i=0}^N \varphi(x_{i+1} - x_i) \left( \frac{d}{dt} x_{i+1} - \frac{d}{dt} x_i \right) \\ &= \sum_{i=1}^N \frac{d}{dt} x_i \left( \varphi(x_{i+1} - x_i) - \varphi(x_i - x_{i-1}) \right) \\ &\quad - \lambda \sum_{i=1}^N \left| \frac{d}{dt} x_i \right|^2 + \sum_{i=0}^N \varphi(x_{i+1} - x_i) \left( \frac{d}{dt} x_{i+1} - \frac{d}{dt} x_i \right). \end{aligned}$$

However, we have

$$\sum_{i=1}^N \frac{d}{dt} x_i \times \varphi(x_i - x_{i-1}) = \sum_{i=0}^N \frac{d}{dt} x_{i+1} \times \varphi(x_{i+1} - x_i).$$

We conclude that

$$\frac{d}{dt} E = -\lambda \sum_{i=1}^N \left| \frac{d}{dt} x_i \right|^2 \leq 0$$

(the right-hand term can be interpreted as the energy dissipated by friction) and therefore  $E(t) \leq E(0)$ . Owing to the properties of the function  $\varphi$ , this estimate can allow us to determine that the solutions are globally defined. We note that

$$\frac{1}{z} \ln(z) = \ln(z) \frac{d}{dz} \ln(z) = \frac{1}{2} \frac{d}{dz} \left( \ln^2(z) \right),$$

which allows us to find  $\Psi(d) = \frac{\gamma a}{2} \ln^2 \left( \frac{d}{a} \right)$  for the model [1.24]. It follows in that case that for all  $i \in \{1, \dots, N\}$  and  $t \geq 0$ , we have

$$\ln^2 \left( \frac{x_{i+1}(t) - x_i(t)}{a} \right) \leq \frac{2E(0)}{\gamma a}.$$

Likewise,  $t \mapsto \left| \frac{d}{dt} x_i(t) \right|^2$  is bounded. The solutions of [1.25] remain in  $\Omega$  for all times and are therefore defined globally.

We will perform numerical simulations of this problem, which will allow us to describe the aggregate formation phenomenon. We fix a time step  $h > 0$  and a final time  $T$  (in such a way that the final time will be reached in  $K = T/h$  iterations):

$$t_0 = 0 < t_1 = h < t_2 = 2h < \dots < t_K = T.$$

We let  $y^k = (x_1^k, \dots, x_N^k)$  denote the numerical unknown at time  $t_k$ . We denote the vector of  $\mathbb{R}^N$  whose  $i$ th component is  $\varphi(x_{i+1} - x_i) - \varphi(x_i - x_{i-1})$  as  $\Phi(y)$ , and we continue to use the convention  $x_0 = 0, x_{N+1} + 1$ . Given  $y^0, y^1$ , we use the following scheme:

$$y^{k+1} = 2y^k - y^{k-1} + h^2\Phi(y^k) - \lambda h(y^k - y^{k-1}). \quad [1.27]$$

Note that the second derivative in [1.25] is consistently approached at order 2, and the friction term is consistent at order 1. We may ask ourselves how to interpret the scheme when we write the problem in the form of a first-order system as in [1.26]. The scheme

$$y^{k+1} = y^k + h\xi^{k+1}, \quad \xi^{k+1} = \xi^k + h\Phi(y^k) - \lambda\xi^k$$

for [1.26] indeed gives [1.27] for the system [1.25]. Although some terms are processed implicitly, the scheme does not require a system resolution stage (this would still be the case if we write  $\lambda\xi^{n+1}$  when updating  $\xi$ ). We will return to the motivation for this algorithm, which differs from the classic Euler method, in section 1.3, which deals with the Hamiltonian structure of the interaction terms.

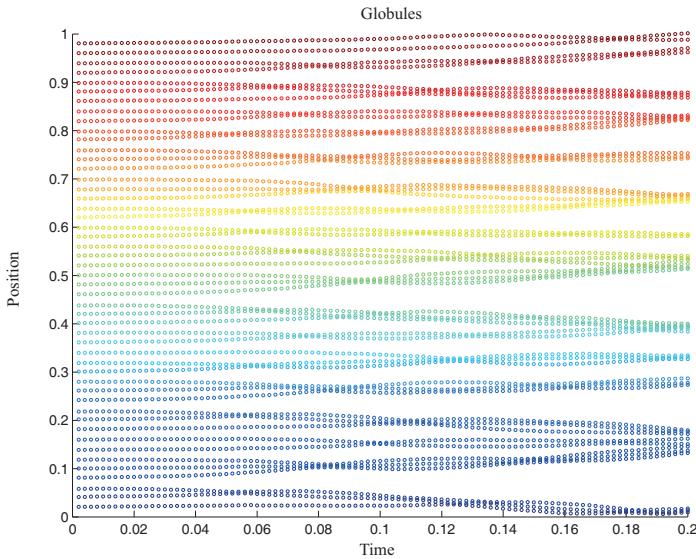
A typical simulation is shown in Figure 1.4. Here, we have taken  $N = 49$  particles, initially dispersed randomly along  $]0, 1[$ , with velocities equal to zero:

$$y_i = \frac{i + 0.1\rho_i}{N + 1}$$

with  $\rho_i$  representing a random variable uniformly distributed on  $[-1, +1]$ . The parameters are fixed as follows:

$$h = 0.002, \quad K = 100, \quad \gamma = 0.5, \quad \lambda = 10, \quad a = 0.004.$$

We observe the formation of aggregates gathering several particles (note that the trajectories do not cross one another). Figure 1.5 shows the evolution of the discrete energy. It is not decreasing because the numerical scheme does not preserve that property of the continuous model. However, we do observe a tendency for energy to decrease in time. We can also note that the standard Euler method does not provide an adequate result in these conditions (the numerical solution does not remain in the interval  $[0, 1]$  and takes outlying values, even with much smaller time steps).



**Figure 1.4.** Red blood cell agglomeration: example of trajectories for 49 particles. For the color version of this figure, see [www.iste.co.uk/goudon/mathmodel.zip](http://www.iste.co.uk/goudon/mathmodel.zip)

We now consider equilibrium solutions of the system [1.25]. For those solutions, the velocity is equal to zero,  $\xi_i = 0$ , and the forces are at equilibrium,  $\varphi(x_{i+1} - x_i) = \varphi(x_i - x_{i-1})$  for all  $i \in \{0, \dots, N\}$ . The discussion is based on an analysis of the graph of the function  $\varphi$  (see Figure 1.6). We write  $d_i = x_{i+1} - x_i$ , which satisfies

$$\sum_{i=0}^N d_i = 1. \quad [1.28]$$

At equilibrium, we have

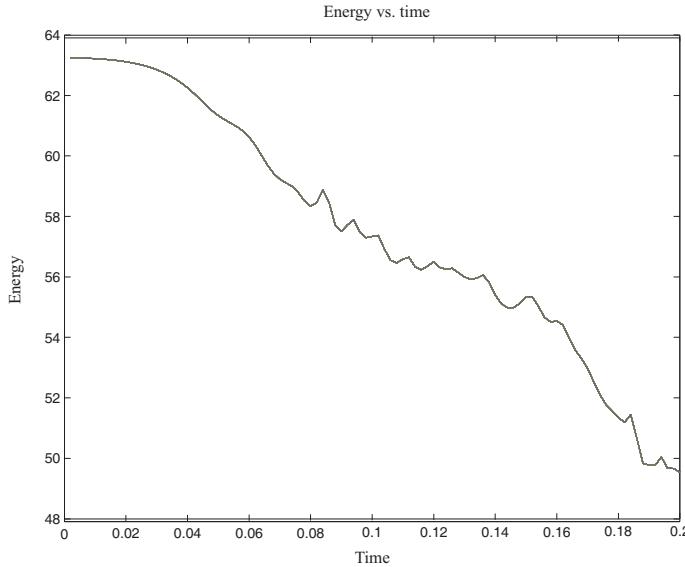
$$\varphi(d_i) = \bar{\varphi}, \text{ constant.} \quad [1.29]$$

We distinguish between two different cases:

- If  $a \geq \frac{1}{N+1}$  (dense population), then in order to satisfy [1.28], at least one of the distances  $d_i$  is less than or equal to  $a$ . Since  $d \mapsto \varphi(d)$  is strictly increasing and negative-valued on  $]0, a]$ , by [1.29], all distances are equal:  $d_i = \bar{d} \leq a$  and, in fact, [1.28] guarantees that  $\bar{d} = \frac{1}{N+1}$ .

- If  $a < \frac{1}{N+1}$  (dispersed population), then in order to satisfy [1.28], at least one of the distances  $d_i$  is greater than  $a$ . Since  $d \mapsto \varphi(d)$  has negative values on  $]0, a]$ , [1.29]

implies that all the  $d_i$  are greater than  $a$ . We note that  $d_i = \frac{1}{N+1}$  for all  $i \in \{0, \dots, N\}$  is still a solution; however, it is not the only one in the situation considered here because  $\varphi$  is not injective on  $]a, +\infty[$ . We can find  $d^* > d_* > a$ , such that  $\varphi(d_*) = \varphi(d^*)$ , and any configuration with  $N_*$  distances equal to  $d_*$  and  $N^*$  distances equal to  $d^*$  and  $N_* + N^* = N$  is a solution.



**Figure 1.5. Red blood cell agglomeration: evolution of energy**

It is interesting to focus on the case with a single free particle ( $N = 1$ , the other two particles are attached to the boundaries  $x = 0$  and  $x = 1$ ). In this case, the dynamic is described by the simple differential system (in  $\mathbb{R}^2$ )

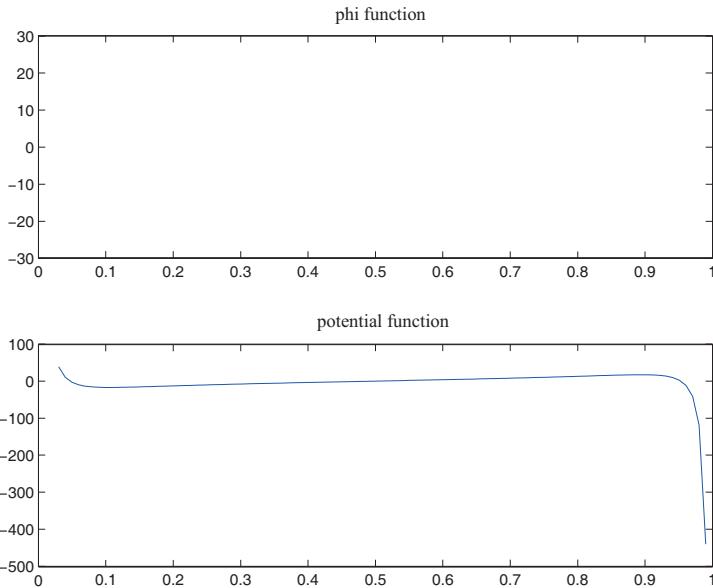
$$\frac{d}{dt} \begin{pmatrix} x \\ \xi \end{pmatrix} = \begin{pmatrix} \xi \\ \varphi(1-x) - \varphi(x) - \lambda \xi \end{pmatrix}.$$

When  $a > 1/2$ ,  $\bar{x} = 1/2$  is the single equilibrium position. There are three equilibrium configurations when  $a < 1/2$ :  $\bar{x} = 1/2$ ,  $\bar{x}$  near 0 or  $\bar{x}$  near 1; the last solutions correspond to the solutions of the nonlinear equation  $\varphi(1-x) - \varphi(x) = 0$ . If  $(\bar{x}, 0)$  denotes an equilibrium solution, the system linearized in the neighborhood of this configuration is written as

$$\frac{d}{dt} \begin{pmatrix} \tilde{x} \\ \tilde{\xi} \end{pmatrix} = \underbrace{\begin{pmatrix} 0 & 1 \\ \Phi'(\bar{x}) - \lambda & 0 \end{pmatrix}}_{:=A} \begin{pmatrix} \tilde{x} \\ \tilde{\xi} \end{pmatrix},$$

with  $\Phi(x) = \varphi(1-x) - \varphi(x)$ . In particular, with [1.24], we obtain:

$$\Phi'(\bar{x}) = -\varphi'(1-x) - \varphi'(x) = \frac{\gamma}{(1-x)^2} \left( \ln \left( \frac{1-x}{a} \right) - 1 \right) + \frac{\gamma}{x^2} \left( \ln \left( \frac{x}{a} \right) - 1 \right).$$



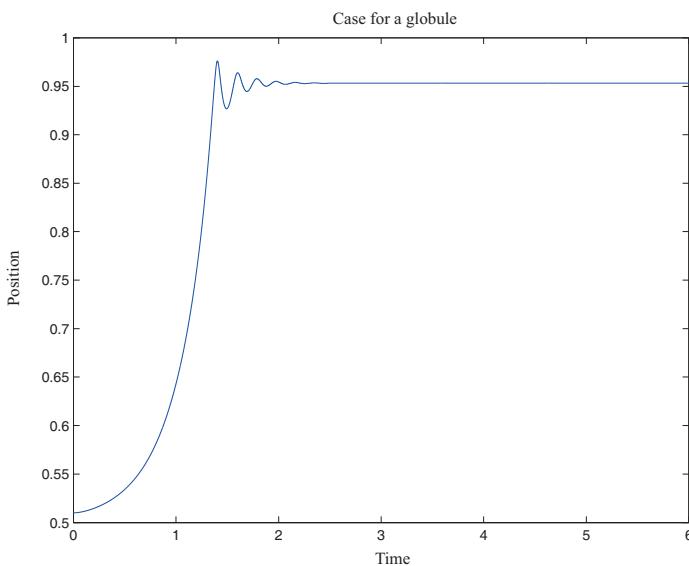
**Figure 1.6.** Graph of the function  $\varphi$  and the associated potential

When  $a > 1/2$ ,  $\bar{x} = 1/2$  is the only equilibrium solution, and this equilibrium is stable because  $\text{Tr}(A) = -\lambda$  is strictly negative and  $\det(A) = -\Phi'(1/2) = +2\varphi'(1/2) = -8\gamma(\ln(\frac{1}{2a}) - 1) > 0$ . Therefore, either the eigenvalues of  $A$  are complex conjugates and have strictly negative real parts, or they are real and both negative. This shows that the equilibrium  $(1/2, 0)$  is linearly stable when  $a > 1/2$ . When  $a < 1/2$ , we proceed to numerical experimentation. We let the system evolve on rather long periods of time. For small values of  $a$ , we observe that  $x(t)$  converges either towards a position near 0, or towards a position near 1, even if the initial value is chosen within a neighborhood of  $1/2$ . We therefore numerically confirm the following points:

- the system linearized on  $(1/2, 0)$  is linearly unstable (complex conjugate eigenvalues with positive real parts or at least one of the values is positive);
- for the system linearized on  $(\bar{x}, 0)$ , the equilibrium position is linearly stable;

- Newton’s algorithm for solving  $\Phi(x) = 0$ , starting from a first iteration “near”  $\bar{x}$ , converges to  $\bar{x}$ .

For Figure 1.7, we have taken  $\lambda = 10$ ,  $\gamma = 3$  and  $a = 0.04$ . The initial position is  $x_0 = 0.51$ . The eigenvalues of the system linearized on  $(1/2, 0)$  are  $-12.8497$  and  $2.8497$ , which confirms linearized instability. The position at the final time is  $x(T) = 0.95327$ , which coincides with the equilibrium position provided by Newton’s method. The eigenvalues of the system linearized at that point are  $-5 \pm 33.5899 i$ , which confirms linearized stability. In practice, we can observe a diversity of situations when we modify the value of the parameter  $a$  and the initial values.

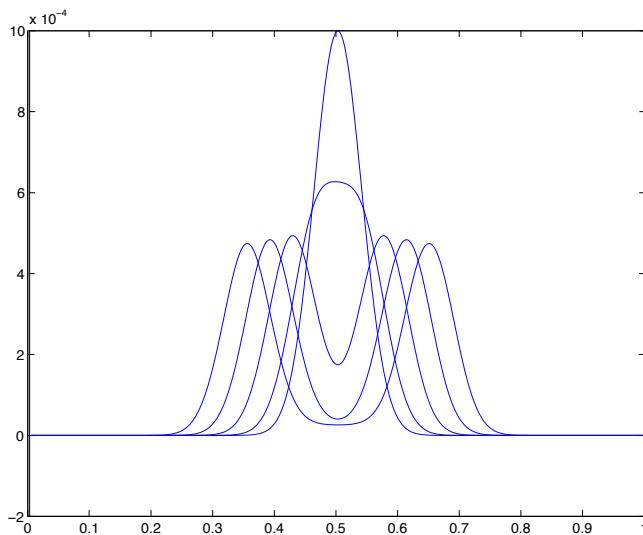


**Figure 1.7.** Single particle model: convergence towards a state of equilibrium

Finally, we perform simulations for a dense situation, with a large number of particles: here with  $a = 0.004$ , we choose  $N = 500$ . We define an initial value as a small perturbation of the solution, for example

$$x_i(0) = \frac{i}{N+1} + \frac{1}{1000} \exp\left(-\frac{i/(N+1) - 1/2}{0.003}\right).$$

We consider  $\gamma = 1/2$  and  $\lambda = 1$  (low damping). Figure 1.8 shows the initial perturbation and its evolution in time. We observe the effect of damping, which levels out maximums, as well as a wave propagation phenomenon: the initial bump separates into two parts of equal size that propagate in opposite directions, one towards the left and the other towards the right, producing a behavior close to that of a solution to the wave equation (with damping) (see section 3.3).



**Figure 1.8.** Blood cell agglomeration: evolution of a disturbance in the equilibrium state

### 1.2.7. SEI model

We are now interested in an epidemiology problem where we seek to describe transmission phenomena for a communicable disease. The population is therefore separated into three categories:

- those at risk, which are the individuals who can be contaminated;
- those exposed, which are the individuals who have already been infected but are not yet contagious;
- those infected, which are individuals who are infected and can transmit the disease to those at risk.

The concentration of individuals in these categories at time  $t \geq 0$  is denoted as  $S(t)$ ,  $E(t)$  and  $I(t)$ , respectively. The infection dynamic is governed by the following rules:

- There is a constant influx of newborns, which enter the category of those at risk. This contribution is described by a constant  $\lambda > 0$ , which therefore appears as a production in the at-risk population.
- Each category is subject to a natural mortality rate, which depends on the category. We denote the mortality rates as  $\gamma, \mu, \nu$  for categories  $S, E, I$ , respectively.

- Individuals exposed to the disease pass on to the infectious category at a certain rate  $\alpha$ .
- Infectious individuals either die or remain healthy at a rate  $\beta$ .
- Encounters between at-risk and infectious individuals make those at risk enter the category of those exposed. We let  $\tau$  denote the rate of those encounters.

Taking these principles into account, the population's evolution is therefore described by the differential system

$$\begin{cases} \frac{d}{dt}S = \lambda - \gamma S - \tau IS + \beta I, \\ \frac{d}{dt}E = \tau IS - (\alpha + \mu)E, \\ \frac{d}{dt}I = \alpha E - (\beta + \nu)I. \end{cases} \quad [1.30]$$

It is an autonomous system; the function

$$f : \begin{pmatrix} S \\ E \\ I \end{pmatrix} \in \mathbb{R}^3 \longmapsto \begin{pmatrix} \lambda - \gamma S - \tau IS + \beta I \\ \tau IS - (\alpha + \mu)E \\ \alpha E - (\beta + \nu)I \end{pmatrix}$$

is  $C^1$  and the Picard–Lindelöf theorem ensures the existence and uniqueness of a solution associated with any initial value  $(S_{\text{Init}}, E_{\text{Init}}, I_{\text{Init}}) \in \mathbb{R}^3$ . In light of the physical interpretation of the unknown functions, those values are positive and the solutions must remain positive. We will also show that the solutions are well defined for all times  $t \geq 0$ .

It is interesting to note that if the population is initially healthy,  $I_{\text{Init}} = E_{\text{Init}} = 0$ ,  $S_{\text{Init}} > 0$ , then it remains healthy for all times and the corresponding solution of [1.30] is simply

$$E(t) = I(t) = 0, \quad S(t) = (1 - e^{-\gamma t}) \frac{\lambda}{\gamma} + e^{-\gamma t} S_{\text{Init}}.$$

This solution is well defined for all times  $t \geq 0$  and converges to the constant state  $(S_\infty, E_\infty, I_\infty) = (\lambda/\gamma, 0, 0)$  when  $t \rightarrow \infty$ . More generally, consider an initial value  $S_{\text{Init}} > 0$ ,  $I_{\text{Init}} \geq 0$ ,  $E_{\text{Init}} \geq 0$ . By writing  $f : x = (x_1, x_2, x_3) \mapsto (f_1(x), f_2(x), f_3(x))$ , we observe that if  $x_j \geq 0$  and  $x_k = 0$ , then  $f_k(x) \geq 0$ . This remark ensures that the solution of [1.30] remains positive:  $S(t) \geq 0, E(t) \geq 0, I(t) \geq 0$ . In order to analyze this property, it is necessary to imagine several cases. Let us first suppose that there exists a  $t_0 > 0$ , such that  $S(t_0) = 0$ , with  $S(t) > 0$  on  $[0, t_0[$

and  $E(t), I(t) \geq 0$  on  $[0, t_0]$ . Then, at  $t_0$ , we have  $\frac{dS}{dt}(t_0) \geq \lambda > 0$ . By continuity,  $t \mapsto S(t)$  remains strictly increasing for  $t$  close enough to  $t_0$ , which contradicts the fact that  $S(t)$  reaches 0 when  $t$  tends towards  $t_0$ . We conclude that  $S$  remains strictly positive. Next, suppose that there exists a  $t_0 > 0$ , such that  $E(t_0) = 0$ , with  $S(t) > 0$  on  $[0, t_0]$  and  $E(t), I(t) \geq 0$  on  $[0, t_0]$ . If  $I(t_0) > 0$ , then  $\frac{dS}{dt}(t_0) > 0$ , and we can reproduce this reasoning to arrive at a contradiction. If  $I(t_0) = 0$ , we explicitly know the solution that satisfies  $S(t) > 0$ , with  $E(t) = I(t) = 0$  for all times  $t \geq t_0$ . The same discussion applies if we suppose there exists a  $t_0 > 0$ , such that  $I(t_0) = 0$ , with  $S(t) > 0$  on  $[0, t_0]$  and  $E(t), I(t) \geq 0$  on  $[0, t_0]$ . Finally, we have

$$\frac{d}{dt}(S + E + I) = \lambda - \underbrace{(\gamma S + \mu E + \nu I)}_{\text{mortality terms}} \leq \lambda.$$

This implies that  $0 \leq S(t) + E(t) + I(t) \leq S_{\text{Init}} + E_{\text{Init}} + I_{\text{Init}} + \lambda t$ , which excludes the possibility of the solution exploding in finite time. Therefore, the solutions of [1.30] are defined for all times  $t \geq 0$ .

It is clear that  $(S_\infty, E_\infty, I_\infty) = (\lambda/\gamma, 0, 0)$  is a solution of equation [1.30], which corresponds to a population in a completely healthy state. More generally, we seek to find  $(\bar{S}, \bar{E}, \bar{I})$ , a stationary solution of [1.30]. In other words, we have

$$\begin{aligned} \alpha \bar{E} &= (\beta + \nu) \bar{I}, \\ \tau \bar{I} \bar{S} &= (\alpha + \mu) \bar{E} = \frac{(\beta + \nu)(\alpha + \mu)}{\alpha} \bar{I}, \\ \lambda &= \gamma \bar{S} + \tau \bar{I} \bar{S} - \beta \bar{I} = \gamma \bar{S} + \left( \frac{(\beta + \nu)(\alpha + \mu)}{\alpha} - \beta \right) \bar{I}. \end{aligned}$$

It follows that  $\bar{I}$  is the root of a simple quadratic equation  $((\beta + \nu)(\alpha + \mu) - \alpha\beta) \bar{I}^2 + \frac{\gamma}{\tau}(\beta + \nu)(\alpha + \mu) \bar{I} = \lambda \alpha \bar{I}$ , where

$$\bar{I} = \frac{\lambda \alpha - \frac{\gamma}{\tau}(\beta + \nu)(\alpha + \mu)}{\nu(\alpha + \mu) + \beta \mu}.$$

This expression provides an acceptable solution ( $\bar{I} > 0$ ) when

$$\rho = \frac{\alpha \lambda \tau}{\gamma(\alpha + \mu)(\beta + \nu)} > 1.$$

The stationary solution  $(\bar{S}, \bar{E}, \bar{I})$  thus obtained corresponds to an epidemiological situation.

We compute

$$\nabla_{S,E,I} f(S, E, I) = \begin{pmatrix} -\gamma - \tau I & 0 & \beta - \tau S \\ \tau I & -(\alpha + \mu) & \tau S \\ 0 & \alpha & -(\beta + \nu) \end{pmatrix}.$$

We write  $M = \nabla_{S,E,I} f(S_\infty, E_\infty, I_\infty) = \nabla_{S,E,I} f(\lambda/\gamma, 0, 0)$ . The eigenvalues of  $M$  are  $z_0 = -\gamma < 0$  and  $z_\pm = \frac{1}{2}(-(\alpha + \mu + \beta + \nu) \pm \sqrt{\Delta})$ , with  $\Delta = ((\alpha + \mu) - (\beta + \nu))^2 + 2\frac{\alpha\tau\lambda}{\gamma} > 0$ . The three eigenvalues are strictly negative when  $\rho < 1$ . In this case, the solutions to the linear system  $\frac{d}{dt}X = MX$  satisfy  $X(t) = e^{Mt}X_{\text{Init}}$  and converge to 0 exponentially fast. This means that, for the linearized problem, the healthy state  $(\lambda/\gamma, 0, 0)$  is stable when  $\rho < 1$ , which describes the situation in which there is no epidemiological state.

The previous result only refers to the system linearized around the equilibrium state  $(\lambda/\gamma, 0, 0)$ . We will now demonstrate the stability of the nonlinear problem [1.30], assuming  $\rho < 1$ . Let  $A > 0$  be a parameter whose value will be fixed later. We calculate

$$\begin{aligned} \frac{d}{dt} \left( \frac{|S - \lambda/\gamma|^2}{2} + A \left( E + \frac{\alpha + \mu}{\alpha} I \right) \right) &= -\gamma |S - \lambda/\gamma|^2 - P_A(S)I, \\ P_A(S) &= \tau S^2 - \left( \tau \frac{\lambda}{\gamma} + \beta + A\tau \right) S + \beta \frac{\lambda}{\gamma} + A \frac{(\alpha + \mu)(\beta + \nu)}{\alpha}. \end{aligned}$$

The function  $S \mapsto P_A(S)$  is a second-order polynomial that satisfies  $P_A(0) > 0$  and  $\lim_{S \rightarrow \infty} P_A(S) = +\infty$ . Therefore,  $P_A(S)$  remains strictly positive for all  $S > 0$  in the case where the polynomial does not have real roots. This condition requires that

$$q(A) = \left( \tau \frac{\lambda}{\gamma} + \beta + A\tau \right)^2 - 4\tau \left( \beta \frac{\lambda}{\gamma} + A \frac{(\alpha + \mu)(\beta + \nu)}{\alpha} \right) < 0.$$

However,  $A \mapsto q(A)$  is a second-order polynomial that satisfies  $q(0) = (\tau\lambda/\gamma - \beta)^2 > 0$  and  $\lim_{A \rightarrow \infty} q(A) = +\infty$ . The polynomial's discriminant is

$$\Delta_q = 4 \frac{\tau^2}{\alpha^2} \left( (\alpha + \mu)(\beta + \nu) - \alpha\tau \frac{\lambda}{\gamma} \right) (\nu(\alpha + \mu) + \beta\mu)$$

which remains strictly positive when  $\rho < 1$ . Thus, the polynomial  $A \mapsto q(A)$  has a positive root  $A_+$ . We therefore fix  $0 < A < A_+$ , in such a way that we can deduce

from the analysis that there exists  $\pi_A > 0$ , such that for all  $S \geq 0$ , we have  $P_A(S) \geq \pi_A > 0$  and arrive at the estimate

$$\begin{aligned} \frac{|S(t) - \lambda/\gamma|^2}{2} + AE(t) + A\frac{\alpha + \mu}{\alpha}I(t) \\ + \gamma \int_0^t |S(\sigma) - \lambda/\gamma|^2 u d\sigma + \pi_A \int_0^t I(\sigma) d\sigma \leq C_0 \end{aligned}$$

where  $C_0 = \frac{|S_{\text{Init}} - \lambda/\gamma|^2}{2} + AE_{\text{Init}} + A\frac{\alpha + \mu}{\alpha}I_{\text{Init}}$ .

In particular, this proves that the functions  $t \mapsto E(t)$ ,  $t \mapsto I(t)$  and  $t \mapsto |S(t) - \lambda/\gamma|^2$ , and therefore  $t \mapsto S(t)$  are uniformly bounded on  $[0, +\infty[$ . Returning to equation [1.30], it follows that the derivative functions  $t \mapsto \frac{d}{dt}E(t)$ ,  $t \mapsto \frac{d}{dt}I(t)$  and  $t \mapsto \frac{d}{dt}S(t)$  are also uniformly bounded on  $[0, +\infty[$ . Moreover, the estimate obtained also shows that  $t \mapsto |S(t) - \lambda/\gamma|^2$  and  $t \mapsto I(t)$  are functions of  $L^1([0, \infty[)$ . In our final discussion of the asymptotic behavior of [1.30], we will use the following general lemma.

**LEMMA 1.10.–** Let  $U : t \geq 0 \mapsto U(t) \in \mathbb{R}$  be a uniformly continuous function integrable on  $[0, \infty)$ . Thus, we have  $\lim_{t \rightarrow \infty} U(t) = 0$ .

**PROOF.–** If we only assume that  $V$  is a continuous integrable function, we can find counterexamples to this statement, with functions that do not have a limit in  $+\infty$ :

$$V(t) = \sum_{n=2}^{\infty} \left( (n^2 t + (1 - n^3) \mathbf{1}_{[n-1/n^2, n]}(t) + (-n^2 t + (1 + n^3) \mathbf{1}_{[n, n+1/n^2]}(t) \right)$$

provides one such counterexample (it is useful to trace the graph of this function and verify that it is continuous and integrable on  $\mathbb{R}$ .) On the other hand, uniform continuity allows us to eliminate this kind of oscillatory behavior. We argue by contradiction by assuming that there exists  $t_n \in \mathbb{R}$  and a strictly increasing sequence  $(t_n)_{n \in \mathbb{N}}$  of positive real numbers, such that  $\lim_{n \rightarrow \infty} t_n = \infty$  and  $|U(t_n)| \geq \epsilon$  for all  $n \in \mathbb{N}$ . Since  $U$  is uniformly continuous, there exists  $\eta > 0$ , such that if  $|t - s| \leq \eta$ , then  $|U(t) - U(s)| \leq \epsilon/2$ . It follows that

$$\int_{t_n - \eta}^{t_n + \eta} |U(s)| ds \geq \int_{t_n - \eta}^{t_n + \eta} ||U(t_n)| - |U(t_n) - U(s)|| ds \geq \frac{\epsilon}{2} \times 2\eta = \eta\epsilon > 0,$$

which would contradict the Cauchy criterion satisfied by the integrable function  $U$ .  $\square$

Since the derivatives of  $S$  and of  $I$  are uniformly bounded, we can apply the lemma to the functions  $t \mapsto |S(t) - \lambda/\gamma|^2$  and  $t \mapsto I(t)$ . We therefore prove that

$$\lim_{t \rightarrow \infty} S(t) = \lambda/\gamma, \quad \lim_{t \rightarrow \infty} I(t) = 0.$$

Finally, we return to the dissipation estimate obtained previously:  $t \mapsto \frac{1}{2}|S(t) - \lambda/\gamma|^2 + AE(t) + A\frac{\alpha+\mu}{\alpha}I(t)$  is a decreasing function on  $[0, \infty[$  and therefore has a limit when  $t \rightarrow \infty$ , denoted as  $\ell \geq 0$ . It follows that  $\lim_{t \rightarrow \infty} E(t) = \ell/A$ . However, by substituting this information into [1.30], we conclude that  $\ell = 0$ . We have thus demonstrated that, when  $\rho < 1$ , the solutions of [1.30] converge when  $t \rightarrow \infty$  to the healthy state  $(\lambda/\gamma, 0, 0)$ .

In terms of numerical simulations, we will see that, in certain cases, it may be interesting to use the implicit Euler scheme rather than the explicit version. Since the function  $f$  is nonlinear, implementing the implicit scheme requires a procedure to find the roots of a certain function. Indeed, by letting  $X$  denote the vector with components  $(S, E, I)$ , the scheme [1.12] is written as  $X^{n+1} = X^n + \Delta t f(X^{n+1})$ . In other words, given  $X^n = (S^n, E^n, I^n)$  and  $\Delta t$ ,  $X^{n+1}$  satisfies the nonlinear equation

$$\Psi(X^{n+1}) = 0,$$

$$\text{where } \Psi : Y \in \mathbb{R}^3 \mapsto \Psi(Y) = Y - X^n - \Delta t f(Y)$$

$$= \begin{pmatrix} Y_1 - S^n - \Delta t(\lambda - \gamma Y_1 - \tau Y_1 Y_3 + \beta Y_3) \\ Y_2 - E^n - \Delta t(\tau Y_1 Y_3 - (\alpha + \mu)Y_2) \\ Y_3 - I^n - \Delta t(\alpha Y_2 - (\beta + \nu)Y_3) \end{pmatrix}.$$

In fact, we will provide an approximation of the roots of the function  $\Psi$  using Newton's method. We construct the sequence defined by

$$Y^{(0)} = X^n = (S^n, E^n, I^n), Y^{(k+1)} = Y^{(k)} - [\nabla_Y \Psi(Y^{(k)})]^{-1} \Psi(Y^k)$$

Iteration  $X^{n+1}$  corresponds to the limit of  $Y^{(k)}$  when  $k \rightarrow \infty$ . More specifically, we introduce the matrix

$$M^{(k)} = \nabla_Y \Psi(Y^{(k)}) = \begin{pmatrix} \partial_{Y_1} \Psi_1(Y^{(k)}) & \partial_{Y_2} \Psi_1(Y^{(k)}) & \partial_{Y_3} \Psi_1(Y^{(k)}) \\ \partial_{Y_1} \Psi_2(Y^{(k)}) & \partial_{Y_2} \Psi_2(Y^{(k)}) & \partial_{Y_3} \Psi_2(Y^{(k)}) \\ \partial_{Y_1} \Psi_3(Y^{(k)}) & \partial_{Y_2} \Psi_3(Y^{(k)}) & \partial_{Y_3} \Psi_3(Y^{(k)}) \end{pmatrix}$$

$$= \mathbb{I} - \Delta t \begin{pmatrix} \gamma Y_1^{(k)} - \tau Y_3^{(k)} & 0 & \beta - \tau Y_1^{(k)} \\ \tau Y_3^{(k)} & -(\alpha + \mu) & \tau Y_1^{(k)} \\ 0 & \alpha & -(\beta + \nu) \end{pmatrix}.$$

In practice, we do not compute the inverse matrix of  $M^{(k)}$  and stop the process after a finite number of steps. Having  $Y^{(k)}$  and therefore this matrix, we proceed in two steps to determine  $Y^{(k+1)}$ :

- solving the linear system  $M^{(k)}Z^{(k)} = \Psi(Y^k)$ ;
- updating the formula  $Y^{(k+1)} = Y^{(k)} - Z^{(k)}$ .

We then employ a stopping criteria, for example by stopping the algorithm when the relative error  $\frac{\|Y^{(k+1)} - Y^{(k)}\|}{\|Y^{(k)}\|}$  reaches a certain limit set previously. We complete this test by conditioning the number of iterations that allows us to detect a possible difficulty in making the method converge.

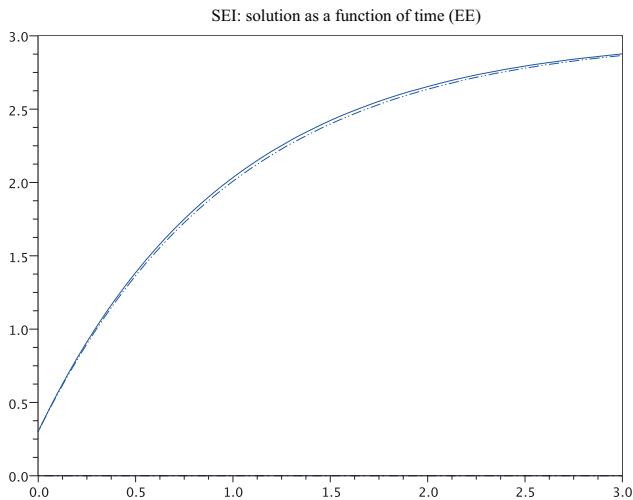
In order to confirm the accuracy of the numerical method, we compare the numerical solution with the known correct solution when  $E_{\text{Init}} = 0 = I_{\text{Init}}$ . The chosen parameters are as follows:  $\lambda = 3$ ,  $\gamma = 1$ ,  $\tau = 1$ ,  $\beta = 1$ ,  $\mu = 5$ ,  $\nu = 1$ ,  $\alpha = 1$  and  $S_{\text{Init}} = 0.3$ . We perform a simulation until the final time  $T = 3$  with a time step  $\Delta t = 0.05$ . Both the explicit and implicit methods provide similar results, which are shown in Figure 1.9. We observe that the exposed and infectious populations indeed remain completely absent throughout the simulation and that the simulation's quality degenerates somewhat for higher simulation times.

We now study a case where the exposed and infectious populations are present. We keep the parameters the same as in the previous simulation, except the initial values  $S_{\text{Init}} = 0.3$ ,  $E_{\text{Init}} = 0.2$  and  $I_{\text{Init}} = 0.4$ . In this case, we have  $\rho = 0.25 < 1$ . Figure 1.10 shows the extinction of the exposed and infectious populations, as well as convergence in the long term towards the healthy state, with  $\lim_{t \rightarrow \infty} S(t) = \lambda/\gamma$  (in order to account for this last result, it is necessary to continue the simulation on longer time spans than those shown in Figure 1.10, whose final time is only  $T = 2$ ).

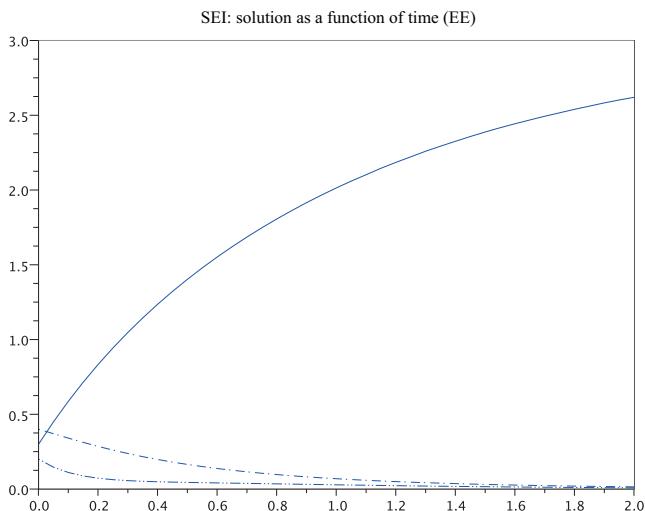
Finally, we study a case where  $\lambda = 13$ ,  $\gamma = 1$ ,  $\tau = 67$ ,  $\beta = 1$ ,  $\mu = 5$ ,  $\nu = 1$  and  $\alpha = 1$ , which leads to  $\rho = 72.5833$ . The initial values are the same, that is,  $S_{\text{Init}} = 0.3$ ,  $E_{\text{Init}} = 0.2$  and  $I_{\text{Init}} = 0.4$ . We can then expect to reach a stationary state with some infectious and exposed populations. Figure 1.11 shows the result of simulations using the explicit and implicit Euler schemes with interval  $\Delta t = 0.05$  until the final time  $T = 1.1$ . We observe that the explicit Euler scheme produces incoherent results after  $t \simeq 0.9$ . In particular, one of the quantities becomes negative. These incoherences become worst in time and the simulation can no longer be continued with these numerical conditions. We therefore use the implicit Euler scheme to explore the solution's behavior on longer spans of time. From Figure 1.12, at the final time  $T = 3$ , we can clearly see that a non-trivial stationary configuration takes hold.

### 1.2.8. A chemotaxis problem

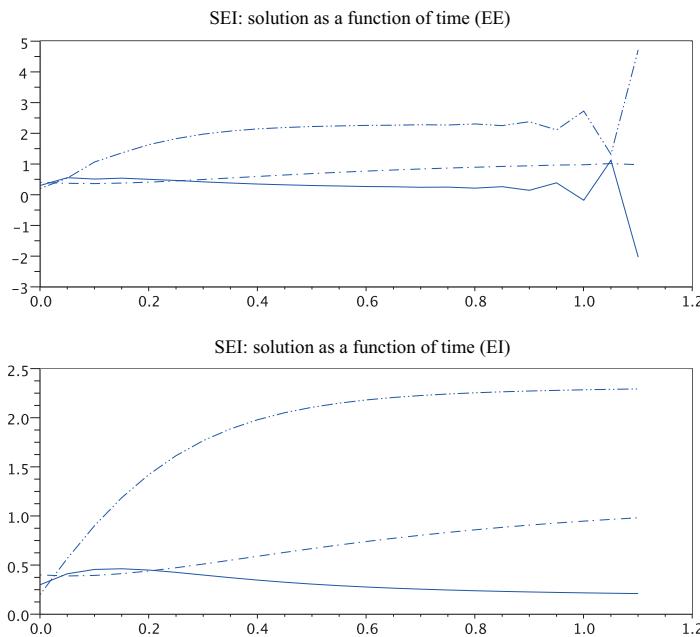
We set out to study a system of differential equations that describes bacterial behavior (for example, *E. coli*). In order to simplify, we assume that the bacteria only move in a straight line (mono-dimensional description) and divide the domain of



**Figure 1.9.** SEI model: comparison with an exact solution



**Figure 1.10.** SEI model: convergence to the healthy state (S represented by a bold line, E represented by a dotted line, I represented by a double dotted line)



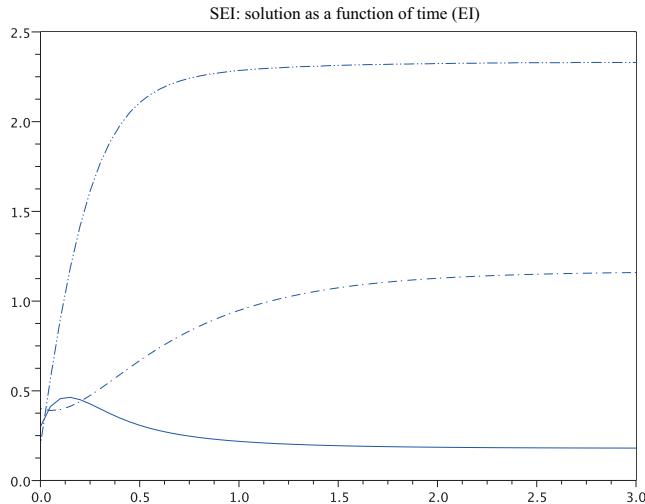
**Figure 1.11.** *SEI model: comparison of the explicit and implicit Euler schemes in an epidemic scenario (S represented by a bold line, E represented by a dotted line, I represented by a double dotted line)*

study into same sized locations, identified by the index  $j \in \{1, \dots, J\}$ . The dynamic is governed by the following principles:

- at each unit of time  $\tau > 0$ , a proportion  $0 < p < 1$  of the bacteria present at location  $j$  migrates towards one of the neighboring locations (so to either  $j - 1$  or  $j + 1$ );
- the bacteria react to an attractive chemical potential. A certain substance is emitted by the bacteria themselves, which move according to the signal they perceive.

We let  $X_j(t)$  denote the concentration of bacteria present at location  $j$ . The chemical attractive effect is described as follows. Each bacteria acts over a long range, but with an amplitude that diminishes as it goes farther away from the source. More specifically, given a decreasing function  $E : ]0, \infty[ \rightarrow ]0, \infty[$ ,  $\lim_{z \rightarrow 0} E(z) = \infty$ , we write

$$U_j = - \sum_{k \neq j} (j - k) E(|j - k|) X_k.$$



**Figure 1.12.** *SEI model: simulation in an epidemic scenario ( $S$  represented by a bold line,  $E$  represented by a dotted line,  $I$  represented by a double dotted line)*

Exchanges between points  $j$  and  $j + 1$  depend on the average value

$$U_{j+1/2} = \frac{U_j + U_{j+1}}{2}.$$

If  $U_{j+1/2} > 0$ , the bacteria migrate towards the location  $j+1$ ; instead, if  $U_{j+1/2} < 0$ , the location  $j$  is reinforced by an inflow of bacteria coming from the location  $j + 1$ . We therefore write

$$X_{j+1/2} = \begin{cases} X_j & \text{if } U_{j+1/2} \geq 0, \\ X_{j+1} & \text{if } U_{j+1/2} < 0. \end{cases}$$

We thus obtain the following differential system:

$$\frac{d}{dt}X_j = \frac{p}{2\tau}(X_{j+1} - 2X_j + X_{j-1}) - (U_{j+1/2}X_{j+1/2} - U_{j-1/2}X_{j-1/2}),$$

where we extend  $X$  by 0 when necessary: we let  $X_0 = 0 = X_{J+1}$ . In condensed form, with  $= (X_1, \dots, X_J)$ , we can write

$$\frac{d}{dt}X = AX - \mathcal{U}(X), \quad A = \frac{p}{2\tau} \begin{pmatrix} -2 & 1 & 0 & \cdots & 0 \\ 1 & -2 & 1 & \cdots & \vdots \\ 0 & 1 & -2 & \cdots & 0 \\ \vdots & \ddots & \ddots & \ddots & 1 \\ 0 & 0 & 1 & -2 & \end{pmatrix}$$

where the components of the function  $\mathcal{U}$  are defined by

$$\begin{aligned} X = (X_1, \dots, X_J) &\mapsto \mathcal{U}(X) \\ &= (U_{3/2}X_{3/2} - U_{1/2}X_{1/2}, \dots, U_{J+1/2}X_{J+1/2} - U_{J-1/2}X_{J-1/2}). \end{aligned}$$

It can be convenient to rewrite this in a clever form as

$$U_{j+1/2}X_{j+1/2} = \frac{1}{2}(U_{j+1/2} + |U_{j+1/2}|)X_j + \frac{1}{2}(U_{j+1/2} - |U_{j+1/2}|)X_{j+1},$$

which allows us to determine the locally Lipschitz continuous nature of the function  $X \mapsto \mathcal{U}(X)$ , thereby proving that the Cauchy problem is well defined, at least locally in time. This formulation also makes it fairly simple to program the Euler schemes that approximate the problem.

In light of their physical interpretation, the unknown  $X_j$ s must remain positive. To ensure this, we fix  $T > 0$ , smaller than the lifespan of the maximal solution, and we designate as  $M$  an upper bound of the  $U_{j+1/2}$  on  $[0, T]$ . We consider the evolution of the negative part of the solution: by denoting  $[X_j]_- = \min(0, X_j) \geq 0$ , we have

$$\begin{aligned} \frac{1}{2} \frac{d}{dt} \sum_{j=1}^J [X_j]_-^2 &= \frac{p}{2\tau} \sum_{j=1}^J (X_{j+1} - X_j - (X_j - X_{j-1})) [X_j]_- - \sum_{j=1}^J \mathcal{U}(X)_j [X_j]_- \\ &= -\frac{p}{2\tau} \sum_{j=1}^J (X_{j+1} - X_j) ([X_{j+1}]_- - [X_j]_-) \\ &\quad - \frac{1}{2} \sum_{j=1}^J (U_{j+1/2} + |U_{j+1/2}|) X_j [X_j]_- - \frac{1}{2} \sum_{j=1}^J (U_{j+1/2} - |U_{j+1/2}|) X_{j+1} [X_j]_- \\ &\quad + \frac{1}{2} \sum_{j=1}^J (U_{j-1/2} + |U_{j-1/2}|) X_{j-1} [X_j]_- + \frac{1}{2} \sum_{j=1}^J (U_{j-1/2} - |U_{j-1/2}|) X_j [X_j]_. \end{aligned}$$

However, on the one hand, we have  $X_j[X_j]_- = [X_j]_-^2$  and, on the other hand, because

$$\frac{1}{2}(\Phi_{j+1/2} - |\Phi_{j+1/2}|) = [\Phi_{j+1/2}]_- \leq 0$$

it makes it possible to bound

$$-(\Phi_{j+1/2} - |\Phi_{j+1/2}|)X_{j+1}[X_j]_- \leq M[X_{j+1}]_-[X_j]_-$$

from above. Since  $(X_{j+1} - X_j)([X_{j+1}]_- - [X_j]_-) \geq 0$ , it follows that

$$\frac{1}{2} \frac{d}{dt} \sum_{j=1}^J [X_j]_-^2 \leq M \sum_{j=1}^J [X_{j+1}]_- [X_j]_- \leq 2M \sum_{j=1}^J [X_j]_-^2$$

using the Cauchy–Schwarz inequality. Thus, Grönwall’s lemma leads to  $\sum_{j=1}^J [X_j(t)]_-^2 \leq e^{2Mt} \sum_{j=1}^J [X_{\text{Init},j}]_-^2$ . Therefore, if the components of the initial value are non-negative, we have  $[X_{\text{Init},j}]_- = 0$  and the upper bound cancels out, so we infer that  $X_j(t) \geq 0$  for all  $0 \leq t \leq T$ .

Let us now study the behavior of this differential system numerically. Given  $\Delta t > 0$ , the explicit Euler scheme is written as

$$Y^{n+1} = Y^n + \Delta t A Y^n - \Delta t \mathcal{U}(Y^n)$$

where  $Y^n = (Y_1^n, \dots, Y_J^n)$  must approach  $X(n\Delta t) = (X_1(n\Delta t), \dots, X_J(n\Delta t))$  and we still assume that  $Y_0^n = 0 = Y_{J+1}^n$ . Even though the role of the nonlinear term  $\mathcal{U}(X)$  might not be evident at first glance, we can still note that the linear term is clearly dissipative because  $A\xi \cdot \xi = -\sum_{j=1}^J |\xi_j - \xi_{j+1}|^2 \leq 0$ . It is therefore worthwhile to make the processing of that term implicit, especially because it does not require finding the root of a complicated function, but only finding the inverse of the linear system. This approach can also be motivated in terms of preserving the solutions’ non-negative values. Indeed, let us consider the case where  $\mathcal{U}(X) = 0$ , so the explicit Euler scheme leads to

$$Y_j^{n+1} = \left(1 - \frac{p}{\tau} \Delta t\right) Y_j^n + \frac{p}{2\tau} \Delta t Y_{j-1}^n + \frac{p}{2\tau} \Delta t Y_{j+1}^n.$$

Therefore, non-negative values are preserved at each step in time when  $Y_j^{n+1}$  is a convex combination of  $Y_j^n$ ,  $Y_{j-1}^n$  and  $Y_{j+1}^n$ , which requires a restriction on the time step  $\Delta t < \frac{\tau}{p}$ . When the ratio  $\tau/p$  is “small” (for example, when the unit of time

$\tau$  is small compared with the observation scale), this condition is actually restrictive and can have a considerable impact on the computation time necessary to evaluate the solution up to a given final time  $T > 0$ . The implicit Euler scheme for this simplified problem where  $\mathcal{U}(X) = 0$  is

$$\left(\mathbb{I} - \frac{p}{2\tau}\Delta t A\right)Y^{n+1} = Y^n.$$

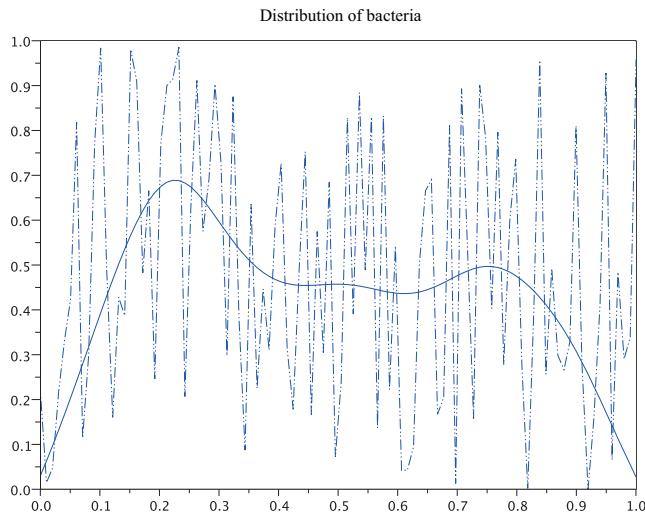
The fact that the components of  $Y^{n+1}$  remain non-negative if those of  $Y^n$  are also non-negative is related to a remarkable property of the matrix  $(\mathbb{I} - \Delta t A)$ : it is an  $M$ -matrix for all  $\Delta t > 0$ . We will study the details of this question when the matrix  $A$  will be involved in the discretization of certain partial differential equations (see Chapter 2 and section 3.1). For the problem at hand, we therefore recommend using a *semi-implicit* strategy with the scheme

$$\left(\mathbb{I} - \frac{p}{2\tau}\Delta t A\right)Y_j^{n+1} = Y_j^n - \Delta t \mathcal{U}(Y^n).$$

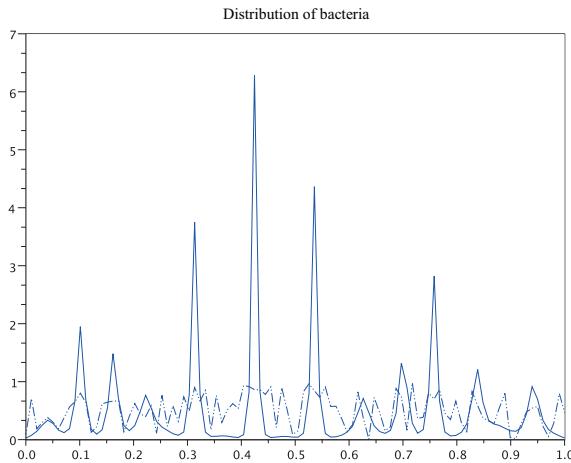
Clearly, the “stability” difficulties can arise from the nonlinear term  $\mathcal{U}(Y) \neq 0$ , whose analysis is a much more delicate matter. However, this scheme has the advantage of remaining simple and overcoming an immediately sensitive difficulty associated with the explicit Euler scheme.

This very simple model allows us to show the rich behavior, depending on the parameter  $p/\tau$ , the amplitude of the connecting term  $\mathcal{U}$  and the initial value. For the simulations, we take  $E(z) = C/|z|^3$ ,  $C > 0$  and a random initial value in  $[0, 1]$ . We begin with a trial at  $p = 0.95$ ,  $\tau = 1$ ,  $C = 0.1$ . Here we have no restrictions on  $\Delta t$ . We take  $\Delta t = 0.1$  and the final time is  $T = 10$ . We do not find any notable differences between the results of the explicit Euler scheme and those of the semi-implicit scheme. Figure 1.13 shows a relatively smooth profile, where the maximal amplitude remains below 1. The situation is radically different when we take  $C = 1$ , and keep the other parameters the same: Figure 1.14 shows the appearance of “peaks” with very large amplitudes.

We perform the same simulations now with  $\tau = 0.03$ . For  $\Delta t = 0.1$ , the stability condition that guarantees that the results are non-negative is violated and we can effectively see that the explicit Euler scheme does not work. The semi-implicit scheme allows us to perform the simulation in those conditions. Figure 1.15 shows the result obtained with  $C = 0.1$ : we see a regular profile, which has a tendency to be smoothed out over time. Figures 1.16 and 1.17 correspond to the same data but with  $C = 8$  and  $C = 10$ . They show peaks with a very localized density. If we increase  $C$  further, if we increase the amplitude of the initial value, or if we perform simulations for higher final times, we will encounter stability difficulties due to these singular structures and to the nonlinear term.



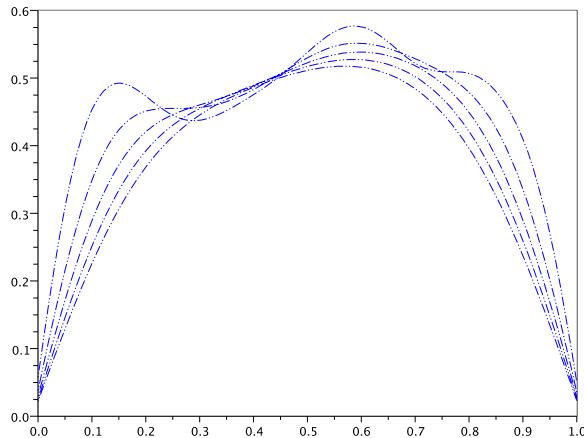
**Figure 1.13.** Chemotaxis: initial value (dotted) and solution at the final time  $T = 10$  for  $p = 0.95$ ,  $\tau = 1$ ,  $C = 0.1$



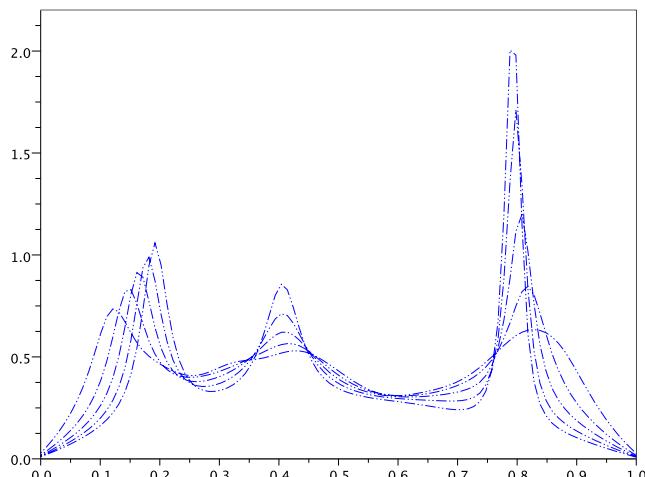
**Figure 1.14.** Chemotaxis: initial value (dotted) and solution at the final time  $T = 10$  for  $p = 0.95$ ,  $\tau = 1$ ,  $C = 1$

The mathematical study of these fascinating phenomena and their biological implications are the object of intense research activity: the one presented here is a discrete version of a model proposed by Keller–Segel [KEL 70]. Original results can

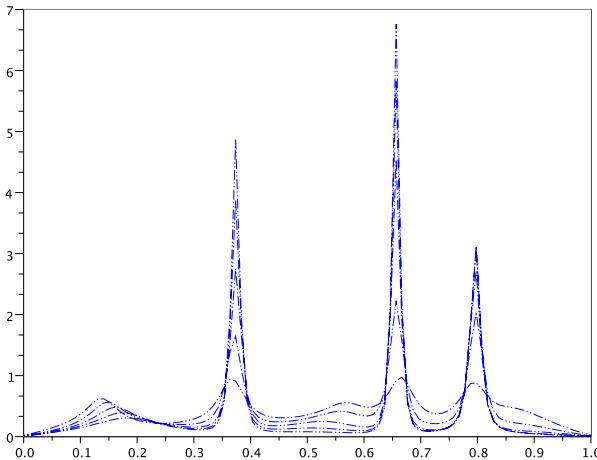
be found in [JÄG 92, RAS 95] as well as in the review articles [HOR 03, HOR 04]. Very similar problems are also at play in astrophysics with models of galaxy formation under the effect of gravitational forces [CHA 43].



**Figure 1.15.** Chemotaxis: solutions at different times for  $p = 0.95, \tau = 0.03, C = 0.1$



**Figure 1.16.** Chemotaxis: solutions at different times  $T \leq 10$  for  $p = 0.95, \tau = 0.03, C = 8$



**Figure 1.17.** Chemotaxis: solutions at different times  $T \leq 10$  for  $p = 0.95, \tau = 0.03, C = 10$

### 1.3. Hamiltonian problems

A very large class of differential systems are of the following form:

$$\frac{d}{dt} \begin{pmatrix} q \\ p \end{pmatrix} = \begin{pmatrix} \nabla_p H(q, p) \\ -\nabla_q H(q, p) \end{pmatrix}, \quad [1.31]$$

for a certain function  $(q, p) \in \mathbb{R}^D \times \mathbb{R}^D \mapsto H(q, p) \in \mathbb{R}$ , the *Hamiltonian* of the system. We let  $\nabla_p H(q, p) = (\partial_{p_1} H(q, p), \dots, \partial_{p_D} H(q, p))$  and  $\nabla_q H(q, p) = (\partial_{q_1} H(q, p), \dots, \partial_{q_D} H(q, p))$ . We will assume that

$$H \text{ is a function of class } C^2 \text{ on the domain } \mathcal{U} \subset \mathbb{R}^D \times \mathbb{R}^D. \quad [1.32]$$

This assumption allows us to apply the Picard–Lindelöf theorem, which ensures the local existence and uniqueness of a solution for each initial value  $(t_{\text{Init}}, q_{\text{Init}}, p_{\text{Init}})$ . Then, the key statement is related to the fact that the Hamiltonian is preserved: for every solution of [1.31], we have

$$\frac{d}{dt} H(q(t), p(t)) = \nabla_q H(q(t), p(t)) \cdot \frac{d}{dt} q(t) + \nabla_p H(q(t), p(t)) \cdot \frac{d}{dt} p(t) = 0.$$

The properties of the function  $H$  often allows us to draw estimates on the solution that guarantee its global existence from the conservation  $H(q(t), p(t)) = H(q_{\text{Init}}, p_{\text{Init}})$ .

This kind of problems arise naturally in mechanics. We present a classical example as follows. Let us consider a particle subject to a force field due to a potential  $\Phi$ . Letting  $m$  denote the mass of the particle,  $q(t)$  its position at time  $t$  and  $p(t)$  its velocity, the fundamental principle of dynamics leads to

$$\frac{d}{dt}q(t) = p(t), \quad m\frac{d}{dt}q(t) = -\nabla_q\Phi(q(t)).$$

The corresponding Hamiltonian is

$$H(q, p) = m\frac{|p|^2}{2} + \Phi(q),$$

the sum of the kinetic energy  $m\frac{|p|^2}{2}$  and the potential energy  $\Phi(q)$ . We say that the potential  $\Phi$  is confining when there exists  $C > 0$ , such that for every  $q$ , we have  $\Phi(q) \geq -C$ . In this case, the system is globally well defined. Indeed,  $m\frac{|p(t)|^2}{2} \leq H(q(t), p(t)) + C = H(q_{\text{Init}}, p_{\text{Init}}) + C$  implies that the velocity remains bounded in time. Thus, we have  $q(t) = q_{\text{Init}} + \int_0^t p(s) ds$ , which allows us to establish that  $|q(t)| \leq M(1+t)$  for a certain constant  $M > 0$ . These estimates exclude the possibility of explosions in finite time. In what follows, it will be convenient to express [1.31] in the form

$$\frac{d}{dt} \begin{pmatrix} p \\ q \end{pmatrix} = \mathbb{J} \nabla_{q,p} H(q, p), \quad \mathbb{J} = \begin{pmatrix} 0 & \mathbb{I} \\ -\mathbb{I} & 0 \end{pmatrix}$$

with  $\nabla_{q,p} H(q, p) = (\partial_{q_1} H(q, p), \dots, \partial_{q_D} H(q, p), \partial_{p_1} H(q, p), \dots, \partial_{p_D} H(q, p))$ . The properties of the matrix  $\mathbb{J}$  play a crucial role in the analysis of Hamiltonian systems. A direct calculation allows us to determine the following properties.

**LEMMA 1.11.**— The matrix  $\mathbb{J}$  satisfies

- i)  $\det(\mathbb{J}) = 1$ ;
- ii)  $\mathbb{J}^\top = -\mathbb{J}$ ,  $\mathbb{J}^2 = -\mathbb{I}$ ,  $\mathbb{J}^{-1} = \mathbb{J}^\top = -\mathbb{J}$ .

Beyond energy conservation, systems of the form [1.31] satisfy another remarkable property: the conservation of volumes in the phase space. Indeed, we consider a domain in the phase space  $D_0 \subset \mathbb{R}^D \times \mathbb{R}^D$  with bounded volume

$$\mathcal{V}_0 = \int_{D_0} dq_0 dp_0 < \infty.$$

For  $t > 0$ , we focus on its image by the flow [1.31]

$$D_t = \{(q(t), p(t)), \text{ where } t \mapsto (q(t), p(t)) \text{ is a solution of [1.31]}$$

$$\text{with } (q(0), p(0)) = (q_0, p_0) \in D_0\}.$$

Thus, the volume of that domain is given by (Chapter 4 in [GOU 11]):

$$\mathcal{V}_t = \int_{D_t} dq dp = \int_{D_0} |\det(\mathcal{J}(t, q_0, p_0))| dq_0 dp_0,$$

where  $\mathcal{J}(t, q_0, p_0)$ ) denotes the Jacobian matrix of the change of variable  $(q, p) \mapsto (q_0, p_0)$ , that is

$$\mathcal{J}(t, q_0, p_0) = \begin{pmatrix} \partial_{q_{0,1}} q_1 & \dots & \partial_{q_{0,D}} q_1 & \partial_{p_{0,1}} q_1 & \dots & \partial_{p_{0,D}} q_1 \\ \partial_{q_{0,1}} q_2 & \dots & \partial_{q_{0,D}} q_2 & \partial_{p_{0,1}} q_2 & \dots & \partial_{p_{0,D}} q_2 \\ \vdots & \vdots & & & & \vdots \\ \partial_{q_{0,1}} q_D & \dots & \partial_{q_{0,D}} q_D & \partial_{p_{0,1}} q_D & \dots & \partial_{p_{0,D}} q_D \\ \partial_{q_{0,1}} p_1 & \dots & \partial_{q_{0,D}} p_1 & \partial_{p_{0,1}} p_1 & \dots & \partial_{p_{0,D}} p_1 \\ \partial_{q_{0,1}} p_2 & \dots & \partial_{q_{0,D}} p_2 & \partial_{p_{0,1}} p_2 & \dots & \partial_{p_{0,D}} p_2 \\ \vdots & \vdots & & & & \vdots \\ \partial_{q_{0,1}} p_D & \dots & \partial_{q_{0,D}} p_D & \partial_{p_{0,1}} p_D & \dots & \partial_{p_{0,D}} p_D \end{pmatrix} (t).$$

**PROPOSITION 1.5.–** The solutions of [1.31] satisfy

- i) for all  $t \geq 0$ ,  $\det(\mathcal{J}(t)) = 1$ , so  $\mathcal{V}_t = \mathcal{V}_0$  (conservation of volume);
- ii) for all  $t \geq 0$ ,  $\mathcal{J}(t)^T \mathbb{J} \mathcal{J}(t) = \mathbb{J}$  (the matrix  $\mathcal{J}(t)$  is *symplectic*).

**PROOF.–** We introduce the vector  $U(q, p) = \mathbb{J} \nabla_{q,p} H(q, p) = (\nabla_p H(q, p), -\nabla_q H(q, p))$  and let  $X = (q, p)$ ,  $x_0 = (q_0, p_0)$ . We observe that

$$\begin{aligned} \frac{d}{dt} \mathcal{J}_{nm}(t) &= \partial_{x_{0,m}} \frac{d}{dt} X_n(t) = \partial_{x_{0,m}} [U_n(X(t))] = \sum_{k=1}^{2D} \partial_{X_k} U_n(X(t)) \partial_{x_{0,m}} X_k(t) \\ &= \sum_{k=1}^{2D} \nabla_X U(X(t))_{nk} \mathcal{J}_{km}(t) = [\nabla_X U \mathcal{J}(t)]_{nm} \end{aligned}$$

where

$$\nabla_X U = \begin{pmatrix} \partial_{X_1} U_1 & \partial_{X_2} U_1 & \dots & \partial_{X_{2D}} U_1 \\ \partial_{X_1} U_2 & \partial_{X_2} U_2 & \dots & \partial_{X_{2D}} U_2 \\ \vdots & \vdots & & \vdots \\ \partial_{X_1} U_{2D} & \partial_{X_2} U_{2D} & \dots & \partial_{X_{2D}} U_{2D} \end{pmatrix}.$$

In the case where  $H(q, p) = \frac{|p|^2}{2} + \Phi(q)$ , we have

$$\nabla_X U = \begin{pmatrix} 0 & \mathbb{I} \\ -D_q^2(\Phi) & 0 \end{pmatrix},$$

with  $D_q^2\Phi$  of the Hessian matrix for  $\Phi$ . In general, returning to the variables  $(q, p)$ , we obtain

$$\nabla_X U(X) = \mathbb{J} D_{q,p}^2 H(q, p).$$

It follows that

$$\begin{aligned} \operatorname{div}_X(U) &= \partial_{q_1} U_1 + \dots + \partial_{q_D} U_D + \partial_{p_1} U_{D+1} + \dots + \partial_{p_D} U_{2D} = \operatorname{tr}(\nabla_X U) \\ &= \nabla_q \cdot \nabla_p H + \nabla_p \cdot (-\nabla_q H) = \sum_{i=1}^D (\partial_{q_i} \partial_{p_i} H - \partial_{p_i} \partial_{q_i} H) = 0. \end{aligned}$$

However, for an invertible matrix  $A$ , the derivative of the determinant mapping is defined by

$$\det'(A)(H) = \operatorname{tr}(A^{-1} H) \det(A).$$

In order to determine this relation, on the one hand, we write  $\det(A + H) = \det(A(I + A^{-1}H)) = \det(A)\det(\mathbb{I} + A^{-1}H)$  and, on the other hand, we show that  $\det(\mathbb{I} + \tilde{H}) = 1 + \operatorname{tr}(\tilde{H}) + R(\tilde{H})$ , where the remainder satisfies  $\lim_{\|\tilde{H}\| \rightarrow 0} \frac{\|R(\tilde{H})\|}{\|\tilde{H}\|} = 0$ . The last property can be obtained by mathematical induction on dimension  $N$ . It is clear that it is satisfied in dimension  $N = 1$ . Let us assume that it is satisfied for an integer  $N$  and consider a matrix  $H$  of  $\mathcal{M}_{N+1}(\mathbb{C})$ . We decompose  $\mathbb{I}_{N+1} + H$  in the form

$$\mathbb{I}_{N+1} + H = \left[ \begin{array}{c|cccc} 1+h & t_1 & t_2 & \dots & t_N \\ \hline s_1 & & & & \\ s_2 & & & & \\ \vdots & & & & \\ s_{N-1} & & & & \mathbb{I}_N + \tilde{H} \\ s_N & & & & \end{array} \right]$$

with  $\tilde{H} \in \mathcal{M}_N(\mathbb{C})$ . We develop the determinant with respect to the first row to obtain  $\det(I + H) = (1 + h)\det(\mathbb{I}_N + \tilde{H}) + r(H)$ , where the remainder  $r(H)$  indeed satisfies the desired estimate. However, according to the induction hypothesis, we have  $(1 + h)\det(\mathbb{I}_N + \tilde{H}) = (1 + h)(1 + \operatorname{tr}(\tilde{H}) + \tilde{R}(\tilde{H}))$ , which is indeed equal to  $1 + (h + \operatorname{tr}(\tilde{H})) = 1 + \operatorname{tr}(H) + \text{remainder terms that satisfy the estimate as desired}$ .

Therefore, we have

$$\frac{d}{dt} \det(\mathcal{J}) = \text{tr}(\mathcal{J}^{-1} \nabla_X U \mathcal{J}) \det(\mathcal{J}) = \text{div}_X(U) \det(\mathcal{J}) = 0.$$

So  $t \mapsto \det(\mathcal{J}(t))$  is constant. Since for  $t = 0$  we have  $\mathcal{J}(0) = \mathbb{I}$ , we conclude that  $\det(\mathcal{J}(t)) = 1$ .

We use the same observations to demonstrate iii). Indeed, we have

$$\begin{aligned} \frac{d}{dt} [\mathcal{J}^\top \mathbb{J} \mathcal{J}] &= [\nabla_X U \mathcal{J}]^\top \mathbb{J} \mathcal{J} + \mathcal{J}^\top \mathbb{J} [\nabla_X U \mathcal{J}] = \mathcal{J}^\top (\nabla_X U^\top \mathbb{J} + \mathbb{J} \nabla_X U) \mathcal{J} \\ &= \mathcal{J}^\top ((\mathbb{J} D^2 H)^\top \mathbb{J} + \mathbb{J} \mathbb{J} D^2 H) \mathcal{J} = \mathcal{J}^\top (D^2 H(-\mathbb{J}) \mathbb{J} + \mathbb{J} \mathbb{J} D^2 H) \mathcal{J} = 0 \end{aligned}$$

using the fact that  $D^2 H$  is symmetric and lemma 1.11. We therefore conclude that  $\mathcal{J}(t)^\top \mathbb{J} \mathcal{J}(t) = \mathcal{J}(0)^\top \mathbb{J} \mathcal{J}(0) = \mathbb{J}$ . We also note that this property implies ii) since it leads to  $\det(\mathbb{J}) = 1 = \det(\mathcal{J}(t)^\top \mathcal{J} \mathcal{J}(t)) = \det(\mathbb{J}) [\det(\mathcal{J}(t))]^2$  and therefore  $[\det(\mathcal{J}(t))]^2 = 1$ . However, for  $t = 0$ , we have  $\det(\mathcal{J}(0)) = 1$ . It follows, by continuity, that  $\det(\mathcal{J}(t)) = 1$  for all  $t \geq 0$ .  $\square$

The typical example of the Hamiltonian problem is given by the description of a pendulum's movement, which we will now describe.

### 1.3.1. The pendulum problem

Consider a pendulum subject to the action of its own weight. The pendulum's position is tracked by the vector  $OM(t)$  with coordinates  $\ell(\sin(\theta(t)), \cos(\theta(t)))$  in the canonical basis  $(e_x, e_y)$ , with  $\ell$  being the cord length. Newton's second law of motion leads to the vectorial relation

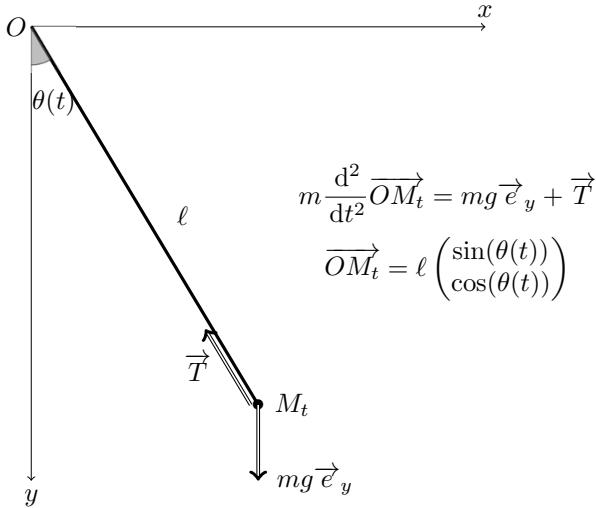
$$m \frac{d^2}{dt^2} OM(t) = m g e_y + T = \ell \begin{pmatrix} \frac{d^2 \theta(t)}{dt^2} \cos(\theta(t)) - \left( \frac{d\theta(t)}{dt} \right)^2 \sin(\theta(t)) \\ - \frac{d^2 \theta(t)}{dt^2} \sin(\theta(t)) - \left( \frac{d\theta(t)}{dt} \right)^2 \cos(\theta(t)) \end{pmatrix}$$

where  $g > 0$  is the acceleration of gravity,  $m$  is the pendulum's mass and  $T$  is the traction exerted by the cord, which is a force carried by the vector  $OM$ .

We project this relation in the direction perpendicular to  $OM(t)$ . This makes it possible to get rid of the traction force. The following formula is therefore obtained

by multiplying the first component by  $\cos(\theta(t))$  and the second one by  $-\sin(\theta(t))$  and then performing the sum:

$$\frac{d^2\theta(t)}{dt^2} = -\frac{g}{\ell} \sin(\theta(t)). \quad [1.33]$$



**Figure 1.18. Simple pendulum**

We note that  $\frac{g}{\ell}$  is homogeneous to the square of a frequency (that is to  $1/\text{time}^2$ ). We let  $\xi(t) = \frac{d\theta(t)}{dt}$  and  $\Theta(t) = (\theta(t), \xi(t))$  in such a way that this second-order equation can be rewritten like a first-order system

$$\frac{d}{dt}\Theta(t) = \begin{pmatrix} \xi(t) \\ -\frac{g}{\ell} \sin(\theta(t)) \end{pmatrix} = f(\Theta(t)).$$

This system is autonomous: the function  $f : (\theta, \xi) \in \mathbb{R}^2 \mapsto (\xi, -\frac{g}{\ell} \sin(\theta)) \in \mathbb{R}^2$  is  $C^1$  and its derivatives are uniformly bounded. Therefore, the Cauchy problem has a unique solution defined for all times. It is a Hamiltonian system associated with the function

$$H(\theta, \xi) = \frac{1}{2}|\xi|^2 + \frac{g}{\ell}(1 - \cos(\theta)).$$

The quantity  $H(\theta(t), \xi(t))$  is conserved throughout time.

For small oscillations  $|\theta(t)| \ll 1$ , we can approximate  $\sin(\theta(t))$  by  $\theta(t)$ . We can therefore expect that the solution of [1.33] remains near the solution  $\theta_L$  of the *linear* problem

$$\frac{d^2\theta_L(t)}{dt^2} = -\frac{g}{\ell}\theta_L(t). \quad [1.34]$$

This equation can be expressed in the form of a first-order linear system

$$\frac{d}{dt} \begin{pmatrix} \theta_L(t) \\ \xi_L(t) \end{pmatrix} = \begin{pmatrix} 0 & 1 \\ -g/\ell & 0 \end{pmatrix} \begin{pmatrix} \theta_L(t) \\ \xi_L(t) \end{pmatrix},$$

which can be associated with the following Hamiltonian

$$H_L(\theta, \xi) = \frac{1}{2} \left( |\xi|^2 + \frac{g}{\ell} |\theta|^2 \right).$$

Of course, replacing the problem [1.33] with the problem [1.34] is an approximation, which relies on the assumption of small derivations with respect to the equilibrium position  $\theta = 0$ . This approximation can be verified with the following statement.

**PROPOSITION 1.6.–** Let  $\varepsilon > 0$  and  $f_1, f_2$  be two functions defined on  $\mathbb{R}^+ \times \mathbb{R}^D$ , uniformly  $L$ -Lipschitz in the state variable and such that  $\|f_1 - f_2\|_\infty \leq \varepsilon$ . Consider the solutions  $x_j$  of  $\frac{d}{dt}x_j(t) = f_j(t, x_j(t))$ , with  $x_j(0) = x_{j,\text{Init}}$ . Then, we have

$$|x_1(t) - x_2(t)| \leq |x_{1,\text{Init}} - x_{2,\text{Init}}| e^{Lt} + \frac{\epsilon}{L} (e^{Lt} - 1).$$

**PROOF.–** We roughly estimate

$$\begin{aligned} |x_1(t) - x_2(t)| &= \left| x_{1,\text{Init}} - x_{2,\text{Init}} + \int_0^t (f_1(s, x_1(s)) - f_2(s, x_2(s))) ds \right| \\ &\leq |x_{1,\text{Init}} - x_{2,\text{Init}}| + \int_0^t |f_1(s, x_1(s)) - f_1(s, x_2(s))| ds \\ &\quad + \int_0^t |f_1(s, x_2(s)) - f_2(s, x_2(s))| ds \\ &\leq |x_{1,\text{Init}} - x_{2,\text{Init}}| + L \int_0^t |x_1(s) - x_2(s)| ds + \epsilon \int_0^t ds. \end{aligned}$$

Using Grönwall's lemma, we obtain

$$\begin{aligned} |x_1(t) - x_2(t)| &\leq |x_{1,\text{Init}} - x_{2,\text{Init}}| e^{Lt} + \epsilon \int_0^t e^{L(t-s)} ds \\ &= |x_{1,\text{Init}} - x_{2,\text{Init}}| e^{Lt} + \frac{\epsilon}{L} (e^{Lt} - 1). \end{aligned}$$

This estimate determines how close are the solutions of differential equations whose second members are close. However, we see that this estimate is based on the fact that  $\sin(\theta)$  and  $\theta$  are close when  $\theta$  is small. Given  $\epsilon > 0$ , we can find a small enough  $0 < \theta_0 < \pi/2$ , such that for all  $|\theta| \leq \theta_0$ , we have  $|\sin(\theta) - \theta| \leq \epsilon$ . We assume that the pendulum is released with no initial speed. Thus, conservation of energy implies that

$$\frac{1}{2} \left| \frac{d}{dt} \theta \right|^2 + \frac{g}{\ell} (1 - \cos(\theta)) = \frac{g}{\ell} (1 - \cos(\theta_0)), \quad \frac{1}{2} \left| \frac{d}{dt} \theta_L \right|^2 + \frac{g}{\ell} \frac{\theta_L^2}{2} = \frac{g}{\ell} \frac{\theta_0^2}{2}.$$

It follows that  $|\theta(t)| \leq \theta_0$ ,  $|\theta_L(t)| \leq \theta_0$ . We infer that

$$|\theta(t) - \theta_L(t)| + \left| \frac{d}{dt} \theta(t) - \frac{d}{dt} \theta_L(t) \right| \leq \epsilon e^t.$$

The solutions of [1.33] and [1.34] remain close for only short periods of time.  $\square$

The solutions of [1.34] are of the form

$$\theta_L(t) = A \cos(\omega t) + B \sin(\omega t), \quad \omega = \sqrt{\frac{g}{\ell}}$$

where  $A$  and  $B$  are determined with the initial conditions

$$A = \theta_L(0), \quad B = \frac{\frac{d}{dt} \theta_L(0)}{\omega}.$$

In particular, the solution of [1.34] satisfies

$$\frac{d}{dt} (|\xi_L(t)|^2 + \frac{g}{\ell} |\theta_L(t)|^2) = 0$$

and the solutions describe ellipses in the phase plane  $(\theta, \xi)$ . We therefore do not need a numerical scheme to know the solution of [1.34]. Nevertheless, it is interesting to study the behavior of schemes, for which we have determined convergence, on this simple example, and to note that they can have different properties with respect to energy criteria.

In order to analyze the behavior of numerical schemes, it is convenient to modify the time scale, by letting

$$s = \omega t.$$

We thus work with a variable  $s$  with no dimension. We write  $\bar{\theta}_L(s) = \theta_L(s/\omega)$  and  $\bar{\xi}_L(s) = \frac{1}{\omega}\xi_L(s/\omega)$ . We obtain

$$\frac{d}{ds} \begin{pmatrix} \bar{\theta}_L \\ \bar{\xi}_L \end{pmatrix} = \mathbb{J} \begin{pmatrix} \bar{\theta}_L \\ \bar{\xi}_L \end{pmatrix}, \quad \mathbb{J} = \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix},$$

that is to say [1.34] where we have replaced  $g/\ell$  with 1. With these variables, the trajectories describe circles in the phase plane  $(\bar{\theta}, \bar{\xi})$  because conservation of energy is expressed as  $\frac{d}{ds}|\bar{\Theta}_L(s)|^2$ . Henceforth, we will discuss the behavior of numerical schemes with  $g/\ell = 1$  and abandon the notation  $\bar{\theta}_L, \bar{\xi}_L$ . The explicit Euler scheme [1.11] for [1.34] takes the form

$$\Theta_{EE}^{n+1} = (\mathbb{I} + \Delta t \mathbb{J}) \Theta_{EE}^n.$$

so

$$|\Theta_{EE}^{n+1}|^2 = |\Theta_{EE}^n|^2 + \Delta t^2 |\mathbb{J} \Theta_{EE}^n|^2 + 2\Delta t \mathbb{J} \Theta_{EE}^n \cdot \Theta_{EE}^n = (1 + \Delta t^2) |\Theta_{EE}^n|^2$$

whereas with the implicit Euler scheme [1.12], we have

$$(\mathbb{I} - \Delta t \mathbb{J}) \Theta_{EI}^{n+1} = \Theta_{EI}^n$$

and

$$\begin{aligned} |(\mathbb{I} - \Delta t \mathbb{J}) \Theta_{EI}^{n+1}|^2 &= |\Theta_{EI}^{n+1}|^2 + \Delta t^2 |\mathbb{J} \Theta_{EI}^{n+1}|^2 - 2\Delta t \mathbb{J} \Theta_{EI}^{n+1} \cdot \Theta_{EI}^{n+1} \\ &= (1 + \Delta t^2) |\Theta_{EI}^{n+1}|^2 = |\Theta_{EI}^n|^2. \end{aligned}$$

Therefore, the explicit Euler scheme makes energy increase; it is energetically unstable, whereas the explicit Euler scheme dissipates energy, so it is energetically stable. For both cases, this is somewhat unsatisfying, since energy is conserved in the continuous problem. The Crank–Nicolson scheme

$$\left(\mathbb{I} - \frac{\Delta t}{2} \mathbb{J}\right) \Theta_{CN}^{n+1} = \left(\mathbb{I} + \frac{\Delta t}{2} \mathbb{J}\right) \Theta_{CN}^n$$

fulfills the criterion: it preserves energy because  $|\Theta_{CN}^{n+1}|^2 = |\Theta_{CN}^n|^2$ , with the same numerical cost as the implicit Euler scheme.

This discussion allows us to effectively understand the behavior and the properties of the different schemes, even if for the linear problem, its value is limited, since, as mentioned earlier, the exact solution is well known. With these remarks in mind, we can return to the nonlinear problem. In order to construct well-performing schemes,

we will seek to preserve, or approximate, the conservation of energy and the geometric properties of the continuous problem described in proposition 1.5. We write

$$q^{n+1} = q^n + \Delta t \nabla_p H(q^{n+1}, p^n),$$

$$p^{n+1} = p^n - \Delta t \nabla_q H(q^{n+1}, p^n).$$

This scheme is consistent of order 1. It is of an implicit nature and, in general, requires appealing to a fixed point method. However, in several practical situations, as in the pendulum problem, we simply have  $H(q, p) = h(q) + \Phi(p)$ . In this case, the scheme is written in an explicit form because we have

$$q^{n+1} = q^n + \Delta t \nabla_p h(p^n),$$

which updates  $q^{n+1}$ , so

$$p^{n+1} = p^n - \Delta t \nabla_q \Phi(q^{n+1}).$$

For the linear pendulum, this gives

$$\theta^{n+1} = \theta^n + \Delta t \xi^n, \quad \xi^{n+1} = \xi^n - \Delta t \theta^{n+1},$$

and for the nonlinear pendulum,

$$\theta^{n+1} = \theta^n + \Delta t \xi^n, \quad \xi^{n+1} = \xi^n - \Delta t \sin(\theta^{n+1}).$$

(Recall that the temporal scale has been modified in order to replace  $g/\ell$  with 1). The scheme does not exactly preserve the Hamiltonian  $H$ . However, for the approximated Hamiltonian defined by

$$\tilde{H}(q, p) = H(q, p) + \frac{\Delta t}{2} \nabla_p H(q, p) \cdot \nabla_q H(q, p)$$

we manage to show, with the help of the somewhat tedious Taylor expansion, that<sup>11</sup>  $\tilde{H}(q^{n+1}, p^{n+1}) = \tilde{H}(q^n, p^n) + \mathcal{O}(\Delta t^3)$ . Thus, finally,

$$|\tilde{H}(q^n, p^n) - \tilde{H}(q^0, p^0)| \leq C \Delta t^2.$$

In the case of the linear pendulum, this approximated Hamiltonian  $\tilde{H}$  is, in fact, exactly preserved. Indeed, the scheme is written in the form

$$\begin{pmatrix} 1 & 0 \\ \Delta t & 1 \end{pmatrix} \Theta^{n+1} = \begin{pmatrix} 1 & \Delta t \\ 0 & 1 \end{pmatrix} \Theta^n \quad [1.35]$$

---

<sup>11</sup> We let  $\mathcal{O}(h)$  denote a remainder term that can be dominated above by  $h$ .

or, developing  $\xi^{n+1} = \xi^n - \Delta t(\theta^n + \Delta t\xi^n) = (1 - \Delta t^2)\xi^n - \Delta t\theta^n$ ,

$$\Theta^{n+1} = S\Theta^n, \quad S = \begin{pmatrix} 1 & \Delta t \\ -\Delta t & 1 - \Delta t^2 \end{pmatrix}. \quad [1.36]$$

An interesting fact is that the matrix  $S$  satisfies  $S^\top \mathbb{J} S = \mathbb{J}$ : we say that it is a *symplectic matrix* and that the scheme [1.36] is symplectic. Moreover, with [1.35], we compute

$$\begin{aligned} \begin{pmatrix} 1 & 0 \\ \Delta t & 1 \end{pmatrix} \Theta^{n+1} \cdot \Theta^{n+1} &= \begin{pmatrix} 1 & \Delta t \\ 0 & 1 \end{pmatrix} \Theta^n \cdot \Theta^{n+1} = \Theta^n \cdot \begin{pmatrix} 1 & 0 \\ \Delta t & 1 \end{pmatrix} \Theta^{n+1} \\ &= \Theta^n \cdot \begin{pmatrix} 1 & \Delta t \\ 0 & 1 \end{pmatrix} \Theta^n = \begin{pmatrix} 1 & 0 \\ \Delta t & 1 \end{pmatrix} \Theta^n \cdot \Theta^n \\ &= |\theta^n|^2 + |\xi^n|^2 + \Delta t\theta^n\xi^n = \tilde{H}(\theta^n, \xi^n) \\ &= |\theta^{n+1}|^2 + |\xi^{n+1}|^2 + \Delta t\theta^{n+1}\xi^{n+1} = \tilde{H}(\theta^{n+1}, \xi^{n+1}). \end{aligned}$$

Therefore, the analysis of numerical schemes for Hamiltonian systems relies on an adequate understanding of the properties of symplectic matrices.

### 1.3.2. Symplectic matrices; symplectic schemes

We begin by presenting antisymmetric bilinear forms and matrices.

**DEFINITION 1.7.**— Let  $f : \mathbb{C}^N \times \mathbb{C}^N \rightarrow \mathbb{C}$  be a bilinear form. We say that

- $f$  is *alternating* if for all  $x \in \mathbb{C}^N$ ,  $f(x, x) = 0$ ;
- $f$  is *antisymmetric* if for all  $x, y \in \mathbb{C}^N$ ,  $f(x, y) = -f(y, x)$ ;
- an antisymmetric form  $f$  is *non-degenerate* if for all  $x \in \mathbb{C}^N \setminus \{0\}$ , the set  $\{y \in \mathbb{C}^N, f(x, y) = 0\}$  is not all  $\mathbb{C}^N$ .

**LEMMA 1.12.**— An alternating form is antisymmetric.

**PROOF.**— We compute  $f(x+y, x+y) = 0 = f(x, x) + f(y, y) + f(x, y) + f(y, x) = 0 + f(x, y) + f(y, x) = 0$ .  $\square$

**PROPOSITION 1.7.**— Let  $A \in \mathcal{M}_N(\mathbb{C})$  be an antisymmetric matrix:  $A^\top = -A$ . We can associate it with the alternating form

$$f(x, y) = \sum_{j,k=1}^N A_{j,k} x_k y_j.$$

Reciprocally, given an alternating form and a basis  $(a_1, \dots, a_N)$ , we associate them with an antisymmetric matrix  $A \in \mathcal{M}_N(\mathbb{C})$  that represents the form in that basis with the relation

$$A_{j,k} = f(a_j, a_k).$$

We say the matrix  $A$  is non-degenerate when the associated form  $f$  is non-degenerate, which is equivalent to saying that 0 is not an eigenvalue of  $A$ .

If  $(b_1, \dots, b_N)$  is another basis of  $\mathbb{C}^N$ , and  $B$  is the matrix associated with  $f$  in that basis, then there exists an invertible matrix  $P$ , such that  $B = PAP^\top$ .

PROOF.– Let  $P$  be the change-of-basis matrix, such that  $b_j = \sum_{k=1}^N P_{j,k} a_k$ . A vector  $x \in \mathbb{C}^N$  is expressed in the form

$$\begin{aligned} x &= \sum_{k=1}^N \alpha_k a_k = \sum_{j=1}^N \beta_j b_j \\ &= \sum_{k,j=1}^N \beta_j P_{j,k} a_k. \end{aligned}$$

We therefore have  $\alpha_k = \sum_{j=1}^N P_{j,k} \beta_j$ . So

$$f(x, x') = \sum_{j,k=1}^N A_{j,k} \alpha_k \alpha'_j = \sum_{j,k,\ell,m=1}^N A_{j,k} P_{\ell,k} \beta_\ell P_{m,j} \beta'_m = \sum_{\ell,m=1}^N B_{m,\ell} \beta_\ell \beta'_m.$$

We infer that

$$B_{m,\ell} = \sum_{j,k=1}^N A_{j,k} P_{\ell,k} P_{m,j}$$

that is to say,  $B = PAP^\top$ . □

LEMMA 1.13.– Let  $A \in \mathcal{M}_N(\mathbb{C})$  be an antisymmetric matrix.

- i) If  $N = 2p + 1$  is odd, then  $\det(A) = 0$ ;
- ii) If  $N = 2p$  is even and the coefficients of  $A$  are real, then  $\det(A) \geq 0$ , and  $\det(A) > 0$  when  $A$  is non-degenerate;
- iii) If  $N = 2p$  is even, then  $\det(A) \geq 0$  is written as the square of a homogeneous polynomial of degree  $p = N/2$  in the coefficients of  $A$ , the *Pfaffian* of  $A$ .

PROOF.– If  $A$  is antisymmetric, we have  $\det(A) = \det(A^\top) = \det(-A) = (-1)^N \det(A)$ . When  $N$  is odd, this relation implies that  $\det(A) = 0$ . When  $N = 2p$  and  $A$  is real, its eigenvalues are purely imaginary and two-by-two conjugates. Since  $\lambda\bar{\lambda} = |\lambda|^2 \geq 0$ , we obtain  $\det(A) = \prod_{\lambda \in \sigma(A)} \lambda \geq 0$ . If  $A$  is non-degenerate, then 0 is not in its spectrum and therefore  $\det(A) > 0$ . Clearly, the subtle point here is to generalize the statement and to justify iii). This will involve showing some intermediate results.  $\square$

LEMMA 1.14.– Let  $A \in \mathcal{M}_{2p}(\mathbb{C})$  be an antisymmetric matrix. If  $A$  is invertible, then there exists an invertible matrix  $P$ , such that

$$A = P^\top \begin{pmatrix} \mathbb{J} & 0 & \cdots & 0 \\ 0 & -\mathbb{J} & & \\ \vdots & & \ddots & \\ 0 & & & 0 \\ 0 & \cdots & 0 & \mathbb{J} \end{pmatrix} P.$$

More generally, if  $A \in \mathcal{M}_N(\mathbb{C})$  is an antisymmetric matrix of rank  $2p$ , then there exists an invertible matrix  $P$ , such that

$$A = P^\top \begin{pmatrix} \mathbb{J} & 0 & \cdots & 0 \\ 0 & -\mathbb{J} & & \\ \vdots & & \ddots & \\ 0 & & & 0 \\ 0 & \cdots & 0 & \mathbb{J} \\ \hline & 0 & & 0 \\ & & 0 & 0 \end{pmatrix} P$$

where the set of non-zero blocks forms a sub-matrix of size  $2p \times 2p$ .

PROOF.– If  $A$  is exactly zero, then the result is clearly satisfied. We therefore assume that  $A \neq 0$  and consider the associated alternating form on  $\mathbb{R}^N \times \mathbb{R}^N$ . We will see that by the change of basis, we can express  $f$  using the matrix  $\mathbb{J}$ . The form  $f$  is not identically zero, so there exist  $v, w \in \mathbb{R}^N$ , such that  $f(v, w) \neq 0$ . We can choose

these vectors in such a way that  $f(v, w) = -f(w, v) = 1$ . Let  $x = \alpha v + \beta w$ , with  $\alpha, \beta \in \mathbb{C}$ . We obtain

$$f(x, v) = f(\alpha v + \beta w, v) = \beta f(w, v) = -\beta,$$

$$f(x, w) = f(\alpha v + \beta w, w) = \alpha f(v, w) = \alpha.$$

In particular, we infer that  $\{v, w\}$  are linearly independent: if  $x = \alpha v + \beta w = 0$ , then  $f(x, v) = f(0, v) = 0 = f(0, w) = f(x, w) = \alpha = \beta$ .

We write  $E_1 = \text{Span}\{v, w\}$  and  $Y_1 = \{y \in \mathbb{C}^N, f(y, x) = 0 \forall x \in E_1\}$ . The restriction  $f|_{E_1 \times E_1}$  is associated, relative to the basis  $(v, w)$ , with the matrix

$$\begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix}.$$

We will show that  $E_1 \oplus Y_1 = \mathbb{C}^N$ . Let  $y \in E_1 \cap Y_1$ . Thus, on the one hand, we have  $f(y, v) = f(y, w) = 0$ , since  $y \in Y_1$ , and on the other hand, since  $y \in E_1$ ,  $y = f(x, w)v - f(x, v)w$ . We conclude that  $y = 0$ , so  $E_1 \cap Y_1 = \{0\}$ . Let  $u \in \mathbb{C}^N$ . We write  $x = f(u, w)v - f(u, v)w \in E_1$  and let  $y = u - x$ . Note that  $f(y, w) = f(u, w) - f(x, v) = f(u, w) - f(u, w)f(v, w) = f(u, w) - f(u, w) = 0$  and  $f(y, v) = f(u, v) - f(x, v) = f(u, v) + f(u, w)f(w, v) = f(u, w) - f(u, w) = 0$ , that is to say,  $y \in Y_1$ . This proves that any vector in  $\mathbb{C}^N$  can be uniquely decomposed in the form of the sum of an element from  $E_1$  and an element from  $Y_1$ . Let  $(y_1, \dots, y_{N-2})$  be a basis of  $Y_1$ . The alternating form  $f$  can be expressed in the basis  $(v, w, y_1, \dots, y_{N-2})$  using the matrix

$$\left( \begin{array}{cc|cccc} 0 & 1 & 0 & \dots & 0 \\ -1 & 0 & 0 & \dots & 0 \\ \hline 0 & & A_1 & & & \end{array} \right)$$

where  $A_1$  is an antisymmetric matrix of  $\mathcal{M}_{N-2}(\mathbb{C})$ . We distinguish between two situations. Either  $A_1 = 0$  and the proof is finished, or  $A_1 \neq 0$  represents an alternating bilinear form  $f_1 = f|_{Y_1 \times Y_1}$  on a space of dimension  $N - 2$  to which we can meaningfully reapply these arguments.

If the rank of  $A$  is  $N = 2p$ , then  $A$  is invertible, as well as  $A_1$ . The construction can be repeated  $p$  times and allows us to decompose  $\mathbb{C}^N = E_1 \oplus E_2 \oplus \dots \oplus E_p$ , with

$E_k = \text{Span}\{v_k, w_k\}$ ,  $f(v_k, w_k) = -f(w_k, v_k) = 1$ . In the basis  $(v_1, w_1, \dots, v_p, w_p)$ , the matrix associated with  $f$  is indeed

$$\begin{pmatrix} 0 & 1 \\ -1 & 0 \\ & \ddots \\ & & 0 & 1 \\ & & -1 & 0 \\ & & & \ddots \\ & & & & 0 & 1 \\ & & & & -1 & 0 \end{pmatrix}.$$

If  $\text{rang}(A) = 2p < N$ , then the construction stops after  $p$  steps and gives way to a decomposition of the form  $\mathbb{C}^N = E_1 \oplus \dots \oplus E_p \oplus \tilde{Y}$ , with  $E_k = \text{Span}\{v_k, w_k\}$ ,  $f(v_k, w_k) = -f(w_k, v_k) = 1$  and  $f|_{\tilde{Y} \times \tilde{Y}} = 0$ . We therefore construct a basis of  $\mathbb{C}^N$  in which  $f$  is expressed with the matrix

$$\begin{pmatrix} \mathbb{J} & 0 & \cdots & 0 \\ 0 & \ddots & & 0 \\ \vdots & & \ddots & 0 \\ 0 & \cdots & 0 & \mathbb{J} \\ \hline & 0 & & 0 \\ & & \ddots & 0 \\ & & & 0 \end{pmatrix}, \quad \mathbb{J} = \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix}.$$

It is important to recall that  $P$  is not necessarily an orthogonal matrix: in general,  $P^{-1} \neq P^\top$ .  $\square$

Let  $A \in \mathcal{M}_{2p}(\mathbb{C})$  be an antisymmetric, invertible matrix. Let  $\mathcal{J}$  denote the matrix of  $\mathcal{M}_{2p}(\mathbb{C})$  whose diagonal blocks are defined by the matrix  $\mathbb{J}$ . We write  $A = P^\top \mathcal{J} P$ . It follows that  $\det(A) = (\det(P))^2 \det(\mathcal{J}) = (\det(P))^2$ . Therefore, the determinant of  $A$  can be written as the square of a function of the coefficients of  $A$ , this function being the determinant of the change-of-basis matrix  $P$ . We will see that this function depends polynomially on the coefficients of the matrix  $A$ .

In order to demonstrate this property, we return to this demonstration from a slightly different perspective. We let  $\mathcal{A} = \{a_{i,j}, j > i\}$  denote the set of  $N(N-1)/2$  coefficients that determine the matrix  $A$  (since it is antisymmetric). The determinant  $\det(A)$  is a polynomial with integer coefficients in the elements of  $\mathcal{A}$ :

$\det(A) \in \mathbb{Z}[\mathcal{A}]$ . Therefore, we can interpret the matrix  $A$  as a matrix of coefficients in the field  $\mathbb{K}$  of rational fractions with rational coefficients, whose indeterminate element is given by the elements of  $\mathcal{A}$ . The resulting matrix  $P$  is an element of  $\mathrm{GL}(\mathbb{K})$ ; its coefficients are elements of  $\mathbb{K}$ . The determinant  $\det(P)$  is a rational fraction  $p/q$ , where  $p$  and  $q$  are polynomials in the elements of  $\mathcal{A}$  with rational coefficients. Finally, we will use the following statement.

LEMMA 1.15.– Let  $\mathcal{H}$  be a field. Let  $m$  be a polynomial on  $\mathcal{H}$ , such that  $m$  can be written as the square of a rational function  $r$ . Then,  $r$  is, in fact, a polynomial.

PROOF.– We begin by studying polynomials and rational fractions with one indeterminate element. The proof takes its inspiration from the demonstration of Gauss's lemma in arithmetic. We write  $r(t) = \frac{p(t)}{q(t)}$ , where  $p$  and  $q$  have no common roots, and  $q$  is unitary. Then, Bezout's theorem (see [ARN 87, Th. VII.2.4] or [PER 96, Th. 3.25 & Cor. 3.26]) ensures the existence of two polynomials,  $u$  and  $v$ , such that  $p(t)u(t) + q(t)v(t) = 1$ . It follows that  $p(t) = u(t)(p(t))^2 + v(t)q(t)p(t) = u(t)m(t)(q(t))^2 + v(t)p(t)q(t) = q(t)(u(t)m(t)q(t) + v(t)p(t))$ , which proves that  $q$  divides  $p$ . Since  $p$  and  $q$  have no common roots, it follows that  $q(t) = 1$ , so  $r = p$  is a polynomial.

Given this result in one variable, we generalize it for  $M > 1$  variables. Indeed, considering the fixed variables  $m_j$  for  $j \neq i$ , we apply this reasoning to the function  $\phi_i : \mu \mapsto r(m_1, \dots, m_{i-1}, \mu, m_{i+1}, \dots, m_M)$ : it is a polynomial in the variable  $\mu$ . In particular, there exists an integer  $N_i$ , such that  $\frac{d^{N_i}}{d\mu^{N_i}}\phi_i = 0 = \partial_{m_i}^{N_i}r = 0$ . This conclusion is valid for all  $i \in \{1, \dots, M\}$ . We infer that  $r$  is a polynomial in the  $M$  variables  $m_1, \dots, m_M$ .  $\square$

THEOREM 1.15.– Let  $N = 2p$  be an even integer. For any antisymmetric matrix  $A \in \mathcal{M}_N(\mathbb{C})$ , there exists a unique polynomial  $\mathrm{Pf}$  with integer coefficients, such that  $\det(A) = (\mathrm{Pf}(A))^2$  and, in the case  $A = \mathbb{J}$ , we have  $\mathrm{Pf}(\mathbb{J}) = 1$ . Moreover, for any matrix  $M \in \mathcal{M}_N(\mathbb{C})$ , we have  $\mathrm{Pf}(M^\top A M) = \mathrm{Pf}(A) \det(M)$ .

PROOF.– If  $A$  is antisymmetric, then for any matrix  $M \in \mathcal{M}_N(\mathbb{C})$ , the matrix  $M^\top A M$  is also antisymmetric because  $(M^\top A M)^\top = M^\top A^\top M^{\top\top} = -M^\top A M$ . We can therefore evaluate the Pfaffian of  $M^\top A M$  to obtain  $\det(M^\top A M) = (\det(M))^2 \det(A) = (\det(M))^2 (\mathrm{Pf}(A))^2 = (\mathrm{Pf}(M^\top A M))^2$ . It follows that  $\mathrm{Pf}(M^\top A M) = \pm \det(M) \mathrm{Pf}(A)$ . In particular, this relation is satisfied for  $M = \mathbb{I}$ , which leads to  $\mathrm{Pf}(M^\top A M) = +\det(M) \mathrm{Pf}(A)$ .  $\square$

DEFINITION 1.8.– We say that a real matrix  $S \in \mathcal{M}_{2d}$  is *symplectic* if it satisfies

$$S^\top \mathbb{J} S = \mathbb{J}. \quad [1.37]$$

PROPOSITION 1.8.– The determinant of a symplectic matrix is equal to 1.

This follows from the definition stating that a symplectic matrix  $S$  satisfies  $\det(S^\top \mathbb{J} S) = \det(\mathbb{J}) = \det(\mathbb{J})[\det(S)]^2$  so  $[\det(S)]^2 = 1$ . This implies that  $S$  is invertible. It remains to show that the determinant cannot be negative. Several approaches are possible, which are all equally interesting. We can provide an argument based on the Pfaffian. Indeed, we have  $\text{Pf}(S^\top \mathbb{J} S) = +\det(S)\text{Pf}(\mathbb{J}) = \det(S) = \text{Pf}(\mathbb{J}) = 1$ . A different approach involves analyzing the spectral properties of symplectic matrices.

**LEMMA 1.16.**— Let  $S$  be a symplectic matrix. Then,

- i)  $S^\top$  is also symplectic;
- ii)  $S$  and  $S^{-1} = -\mathbb{J} S^\top \mathbb{J}$  are similar.

**PROOF.**— We invert the relation [1.37]

$$\mathbb{J}^{-1} = -\mathbb{J} = (S^\top \mathbb{J} S)^{-1} = S^{-1} \mathbb{J}^{-1} (S^{-1})^\top = -S^{-1} \mathbb{J} (S^{-1})^\top.$$

It follows that

$$S \mathbb{J} S^\top = S (S^{-1} \mathbb{J} (S^{-1})^\top) S^\top = \mathbb{J},$$

which proves that  $S^\top$  is symplectic. However, we also have

$$S^{-1} = \mathbb{J}^{-1} \mathbb{J} S^{-1} = \mathbb{J}^{-1} (S^\top \mathbb{J} S) S^{-1} = \mathbb{J}^{-1} S^\top \mathbb{J}.$$

So  $S^{-1}$  is similar to  $S^\top$ . The following general statement allows us to conclude ii).  $\square$

**LEMMA 1.17.**— Let  $M \in \mathcal{M}_N(\mathbb{C})$ . Then,  $M$  and  $M^\top$  are similar: there exists an invertible matrix  $P \in \mathcal{M}_N(\mathbb{C})$ , such that  $M = P M^\top P^{-1}$ . The result is also true for  $M \in \mathcal{M}_N(\mathbb{R})$ , with a real-coefficient change-of-basis matrix  $P$ .

**PROOF.**— The proof uses the Jordan form: for  $M \in \mathcal{M}_N(\mathbb{C})$ , there exists an invertible matrix  $Q \in \mathcal{M}_N(\mathbb{C})$ , such that  $M = Q \mathcal{J}_M Q^{-1}$ , where  $Q$  is written in blocks

$$Q = \begin{pmatrix} \mathcal{J}_1 & 0 & \cdots & 0 \\ 0 & \mathcal{J}_2 & & 0 \\ \vdots & & \ddots & \vdots \\ 0 & \cdots & 0 & \mathcal{J}_r \end{pmatrix}$$

with  $\mathcal{J}_k$  of the form

$$\mathcal{J}_k = \begin{pmatrix} \lambda & 1 & 0 & \cdots & 0 \\ 0 & \ddots & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & \lambda & 1 \end{pmatrix}$$

where  $\lambda$  is an eigenvalue of  $M$ .

The transposition operation is equivalent to multiplying by the matrix

$$\mathbb{T} = \begin{pmatrix} 0 & \cdots & 0 & 1 \\ \vdots & \ddots & \vdots & 0 \\ 0 & \cdots & 0 & 0 \\ 1 & 0 & \cdots & 0 \end{pmatrix}$$

that is to say,  $M^\top = \mathbb{T} M \mathbb{T}$ . Note that  $\mathbb{T}^{-1} = \mathbb{T}$ . Therefore,  $M^\top = (Q^{-1})^\top \mathcal{J}_M^\top Q^\top = (Q^{-1})^\top \mathbb{T} \mathcal{J}_M \mathbb{T} Q = (Q^{-1})^\top \mathbb{T} Q^{-1} M Q \mathbb{T} Q^\top$ . We write  $P = Q \mathbb{T} Q^\top$  in such a way that  $M^\top = P^{-1} M P$ .

This result applies when the matrix  $M$  has real coefficients, but the resulting matrix  $P$  can have complex coefficients. We can still transform  $P$  in order to construct a real-valued change-of-basis matrix. Indeed, let  $A$  and  $B$  be two real matrices, such that  $A = PBP^{-1}$ , with  $P \in \mathcal{M}_N(\mathbb{C})$ . We decompose  $P = P_R + iP_I$ , where  $P_R$  and  $P_I$  have real coefficients. We have  $AP = PB$ . Therefore, by decomposing into real and complex parts,  $AP_R = P_RB$  and  $AP_I = BP_I$ . We seek to construct a new change-of-basis matrix  $\tilde{P}$  as a linear combination of  $P_I$  and  $P_R$ . We introduce the function  $f : z \in \mathbb{C} \mapsto f(z) = \det(P_R + zP_I)$ . This function is a polynomial with real coefficients, which is not identical to zero because  $f(i) = \det(P) \neq 0$ . There exists a real number  $t$ , such that  $f(t) \neq 0$ . We write  $\tilde{P} = P_R + tP_I$ . This matrix with real coefficients is invertible and satisfies  $A\tilde{P} = \tilde{P}B$ .  $\square$

**NOTE.** – Note that two matrices with the same characteristic polynomial and the same minimal polynomial are not necessarily similar. Indeed, for matrices whose dimension

$N$  is greater than 3, it is possible to find the matrices whose Jordan forms are not the same. For example,

$$A = \begin{pmatrix} \lambda & 1 & 0 & 0 \\ 0 & \lambda & 0 & 0 \\ 0 & 0 & \lambda & 1 \\ 0 & 0 & 0 & \lambda \end{pmatrix} \quad \text{and} \quad B = \begin{pmatrix} \lambda & 0 & 0 & 0 \\ 0 & \lambda & 0 & 0 \\ 0 & 0 & \lambda & 1 \\ 0 & 0 & 0 & \lambda \end{pmatrix}.$$

For the above two matrices, the characteristic polynomial is  $(X - \lambda)^4$  and the minimal polynomial is  $(X - \lambda)^2$ . However,  $\dim(A - \lambda\mathbb{I}) = 2$  and  $\dim(B - \lambda\mathbb{I}) = 3$ . By finding the criteria that allow us to decide whether two matrices are similar, notions of similarity invariants can be introduced (see section 6.3 in [SER 01]).

LEMMA 1.18.– Let  $S$  be a symplectic matrix and  $\lambda$  an eigenvalue of  $S$ . Then,  $\bar{\lambda}$ ,  $1/\lambda$  and  $1/\bar{\lambda}$  are also eigenvalues of  $S$ .

PROOF.– We have seen that a symplectic matrix is invertible, so  $\lambda \neq 0$ . The matrix  $S$  is real, so its complex eigenvalues are two-by-two conjugates. The eigenvalues of  $S^{-1}$  are of the form  $1/\lambda$ , with  $\lambda$  being an eigenvalue of  $S$ . Now,  $S$  and  $S^{-1}$  are similar, so they have the same characteristic polynomial and  $1/\lambda$ , the eigenvalue of  $S^{-1}$ , is also the eigenvalue of  $S$ .  $\square$

COROLLARY 1.9.– Let  $P_S$  be the characteristic polynomial of a symplectic matrix  $S \in \mathcal{M}_{2d}(\mathbb{R})$ . Thus, for all  $\lambda \in \mathbb{C} \setminus \{0\}$ , we have

$$\det(S) P_S(\lambda) = (\lambda)^{2d} P_S(1/\lambda).$$

PROOF.– We consider that  $S$  and  $S^{-1}$  have the same characteristic polynomial:

$$\begin{aligned} P_S(\lambda) &= \det(S - \lambda\mathbb{I}) = \det(S^{-1} - \lambda\mathbb{I}) = \det(S^{-1} - \lambda S^{-1}S) \\ &= \det(S^{-1})\det(\mathbb{I} - \lambda S) \\ &= \frac{1}{\det(S)} \det\left(\lambda\left(\frac{1}{\lambda}\mathbb{I} - S\right)\right) = \frac{1}{\det(S)} \lambda^{2d} \det\left(\frac{1}{\lambda}\mathbb{I} - S\right) \\ &= \frac{1}{\det(S)} \lambda^{2d} P_S(1/\lambda). \quad \square \end{aligned}$$

For the analysis, we introduce the following vector subspaces: for  $\lambda$ , an eigenvalue of  $S$  and  $k \in \mathbb{N} \setminus \{0\}$ , we define

$$N_k(\lambda) = \text{Ker}(S - \lambda\mathbb{I})^k,$$

which are just the generalized eigenspaces of  $S$ . In particular,  $N_k(\lambda) \subset N_{k+1}(\lambda)$  and there exists  $k_\lambda \in \mathbb{N} \setminus \{0\}$ , the order of algebraic multiplicity for the eigenvalue  $\lambda$ , such that  $N(\lambda) = \bigcup_k N_k(\lambda) = N_{k_\lambda}(\lambda)$ . For any eigenvalue  $\lambda$  of  $S$ ,  $1/\lambda$  is an eigenvalue of

$S^{-1}$ , and we note  $\tilde{N}_k(\lambda) = \text{Ker}(S^{-1} - \frac{1}{\lambda}\mathbb{I})^k$  as well as  $\tilde{N}(\lambda) = \bigcup_k \tilde{N}_k(\lambda)$ . Finally, we write

$$f(x, y) = \mathbb{J}x \cdot y = \sum_{k, \ell} \mathbb{J}_{k, \ell} x_\ell y_k.$$

LEMMA 1.19.– For all eigenvalues  $\lambda$  and  $\mu$  of  $S$ , such that  $\mu \neq 1/\lambda$ , we have  $f(N_1(\lambda), N_1(\mu)) = 0$ .

PROOF.– Let  $x \in N_1(\lambda)$  and  $y \in N_1(\mu)$ :  $Sx = \lambda x$  and  $Sy = \mu y$ . We therefore also have  $S^{-1}x = \frac{1}{\lambda}x$ . Considering that  $S$  is symplectic, the computation of the bilinear form  $f$  leads to

$$\mu f(x, y) = f(x, \mu y) = f(x, Sy) = S^\top \mathbb{J}x \cdot y = S^\top J \frac{Sx}{\lambda} \cdot y = \frac{1}{\lambda} \mathbb{J}x \cdot y = \frac{1}{\lambda} f(x, y).$$

In other words,  $(\mu - \frac{1}{\lambda})f(x, y) = 0$ . Since  $\mu \neq \frac{1}{\lambda}$ , it follows that  $f(x, y) = 0$ .  $\square$

LEMMA 1.20.– For each eigenvalue  $\lambda$  of  $S$  and  $k \in \mathbb{N} \setminus \{0\}$ , we have  $N_k(\lambda) = \tilde{N}_k(\lambda)$ .

PROOF.– If  $x$  is an eigenvector of  $S$  for the eigenvalue  $\lambda$ , then  $Sx = \lambda x$  also leads to  $S^{-1}x = \frac{1}{\lambda}x$ . We infer that  $N_1(\lambda) \subset \tilde{N}_1(\lambda)$ . However,  $S$  and  $S^{-1}$  have symmetric roles. Therefore, we will also show the opposite inclusion. We conclude that  $N_1(\lambda) = \tilde{N}_1(\lambda)$ .

Suppose that the result is obtained for an integer  $k$ . We consider an  $x \in N_{k+1}(\lambda)$  in such a way that  $(S - \lambda\mathbb{I})^{k+1}x = (S - \lambda\mathbb{I})^k(S - \lambda\mathbb{I})x = 0$ , which shows that  $(S - \lambda\mathbb{I})x \in N_k(\lambda)$ . So, the equality  $N_k(\lambda) = \tilde{N}_k(\lambda)$  leads to  $(S^{-1} - \frac{1}{\lambda}\mathbb{I})^k(S - \lambda\mathbb{I})x = 0 = (S - \lambda\mathbb{I})(S^{-1} - \frac{1}{\lambda}\mathbb{I})^kx$ . Therefore, we have  $(S^{-1} - \frac{1}{\lambda}\mathbb{I})^kx \in N_1(\lambda) = \tilde{N}_1(\lambda)$ , that is to say,  $(S^{-1} - \frac{1}{\lambda}\mathbb{I})(S^{-1} - \frac{1}{\lambda}\mathbb{I})^kx = 0 = (S^{-1} - \frac{1}{\lambda}\mathbb{I})^{k+1}x$ . We have shown that  $N_{k+1}(\lambda) \subset \tilde{N}_{k+1}(\lambda)$ . We show the equality of those sets by noting that  $S$  and  $S^{-1}$  have symmetric roles.  $\square$

COROLLARY 1.10.– For all eigenvalues  $\lambda$  and  $\mu$  of  $S$ , such that  $\mu \neq 1/\lambda$ , we have  $f(N(\lambda), N(\mu)) = 0$ .

PROOF.– We will show this result by induction. Lemma 1.19 gives us the identity  $f(N_k(\lambda), N_k(\mu)) = 0$  for  $k = 1$ . Let us then suppose that  $f(N_k(\lambda), N_k(\mu)) = 0$  and

we will show the heredity of the property. Consider  $x \in N_{k+1}(\lambda)$  and  $y \in N_k(\mu)$ :  $(S - \lambda\mathbb{I})^{k+1}x = 0 = (S - \mu\mathbb{I})^k y$ . We have

$$\begin{aligned} f(x, (S - \mu)^k y) &= 0 = f\left(x, \left(S - \frac{1}{\lambda}\mathbb{I} + \left(\frac{1}{\lambda} - \mu\right)\mathbb{I}\right)y\right) \\ &= \sum_{j=0}^k C_k^j \left(\frac{1}{\lambda} - \mu\right)^{k-j} f\left(x, \left(S - \frac{1}{\lambda}\mathbb{I}\right)^j y\right). \end{aligned}$$

Now, by using the relation  $S^\top \mathbb{J} = \mathbb{J} S^{-1}$ , we show that

$$\begin{aligned} f(x, (S - \frac{1}{\lambda}\mathbb{I})^j y) &= \mathbb{J}x \cdot (S - \frac{1}{\lambda}\mathbb{I})^j y = (S^\top - \frac{1}{\lambda}\mathbb{I})^j \mathbb{J}x \cdot y \\ &= \mathbb{J}(S^{-1} - \frac{1}{\lambda}\mathbb{I})^j x \cdot y = f((S^{-1} - \frac{1}{\lambda}\mathbb{I})^j x, y). \end{aligned}$$

Therefore, we obtain

$$f(x, (S - \mu)^k y) = 0 = \sum_{j=0}^k C_k^j \left(\frac{1}{\lambda} - \mu\right)^{k-j} f\left(\left(S^{-1} - \frac{1}{\lambda}\mathbb{I}\right)^j x, y\right).$$

However, using lemma 1.20,  $x \in N_{k+1}(\lambda) = \tilde{N}_{k+1}(\lambda)$  implies that

$$(S^{-1} - \frac{1}{\lambda}\mathbb{I})^{k+1-j} (S^{-1} - \frac{1}{\lambda}\mathbb{I})^j x = 0.$$

Therefore, for all  $j \in \{1, \dots, k\}$ , we have  $(S^{-1} - \frac{1}{\lambda}\mathbb{I})^j x \in \tilde{N}_{k+1-j}(\lambda) \subset \tilde{N}_k(\lambda) = N_k(\lambda)$ . Using the recurrence hypothesis, we obtain

$$f(x, (S - \mu)^k y) = 0 = \left(\frac{1}{\lambda} - \mu\right)^k f(x, y).$$

Since  $m \neq 1/\lambda$ , we conclude that  $f(x, y) = 0$  and, more generally, that  $f(N_{k+1}(\lambda), N_k(\mu)) = 0$ . With this intermediate result obtained, we can repeat the same argument in order to finally show that  $f(N_{k+1}(\lambda), N_{k+1}(\mu)) = 0$ .  $\square$

**COROLLARY 1.11.** – If  $\pm 1$  is an eigenvalue of  $S$ , then  $N(\pm 1)$  has an even dimension and the order of algebraic multiplicity of  $\pm 1$  is even.

**PROOF.** – Jordan's theorem makes it possible to decompose  $\mathbb{R}^{2d}$  into direct sums of generalized eigenspaces, where we distinguish the spaces  $N(\lambda)$ ,  $N(1/\lambda)$  for  $\lambda \neq \pm 1$  and  $N(\pm 1)$ . Let us suppose that  $\pm 1$  is an eigenvalue of  $S$ . Then, the restriction  $g = f|_{N(\pm 1) \times N(\pm 1)}$  defines an alternating bilinear form on  $N(\pm 1)$ . Suppose that  $g$  is

degenerate: there exists an  $x \neq 0$  in  $N(\pm 1)$ , such that  $f(x, N(\pm 1)) = 0$ . However, this implies that  $f$  is degenerate on  $\mathbb{R}^{2d}$  because every vector  $y \in \mathbb{R}^{2d}$  can be written in the form  $y_1 + y_\perp$ , where  $y_1 \in N(\pm 1)$  and  $y_\perp$  belongs to the sum of the other generalized eigenspaces in such a way that, using corollary 1.11,  $f(x, y) = f(x, y_1 + y_\perp) = f(x, y_\perp) = 0$ . This conclusion is absurd, so  $g$  is a non-degenerate alternating bilinear form on  $N(\pm 1)$ . Its associated matrix has a non-zero determinant, which implies that  $\dim(N(\pm 1))$  is even.  $\square$

Finally, it remains to find that  $\det(S)$ , for which we have just established  $|\det(S)| = 1$ , is expressed as the product of the eigenvalues of  $S$ . If  $\lambda \in \mathbb{C} \setminus \{\pm 1\}$  is an eigenvalue, then  $\bar{\lambda}$  is also one, and the product of those eigenvalues  $|\lambda|^2$  is strictly positive. If  $\pm 1$  is an eigenvalue, then its order of multiplicity is even. We conclude that  $\det(S) > 0$  and therefore  $\det(S) = +1$ .

**NOTE.**— The fact that  $S$  and  $S^{-1}$  are similar is not enough to reach a conclusion about the parity of the multiplicity of the eigenvalues  $\pm 1$ , as shown by the example of the matrix  $\text{diag}(1, -1)$

We consider a Hamiltonian problem [1.31], which we seek to approximate using a numerical scheme. The example of the linear pendulum shows us that the explicit Euler scheme

$$q^{n+1} = q^n + \Delta t \nabla_p H(q^n, p^n), \quad p^{n+1} = p^n - \Delta t \nabla_q H(q^n, p^n)$$

will certainly not have the right qualitative properties, even though it will be convergent. This is also true for the implicit Euler scheme

$$q^{n+1} = q^n + \Delta t \nabla_p H(q^{n+1}, p^{n+1}), \quad p^{n+1} = p^n - \Delta t \nabla_q H(q^{n+1}, p^{n+1})$$

for which we face the difficulty of inverting a nonlinear system of equations. In order to construct better performing schemes, we will seek to reproduce the properties of the continuous problem at the discrete level (see proposition 1.5). We therefore write the numerical scheme in an abstract form

$$X = (p, q), \quad X^{n+1} = \Psi_{\Delta t}(X^n).$$

We let  $\nabla_X \Psi_{\Delta t}$  denote the Jacobian matrix of that mapping. The symplectic structure of the problem is preserved if  $\nabla_X \Psi_{\Delta t}$  is a symplectic matrix; that is, if it satisfies

$$\nabla_X \Psi_{\Delta t}^\top \mathbb{J} \nabla_X \Psi_{\Delta t} = \mathbb{J}.$$

In light of proposition 1.8, in this case, we have  $\det(\nabla_X \Psi_{\Delta t}) = 1$ . Such a scheme is called a *symplectic scheme*.

The symplectic Euler scheme, which we have already described for the pendulum problem, takes the general form

$$q^{n+1} = q^n + \Delta t \nabla_p H(q^{n+1}, p^n),$$

$$p^{n+1} = p^n - \Delta t \nabla_q H(q^{n+1}, p^n).$$

We have seen that this scheme is, in fact, explicit in the case where  $H(q, p) = h(p) + \Phi(q)$ . In the general case, for fixed  $(q, p)$  and  $\Delta t > 0$ , we can apply the implicit function theorem 1.4 and define  $(q^*, p^*) = \Psi_{\Delta t}(q, p)$ , the solution of

$$q^* = q + \Delta t \nabla_p H(q^*, p), \quad p^* = p - \Delta t \nabla_q H(q^*, p).$$

We differentiate with respect to the data to obtain the relations

$$\nabla_q q^* = \mathbb{I} - \Delta t D_{pq}^2 H \nabla_q q^*, \quad \nabla_p q^* = -\Delta t (D_{pp}^2 H + D_{pq}^2 H \nabla_p q^*),$$

$$\nabla_q p^* = \Delta t D_{qq}^2 H \nabla_q q^*, \quad \nabla_p p^* = \mathbb{I} + \Delta t (D_{pq}^2 H + D_{qq}^2 H \nabla_q q^*),$$

which we reorganize in the form

$$\nabla_q q^* = (\mathbb{I} + \Delta t D_{pq}^2 H)^{-1} \quad \nabla_p q^* = -\Delta t (\mathbb{I} + \Delta t D_{pq}^2 H)^{-1} D_{pp}^2 H,$$

$$\nabla_q p^* = \Delta t D_{qq}^2 H (\mathbb{I} + \Delta t D_{pq}^2 H)^{-1},$$

$$\nabla_p p^* = \mathbb{I} + \Delta t D_{pq}^2 H - \Delta t^2 D_{qq}^2 H (\mathbb{I} + \Delta t D_{pq}^2 H)^{-1} D_{pp}^2 H.$$

We then calculate

$$\begin{pmatrix} \nabla_p p^* & \nabla_p q^* \\ \nabla_q p^* & \nabla_q q^* \end{pmatrix} \mathbb{J} \begin{pmatrix} \nabla_p p^* & \nabla_q p^* \\ \nabla_p q^* & \nabla_q q^* \end{pmatrix} = \begin{pmatrix} 0 & A \\ -A & 0 \end{pmatrix}$$

with  $A = \nabla_p p^* \nabla_q q^* - \nabla_q p^* \nabla_p q^*$ . By developing this expression, we confirm that  $A = \mathbb{I}$ , which proves that the scheme is symplectic.

**NOTE.–** The symplectic Euler scheme is consistent of order 1. There is a fairly simple variant that makes it possible to reach order 2. It is *Verlet's scheme*, which has the following form: given  $q_n, p_n$ , we write

$$q^* = q^n + \frac{\Delta t}{2} \nabla_p H(q^*, p^n),$$

$$p^{n+1} = p^n - \frac{\Delta t}{2} (\nabla_q H(q^*, p^n) + \nabla_q H(q^*, p^{n+1})),$$

$$q^{n+1} = q^n + \frac{\Delta t}{2} \nabla_p H(q^*, p^{n+1}).$$

In the case where  $H(q, p) = \frac{p^2}{2} + \Phi(q)$ , by rewriting the scheme using only the variable  $q$ , we obtain  $q^{n+1} - 2q^* + q^n = -\frac{h^2}{2} \nabla_q \Phi(q^*)$ , which is indeed a natural discretization of the second-order equation  $\frac{d^2}{dt^2} q = -\nabla_q \Phi(q)$ .

### 1.3.3. Kepler problem

We describe the evolution of two bodies subject to gravity forces, one of which is far more massive than the other, like a planet and its sun. We construct a model in  $\mathbb{R}^3$  where the sun is at the origin and  $t \mapsto q(t) \in \mathbb{R}^3$  denotes the planet's position in the reference frame, while  $t \mapsto p(t) \in \mathbb{R}^3$  denotes its velocity. The gravity force exerted by the sun on the planet is in the planet–sun direction, its intensity is proportional to the product of their masses, and it derives from a potential that is inversely proportional to the distance between them. The proportionality constant  $\mathcal{G}$  is the gravitational constant. We therefore obtain the differential system

$$\frac{d}{dt} q = p, \quad m \frac{d}{dt} p = -mM\mathcal{G} \frac{q}{|q|^3}$$

where  $M$  and  $m$  are the masses of the heavy and light bodies, respectively. This is a Hamiltonian system associated with

$$H(q, p) = \frac{|p|^2}{2} - \frac{\mathcal{G}M}{|q|}.$$

In particular,  $t \mapsto H(q(t), p(t))$  is preserved in time and the flow is symplectic: the Jacobian of the flow satisfies  $\mathcal{J}^\top \mathbb{J} \mathcal{J} = \mathbb{J}$ . For analyzing the trajectories, a brief review of elementary geometry is required.

#### 1.3.3.1. Review of conic sections and vector product

**DEFINITION 1.9.–** In a plane  $\mathcal{P}$ , we consider a line  $\mathcal{D}$  and a point  $F \notin \mathcal{D}$ . Let  $e > 0$ . The *conic section with directrix  $\mathcal{D}$ , focus  $F$  and eccentricity  $e$*  refers to the set of points  $M$  on the plane  $\mathcal{P}$ , such that

$$\text{dist}(M, F) = e \text{ dist}(M, \mathcal{D}).$$

For  $0 \leq e < 1$ , the resulting set is an *ellipse* (a circle whenever  $e = 0$ ); in this case, it is a closed and bounded curve. For  $e > 1$ , the resulting set is a *hyperbola*; for  $e = 1$ , it is a *parabola*.

Let  $u, v, w$  be three vectors in  $\mathbb{R}^3$ . Consider the set  $\mathcal{E} = \{x \in \mathbb{R}^3, x = r_1 u + r_2 v + r_3 w, 0 \leq r_j \leq 1\}$ . We calculate the volume of  $\mathcal{E}$  using a change of variables  $x = (x_1, x_2, x_3) \in \mathcal{E} \mapsto (r_1, r_2, r_3) \in [0, 1]^3$ . In this

case, the change of variables is a simple change of basis. We let  $P$  denote the matrix whose columns give the coordinates of  $u, v, w$  in the canonical basis, respectively. Then, the change of variables is given by

$$\begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = P \begin{pmatrix} r_1 \\ r_2 \\ r_3 \end{pmatrix}$$

and we have (see [GOU 11, Theorem 4.8])

$$\text{vol}(\mathcal{E}) = \int_{\mathcal{E}} dx_1 dx_2 dx_3 = \int_{[0,1]^3} |\det(P)| dr_1 dr_2 dr_3 = |\det(P)|.$$

We write  $\det(P) = \det(u, v, w)$ . For fixed  $u, v, w \mapsto \det(u, v, w)$  is a linear form. Therefore, there exists a unique vector, denoted as  $u \wedge v$  and called the *vector product of  $u$  and  $v$* , such that for all  $w \in \mathbb{R}^3$  we have

$$\det(u, v, w) = (u \wedge v) \cdot w.$$

In particular, we have

$$\det(u, v, u \wedge v) = |u \wedge v|^2,$$

which therefore corresponds to the volume of the domain of  $\mathbb{R}^3$  given by the vectors  $u, v$  and  $u \wedge v$ . Let  $\mathcal{A}$  be the area of the plane domain determined by the vectors  $u$  and  $v$ , that is, the set  $\{r_1 u + r_2 v, 0 \leq r_j \leq 1\}$ . We therefore obtain  $\det(u, v, u \wedge v) = |u \wedge v|^2 = \mathcal{A}|u \wedge v|$  and finally

$$|u \wedge v| = \mathcal{A} = |u| |v| \sin(\theta)$$

where  $\theta$  represents the angle formed by the vectors  $u$  and  $v$ .

Thus, for the Kepler problem, the area of the triangle defined by the vectors  $q(t)$  and  $q(t + \Delta t)$  is

$$\frac{1}{2} |q(t) \wedge q(t + \Delta t)| \simeq \frac{1}{2} |q(t) \wedge (q(t) + \frac{d}{dt} q(t) \Delta t)| = \frac{1}{2} |q(t) \wedge \frac{d}{dt} q(t)| \Delta t,$$

with an error of order  $\mathcal{O}(\Delta t^2)$ . We evaluate the area swept by the vector  $q(t)$  when  $t$  runs through the segment  $[t_0, t_1]$  with a Riemann sum

$$\begin{aligned} \text{Area} &= \lim_{N \rightarrow \infty} \left\{ \frac{t_1 - t_0}{N} \sum_{k=0}^{N-1} \frac{1}{2} \left| q\left(t_0 + k \frac{t_1 - t_0}{N}\right) \wedge \frac{d}{dt} q\left(t_0 + k \frac{t_1 - t_0}{N}\right) \right| \right. \\ &\quad \left. + \mathcal{O}\left(\frac{t_1 - t_0}{N}\right) \right\}. \end{aligned}$$

We infer that this area is, in fact, given by the integral

$$\frac{1}{2} \int_{t_0}^{t_1} \left| q(t) \wedge \frac{d}{dt} q(t) \right| dt.$$

NOTE.– A simple, albeit tedious, calculation allows us to identify the coordinates of  $u \wedge v$  as a function of those of  $u$  and  $v$ :

$$u \wedge v = \begin{pmatrix} u_2 v_3 - u_3 v_2 \\ -u_1 v_3 + u_3 v_1 \\ u_1 v_2 - u_2 v_1 \end{pmatrix}.$$

The analysis of Kepler's equations relies on identifying invariant quantities. We introduce the *kinetic moment*  $\ell(t) = q(t) \wedge \frac{d}{dt} q(t)$ . Next, we define the *Laplace–Runge–Lenz vector*  $A(t) = \frac{1}{\mu} \frac{d}{dt} q(t) \wedge \ell(t) - \frac{q(t)}{|q(t)|}$ . We observe that these vectors remain constant in time.

LEMMA 1.21.– We have  $\frac{d}{dt} \ell(t) = 0 = \frac{d}{dt} A(t)$ . Moreover, we have  $|A| = 1 + 2E \frac{|\ell|^2}{\mu^2}$ , where  $E$  is the energy  $E = \frac{1}{2} |\frac{d}{dt} q(t)|^2 - \frac{\mu}{|q(t)|} = \frac{1}{2} |\frac{d}{dt} q(0)|^2 - \frac{\mu}{|q(0)|}$ .

PROOF.– Indeed, we have

$$\begin{aligned} \frac{d}{dt} \ell(t) &= \frac{d}{dt} q(t) \wedge \frac{d}{dt} q(t) + q(t) \wedge \ddot{q}(t) \\ &= \frac{d}{dt} q(t) \wedge \frac{d}{dt} q(t) - \frac{\mu}{|q(t)|^3} q(t) \wedge q(t) = 0. \end{aligned}$$

This proves that the trajectory remains within a fixed plane and that the area swept by the vector  $q(t)$  between two instants  $t_0$  and  $t_0 + s$  depends only on  $s$ . Next, we use the formula

$$a \wedge (b \wedge c) = a \cdot c b - a \cdot b c$$

to write

$$\begin{aligned} \frac{d}{dt} A(t) &= \frac{1}{\mu} \ddot{q}(t) \wedge \ell - \frac{\frac{d}{dt} q(t)}{|q(t)|} + \frac{q(t) \otimes q(t)}{|q(t)|^3} \frac{d}{dt} q(t) \\ &= -\frac{q(t)}{|q(t)|^3} \wedge (q(t) \wedge \frac{d}{dt} q(t)) + \frac{q(t) \cdot \frac{d}{dt} q(t)}{|q(t)|^3} q(t) = 0. \end{aligned}$$

The norm of that vector is thus evaluated as follows:

$$|A|^2 = \frac{1}{\mu^2} |\frac{d}{dt} q(t) \wedge \ell|^2 + \frac{q(t) \cdot q(t)}{|q(t)|^2} - 2 \frac{q(t)}{|q(t)|} \cdot \frac{1}{\mu} \frac{d}{dt} q(t) \wedge \ell.$$

Now, since  $\frac{d}{dt}q(t)$  and  $\ell = q(t) \wedge \frac{d}{dt}q(t)$  are orthogonal, we have  $|\frac{d}{dt}q(t) \wedge \ell| = |\frac{d}{dt}q(t)||\ell|$  while  $q(t) \cdot (\frac{d}{dt}q(t) \wedge \ell) = (q(t) \wedge \frac{d}{dt}q(t)) \cdot \ell = \ell \cdot \ell = |\ell|^2$ . It follows that

$$\begin{aligned}|A|^2 &= \frac{1}{\mu^2} \left| \frac{d}{dt}q(t) \right|^2 |\ell|^2 + 1 - \frac{2|\ell|^2}{\mu|q(t)|} \\&= 1 + \frac{|\ell|^2}{\mu^2} \left( \left| \frac{d}{dt}q(t) \right|^2 - \frac{2\mu}{|q(t)|} \right) = 1 + 2E \frac{|\ell|^2}{\mu^2},\end{aligned}$$

which is therefore a positive quantity (even if we can have  $E < 0$ ). However, we will distinguish the case where  $E > 0$  and therefore  $|A| > 1$ , and the case in which  $E < 0$  and thus  $|A| < 1$ .  $\square$

We thus calculate  $q(t) \cdot A = \frac{1}{\mu}q(t) \cdot (\frac{d}{dt}q(t) \wedge \ell) - q(t) \cdot \frac{q(t)}{|q(t)|} = \frac{1}{\mu}(q(t) \wedge \frac{d}{dt}q(t)) \cdot \ell - |q(t)| = \frac{|\ell|^2}{\mu} - |q(t)|$ , which leads to the relation

$$|q(t)| = |A| \left( \frac{|\ell|^2}{\mu|A|} - q(t) \cdot \frac{A}{|A|} \right). \quad [1.38]$$

It remains to show that this equality is indeed the equation of a conic section. We introduce an orthonormal vector system  $(O, i, j)$ . We identify  $q(t)$  with the vector  $OQ(t)$ , where  $Q(t)$  denote the position of the planet at the instant  $t$  in that frame, and we have  $|q(t)| = \text{dist}(O, Q(t))$ . Let  $F$  be the point with coordinates  $(\frac{|\ell|^2}{\mu|A|}, 0)$ . We let  $\mathcal{D}$  denote the line that passes through  $F$  whose direction vector is orthogonal to  $A$ . The distance from  $Q(t)$  to  $\mathcal{D}$  is thus given by  $\text{dist}(Q(t), \mathcal{D}) = \min\{|MQ(t)|, M \in \mathcal{D}\} = |OF| + |P(t)Q(t)|$ , where  $P(t)$  is the orthogonal projection of  $Q(t)$  on the line that passes through  $O$ , in a direction orthogonal to  $A$ . Therefore,

$$\text{dist}(Q(t), \mathcal{D}) = \frac{|\ell|^2}{\mu|A|} - OQ(t) \cdot \frac{A}{|A|}.$$

We can also describe  $Q(t)$  using the orthonormal basis for which  $\frac{A}{|A|}$  is the first vector. We write  $x(t) = q(t) \cdot \frac{A}{|A|}$  and let  $y(t)$  denote the second coordinate of  $Q(t)$  in that basis. The relation [1.38] can be written as  $\sqrt{x(t)^2 + y(t)^2} = |A|(K - x(t))$ , with  $K = \frac{|\ell|^2}{\mu|A|} = \frac{\mu}{2E}(|A|^2 - 1)$ . Let  $z = K \frac{|A|^2}{|A|^2 - 1}$ . We write  $X(t) = x(t) - z$  and  $Y(t) = y(t)$ . We obtain

$$(X(t) + z)^2 + Y(t)^2 = |A|^2(K - X(t) - z)^2 = |A|^2(X(t) + z/|A|^2)^2,$$

which leads to

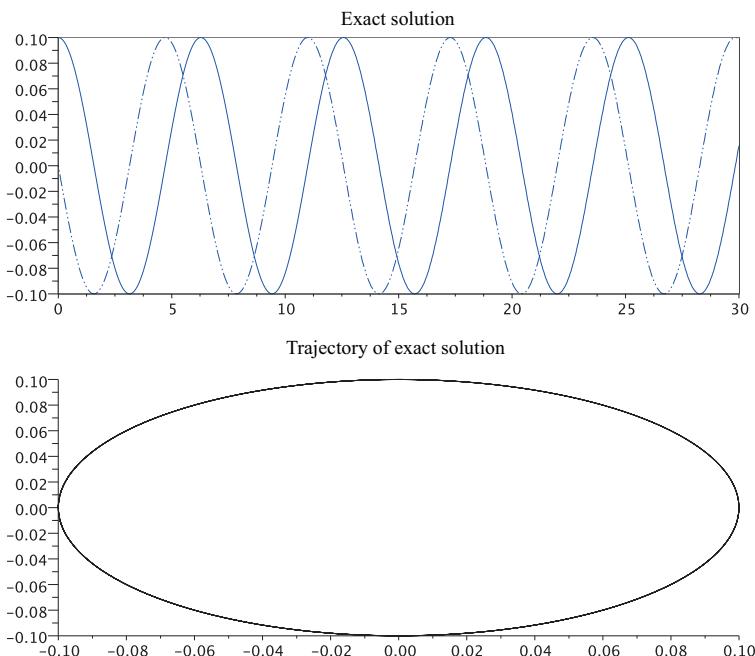
$$X(t)^2 \frac{(1 - |A|^2)^2}{|A|^2 K^2} + Y(t)^2 \frac{1 - |A|^2}{|A|^2 K^2} = 1.$$

Indeed, we find the Cartesian equation for conic sections (ellipse for  $|A| < 1$ ; hyperbola for  $|A| > 1$ ).

### 1.3.4. Numerical results

We compare the performance of different numerical schemes when solving Hamiltonian problems. For this purpose, we use the explicit and implicit Euler schemes, the symplectic Euler scheme, which is a well-adapted variant for Hamiltonian problems, and the second-order Runge–Kutta scheme (RK2).

We begin by considering the linear pendulum [1.34], for which we can compare the solutions that are obtained from these schemes with the exact solution that is known in this case. As observed, the mass  $m$  of the pendulum does not appear in the equation. Moreover, with a change of time scale, we can consider the case of  $g/\ell = 1$ . For these tests, the initial values are  $\theta_{L,\text{Init}} = 0.1$  and  $\xi_{L,\text{Init}} = 0$ . The value of the associated Hamiltonian is thus 0.005. The exact solution is shown in Figure 1.19.



**Figure 1.19.** Exact solution for linear pendulum: Above  $\theta_L$  (bold line) and  $\xi_L$  (dotted line) as a function of time; below, the trajectory in the phase plane

Figure 1.20 shows the comparison of the solutions that are provided by the four schemes until the final time  $T = 30$ , using the same time step  $\Delta t = 0.04$ . The variables are amplified by the explicit Euler scheme, reduced by the implicit Euler scheme; the RK2 and symplectic schemes reproduce the solution fairly accurately. These conclusions are confirmed by the phase portraits shown in Figure 1.21, as well as the evolution of the discrete Hamiltonians in Figure 1.22, which are consistent with the above analysis: an increase in the Hamiltonian with the implicit Euler scheme, a decrease in the Hamiltonian with the explicit Euler scheme, oscillations in the Hamiltonian and preservation of the approximated Hamiltonian with the symplectic scheme; variations observed with the RK2 scheme seem to be completely acceptable. Figure 1.23 shows this comparison more relatively: we bring the simulation up to the final time  $T = 60$  with the time step  $\Delta t = 0.33$ . Hamiltonian's amplification by the RK2 scheme becomes perceptible on a scale of time as long as this one, and the deviations with respect to the exact trajectory become very much visible, even though the method is consistent of order 2. In order to obtain acceptable results, it would be necessary to reduce the time step and therefore to increase the computation cost significantly. Even though it is only of order 1, the symplectic Euler method takes a neat advantage for this kind of “long-term” simulations (note, however, that the trajectory in the phase plane is slightly deformed, with an error of order  $\mathcal{O}(\Delta t)$ ). Finally, Figure 1.25 shows the comparison of the trajectories' evolution for different initial values, which allows us to evaluate variations of volume in the phase space that are produced by the different methods. The final time is  $T = 10$  and the time step is fixed at  $\Delta t = 0.4$ . The initial values are  $(\theta_{\text{Init}}, \xi_{\text{Init}}) = (0.08, -0.02), (0.1, -0.02), (0.12, 0), (0.08, 0), (0.1, 0, 02)$  and  $(0.12, 0.02)$ , respectively. The results can be compared with the exact trajectories shown in Figure 1.24. The Euler schemes produce an expansion or restriction of the volume. We observe that the positions given by the RK2 scheme with that (relatively large) time step correspond poorly with the exact solution.

Finally, as we know the exact solution, the results of the convergence analysis can be verified experimentally. For the same physical data, we test the Euler explicit, implicit, symplectic and RK2 schemes up to the final time  $T = 5$ , with different time steps  $\Delta t = 0.0 \times 2^{-k}$ ,  $k \in \{0, \dots, 8\}$ . We then evaluate  $\ln(\max_n(|\theta_n - \theta(n\Delta t)| + |\xi_n - \xi(n\Delta t)|))$ . Figure 1.26 shows the evolution of those quantities as a function of  $\ln(\Delta t)$ . We can clearly see that the RK2 scheme leads to a greater slope, consistent with its higher convergence order. However, it is, in fact, not necessarily a very good scheme for that problem: adequate results require a rather small time steps, and its behavior for longer times is not satisfying. Although it is only of order 1, the symplectic Euler scheme performs better under these criteria.

We now consider the study of the nonlinear problem [1.33], where we assume that  $g/\ell = 1$ . In view of the results in the linear case, we can expect the implicit Euler scheme not to hold a great advantage over the explicit Euler scheme for this problem. We will therefore compare the explicit schemes that are simpler to implement. The initial values are  $\theta_{\text{Init}} = 1.2$  and  $\xi_{\text{Init}} = 0$ . We perform a first simulation until the final time  $T = 30$ , with intervals  $\Delta t = 0.25$ . Figure 1.27 shows

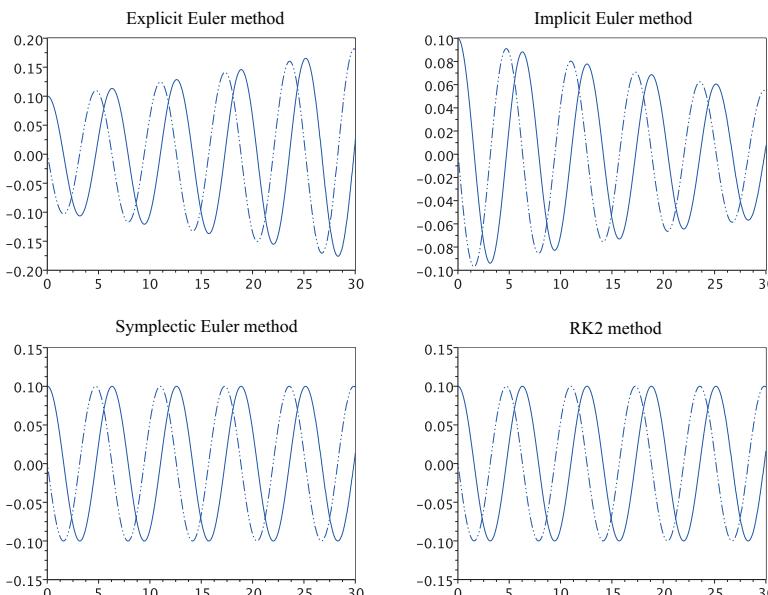
Hamiltonian's evolution with the explicit Euler, symplectic Euler and RK2 schemes. The explicit Euler scheme is immediately discriminated in these conditions, since the Hamiltonian has values that are very far from the theoretical value. The variations for the RK2 scheme are acceptable for this span of time (they are indeed less than  $\Delta t$ ); however, they are strictly increasing and suggest undesirable behavior in longer spans of time. The variations for the symplectic scheme have a greater amplitude, but the discrete Hamiltonian oscillates around the theoretical value. The approximated Hamiltonian is not precisely preserved, unlike in the linear case, but the values that are produced by it oscillate with a very small amplitude (which is shown to remain within  $\mathcal{O}(\Delta t^2)$ ). Figure 1.28 shows the evolution of the numerical solution through time and Figure 1.29 shows the trajectories in the phase plane  $(\theta, \xi)$ : it is clear that the solution produced by the explicit Euler scheme is not acceptable, whereas the solutions for the RK2 and symplectic Euler schemes are very close in those conditions. However, we see that perceivable gaps appear in the phase portrait using the RK2 method. We then test the behavior for the symplectic Euler and RK2 schemes in longer spans of time. The data remain the same, except for the final time and the time step, which are now set to  $T = 140$  and  $\Delta t = 0.48$ . Figure 1.30 shows Hamiltonian's evolution and Figure 1.31 shows the evolution of the solutions as a function of time and phase portraits. On this time scale, the results of the RK2 scheme become incoherent, whereas those of the symplectic scheme remain completely satisfying, as suggested in the linear case. We finish these numerical experimentations with an evaluation of the order of convergence for the different methods. In the nonlinear case, we do not have an explicit expression of the solution to the Cauchy problem, and therefore we do not have access to  $y_n - y(n\Delta t)$ . So we compare the numerical solutions obtained with different time intervals. More specifically, given  $\Delta t_0 > 0$ , we produce simulations with steps of the form  $\Delta t_k = 2^{-k}\Delta t_0$ , where  $k$  varies. With a small abuse of notation, we write  $y_{\Delta t_k}$  and  $y_{\Delta t_{k+1}}$ , the discrete solutions corresponding to consecutive exponents, which we define on the same discretization grid (either by not taking all points in the fine-grained grid, or by interpolating the coarse grid to the fine-grained grid). We thus evaluate  $y_{\Delta t_{k+1}} - y_{\Delta t_k}$  and report the values obtained as a function of  $k$ . Figure 1.32 shows the error curves with the final time  $T = 50$ , a first coarse grid defined by  $\Delta t_0 = 1/10$  and  $k \in \{0, \dots, 6\}$ . We observe that the RK2 scheme is of order 2 and the explicit and symplectic Euler schemes are of order 1. However, the higher order is not enough to make RK2 a “good” method for the problem, unless we consider only short time spans and work with a relatively small time step.

Finally, we study the Kepler system numerically. Again, we compare only the explicit Euler, symplectic Euler and RK2 schemes. The data are such that  $\mathcal{G}M = 1$ ,  $q_{\text{init}} = (0.65, 0)$ ,  $p_{\text{Init}} = (0, 0.7)$ . The simulation is performed until  $T = 15$  with the step  $\Delta t = 0.01$ . The position  $q(t)$  is shown in Figure 1.33, while Figure 1.34 shows Hamiltonian's evolution in time. As can be expected, the explicit Euler scheme very quickly provides unacceptable results: the numerical trajectory immediately departs from the expected ellipse and the Hamiltonian of the initial value. The RK2 scheme

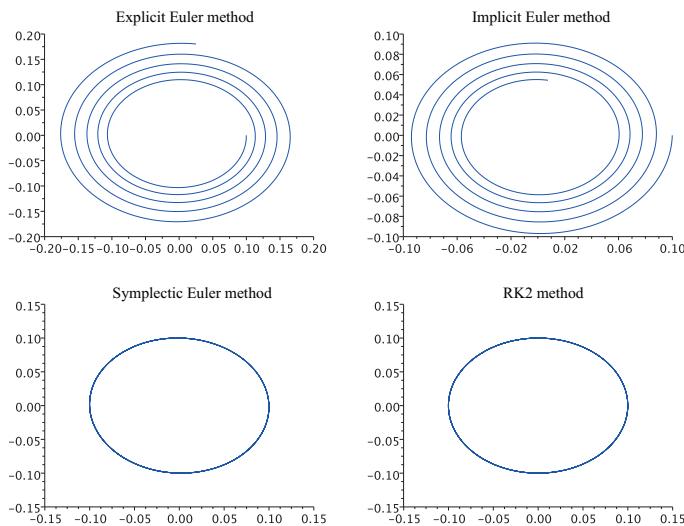
provides relevant results only for short time spans. The Hamiltonian computed with the symplectic scheme oscillates around the theoretical value, and the trajectory remains within a bounded domain, relatively close to the ellipse predicted by the theory. We note that the symplectic scheme also preserves the kinetic moment. By writing  $\ell = q \wedge p$ , we have

$$\begin{aligned}\ell^{n+1} &= (q^n + \Delta t p^n) \wedge \left( p^n - \Delta t \mathcal{G}M \frac{q^{n+1}}{|q^{n+1}|^3} \right) \ell^n \\ &\quad - \Delta t \mathcal{G}M (q^n + \Delta t p^n) \wedge \frac{q^n - \Delta t p^n}{|q^n - \Delta t p^n|^3} = \ell^n.\end{aligned}$$

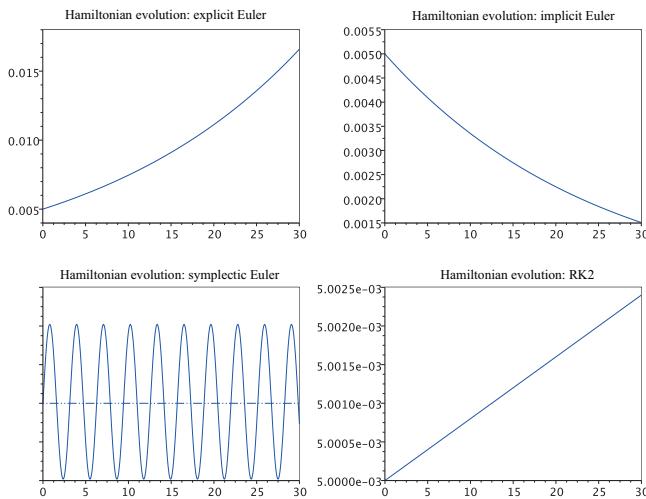
The explicit Euler and RK2 schemes do not preserve this quantity, which can be verified numerically.



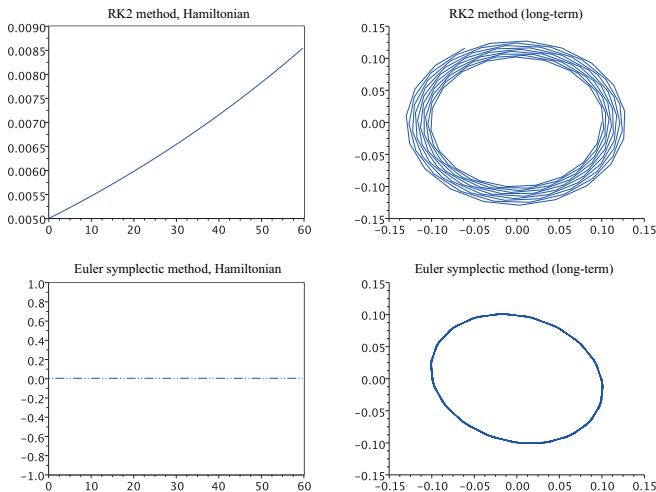
**Figure 1.20.** Comparison of explicit Euler, implicit Euler, symplectic Euler and RK2 numerical solutions for the linear pendulum problem:  $\theta_L$  (bold line) and  $\xi_L$  (dotted line) as a function of time



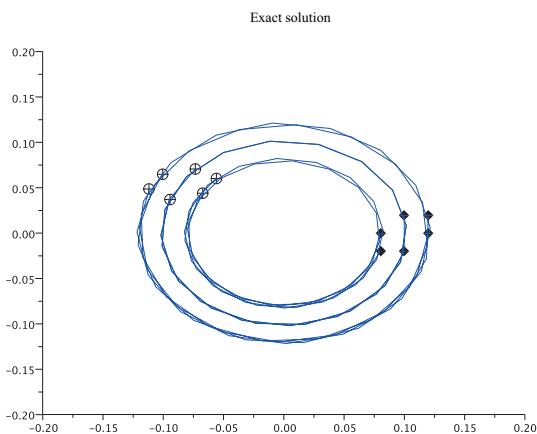
**Figure 1.21.** Comparison of explicit Euler, implicit Euler, symplectic Euler and RK2 numerical solutions for the linear pendulum problem: trajectories in the phase plane



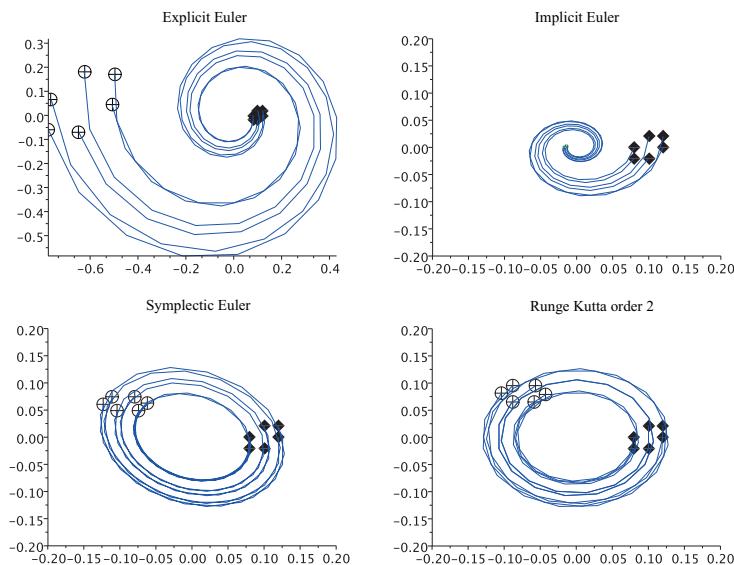
**Figure 1.22.** Comparison of explicit Euler, implicit Euler, symplectic Euler and RK2 numerical solutions for the linear pendulum problem: evolution of the discrete Hamiltonian as a function of time



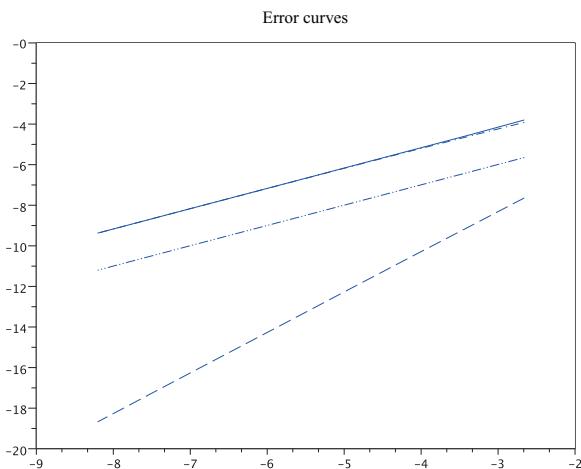
**Figure 1.23.** Long-term linear pendulum simulation: comparison of RK2 and symplectic Euler schemes – Hamiltonian (approximated by the symplectic Euler scheme) and trajectories in the phase plane



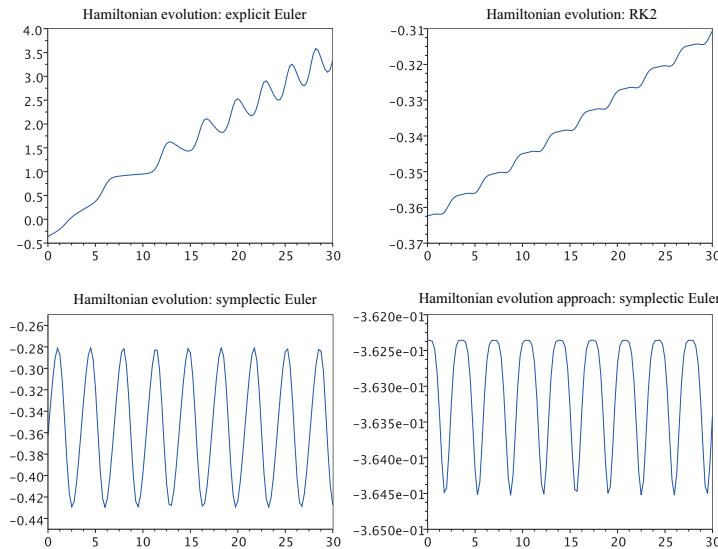
**Figure 1.24.** Linear pendulum: trajectories in the phase plane – exact solution



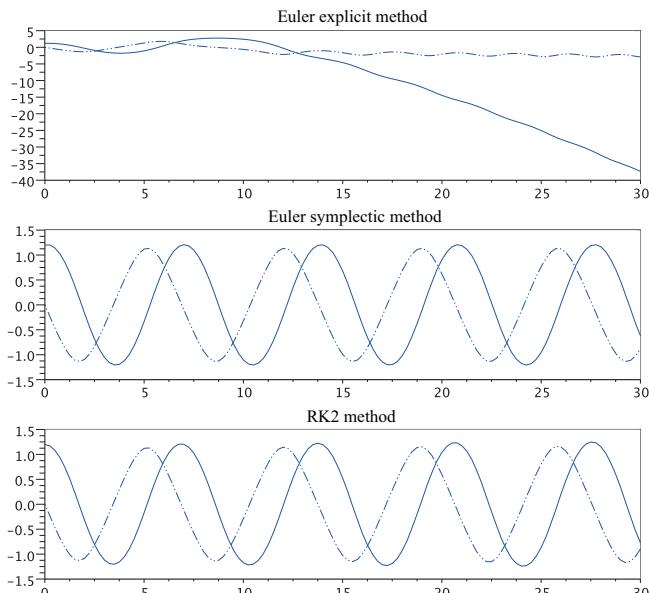
**Figure 1.25.** Linear pendulum: trajectories in the phase plane, comparison of Euler explicit, implicit, symplectic and RK2 schemes. The initial values are indicated by rectangles and the final solutions by circles



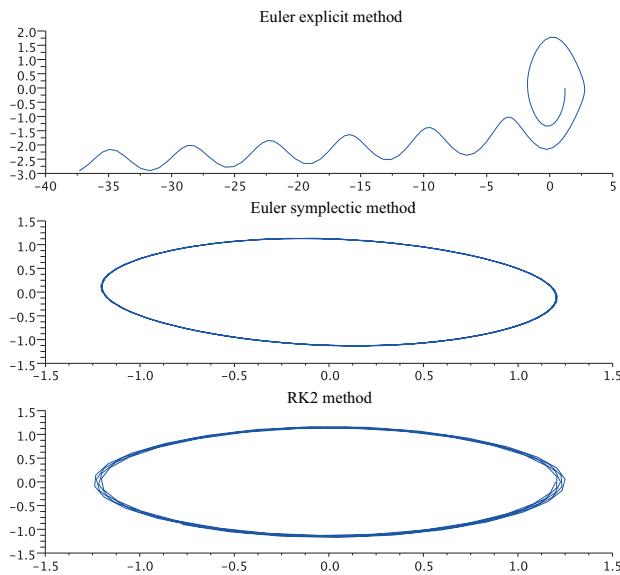
**Figure 1.26.** Linear pendulum: error evaluation (log scale) as a function of  $\ln(\Delta t)$  – explicit Euler (bold line), implicit Euler (dotted), symplectic Euler (dotted-dashed) and RK2 (dashed)



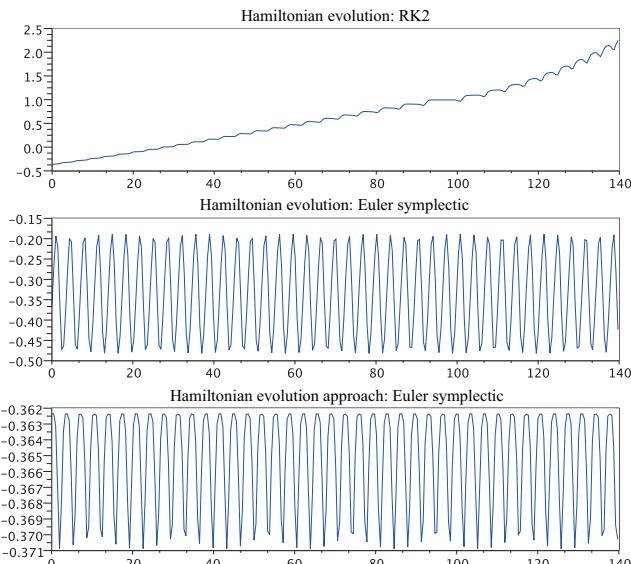
**Figure 1.27.** Nonlinear pendulum: evaluation of the Hamiltonian in time for the explicit Euler, symplectic Euler and RK2 schemes



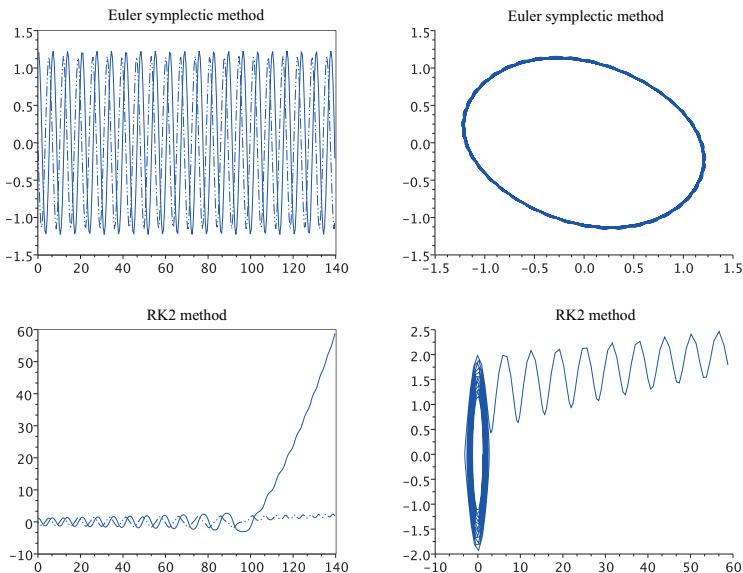
**Figure 1.28.** Comparison of explicit Euler, symplectic Euler and RK2 numerical solutions for the nonlinear pendulum problem:  $\theta_L$  (bold line) and  $\xi_L$  (dotted line) as a function of time



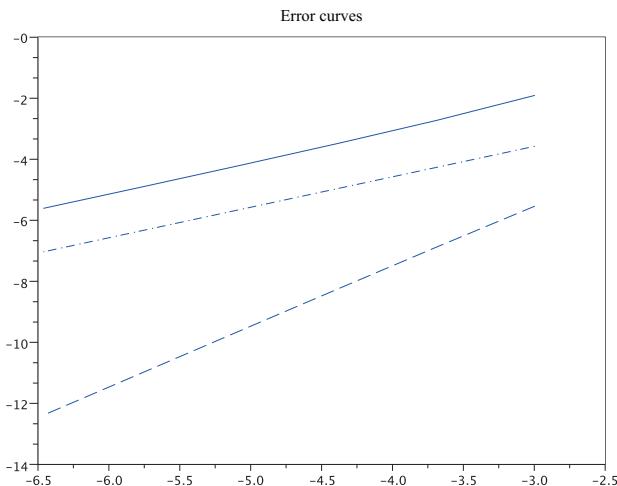
**Figure 1.29.** Comparison of explicit Euler, symplectic Euler and RK2 numerical solutions for the nonlinear pendulum problem:  $\theta_L$  (bold line) and  $\xi_L$  (dotted line) as a function of time



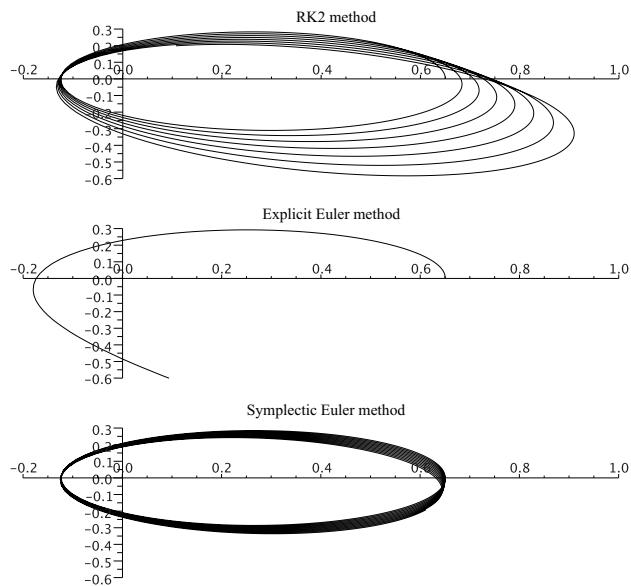
**Figure 1.30.** Simulation of the non linear pendulum for large times: comparison of symplectic Euler and RK2 scheme; evolution of the Hamiltonian and trajectories in the phase plane



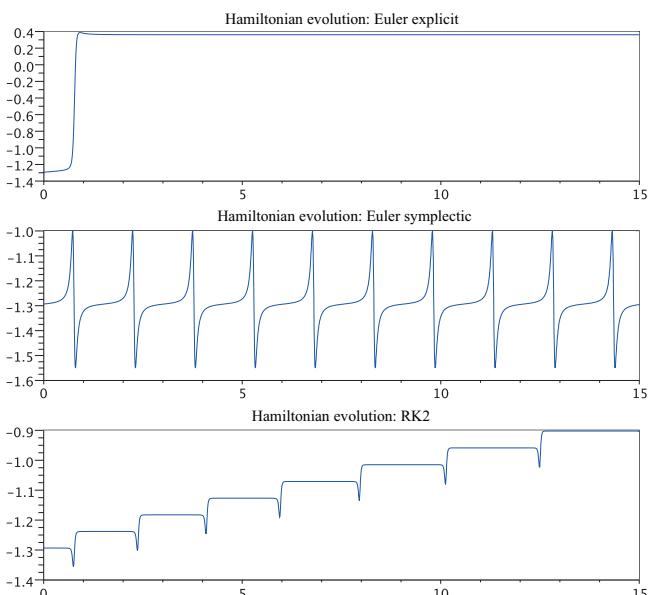
**Figure 1.31.** Simulation of the nonlinear pendulum for long periods of time: comparison of symplectic Euler and RK2 schemes – solution as a function of time and trajectories in the phase plane



**Figure 1.32.** Nonlinear pendulum: error evaluation (log scale) as a function of  $\ln(\Delta t)$ : explicit Euler (bold line), symplectic Euler (dotted-dashed line) and RK2 (dashed)



**Figure 1.33.** Kepler problem: representation of positions



**Figure 1.34.** Kepler problem: evolution of the Hamiltonian in time



---

## Numerical Simulation of Stationary Partial Differential Equations: Elliptic Problems

---

### 2.1. Introduction

In this chapter we will study *stationary* equations; that is, those in which there is only a single space variable  $x \in \Omega \subset \mathbb{R}^D$ , unlike in evolution problems, where the unknown also depends on the time variable  $t \geq 0$ . In what follows, the domain  $\Omega$  is either all of  $\mathbb{R}^D$ , or a regular bounded open set, in the sense that its boundary is given by a regular function, say, of class  $C^\infty$ , for the sake of simplicity. In this case, at each point  $x \in \partial\Omega$ , we can define a vector  $\nu(x)$  that will be normal to  $\partial\Omega$  in  $x$  and will point to the outside of  $\Omega$ . In general, we are interested in partial differential equations<sup>1</sup> of the form

$$P(x, \partial_x)u = f \quad \text{for } x \in \Omega,$$

where  $\partial_x = (\partial_{x_1}, \dots, \partial_{x_D})$  and, for any given  $x, \xi \in \mathbb{R}^D \mapsto P(x, \xi)$  is a polynomial. The problem is completed by any suitable conditions on the boundaries of  $\partial\Omega$ . Following certain structure assumptions on  $P$ , analysis theorems provide the functional framework for which we can justify the existence and uniqueness of solutions to this kind of equation. In some cases, we also have qualitative information: estimates satisfied by the solutions, gain of regularity with respect to the data, etc.. The numerical solutions should be able to reproduce those properties as much as possible. More specifically, we will focus on the following model problem (which is a typical *elliptic problem*):

$$\lambda u - \nabla \cdot (k(x)\nabla u) = f \quad \text{for } x \in \Omega, \tag{2.1}$$

---

<sup>1</sup> In what follows, we indifferently write  $\partial_{x_i}$  or  $\frac{\partial}{\partial x_i}$  for the differentiation operator in the  $i$ th direction.

where  $\lambda \geq 0$ ,  $f$  is given, let us say in  $L^2(\Omega)$ , and the function  $k : x \in \Omega \mapsto k(x) \in \mathcal{M}_N(\mathbb{R})$  is such that there exist  $\kappa, K > 0$  satisfying and for all  $x \in \Omega, \xi \in \mathbb{R}^N$

$$k(x)\xi \cdot \xi \geq \kappa|\xi|^2, \quad |k(x)\xi| \leq K|\xi|.$$

When  $\lambda = 0$ , equation [2.1] is known as the *Poisson equation* (which is non-homogeneous when the coefficient  $k$  varies as a function of the position  $x$ ). We use the following differential operators:

– *gradient*: for a function  $u : x \in \mathbb{R}^D \mapsto u(x) \in \mathbb{R}$ ,  $\nabla u(x)$  is the vector of  $\mathbb{R}^D$  whose components are the partial derivatives  $\frac{\partial u}{\partial x_j}(x)$ ;

– *divergence*: for a function  $V : x \in \mathbb{R}^D \mapsto V(x) = (V_1(x), \dots, V_D(x)) \in \mathbb{R}^D$ ,  $\nabla \cdot V(x)$  is the scalar  $\sum_{j=1}^D \frac{\partial}{\partial x_j} V_j(x)$ . (This is the trace of the Jacobian matrix of  $V$  evaluated at  $x$ .)

In components, equation [2.1] is written as  $\lambda u - \sum_{i,j=1}^D \frac{\partial}{\partial x_i} (k_{ij}(x) \frac{\partial}{\partial x_j} u) = f$ . An interesting particular case assumes  $k(x) = k\mathbb{I}$ , with  $k > 0$ ; then the equation becomes

$$\lambda u - k\Delta u = f$$

where  $\Delta$  represents the (*Laplace*) differential operator

$$\Delta u(x) = \left( \frac{\partial^2 u}{\partial x_1^2} + \dots + \frac{\partial^2 u}{\partial x_D^2} \right)(x).$$

The problem is completed by prescribing conditions on the boundary  $\partial\Omega$ . We consider the specific cases of:

– (homogeneous) Dirichlet conditions:  $u|_{\partial\Omega} = 0$ ;

– (homogeneous) Neumann conditions:  $k\partial_\nu u|_{\partial\Omega} = k\nabla u \cdot \nu|_{\partial\Omega} = 0$ , where we recall that  $\nu(x)$  denotes the unitary outward vector normal to  $\partial\Omega$  at the point  $x \in \partial\Omega$ .

The direct use of general functional analysis theorems together with the set up of an appropriate framework ensures the existence and uniqueness of the solution of [2.1]. More precisely, the Lax–Milgram theorem (see [GOU 11, corollary 5.27]) guarantees that, for all  $\lambda \geq 0$  with Dirichlet conditions, or for all  $\lambda > 0$  with Neumann conditions, [2.1] has a unique solution  $u \in L^2(\Omega)$ , with  $\nabla_x u \in L^2(\Omega)$ . Additionally, if the data, coefficients and second members, are more regular,  $k \in C^m(\overline{\Omega})$ , and  $f$  is such that for each  $D$ -uplet  $\alpha = (\alpha_1, \dots, \alpha^D) \in \mathbb{N}^D$  with  $|\alpha| = \alpha_1 + \dots + \alpha_D \leq m$ ,  $\partial^\alpha f = \partial_{x_1}^{\alpha_1} \dots \partial_{x_D}^{\alpha_D} f \in L^2(\Omega)$ , then the solution  $u$  satisfies  $\partial^\beta u \in L^2(\Omega)$  for each  $D$ -uplet of length  $|\beta| \leq m + 2$  (see [EVA 98, Chapter 6.3]). Finally, we can

demonstrate the following *maximum principle*: if  $0 \leq f(x) \leq M$  for almost all  $x \in \Omega$ , then  $0 \leq u(x) \leq M/\lambda$  for almost all  $x \in \Omega$ .

The numerical methods used to approximate the solutions to this kind of problem seek to construct a sequence  $(u_h)_{h>0}$ , where the parameter  $h > 0$  measures the approximation “effort”. The somewhat vague notion of “effort” is understood as the computation burden: the smaller  $h$  is, the greater the number of unknowns that must be processed, and the greater the memory and time required to perform the computations. The elements of this sequence  $(u_h)_{h>0}$  are vectors for which the number of components depends on  $h$ , and those components allow us to define a function meant to approximate the solution  $x \mapsto u(x)$  of the continuous problem:

- using a certain procedure, which can be computed effectively, we define a vector  $u_h \in \mathbb{R}^{N_h}$ , with  $\lim_{h \rightarrow 0} N_h = \infty$ ;
- that vector’s components make it possible to define a function  $x \in \Omega \rightarrow \tilde{u}_h(x)$ ;
- we seek to show that, for a certain norm  $\|\cdot\|$ , we have  $\lim_{h \rightarrow 0} \|u - \tilde{u}_h\| = 0$ , where  $u$  is the solution of [2.1].

Here we consider functional spaces of infinite dimension. Our choice of norm is therefore not trivial: convergence can be determined for some norms, but not for others, which implies, for example, greater or lesser regularity properties (convergence in the norm  $L^2$  or  $L^1$ , for example, is less “demanding” than convergence in the norm  $L^\infty$ ). We can vaguely distinguish between three families of numerical methods for these problems, which are based on different ways of seeing and interpreting the equation.

– *Finite difference* (FD) methods are “cross-country” methods, which are the easiest to understand and implement, at first glance. We introduce a subdivision of the domain  $\Omega$  in the form of cubes: the interval  $h > 0$  measures how fine-grained this subdivision will be. We denote the nodes of this discretization as  $x_i$ , so the numerical unknown  $u_i$  can be seen as an approximation of the value  $u(x_i)$  of the solution at the point  $x_i$ . The derivatives are approximated with differential quotients. The scheme takes the general form  $\sum_{j=1}^{N_h} \omega_{ij} u_j = f_i$ , where  $f_i = f(x_i)$ . (Note that this assumes there is sufficient regularity on  $f$ , for example  $f$  is continuous.) The weights  $\omega_{ij}$  are constructed in such a way that the consistency error tends to 0 when the grid becomes more fine-grained<sup>2</sup>:  $\lim_{h \rightarrow 0} (\sum_{j=1}^{N_h} \omega_{ij} u(x_j) - f(x_i)) = 0$ .

– *Finite volume* (FV) methods are based on the integration of the equation on control volumes that partition the domain  $\Omega$ . These control volumes can, in principle, have any shape (although the analysis of the schemes often imposes restrictions on

---

<sup>2</sup> It is important to note that  $u_i \neq u(x_i)$ : the first does satisfy  $\sum_{j=1}^{N_h} \omega_{ij} u_j - f_i = 0$ , but, in general, the evaluations of the unknown  $u$  at the discretization points are such that  $\sum_{j=1}^{N_h} \omega_{ij} u(x_j) - f_i \neq 0$ . This distinction will be discussed later.

those shapes). This approach works on equations with *conservative* form, where the differential term is expressed as  $\nabla \cdot (\dots)$ . Here, the numerical unknown  $u_i$  can be interpreted as an approximation of the average of the solution  $u$  on the  $i$ th control volume. The key notion for FV methods is that of *numerical flux*, which defines a relevant approximation of the values of  $(\dots) \cdot \nu$  at the control volume interfaces.

– *Finite element* (FE) methods use a *variational structure*, which is not suitable to handle all problems (but which works for equations of the form [2.1]). Typically, the solution of the partial differential equation can be reinterpreted as “finding  $u \in H$ , a certain Hilbert space, that satisfies the relation  $a(u, v) = \ell(v)$  for all  $v \in H$ ”, where  $a$  and  $\ell$  are a bilinear form and a linear form, respectively, both continuous on  $H$ . The task involves the construction of a space  $H_h$  of finite dimension  $N_h$ , an “approximation” of the original space  $H$ . We will solve the problem  $a(\tilde{u}_h, v_h) = \ell(v_h)$  for all  $v_h \in H_h$ , since it is a finite-dimensional problem that has a simple matrix expression. For example, we can construct the space  $H_h$  by restricting ourselves to a finite number of Fourier modes (spectral methods) or by considering spaces of piecewise polynomial functions. Here, the convergence evaluation of the method depends on the structure of the form  $a$  and the construction of the approximation space  $H_h$ .

In fact, the differences between these methods tend to be smoothed out in modern presentations. The most advanced methods are often “hybrid”, their principles and interpretations can combine several points of view (discontinuous Galerkin (DG) methods [HES 08], residual distribution (RD) methods [ABG 10, ABG 09], etc.). In what follows, it will be necessary to consider the regularity assumptions imposed (sometimes implicitly) on the solution to the original PDE in order to provide the order of convergence for certain numerical methods: in several cases, consistency analysis assumes far more regularity on the solution  $x \mapsto u(x)$  than that which is “naturally” available.

### 2.1.1. The 1D model problem; elements of modeling and analysis

We will focus on mono-dimensional scenarios,  $\Omega = ]0, 1[$ , and [2.1] with Dirichlet conditions becomes

$$\lambda u - \frac{d}{dx} \left( k \frac{d}{dx} u \right) = f \quad \text{for } x \in ]0, 1[, \quad u(0) = 0 = u(1) \quad [2.2]$$

with  $k : ]0, 1[ \rightarrow \mathbb{R}$  satisfying  $0 < \kappa \leq k(x) \leq K < \infty$ , for all  $x \in ]0, 1[$ .

Whenever  $\lambda = 0$ , the problem can be triggered by the description of heat fluxes within a rod, represented by the segment  $]0, 1[$ . The ends of the bar are kept at a constant temperature, and the units are chosen so that that reference temperature is 0. The second member of the equation  $f$  describes the heat sources on the rod and

$x \mapsto k(x) > 0$  is the thermal conductivity coefficient, which depends on the position on the rod if it is composed of a heterogeneous material. If we let  $q(x)$  denote the heat flux in a section of the abscissa  $x \in ]0, 1[$ , then between two positions  $x$  and  $x + dx$ , the difference in heat fluxes is balanced by the sources

$$q(x + dx) - q(x) = \int_x^{x+dx} f(y) dy.$$

We describe this relation with  $dx > 0$  and then we let  $dx$  tend to 0 to obtain

$$\lim_{dx \rightarrow 0} \frac{q(x + dx) - q(x)}{dx} = \frac{d}{dx} q(x) = \lim_{dx \rightarrow 0} \frac{1}{dx} \int_x^{x+dx} f(y) dy = f(x)$$

by the definition of the derivative and using the Lebesgue theorem<sup>3</sup>. Equation [2.2] can therefore be obtained using the Fourier law that connects the heat flux with the temperature

$$q(x) = -k(x) \frac{d}{dx} u(x).$$

We note that the heat flux is opposite to the variations in temperature: if the temperature is increasing, the heat flux is negative and vice versa. As a result, heat fluxes will heat the cold areas and cool the hot ones.

Before detailing the analysis of equation [2.2], let us make some simple remarks:

- If  $\lambda > 0$ , we can always go back to the case of  $\lambda = 1$  by noting that  $v = \lambda u$  satisfies  $v - (\frac{k}{\lambda} v')' = f$ , with  $v(0) = v(1) = 0$ .
- If we impose *heterogeneous* Dirichlet conditions  $u(0) = u_0$  and  $u(1) = u_1$ , we can go to the case of [2.2] by modifying the source term. Indeed, we write  $\varphi(x) = \gamma x + \eta$ , where  $\gamma$  and  $\eta$  are chosen such that  $\varphi(0) = u_0$  and  $\varphi(1) = u_1$ . In other words, we have  $\varphi(x) = (u_1 - u_0)x + u_0$ . Let  $v : x \mapsto v(x) = u(x) - \varphi(x)$ . This function satisfies  $v(0) = 0 = v(1)$  and  $\lambda v - (kv')' = \lambda u - (ku')' - (\lambda \varphi - (k\varphi')') = f - \lambda \varphi + (u_1 - u_0)k'$ .

– It is very important to distinguish between the *Cauchy problem for the ordinary differential equation*

$$\lambda u - \frac{d}{dx} \left( k \frac{d}{dx} u \right) = f, \quad \text{given } u(0) \text{ and } \frac{d}{dx} u(0),$$

---

<sup>3</sup> The result is quite direct if we assume that  $f$  is continuous; if we only assume that  $f$  is integrable, convergence is satisfied for almost all  $x \in ]0, 1[$ , see for example [RUD 87, theorem 7.7].

and the *boundary value problem for the partial differential equation*<sup>4</sup>

$$\lambda u - \frac{d}{dx} \left( k \frac{d}{dx} u \right) = f, \quad \text{given } u(0) \text{ and } u(1).$$

– In the one-dimensional case, we can express the solution directly as a function of the data, integrating equation [2.2]. Consider, for example, a case where  $\lambda = 0$ . We first obtain  $k(x) \frac{d}{dx} u(x) = A - \int_0^x f(z) dz$ , and then the general form

$$u(x) = B + \int_0^x \frac{1}{k(y)} \left( A - \int_0^y f(z) dz \right) dy. \quad [2.3]$$

The constants  $A, B$  are determined by the boundary conditions. If we consider the homogeneous Dirichlet conditions  $u(0) = 0 = u(1)$ , then

$$B = 0, \quad A = \left( \int_0^1 \frac{dy}{k(y)} \right)^{-1} \int_0^1 \frac{1}{k(y)} \left( \int_0^y f(z) dz \right) dy. \quad [2.4]$$

In particular, note that even without assuming the continuity of  $k$  and  $f$ , the solution thus obtained with formula [2.3] is continuous on  $[0, 1]$  (assuming only  $f \in L^1([0, 1])$  and  $k$  measurable, bounded from below by a strictly positive constant). This is a specific property of the 1D case, which will be made explicit in our discussion of lemma 2.1. We can also verify the *maximum principle*: if  $f$  takes non-negative values, then  $u$  is also non-negative on  $[0, 1]$ . Suppose that  $u$ , which is continuous, is negative on  $[a, b] \subset [0, 1]$ , with  $u(a) = u(b) = 0$ . Then, we have

$$\begin{aligned} 0 &\geq \int_a^b f(x)u(x) dx = \int_a^b \frac{d}{dx} \left( \int_0^x f(y) dy - A \right) u(x) dx \\ &\geq - \int_a^b \left( \int_0^x f(y) dy - A \right) \frac{d}{dx} u(x) dx = \int_a^b \frac{1}{k(x)} \left( A - \int_0^x f(y) dy \right)^2 dx \end{aligned}$$

using integration by parts<sup>5</sup>. It follows that the functions  $x \mapsto f(x)u(x)$  and  $x \mapsto A - \int_0^x f(y) dy = k(x) \frac{d}{dx} u(x)$  are exactly zero on  $[a, b]$ . (In fact, the relations that we have just established can be rewritten as  $\int_a^b k(x) \left( \frac{d}{dx} u(x) \right)^2 dx = 0$ , and we will deal with this perspective further below.) This last property implies that  $u$  is constant on the segment  $[a, b]$ , so since  $u(a) = u(b) = 0$ , it is zero on  $[a, b]$ .

– Formulas [2.3]–[2.4] do not, in general, provide an explicit expression of  $u(x)$  as a function of  $x$ , since the computations of the integrals involved in those formulas

<sup>4</sup> Here we have chosen this terminology, although the unknown depends only on the single variable  $x$ .

<sup>5</sup> Integration is justified if  $u$  is  $C^1$ , thanks to the fact that  $k$  is continuous. We will see that the formula remains meaningful for weak solutions lying in the Sobolev space  $H_0^1([0, 1])$ ; that is,  $u$  and  $\frac{d}{dx} u$  have an integrable square.

may not be accomplished with standard integral calculus techniques. One possible approach, for the one-dimensional case, could be to approximate those integrals. We will instead focus on methods that are based on the partial differential equation, which can be extended to the multidimensional case. Nevertheless, in some simple cases, we have an explicit formula for the solution. For instance, for  $k(x) = 1$ ,  $f(x) = 1$ , we find  $u(x) = \frac{x(1-x)}{2}$ .

– Let us finish this list of commentaries by describing a method, for the one-dimensional case, which allows us to also ensure the existence of a solution to the problem

$$-\frac{d}{dx} \left( k(x) \frac{du}{dx} \right) = f(x) \quad [2.5]$$

with the boundary condition

$$u(0) = 0, \quad u(1) = 0, \quad [2.6]$$

and also to calculate an approximation to this solution. This approach provides the opportunity to distinguish between the Cauchy problem and the boundary value problem. We begin by focusing on the *Cauchy problem* where [2.5] is completed by initial data at  $x = 0$  (considered as a “time” variable)

$$u(0) = 0, \quad \frac{d}{dx} u(0) = \alpha.$$

It is a second-order differential equation, which can be rewritten as a first-order system

$$\begin{pmatrix} \frac{d}{dx} & \\ \frac{d}{dx} & \end{pmatrix} \begin{pmatrix} u \\ v \end{pmatrix} = \begin{pmatrix} 0 & 1 \\ 0 & -k'/k \end{pmatrix} \begin{pmatrix} u \\ v \end{pmatrix} + \begin{pmatrix} 0 \\ -f/k \end{pmatrix}$$

assuming here that  $k$  is  $C^1$  and  $f$  is continuous. The Cauchy–Lipschitz theorem ensures the existence and uniqueness of the solution, which we denote as  $x \mapsto u_\alpha(x)$ , of problem [2.5]–[2.6]. As shown above, we have the formula

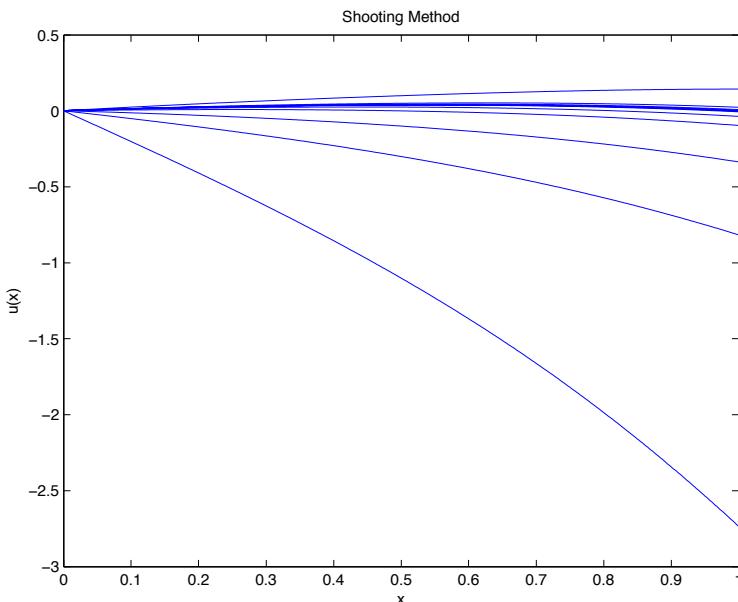
$$u_\alpha(x) = \int_0^x \frac{1}{k(y)} \left( k(0)\alpha - \int_0^y f(z) dz \right) dy. \quad [2.7]$$

In particular, we have

$$u_\alpha(1) = \int_0^1 \frac{1}{k(y)} \left( k(0)\alpha - \int_0^y f(z) dz \right) dy.$$

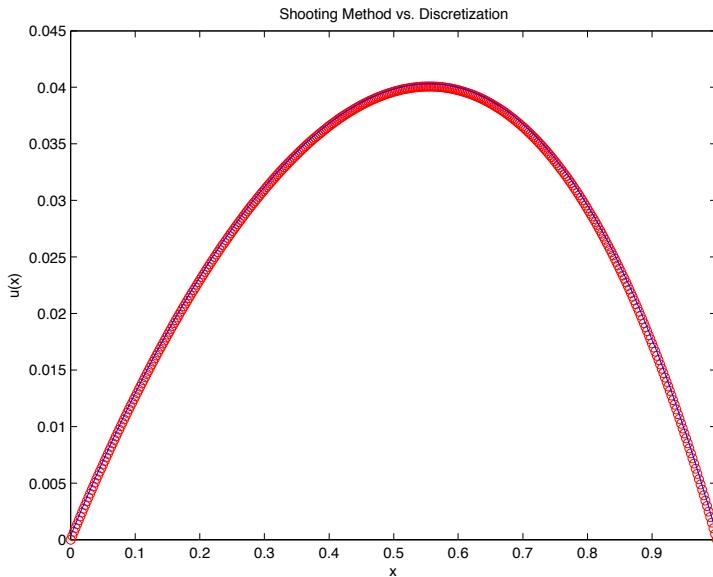
The function  $\alpha \mapsto u_\alpha(1)$  is affine. It follows by the intermediate value theorem that there exists a unique  $\alpha_0 \in \mathbb{R}$ , such that  $u_{\alpha_0}(1) = 0$ .

We thus have a practical procedure, the *shooting method*, to calculate the solution of [2.5]–[2.6]. For a given  $\alpha$ , we solve the Cauchy problem numerically (using a numerical differential equation resolution method or by approximating the integrals [2.7]), and we thus determine  $u_\alpha(1)$ . With values of the parameter  $\alpha > \alpha'$ , such that  $u_\alpha(1) > 0$  and  $u_{\alpha'}(1) < 0$ , we use a dichotomy approach to obtain an approximation of  $\alpha_0$  that guarantees  $u_{\alpha_0}(1) = 0$  (see Figures 2.1 and 2.2). Clearly, for this particular problem, we can directly obtain (an approximation of) the solution by calculating the integrals (or its approximation) involved in [2.3] and [2.4]. A more relevant example is presented in section 2.1.2.



**Figure 2.1.** Shooting method:  $k(x) = 3 + \cos(2.5x)$ ,  $f(x) = 1 + \sin(3x + .5) + .3 \ln(1.7 + x)$ . The first points correspond to  $\alpha = 1$  and  $\alpha = -5$ . Expression [2.7] is evaluated by a simple rectangle method

As mentioned above, the generalization for the multidimensional case involves functional analysis arguments. The key is the fact that the equation has a *variational formulation*. This perspective will also be fruitful for designing numerical schemes. Before considering the general case, we will review some notions in the simple one-dimensional case.



**Figure 2.2.** Comparison of solutions obtained by the shooting method and an approximation using finite differences for the data in Figure 2.1 (400 discretization points): the two curves coincide

The task consists of multiplying [2.2] by a “fairly regular” function  $v$ , such that  $v(0) = v(1) = 0$  and, integrating by parts, we obtain

$$\lambda \int_0^1 u(x)v(x) dx + \int_0^1 k(x) \frac{d}{dx} u(x) \frac{d}{dx} v(x) dx = \int_0^1 f(x)v(x) dx. \quad [2.8]$$

The boundary terms that result from the integration by parts cancel each other out because  $v$  satisfies  $v(0) = v(1) = 0$ . The right-hand term in [2.8] defines, for a given  $f$  in  $L^2([0, 1])$ , a continuous linear form on  $L^2([0, 1])$ :

$$\ell : v \in L^2([0, 1]) \mapsto \int_0^1 f(x)v(x) dx.$$

The left-hand term in [2.8] defines a bilinear form  $a : (u, v) \mapsto a(u, v)$ , which satisfies

- the continuity relation  $|a(u, v)| \leq \lambda \|u\|_{L^2} \|v\|_{L^2} + K \|\frac{d}{dx} u\|_{L^2} \|\frac{d}{dx} v\|_{L^2}$ ;
- the coercivity relation  $a(u, u) = \lambda \int_0^1 |u|^2 dx + \int_0^1 k(x) |\frac{d}{dx} u|^2 dx \geq \lambda \|u\|_{L^2}^2 + \kappa \|\frac{d}{dx} u\|_{L^2}^2$ .

Note that  $a(u, u) \geq 0$  and  $a(u, u)$  cancels out if and only if  $u(x) = 0$  for almost all  $x \in \Omega$ . The last remark applies even in the case of  $\lambda = 0$ : indeed, from  $a(u, u) = 0$ , we infer that  $\frac{d}{dx}u(x) = 0$  is zero for almost all  $x \in ]0, 1[$ , so  $x \mapsto u(x)$  is a constant function. However, the Dirichlet condition  $u(0) = 0 = u(1)$  implies that this constant is 0. The difficulty lies in the construction of an appropriate functional framework to use these observations.

The expressions that we derived could be meaningful if the functions  $u$  and  $v$ , and their derivatives, are square integrable. Note that  $\ell(v)$  and  $a(u, v)$  are well defined, even with discontinuous functions  $f$  and  $k$ . These considerations first lead us to generalize the notion of derivative. If  $u$  is a regular function, say, of class  $C^\infty([0, 1])$ , by integration by parts, we clearly have  $\int_0^1 \frac{d}{dx}u(x)\varphi(x) dx = -\int_0^1 u(x)\frac{d}{dx}\varphi(x) dx$ , for all  $\varphi \in C_c^\infty([0, 1])$ . We use the right-hand expression to define the derivative of a function  $u \in L^2([0, 1])$ , considered as a linear form acting on the functions  $\varphi \in C_c^\infty([0, 1])$ . The functions  $u \in H^1([0, 1])$  are the elements of  $L^2([0, 1])$ , such that there exists a  $C > 0$  satisfying, for all  $\varphi \in C_c^\infty([0, 1])$ , the estimate  $|\int_0^1 u(x)\frac{d}{dx}\varphi(x) dx| \leq C\|\varphi\|_{L^2}$ . By the Riesz's lemma, this estimate means that we can identify the derivative of  $u$  with an element of  $L^2([0, 1])$ . We let  $H^1([0, 1])$  denote the space of functions  $u \in L^2([0, 1])$  satisfying that property. More generally, we can thus define a family of spaces, the *Sobolev spaces*, by requiring integrability properties on the successive derivatives. We define  $H_0^1([0, 1])$  as the complement of  $C_c^\infty([0, 1])$  for the norm

$$\|u\|_{H^1} = \sqrt{\|u\|_{L^2}^2 + \left\| \frac{d}{dx}u \right\|_{L^2}^2}.$$

The space  $H^1([0, 1])$  is completed from  $C^\infty([0, 1])$  for the same norm. The spaces  $H^1([0, 1])$  and  $H_0^1([0, 1])$  along with the norm  $\|\cdot\|_{H^1}$  are Hilbert spaces. One subtle question deals with the meaning given to the traces of those functions at the boundary of the domain (since we recall that the functions in  $L^2$  are only defined almost everywhere). In dimension one, the difficulty is overcome by noting that the elements of that space  $H^1$  are, in fact, continuous functions (or, more precisely, they admit a continuous representative).

**LEMMA 2.1 (Sobolev Embedding Theorem (1D)).**— Let  $u \in H^1([0, 1])$ . Then,  $u \in C^0([0, 1])$ . Moreover, there exists a  $C > 0$ , such that for all  $u \in H^1([0, 1])$ , we have

$$\|u\|_{L^\infty} \leq C\|u\|_{H^1}. \quad [2.9]$$

**LEMMA 2.2 (Poincaré lemma (1D)).**— Let  $u \in H_0^1([0, 1])$ , then we have the following estimate:

$$\|u\|_{L^\infty} \leq \left\| \frac{d}{dx}u \right\|_{L^2}, \quad \|u\|_{L^2} \leq C\left\| \frac{d}{dx}u \right\|_{L^2}.$$

where  $C > 0$  does not depend on  $u$ .

PROOF.– Let  $\varphi \in C^\infty([0, 1])$ . We can write

$$\varphi(x) - \varphi(y) = \int_y^x \frac{d}{dx} \varphi(\xi) d\xi.$$

The Cauchy–Schwarz inequality leads to

$$\begin{aligned} |\varphi(x) - \varphi(y)| &\leq \left| \int_y^x \frac{d}{dx} \varphi(\xi) d\xi \right| \leq \sqrt{|x - y|} \sqrt{\int_0^1 \left| \frac{d}{dx} \varphi(\xi) \right|^2 d\xi} \\ &\leq \sqrt{|x - y|} \|\varphi\|_{H^1}. \end{aligned}$$

By the density of  $C^\infty([0, 1])$  in  $H^1([0, 1])$ , we extend this relation to the functions of  $H^1([0, 1])$ , which proves their continuity:  $H^1([0, 1]) \subset C^0([0, 1])$ . The last estimate implies even a stronger consequence: the embedding of  $H^1([0, 1])$  in  $C^0([0, 1])$  is *compact*. Indeed, if  $(u_n)_{n \in \mathbb{N}}$  is a bounded sequence in  $H^1([0, 1])$ , then there exists  $C > 0$ , such that for all  $n \in \mathbb{N}$ ,  $x, y \in [0, 1]$ , we have  $|u_n(x) - u_n(y)| \leq C\sqrt{|x - y|}$ . The sequence  $(u_n)_{n \in \mathbb{N}}$  is therefore equicontinuous on  $[0, 1]$ . The estimate [2.9] shows that it is also equibounded. Thus, the Arzela–Ascoli theorem [GOU 11, theorem 7.49] implies that  $(u_n)_{n \in \mathbb{N}}$  is compact on  $C^0([0, 1])$ .

To justify [2.9], we will use a duality argument. Let  $\varphi \in C_c^\infty([0, 1])$ . We introduce an auxiliary function  $\theta \in C^\infty([0, 1])$ , such that  $0 \leq \theta(x) \leq 1$  for all  $x \in [0, 1]$ ,  $\theta(x) = 1$  for  $x \in [3/4, 1]$  and  $\theta(x) = 0$  for  $x \in [0, 1/4]$ . We therefore write

$$\psi(x) = \int_0^x \varphi(y) dy - \theta(x) \int_0^1 \varphi(y) dy.$$

We have

$$\psi \in C_c^\infty([0, 1]), \quad |\psi(x)| \leq K \|\varphi\|_{L^1}$$

(for a certain constant  $K > 0$  that depends on the auxiliary function  $\theta$ ). For  $u \in H^1([0, 1])$ , we get

$$\begin{aligned} \left| \int_0^1 u(x) \varphi(x) dx \right| &= \left| \int_0^1 u(x) \frac{d}{dx} \psi(x) dx + \int_0^1 \varphi(y) dy \int_0^1 u(x) \frac{d}{dx} \theta(x) dx \right| \\ &= \left| \int_0^1 \frac{d}{dx} u(x) \psi(x) dx + \int_0^1 \varphi(y) dy \int_0^1 u(x) \frac{d}{dx} \theta(x) dx \right| \\ &\leq (K + \|\theta\|_{H^1}) \|u\|_{H^1} \|\varphi\|_{L^1}. \end{aligned}$$

Thus,  $u$  is equivalent to a continuous linear form on  $L^1(]0, 1[)$  and therefore, by Riesz's theorem [GOU 11, theorem 4.37], to a function in  $L^\infty(]0, 1[)$  whose norm satisfies [2.9].

To demonstrate lemma 2.2, consider  $\varphi \in C_c^\infty(]0, 1[)$ . Since  $\varphi(0) = 0$ , we obtain the estimate  $|\varphi(x)| \leq \|\frac{d}{dx}\varphi\|_{L^2}$ .  $\square$

With this statement, we can return to the continuity and coercivity relations introduced with [2.8]: for all  $\lambda \geq 0$ , including the case where  $\lambda = 0$ , thanks to lemma 2.2, there exist constants  $m(\lambda), M(\lambda) > 0$ , such that for all  $u, v \in H_0^1(]0, 1[)$ , we have

$$|a(u, v)| \leq M(\lambda)\|u\|_{H^1}\|v\|_{H^1}, \quad a(u, u) \geq m(\lambda)\|u\|_{H^1}^2.$$

The Lax–Milgram theorem now applies, and allows us to demonstrate

**THEOREM 2.1.–** For all  $\lambda \geq 0$  and all  $f \in L^2(]0, 1[)$ , problem [2.2] with homogeneous Dirichlet conditions has a unique solution  $u \in H_0^1(]0, 1[)$  in the sense that for all  $v \in H_0^1(]0, 1[)$ , we have  $a(u, v) = \ell(v)$ . Note that

$$\|u\|_{H^1} \leq \frac{1}{m(\lambda)}\|f\|_{L^2}.$$

The estimate can be obtained simply by taking  $v = u$  in the variational formulation, so that  $m(\lambda)\|u\|_1^2 \leq a(u, u) = \ell(u) \leq \|f\|_{L^2}\|u\|_{L^2} \leq \|f\|_{L^2}\|u\|_{H^1}$ . The problem takes a different nature when we work with Neumann conditions  $\frac{d}{dx}u(0) = 0 = \frac{d}{dx}u(1)$ . Assuming that  $u$  is a fairly regular solution of  $\lambda u - \frac{d}{dx}(k \frac{d}{dx}u) = f$  with those boundary conditions, given that the integration by parts is valid, we obtain  $a(u, v) = \ell(v)$  for all  $v \in C^\infty([0, 1])$ , with the same bilinear form  $a$  and the linear form  $\ell$  defined on  $H^1(]0, 1[)$ . Here we do not require the test function  $v$  to cancel out in  $x = 0$  and  $x = 1$ : the boundary terms of the integration by parts cancel out, thanks to the Neumann condition. For  $\lambda > 0$ , we verify that the Lax–Milgram theorem still applies in the space  $H^1(]0, 1[)$ , which allows us to extend theorem 2.1 to this case.

**THEOREM 2.2.–** For all  $\lambda > 0$  and all  $f \in L^2(]0, 1[)$ , problem [2.2] along with homogeneous Neumann conditions has a unique solution  $u \in H^1(]0, 1[)$  in the sense that for all  $v \in H^1(]0, 1[)$ , we have  $a(u, v) = \ell(v)$ .

For  $\lambda = 0$ , we can see that the coerciveness of the form  $a$  fails. In particular, note that  $\|\frac{d}{dx}u\|_{L^2} = 0$  implies that  $x \mapsto u(x)$  is a constant function, but the boundary conditions do not imply that this constant is 0. In fact, the problem  $-(ku')' = f$ , with  $u'(0) = 0 = u'(1)$ , has an infinite number of solutions: if  $u$  is a solution, then for all  $C \in \mathbb{R}$ ,  $x \mapsto u(x) + C$  is also a solution. We also observe that the

equation makes sense only if the right hand side  $f$  satisfies the compatibility condition  $\int_0^1 f(x) dx = 0$ . Indeed, this condition is a consequence of the boundary conditions when the equation is integrated. We encounter an analogous difficulty in the periodic case: the problem  $-(ku')' = f$  with 1-periodic conditions does not have a unique solution in  $L^2_\#([0, 1])$ , where the symbol  $\#$  denotes the condition of 1-periodicity, since the constant functions are solutions to the homogeneous equation, with  $f = 0$ . Moreover, by formally integrating the equation, we realize that the second member must average to zero in order for the problem to make sense. We will therefore look for a solution that also averages to zero. The periodic problem can be understood using Fourier analysis, and we are going to show how those difficulties can be overcome. We introduce the Fourier coefficients, for  $n \in \mathbb{Z}$ ,

$$\hat{u}_n = \int_0^1 u(x) e^{-2i\pi nx} dx,$$

and associate  $u \in L^2_\#$  with the series  $\sum_{n \in \mathbb{Z}} \hat{u}_n e^{2i\pi nx}$  (convergence taken to be in  $L^2$ , see [GOU 11, sections 5.3 & 5.4]). Moreover, its derivative is associated with the series  $\sum_{n \in \mathbb{Z} \setminus \{0\}} 2\pi i n \hat{u}_n e^{2i\pi nx}$ . In this context, we thus define the Sobolev space

$$H_\#^1 = \{u \in L^2_\#, (\hat{u}_n)_{n \in \mathbb{Z}} \in \ell^2 \text{ and } (n\hat{u}_n)_{n \in \mathbb{Z}} \in \ell^2\}$$

which is indeed a Hilbert space for the norm defined by

$$\|u\|_{H^1}^2 = \sum_{n \in \mathbb{Z}} (1 + 4\pi^2 n^2) |\hat{u}_n|^2.$$

We write

$$H_{0,\#}^1 = \{u \in H_\#^1, \hat{u}(0) = 0\}$$

which is a closed subspace of  $H_\#^1$ . We therefore have the following Poincaré inequality, which is analogous to lemma 2.2 in this context

if  $\hat{u}_0 = \int_0^1 u(x) dx = 0$ , then

$$\|u(x)\|_{L^2}^2 = \sum_{n \in \mathbb{Z}} |\hat{u}_n|^2 \leq \sum_{n \in \mathbb{Z} \setminus \{0\}} 4\pi^2 n^2 |\hat{u}_n|^2 = \left\| \frac{d}{dx} u(x) \right\|_{L^2}^2.$$

(For other variants of this kind of statement, the reader may also refer to the applications given in Appendix 3.) This observation allows us to apply the Lax–Milgram theorem in the periodic case.

**THEOREM 2.3.**— For all  $\lambda > 0$  and all  $f \in L^2_{\#}$ , problem [2.2] with periodic conditions has a unique solution  $u \in H^1_{\#}$  in the sense that for all  $v \in H^1_{\#}$ , we have  $a(u, v) = \ell(v)$ . If  $\lambda = 0$  for all  $f \in L^2_{\#}$ , such that  $\hat{f}(0) = 0$ , problem [2.2] with periodic conditions has a unique solution  $u \in H^1_{0,\#}$  in the sense that for all  $v \in H^1_{0,\#}$ , we have  $a(u, v) = \ell(v)$ .

Returning to problem [2.2] with Neumann conditions, we must thus find a technical means to eliminate constant solutions, as accomplished by introducing  $H^1_{0,\#}$  in the periodic case. This involves the introduction of the appropriate quotient spaces and next, it translates into the numerical processing of the equation.

The variational formulation of the Dirichlet problem also allows us to justify the maximum principle.

**PROPOSITION 2.1.**— Let  $\lambda \geq 0$ . Let  $f$  be a measurable function, such that for almost all  $x \in ]0, 1[$ , we have  $0 \leq f(x) \leq M < \infty$ . Then, the solution  $u \in H^1_0(]0, 1[)$  of [2.2] satisfies  $0 \leq u(x) \leq M/\lambda$  for almost all  $x \in ]0, 1[$ .

**PROOF.**— Let  $\Psi : \mathbb{R} \rightarrow \mathbb{R}$  be a  $C^1$  function, such that  $\Psi(0) = 0$  and  $\Psi' \in L^\infty(\mathbb{R})$ . By assuming that

$$\Psi(u) \in H^1_0(]0, 1[), \quad \frac{d}{dx}\Psi(u) = \Psi'(u)\frac{d}{dx}u, \quad [2.10]$$

we may take  $v = \Psi(u)$  in [2.8] to obtain

$$\lambda \int_0^1 \Psi(u)u \, dx + \int_0^1 \Psi'(u) \left| \frac{d}{dx}u \right|^2 \, dx = \int_0^1 f\Psi(u) \, dx.$$

We choose the function  $\Psi$  in such a way that  $\Psi$  is strictly increasing on  $]-\infty, 0[$  and  $\Psi(z) = 0$  for  $z \geq 0$ . Since  $\Psi(u) \leq 0$ ,  $u\Psi(u) \geq 0$  and  $\Psi'(u) \geq 0$ , if  $f$  is positive, then for such a function we obtain

$$0 \leq \lambda \int_0^1 \Psi(u)u \, dx + \int_0^1 \Psi'(u) \left| \frac{d}{dx}u \right|^2 \, dx = \int_0^1 f\Psi(u) \, dx \leq 0.$$

We distinguish between the following two situations:

–  $\lambda > 0$ : in this case,  $u\Psi(u) = 0$  almost everywhere, which implies that  $u(x) \geq 0$  for almost all  $x \in ]0, 1[$ ;

–  $\lambda = 0$ : in this case, we write  $Z(s) = \int_0^s \sqrt{\Psi'(\sigma)} \, d\sigma$ . Note that  $Z(s) = 0$  for all  $s \geq 0$  and  $Z(s) < 0$  for all  $s < 0$ . It follows from these manipulations that  $\Psi'(u(x)) \left| \frac{d}{dx}u(x) \right|^2 = \left| \frac{d}{dx}Z(u(x)) \right|^2$  cancels out for almost all  $x \in ]0, 1[$ . Now,  $Z(u) \in H^1_0(]0, 1[)$  and the Poincaré inequality thus imply that  $Z(u(x)) = 0$ , so  $u(x) \geq 0$ , for almost all  $x \in ]0, 1[$ .

Bounding  $u$  from above is accomplished with the same argument, by considering in this case a function  $\Psi_k$  that cancels out on  $] -\infty, k]$  and is strictly increasing on  $[k, +\infty[$ . We write  $M = \|f\|_\infty$  and note that  $\Psi_{M/\lambda}(0) = 0$ , such that  $\Psi_{M/\lambda}(u) \in H_0^1(]0, 1[)$ . We thus obtain

$$\begin{aligned} & \lambda \int_0^1 \underbrace{\Psi_{M/\lambda}(u)(u - M/\lambda)}_{\geq 0} dx + \int_0^1 \underbrace{\Psi'_{M/\lambda}(u) \left| \frac{d}{dx} u \right|^2}_{\geq 0} dx \\ &= \int_0^1 \underbrace{(f - M)\Psi_{M/\lambda}(u)}_{\leq 0} dx. \end{aligned}$$

We infer that  $x \mapsto \Psi_{M/\lambda}(u(x))(u(x) - M/\lambda)$  cancels out on all  $]0, 1[$ , so that  $u(x) \leq M/\lambda$ . Whenever  $\lambda = 0$ , we can still show that  $\|u\|_\infty \leq C\|f\|_\infty$ , for a certain constant  $C > 0$ , as a result of lemma 2.1 and theorem 2.1.

It remains to show [2.10]. The problem lies in extending the chain rule formula to functions in  $H^1$ . Let us first consider  $\phi \in C_c^\infty(]0, 1[)$ . Then,  $\Psi(\phi) \in C_c^1(]0, 1[)$  satisfies  $\frac{d}{dx}\Psi(\phi) = \Psi'(\phi)\frac{d}{dx}\phi$ . Let  $u \in H_0^1(]0, 1[)$  and  $(\phi_n)_{n \in \mathbb{N}}$  be a sequence of elements of  $C_c^\infty(]0, 1[)$  that converges towards  $u$  in  $H_0^1(]0, 1[)$ . In particular,  $(\phi_n)_{n \in \mathbb{N}}$  converges to  $u$  in  $L^2(]0, 1[)$  and, even if it involves extracting a subsequence, we may assume that  $\lim_{n \rightarrow \infty} \phi_n(x) = u(x)$  for almost all  $x \in ]0, 1[$  and that the functions  $\phi_n$  are dominated by a square-integrable function (see [GOU 11, lemma 3.31]). Since  $|\Psi(\phi_n) - \Psi(u)| \leq \|\Psi'\|_\infty |\phi_n - u|$ , it follows that  $\Psi(\phi_n)$  converges to  $\Psi(u)$  in  $L^2(]0, 1[)$ . Furthermore,  $\Psi'(\phi_n)$  converges almost everywhere to  $\Psi'(\phi)$  and remains uniformly bounded on  $L^\infty(]0, 1[)$ . Finally,  $\frac{d}{dx}\phi_n$  converges to  $\frac{d}{dx}u$  in  $L^2(]0, 1[)$ . It follows that the product  $\Psi'(\phi_n)\frac{d}{dx}\phi_n$  converges to  $\Psi'(u)\frac{d}{dx}u$  in  $L^2(]0, 1[)$ . Let  $\eta \in C_c^\infty(]0, 1[)$ . We have

$$\begin{aligned} - \int_0^1 \Psi(\phi_n) \frac{d}{dx} \eta dx &= \int_0^1 \Psi'(\phi_n) \frac{d}{dx} \phi_n \eta dx \\ &\xrightarrow{n \rightarrow \infty} - \int_0^1 \Psi(u) \frac{d}{dx} \eta dx = \int_0^1 \Psi'(u) \frac{d}{dx} u \eta dx \end{aligned}$$

which means that  $\frac{d}{dx}\Psi(u) = \Psi'(u)\frac{d}{dx}u \in L^2(]0, 1[)$ . □

### 2.1.2. A radiative transfer problem

In order to illustrate the notions introduced, we will consider the following nonlinear problem:

$$\begin{cases} \sigma T^4(x) - \frac{d^2}{dx^2}T(x) = \sigma T_{\text{ref}}^4(x) & \text{for } 0 < x < 1, \\ T(0) = 0 = T(1). \end{cases} \quad [2.11]$$

This equation is also motivated by questions from thermal science. The source  $f = \sigma(T_{\text{ref}}^4 - T^4)$  describes an energy transfer through radiation at temperature  $T$ : this medium emits energy at temperature  $T_{\text{ref}}$  and absorbs ambient energy related to the temperature  $T$ . The nonlinearity is related to the Stefan–Boltzmann emission law. Indeed, a body emits and absorbs energy in the form of electromagnetic radiations, with a frequency distribution that depends on the temperature. More specifically, this energy is emitted or absorbed in the form of a flux of photons whose distribution follows Planck's law

$$I(x, \nu) = \frac{2h}{c^2} \frac{\nu^3}{e^{h\nu/kT(x)} - 1}.$$

Here,  $\nu > 0$  is the frequency of photons emitted or absorbed,  $h$  is Planck's constant,  $k$  is Boltzmann's constant and  $c$  is the speed of light. The total energy is obtained by summing over the frequencies

$$E(x) = \int_0^\infty I(x, \nu) d\nu = \frac{2k^4}{c^2 h^3} T(x)^4 \int_0^\infty \frac{\mu^3}{e^\mu - 1} d\mu.$$

where we have used the change of variables  $\mu = \frac{h\nu}{kT}$ . In order to calculate the integral that appears in this expression, we use the series expansion

$$\frac{1}{e^\mu - 1} = e^{-\mu} \times \frac{1}{1 - e^{-\mu}} = e^{-\mu} \sum_{n=0}^{\infty} e^{-n\mu} = \sum_{n=1}^{\infty} e^{-n\mu}.$$

By Fubini's theorem, and since the expressions are non-negative, we can permute the sum and integral and write it as

$$E(x) = \frac{2k^4}{c^2 h^3} T(x)^4 \times \sum_{n=1}^{\infty} \int_0^\infty \mu^3 e^{-n\mu} d\mu.$$

With three successive integrations by parts, we reach

$$E(x) = \frac{2k^4}{c^2 h^3} T(x)^4 \times \sum_{n=1}^{\infty} \frac{6}{n^3} \int_0^\infty e^{-n\mu} d\mu = \frac{2k^4}{c^2 h^3} T(x)^4 \times 6 \sum_{n=1}^{\infty} \frac{1}{n^4}.$$

We express this last sum by recognizing the Fourier coefficients of the function  $\psi : x \in [0, 1] \mapsto x(1-x)$ , extended on  $\mathbb{R}$  by 1-periodicity. Indeed, we find

$$\widehat{\psi}(0) = \int_0^1 x(1-x) dx = 1/2 - 1/3 = 1/6,$$

and for  $k \neq 0$

$$\begin{aligned}\widehat{\psi}(k) &= \int_0^1 x(1-x)e^{-2i\pi kx} dx = \int_0^1 (1-2x)\frac{e^{-2i\pi kx}}{2i\pi k} dx \\ &= \int_0^1 2\frac{e^{-2i\pi kx}}{-(2i\pi k)^2} dx = \frac{1}{2\pi^2 k^2}.\end{aligned}$$

It follows that

$$\begin{aligned}\|\psi\|_{L^2}^2 &= \int_0^1 x^2(1-x)^2 dx = \frac{1}{30} \\ &= \sum_{k=-\infty}^{+\infty} |\widehat{\psi}(k)|^2 = |\widehat{\psi}(0)|^2 + 2\sum_{k=1}^{+\infty} |\widehat{\psi}(k)|^2 = \frac{1}{36} + \frac{1}{2\pi^4} \sum_{k=1}^{+\infty} \frac{1}{k^4}\end{aligned}$$

and therefore  $\sum_{k=1}^{+\infty} \frac{1}{k^4} = \frac{\pi^4}{90}$ . Finally, we find  $E(x) = \sigma T(x)^4$ , with  $\sigma = \frac{2\pi^4 k^4}{15c^2 h^3}$ , which is the Stefan–Boltzmann constant.

### 2.1.2.1. Uniqueness given existence

We suppose there are two solutions in  $H_0^1([0, 1])$ , which we denote as  $T$  and  $S$ . In particular, these functions are continuous by lemma 2.1. Then, we have

$$\sigma(T^4 - S^4) - \frac{d^2}{dx^2}(T - S) = 0,$$

with  $T(0) = S(0) = 0 = T(1) = S(1)$ . We multiply by  $T - S$  and integrate on  $[0, 1]$  to obtain

$$\sigma \int_0^1 (T^4 - S^4)(T - S) dx + \int_0^1 \left| \frac{d}{dx}(T - S) \right|^2 dx = 0.$$

Now,  $(T^4 - S^4)(T - S)$  is a non-negative quantity. It follows that  $T(x) = S(x)$  almost everywhere on  $[0, 1]$ .

### 2.1.2.2. A priori estimates

To be consistent with the physical interpretation, we expect to have  $T \geq 0$ . In order to ensure this property, we rewrite the equation in the form

$$\sigma f(T) - \frac{d^2}{dx^2}T = \sigma T_{\text{ref}}^4, \quad [2.12]$$

with

$$f(T) = |T|^3 T.$$

We proceed as in the proof of proposition 2.1. If  $T \in H_0^1(]0, 1[)$  is a solution of [2.12], then we have

$$\sigma \int_0^1 f(T) \Psi(T) dx + \int_0^1 \Psi'(T) \left| \frac{d}{dx} T \right|^2 dx = \int_0^1 \sigma T_{\text{ref}}^4 \Psi(T) dx$$

for  $\Psi$  in  $C^1$ , such that  $\Psi(0) = 0$  and  $\Psi'$  is bounded. First, we take  $\Psi$  increasing, such that  $z\Psi(z) > 0$  for all  $z < 0$  and  $\Psi(z) = 0$  if  $z \geq 0$ , we have  $f(T)\Psi(T) \geq 0$  while  $\sigma T_{\text{ref}}^4 \Psi(T) \leq 0$  because  $\sigma T_{\text{ref}}^4 \geq 0$ . It follows that  $T(x) \geq 0$ , so  $T$  indeed satisfies [2.11]. Next, we write  $M = \|T_{\text{ref}}\|_\infty$  and consider  $\Psi_M$ , such that  $\Psi_M(z) = 0$  for  $z \leq M$  and  $\Psi_M(z) > 0$  for  $z > M$ . Therefore,

$$\begin{aligned} \sigma \int_0^1 \underbrace{(f(T) - M^4) \Psi_M(T)}_{\geq 0} dx + \int_0^1 \underbrace{\Psi'_M(T) \left| \frac{d}{dx} (T) \right|^2}_{\geq 0} dx \\ = \sigma \int_0^1 (T_{\text{ref}}^4 - M^4) \Psi_M(T) dx \leq 0, \end{aligned}$$

which allows us to conclude that  $T(x) \leq M$  is satisfied on  $]0, 1[$ .

### 2.1.2.3. Existence of solutions

We construct a solution of [2.11] as a fixed point of the mapping  $\Phi : T \mapsto \mathcal{T}$ , where  $\mathcal{T}$  is a solution to the linear problem

$$\sigma T^3 \mathcal{T} - \frac{d^2}{dx^2} \mathcal{T} = \sigma T_{\text{ref}}^4 \quad \text{for } 0 < x < 1$$

with Dirichlet conditions. The existence of a fixed point is the result of a general statement concerning *compact* mappings, which is detailed in Appendix 4. We begin by recalling the following definition.

**DEFINITION 2.1.–** Let  $E$  and  $F$  be two normed vector spaces. We say that a (linear or nonlinear) mapping  $f : \mathcal{A} \subset E \rightarrow F$  is *compact* if for every bounded set  $B \subset \mathcal{A}$ ,  $\overline{f(B)}$  is a compact set of  $F$ . In other words, for every sequence  $(x_n)_{n \in \mathbb{N}}$  that is bounded in  $\mathcal{A}$ , we can extract a subsequence  $(x_{n_k})_{k \in \mathbb{N}}$ , such that  $(f(x_{n_k}))_{k \in \mathbb{N}}$  converges in  $F$ .

**THEOREM 2.4 (Schauder Theorem).–** Let  $B$  be a non-empty, convex, closed and bounded set in a normed vector space  $E$ . Moreover, let  $f : B \rightarrow B$  be a continuous and compact mapping. Then,  $f$  has a fixed point in  $B$ .

We seek to find a fixed point of the mapping  $\Phi$  in the set

$$\mathcal{C} = \{u \in C^0([0, 1]), u(0) = 0 = u(1), 0 \leq u(x) \leq R \text{ for all } x \in [0, 1]\}$$

where  $R > 0$  is a constant that will be determined later. This is a non-empty, convex, closed and bounded set of  $C^0([0, 1])$ . A direct application of the Lax–Milgram theorem (which reproduces the proofs of theorem 2.1 and proposition 2.1) ensures that for all  $T \in \mathcal{C}$ , we can indeed define a unique function  $\Phi(T) = \mathcal{T} \in H_0^1([0, 1])$  that is positive and, by lemma 2.1, also continuous on  $[0, 1]$ . Moreover, we have

$$\sigma \int_0^1 T^3 |\mathcal{T}|^2 dx + \int_0^1 \left| \frac{d}{dx} \mathcal{T} \right|^2 dx = \sigma \int_0^1 T_{\text{ref}}^4 \mathcal{T} dx.$$

The first integral of the left-hand term is non-negative, whereas the right-hand term can be bounded from above using the Cauchy–Schwarz inequality, and then the Poincaré lemma 2.2, by  $\sigma \|T_{\text{ref}}^4\|_{L^2} \|\mathcal{T}\|_{L^2} \leq C \sigma \|T_{\text{ref}}^4\|_{L^2} \left\| \frac{d}{dx} \mathcal{T} \right\|_{L^2}$ . We can infer that

$$\|\mathcal{T}\|_{L^\infty} \leq \sqrt{\int_0^1 \left| \frac{d}{dx} \mathcal{T} \right|^2 dx} \leq C \sigma \|T_{\text{ref}}^4\|_{L^2},$$

using lemma 2.2. Therefore, with  $R \geq C \sigma \|T_{\text{ref}}^4\|_{L^2}$ ,  $\Phi$  is indeed a mapping of  $\mathcal{C}$  onto itself. In fact, this argument also shows that for all  $T \in \mathcal{C}$ , the  $H^1$  norm of the solution  $\Phi(T) = \mathcal{T}$  is bounded from above by  $R$ . As noted throughout the proof of lemma 2.1, it follows (from the Arzela–Ascoli theorem) that  $\Phi$  is a *compact* mapping of  $\mathcal{C}$  onto  $\mathcal{C}$ . It remains to show that  $\Phi$  is continuous in order to conclude by using Schauder’s theorem that there exists a  $T \in \mathcal{C}$ , such that  $\Phi(T) = T$ , which is therefore the non-negative solution of [2.12] (and of [2.11]!).

More generally, we consider a sequence  $(\lambda_n)_{n \in \mathbb{N}}$  of non-negative and continuous functions, such that  $\lambda_n$  converges uniformly to  $\lambda$  on  $[0, 1]$ . We associate it with the sequence  $(u_n)_{n \in \mathbb{N}}$  of solutions for

$$\lambda_n u_n - \frac{d^2}{dx^2} u_n = f \quad \text{for } 0 < x < 1$$

with Dirichlet conditions and where  $f \in L^2([0, 1])$ . We let  $u$  denote the solution to the problem associated with the limit  $\lambda$ . These functions are elements of  $H_0^1([0, 1])$ , and we have

$$\begin{aligned} \int_0^1 \lambda_n |u_n - u|^2 dx + \int_0^1 \left| \frac{d}{dx} (u_n - u) \right|^2 dx &= \int_0^1 (\lambda - \lambda_n) u (u_n - u) dx \\ &\leq \|\lambda_n - \lambda\|_{L^\infty} \|u\|_{L^2} \|u_n - u\|_{L^2}. \end{aligned}$$

By lemma 2.1 and lemma 2.2, it follows that

$$\|u_n - u\|_{L^\infty} \leq \left\| \frac{d}{dx}(u_n - u) \right\|_{L^2} \leq C \|u\|_{L^2} \|\lambda_n - \lambda\|_{L^\infty},$$

which proves that  $u_n$  converges to  $u$  on  $H^1$  and in  $C^0$ .

#### 2.1.2.4. Shooting method

Unlike in the example that we studied in the introduction, the solution of problem [2.11] does not have a simple expression. We will seek to approximate it by using the shooting method. In order to justify the effectiveness of the method, we will use some of the information that we know about the solution. Indeed, we know that the desired solution  $T$  is bounded from above by a certain constant  $M > 0$ . Also, we will replace the nonlinearity  $\sigma T^4$  with

$$f_M(T) = \begin{cases} \sigma T^4 & \text{if } |T| \leq M, \\ \sigma M^4 + \zeta_M(T) & \text{if } |T| \geq M, \end{cases}$$

where  $\zeta_M$  is a function of class  $C^\infty$ , such that  $0 \leq \zeta_M(z) \leq 1$ ,  $\zeta(0) = 0$ ,  $0 \leq \zeta'_M(z) \leq C_M$  and  $f_M$  is  $C^\infty$ . This modification only holds a technical advantage: by adopting the above estimate analysis for the solution of  $f_M(u) - \frac{d^2}{dx^2}u = \sigma T_{\text{ref}}^4$ , with Dirichlet conditions, we can show that  $0 \leq u(x) \leq M$ . Consider the differential system

$$\frac{d}{dt} \begin{pmatrix} T \\ S \end{pmatrix} (t) = \begin{pmatrix} S(t) \\ f_M(T(t)) - \sigma T_{\text{ref}}^4(t) \end{pmatrix}.$$

Assuming that  $t \mapsto T_{\text{ref}}(t)$  is continuous, the mapping

$$\begin{aligned} F : [0, 1] \times \mathbb{R} \times \mathbb{R} &\longrightarrow \mathbb{R}^2 \\ (t, X, Y) &\longmapsto \begin{pmatrix} Y \\ f_M(X) \end{pmatrix} - \begin{pmatrix} 0 \\ \sigma T_{\text{ref}}^4(t) \end{pmatrix} \end{aligned}$$

is indeed continuous and uniformly Lipschitz continuous in the variables  $X, Y$ . For all  $\alpha \in \mathbb{R}$ , the system therefore has a single solution  $(T_\alpha, S_\alpha) : [0, 1] \rightarrow \mathbb{R}^2$ , defined on  $[0, 1]$ , which meets the Cauchy data  $T_\alpha(0) = 0, S_\alpha(0) = \alpha$ . The shooting method involves finding  $\bar{\alpha}$ , such that  $T_{\bar{\alpha}}(1) = 0$ . By the continuity of the solutions of a differential system with respect to the data (see, for example, [ARN 88, Chapter 4, section 31] or [BEN 10, lemma II.3]), the mapping  $\alpha \mapsto T_\alpha(1)$  is continuous. More specifically, the Grönwall lemma shows that  $|T_\alpha(1) - T_\beta(1)| \leq e^L |\alpha - \beta|$ , where  $L$  designates the Lipschitz constant for the function  $F$ . It remains to show that  $T_\alpha(1)$  can take negative and positive values. On the one hand, we have

$$T_\alpha''(t) = f_M(T_\alpha(t)) - \sigma T_{\text{ref}}^4 \geq -\sigma T_{\text{ref}}^4 \geq -\sigma M^4$$

so

$$T'_\alpha(t) \geq \alpha - \sigma M^4 t$$

then

$$T_\alpha(t) \geq \alpha t - \sigma M^4 \frac{t^2}{2}.$$

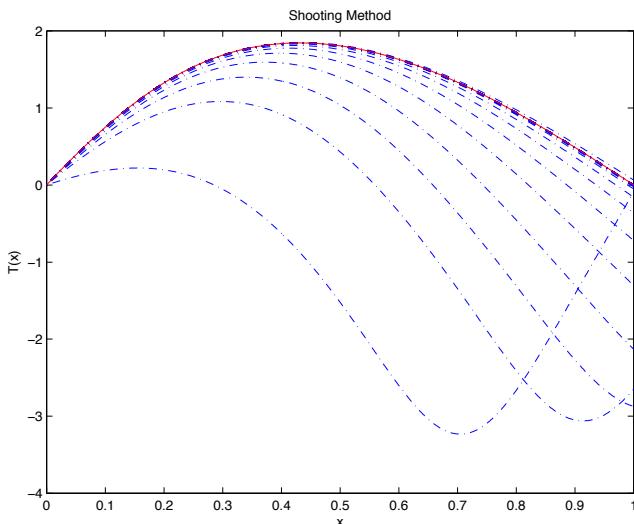
It follows that  $T_\alpha(1) > 0$  when  $\alpha > \frac{\sigma M^4}{2}$ . On the other hand, with the same argument, we obtain

$$T''_\alpha(t) = f_M(T_\alpha(t)) - \sigma T_{\text{ref}}^4 \leq f_M(T_\alpha(t)) \leq \sigma M^4 + 1$$

and finally

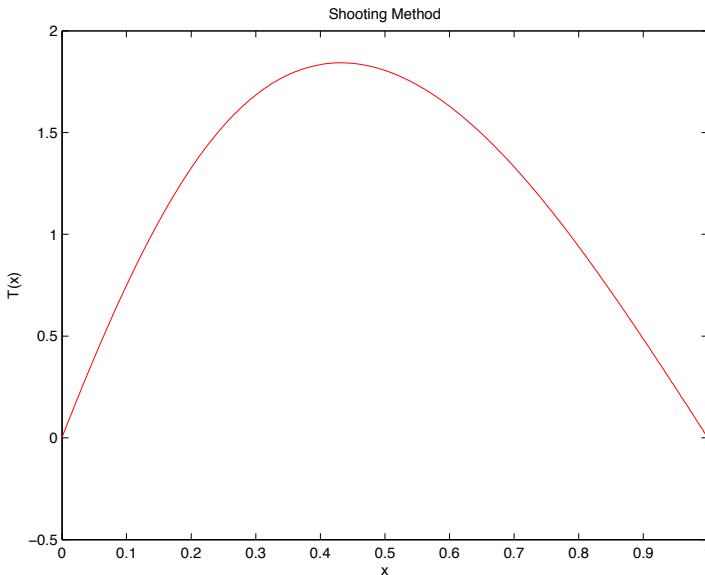
$$T_\alpha(t) \leq \alpha t + (\sigma M^4 + 1) \frac{t^2}{2}.$$

In particular,  $T_\alpha(1) < 0$  when  $\alpha < -\frac{1+\sigma M^4}{2}$ . The intermediate value theorem guarantees the existence of  $\bar{\alpha}$ , such that  $T_{\bar{\alpha}}(1) = 0$ .



**Figure 2.3.** Solution of the radiation transfer problem using the shooting method: successive iterations

We perform simulations with the data  $\sigma = 1.4$  and  $T_{\text{ref}}(x) = 1 + \sin(3x + .5) + .3 * \ln(1.7 + x)$ . The differential system is solved by the explicit Euler scheme. We find  $\bar{\alpha} = 8.1173$ . Figure 2.3 shows some iterations and Figure 2.4 shows the graph of the solution obtained.



**Figure 2.4.** Solution of the radiation transfer problem using the shooting method: solution obtained

#### 2.1.2.5. Interpretation as a minimization problem

It can also be interesting to interpret equation [2.11] as the solution of the minimization problem

$$\min \{ \mathcal{J}(u), u \in H_0^1(]0, 1[) \},$$

$$\text{where } \mathcal{J}(u) = \int_0^1 \left( \frac{1}{2} \left| \frac{du}{dx}(x) \right|^2 + \sigma \left( \frac{|u|^5}{5} - u T_{\text{ref}}^4 \right)(x) \right) dx.$$

Using lemma 2.1, we will confirm that the functional  $\mathcal{J}$  is well defined on  $H_0^1(]0, 1[)$  and strictly convex. More specifically, we note that

$$\begin{aligned} \mathcal{J}(u + h) - \mathcal{J}(u) &= \int_0^1 \left( \frac{du}{dx} \frac{dh}{dx}(x) + \sigma(|u|^3 u - T_{\text{ref}}^4) h(x) \right) dx + R(u, h), \\ R(u, h) &= \int_0^1 \left( \frac{1}{2} \left| \frac{dh}{dx}(x) \right|^2 + \sigma \int_0^1 (1 - \theta) 4|u + \theta h|^3 |h|^2(x) d\theta \right) dx \end{aligned}$$

where  $\lim_{\|h\|_{H_0^1} \rightarrow 0} \frac{R(u, h)}{\|h\|_{H_0^1}} = 0$ . Moreover,

$$\mathcal{J}(u) \geq \int_0^1 \left( \frac{1}{2} \left| \frac{du}{dx}(x) \right|^2 - \sigma u T_{\text{ref}}^4(x) \right) dx \geq \|u\|_{H_0^1} (\|u\|_{H_0^1} - \sigma \|T_{\text{ref}}^4\|_{L^2}), [2.13]$$

implies that  $\lim_{\|h\|_{H_0^1([0,1])} \rightarrow 0} \mathcal{J}(u) = +\infty$ . We infer from the computation that  $\mathcal{J}$  has a minimum. Indeed, if  $(u_n)_{n \in \mathbb{N}}$  designates a minimizing sequence, then  $(\mathcal{J}(u_n))_{n \in \mathbb{N}}$  is bounded in  $\mathbb{R}$ . It follows that that sequence  $(u_n)_{n \in \mathbb{N}}$  is bounded in  $L^5([0, 1])$  and in the separable Hilbert space  $H_0^1([0, 1])$ . It therefore has a subsequence that weakly converges in those spaces (see [GOU 11, corollary 7.32]). We can therefore assume that  $u_n \rightharpoonup T$  weakly in  $L^5([0, 1])$  and  $\frac{du_n}{dx} \rightharpoonup \frac{dT}{dx}$  weakly in  $L^2([0, 1])$ . We use the convexity inequalities

$$\begin{aligned} \int_0^1 \left| \frac{du_n}{dx} \right|^2 dx &= \int_0^1 \left| \frac{d(u_n - u)}{dx} \right|^2 dx + 2 \int_0^1 \frac{du}{dx} \frac{dx(u_n - u)}{dx} dx + \int_0^1 \left| \frac{du}{dx} \right|^2 dx \\ &\geq 2 \int_0^1 \frac{du}{dx} \frac{dx(u_n - u)}{dx} dx + \int_0^1 \left| \frac{du}{dx} \right|^2 dx \end{aligned}$$

and

$$\int_0^1 |u_n|^q dx \geq \int_0^1 |u|^q dx + q \int_0^1 |u|^{q-2} u (u_n - u) dx$$

where  $|u|^{q-2} u \in L^{q'}([0, 1])$  since  $(q-1)q' = q$ . By making  $n \rightarrow \infty$ , we thus obtain  $\lim_{n \rightarrow \infty} \mathcal{J}(u_n) \geq \mathcal{J}(T)$ , so  $T$  is a minimizer of  $\mathcal{J}$ . Since  $\mathcal{J}$  is strictly convex, this minimizer is unique. Finally, [2.13] shows that the minimizer  $T$  is characterized by

$$\int_0^1 \left( \frac{dT}{dx} \frac{dh}{dx}(x) + \sigma (|T|^3 T - T_{\text{ref}}^4) h(x) \right) dx = 0$$

for all  $h \in H_0^1([0, 1])$ . In other words,  $T$  is a solution of [2.12].

### 2.1.3. Analysis elements for multidimensional problems

The approach detailed in section 2.1.1 can be adapted to higher dimensions. The analysis involves the preliminary introduction of appropriate functional spaces. We will therefore let  $H^1(\Omega)$  denote the set of square-integrable functions  $u : \Omega \subset \mathbb{R}^D \rightarrow$

$\mathbb{R}$ , such that there exists a constant  $C > 0$  that, for all  $\varphi \in C_c^\infty(\Omega)$  and all  $i \in \{1, \dots, D\}$ , satisfies

$$\left| \int_{\Omega} u(x) \partial_{x_i} \varphi(x) dx \right| \leq C \|\varphi\|_{L^2(\Omega)}.$$

This allows us to define the gradient of  $u$  by duality as the vector whose components  $\partial_{x_i} u$  are elements of  $L^2(\Omega)$ . This space is equipped with the norm

$$\|u\|_{H^1} = \left( \|u\|_{L^2}^2 + \sum_{i=1}^D \int_{\Omega} |\partial_{x_i} u|^2 dx \right)^{1/2}$$

which makes it a Hilbert space. We define  $H_0^1(\Omega)$  as the complement of  $C_c^\infty(\Omega)$  for this norm. This space can be interpreted as the set of functions of  $H^1(\Omega)$  “that cancel out on  $\partial\Omega$ ”. This statement is ambiguous unless the precise meaning of the trace of elements of  $H^1(\Omega)$  on  $\partial\Omega$  is specified: in dimension  $D > 1$ , this point represents a real difficulty because the elements are not necessarily continuous functions. The definition of the traces of elements of  $H^1(\Omega)$  is a delicate and deep subject that goes beyond the scope of this book, and here we provide only an intuitive sketch. (The reader may refer to [EVA 98, Chapter 5] or [ZUI 02].) Equation [2.1] can thus be understood in the following variational form:

Finding  $u \in H$ , such that for all  $v \in H$ , we have  $a(u, v) = \ell(v)$ ,

where:

–  $H$  is  $H_0^1(\Omega)$  for the case of Dirichlet conditions, or  $H^1(\Omega)$  for Neumann conditions;

–  $a$  is the bilinear form

$$a(u, v) = \lambda \int_{\Omega} u(x) v(x) dx + \int_{\Omega} k(x) \nabla_x u(x) \cdot \nabla_x v(x) dx.$$

Note that  $a$  satisfies the continuity property  $|a(u, v)| \leq (\lambda + K) \|u\|_{H^1} \|v\|_{H^1}$ ;

–  $\ell$  is the linear form

$$\ell(v) = \int_{\Omega} f(x) v(x) dx,$$

which satisfies the continuity property  $|\ell(v)| \leq \|f\|_{L^2} \|v\|_{L^2} \leq \|f\|_{L^2} \|v\|_{H^1}$ .

To apply the Lax–Milgram theorem, a coercivity estimate is also required to be confirmed. It can be obtained relatively directly whenever  $\lambda > 0$  since

$$a(u, u) \geq \lambda \|u\|_{L^2}^2 + \kappa \|\nabla u\|_{L^2}^2 \geq \min(\lambda, \kappa) \|u\|_{H^1}^2.$$

When  $\lambda = 0$ , we have only  $a(u, u) \geq \kappa \|\nabla u\|_{L^2}^2$ . We must therefore show something analogous to lemma 2.2, which only applies to the space  $H_0^1(\Omega)$ .

**LEMMA 2.3** (Poincaré lemma (general case)).— We assume that  $\Omega \subset \mathbb{R}^D$  is an open set, bounded at least in one direction. Then, there exists a  $C > 0$ , such that for all  $u \in H_0^1(\Omega)$ , we have

$$\|u\|_{L^2} \leq C \|\nabla u\|_{L^2}.$$

**PROOF.**— We will demonstrate this relation for regular functions and conclude by density. Let  $\varphi \in C_c^\infty(\Omega)$ . We assume that  $\Omega$  is bounded in the direction  $\omega \in \mathbb{S}^{D-1}$ : there exists  $R > 0$ , such that for all  $y \in \Omega$ ,  $|y \cdot \omega| \leq R$ . We write

$$\varphi(x) = \int_{-\infty}^0 \frac{d}{ds} \varphi(x + s\omega) ds = \int_{-\infty}^0 \nabla \varphi(x + s\omega) \cdot \omega ds,$$

which leads to

$$|\varphi(x)| \leq \int_{-\infty}^{\infty} |\nabla \varphi(x + s\omega)| ds$$

where the integration domain can, in fact, be restricted to  $|s| \leq |x \cdot \omega| + |s + x \cdot \omega| \leq 2R$ . Using the Cauchy–Schwarz inequality, it follows that

$$\begin{aligned} \int_{\Omega} |\varphi(x)|^2 dx &\leq \int_{\Omega} \left( \int_{-2R}^{2R} ds \times \int_{-2R}^{2R} |\nabla \varphi(x + s\omega)|^2 ds \right) dx \\ &\leq 4R \int_{-2R}^{2R} \left( \int_{\Omega} |\nabla \varphi(x + s\omega)|^2 dx \right) ds = 16R^2 \|\nabla \varphi\|_{L^2}^2. \end{aligned} \quad \square$$

Finally, we obtain the following general result.

**THEOREM 2.5.**— Let  $\Omega \subset \mathbb{R}^D$  be an open set bounded in at least one direction. Let  $\lambda \geq 0$  (respectively  $\lambda > 0$ ) and  $f \in L^2(\Omega)$ . Then, problem [2.1] has a unique solution  $u \in H_0^1(\Omega)$  (respectively  $u \in H^1(\Omega)$ ) in the sense that for all  $v \in H_0^1(\Omega)$  (respectively  $v \in H^1(\Omega)$ ), we have  $a(u, v) = \ell(v)$ .

Moreover, there exists  $C > 0$  independent of the data  $f$ , such that  $\|u\|_{H^1} \leq C \|f\|_{L^2}$ . If  $0 \leq f(x) \leq M$  for almost all  $x \in \Omega$ , then  $0 \leq u(x) \leq M/\lambda$ .

The estimates on the solution can be shown as in dimension one, by choosing an appropriate test function. The noticeable point is a regularization effect: for a data  $f$  that is only  $L^2$ , the solution and its derivatives are  $L^2$  functions. We know that when  $\Omega$  is bounded, the embedding  $H^1(\Omega) \subset L^2(\Omega)$  is compact (see [GOU 11, corollary 7.58]). Then,  $f \mapsto u$ , which associates with data  $f$  the solution  $u \in H^1(\Omega)$  of [2.1], is a compact operator (see definition 2.1), which implies remarkable spectral properties (see [BRÉ 05, Chapter 6]).

**PROPOSITION 2.2.** – Let  $\Omega$  be a regular bounded open set of  $\mathbb{R}^D$ . Then, there exists a sequence  $(\lambda_k)_{k \in \mathbb{N}}$  of positive real numbers and a sequence  $(\psi_k)_{k \in \mathbb{N}}$  of elements of  $H_0^1(\Omega)$ , such that

$$-\Delta \psi_k = \lambda_k \psi_k, \quad \lim_{k \rightarrow \infty} \lambda_k = +\infty$$

and the set  $\{\psi_k, k \in \mathbb{N}\}$  forms an orthonormal basis for  $H_0^1(\Omega)$ .

**NOTE 2.1.** – Whenever  $\Omega = ]0, 1[$ , solving  $-\psi'' = \lambda_n \psi$ ,  $\psi(0) = 0 = \psi(1)$  leads to  $\lambda_n = \pi^2 n^2$ ,  $n \in \mathbb{N} \setminus \{0\}$  and  $\psi_n(x) = \sin(\pi n x)$ . In the periodic case, we also find  $\lambda_n = 4\pi^2 n^2$ ,  $\psi_n(x) = e^{2i\pi n x}$ .

## 2.2. Finite difference approximations to elliptic equations

### 2.2.1. Finite difference discretization principles

We will focus only on the one-dimensional case and thus consider problem [2.2] where we assume that

- h1)  $\lambda \geq 0$ ;
- h2)  $x \mapsto f(x) \in C^0([0, 1])$  is given;
- h3)  $x \mapsto k(x) \in C^0([0, 1])$  is given and there exist  $\kappa, K > 0$ , such that for all  $x \in [0, 1]$ , we have  $0 < \kappa \leq k(x) \leq K$ .

The finite difference method is constructed as follows:

- a) Consider a subdivision of the segment  $[0, 1]$  for which we denote the discretization points as  $x_j$ :

$$0 = x_0 < x_1 < \dots < x_j < x_{j+1} < \dots < x_{J+1} = 1.$$

We write

$$h_{j+1/2} = x_{j+1} - x_j$$

and

$$h = \max \{h_{j+1/2}, j \in \{0, \dots, J\}\}$$

which measure how “fine-grained” the discretization is. Note that  $J$  and  $h$  are related; in the case of a *uniform grid*, the space step is constant  $h_j = h$  for all  $j \in \{0, \dots, J\}$ , and we simply have  $h = \frac{1}{J+1}$ .

b) The numerical unknown is a vector  $U = (u_1, \dots, u_J) \in \mathbb{R}^J$  whose components are interpreted as potential approximations of the unknown  $u$  of the continuous problem [2.2] at the discretization points  $x_1, \dots, x_J$ .

c) The scheme is constructed by writing a linear system  $\mathcal{A}U = F$ , where  $F = (f(x_1), \dots, f(x_J)) \in \mathbb{R}^D$  is given. (Note that  $F$  is well defined since we assume that  $f$  is continuous.) The coefficients  $a_{i,j}$  of the matrix  $\mathcal{A}$  are precisely defined so that by writing  $\mathcal{U} = (u(x_1), \dots, u(x_J))$ , where  $u$  is the solution of the continuous problem [2.2], the *consistency error*

$$E_i = (\mathcal{A}\mathcal{U})_i - F_i \quad [2.14]$$

tends to 0 when  $h \rightarrow 0$ .

There are several constructions that fulfill criterion c). In order to limit the complexity of the linear system that needs to be solved, we will construct a *sparse* matrix  $\mathcal{A}$ , that is to say, with several coefficients equal to zero. Therefore, it would be very natural to attempt to approach  $\lambda u(x_i)$  by only employing the value of  $u$  at the point  $x_i$  and to approximate the derivatives by only using neighboring discretization points  $x_{i \pm 1}$ . We thus decompose

$$\mathcal{A} = \lambda \mathbb{I} + \mathbb{A}.$$

By denoting the coefficients of the matrix  $\mathbb{A}$  as  $\alpha_{i,j}$ , we assume that

$$\alpha_{i,j} = 0 \quad \text{if } |i - j| > 1.$$

It remains to establish that the consistency error [2.14] tends to 0 when  $h \rightarrow 0$ , that is, with  $x \mapsto u(x)$  as the solution to [2.2],

$$\lambda u(x_i) - (\alpha_{i,i-1}u(x_{i-1}) + \alpha_{i,i}u(x_i) + \alpha_{i,i+1}u(x_{i+1})) - f(x_i) \xrightarrow[h \rightarrow 0]{} 0. \quad [2.15]$$

Let us begin by studying the simplest case where  $x \mapsto k(x) = k > 0$ . Then,  $(\alpha_{i,i-1}u(x_{i-1}) + \alpha_{i,i}u(x_i) + \alpha_{i,i+1}u(x_{i+1}))$  should approach  $ku''(x)$ . We obtain

$$(\alpha_{i,i-1}u(x_{i-1}) + \alpha_{i,i}u(x_i) + \alpha_{i,i+1}u(x_{i+1}))$$

$$\begin{aligned}
&= (\alpha_{i,i-1} + \alpha_{i,i} + \alpha_{i,i+1})u(x_i) \\
&\quad + (\alpha_{i,i-1}(x_{i-1} - x_i) + \alpha_{i,i+1}(x_{i+1} - x_i))u'(x_i) \\
&\quad + \left( \alpha_{i,i-1} \frac{(x_i - x_{i-1})^2}{2} + \alpha_{i,i+1} \frac{(x_{i+1} - x_i)^2}{2} \right) u''(x_i) + \mathcal{O}(h^3).
\end{aligned}$$

It therefore leads to the following linear system to identify the coefficients  $\alpha_{i,j}$

$$\begin{aligned}
\alpha_{i,i-1} + \alpha_{i,i} + \alpha_{i,i+1} &= 0, \\
\alpha_{i,i-1}(x_{i-1} - x_i) + \alpha_{i,i+1}(x_{i+1} - x_i) &= 0, \\
\alpha_{i,i-1} \frac{(x_i - x_{i-1})^2}{2} + \alpha_{i,i+1} \frac{(x_{i+1} - x_i)^2}{2} &= k.
\end{aligned}$$

With  $h_{i+1/2} = x_{i+1} - x_i$ , it follows that

$$\alpha_{i,i-1} = \alpha_{i,i+1} \frac{h_{i+1/2}}{h_{i-1/2}} \quad \text{and} \quad \alpha_{i,i+1}(h_{i+1/2}^2 + h_{i+1/2}h_{i-1/2}) = 2k.$$

Finally, we obtain

$$\begin{aligned}
\alpha_{i,i-1} &= \frac{2k}{h_{i-1/2}} \frac{1}{h_{i-1/2} + h_{i+1/2}}, \\
\alpha_{i,i+1} &= \frac{2k}{h_{i+1/2}} \frac{1}{h_{i-1/2} + h_{i+1/2}}, \\
\alpha_{i,i} &= -\frac{2k}{h_{i-1/2} + h_{i+1/2}} \left( \frac{1}{h_{i-1/2}} + \frac{1}{h_{i+1/2}} \right).
\end{aligned} \tag{2.16}$$

In the case where the grid is uniform with step  $h$ , the numerical unknowns are determined by the solution of the linear system

$$\left( \lambda \mathbb{I} - \frac{k}{h^2} \begin{pmatrix} -2 & 1 & 0 & \cdots & 0 \\ 1 & -2 & 1 & \ddots & \vdots \\ 0 & 1 & -2 & \ddots & 0 \\ \vdots & \ddots & \ddots & \ddots & 1 \\ 0 & 0 & 0 & 1 & -2 \end{pmatrix} \right) U = F. \tag{2.17}$$

We follow the same steps for a variable (but regular) coefficient  $x \mapsto k(x)$ . To this end, we develop  $(ku')' = k'u' + ku''$ . The discretization coefficients  $\alpha_{i,j}$  are thus defined by the linear system

$$\alpha_{i,i-1} + \alpha_{i,i} + \alpha_{i,i+1} = 0 \quad (\text{coefficient with } u(x_i)),$$

$$-\alpha_{i,i-1}h_{i-1/2} + \alpha_{i,i+1}h_{i+1/2} = k'(x_i) \quad (\text{coefficient with } u'(x_i)),$$

$$\alpha_{i,i-1}\frac{h_{i-1/2}^2}{2} + \alpha_{i,i+1}\frac{h_{i+1/2}^2}{2} = k(x_i) \quad (\text{coefficient with } u''(x_i)).$$

It follows that

$$\begin{aligned} \alpha_{i,i-1} &= \frac{\alpha_{i,i+1}h_{i+1/2} - k'(x_i)}{h_{i-1/2}}, \\ \alpha_{i,i+1}\left(\frac{h_{i+1/2}^2}{2} + \frac{h_{i-1/2}^2}{2}\frac{h_{i+1/2}}{h_{i-1/2}}\right) &= k(x_i) + k'(x_i)\frac{h_{i-1/2}^2}{2h_{i-1/2}} \end{aligned}$$

that is to say

$$\begin{aligned} \alpha_{i,i+1} &= \frac{1}{h_{i+1/2} + h_{i-1/2}}\left(\frac{2k(x_i)}{h_{i+1/2}} + k'(x_i)\frac{h_{i-1/2}}{h_{i+1/2}}\right), \\ \alpha_{i,i-1} &= \frac{1}{h_{i+1/2} + h_{i-1/2}}\left(\frac{2k(x_i)}{h_{i-1/2}} - k'(x_i)\frac{h_{i+1/2}}{h_{i-1/2}}\right), \\ \alpha_{i,i} &= -\alpha_{i,i-1} - \alpha_{i,i+1} \\ &= -\frac{1}{h_{i+1/2} + h_{i-1/2}}\left(2k(x_i)\left(\frac{1}{h_{i+1/2}} + \frac{1}{h_{i-1/2}}\right) \right. \\ &\quad \left. + k'(x_i)\left(\frac{h_{i+1/2}}{h_{i-1/2}} - \frac{h_{i-1/2}}{h_{i+1/2}}\right)\right). \end{aligned}$$

For a uniform grid with step  $h$ , this becomes

$$\alpha_{i,i} = -2\frac{k(x_i)}{h^2}, \quad \alpha_{i,i+1} = \frac{k(x_i)}{h^2} + \frac{k'(x_i)}{2h}, \quad \alpha_{i,i-1} = \frac{k(x_i)}{h^2} - \frac{k'(x_i)}{2h} \quad [2.18]$$

and the scheme is expressed as

$$\lambda u_i - \frac{k(x_i)}{h^2}(u_{i-1} - 2u_i + u_{i+1}) - \frac{k'(x_i)}{2h}(u_{i+1} - u_{i-1}) = f(x_i).$$

Note that when the coefficient  $k$  or the grid step is variable, the matrix associated with the discrete problem is not symmetric<sup>6</sup>. This is a shortcoming in two respects:

- on the one hand, the discrete equation comes from a continuous problem that has symmetric properties (the bilinear form  $a$  of the Lax–Milgram theorem is symmetric for the case studied here); the discretization has therefore brought about a loss of structure;

- on the other hand, we have specific numerical methods for solving symmetric systems of linear equations more efficiently (conjugate gradient method, see section A1.3).

We propose a different discretization for variable coefficients on a uniform grid with step  $h > 0$ . The idea is to approximate the derivative of a given regular function  $\varphi$  with a centered differential quotient inspired by the relation

$$\lim_{h \rightarrow 0} \frac{\varphi(x + h/2) - \varphi(x - h/2)}{h} = \varphi'(x).$$

More specifically, we have the estimate

$$\left| \frac{\varphi(x + h/2) - \varphi(x - h/2)}{h} - \varphi'(x) \right| \leq \|\varphi''\|_{L^\infty} h^2,$$

which is a better estimate than would be obtained with right-centered approximations  $\frac{\varphi(x+h)-\varphi(x)}{h}$  or left-centered approximations  $\frac{\varphi(x)-\varphi(x-h)}{h}$ . We use this idea to approximate  $(ku')'$  by using the values of  $ku'$  at the points  $x \pm h/2$ , and then approximating  $u'(x \pm h/2)$  using the same principle with the values of  $u$  at the points  $x_{i-1}, x_i, x_{i+1}$ . The scheme obtained along this line of thought takes the form of the following linear system

$$\lambda u_i - \frac{1}{h} \left( k(x_{i+1/2}) \frac{u_{i+1} - u_i}{h} - k(x_{i-1/2}) \frac{u_i - u_{i-1}}{h} \right) = f(x_i)$$

for  $i \in \{1, \dots, J\}$  and with the convention  $u_0 = 0 = u_{J+1}$ . Assuming that  $a$  is a regular function, this approximation is consistent with order  $\mathcal{O}(h^2)$ . We will see that this rationale becomes similar to finite volume techniques by giving a key role to the interfaces  $x_{i \pm 1/2}$ . An advantage of this discretization is that it preserves the problem's symmetry. However, we will also see (in section 2.3.1) that for a non-uniform grid, this scheme is no longer consistent, even though it is possible to show that it is convergent!

**NOTE 2.2 (Non-homogeneous conditions).**— It is interesting to write an “extended” form of problem [2.17] by dealing with a discrete unknown  $\bar{U}$  of size  $(N + 2)$  which

---

<sup>6</sup> For a constant coefficient  $k$ , this remark can become more relative: in fact, the matrix  $A$  is symmetric for the scalar product  $\langle u | v \rangle = \sum_{i=1}^J u_i v_i \frac{h_{i+1/2} + h_{i-1/2}}{2}$ . We can verify that, indeed,  $\langle Au | v \rangle = \langle u | Av \rangle$ , whereas  $A$  is not symmetric for the usual scalar product on  $\mathbb{R}^N$ .

includes the boundary terms  $\bar{U}_0$  and  $\bar{U}_{J+1}$ . In that case, we let  $\bar{\mathbb{I}}$  denote the identity matrix of size  $(J+2) \times (J+2)$ , and write

$$\left( \lambda \bar{\mathbb{I}} - \frac{k}{h^2} \begin{pmatrix} 1 & 0 & & & & & 0 \\ 1 & -2 & 1 & 0 & & & 0 \\ 0 & 1 & & & & & \\ \vdots & & & & & & \\ 0 & & & & & & \\ 0 & & & & & & \\ 0 & & & & & & \end{pmatrix} \right) \bar{U} = \bar{F}.$$

The extended second member  $\bar{F}$  is constructed by taking into account the boundary conditions:  $\bar{F}_0 = 0$ ,  $\bar{F}_{J+1} = 0$  and  $\bar{F}_j = F_j = f(x_j)$  for  $j \in \{1, \dots, J\}$ . We can easily take into account non-homogeneous Dirichlet data by modifying the extremes of the second member  $\bar{F}$ . (Precisely, using the Dirichlet conditions  $u(0) = \alpha$  and  $u(1) = \beta$ , we write  $\bar{F}_0 = (\lambda - \frac{k}{h^2})\alpha$ ,  $\bar{F}_{J+1} = (\lambda - \frac{k}{h^2})\beta$ .) Note that the resulting matrix is no longer symmetric. One way of restoring symmetry is to introduce a penalty term

$$\left( \lambda \bar{\mathbb{I}} - \frac{k}{h^2} \begin{pmatrix} 1/\epsilon & 1 & & & & & 0 \\ 1 & -2 & 1 & 0 & & & 0 \\ 0 & 1 & & & & & \\ \vdots & & & & & & \\ 0 & & & & & & \\ 0 & & & & & & \\ 0 & & & & & & 1/\epsilon \end{pmatrix} \right) \bar{U} = \bar{F}^\epsilon,$$

where we also modify the extremes of the second member  $\bar{F}_0^\epsilon = \frac{1}{\epsilon}\bar{F}_0$ ,  $\bar{F}_{J+1}^\epsilon = \frac{1}{\epsilon}\bar{F}_{J+1}$  for a parameter  $0 < \epsilon \ll 1$ .

**NOTE 2.3 (Neumann boundary conditions).**— When we use Neumann boundary conditions, the matrix  $\mathbb{A}$  must be modified. For Neumann conditions and a constant coefficient  $k$ , with a constant grid step  $h$ , we obtain

$$\mathbb{A}^{\text{Neumann}} = -\frac{k}{h^2} \begin{pmatrix} -1 & 1 & 0 & \cdots & 0 \\ 1 & -2 & & & \\ 0 & & -2 & 1 & \\ \vdots & & & & 0 \\ 0 & \cdots & 0 & 1 & -1 \end{pmatrix}.$$

Indeed, the condition  $u'(0) = 0 = u'(1)$  becomes  $u_1 - u_0 = 0 = u_{J+1} - u_J$  at the discrete level. The last formula takes its cue from the fact that  $\frac{u(x+h) - u(x)}{h} - u'(x) = \mathcal{O}(h)$ : it provides a consistent first-order approximation of  $u'$ . We can improve the scheme by using an approximation, that uses the equation satisfied by  $u$ , which is second-order consistent. Indeed, we have seen that the construction of the scheme relies, within the calculation domain, on an approximation of the equation that is second-order consistent. A “naïve” treatment of the Neumann condition can therefore alter the global quality of the approximation. We use the development

$$\begin{aligned} u(x+h) &= u(x) + u'(x)h + \underbrace{u''(x)\frac{h^2}{2}}_{h^2\epsilon(x,h)} + h^2\epsilon(x,h), \text{ with } \lim_{h \rightarrow 0} \epsilon(x,h) = 0, \\ &= \frac{1}{k}(\lambda u(x) - f(x)) \end{aligned} \quad \text{by [2.2]}$$

which defines  $u_0$  and  $u_{J+1}$  by the following relations:

$$u_1 = u_0 + \frac{h^2}{2k}(\lambda u_0 - f(0)), \quad u_J = u_{J+1} + \frac{h^2}{2k}(\lambda u_{J+1} - f(1)).$$

**NOTE 2.4 (Periodic conditions).**— For periodic conditions, we must also modify the discretization matrix in order to take into account the behavior imposed on the boundaries. We obtain

$$\mathbb{A}^{\text{Perio}} = -\frac{k}{h^2} \begin{pmatrix} 2 & -1 & 0 & \cdots & 0 & -1 \\ -1 & 2 & -1 & \cdots & 0 & 0 \\ 0 & & 2 & -1 & \cdots & 0 \\ \vdots & & & & & \vdots \\ 0 & \cdots & 0 & 2 & -1 & \end{pmatrix}$$

because, for the discrete problem, the periodicity condition requires that  $u_0 = u_J$  and  $u_{J+1} = u_1$ .

### 2.2.2. Analysis of the discrete problem

We now analyze the discrete problem [2.17], which can be written in the form  $\mathcal{A}U = F$ , with  $\mathcal{A} = \lambda I + \mathbb{A}$  and

$$\mathbb{A} = -\frac{k}{h^2} \begin{pmatrix} -2 & 1 & 0 & \cdots & 0 \\ 1 & -2 & 1 & \ddots & \vdots \\ 0 & 1 & -2 & \ddots & 0 \\ \vdots & \ddots & \ddots & \ddots & 1 \\ 0 & \cdots & 0 & 1 & -2 \end{pmatrix}. \quad [2.19]$$

This equation therefore corresponds to the finite difference discretization of the model problem [2.2], for a constant coefficient  $k$  and a uniform grid of step  $h$ . Of course, the first question to ask is whether the linear system has a solution.

**PROPOSITION 2.3.–** When  $\lambda > 0$ , the matrix  $\mathcal{A}$  has a strictly dominant diagonal, so it is invertible.

**PROOF.–** When  $\lambda > 0$ , we find that, indeed,  $\mathcal{A}_{i,i} > \sum_{j \neq i} |\mathcal{A}_{i,j}|$  for all  $i \in \{1, \dots, J\}$ . Suppose there exists a vector  $x \neq 0$ , such that  $\mathcal{A}x = 0$ . At the possible cost of renormalizing  $x$ , we assume that  $x_i = \max \{|x_j|, j \in \{1, \dots, J\}\} = 1$ . So we have

$$\mathcal{A}_{i,i}x_i = -\sum_{j \neq i} \mathcal{A}_{i,j}x_j = \mathcal{A}_{i,i} > \sum_{j \neq i} |\mathcal{A}_{i,j}|$$

which contradicts the fact that

$$\left| -\sum_{j \neq i} \mathcal{A}_{i,j}x_j \right| \leq \sum_{j \neq i} |\mathcal{A}_{i,j}| |x_j| \leq \sum_{j \neq i} |\mathcal{A}_{i,j}|.$$

The matrix  $\mathcal{A}$  is therefore invertible. □

For  $\lambda = 0$ , the matrix  $\mathcal{A}$  does not have a strictly dominant diagonal: the diagonal coefficients are positive and the extra-diagonal coefficients are negative or zero, but we can have  $\mathcal{A}_{i,i} + \sum_{j \neq i} \mathcal{A}_{i,j} = 0$ . However, the matrix  $\mathcal{A}$  is still invertible.

**PROPOSITION 2.4.–** When  $\lambda = 0$ , the matrix  $\mathcal{A} = \mathbb{A}$  is invertible.

**PROOF.–** The argument is inspired by the derivation of an *a priori* estimate for the continuous problem (see theorem 2.1). We compute

$$\begin{aligned}\mathbb{A}U \cdot U &= \frac{k}{h^2} \sum_{i=1}^J (2u_i^2 - u_{i+1}u_i - u_{i-1}u_i) \\ &= \frac{k}{h^2} \sum_{i=1}^J (u_i - u_{i+1})u_i + \frac{k}{h^2} \sum_{i=1}^J (u_i - u_{i-1})u_i \\ &= \frac{k}{h^2} \sum_{i=1}^J (u_i - u_{i+1})^2 \geq 0.\end{aligned}$$

Moreover, if  $\mathbb{A}U \cdot U = 0$ , then for all  $i \in \{1, \dots, J\}$ , we have  $u_{i+1} = u_i$ . Now, the boundary conditions (which we have used in the foregoing manipulations) require that  $u_0 = 0 = u_{J+1} = 0$ , which finally implies that  $U = 0$ . This justifies the fact that  $\mathbb{A}$  is invertible.  $\square$

In particular, this analysis shows that the matrix  $\mathcal{A}$  is *symmetric positive definite* (with a constant space step). This observation is, in practice, important because it allows us to resort to better performing methods for solving systems of linear equations (Cholesky or conjugate gradient algorithms, see [LAS 04, Chapter 4] and section A1.3).

**NOTE 2.5** (Neumann boundary conditions and periodic conditions).– Matrices that correspond to Neumann or periodic conditions (see notes 2.3 and 2.4) are not invertible. The matrices  $\mathbb{A}^{\text{Neumann}}$  and  $\mathbb{A}^{\text{Perio}}$  have a non-trivial kernel spanned by the vector  $(1, \dots, 1)$ . This is consistent with the analysis of the continuous problem for which the constants are solutions of the homogeneous equation. In practice, we seek solutions whose average is zero (or that have a fixed value at a given point). This leads us to replace one of the original matrix's rows with a different expression, which corresponds to the constraint and makes it invertible again. In general, this manipulation makes the matrix stop being symmetric, which makes it impossible to resort to specific and effective methods for solving symmetric linear systems. In section 2.8, we will study an alternative.

Studying the continuous problem has revealed that for positive data  $f$ , the solution  $u$  is positive (maximum principle). It is important to make sure that the discrete problem preserves this property. Such a result is related to the structure of the matrix  $\mathcal{A}$ .

**DEFINITION 2.2.**— We say that a matrix  $M \in \mathcal{M}_J(\mathbb{R})$  is *positive*<sup>7</sup> if its coefficients  $m_{i,j}$  are positive or equal to zero.

**DEFINITION 2.3.**— We say that a matrix  $A \in \mathcal{M}_J(\mathbb{R})$  is *monotone* if  $A$  is invertible and the matrix  $A^{-1}$  is positive. We also say that  $A$  is an *M-matrix*.

**THEOREM 2.6.**— A matrix  $A \in \mathcal{M}_J(\mathbb{R})$  satisfies

$$(*) \quad \text{if the components of } Ax \text{ are } \geq 0, \text{ then the components of } x \text{ are also } \geq 0$$

if and only if  $A$  is monotone.

**PROOF.**— Assume that the condition  $(*)$  is satisfied. Let  $x \in \mathbb{R}^J$ , such that  $Ax = 0$ . Then, the components of  $Ax$  are both positive and negative and  $(*)$  therefore implies that the components of  $x$  and those of  $(-x)$  are positive or zero. It follows that, in fact,  $x = 0$ . Therefore,  $A$  is invertible. Let us suppose that  $Ax = y$  has positive components. Then,  $x = A^{-1}y$  has positive components by  $(*)$ . In particular, by taking  $y = e_i$ , the elements of the canonical basis for  $\mathbb{R}^J$ ,  $A^{-1}e_i$ , which corresponds to the  $i$ th column of  $A^{-1}$ , have positive components. We conclude that  $A^{-1}$  is positive.

Reciprocally, if  $A$  is monotone, and  $Ax = y$  has positive components, then  $x = A^{-1}y$  also has positive components because the coefficients of  $A^{-1}$  are positive.  $\square$

**THEOREM 2.7.**— For all  $\lambda \geq 0$ , the matrix  $\mathcal{A}$  is monotone.

**PROOF.**— Suppose that  $\mathcal{A}u = F$  has positive components and that there exists  $i \in \{1, \dots, J\}$ , such that  $u_i = \min \{u_j, j \in \{0, \dots, J+1\}\}$ , where we recall that  $u_0 = u_{J+1} = 0$  (note that under this assumption,  $u_i$  is an interior point). Then, we have

$$\lambda u_i + \underbrace{\frac{k}{h^2}(u_i - u_{i-1})}_{\leq 0} + \underbrace{\frac{k}{h^2}(u_i - u_{i+1})}_{\leq 0} \geq 0.$$

If  $\lambda > 0$ , this implies directly that  $u_i \geq 0$ . If  $\lambda = 0$ , we can infer from the relation that  $u_i = u_{i+1} = u_{i-1}$ . Repeating the same reasoning, we conclude that  $u$  is a constant vector with  $u_i = u_0 = u_{J+1} = 0$ , that is,  $u = 0$ . However, this would imply that  $\mathcal{A}u = F = 0$ . It follows that if  $F$  has positive components with  $F \neq 0$ , then the vector  $u \in \mathbb{R}^J$ , which is a solution of  $\mathcal{A}u = F'$  has only strictly positive components.  $\square$

**LEMMA 2.4.**— For a matrix  $M \in \mathcal{M}_J(\mathbb{R})$ , we let  $\|M\|_\infty = \sup \{|Mx|_\infty, x \in \mathbb{R}^J, |x|_\infty = 1\}$ , with  $|\cdot|_\infty$ , denoting the infinite norm in  $\mathbb{R}^J$ :  $|x|_\infty = \sup \{|x_i|, i \in \{1, \dots, J\}\}$ . We have  $\|\mathcal{A}^{-1}\|_\infty \leq \frac{1}{8k}$ .

---

<sup>7</sup> It is important that this definition is not confused with the positive character of the quadratic form associated with a symmetric matrix.

PROOF.– We begin by noting that  $\|\mathbb{A}^{-1}\|_\infty = |\mathbb{A}^{-1}e|_\infty$ , where  $e = (1, \dots, 1)$ . This is because  $\mathbb{A}^{-1}$  is monotone. For  $x, y \in \mathbb{R}^J$ , we designate  $|x|$  as the vector of components  $|x_i|$  and write  $x \geq y$  if for all  $i \in \{1, \dots, J\}$ , we have  $x_i \geq y_i$ . Then, for every vector  $x$ , such that  $|x|_\infty = 1$ , since  $\mathbb{A}$  is monotone, we have  $\mathbb{A}^{-1}e \geq \mathbb{A}^{-1}|x| \geq |\mathbb{A}^{-1}x|$ . Now,  $\mathbb{A}z = e$  corresponds to the discretization of the equation

$$-k\mathcal{U}''(x) = 1 \quad \text{pour } x \in ]0, 1[, \quad \mathcal{U}(0) = 0 = \mathcal{U}(1).$$

The solution of this equation is known explicitly:  $\mathcal{U}(x) = \frac{x(1-x)}{2k}$ . Since it is a second-order polynomial, the consistency error is zero: we have exactly

$$-k \frac{\mathcal{U}(x_{i+1}) - 2\mathcal{U}(x_i) + \mathcal{U}(x_{i-1})}{h^2} = 1.$$

In other words, the solution of  $\mathbb{A}z = e$  is given by  $z_i = \mathcal{U}(x_i) = \frac{ih(1-ih)}{2k}$ . It follows that

$$\|\mathbb{A}^{-1}\|_\infty = |\mathbb{A}^{-1}e|_\infty = |z|_\infty \leq \sup_{x \in [0, 1]} \frac{x(1-x)}{2k} = \frac{1}{8k}. \quad \square$$

We can interpret this estimate as a stability property. Therefore, combining the stability and consistency properties of the finite difference scheme, we can arrive at the convergence of discrete solutions towards the solution of the continuous problem.

**THEOREM 2.8.–** Let  $x \mapsto u(x)$  be the solution of problem [2.2], with constant  $\lambda = 0$  and  $k > 0$ . For all  $J \in \mathbb{N} \setminus \{0\}$ , we associate it with the vector  $\bar{u} \in \mathbb{R}^J$  whose components are  $u(x_1), \dots, u(x_J)$ , where  $x_i = ih$ ,  $h = 1/(J+1)$ . We let  $u^h \in \mathbb{R}^J$  denote the solution of the linear system  $\mathbb{A}u^h = F$ , where  $F = (f(x_1), \dots, f(x_J))$ . Thus, we have

$$|u^h - \bar{u}|_\infty \leq \frac{\|u^{(4)}\|_\infty}{96} h^2.$$

PROOF.– Let  $R$  designate the vector of consistency errors

$$R_i = -k \frac{u(x_{i+1}) - 2u(x_i) + u(x_{i-1})}{h^2} - f(x_i) = (\mathbb{A}\bar{u} - F)_i.$$

Then, we have

$$\mathbb{A}(u^h - \bar{u}) = F - \mathbb{A}\bar{u} = -R.$$

Therefore,

$$|u^h - \bar{u}|_\infty \leq \|\mathbb{A}^{-1}\|_\infty |R|_\infty = \frac{1}{8} |R|_\infty.$$

The consistency error can therefore be estimated by using the Taylor expansion of  $u$  and we obtain  $|R|_\infty \leq h^2 \|u^{(4)}\|_\infty / 12$ . Note that this estimate requires higher-order derivatives of the solution  $u$ . This assumes that the solution is regular, beyond mere  $H^1$  regularity ensured by the functional arguments discussed above.  $\square$

We provide an analogous result for the problem with  $\lambda > 0$ .

**THEOREM 2.9.**— Let  $\lambda > 0$  and  $k > 0$  be a constant. The matrix  $\mathcal{A} = (\lambda\mathbb{I} + \mathbb{A})$  satisfies  $\|\mathcal{A}^{-1}\|_\infty \leq \frac{1}{8k}$ . Let  $x \mapsto u(x)$  be the solution of problem [2.2]. For all  $J \in \mathbb{N} \setminus \{0\}$ , we associate it with a vector  $\bar{u} \in \mathbb{R}^J$  whose components are  $u(x_1), \dots, u(x_J)$ , where  $x_i = ih$ ,  $h = 1/(J+1)$ . Let  $u^h \in \mathbb{R}^J$  denote the solution of the linear system  $\mathcal{A}u^h = F$ , where  $F = (f(x_1), \dots, f(x_J))$ . Then, we have

$$|u^h - \bar{u}|_\infty \leq \frac{\|u^{(4)}\|_\infty}{96} h^2.$$

**PROOF.**— Surprisingly, the stability estimate for  $\mathcal{A} = \lambda\mathbb{I} + \mathbb{A}$  is a corollary of the case when  $\lambda = 0$ . The trick is to write

$$\mathbb{A}^{-1} - \mathcal{A}^{-1} = \mathbb{A}^{-1} \mathcal{A} \mathcal{A}^{-1} - \mathbb{A}^{-1} \mathbb{A} \mathcal{A}^{-1} = \mathbb{A}^{-1} (\mathcal{A} - \mathbb{A}) \mathcal{A}^{-1} = \lambda \mathbb{A}^{-1} \mathcal{A}^{-1}.$$

Now,  $\mathbb{A}^{-1}$  and  $\mathcal{A}^{-1}$  are both matrices with positive coefficients as a result of theorem 2.7. It follows that all the coefficients of  $\mathbb{A}^{-1} - \mathcal{A}^{-1}$  are also positive and, therefore, with the notation of lemma 2.4, that  $\mathbb{A}^{-1}e \geq \mathcal{A}^{-1}e$ . We conclude that

$$\|\mathbb{A}\|_\infty = |\mathbb{A}^{-1}e|_\infty \geq |\mathcal{A}^{-1}e|_\infty = \|\mathcal{A}^{-1}\|_\infty.$$

Finally, it remains to analyze the consistency error and reproduce the arguments for the proof of theorem 2.8.  $\square$

Let us finish with a few remarks on the *spectral properties* of the matrix  $\mathbb{A}$ , in the case where  $k$  is constant. Let us first note that the spectrum of the operator  $-k \frac{d^2}{dx^2}$  with Dirichlet conditions on  $[0, 1]$  is  $\{\pi^2 n^2, n \in \mathbb{N}\}$ . Indeed, let  $\phi$  be the solution of  $-k \frac{d^2}{dx^2} \psi = \lambda \psi$ . By integrating by parts, we note that  $k \int_0^1 |\psi'(x)|^2 dx = \lambda \int_0^1 |\psi(x)|^2 dx \geq 0$ , which implies that necessarily  $\lambda \geq 0$ . Now, the solutions of  $-k \frac{d^2}{dx^2} \psi = \lambda \psi$ , with  $\lambda \geq 0$ , are of the form  $A \cos(\sqrt{\lambda/k} x) + B \sin(\sqrt{\lambda/k} x)$ , with  $A, B \in \mathbb{R}$ . Dirichlet conditions ensure that  $A = 0$  and  $\sqrt{\lambda/k} = \pi n$  for  $n \in \mathbb{Z}$ . We conclude that the eigenvalues are of the form  $k\pi^2 n^2$ , for  $n \in \mathbb{N} \setminus \{0\}$ , and the associated eigenspaces are given by the functions  $\psi_n : x \mapsto \sin(\pi n x)$ .

Let  $J \in \mathbb{N} \setminus \{0\}$ . For  $n, j \in \mathbb{N}$ , we write  $\psi_{n,J}^j = \sin(\pi j n / (J+1))$ , which satisfies

$$\psi_{n,J}^0 = 0 = \psi_{n,J}^{J+1}, \quad \psi_{n,J+1}^{j+\ell(J+1)} = (-1)^\ell \psi_{n,J}^j \quad \text{for any } \ell \in \mathbb{N}$$

and for  $j \in \{1, \dots, J\}$ ,  $\psi_{n,J}^j = \psi_n(j/(J+1))$ .

We thus evaluate the matrix  $\mathbb{A}$  on the vector  $\psi_{n,J} = (\psi_{n,J}^1, \dots, \psi_{n,J}^J)$ . We have

$$\begin{aligned} k(J+1)^2(-\psi_{n,J}^{j+1} - \psi_{n,J}^{j-1} + 2\psi_{n,J}^j) \\ = k(J+1)^2 \operatorname{Im}\left(2e^{i\pi j n / (J+1)}\left(1 - \cos\left(\pi \frac{n}{J+1}\right)\right)\right) \\ = 4k(J+1)^2 \sin^2\left(\frac{\pi n}{2(J+1)}\right) \sin\left(\frac{\pi j n}{J+1}\right) \\ = 4k(J+1)^2 \sin^2\left(\frac{\pi n}{2(J+1)}\right) \psi_{n,J}^j. \end{aligned}$$

Therefore, the matrix  $\mathbb{A} \in \mathcal{M}_J$  has  $J$  distinct eigenvalues

$$\lambda_1 = 4k(J+1)^2 \sin^2\left(\frac{\pi}{2(J+1)}\right), \dots, \lambda_J = 4k(J+1)^2 \sin^2\left(\frac{J\pi}{2(J+1)}\right)$$

associated with the eigenvectors  $\psi_{n,J}$ ,  $n \in \{1, \dots, J\}$ .

In particular, note that the conditioning of  $\mathbb{A}$  degrades when the discretization step tends to 0 because

$$\operatorname{cond}(\mathbb{A}) = \lambda_J / \lambda_1 \underset{J \rightarrow \infty}{\sim} 4J^2 / \pi^2.$$

In practice, poor conditioning is not necessarily all that bad: we often use “regular” second members, in the sense that they are described by a small number of basis function  $\psi_n$  (this is to say, we have an estimate of the type  $\|f - \sum_{n=0}^N c_f(n) \psi_n\|_{L^2} \leq CN^{-r}$ , with  $c_f(n) = \int_0^1 f(x) \psi_n(x) dx$ ). The stability estimate is therefore far more favorable than that involving conditioning.

**NOTE 2.6.–** The construction of the approximation can be generalized easily to higher dimensions by arguing coordinate by coordinate. For example, in dimension 2, the equation

$$-\Delta u = (\partial_x^2 + \partial_y^2)u = f, \quad u|_{\partial\Omega} = 0$$

can be approximated with the linear system

$$-\frac{1}{\Delta x^2}(u_{i+1,j} - 2u_{i,j} + u_{i-1,j}) - \frac{1}{\Delta y^2}(u_{i,j+1} - 2u_{i,j} + u_{i,j-1}) = f(i\Delta x, j\Delta y). [2.20]$$

When the problem is set on the domain  $[0, L] \times [0, M]$ , with

$$(I+1)\Delta x = L \quad \text{et} \quad (J+1)\Delta y = M,$$

these equations are satisfied for  $i \in \{1, \dots, I\}$ ,  $j \in \{1, \dots, J\}$  with Dirichlet conditions  $u_{0,j} = u_{I+1,j} = 0 = u_{i,0} = u_{i,J+1}$ . In practice, we construct a system of size  $IJ \times IJ$  in the usual matrix form. The construction is as follows: given the quantities  $v_{i,j}$ , where  $i \in \{1, \dots, I\}$ ,  $j \in \{1, \dots, J\}$ , we construct a vector  $V \in \mathbb{R}^{IJ}$  by writing  $V_{i+(j-1)J} = v_{i,j}$ , which can be summarized by writing in coordinates

$$\tilde{V}_j = \begin{pmatrix} v_{1,j} \\ v_{2,j} \\ \vdots \\ v_{I,j} \end{pmatrix} \in \mathbb{R}^I, \quad V = \begin{pmatrix} \tilde{V}_1 \\ \tilde{V}_2 \\ \vdots \\ \tilde{V}_J \end{pmatrix} \in \mathbb{R}^{IJ}$$

We write  $K = i + (j-1)J$ , which spans  $\{1, \dots, IJ\}$  when  $i$  and  $j$  pass through  $\{1, \dots, I\}$  and  $\{1, \dots, J\}$ , respectively. In this way, we define the second member  $F$  using the values  $f(i\Delta x, j\Delta y)$ . Relation [2.20] now takes the generic form

$$-\frac{1}{\Delta x^2}U_{K-1} - \frac{1}{\Delta x^2}U_{K+1} + 2\left(\frac{1}{\Delta x^2} + \frac{1}{\Delta y^2}\right)U_K - \frac{1}{\Delta y^2}U_{K+J} - \frac{1}{\Delta y^2}U_{K-J} = F_K.$$

Finally, we write it as a matrix problem  $\mathbb{A}U = F$ , where the matrix  $\mathbb{A}$  has a tridiagonal structure by blocks (with  $J$  blocks by rows and columns)

$$\mathbb{A} = \begin{pmatrix} & -\frac{\mathbb{I}}{\Delta y^2} & 0 & \cdots & 0 \\ -\frac{\mathbb{I}}{\Delta y^2} & & & & \vdots \\ 0 & & & & 0 \\ \vdots & & & & \vdots \\ 0 & \cdots & 0 & -\frac{\mathbb{I}}{\Delta y^2} & \mathbb{B} \end{pmatrix}$$

with  $\mathbb{I}$ , the identity matrix on  $\mathbb{R}^I$  and  $\mathbb{B}$ , the following matrix of size  $I \times I$

$$\mathbb{B} = \begin{pmatrix} b & -\frac{1}{\Delta x^2} & 0 & \cdots & 0 \\ -\frac{1}{\Delta x^2} & b & -\frac{1}{\Delta x^2} & \cdots & 0 \\ 0 & -\frac{1}{\Delta x^2} & b & -\frac{1}{\Delta x^2} & \cdots \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & \cdots & 0 & -\frac{1}{\Delta x^2} & b \end{pmatrix}, \quad b = 2\left(\frac{1}{\Delta x^2} + \frac{1}{\Delta y^2}\right).$$

An example for the right-hand side

$$\begin{aligned} f(x, y) = & 20 \exp(-100((x - 0.25)^2 + (y - 0.5)^2)) \\ & + 15 \exp(-120((x - 0.75)^2 + (y - 0.4)^2)) \end{aligned}$$

is shown in Figure 2.5

## 2.3. Finite volume approximation of elliptic equations

### 2.3.1. Discretization principles for finite volumes

We consider the same model problem [2.2], but with a different approach: the idea is to integrate the equation over *control volumes*  $\mathcal{C}_i = [\bar{x}_{i-1/2}, \bar{x}_{i+1/2}]$ , where

$$0 = \bar{x}_{1/2} < \bar{x}_{3/2} < \dots < \bar{x}_{I-1/2} < \bar{x}_{I+1/2} = 1, \quad i \in \{0, \dots, I\}.$$

We write  $\bar{x}_0 = 0$ ,  $\bar{x}_{I+1} = 1$  and for  $i \in \{1, \dots, I\}$ , we set

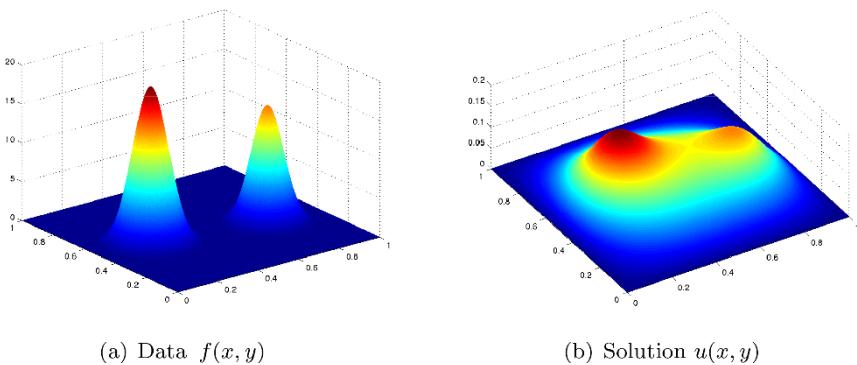
$$\bar{x}_i = \frac{\bar{x}_{i+1/2} + \bar{x}_{i-1/2}}{2}$$

which is the barycenter of the control volume  $\mathcal{C}_i$ . Finally, for  $i \in \{1, \dots, I\}$ , we write

$$h_{i+1/2} = \bar{x}_{i+1} - \bar{x}_i, \quad h_i = \bar{x}_{i+1/2} - \bar{x}_{i-1/2}$$

which are, respectively, the distance between the barycenters of the volumes  $\mathcal{C}_i$  and  $\mathcal{C}_{i+1}$ , and the length of volume  $\mathcal{C}_i$ . In particular, we note that

$$h_{i+1/2} = \frac{h_i}{2} + \frac{h_{i+1}}{2}. \quad [2.21]$$

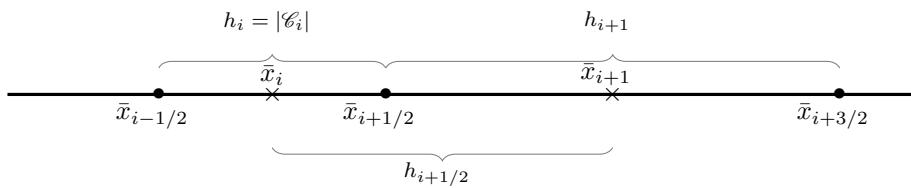


**Figure 2.5.** Solution to the 2D Poisson equation. For the color version of this figure, see [www.iste.co.uk/goudon/mathmodel.zip](http://www.iste.co.uk/goudon/mathmodel.zip)

This notation is specified in Figure 2.6.

We have

$$\lambda \int_{\mathcal{C}_i} u(y) \, dy - \left( k(\bar{x}_{i+1/2}) \frac{d}{dx} u(\bar{x}_{i+1/2}) - k(\bar{x}_{i-1/2}) \frac{d}{dx} u(\bar{x}_{i-1/2}) \right) = \int_{\mathcal{C}_i} f(y) \, dy.$$



**Figure 2.6.** Finite volume discretization: control volumes are the segments  $\mathcal{C}_i = [\bar{x}_{i-1/2}, \bar{x}_{i+1/2}]$

The numerical unknown  $U = (u_1, \dots, u_I) \in \mathbb{R}^I$  is understood to give an approximation of the mean values

$$\frac{1}{h_i} \int_{\mathcal{C}_i} u(y) \, dy.$$

This leads us to write the scheme in the form

$$\lambda h_i u_i + (F_{i+1/2} - F_{i-1/2}) = h_i \bar{f}_i \quad [2.22]$$

where

$$\bar{f}_i = \frac{1}{h_i} \int_{\mathcal{C}_i} f(y) dy.$$

We naturally associate the vector  $U = (u_1, \dots, u_I)$  with the following function that is constant over each control volume

$$u_h(x) = \sum_{i=1}^I u_i \mathbf{1}_{\mathcal{C}_i}(x).$$

The key notion of FV methods is the *numerical flux*: the point is to find a relevant definition of  $F_{i+1/2}$ , as a function of the components  $U_1, \dots, U_I$ . This quantity is designed to be an approximation of the flux at the interface  $\bar{x}_{i+1/2}$  between the control volumes  $\mathcal{C}_i$  and  $\mathcal{C}_{i+1}$ . For the solution  $x \mapsto u(x)$  of the continuous problem, the flux is

$$\mathcal{F}_{i+1/2} = -k(\bar{x}_{i+1/2}) \frac{d}{dx} u(\bar{x}_{i+1/2}).$$

We approximate this quantity using a differential quotient that leads to the formula

$$F_{i+1/2} = -k(\bar{x}_{i+1/2}) \frac{u_{i+1} - u_i}{h_{i+1/2}}.$$

We have considered a formula of the same kind for the finite difference approximation on a regular grid at the end of section 2.2.1. This definition makes sense for the interior points, where  $i \in \{1, \dots, I\}$ ; it must be modified for interfaces on the boundary. At  $x = 0$  and  $x = 1$ , the function  $x \mapsto u(x)$  vanishes. We interpret this condition as fixing the values  $u_0$  and  $u_{I+1}$  at 0 in control volumes of size zero, which are concentrated at the boundary points  $x = 0 = \bar{x}_{1/2} = \bar{x}_0$  and  $x = 1 = \bar{x}_{I+1/2} = \bar{x}_{I+1}$ . We thus obtain

$$F_{1/2} = -2k(0) \frac{u_1}{h_1}, \quad F_{I+1/2} = +2k(1) \frac{u_I}{h_I}.$$

Finally, the discrete problem comes down to solving the linear system

$$(\lambda \text{diag}(h_1, \dots, h_I) + \bar{\mathbb{A}})u = \text{diag}(h_1, \dots, h_I)\bar{f}$$

with  $\bar{f} = (\bar{f}_1, \dots, \bar{f}_I)$  and

$$\bar{\mathbb{A}}_{i,j} = \bar{\mathbb{A}}_{j,i},$$

$$\bar{\mathbb{A}}_{i,j} = 0 \quad \text{if } |i - j| > 1,$$

$$\bar{\mathbb{A}}_{i,i} = \frac{k(\bar{x}_{i+1/2})}{h_{i+1/2}} + \frac{k(\bar{x}_{i-1/2})}{h_{i-1/2}}, \quad i \in \{2, \dots, I-1\}$$

$$\bar{\mathbb{A}}_{i,i+1} = -\frac{k(\bar{x}_{i+1/2})}{h_{i+1/2}}, \quad i \in \{2, \dots, I-1\},$$

$$\bar{\mathbb{A}}_{I,I} = +\frac{k(\bar{x}_{I-1/2})}{h_{I-1/2}} + 2\frac{k(1)}{h_I},$$

$$\bar{\mathbb{A}}_{1,1} = +\frac{k(\bar{x}_{1/2})}{h_{1/2}} + 2\frac{k(0)}{h_1}.$$

We should pay special attention to the location of the discretization coefficients  $h_i$  in the resulting system [2.22]. In particular, it would not be correct to divide the equations [2.22] by  $h_i$ : the corresponding linear system would not be symmetric (for a general grid). It is also useful to compare the expression of the finite volume scheme with that of the finite difference scheme, even for a constant  $k$ . Let  $h > 0$  denote the grid's step, which is assumed to be uniform (except for the boundary points, as we will see in more detail). Therefore, for the interior points  $i \in \{2, \dots, I-1\}$ , we have

$$\lambda h u_i^{VF} - \frac{k}{h}(u_{i+1}^{VF} - 2u_i^{VF} + u_{i-1}^{VF}) = h\bar{f}_i$$

for the finite volume and

$$\lambda u_i^{DF} - \frac{k}{h^2}(u_{i+1}^{DF} - 2u_i^{DF} + u_{i-1}^{DF}) = f(\bar{x}_i)$$

for the finite difference scheme. The expression is different at the boundaries, where it should be noted that  $\bar{x}_{3/2} - \bar{x}_{1/2} = h$ ,  $\bar{x}_1 - \bar{x}_0 = h/2$ , with  $\bar{x}_0 = \bar{x}_{1/2} = 0$ . For the finite volume scheme, we have

$$\lambda h u_1^{VF} - \underbrace{\frac{k}{h}(u_2^{VF} - 3u_1^{VF})}_{F_{3/2}-F_{1/2}} = h\bar{f}_1.$$

For the finite difference scheme (see [2.16], with  $h_{3/2} = h$ ,  $h_{1/2} = h/2$ ), we obtain

$$\lambda u_1^{DF} - \frac{4k}{3h^2}(u_2^{DF} - 3u_1^{DF}) = f(h/2).$$

An analogous situation presents itself for the last discretization element:

$$\lambda h u_I^{VF} - \frac{k}{h}(u_{I-1}^{VF} - 3u_I^{VF}) = h \bar{f}_I \quad vs.$$

$$\lambda u_I^{DF} - \frac{4k}{3h^2}(u_{I-1}^{DF} - 3u_I^{DF}) = f(1 - h/2).$$

In particular, we observe that, since the finite difference grid is not uniform, the associated matrix is not symmetric (for the usual scalar product). By writing  $F_{i+1/2} = \frac{u_{i+1} - u_i}{h_{i+1/2}}$ , the finite volume scheme is itself defined by

$$\underbrace{\lambda h_i u_i + F_{i+1/2} - F_{i-1/2}}_{[(\lambda \text{diag}(h_1, \dots, h_I) + \bar{\mathbb{A}})u]_i} = h_i f_i. \quad [2.23]$$

The matrix for that system is symmetric (by construction) and positive definite because

$$\begin{aligned} (\lambda \text{diag}(h_1, \dots, h_I) + \bar{\mathbb{A}})u \cdot u &= \lambda \sum_{i=0}^I h_i u_i^2 + \sum_{i=0}^I F_{i+1/2} (u_{i+1} - u_i) \\ &= \lambda \sum_{i=0}^I h_i u_i^2 + \sum_{i=0}^I k(x_{i+1/2}) \frac{(u_{i+1} - u_i)^2}{h_{i+1/2}} \geq 0. \end{aligned}$$

This quantity equals zero if and only if we have  $u_{i+1} = u_i$  for each index  $i$ , that is to say, taking into account the boundary conditions,  $u = 0$ .

It is therefore tempting to interpret  $u_i$  as an approximation of  $u(\bar{x}_i)$ , as in the analysis carried out for finite difference schemes. However, the scheme obtained in that analysis is not, in general, consistent in the sense required for finite differences. Indeed, let us study the case where  $k(x) = k > 0$  is constant. If we apply the formula that defines the scheme for values  $u(\bar{x}_i)$ , we obtain

$$\begin{aligned} \lambda u(\bar{x}_i) - \frac{k}{h_i} \left( \frac{u(\bar{x}_{i+1}) - u(\bar{x}_i)}{h_{i+1/2}} - \frac{u(\bar{x}_i) - u(\bar{x}_{i-1})}{h_{i-1/2}} \right) \\ = \lambda u(\bar{x}_i) - \frac{k}{h_i} \left( u'(\bar{x}_i) + u''(\bar{x}_i) \frac{h_{i+1/2}}{2} - u'(\bar{x}_i) + u''(\bar{x}_i) \frac{h_{i-1/2}}{2} \right) + \mathcal{O}(h) \\ = \lambda u(\bar{x}_i) - k \frac{h_{i+1/2} + h_{i-1/2}}{2h_i} u''(\bar{x}_i) + \mathcal{O}(h). \end{aligned}$$

Therefore, the consistency error does not necessarily tend to 0 when  $h \rightarrow 0$ . For example, we can define a grid where  $h_{2p} = h$  and  $h_{2p+1} = h/2$ , in which case  $h_{i+1/2} + h_{i-1/2} = 3h/2$  and the ratio  $\frac{h_{i+1/2} + h_{i-1/2}}{2h_i}$  alternatively takes the values  $3/4$  and  $3/2$ . However, that sense of consistency is satisfied when the grid is uniform, so that  $h_{i+1/2} = h_i = h$ .

Nevertheless, we can show that the finite volume scheme converges. By definition, the right-hand term of [2.23] is equal to

$$\int_{\bar{x}_{i-1/2}}^{\bar{x}_{i+1/2}} f(y) dy = \lambda h_i \bar{u}_i + \mathcal{F}_{i+1/2} - \mathcal{F}_{i-1/2}$$

where we denote

$$\bar{u}_i = \frac{1}{h_i} \int_{\bar{x}_{i-1/2}}^{\bar{x}_{i+1/2}} u(y) dy.$$

Note that, on the one hand,

$$|\bar{u}_i - u(\bar{x}_i)| = \left| \frac{1}{h_i} \int_{\bar{x}_{i-1/2}}^{\bar{x}_{i+1/2}} (u(y) - u(\bar{x}_i)) dy \right| \leq Ch,$$

and, on the other hand, using the expansion of  $u$  in  $x_{i+1/2}$ ,

$$\begin{aligned} & \left| u'(\bar{x}_{i+1/2}) - \frac{u(\bar{x}_{i+1}) - u(\bar{x}_i)}{h_{i+1/2}} \right| \\ &= \left| u'(\bar{x}_{i+1/2}) - \frac{u(\bar{x}_{i+1}) - u(\bar{x}_{i+1/2}) + u(\bar{x}_{i+1/2}) - u(\bar{x}_i)}{h_{i+1/2}} \right| \leq Ch \end{aligned} \quad [2.24]$$

where  $C$  depends on  $\|u\|_{C^2}$ . In what follows, we will again denote such a quantity as  $C$ , even if the value may change from one line to another. The second inequality employs the fact that  $\bar{x}_{i+1} - \bar{x}_{i+1/2} = h_{i+1}/2$ ,  $\bar{x}_{i+1/2} - \bar{x}_i = h_i/2$  and  $h_{i+1/2} = \bar{x}_{i+1} - \bar{x}_i = h_{i+1}/2 + h_i/2$ . We introduce the notation

$$\begin{aligned} e_i &= u_i - u(\bar{x}_i), \\ R_{i+1/2} &= k(\bar{x}_{i+1/2}) \frac{u(\bar{x}_{i+1}) - u(\bar{x}_i)}{h_{i+1/2}} - \mathcal{F}_{i+1/2} \\ &= k(\bar{x}_{i+1/2}) \left( \frac{u(\bar{x}_{i+1}) - u(\bar{x}_i)}{h_{i+1/2}} - u'(\bar{x}_{i+1/2}) \right), \end{aligned}$$

and we have

$$\begin{aligned} \lambda h_i(u_i - \bar{u}_i) + (F_{i+1/2} - \mathcal{F}_{i+1/2}) - (F_{i-1/2} - \mathcal{F}_{i-1/2}) &= 0 \\ &= \lambda h_i e_i + k(\bar{x}_{i+1/2}) \frac{e_{i+1} - e_i}{h_{i+1/2}} - k(\bar{x}_{i-1/2}) \frac{e_i - e_{i-1}}{h_{i-1/2}} \\ &\quad + \lambda h_i(u(\bar{x}_i) - \bar{u}_i) + R_{i+1/2} - R_{i-1/2}. \end{aligned}$$

We multiply this expression by  $e_i$  and add to obtain

$$\begin{aligned} \lambda \sum_i h_i |e_i|^2 + \sum_i k(\bar{x}_{i+1/2}) \frac{|e_{i+1} - e_i|^2}{h_{i+1/2}} \\ = \sum_i R_{i+1/2}(e_{i+1} - e_i) + \lambda \sum_i h_i(\bar{u}_i - u(x_i))e_i. \end{aligned}$$

Using the Cauchy–Schwarz inequality, the right-hand term can be dominated by

$$\begin{aligned} &\sqrt{\sum_i h_{i+1/2} |R_{i+1/2}|^2} \sqrt{\sum_i \frac{|e_{i+1} - e_i|^2}{h_{i+1/2}}} + \lambda \sqrt{\sum_i h_i |\bar{u}_i - u(x_i)|^2} \sqrt{\sum_i h_i |e_i|^2} \\ &\leq Ch \left( \sqrt{\sum_i \frac{|e_{i+1} - e_i|^2}{h_{i+1/2}}} + \sqrt{\lambda \sum_i h_i |e_i|^2} \right) \end{aligned}$$

because  $\sum h_{i+1/2}$  and  $\sum h_i$  are bounded above by the size of the domain. Since  $\sqrt{a} + \sqrt{b} \leq \sqrt{2}\sqrt{a+b}$ , and because  $k(x) \geq \kappa > 0$ , we conclude that

$$\left( \lambda \sum_i h_i |e_i|^2 + \sum_i \frac{|e_{i+1} - e_i|^2}{h_{i+1/2}} \right)^{1/2} \leq Ch, \quad [2.25]$$

which can be interpreted as an estimate on a discrete  $H^1$  norm of the error. Note that  $e_i = \sum_{j=0}^{i-1} (e_{j+1} - e_j)$  because  $e_0 = 0$ , which enables us to obtain the discrete Poincaré inequality

$$\max_{i \in \{0, \dots, I\}} |e_i| \leq C \left( \sum_j \frac{|e_{j+1} - e_j|^2}{h_{j+1/2}} \right)^{1/2}.$$

Finally, we infer that  $\max_{i \in \{0, \dots, I\}} |e_i| \leq Ch$ .

Note that when using a general grid, the order of convergence is hampered: if we assume that the grid is uniform of step  $h$ , then for [2.25], we obtain an estimate in

$\mathcal{O}(h^2)$ . This results from the fact that in [2.24] the second-order term of the expansion for  $\frac{u(\bar{x}_{i+1/2})}{2h_{i+1/2}}((\bar{x}_{i+1} - \bar{x}_{i+1/2})^2 - (\bar{x}_i - \bar{x}_{i+1/2})^2)$  is zero for that grid. We thus obtain an estimate of the form  $|u'(\bar{x}_{i+1/2}) - \frac{u(\bar{x}_{i+1}) - u(\bar{x}_i)}{h}| \leq Ch^2$ . More generally, we should note that the key ingredient for the convergence proof lies in the estimate of the remainder  $R_{i+1/2}$ : although the discrete equation is not consistent, the fluxes will be consistent. As in the FD scheme, we establish the maximum principle and the  $L^\infty$  stability. Assume that  $f \geq 0$ ,  $f \neq 0$ . Let  $i_0$  be an index, such that  $u_{i_0} = \min\{u_0, \dots, u_{I+1}\}$ , with the convention  $u_0 = 0 = u_{I+1}$ . Suppose that  $i_0$  is an interior point: if  $i_0 \in \{1, \dots, I\}$ , then we have

$$0 \leq h_{i_0} f_{i_0} = \lambda h_{i_0} u_{i_0} - \frac{k_{i_0+1/2}}{h_{i_0+1/2}} \underbrace{(u_{i_0+1} - u_{i_0})}_{\geq 0} + \frac{k_{i_0-1/2}}{h_{i_0-1/2}} \underbrace{(u_{i_0} - u_{i_0-1})}_{\leq 0}.$$

Whenever  $\lambda > 0$ , this proves that  $u_{i_0} \geq 0$ . If  $\lambda = 0$ , we can infer from this relation that  $u_{i_0} = u_{i_0 \pm 1}$ . We thus obtain the result that  $f$  is exactly zero, which is a contradiction. We have  $u_j > 0$  for all  $j \in \{1, \dots, I\}$ . As a result, the matrix  $\lambda \text{diag}(h_1, \dots, h_I) + \bar{\mathbb{A}}$  is monotone. The  $L^\infty$  stability lies in the fact that we know a specific solution whenever there is a constant second member  $f$ . If we know  $w$ , such that  $\bar{\mathbb{A}}w_j = h_j$ , then the solution  $u$  of  $\bar{\mathbb{A}}u = hf$  satisfies  $\bar{\mathbb{A}}(\|f\|_\infty w - u)_j = h_j(\|f\|_\infty - f_j) \geq 0$ . We infer that  $\|f\|_\infty w_j \geq u_j$  for all  $j$ , so  $\|u\|_\infty \leq \|f\|_\infty \|w\|_\infty$ . Finally, for constant coefficients and whenever  $\lambda = 0$ , we can verify that  $w_j = \frac{1}{2}(\bar{x}_j)(1 - \bar{x}_j) + \frac{h_j^2}{8}$  satisfies  $\bar{\mathbb{A}}w_j = h_j$ .

### 2.3.2. Discontinuous coefficients

When the coefficient  $k$  is discontinuous on an interface  $x = \bar{x}_{i+1/2}$ , there is an ambiguity in defining the value of the numerical coefficient  $k_{i+1/2}$ . The idea is to reproduce the formulas from the continuous case. We write  $k^\pm = \lim_{h \rightarrow 0^\pm} k(\bar{x}_{i+1/2} + h)$ . For  $f \in L^2$ , we have seen that the flux  $ku'$  is continuous: when  $k$  has a jump,  $u'$  also has a jump in such a way that the product remains continuous ( $u$  is continuous but not  $C^1$  in that case; see the example described in Figure 2.16). We thus have

$$\begin{aligned} \lim_{h \rightarrow 0^+} k(\bar{x}_{i+1/2} + h) \frac{u(\bar{x}_{i+1/2} + h) - u(\bar{x}_{i+1/2})}{h} \\ = \lim_{h \rightarrow 0^-} k(\bar{x}_{i+1/2} + h) \frac{u(\bar{x}_{i+1/2} + h) - u(\bar{x}_{i+1/2})}{h}. \end{aligned}$$

We first seek  $u_{i+1/2}$  such that the discrete analogue is satisfied

$$k^+ \frac{u_{i+1} - u_{i+1/2}}{x_{i+1} - x_{i+1/2}} = k^- \frac{u_{i+1/2} - u_i}{x_{i+1/2} - x_i}.$$

We let  $F_{i+1/2}$  denote this common value, which will be the value of the numerical flux at the interface  $x_{i+1/2}$ . In other words, we have

$$u_{i+1/2} = \frac{u_{i+1}k^+/h_{i+1} + u_i k^-/h_i}{k^+/h_{i+1} + k^-/h_i}.$$

Then,  $k_{i+1/2}$  is defined, so that  $F_{i+1/2} = k_{i+1/2} \frac{u_{i+1} - u_i}{h_{i+1/2}}$ . We thus obtain the following expression for the numerical flux

$$F_{i+1/2} = \frac{2k^+k^-}{h_{i+1}h_i} \frac{h_{i+1/2}}{k^+/h_{i+1} + k^-/h_i} \frac{u_{i+1} - u_i}{h_{i+1/2}},$$

this is to say,

$$k_{i+1/2} = \frac{2h_{i+1/2}}{h_i/k^- + h_{i+1}/k^+}. \quad [2.26]$$

When  $h_i = h_{i+1/2} = h$ , this coefficient is just the harmonic mean  $\frac{2}{1/k^- + 1/k^+}$ .

We have seen that the key point in establishing convergence of the finite volume method lies in the consistency property of the fluxes, which is equivalent to showing that

$$\lim_{h \rightarrow 0} R_{i+1/2} = \lim_{h \rightarrow 0} k_{i+1/2} \frac{u(\bar{x}_{i+1}) - u(\bar{x}_i)}{h_{i+1/2}} - w(\bar{x}_{i+1/2}) = 0,$$

where  $w(x) = (k \frac{d}{dx} u)(x)$ . Even though  $k$  is discontinuous, the regularity of the second member  $f$  implies that  $w$  is a regular function (at least continuous for  $f \in L^2$ , see lemma 2.1, and we will specify the required degree of regularity for the proof). We rewrite the desired quantity in the form

$$R_{i+1/2} = \frac{k_{i+1/2}}{h_{i+1/2}} \int_{\bar{x}_i}^{\bar{x}_{i+1}} \frac{w(y)}{k(y)} dy - w(\bar{x}_{i+1/2}).$$

Now, we have

$$w(y) = w(\bar{x}_{i+1/2}) + \int_0^1 w'(\bar{x}_{i+1/2} + \theta(y - \bar{x}_{i+1/2}))(y - \bar{x}_{i+1/2}) d\theta$$

which leads to

$$R_{i+1/2} = w(\bar{x}_{i+1/2}) \left( \frac{k_{i+1/2}}{h_{i+1/2}} \int_{\bar{x}_i}^{\bar{x}_{i+1}} \frac{dy}{k(y)} - 1 \right) + r_h$$

where the remainder  $r_h$  indeed tends to 0 when  $h \rightarrow 0$  because

$$\begin{aligned} |r_h| &= \left| \frac{k_{i+1/2}}{h_{i+1/2}} \int_{\bar{x}_i}^{\bar{x}_{i+1}} \frac{1}{k(y)} \int_0^1 w'(\bar{x}_{i+1/2} + \theta(y - \bar{x}_{i+1/2}))(y - \bar{x}_{i+1/2}) d\theta dy \right| \\ &\leq \frac{K}{\kappa} \|w'\|_{L^\infty} h. \end{aligned}$$

Therefore, for all interfaces  $x = x_{i+1/2}$ , we could set

$$k_{i+1/2} = \frac{h_{i+1/2}}{\int_{\bar{x}_i}^{\bar{x}_{i+1}} \frac{dy}{k(y)}}.$$

If  $x \mapsto k(x)$  satisfies  $k(x) = k^-$  on  $]x_i, x_{i+1/2}[$  and  $k(x) = k^+$  on  $]x_{i+1/2}, x_{i+1}[$ , we obtain  $k_{i+1/2} = \frac{h_{i+1/2}}{h_i/(2k^-) + h_{i+1}/(2k^+)}$ . More generally, suppose that  $k$  is  $C^1$  on both sides, with  $\lim_{x \rightarrow x_{i+1/2}^\pm} k = k^\pm$ . Thus, we also have

$$\begin{aligned} \frac{1}{h_{i+1/2}} \int_{\bar{x}_i}^{\bar{x}_{i+1}} \frac{dy}{k(y)} &= \frac{1}{h_{i+1/2}} \left( \int_{\bar{x}_i}^{\bar{x}_{i+1/2}} \frac{dy}{k(y)} + \int_{\bar{x}_{i+1/2}}^{\bar{x}_{i+1}} \frac{dy}{k(y)} \right) \\ &= \frac{h_i/k^- + h_{i+1}/k^+}{2h_{i+1/2}} \\ &\quad + \frac{1}{h_{i+1/2}} \int_{\bar{x}_i}^{\bar{x}_{i+1/2}} \left( \frac{1}{k(y)} - \frac{1}{k^-} \right) dy + \frac{1}{h_{i+1/2}} \int_{\bar{x}_{i+1/2}}^{\bar{x}_{i+1}} \left( \frac{1}{k(y)} - \frac{1}{k^+} \right) dy \\ &= \frac{h_i/k^- + h_{i+1}/k^+}{2h_{i+1/2}} + s_h, \quad \text{avec } |s_h| \leq Ch, \end{aligned}$$

which justifies defining  $k_{i+1/2}$  by [2.26].

### 2.3.3. Multidimensional problems

The generalization of finite volume methods to solve multidimensional elliptic problems leads to difficulties related to the grid's geometry. To illustrate those difficulties, we will limit ourselves to the two-dimensional case and cover the domain under study,  $\Omega \subset \mathbb{R}^2$ , with a collection of convex polygons  $\Omega = \bigcup_{j=1}^J T_j$ . These polygons represent the control volumes. For each polygon  $T_j$ , we select a “representative point”  $G_j$ . This point can be the barycenter of  $T_j$ , but there may be more practical choices. We assume that the grid is *conforming*: if for two control volumes  $T_i$  and  $T_j$  we have  $T_i \cap T_j \neq \emptyset$ , then the intersection is one of the sides of  $T_i$  and  $T_j$ . We still obtain the scheme by minimizing the formula obtained by

integrating the equation on the control volumes. Let us consider the simple Laplace problem with Dirichlet conditions

$$\lambda u - \Delta u = f \in L^2(\Omega), \quad u|_{\partial\Omega} = 0.$$

The finite volume scheme is inspired by the relations

$$\frac{\lambda}{|T_j|} \int_{T_j} u(y) dy - \frac{1}{|T_j|} \int_{\partial T_j} \nabla_x u(y) \cdot n_j(y) d\sigma(y) = \frac{1}{|T_j|} \int_{T_j} f(y) dy,$$

where  $n_j(y)$  denotes the outward unit normal vector at the point  $y \in \partial T_j$  and  $d\sigma$  is the Lebesgue measure on  $\partial T_j$ . We can define a simple approximation of  $\nabla_x u(y) \cdot n_j$  by imposing the grid a specific structure. Therefore, we say the grid is *admissible* if for all volumes  $T_j, T_k$  with a common edge  $\mathcal{A}_{jk}$ , the segment  $G_j G_k$  is orthogonal to  $\mathcal{A}_{jk}$ . In this case,

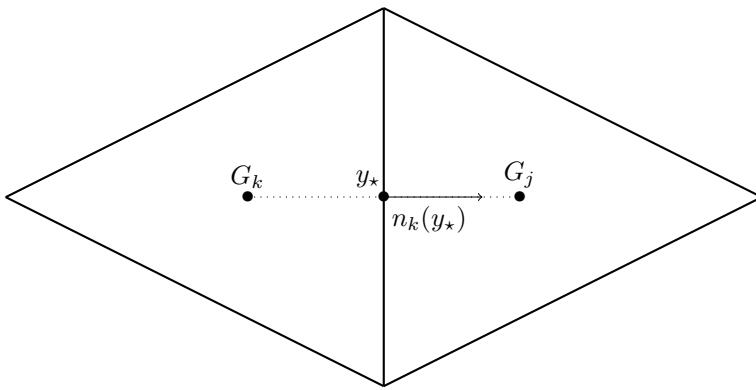
$$\nabla_u(y_*) \cdot n(y_*) \quad \text{is approximated by} \quad \frac{u(G_k) - u(G_j)}{|G_k G_j|}, \quad [2.27]$$

with  $G_j G_k \cap \mathcal{A}_{jk} = \{y_*\}$ . However, the structure assumption is a very strong restriction on the grid. It is satisfied for rectangular control volumes or a grid made up of equilateral triangles, where  $G_j$  is the barycenter of the volume  $T_j$ . For more general grids, this condition might be violated: as a result, the fluxes defined by [2.27] are not consistent, and the scheme does not converge to the solution of the continuous problem [FAI 92]. Moreover, if the Laplace problem is replaced by

$$-\nabla \cdot (k(x) \nabla u) = f \in L^2(\Omega), \quad u|_{\partial\Omega} = 0,$$

with  $k$  a matrix of  $M_N(\mathbb{R})$ , such that  $0 < \kappa |\xi|^2 \leq k(x)\xi \leq K |\xi|^2$  with  $\kappa, K > 0$ , this approximation is not enough: it only provides an approximation of the gradient of  $u$  in the direction of the normal vector  $n(y)$ , while we now need to know  $\nabla u$  in the transverse direction to evaluate  $\int_{\partial T_j} k \nabla u(y) \cdot n_j(y) d\sigma(y)$ .

These questions are strongly motivated by applications related to the simulation of flows in porous media. For those applications, methods based on finite difference or finite element approximations have reached certain limits, especially for very heterogeneous environments. The subject is an area of active research, especially since the developments of [COU 99, DOM 05], which have provided clues to manage those difficulties in constructing an effective approximation of  $\int_{\partial T} k \nabla u \cdot n(y) d\sigma(y)$  on control volume interfaces. The reference work [EYM 00] provides a more complete overview of these finite volume methods and their analysis.



**Figure 2.7.** Example of an admissible grid: the segment  $G_kG_j$  cuts the edge common to  $T_k$  and  $T_j$  at a right angle

## 2.4. Finite element approximations of elliptic equations

### 2.4.1. $\mathbb{P}_1$ approximation in one dimension

The finite element (FE) method uses the *variational formulation* of problem [2.2]: we have seen that [2.2] leads to the relation

$$\lambda \int_0^1 u(x)\varphi(x) dx + \int_0^1 k(x) \frac{d}{dx}u(x) \frac{d}{dx}\varphi(x) dx = \int_0^1 f(x)\varphi(x) dx, \quad [2.28]$$

which is satisfied for all  $\varphi \in C^1([0, 1])$ , such that  $\varphi(0) = 0 = \varphi(1)$ . The idea is to use a restricted class of such test functions  $\varphi$  that describe a finite-dimensional vector space; we thus seek an approximation of the solution  $u$  as an element of that space. We continue to use a partition of  $[0, 1]$

$$x_0 = 0 < x_1 < \dots < x_{I+1} = 1$$

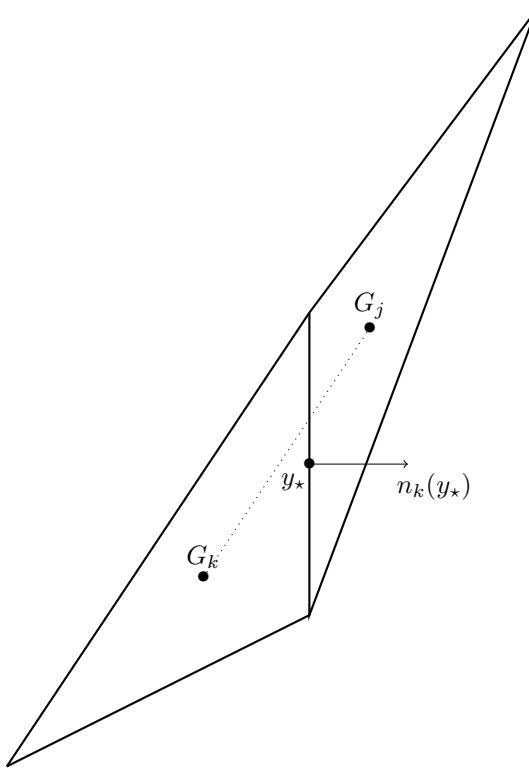
and we use the notation

$$h_{i+1/2} = x_{i+1} - x_i, \quad h = \max \{h_{i+1/2}, i \in \{0, \dots, I\}\}.$$

We now define a finite-dimensional space meant to approximate the “natural” functional space for [2.28]. This approximation space is given by

$$V_h = \{\varphi \in C^0([0, 1]), \quad \varphi(0) = 0 = \varphi(1), \quad \varphi|_{\mathcal{I}_j} \in \mathbb{P}_1 \text{ for all } j \in \{0, \dots, I\}\}$$

where  $\mathcal{I}_j = [x_j, x_{j+1}]$  (note that  $\bigcup_{j=0}^I \mathcal{I}_j = [0, 1]$ ) and  $\mathbb{P}_1$  denotes the set of polynomial of degree  $\leq 1$ .

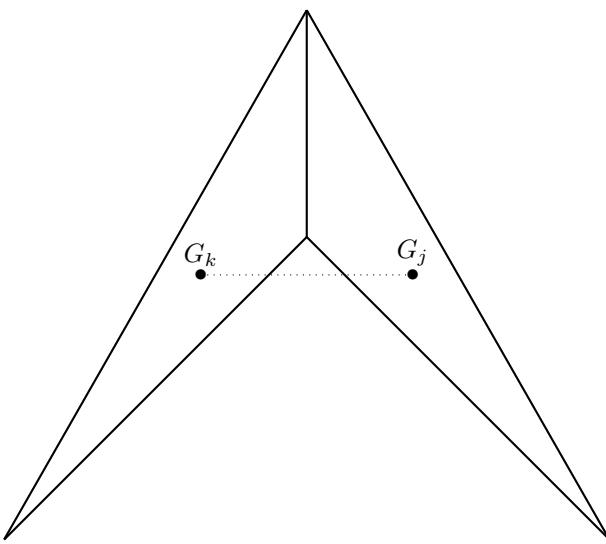


**Figure 2.8.** Example of an non admissible grid: the segment  $G_kG_j$  does not intersect the edge common to  $T_k$  and  $T_j$  at a right angle; the numerical unknowns contained in  $G_j$  and  $G_k$  are not enough to derive a discrete derivative in the direction of the normal vector  $n(y_*)$

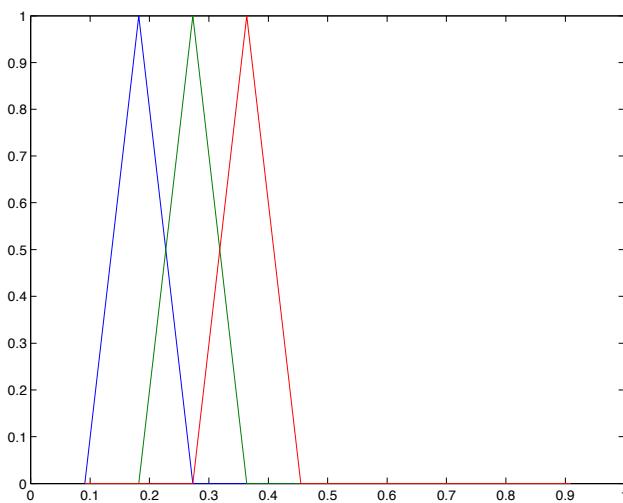
LEMMA 2.5.– The set  $V_h$  is a vector space of dimension  $I$  spanned by the functions

$$x \mapsto \varphi_j(x) = \begin{cases} \frac{x - x_{j-1}}{x_j - x_{j-1}} & \text{for } x_{j-1} \leq x \leq x_j, \\ \frac{x_{j+1} - x}{x_{j+1} - x_j} & \text{for } x_j \leq x \leq x_{j+1}, \\ 0 & \text{elsewhere} \end{cases}$$

with  $j \in \{1, \dots, I\}$ .



**Figure 2.9.** Example of an non admissible grid: the segment  $G_kG_j$  does not intersect the edge common to  $T_k$  and  $T_j$



**Figure 2.10.** Example of  $\mathbb{P}_1$  basis functions functions. For the color version of this figure, see [www.iste.co.uk/goudon/mathmodel.zip](http://www.iste.co.uk/goudon/mathmodel.zip)

PROOF.– The mapping

$$\Phi : V_h \longrightarrow \mathbb{R}^I,$$

$$u \longmapsto (u(x_1), \dots, u(x_I))$$

is linear and bijective. Indeed, on each interval  $\mathcal{I}_j$ ,  $j$  describing  $\{0, \dots, I\}$ , a function  $u \in V_h$  is of the form  $\alpha_j x + \beta_j$ . If  $\Phi(u) = 0$ , then it follows from  $u(0) = 0 = u(x_1)$  that  $\beta_0 = 0 = \alpha_0 = 0$ , and, next, that all coefficients  $\alpha_j$  and  $\beta_j$  are null; it proves that  $\Phi$  is injective. Then, given  $y \in \mathbb{R}^I$ , we construct the coefficients  $\alpha_j, \beta_j$ , such that  $u \in V_h$  satisfies  $\Phi(u) = y$ :  $\beta_0 = 0$ ,  $\alpha_0 = \frac{y_1}{x_1}$ ,  $\alpha_j = \frac{y_{j+1}-y_j}{x_{j+1}-x_j}$  and  $\beta_j = y_j - \frac{y_{j+1}-y_j}{x_{j+1}-x_j} x_j$ ... (using the convention  $y_0 = 0 = y_{I+1}$  to take into account the Dirichlet condition), in order to show that  $\Phi$  is surjective<sup>8</sup>. Now, the  $I$  functions  $\varphi_i \in V_h$  are linearly independent and therefore make up a basis of  $V_h$ .  $\square$

Note that [2.28] is still satisfied by the  $\% \phi_j$ s: we integrate over the segments  $[x_i, x_{i+1}]$  and conclude using the continuity of  $\varphi_i$  and  $x \mapsto k(x) \frac{d}{dx} u(x)$ . We thus seek to define an approximation to the solution of [2.2] of the form

$$u_h(x) = \sum_{j=1}^I u_j \varphi_j(x) \in V_h.$$

The numerical unknown  $U = (u_1, \dots, u_I)$  now corresponds to the coefficients of  $u_h$  in that expansion. We define them by requiring that

$$\lambda \int_0^1 u_h(x) \varphi_j(x) dx + \int_0^1 k(x) \frac{d}{dx} u_h(x) \frac{d}{dx} \varphi_j(x) dx = \int_0^1 f(x) \varphi_j(x) dx \quad [2.29]$$

is satisfied for all  $j \in \{1, \dots, I\}$ . This is equivalent to saying that the vector  $U$  is a solution to the linear system

$$AU = F, \quad F_i = \int_0^1 f(x) \varphi_i(x) dx,$$

$$A_{i,j} = \lambda \int_0^1 \varphi_i(x) \varphi_j(x) dx + \int_0^1 k(x) \frac{d}{dx} \varphi_i(x) \frac{d}{dx} \varphi_j(x) dx.$$

---

<sup>8</sup> In fact, this is equivalent to writing  $u(x) = \sum_{j=0}^I \mathbf{1}_{\mathcal{I}_j} \left( \frac{y_{j+1}-y_j}{x_{j+1}-x_j} (x - x_j) + y_j \right) = \sum_{j=1}^I y_j \varphi_j(x)$ .

The matrix of coefficients  $\int_0^1 \varphi_i(x)\varphi_j(x) dx$  is called the *mass matrix*, and the matrix of coefficients  $\int_0^1 \frac{d}{dx}\varphi_i(x)\frac{d}{dx}\varphi_j(x) dx$  is called the *rigidity matrix*. The matrix  $A$  is invertible, including for  $\lambda = 0$ , since for all  $\xi \in \mathbb{R}^I \setminus \{0\}$ , we have

$$\begin{aligned} A\xi \cdot \xi &= \lambda \int_0^1 \left( \sum_{i=1}^I \varphi_i(x)\xi_i \right)^2 dx + \int_0^1 k(x) \left( \sum_{i=1}^I \frac{d}{dx}\varphi_i(x)\xi_i \right)^2 dx \\ &\geq \kappa \int_0^1 \left( \sum_{i=1}^I \frac{d}{dx}\varphi_i(x)\xi_i \right)^2 dx \geq 0, \end{aligned} \quad [2.30]$$

and, in fact, this quantity is strictly positive (for any  $\xi \neq 0$ ). The key to demonstrating this property consists of noting that

$$\frac{d}{dx}\varphi_j(x) = \begin{cases} \frac{1}{x_j - x_{j-1}} & \text{for } x_{j-1} \leq x \leq x_j, \\ \frac{-1}{x_{j+1} - x_j} & \text{for } x_j \leq x \leq x_{j+1}, \\ 0 & \text{elsewhere} \end{cases}$$

and that the segments  $[0, x_1], [x_I, 1]$  only intersect the support of one of those basis functions. Let  $\xi \in \text{Ker}(A)$ . It follows from [2.30] that  $\sum_{i=1}^I \xi_i \frac{d}{dx}\varphi_i(x) = 0$  for almost all  $x \in [0, 1]$ . With the convention  $\xi_0 = 0 = \xi_{I+1}$ , we can rewrite this relation as

$$\sum_{i=0}^I (\xi_{i+1} - \xi_i) \frac{\mathbf{1}_{\mathcal{I}_i}(x)}{x_{i+1} - x_i} = 0,$$

which implies  $\xi_1 = \xi_2 = \dots = \xi_I = 0$ . Therefore, the functions  $\frac{d}{dx}\varphi_i$  are linearly independent. Moreover, note that [2.29] means that  $u_h$  is simply the projection of  $u$  on  $V_h$  for the scalar product

$$\langle u|v \rangle = \lambda \int_0^1 u(x)v(x) dx + \int_0^1 k(x) \frac{d}{dx}u(x) \frac{d}{dx}v(x) dx$$

because we have  $\langle u - u_h | \varphi_i \rangle = 0$  for all  $i \in \{1, \dots, I\}$ .

NOTE 2.7.– For a uniform grid with step  $h > 0$ , we have

$$\varphi_i(x) = \frac{x - (i-1)h}{h} \mathbf{1}_{(i-1)h \leq x \leq ih} + \frac{(i+1)h - x}{h} \mathbf{1}_{ih \leq x \leq (i+1)h}$$

and with a constant coefficient  $k(x) = k > 0$ , we find

$$A = \frac{\lambda h}{6} \begin{pmatrix} 4 & 1 & 0 & \cdots & 0 \\ 1 & -2 & 1 & 0 & \cdots \\ 0 & 1 & -2 & 1 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 1 & 4 & \end{pmatrix} - \frac{k}{h} \begin{pmatrix} -2 & 1 & 0 & \cdots & 0 \\ 1 & -2 & 1 & 0 & \cdots \\ 0 & 1 & -2 & 1 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 1 & -2 & \end{pmatrix}$$

In order to justify the validity of this approach, we begin by determining an interpolation result.

**PROPOSITION 2.5.** – Let  $g \in C^2([0, 1])$  verify  $g(0) = 0 = g(1)$ . We write  $\tilde{g}_h(x) = \sum_{i=1}^I g(x_i) \varphi_i(x)$ . Therefore, there exists a  $C > 0$ , such that

$$\sup_{x \in [0, 1]} |\tilde{g}_h(x) - g(x)| \leq C h^2, \quad \sup_{x \in [0, 1]} \left| \frac{d}{dx} \tilde{g}_h(x) - \frac{d}{dx} g(x) \right| \leq C h.$$

**PROOF.** – For all  $x_i \leq x < x_{i+1}$ , we have

$$\frac{d}{dx} \tilde{g}_h(x) = \frac{1}{h_{i+1/2}} (g(x_{i+1}) - g(x_i)) = \frac{1}{h_{i+1/2}} \int_{x_i}^{x_{i+1}} \frac{d}{dz} g(z) dz.$$

It follows that

$$\frac{d}{dx} \tilde{g}_h(x) - \frac{d}{dx} g(x) = \int_0^1 \left( \frac{d}{dx} g(x_i + \theta h_{i+1/2}) - \frac{d}{dx} g(x) \right) d\theta$$

whose absolute value is bounded above by  $2 \sup_{\xi} \left| \frac{d^2}{dx^2} g(\xi) \right| h$ . Similarly, we determine an upper bound for  $\tilde{g}_h - g$  with  $C h^2$ .  $\square$

Using lemma 2.2, we obtain

$$\begin{aligned} \|u_h - u\|_{L^\infty}^2 &\leq \frac{\lambda}{\kappa} \|u_h - u\|_{L^2} + \|u'_h - u'\|_{L^2}^2 \\ &\leq \frac{1}{\kappa} \left( \lambda \int_0^1 (u_h - u)^2(x) dx + \int_0^1 k(x)(u'_h - u')^2(x) dx \right) \\ &\leq \frac{1}{\kappa} \inf_{\varphi \in V_h} \left( \lambda \int_0^1 (\varphi - u)^2(x) dx + \int_0^1 k(x)(\varphi' - u')^2(x) dx \right) \end{aligned}$$

according to the interpretation of  $u_h$  by means of orthogonal projection

$$\leq \frac{1}{\kappa} \left( \lambda \int_0^1 (\tilde{u}_h - u)^2(x) dx + \int_0^1 k(x)(\tilde{u}'_h - u')^2(x) dx \right)$$

since  $\tilde{u}_h \in V_h$

$$\leq \frac{\lambda}{\kappa} \|\tilde{u}_h - u\|_{L^\infty}^2 + \frac{K}{\kappa} \|\tilde{u}'_h - u'\|_{L^\infty}^2 \leq \frac{\lambda + K}{\kappa} C h^2$$

owing to proposition 2.5.

**THEOREM 2.10.**— There exists a constant  $C > 0$ , such that the error between the approximated solution  $u_h$  and the solution  $u$  of [2.2] satisfies

$$\|u_h - u\|_{L^\infty} \leq \|u'_h - u'\|_{L^2} \leq C h.$$

We note that the resulting estimate seems less favorable than that obtained from the finite difference discretization, even though the matrices corresponding to the two methods are very similar. However, it should be noted in this interpretation that the norm required on the solution to prove the consistency estimate is also different:  $\sup_\xi |\frac{d^2}{dx^2} g(\xi)|$  instead of  $\sup_\xi |\frac{d^4}{dx^4} g(\xi)|$ . Finally, the finite element method favors variational formulations, and by using the underlying functional spaces, it is possible to obtain more fine-grained estimations: the  $L^\infty$  framework is not well adapted to the derivation of error estimates for the Finite Element method, in contrast to the  $L^2$  framework.

#### 2.4.2. $\mathbb{P}_2$ approximations in one dimension

In order to increase the order of the approximation, the idea is to use higher-order polynomials. For example, with second-order polynomials, the basis functions are obtained using appropriate transformations of the functions

$$\psi_0(x) = 2(x - 1/2)(x - 1),$$

$$\psi_{1/2}(x) = -4x(x - 1),$$

$$\psi_1(x) = 2x(x - 1/2),$$

which are shown in Figure 2.12. From a conceptual point of view, modifying the basis functions in this way does not present new difficulties. Instead, its implementation is somewhat more delicate: we do not have an easy expression for the matrix of finite elements, as discussed in note 2.7 for the elements  $\mathbb{P}_1$ .

We use a more systematic angle to construct the matrix, which can be adapted to higher dimensions. We let  $\psi_i$  denote the basis functions. The function  $\psi_i$  is piecewise polynomial, in this case of degree 2. It is equal to 1 at the grid nodes  $i$  and to zero elsewhere. To obtain the element  $A_{ij}$  of the finite element matrix, we must evaluate integrals of the form

$$\sum_{\ell} \int_{I_{\ell}} k(x) \psi'_i(x) \psi'_j(x) dx$$

where the sum is taken on all segments  $I_{\ell}$  that have end nodes  $i$  and  $j$ . In order to construct the matrix, it is necessary to have a numbering of the elements as well as a correspondence table to identify what element a given node belongs to and of the grid nodes.

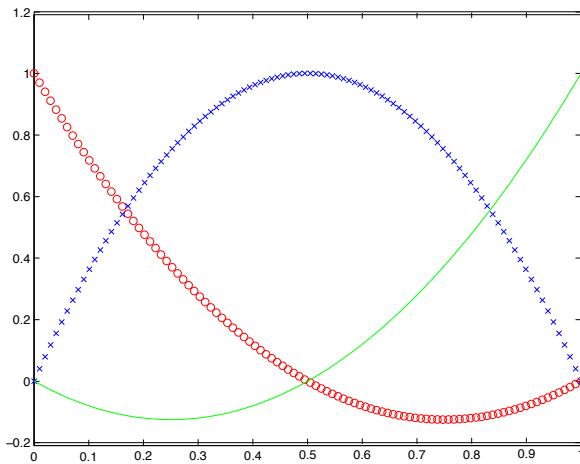


**Figure 2.11.**  $\mathbb{P}_2$  Discretization nodes

For elements  $\mathbb{P}_2$  in dimension 1, we can proceed as follows. We provide a discretization of the segment  $[0, 1]$ , using the points  $x_0 = 0 < x_1 < \dots < x_J < x_{J+1} = 1$ . The basis functions also involve the barycenters  $x_{j+1/2} = (x_j + x_{j+1})/2$ ,  $j \in \{0, \dots, J\}$  (see Figure 2.11). The unknowns and the basis functions are thus attached to the nodes  $(x_j, x_{j+1/2}, x_{j+1})$ : we denote as  $U = (u_0, u_{1/2}, u_1, \dots, u_J, u_{J+1/2}, u_{J+1})$  the vector of numerical unknowns corresponding to the nodes  $Y = (x_0, x_{1/2}, x_1, \dots, x_J, x_{J+1/2}, x_{J+1})$ . We thus have  $(J + 1)$  segments, each involving three nodes; the vectors  $U$  and  $Y$  have  $(2J + 3)$  coordinates ( $(J + 2)$  interfaces and  $(J + 1)$  barycenters). To construct the matrix, we sweep the segments  $k \in \{1, \dots, J + 1\}$ , and adjust the contributions of the nodes indexed as  $2k + i - 2$ ,  $2k + j - 2$ , where  $i, j$  take their value in  $\{1, 2, 3\}$ .

For a problem, with constant coefficient we must evaluate the integrals

$$\begin{aligned} \int_0^1 \psi'_0(x) \psi'_0(x) dx, \quad \int_0^1 \psi'_0(x) \psi'_{1/2}(x) dx, \quad \int_0^1 \psi'_0(x) \psi'_1(x) dx, \\ \int_0^1 \psi'_{1/2}(x) \psi'_0(x) dx, \quad \int_0^1 \psi'_{1/2}(x) \psi'_{1/2}(x) dx, \quad \int_0^1 \psi'_{1/2}(x) \psi'_1(x) dx, \\ \int_0^1 \psi'_1(x) \psi'_0(x) dx, \quad \int_0^1 \psi'_1(x) \psi'_{1/2}(x) dx, \quad \int_0^1 \psi'_1(x) \psi'_1(x) dx. \end{aligned}$$



**Figure 2.12.**  $\mathbb{P}_2$  basis functions. For the color version of this figure, see [www.iste.co.uk/goudon/mathmodel.zip](http://www.iste.co.uk/goudon/mathmodel.zip)

We find

$$M_{\text{loc}} = \frac{1}{3} \begin{pmatrix} 7 & -8 & 1 \\ -8 & 16 & -8 \\ 1 & -8 & 7 \end{pmatrix}.$$

The construction of the Laplace matrix is as follows:

Start with  $A = 0$ .

For  $k = 1$  at  $k = (J + 1)$ ,

for  $i = 1$  at  $i = 3$  and  $j = 1$  at  $j = 3$

$I = 2k + i - 2$ ,  $J = 2k + j - 2$ ,

$A(I, J) = A(I, J) + M_{\text{loc}}(i, j)$ .

Bring it to the scale<sup>9</sup>  $A = \frac{1}{h}A$

Take into account the boundary conditions.

---

<sup>9</sup> Here  $h > 0$  designates the space step, which we assume is uniform ( $x_{j+1} - x_j = x_{j+1/2} - x_{j-1/2} = h$ ). The basis functions are defined on the grid by the relation  $\psi_h(y) = \psi(\frac{y-x_j}{h})$  with  $y \in [x_j, x_{j+1}]$ . The factor  $1/h$  comes from the combination of the fact that  $\psi'_h(y) = \frac{1}{h}\psi'(\frac{y-x_j}{h})$  and the fact that we go from integrals on  $[x_j, x_{j+1}]$  to integrals on  $[0, 1]$  using the change of variable  $y = x_j + xh$ .

We use a similar rationale to construct the data  $F$ , calculating the integrals  $\int f(x)\psi_i(x) dx$ . In this regard, these integrals cannot, in general, be computed exactly. It is necessary to use approximated formulas; however, they must themselves be chosen so as to guarantee a high quality approximation (it would be meaningless to use  $\mathbb{P}_2$  elements if the integrals that define the matrix system are evaluated using a simple rectangle method!). These questions are discussed in Appendix 2.

### 2.4.3. Finite element methods, extension to higher dimensions

In order to generalize the method to higher dimensions, we proceed as follows. In fact, we can develop an abstract framework that allows us to analyze finite element numerical schemes. We have two working functional spaces  $V$  and  $H$ . They are real Hilbert spaces with the respective scalar products  $\langle \cdot | \cdot \rangle_V$  and  $\langle \cdot | \cdot \rangle_H$  and the associated norms, denoted as  $\| \cdot \|_V$  and  $| \cdot |_H$ . We assume that  $V \subset H$ , given that the embedding is continuous and dense. Moreover, using the Riesz theorem, we can identify  $H$  and its dual space  $H'$  (we can therefore no longer identify  $V$  and  $V'$ ; see [BRÉ 05]). Let  $f \in H$  and  $a$  be a coercive and continuous bilinear form on  $V \times V$ . By the Lax–Milgram theorem, there exists a unique element  $u \in V$ , such that for all  $v \in V$ , the relation  $a(u, v) = \langle f | v \rangle_H$  holds. The typical example consists of taking  $V = H_0^1(\Omega)$ ,  $H = L^2(\Omega)$  and  $V' = H^{-1}(\Omega)$ , and the functional framework effectively allows us to deal with elliptic problems [2.1].

The construction of the approximation relies on the fact that we have a sequence of subspaces  $V_h \subset V$ , with finite dimensions  $N_h = \dim(V_h) < \infty$ . Then, for any fixed  $h$ , there exists a unique element  $u_h \in V_h$ , such that for all  $v \in V_h$ , the relation  $a(u_h, v) = \langle f | v \rangle_H$  holds. In practice, determining  $u_h$  involves solving a linear system. Indeed, let  $\{e_1, \dots, e_{N_h}\}$  be a basis for the subspace  $V_h$ , which can be assumed as orthonormal in  $V$  (so  $\langle e_i | e_j \rangle_V = \delta_{ij}$ ). Thus, we decompose  $u_h$  in this basis  $u_h = \sum_{j=1}^{N_h} u_j e_j$  and the coefficients  $u_1, \dots, u_{N_h}$  are characterized by

$$a\left(\sum_{j=1}^{N_h} u_j e_j, e_i\right) = \sum_{j=1}^{N_h} u_j a(e_j, e_i) = \langle f | e_i \rangle_H$$

for all  $i \in \{1, \dots, N_h\}$  (we have taken  $v = e_i$  for each basis vector in the variational formulation). This problem can be expressed in the form  $A_h U = b$ , where  $b_i = \langle f | e_i \rangle_H$ ,  $U = (u_1, \dots, u_{N_h})$ , and the coefficients of the matrix  $A_h$  are defined by  $(A_h)_{i,j} = a(e_j, e_i)$ .

**PROPOSITION 2.6.–** The matrix  $A_h$  is invertible.

PROOF.– This is a result of the coerciveness of the bilinear form  $a$ . Indeed, for all  $\xi \in \mathbb{R}^{N_h}$ , we have

$$A_h \xi \cdot \xi = \sum_{i,j=1}^{N_h} a(e_j, e_i) \xi_j \xi_i = a\left(\sum_{j=1}^{N_h} \xi_j e_j, \sum_{i=1}^{N_h} \xi_i e_i\right) \geq \kappa \left\| \sum_{j=1}^{N_h} \xi_j e_j \right\|_V^2 = \kappa |\xi|^2.$$

This inequality proves that  $\text{Ker}(A_h) = \{0\}$ .  $\square$

NOTE 2.8.– In the case where the form  $a$  is *symmetric*,  $u$  is characterized as a solution to the following minimization problem [GOU 11, corollary 5.27]:

$$\frac{1}{2}a(u, u) - \langle f | u \rangle_H = \inf \left\{ \frac{1}{2}a(v, v) - \langle f | v \rangle_H, v \in V \right\}.$$

Likewise, the approximate solution  $u_h$  is a solution to the same minimization problem set on  $V_h$ , which defines  $U = (u_1, \dots, u_{N_h})$  as a minimizer on  $\mathbb{R}^{N_h}$  of  $x \mapsto \frac{1}{2}A_h x \cdot x - b \cdot x$ . In this symmetric case,  $a$  defines a inner product on  $V$  and  $u_h$  can be interpreted as the orthogonal projection of  $u$  for that inner product associated with  $a$ , on the subspace  $V_h$ . Thus, the matrix  $A_h$  is symmetric and positive definite.

By construction, we directly obtain a stability property.

LEMMA 2.6.– The sequence of approximated solutions  $(u_h)_{h>0}$  is bounded in  $V$ , independently of  $h$ .

PROOF.– We use  $v = u_h \in V_h$  in the variational formulation of the approximated problem. The continuity and coerciveness of  $a$  thus lead to  $\kappa \|u_h\|_V^2 \leq a(u_h, u_h) = \langle f | u_h \rangle_H \leq \|f\|_H \|u_h\|_V$ , so finally  $\|u_h\|_V \leq \|f\|_H / \kappa$  (like the estimation satisfied by the solution  $u$  of the original problem; see theorem 2.5).  $\square$

The consistency of the method involves constructing a “good” approximation space  $V_h$ . Thus, we will be able to estimate  $\|u - u_h\|_V$ : the error tends to 0 when  $h \rightarrow 0$  if the spaces  $V_h$  “resemble  $V$ ” when  $h \rightarrow 0$  and  $\lim_{h \rightarrow 0} \dim(V_h) = \infty$ .

DEFINITION 2.4.– We say that the approximation method converges if we have  $\lim_{h \rightarrow 0} \|u - u_h\|_V = 0$ . The method is of order  $k$  if there exists  $C > 0$ , such that  $\|u - u_h\|_V \leq Ch^k$ .

The key tool for analyzing these methods is

LEMMA 2.7 (Céa’s lemma).– There exists a constant  $C > 0$  that depends only on the bilinear form  $a$ , such that

$$\|u - u_h\|_V \leq C \inf \{ \|u - v_h\|_V, v_h \in V_h \} = C \|u - P_h u\|_V = C \text{dist}(u, V_h)$$

where  $P_h u$  denotes the orthogonal projection of  $u$  on  $V_h$  (for the inner product  $\langle \cdot | \cdot \rangle_V$ ).

PROOF.– Let  $v_h \in V_h$ . Then,  $u_h - v_h \in V_h \subset V$ , so that

$$\begin{aligned} a(u - u_h, u_h - v_h) &= a(u, u_h - v_h) - a(u_h, u_h - v_h) \\ &= \langle f | u_h - v_h \rangle - \langle f | u_h - v_h \rangle = 0 \\ &= a(u - u_h, u_h - u) + a(u - u_h, u - v_h). \end{aligned}$$

It follows that

$$\kappa \|u - u_h\|_V^2 \leq a(u - u_h, u - u_h) = a(u - u_h, u - v_h) \leq K \|u - u_h\|_V \|u - v_h\|_V.$$

We conclude that  $C = K/\kappa$  by taking the infimum on  $v_h \in V_h$ . This result indeed expresses a consistency property; the error  $u - u_h$  is bounded above by  $C \times$ , the distance from  $u$  to the point  $V_h$  that is closest to  $u$ . Furthermore, note that the property  $V_h \subset V$  is essential in this context; we say that the approximation space is *conformal*.

When  $a$  is symmetric, we can improve the estimate because in this case

$$\begin{aligned} a(u - v_h, u - v_h) &= a(u - u_h, u - u_h) + a(u_h - v_h, u_h - v_h) \\ &\quad + 2 \underbrace{a(u - u_h, u_h - v_h)}_{=0 \text{ since } u_h - v_h \in V_h \subset V} \\ &\geq a(u - u_h, u - u_h). \end{aligned}$$

This proves that  $a(u - u_h, u - u_h) = \inf\{a(u - v_h, u - v_h), v_h \in V_h\}$ , and therefore we are led to the estimate  $\alpha \|u - u_h\|_V^2 \leq K \|u - v_h\|_V^2$  for all  $v_h \in V_h$ . In fact, in the symmetric case, as mentioned above,  $u_h$  is the orthogonal projection of  $u$  on  $V_h$  for the inner product associated with  $a$ .  $\square$

The convergence of the method is therefore ensured by the following statement.

**THEOREM 2.11.–** We assume there exists

- a subspace  $\mathcal{V}$ , which is dense in  $V$ ;
- a mapping  $r_h : \mathcal{V} \rightarrow V_h$ , such that for all  $v \in \mathcal{V}$ , we have  $\lim_{h \rightarrow 0} \|v - r_h(v)\|_V = 0$ .

Then, the variational approximation method converges.

PROOF.– Let  $u \in V$  and  $\varepsilon > 0$ . By density, there exists a  $v_\varepsilon \in \mathcal{V}$ , such that  $\|u - v_\varepsilon\|_V \leq \varepsilon$ . Moreover, since  $\varepsilon$  is fixed, there exists a  $h(\varepsilon) > 0$ , such that if  $0 < h < h(\varepsilon)$ , then  $\|v_\varepsilon - r_h(v_\varepsilon)\| \leq \varepsilon$ . It follows that

$$\begin{aligned}\|u - u_h\|_V &\leq C \inf \{\|u - v_h\|_V, v_h \in V_h\} \leq C \|u - r_h(v_\varepsilon)\|_V \\ &\leq C (\|u - v_\varepsilon\|_V + \|v_\varepsilon - r_h(v_\varepsilon)\|_V) \leq 2C\varepsilon\end{aligned}$$

given that  $0 < h < h(\varepsilon)$ .  $\square$

Finally, we introduce the adjoint form

$$a_\star(u, v) = a(v, u),$$

which also satisfies the assumptions of the Lax–Milgram theorem. As a result, for all  $\phi \in H$ , there exists a unique element  $v_\phi \in V$ , which, for all  $v \in V$  satisfies  $a_\star(v_\phi, v) = \langle \phi | v \rangle_H$ .

LEMMA 2.8 (Aubin–Nitsche lemma).– The following relation is satisfied:

$$\|u - u_h\|_H \leq M \sup_{\phi \in H \setminus \{0\}} \left\{ \frac{1}{\|\phi\|_H} \inf_{v_h \in V_h} \|v_\phi - v_h\| \right\} \|u - u_h\|_V$$

PROOF.– We have

$$\begin{aligned}\|u - u_h\|_H &= \sup_{\phi \in H \setminus \{0\}} \left\{ \frac{\langle u - u_h, \phi \rangle_H}{\|\phi\|_H} \right\} \\ &= \sup_{\phi \in H \setminus \{0\}} \left\{ \frac{a_\star(v_\phi, u - u_h)}{\|\phi\|_H} \right\} \quad \text{since } u - u_h \in V \\ &= \sup_{\phi \in H \setminus \{0\}} \left\{ \frac{a(u - u_h, v_\phi)}{\|\phi\|_H} \right\} \\ &= \sup_{\phi \in H \setminus \{0\}} \left\{ \frac{a(u - u_h, v_\phi - v_h)}{\|\phi\|_H} \right\} \\ &\quad \text{for all } v_h \in V_h \subset V \text{ since } a(u, v_h) = a(u_h, v_h) \\ &\leq M \sup_{\phi \in H \setminus \{0\}} \left\{ \frac{\|u - u_h\|_V \|v_\phi - v_h\|_V}{\|\phi\|_H} \right\} \quad \text{by continuity of } a.\end{aligned}$$

This relation applies for all  $v_h \in V_h$ . In particular, it applies for  $P_h v_\phi$ , the orthogonal projection of  $v_\phi$  onto  $V_h$ , which satisfies

$$\|v_\phi - P_h v_\phi\|_V = \inf \{\|v_\phi - v_h\|_V, v_h \in V_h\}.$$

$\square$

This statement allows us to improve the error estimates, at the price of considering a weaker norm (a “small”  $L^2$  norm instead of the “large”  $H^1$  norm). Indeed, if we manage to show that  $\|u - u_h\|_V \leq C\|f\|_H h$ , then an estimate of the same kind applies for the adjoint problem, for which we can also define a variational approximation:  $\|v_\phi - v_{\phi,h}\|_V \leq C\|\phi\|_H h$ . Now, we have  $\inf \{\|v_\phi - v_h\|_V, v_h \in V_h\} \leq \|v_\phi - v_{\phi,h}\|_V$ . It follows that

$$\|u - u_h\|_H \leq M \sup_{\phi \in H \setminus \{0\}} \left\{ \frac{\|v_\phi - v_{\phi,h}\|_V}{\|\phi\|_H} \right\} \|u - u_h\|_V \leq C\|f\|_H h^2.$$

With the finite element methods, we have two ways for reducing the error: making the grid more fine-grained or increasing the degree of the polynomials which define the approximation locally. The effectiveness of these two “cursors” can depend on the considered; application see, for example, the comments in [AIN 04].

## 2.5. Numerical comparison of FD, FV and FE methods

We present a series of numerical tests in which the performance of the methods is evaluated in cases where we know an explicit solution of the problem

$$\begin{cases} -\frac{d}{dx} \left( k(x) \frac{du}{dx} \right) = f(x), & x \in ]0, 1[, \\ u(0) = 0 = u(1). \end{cases} \quad [2.31]$$

– *Example 1: constant coefficient.*  $k(x) = 1$ ,  $f(x) = e^x$ . The exact solution is

$$u_{\text{exact}}(x) = (e-1)x - e^x + e.$$

– *Example 2: variable coefficient.*  $k(x) = 1/\cosh(\sin(\pi x))$ ,  $f(x) = \pi^2 \sin(\pi x)$ . The exact solution is

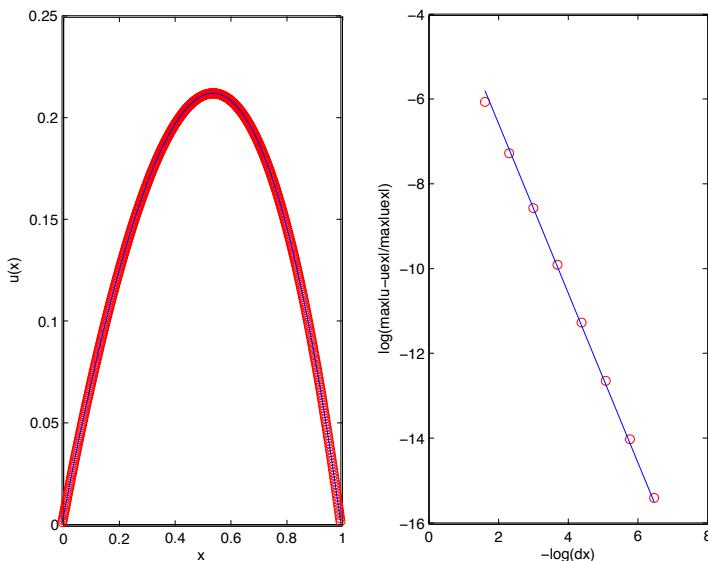
$$u_{\text{exact}}(x) = \sinh(\sin(\pi x)).$$

– *Example 3: discontinuous coefficient.*  $k(x) = \mathbf{1}_{0 \leq x \leq 1/4} + \mathbf{1}_{3/4 \leq x \leq 1} + 2\mathbf{1}_{1/4 \leq x \leq 3/4}$ ,  $f(x) = 1$ . The exact solution is

$$u_{\text{exact}}(x) = \frac{x(1-x)}{2} (\mathbf{1}_{0 \leq x \leq 1/4} + \mathbf{1}_{3/4 \leq x \leq 1}) + \left( \frac{3}{64} \frac{x(1-x)}{4} \right) \mathbf{1}_{1/4 \leq x \leq 3/4}.$$

We compare the methods on grids with constant steps and grids constructed with randomly chosen steps, following a uniform law (except for the boundary points,  $x_i$

are taken at random and ordered according to a uniform law on  $[0, 1]$ ). For all methods, in the case of a uniform grid with continuous coefficients (Examples 1 and 2), we observe convergence of order 2 (see Figures 2.13 and 2.14). This is also the case for a random grid (see Figures 2.15 and 2.18). With the  $\mathbb{P}_2$  method, we improve the order of convergence, as shown in Figure 2.20. For discontinuous data, as in Example 3, the solution is no longer  $C^1$  and the estimates are no longer justified by the developments presented in the consistency analysis. In practice, we observe a decrease in the order of convergence (see Figures 2.16, 2.17 and 2.19). This decrease is compensated, especially on a uniform grid, by defining the fluxes using the harmonic mean of the coefficients on both sides of the interface (see Figures 2.16 and 2.17).



**Figure 2.13.** Example 1: numerical and exact solutions; error curve with  $L^\infty$  norm for the FD method on a regular grid (log-log scale, line with slope 2). For the color version of this figure, see [www.iste.co.uk/goudon/mathmodel.zip](http://www.iste.co.uk/goudon/mathmodel.zip)

## 2.6. Spectral methods

We present another method for solving problem [2.2], which is still understood as a projection on a finite-dimensional space. We limit ourselves in this presentation to the case where  $\lambda = 0$ . This method uses the existence of Hilbertian bases of  $H_0^1([0, 1])$ ,

the functional space that we identified for the analysis of the problem. Let  $\{e_n, n \in \mathbb{N}\}$  be a family of functions on  $[0, 1]$ , such that

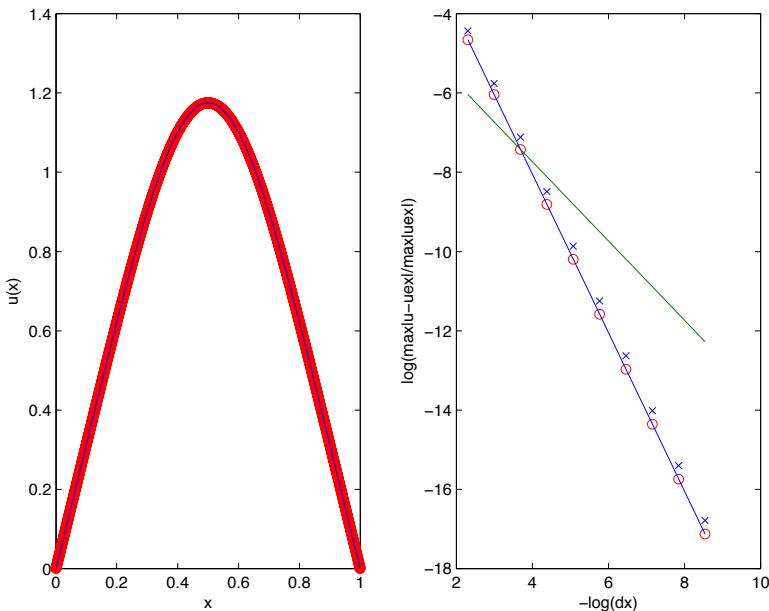
$$e_n \in H^1(]0, 1[), \quad e_n(0) = e_n(1),$$

$$\int_0^1 e_n(x)e_m(x) dx + \int_0^1 \frac{d}{dx}e_n(x) \frac{d}{dx}e_m(x) dx = \delta_{nm},$$

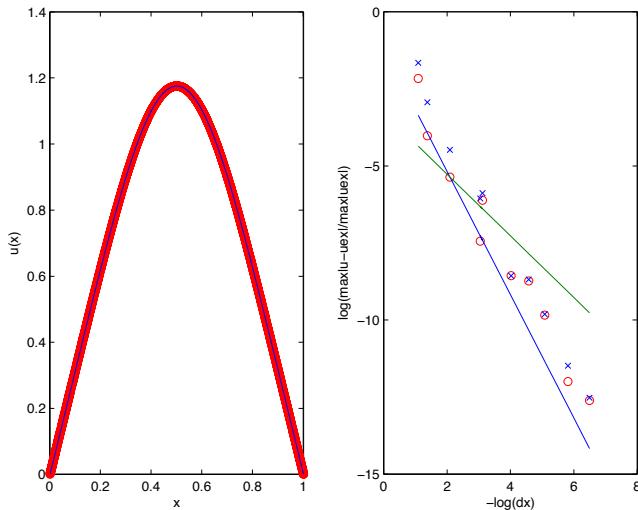
and for all  $f \in H_0^1(]0, 1[)$ , we have

$$\lim_{N \rightarrow \infty} \left\| f(x) - \sum_{n=0}^N \eta_n e_n(x) \right\|_{H_0^1} = 0,$$

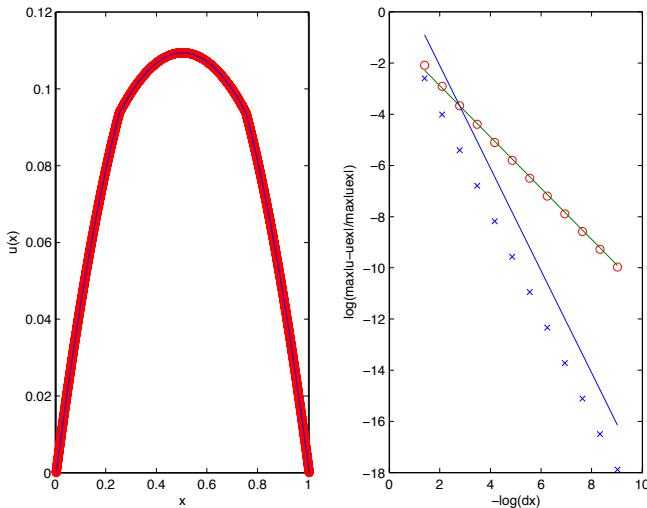
$$\eta_n = \int_0^1 \left( f(x)e_n(x) + \frac{d}{dx}f(x) \frac{d}{dx}e_n(x) \right) dx.$$



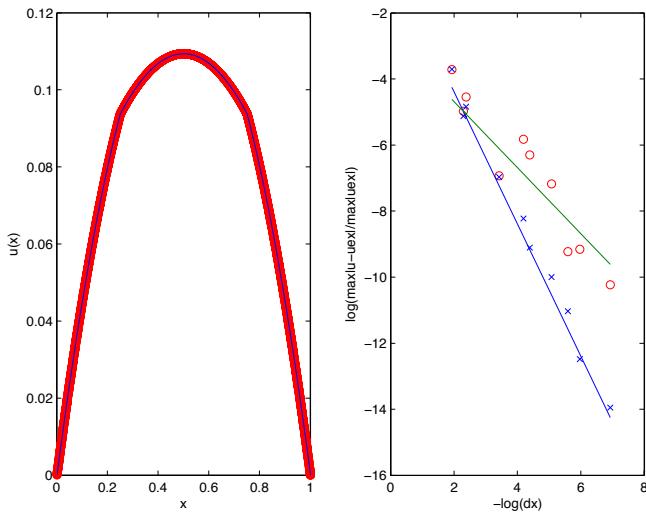
**Figure 2.14.** Example 2: FV, FD numerical and exact solutions; error curve for  $L^\infty$  norm for FD ( $x$ ) and FV ( $o$ ) methods on a regular grid (log–log scale, line with slope 2). For the color version of this figure, see [www.iste.co.uk/goudon/mathmodel.zip](http://www.iste.co.uk/goudon/mathmodel.zip)



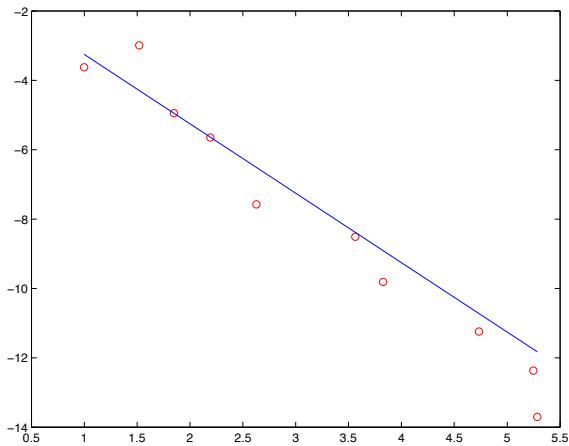
**Figure 2.15.** Example 2: exact, numerical FV and modified FV solutions; error curve in  $L^\infty$  norm for FV (o) and modified FV (x) on a random grid (log–log scale, lines with slope 1 [green] and 2 [blue]). For the color version of this figure, see [www.iste.co.uk/goudon/mathmodel.zip](http://www.iste.co.uk/goudon/mathmodel.zip)



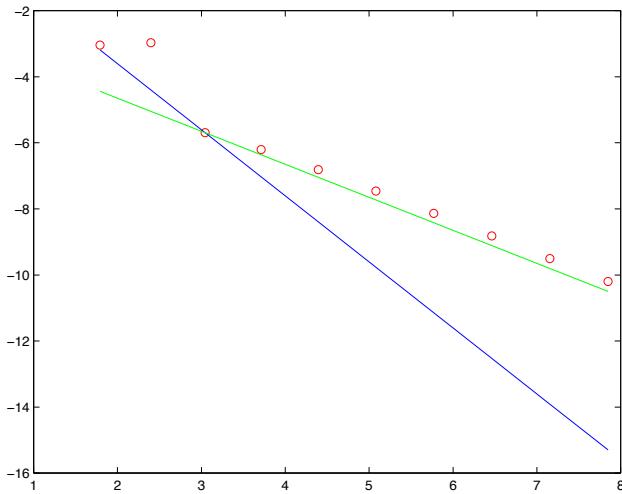
**Figure 2.16.** Example 3: exact, numerical FV and modified FV solutions; error curve in  $L^\infty$  norm for FV (o) and modified FV (x) on a regular grid (log–log scale, lines with slope 1 [green] and 2 [blue]). For the color version of this figure, see [www.iste.co.uk/goudon/mathmodel.zip](http://www.iste.co.uk/goudon/mathmodel.zip)



**Figure 2.17.** Example 3: exact, numerical FV and modified FV solutions; error curve in  $L^\infty$  norm for FV (o) and modified FV (x) on a regular grid (log–log scale, lines with slope 1 [green] and 2 [blue]). For the color version of this figure, see [www.iste.co.uk/goudon/mathmodel.zip](http://www.iste.co.uk/goudon/mathmodel.zip)



**Figure 2.18.** Example 2: error curve in  $L^\infty$  norm for the FE method on a random grid (log–log scale, line with slope 2). For the color version of this figure, see [www.iste.co.uk/goudon/mathmodel.zip](http://www.iste.co.uk/goudon/mathmodel.zip)



**Figure 2.19.** Example 3: error curve in  $L^\infty$  norm for the FE method on a regular grid (log–log scale, lines with slopes 1 [green] and 2 [blue]). For the color version of this figure, see [www.iste.co.uk/goudon/mathmodel.zip](http://www.iste.co.uk/goudon/mathmodel.zip)

We construct an approximation of the form  $u^{(N)}(x) = \sum_{j=0}^N u_j^{(N)} e_j(x)$  by defining the coefficients  $u_j^{(N)}$  as solutions to the linear system

$$\sum_{j=0}^N \int_0^1 k(x) u_j^{(N)} \frac{d}{dx} e_j(x) \frac{d}{dx} e_k(x) dx = \int_0^1 f(x) e_k(x) dx \text{ for all } k \in \{1, \dots, N\}.$$

Therefore,  $u^{(N)}$  is the projection of  $u$  onto  $V^{(N)} = \text{Span}\{x \mapsto e_n(x), n \in \{1, \dots, N\}\}$  for the scalar product defined by

$$a(u, v) = \int_0^1 k(x) \frac{d}{dx} u(x) \frac{d}{dx} v(x) dx.$$

Note that  $(u, v) \mapsto a(u, v)$  is well defined on the set  $\mathcal{C} = \{u \in C^1([0, L]), u(0) = 0 = u(L)\}$ ; it is a symmetric form on that set. Moreover, the quadratic form  $u \mapsto a(u, u)$  is positive definite on  $\mathcal{C}$ . We therefore have pre-Hilbertian space. Finally,  $V^{(N)}$  is a finite-dimensional subspace, so it is complete for the norm defined by  $a$ . We have  $a(u, e_k) = \int_0^1 f(x) e_k(x) dx$ , so  $a(u^{(N)} - u, v) = 0$  for all  $v \in V^{(N)}$ , which, indeed, characterizes  $u^{(N)}$  as the

orthogonal projection of  $u$  on  $V^{(N)}$  for  $a(\cdot, \cdot)$ . The value of this approach is shown in the following theorem.

**THEOREM 2.12.**— For all  $f \in L^2([0, 1])$ , when  $N \rightarrow \infty$ ,  $u^{(N)}$  converges to  $u$ , the solution of [2.2] in  $H_0^1([0, 1])$ .

**PROOF.**— By definition, we have

$$a(u^{(N)} - u, u^{(N)} - u) = \inf_{v \in V^{(N)}} a(v - u, v - u) \leq a(\phi - u, \phi - u) \leq K \|\phi' - u'\|_{L^2},$$

which can be dominated by  $\|\phi - u\|_{H_0^1}$  for all  $\phi \in V^{(N)}$ . In particular, this is true for  $x \mapsto \phi(x) = \sum_{n=1}^N u_n e_n(x)$ , the element of  $V^{(N)}$  whose  $N$  coefficients are given by the Fourier coefficients of  $u$  in basis  $\{e_n, n \in \mathbb{N}\}$ . In other words,  $\phi$  is the orthogonal projection of  $u$  onto  $V^{(N)}$  for the inner product  $H_0^1$ , and no longer for the product associated with the form  $a$ . We can infer that  $\lim_{N \rightarrow \infty} a(u^{(N)} - u, u^{(N)} - u) = 0$ . The relations

$$\kappa \left\| \frac{d}{dx} u^{(N)} - \frac{d}{dx} u \right\|_{L^2}^2 \leq a(u^{(N)} - u, u^{(N)} - u)$$

and

$$|u^{(N)}(x) - u(x)| = \left| \int_0^x \left( \frac{d}{dx} u^{(N)}(y) - \frac{d}{dx} u(y) \right) dy \right| \leq \left\| \frac{d}{dx} u^{(N)} - \frac{d}{dx} u \right\|_{L^2}$$

allow us to conclude that  $u^{(N)}$  converges uniformly to  $u$  on  $[0, 1]$  and in  $H_0^1([0, 1])$ . (For higher dimensions, we can only show convergence in  $H_0^1(\Omega)$  using this argument.) Note that the convergence speed is controlled by the approximation of  $u$  by its  $N$ th order Fourier expansion.  $\square$

One particular case, which is adapted to problem [2.2], consists of taking

$$e_n(x) = \sin(n\pi x).$$

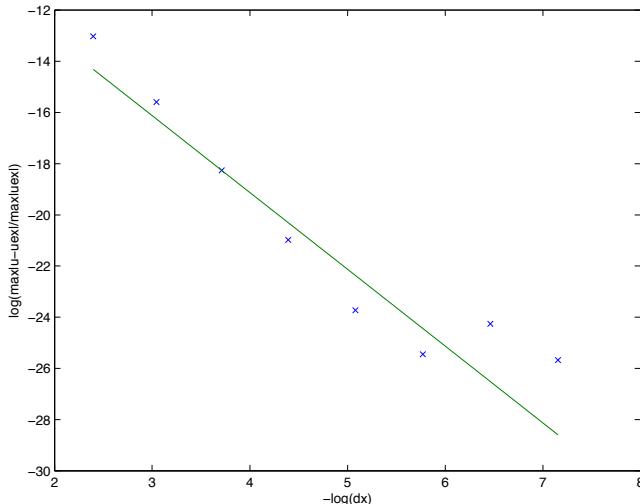
To ensure that these functions form a basis of  $L^2([0, 1])$ , we extend the functions of  $L^2([0, 1])$  on  $] -1, 1[$  by imparity, then on all of  $\mathbb{R}$  by 2-periodicity. We thus use the fact that  $\{\frac{1}{2}e^{i\pi nx}, n \in \mathbb{Z}\}$  is an hilbertian basis of  $L_\#^2(-1, 1)$  (see [GOU 11, section 5.4]). The Fourier expansion of an odd function  $f \in L_\#^2(-1, 1)$  is, in fact, written as

$$f(t) = \sum_{n=0}^{\infty} f_n \sin(n\pi t), \quad f_n = 2 \int_0^1 f(x) \sin(n\pi x) dx.$$

We have seen elsewhere that these functions  $e_n$  correspond to eigenfunctions in  $H_0^1([0, 1])$  of the operator  $-\frac{d^2}{dx^2}$ . These comments motivate the search for an

approximated solution to the problem [2.2] in the form of a trigonometric polynomial  $u^{(N)}(x) = \sum_{j=1}^N u_j \sin(j\pi x)$ . The numerical unknown  $\mathbb{U}^{(N)} = (u_1^{(N)}, \dots, u_N^{(N)}) \in \mathbb{R}^N$  is obtained as a solution to the linear system

$$\begin{cases} \mathbb{A}^{(N)} \mathbb{U}^{(N)} = \mathbb{S}^{(N)}, \\ \mathbb{A}_{j\ell}^{(N)} = 2\pi^2 j\ell \int_0^1 k(x) \cos(\pi jx) \cos(\pi \ell x) dx, \quad j, \ell \in \{1, \dots, N\}, \\ \mathbb{S}_j^{(N)} = 2 \int_0^1 \sin(\pi jx) f(x) dx, \quad j \in \{1, \dots, N\}. \end{cases} \quad [2.32]$$



**Figure 2.20.** Example 1: error curve in  $L^\infty$  norm for the FE- $\mathbb{P}_2$  method on a uniform grid (log–log scale, line with slope 3). For the color version of this figure, see [www.iste.co.uk/goudon/mathmodel.zip](http://www.iste.co.uk/goudon/mathmodel.zip)

The matrix  $\mathbb{A}^{(N)}$  is symmetric. We verify that it is positive definite because for any vector  $\xi \in \mathbb{R}^N \setminus \{0\}$ , we have

$$\begin{aligned} \mathbb{A}^{(N)} \xi \cdot \xi &= \sum_{j,\ell=1}^N 2\pi^2 j\ell \int_0^1 k(x) \cos(\pi jx) \cos(\pi \ell x) \xi_j \xi_\ell dx \\ &= 2\pi^2 \int_0^1 k(x) \left( \sum_{j=1}^N j \cos(\pi jx) \xi_j \right)^2 dx > 0, \end{aligned}$$

because the functions  $x \mapsto j \cos(\pi jx)$  form a linearly independent family in  $L^2([0, 1])$ . The general analysis, presented above, shows that the convergence speed is controlled by

$$\int_0^2 \left| \frac{d}{dx} u(x) - \sum_{n=1}^N n\pi u_n \cos(n\pi x) \right|^2 dx.$$

In other words, the convergence speed is governed by how well  $u'$  is approached using its Fourier cosine series. We know that the quality of this approximation depends on the regularity of  $u'$ : according to the regularity of the data  $f$  and the coefficient  $k$ , if we can show that the solution  $u$  is regular (say of class  $C^k$  for some  $k > 0$ ), then the approximation will be better. This method thus seems extremely attractive. However, when compared with other approximation methods, in practice, its efficiency is affected by the following facts:

- The matrix  $\mathbb{A}^{(N)}$  is, in general, “full”: it has few zeros<sup>10</sup>, unlike matrices resulting from discretization through finite differences, finite volumes or elements, which are very sparse and also have a very specific structure (tridiagonal matrices). The sparse structure can be used to radically improve the efficiency of algorithms for solving the underlying systems, by avoiding numerous unnecessary operations. Here, we cannot benefit from such improvements.

- Computing the coefficients of the matrix  $\mathbb{A}^{(N)}$  and of the right hand side requires evaluating integrals. In general, these cannot be computed exactly, and it is necessary to resort to integral approximation methods. Therefore, we do not actually solve  $\mathbb{A}^{(N)} \widetilde{\mathbb{U}}^{(N)} = \mathbb{S}^{(N)}$ , but rather  $\widetilde{\mathbb{A}^{(N)}} \widetilde{\mathbb{U}^{(N)}} = \widetilde{\mathbb{S}^{(N)}}$ , where the coefficients of  $\widetilde{\mathbb{A}^{(N)}}$  and  $\widetilde{\mathbb{S}^{(N)}}$  are obtained using formulas of the type

$$\sum_{\lambda=1}^{\Lambda_h} 2\pi^2 j\ell \omega_\lambda k(\lambda h) \cos(\pi j \lambda h) \cos(\pi \ell \lambda h), \quad \sum_{\lambda=1}^{\Lambda_h} 2\omega_\lambda \sin(\pi j \lambda h) f(\lambda h),$$

for a given parameter  $h > 0$ . However, the matrix  $\mathbb{A}^{(N)}$  is, in general, ill-conditioned: errors due to the approximation made on the coefficients, even though they may remain relatively small, can thus produce a significant error and  $\widetilde{\mathbb{U}^{(N)}}$  becomes notably different from the desired solution  $\mathbb{U}^{(N)}$ .

We can test this method with two types of data:

---

<sup>10</sup> Unless we can identify a specific basis formed by orthogonal eigenvectors for the operator  $-\partial_x(k(x)\partial_x \cdot)$  along with Dirichlet conditions; in that case the matrix  $\mathbb{A}^{(N)}$  would be diagonal.

– Example 1: regular data.

$$k(x) = 2 + \cos(5.3x),$$

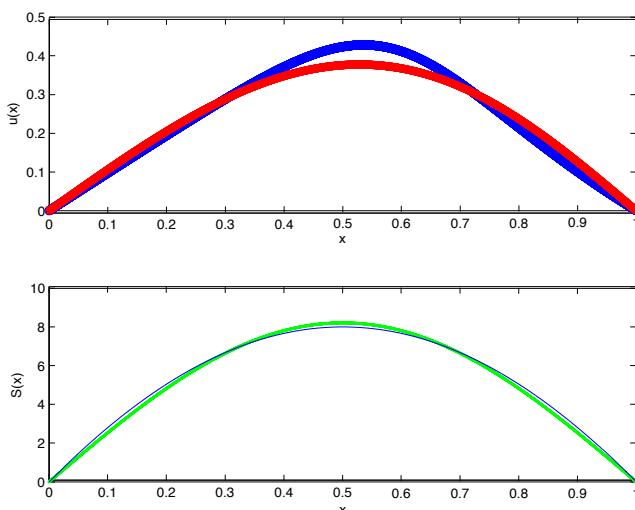
$$f(x) = 100 \times (\ln(13/4) - \ln(3 + (x - 1/2)^2)).$$

– Example 2: discontinuous data.

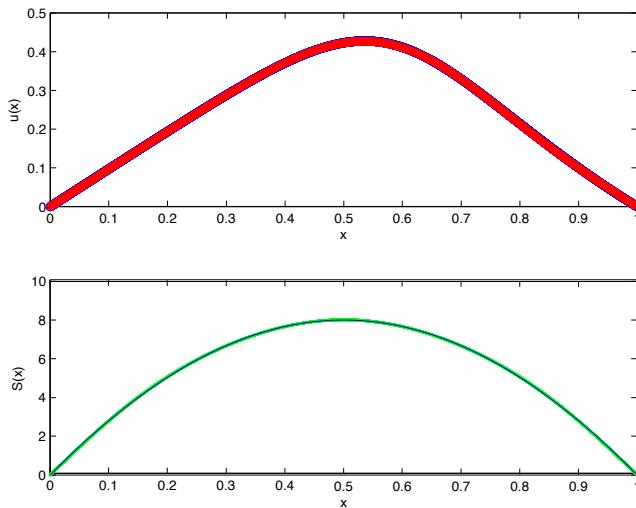
$$k(x) = \mathbf{1}_{0 \leq x \leq 1/2} + 4 \times \mathbf{1}_{1/2 < x \leq 1},$$

$$f(x) = 3 \times \mathbf{1}_{1/8 < x \leq 1/2} + 4 \times \mathbf{1}_{1/2 < x \leq 7/8}.$$

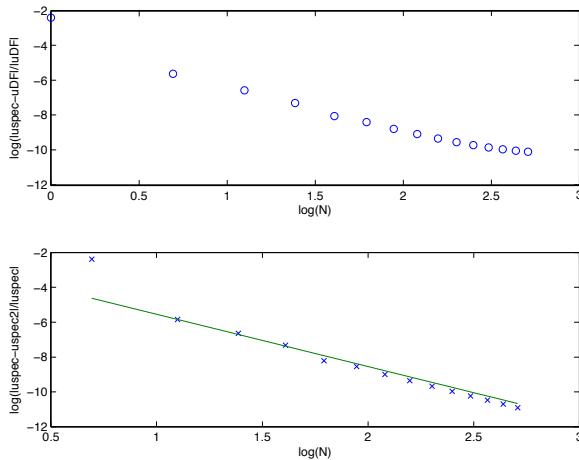
We compare the results obtained using the spectral method with a standard finite difference simulation, performed with a very fine-grained grid (see Figures 2.21–2.22 and 2.24–2.27). For Example 1, the solution remains very regular, whereas for Example 2, the  $C^1$  regularity of the solution is lost. In the regular case, the solution of the boundary value problem is well reproduced in qualitative terms with a very small number of modes; it is necessary to use more modes in the discontinuous case. From Figures 2.24–2.27, we can see that the Fourier series approximation of the data has oscillations and some over/undershoots around the discontinuities: (the approximation exceeds noticeably the extreme values of the data) this is known as the Gibbs phenomenon. Figures 2.23 and 2.28 show, on the one hand, the relative error with respect to the reference Finite Difference solution and, on the other hand, the evolution of the relative error as we make the number of modes vary. Note that this last quantity decreases less quickly when the solution is more regular.



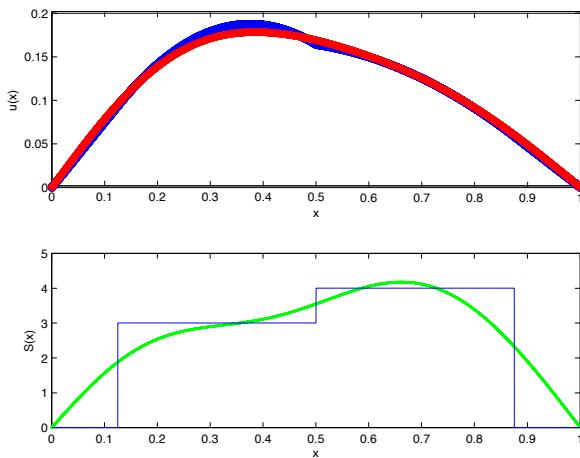
**Figure 2.21.** Example 1: spectral versus FD methods (top) and approximation of the data  $f$  (bottom) with two Fourier modes. For the color version of this figure, see [www.iste.co.uk/goudon/mathmodel.zip](http://www.iste.co.uk/goudon/mathmodel.zip)



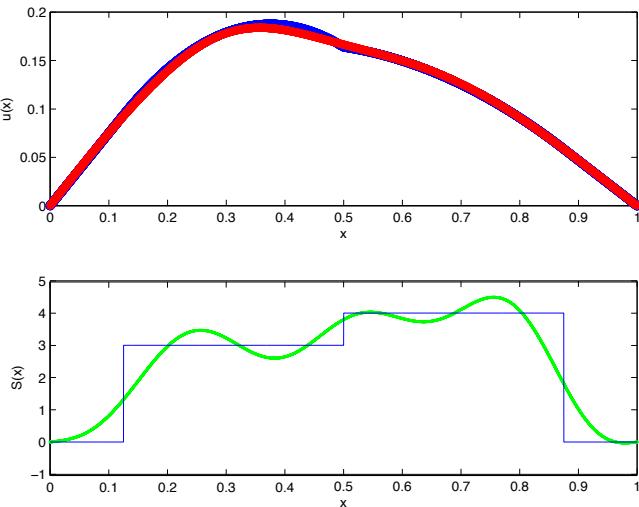
**Figure 2.22.** Example 1: spectral versus FD methods (top) and approximation of the data  $f$  (bottom) with six Fourier modes. For the color version of this figure, see [www.iste.co.uk/goudon/mathmodel.zip](http://www.iste.co.uk/goudon/mathmodel.zip)



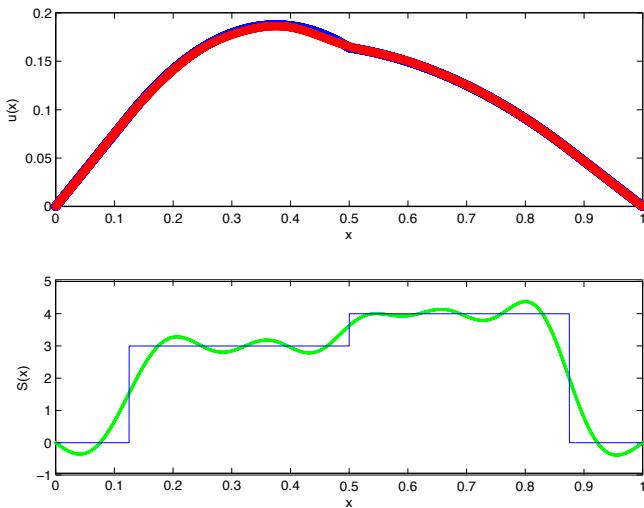
**Figure 2.23.** Example 1: relative error with respect to a FD numerical solution (top) and relative error between two solutions produced with different numbers of Fourier modes, line with slope 3 (bottom). For the color version of this figure, see [www.iste.co.uk/goudon/mathmodel.zip](http://www.iste.co.uk/goudon/mathmodel.zip)



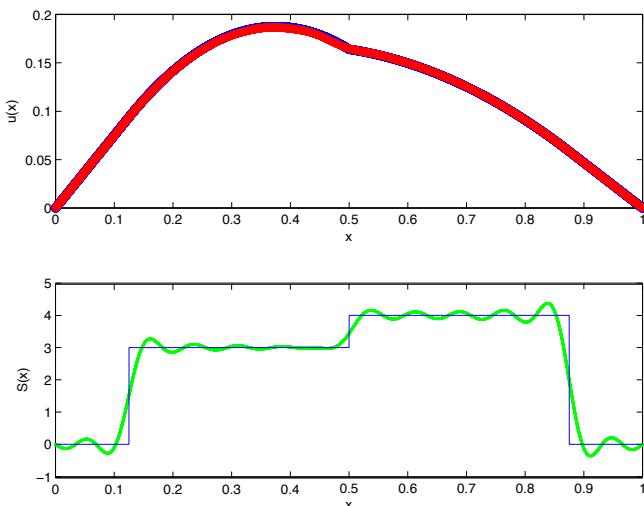
**Figure 2.24.** Example 2: spectral versus FD method (top) and approximation of the data  $f$  (bottom) with four Fourier modes. For the color version of this figure, see [www.iste.co.uk/goudon/mathmodel.zip](http://www.iste.co.uk/goudon/mathmodel.zip)



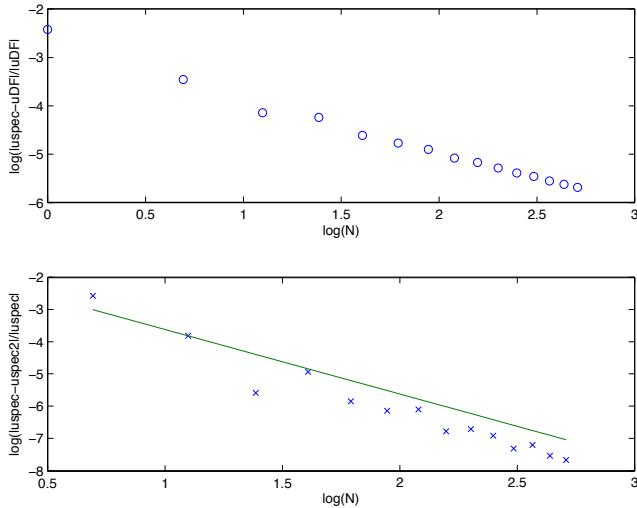
**Figure 2.25.** Example 2: spectral versus FD method (top) and approximation of the data  $f$  (bottom) with eight Fourier modes. For the color version of this figure, see [www.iste.co.uk/goudon/mathmodel.zip](http://www.iste.co.uk/goudon/mathmodel.zip)



**Figure 2.26.** Example 2: spectral versus FD method (top) and approximation of the data  $f$  (bottom) with 14 Fourier modes. For the color version of this figure, see [www.iste.co.uk/goudon/mathmodel.zip](http://www.iste.co.uk/goudon/mathmodel.zip)



**Figure 2.27.** Example 2: spectral versus FD method (top) and approximation of the data  $f$  (bottom) with 26 Fourier modes. For the color version of this figure, see [www.iste.co.uk/goudon/mathmodel.zip](http://www.iste.co.uk/goudon/mathmodel.zip)



**Figure 2.28.** Example 2: Relative error with respect to a FD numerical solution (top) and relative error between two solutions produced with different numbers of Fourier modes, line with slope 2 (bottom). For the color version of this figure, see [www.iste.co.uk/goudon/mathmodel.zip](http://www.iste.co.uk/goudon/mathmodel.zip)

## 2.7. Poisson–Boltzmann equation; minimization of a convex function, gradient descent algorithm

A point-based charge  $q$  located at a point  $M$  creates a potential  $q/|AM|$  at a point  $A$ . We now consider a charge density  $x \mapsto \rho(x)$ , which we assume is supported in  $B(0, R)$ . Then, the potential at point  $A$  is given by the integral

$$\Phi(A) = \int_{B(0, R)} \frac{\rho(x)}{|AM(x)|} dx.$$

To simplify, we assume that the charge distribution has spherical symmetry:  $\rho$  depends only on  $r = |OM|$ , where  $O$  denotes a certain origin point. In order to evaluate the integral using a change of variables, we note that  $AM = AO + OM$ , so  $|AM|^2 = |AO|^2 + |OM|^2 + 2AO \cdot OM = a^2 + r^2 - 2ra \cos(\theta)$ , having constructed the reference frame  $O, e_1, e_2, e_3$ , so that  $OA = ae_3$  and

$$OM = r \cos(\theta)e_3 + r \sin(\theta) \cos(\psi)e_2 + r \sin(\theta) \sin(\psi)e_1.$$

Therefore,

$$\begin{aligned}
 \Phi(a) &= \int_0^R \int_0^\pi \int_0^{2\pi} \frac{\rho(r)}{\sqrt{a^2 + r^2 - 2ar \cos(\theta)}} d\psi \sin(\theta) d\theta r^2 dr \\
 &= 2\pi \int_0^R r^2 \rho(r) \left( \int_0^\pi \frac{\sin(\theta) d\theta}{\sqrt{a^2 + r^2 - 2ar \cos(\theta)}} \right) dr \\
 &= 2\pi \int_0^R r^2 \rho(r) \frac{1}{ar} \left( r^2 + a^2 - 2ar \cos(\theta) \right)^{1/2} \Big|_{\theta=0}^\theta dr \\
 &= \frac{2\pi}{a} \int_0^R r \rho(r) \left( \sqrt{(r+a)^2} - \sqrt{(r-a)^2} \right) dr \\
 &= \frac{2\pi}{a} \int_0^R r \rho(r) (r+a - |r-a|) dr.
 \end{aligned}$$

We distinguish between two cases. If  $a > R$ , then we obtain  $\Phi(a) = \frac{2\pi}{a} \int_0^R 2r^2 \rho(r) dr = q/a$ , with  $q = \frac{4\pi}{a} \int_0^R 2r^2 \rho(r) dr = \int \rho(x) dx$ . If  $a < R$  (which applies at all points if  $R = +\infty$ ), then we obtain

$$\Phi(a) = \frac{2\pi}{a} \int_0^a 2r^2 \rho(r) dr + \frac{2\pi}{a} \int_a^R 2ar \rho(r) dr.$$

We note in this situation that  $\Phi$  satisfies the differential equation

$$\Phi''(a) + \frac{2}{a} \Phi'(a) = -4\pi \rho(a).$$

We can generalize the calculation by using the fact that  $x \mapsto \frac{1}{|x|}$  is proportional to the elementary solution of

$$-\Delta = -(\partial_{x_1}^2 + \partial_{x_2}^2 + \partial_{x_3}^2)$$

in  $\mathbb{R}^3$ . (It is good to know the result of broader generalization; see Appendix 5.) This means that

$$-\Delta \left( \frac{1}{|x|} \right) = C \delta(x=0),$$

for a certain constant  $C > 0$ , that is to say, for all functions  $\varphi \in C_c^\infty(\mathbb{R}^3)$ , we have

$$-\int \frac{1}{|x|} \Delta \varphi(x) dx = C \varphi(0).$$

The left-hand term is well defined because  $x \mapsto \frac{1}{|x|}$  is locally integrable on  $\mathbb{R}^3$  (since the integral  $\int_0^R \frac{r^2 dr}{r}$  is well defined for all  $0 < R < \infty$ ), and it can be evaluated as follows

$$\begin{aligned} \int \frac{1}{|x|} \Delta \varphi(x) dx &= \lim_{\epsilon \rightarrow 0} \int_{|x| \geq \epsilon} \frac{1}{|x|} \Delta \varphi(x) dx \\ &= \lim_{\epsilon \rightarrow 0} \left( \int_{|x| \geq \epsilon} \Delta \left( \frac{1}{|x|} \right) \varphi(x) dx \right. \\ &\quad \left. - \int_{|x|=\epsilon} \nabla \left( \frac{1}{|x|} \right) \cdot \nu(x) \varphi(x) d\sigma(x) - \int_{|x|=\epsilon} \frac{1}{|x|} \nabla \varphi(x) \cdot \nu(x) d\sigma(x) \right). \end{aligned}$$

Here  $\nu(x)$  denotes the unit vector normal to the sphere of radius  $\epsilon$  at the point  $x$ , and pointing outward the domain  $\{|x| \geq \epsilon\}$ , that is to say,  $\nu(x) = -\frac{x}{|x|}$ , and  $d\sigma(x)$  is the Lebesgue measure on that sphere, which can be expressed in the form  $\epsilon^2 d\omega$ ,  $\omega = \frac{x}{|x|} \in \mathbb{S}^2$ . Moreover, for a radial function  $f(x) = \tilde{f}(|x|)$ , we have  $\Delta f(x) = ((\partial_r^2 + \frac{2}{r} \partial_r) \tilde{f})(|x|)$ . In particular, for  $r \geq \epsilon > 0$ , we note that

$$\left( \partial_r^2 + \frac{2}{r} \partial_r \right) \frac{1}{r} = \frac{2}{r^3} + \frac{2}{r} \left( -\frac{1}{r^2} \right) = 0.$$

Finally, we have  $\nabla \left( \frac{1}{|x|} \right) = -\frac{x}{|x|^3}$ . It follows that

$$\begin{aligned} \int \frac{1}{|x|} \Delta \varphi(x) dx &= \lim_{\epsilon \rightarrow 0} \left( 0 - \epsilon^2 \int_{\mathbb{S}^2} \frac{1}{\epsilon^2} \varphi(\epsilon \omega) d\omega - \epsilon^2 \int_{\mathbb{S}^2} \frac{1}{\epsilon} \nabla \varphi(\epsilon \omega) \cdot \omega d\omega \right) \\ &= -|\mathbb{S}^2| \varphi(0) = -4\pi \varphi(0). \end{aligned}$$

As a result, if we associate the charge potential  $\Phi$  defined by

$$\Phi(x) = \frac{1}{4\pi} \int \frac{\rho(y)}{|x-y|} dy$$

with a charge density  $\rho$ , then  $\Phi$  satisfies

$$-\Delta \Phi = \rho.$$

In what follows, we will assume that the charge distribution is itself influenced by potential: the Boltzmann distribution

$$\rho(x) = e^{-\Phi(x)},$$

describes a situation where the charges are in an electrostatic equilibrium state. As a result,  $\Phi$  is a solution of the nonlinear problem

$$-\Delta\Phi = e^{-\Phi}.$$

We complete the problem with Dirichlet boundary conditions on the domain  $\Phi|_{\partial\Omega} = 0$ . Energy considerations allow us to see this problem as related to minimizing the following energy functional:

$$\mathcal{J}(\Phi) = \int_{\Omega} \left( \frac{1}{2} |\nabla\Phi|^2 + e^{-\Phi} \right) dx.$$

We will study this method from a discrete point of view. We use a finite element discretization, whose principles are summarized as follows:

- we use the variational formulation of the problem:

$$\int_{\Omega} \nabla\Phi(x) \cdot \nabla\psi(x) dx = \int_{\Omega} e^{-\Phi(x)} \psi(x) dx,$$

which is satisfied for all test functions  $\psi$ ;

- given  $\{\psi_1, \dots, \psi_{N_h}\}$ , a basis of the chosen finite element space (for example, the “hat functions” of the approximation  $\mathbb{P}_1$ ),  $\Phi$  is approximated by

$$\Phi^h(x) = \sum_{n=1}^{N_h} \Phi_n \psi_n(x).$$

We let  $A$  denote the matrix corresponding to the discretization of  $-\Delta$ , that is to say,

$$A_{i,j} = \int_{\Omega} \nabla_x \psi_j(x) \nabla_x \psi_i(x) dx.$$

We introduce the mapping

$$\mathcal{E} : (\Phi_1, \dots, \Phi_{N_h}) \in \mathbb{R}^{N_h}$$

$$\mapsto \left( \int_{\Omega} \exp \left( - \sum_{n=1}^{N_h} \Phi_n \psi_n(x) \right) \psi_k(x) dx, k \in \{1, \dots, N_h\} \right) \in \mathbb{R}^{N_h}.$$

The Poisson–Boltzmann equation takes on the discrete form

$$A\Phi = \mathcal{E}(\Phi).$$

The energy criteria allows us to consider the functional

$$\mathcal{J} : \Phi \in \mathbb{R}^{N_h} \longmapsto \mathcal{J}(\Phi) = \frac{1}{2} A\Phi \cdot \Phi + b(\Phi)$$

where

$$b(\Phi) = \int_{\Omega} \exp \left( - \sum_{n=1}^{N_h} \Phi_n \psi_n(x) \right) dx.$$

**PROPOSITION 2.7.–** The functional  $\mathcal{J}$  is  $C^2$  and there exists an  $\alpha > 0$ , such that  $\mathcal{J}$  is  $\alpha$ -convex on  $\mathbb{R}^{N_h}$ . It has a unique minimum  $\Phi^*$  that satisfies  $A\Phi^* = \mathcal{E}(\Phi^*)$ .

**PROOF.–** Let us first suppose that  $\mathcal{J}$  is  $\alpha$ -convex, this is to say, for all  $\Phi, \Psi \in \mathbb{R}^{N_h}$ , we have

$$\mathcal{J}(\Phi) \geq \mathcal{J}(\Psi) + \nabla \mathcal{J}(\Psi) \cdot (\Phi - \Psi) + \alpha \|\Phi - \Psi\|^2,$$

and let us begin by showing that  $\mathcal{J}$  has a unique minimum. Let  $(\Phi^{(n)})_{n \in \mathbb{N}}$  be a minimizing sequence, that is to say, such that  $\lim_{n \rightarrow \infty} \mathcal{J}(\Phi^{(n)}) = \inf \{\mathcal{J}(\Psi), \Psi \in \mathbb{R}^{N_h}\}$ . In particular, we have

$$\begin{aligned} \mathcal{J}(\Phi^{(n)}) &\geq \mathcal{J}(\Phi^{(0)}) + \nabla \mathcal{J}(\Phi^{(0)}) \cdot (\Phi^{(n)} - \Phi^{(0)}) + \alpha \|\Phi^{(n)} - \Phi^{(0)}\|^2 \\ &\geq \mathcal{J}(\Phi^{(0)}) + \|\Phi^{(n)} - \Phi^{(0)}\|(\alpha \|\Phi^{(n)} - \Phi^{(0)}\| - \|\nabla \mathcal{J}(\Phi^{(0)})\|). \end{aligned}$$

Since the left-hand term is bounded, it follows from an argument by contradiction that the sequence  $(\Phi^{(n)})_{n \in \mathbb{N}}$  is bounded. By the Bolzano–Weierstrass theorem, we can extract a subsequence  $(\Phi^{(n_k)})_{n_k \in \mathbb{N}}$  that converges:  $\lim_{k \rightarrow \infty} \Phi^{(n_k)} = \Phi^*$ . Then, by continuity, we obtain  $\lim_{n \rightarrow \infty} \mathcal{J}(\Phi^{(n)}) = \mathcal{J}(\Phi^*) = \inf \{\mathcal{J}(\Psi), \Psi \in \mathbb{R}^{N_h}\}$ . This justifies the existence of a minimizer. Note that we also have

$$\nabla \mathcal{J}(\Phi^*) = 0.$$

Indeed, for all  $t > 0$  and all  $\eta \in \mathbb{R}^{N_h}$ , we have  $\mathcal{J}(\Phi^* + t\eta) \geq \mathcal{J}(\Phi^*)$ . It follows that  $\frac{\mathcal{J}(\Phi^* + t\eta) - \mathcal{J}(\Phi^*)}{t} \geq 0$ . Making  $t$  tend to 0, we can see that  $\nabla \mathcal{J}(\Phi^*) \cdot \eta \geq 0$  for all  $\eta \in \mathbb{R}^{N_h}$ . This inequality is thus satisfied for both  $\eta$  and  $-\eta$  for all  $\eta \in \mathbb{R}^{N_h}$ , which therefore implies that  $\nabla \mathcal{J}(\Phi^*) = 0$ . Finally, if we suppose that  $\Phi^*$  and  $\Psi$  are both minimizers, then we obtain

$$\mathcal{J}(\Psi) \geq \mathcal{J}(\Phi^*) + \nabla \mathcal{J}(\Phi^*) \cdot (\Psi - \Phi^*) + \alpha \|\Psi - \Phi^*\|^2 = \mathcal{J}(\Phi^*) + \alpha \|\Psi - \Phi^*\|^2.$$

Now,  $\mathcal{J}(\Psi) = \mathcal{J}(\Phi^*)$ , which implies  $\|\Phi^* - \Psi\| = 0$ : the minimizer is thus unique.

To show that  $\mathcal{J}$  is  $\alpha$ -convex, we first calculate its gradient

$$\nabla \mathcal{J}(\Phi) \cdot \eta = A\Phi \cdot \eta - \sum_{k=1}^{N_h} \int_{\Omega} \exp\left(-\sum_{n=1}^{N_h} \Phi_n \psi_n(x)\right) \psi_k(x) \eta_k \, dx,$$

and then its Hessian matrix

$$\begin{aligned} D^2 \mathcal{J}(\Phi) \eta^{(1)} \cdot \eta^{(2)} &= A\eta^{(1)} \cdot \eta^{(2)} \\ &+ \sum_{k,\ell=1}^{N_h} \int_{\Omega} \exp\left(-\sum_{n=1}^{N_h} \Phi_n \psi_n(x)\right) \psi_k(x) \eta_k^{(1)} \psi_\ell(x) \eta_\ell^{(2)} \, dx. \end{aligned}$$

It follows that

$$D^2 \mathcal{J}(\Phi) \eta \cdot \eta = A\eta \cdot \eta + \int_{\Omega} \exp\left(-\sum_{n=1}^{N_h} \Phi_n \psi_n(x)\right) \left(\sum_{k=1}^{N_h} \psi_k(x) \eta_k\right)^2 \, dx \geq A\eta \cdot \eta.$$

The matrix  $A$  is positive definite; the functional  $\mathcal{J}$  is therefore  $\alpha$ -convex, with  $\alpha$  equal to the smallest eigenvalue of  $A$ .

Finally, we note that with the above formulas, the condition  $\nabla \mathcal{J}(\Phi^*) = 0$  leads to

$$(A\Phi^* - b(\Phi^*)) \cdot \eta = 0.$$

□

These observations motivate the search for a solution using a “descent algorithm”: it is an iterative method where at each step, we follow the direction of the steepest descent, which is that of the gradient. The gradient algorithm proceeds as follows. Let  $\Phi^{(0)} \in \mathbb{R}^{N_h}$  and  $\rho > 0$  be fixed. We construct the sequence

$$\Phi^{(m+1)} = \Phi^{(m)} - \rho \nabla \mathcal{J}(\Phi^{(m)}). \quad [2.33]$$

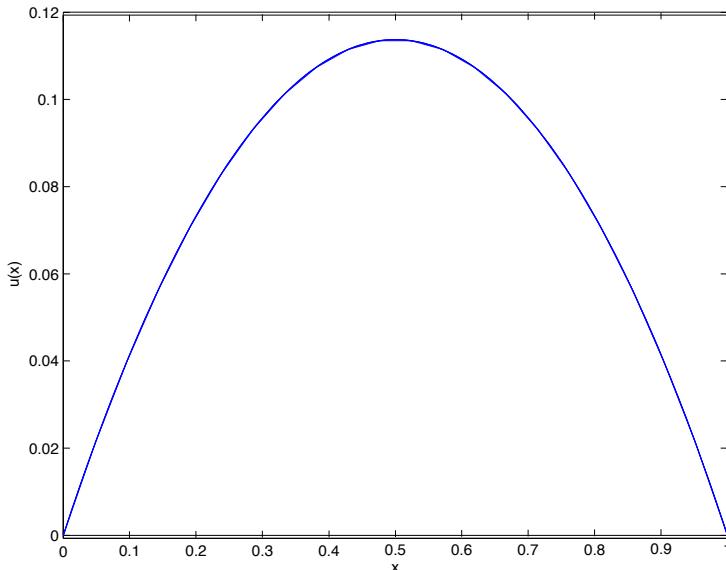
**THEOREM 2.13.**— Let  $\mathcal{J}$  be a  $C^2$ ,  $\alpha$ -convex functional, such that  $\nabla \mathcal{J}$  is  $L$ -Lipschitz. Then, for all  $0 < \rho < 2\alpha/L$ , the sequence  $(\Phi^{(m)})_{m \in \mathbb{N}}$  defined by [2.33] converges to  $\Phi^*$ , the unique minimizer of  $\mathcal{J}$  on  $\mathbb{R}^{N_h}$ .

**PROOF.**— Recall that  $\nabla \mathcal{J}(\Phi^*) = 0$  because  $\Phi^*$  is a minimizer. We write  $\delta^{(m)} = \Phi^{(m)} - \Phi^*$ . We therefore have  $\delta^{(m+1)} = \delta^{(m)} - \rho(\nabla \mathcal{J}(\Phi^{(m)}) - \nabla \mathcal{J}(\Phi^*))$ . Now,

$\mathcal{J}$  is  $\alpha$ -convex, which means that for all  $a, b \in \mathbb{R}^{N_h}$ , we have  $(\nabla \mathcal{J}(a) - \nabla \mathcal{J}(b)) \cdot (a - b) \geq \alpha |a - b|^2$ . It follows that

$$\begin{aligned} |\delta^{(m+1)}|^2 &= |\delta^{(m)} - \rho(\nabla \mathcal{J}(\Phi^{(m)}) - \nabla \mathcal{J}(\Phi^*))|^2 \\ &\leq |\delta^{(m)}|^2 + \rho^2 |\nabla \mathcal{J}(\Phi^{(m)}) - \nabla \mathcal{J}(\Phi^*)|^2 \\ &\quad - 2\rho(\nabla \mathcal{J}(\Phi^{(m)}) - \nabla \mathcal{J}(\Phi^*)) \cdot \delta^{(m)} \\ &\leq |\delta^{(m)}|^2(1 - 2\alpha\rho + L\rho^2). \end{aligned}$$

The condition  $0 < \rho < 2\alpha/L$  ensures that  $0 < (1 - 2\alpha\rho + L\rho^2) < 1$ , from which we conclude that  $\lim_{m \rightarrow \infty} |\delta^{(m)}| = 0$ .  $\square$



**Figure 2.29.** Numerical solutions to the Boltzmann–Poisson problem for different space steps

Figure 2.29 shows the numerical solutions obtained using this algorithm, with a  $\mathbb{P}_1$  method for different space steps. In practice, the algorithm is rather delicate: if the descent step  $\rho$  is too large, the method does not converge and produces outlying results. Moreover, as shown by the convergence analysis and the spectral properties

of the Laplace matrix, the descent steps must be made smaller when the grid is finer. This difficulty is also related to the conditioning of the underlying linear system. As a result, it is necessary to perform a very large number of iterations and not be fooled by a solution that remains close to a different state of the desired solution. Therefore, the efficiency of the method is also strongly dependent on the choice of the initial guess.

## 2.8. Neumann conditions: the optimization perspective

For the problem

$$\begin{aligned} -\frac{d}{dx} \left( k(x) \frac{d}{dx} u(x) \right) &= f(x) \quad \text{for } x \in ]0, 1[, \\ \frac{d}{dx} u(0) &= 0 = \frac{d}{dx} u(1), \end{aligned} \tag{2.34}$$

we previously made the following remarks:

- if a solution exists, say  $u_0$ , it is not unique because for all  $c \in \mathbb{R}$ ,  $x \mapsto u_0(x) + c$  is also a solution;
- integrating the equation, we realize that the data must satisfy  $\int_0^1 f(x) dx = 0$  in order for the equation to make sense;
- these difficulties spread at the discrete level, see note 2.3: the corresponding finite difference matrix for a constant coefficient  $k$  is written as

$$\mathbb{A}^{\text{Neumann}} = -\frac{k}{h^2} \begin{pmatrix} -1 & 1 & 0 & \cdots & 0 \\ 1 & -2 & & & \\ 0 & & & & \\ \vdots & & & & \\ 0 & 0 & 1 & -1 \end{pmatrix} \in \mathcal{M}_N.$$

Note that  $\text{Ker}(\mathbb{A}^{\text{Neumann}}) = \text{Span}(1, 1, \dots, 1)$ .

The following statement makes this last observation more precise.

LEMMA 2.9.– The matrix  $\mathbb{A}^{\text{Neumann}}$  is coercive on  $[\text{Span}(1, 1, \dots, 1)]^\perp$  and there exists a constant  $\alpha > 0$ , such that for all  $x \in \mathbb{R}^N$ , we have

$$\mathbb{A}^{\text{Neumann}} x \cdot x \geq \alpha |x - \langle x \rangle e|^2$$

where  $e = (1, 1, \dots, 1)$  and  $\langle x \rangle = \frac{1}{N} \sum_{i=1}^N x_i = \frac{x \cdot e}{e \cdot e}$ .

PROOF.– We have seen that  $\mathbb{A}^{\text{Neumann}} e = 0$ . Note that

$$\begin{aligned}\mathbb{A}^{\text{Neumann}} x \cdot x &= \frac{k}{h^2} \left( \sum_{i=2}^{N-1} (x_{i+1} - x_i)x_i + (x_2 - x_1)x_1 \right. \\ &\quad \left. - \sum_{i=2}^{N-1} (x_i - x_{i-1})x_i - (x_N - x_{N-1})x_N \right) \\ &= \frac{k}{h^2} \left( \sum_{i=1}^{N-1} (x_{i+1} - x_i)x_i - \sum_{i=1}^{N-1} (x_{i+1} - x_i)x_{i+1} \right) \\ &= \frac{k}{h^2} \sum_{i=1}^{N-1} (x_{i+1} - x_i)^2 \geq 0.\end{aligned}$$

Moreover, this quantity is zero whenever  $x$  is collinear with  $e$ . The function  $y \mapsto \mathbb{A}^{\text{Neumann}} y \cdot y$  is continuous and strictly positive on the compact set  $\{|y| = 1\} \cap \text{Span}(e)^\perp$ . So it is bounded below on that domain by a strictly positive constant  $\alpha$ . It follows that for all  $y \in \mathbb{R}^N$ , such that  $y \cdot e = 0$ , we have  $\mathbb{A}^{\text{Neumann}} y \cdot y \geq \alpha|y|^2$ . Finally, we conclude with the following relations:

$$\begin{aligned}\mathbb{A}^{\text{Neumann}} x \cdot x &= \mathbb{A}^{\text{Neumann}}(x - \langle x \rangle e) \cdot x \\ &= \mathbb{A}^{\text{Neumann}}(x - \langle x \rangle e) \cdot (x - \langle x \rangle e) + 0 \geq \alpha|x - \langle x \rangle e|^2\end{aligned}$$

because  $\mathbb{A}^{\text{Neumann}}$  is symmetric, so  $\mathbb{A}^{\text{Neumann}}(x - \langle x \rangle e) \cdot e = 0$  and  $x - \langle x \rangle e$  is orthogonal to  $e$ .  $\square$

As mentioned before, the same difficulty arises for periodic conditions. One way to restore uniqueness is to impose an additional constraint on the solution, for example searching for the solution with zero mean (that is to say, in a periodic context, with a first Fourier mode equal to zero). We will study this question from a strictly discrete perspective (the extension onto the continuous framework relies on the same arguments, but it uses a more elaborated functional framework). We reinterpret the problem as a minimization problem with the constraint

$$(P) \quad \begin{cases} \text{Minimizing } y \in \mathbb{R}^N \mapsto \mathcal{J}(y) = \frac{1}{2} \mathbb{A} y \cdot y - b \cdot y \\ \text{on the set } \mathcal{C} = \{y \in \mathbb{R}^N, y \cdot e = 0\}, \end{cases}$$

where  $e$  represents the vector  $(1, \dots, 1)$ .

THEOREM 2.14.– The problem  $(P)$  has at least one solution.

PROOF.– The set  $\mathcal{C}$  is a convex, closed and non-empty subset of  $\mathbb{R}^N$ . The functional  $\mathcal{J}$  is continuous and convex on  $\mathbb{R}^N$ , so also on  $\mathcal{C}$ . We will show that it is coercive on  $\mathcal{C}$ : if  $(y_n)_{n \in \mathbb{N}}$  is a sequence of elements of  $\mathcal{C}$ , such that  $\lim_{n \rightarrow \infty} |y_n| = \infty$ , then  $\lim_{n \rightarrow \infty} \mathcal{J}(y_n) = \infty$ . If this statement were false, we could find a subsequence, which we continue to denote  $(y_n)_{n \in \mathbb{N}}$ , such that  $(\mathcal{J}(y_n))_{n \in \mathbb{N}}$  has a finite limit. We can write  $\bar{y}_n = y_n/|y_n|$ . Since this sequence is bounded, at the possible cost of extracting a subsequence, we can assume it has a limit  $\bar{y}$ . In particular, we have  $\bar{y} \in \mathcal{C}$  and

$$\frac{\mathcal{J}(y_n)}{|y_n|^2} = \frac{1}{2} \mathbb{A} \bar{y}_n \cdot \bar{y}_n - \frac{b \cdot \bar{y}_n}{|y_n|} \xrightarrow[n \rightarrow \infty]{} 0 = \frac{1}{2} \mathbb{A} \bar{y} \cdot \bar{y}.$$

This implies that  $\mathbb{A} \bar{y} = 0$ , with  $\bar{y}$  of norm 1 and orthogonal to  $\text{Span}(e) = \text{Ker}(\mathbb{A})$ , which is a contradiction.

Let  $(x_n)_{n \in \mathbb{N}}$  be a minimizing sequence:  $x_n \in \mathcal{C}$  and  $\lim_{n \rightarrow \infty} \mathcal{J}(x_n) = \inf_{y \in \mathcal{C}} \mathcal{J}(y)$ . Since  $\mathcal{J}$  is coercive, this sequence is bounded and we can assume that it has a limit  $x$ . By continuity, we obtain  $x \in \mathcal{C}$  and  $\mathcal{J}(x) = \inf_{y \in \mathcal{C}} \mathcal{J}(y)$ .  $\square$

We now apply the constrained optimization theorem, which ensures the existence of  $\lambda \in \mathbb{R}$ , the Lagrange associated with the constraint, such that  $\nabla \mathcal{J}(x) + \lambda e = 0$ ; in other words, the multiplier pair  $(x, \lambda)$  satisfies the system

$$\begin{pmatrix} \mathbb{A} & e \\ e^\top & 0 \end{pmatrix} \begin{pmatrix} x \\ \lambda \end{pmatrix} = \begin{pmatrix} b \\ 0 \end{pmatrix}. \quad [2.35]$$

We are faced with a linear system of size  $(N + 1) \times (N + 1)$ , which remains symmetric<sup>11</sup>.

**THEOREM 2.15.**– The problem [2.35] has a unique solution.

PROOF.– It suffices to show that the associated matrix is injective, that is to say, the solution of [2.35] with  $b = 0$  is  $x = 0, \lambda = 0$ . Indeed, if  $(x, \lambda)$  satisfies this equation, then  $\mathbb{A}x \cdot x + \lambda e \cdot x = \mathbb{A}x \cdot x = 0$ . This implies that  $x = \alpha e$ , for a certain  $\alpha \in \mathbb{R}$ . However,  $x \cdot e = 0$  becomes  $\alpha e \cdot e = 0$ , which implies  $\alpha = 0$  and therefore  $x = 0$ , and then  $\mathbb{A}x + \lambda e = \lambda e = 0$  gives  $\lambda = 0$ .

We note that in [2.35] no compatibility conditions are required on the right hand side  $b$ . In fact, we find it in the multiplier  $\lambda$ : if  $(x, \lambda)$  is a solution of [2.35], then we have  $\mathbb{A}x \cdot e + \lambda e \cdot e = b \cdot e$ . However, since  $\mathbb{A}$  is symmetric,  $\mathbb{A}x \cdot e = x \cdot \mathbb{A}e = 0$ ,

---

<sup>11</sup> The matrix of the system [2.35] is not however positive definite. Indeed, for a symmetric, positive definite  $M$ , we have in particular the relation  $M e_i \cdot e_i = M_{ii} > 0$  for the canonical basis vectors  $e_1, \dots, e_{N+1}$ , which property is not satisfied by the matrix from system [2.35].

which implies that  $\lambda = \frac{b \cdot e}{e \cdot e}$  and  $\mathbb{A}x = b - \frac{b \cdot e}{e \cdot e}e$ : the right hand side of that equation is orthogonal to  $e$ .  $\square$

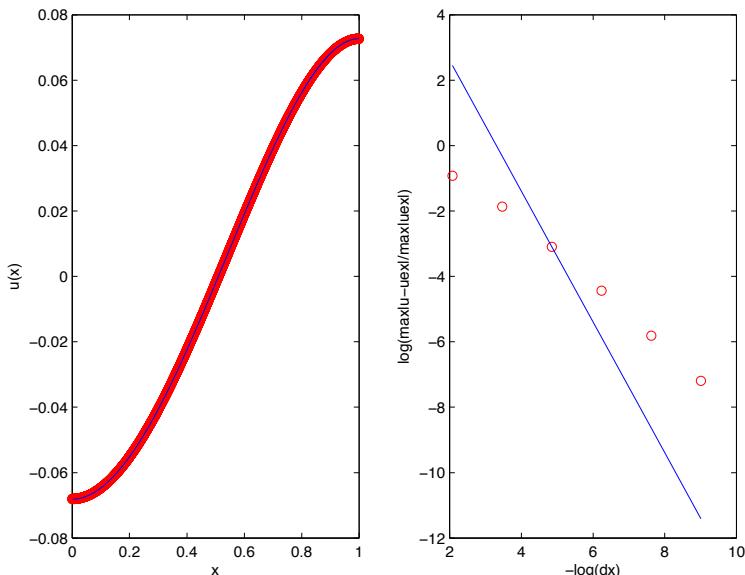
Figure 2.30 shows the results obtained using this approach for the problem

$$-\frac{d^2}{dx^2}u = e^x - \int_0^1 e^y dy, \quad \frac{d}{dx}u(0) = \frac{d}{dx}u(1) = 0.$$

The solution is

$$u_{\text{exact}}(x) = (e-1)\frac{x^2}{2} + x - e^x + \left(\frac{5}{6}(e-1) - \frac{1}{2}\right).$$

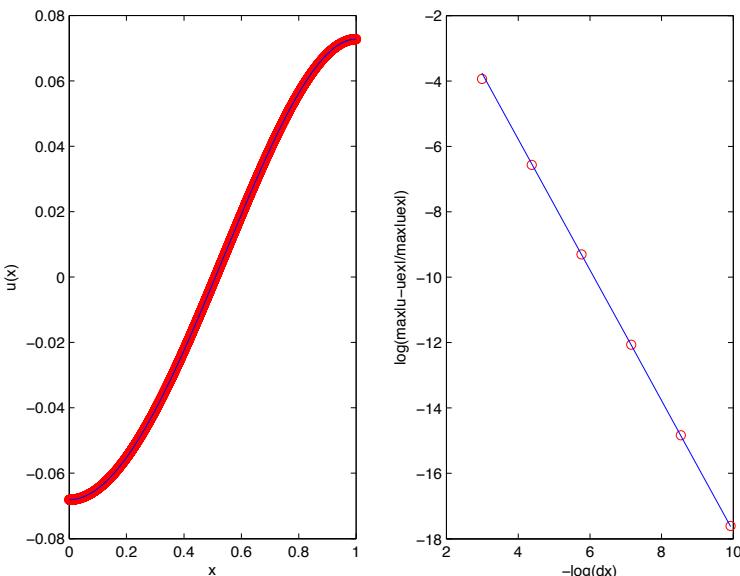
We find only order 1 accuracy. To reach order 2, it is necessary to modify the algorithm to some extent. On the one hand, we modify the treatment of boundary terms, see note 2.3. On the other hand, this discussion relies on an approximation of the integral on  $[0, 1]$  through the rectangle method, so it is only of order 1. It is necessary to replace the multiplication times  $e$  by an operation corresponding to the trapezoid method approximation (see Appendix 2). These manipulations allow us to make the approximation of order 2 (see Figure 2.31).



**Figure 2.30.** Simulation of Neumann problem: order 1 method

## 2.9. Charge distribution on a cord

We consider a tight rope, attached at the ends. At rest, the cord is in a horizontal position, but it deforms if a charge is delivered to it. This charge may cause it to rupture at the fixture points if the tension on it becomes too large. We therefore inquire how to best distribute the charge to make the effect of tension minimal. The model that will be discussed can be seen as a simplified version of several engineering and material resistance problems.



**Figure 2.31.** Simulation of Neumann problem: order 2 method. For the color version of this figure, see [www.iste.co.uk/goudon/mathmodel.zip](http://www.iste.co.uk/goudon/mathmodel.zip)

The cord is described as a one-dimensional medium with each abscissa  $x \in [0, 1]$  we associate the displacement  $u(x) \in \mathbb{R}$  with respect to the equilibrium position. Since the cord is attached at the ends, we have  $u(0) = 0 = u(1)$ . The charge is represented by a density per unit length  $x \mapsto f(x)$ . The mechanical properties of the cord are described with a rigidity coefficient  $x \mapsto k(x) > 0$  that can be variable (and we will see the role that these variations can play in the optimal charge distribution). We consider an element of the cord between the points  $x$  and  $x + dx$ . The length of this element at rest is therefore simply  $dx$ . Once it has been charged, using the Pythagorean theorem (see Figure 2.32), this element has a length close to

$\sqrt{(u(x + dx) - u(x))^2 + (x + dx - x)^2}$ . The energy corresponding to the stretching of the cord is proportional to the increase in length:

$$\mathcal{E} = k(x) \left( \sqrt{(u(x + dx) - u(x))^2 + dx^2} - dx \right).$$

We assume that the function  $x \mapsto u(x)$  is regular enough to use the expansion  $u(x + dx) - u(x) = \frac{\partial}{\partial x} u(x) dx + dx\epsilon(dx)$ , with  $\lim_{h \rightarrow 0} \epsilon(h) = 0$ . Then, we have

$$\mathcal{E} = k(x) \left( \sqrt{\left( \frac{\partial}{\partial x} u(x) + \epsilon(dx) \right)^2 dx^2 + dx^2} - dx \right),$$

and since we are considering infinitesimal elements  $0 < dx \ll 1$ , we can replace  $\mathcal{E}$  with the approximate expression

$$\tilde{\mathcal{E}} = k(x) \left( \sqrt{\left( \frac{\partial}{\partial x} u(x) \right)^2 + 1} - 1 \right) dx.$$

Moreover, the potential energy is the opposite of the work from the external force on that displacement, which is given by

$$-f(x)u(x)dx.$$

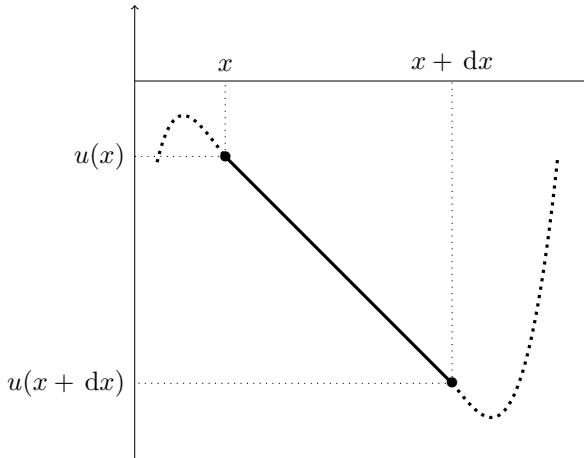
Note that, since  $f$  is a force density by length unit, this quantity has the homogeneity of a “force×length” product, that is to say, mass  $\times \frac{\text{length}^2}{\text{time}^2}$ , which is, indeed, the dimension of an energy. The total energy is obtained by adding all the infinitesimal elements, and we thus write

$$\mathcal{E}_{\text{Tot}}(u) = \int_0^1 \left\{ k(x) \left( \sqrt{\left( \frac{\partial}{\partial x} u(x) \right)^2 + 1} - 1 \right) - f(x)u(x) \right\} dx.$$

Next, we simplify the expression further by assuming that the displacement  $u$  and its derivative are small. We resort to the expansion  $\sqrt{1+y} = 1 + \frac{1}{2}y + y\eta(y)$ , where  $\lim_{y \rightarrow 0} \eta(y) = 0$ . Finally, we focus in this context on the functional

$$\tilde{\mathcal{E}}_{\text{Tot}}(u) = \frac{1}{2} \int_0^1 k(x) \left( \frac{\partial}{\partial x} u(x) \right)^2 dx - \int_0^1 f(x)u(x) dx.$$

The cord's position at equilibrium under charge  $f$  is thus defined as the function  $x \mapsto u(x)$ , which satisfies the boundary conditions  $u(0) = 0$  and  $u(1) = 0$  that minimize  $\tilde{\mathcal{E}}_{\text{Tot}}$ .

**Figure 2.32.** *Cord elongation*

We assume that the data  $f$  is a continuous function, or an element of  $L^2([0, 1])$ . We note that  $\tilde{\mathcal{E}}_{\text{Tot}}(u)$  is well defined for all  $u \in H_0^1([0, 1])$ , a functional space that takes into account the Dirichlet conditions. More precisely, we can rewrite

$$\tilde{\mathcal{E}}_{\text{Tot}}(u) = \frac{1}{2}a(u, u) - \ell(u)$$

where  $a$  is the symmetric bilinear form defined on  $H_0^1([0, 1]) \times H_0^1([0, 1])$  by

$$a(u, v) = \int_0^1 k(x) \frac{\partial}{\partial x} u(x) \frac{\partial}{\partial x} v(x) \, dx$$

and  $\ell$  is the linear form defined on  $H_0^1([0, 1])$  by

$$\ell(u) = \int_0^1 f(x) u(x) \, dx.$$

**PROPOSITION 2.8.–** The function  $u$  minimizes  $\tilde{\mathcal{E}}_{\text{Tot}}$  on  $H_0^1([0, 1])$  if and only if

$$\text{for all } v \in H_0^1([0, 1]), \text{ we have } a(u, v) = \ell(v). \quad [2.36]$$

**PROOF.–** First suppose that  $u$  minimizes  $\tilde{\mathcal{E}}_{\text{Tot}}$ . Then, for all  $h \in H_0^1([0, 1])$  and all  $t > 0$ , we have

$$\tilde{\mathcal{E}}_{\text{Tot}}(u + th) \geq \tilde{\mathcal{E}}_{\text{Tot}}(u).$$

Now, by developing  $a(u+th)$  and using the fact that  $a(u, h) = a(h, u)$ , we observe that

$$\tilde{\mathcal{E}}_{\text{Tot}}(u + th) = \tilde{\mathcal{E}}_{\text{Tot}}(u) + t(a(u, h) - \ell(h)) + \frac{t^2}{2}a(h, h).$$

Since  $\tilde{\mathcal{E}}_{\text{Tot}}(u + th) \geq \tilde{\mathcal{E}}_{\text{Tot}}(u)$ , it follows that

$$t(a(u, h) - \ell(h)) + \frac{t^2}{2}a(h, h) \geq 0.$$

We divide by  $t > 0$  and make  $t$  tend to 0 to obtain  $a(u, h) - \ell(h) \geq 0$ . Since this relation is satisfied for all  $h \in H_0^1([0, 1])$ , we can apply it with  $h = v$  and  $h = -v$  for all elements  $v$  of  $H_0^1([0, 1])$ : we conclude that  $u$  satisfies [2.36].

Reciprocally, if  $u$  satisfies [2.36], then we obtain

$$\tilde{\mathcal{E}}_{\text{Tot}}(u + v) = \tilde{\mathcal{E}}_{\text{Tot}}(u) + \underbrace{a(u, v) - \ell(v)}_{=0} + \underbrace{\frac{1}{2}a(v, v)}_{\geq 0} \geq \tilde{\mathcal{E}}_{\text{Tot}}(u). \quad \square$$

The bilinear form  $a$  is continuous on  $H_0^1([0, 1]) \times H_0^1([0, 1])$  because the Cauchy–Schwarz inequality makes it possible to dominate  $|a(u, v)| \leq \|k\|_\infty \|\partial_x u\|_{L^2} \|\partial_x v\|_{L^2}$ . Moreover, by writing  $\kappa = \min_{x \in [0, 1]} k(x) > 0$ , we obtain  $a(u, u) \geq \kappa \|\partial_x u\|_{L^2}^2$ . By the Poincaré inequality,  $u \mapsto \|\partial_x u\|_{L^2}^2$  defines a norm on  $H_0^1([0, 1])$ , which is equivalent to the usual norm on  $H^1$ ; the form  $a$  is therefore coercive. The Lax–Milgram theorem ensures the existence and uniqueness of a solution  $u \in H_0^1([0, 1])$  of [2.36], which is therefore the solution to the minimization problem. We conclude that [2.36] is just the variational formulation of the boundary problem

$$\begin{aligned} -\frac{\partial}{\partial x} \left( k(x) \frac{\partial}{\partial x} u(x) \right) &= f(x) \quad \text{for } 0 < x < 1, \\ u(0) &= u(1) = 0. \end{aligned} \quad [2.37]$$

Therefore, to every data  $f$ , there corresponds a minimizer  $u$  of  $\tilde{\mathcal{E}}_{\text{Tot}}$ , the solution of [2.37]. With this solution, we associate a *rupture energy* defined from the normal derivatives at the ends  $x = 0$  and  $x = 1$ . More precisely, we write

$$R = \frac{1}{2} \left( k(0) \frac{\partial}{\partial x} u(0) \right)^2 + \frac{1}{2} \left( k(1) \frac{\partial}{\partial x} u(1) \right)^2.$$

This quantity depends on the applied force  $f$ . We will focus on forces that have a specific form characterized by an application point, and then we will look for the optimal position that minimizes rupture energy.

We define the applied force as follows: given a function  $p : \mathbb{R} \rightarrow [0, \infty[$  with compact support in  $[-\ell, +\ell]$ , where  $0 < \ell < 1/2$ , we fix  $a \in ]\ell, 1 - \ell[$  and write

$$f(x) = -p(x - a) \leq 0.$$

The sign means that the force is oriented “downwards”, and we therefore expect to find a negative solution  $u$ . The point  $a$  represents the “center” of the charge. Therefore, for each value of  $a$ , we can associate it with a corresponding solution of [2.37], denoted as  $x \mapsto u^{(a)}(x)$ , as well as the rupture energy

$$R(a) = \frac{1}{2} \left( k(0) \frac{\partial}{\partial x} u^{(a)}(0) \right)^2 + \frac{1}{2} \left( k(1) \frac{\partial}{\partial x} u^{(a)}(1) \right)^2.$$

The question thus lies in finding the value of  $a$  that minimizes  $a \mapsto R(a)$ . Let us begin by studying the simple case of a *homogeneous* cord, for which the rigidity coefficient  $k(x) = \bar{k} > 0$  is constant. In this particular case, we can explicitly calculate  $R$  as a function of  $a$ . Indeed, multiplying [2.37] by  $x$  and integrating, we obtain

$$\begin{aligned} -\bar{k} \frac{\partial}{\partial x} u^{(a)}(1) &= - \int_0^1 p(x - a)x \, dx = - \int_{-\infty}^{+\infty} p(x - a)x \, dx \\ &= - \int_{-\infty}^{+\infty} p(x - a)(x - a) \, dx + a \int_{-\infty}^{+\infty} p(x - a) \, dx \\ &= - \int_{-\infty}^{+\infty} p(y)y \, dy + a \int_{-\infty}^{+\infty} p(y) \, dy \end{aligned}$$

where we have used the fact that  $p$  has compact support to extend the integration domain. We denote this relation as  $\bar{k} \frac{\partial}{\partial x} u^{(a)}(1) = \langle yp \rangle - a \langle p \rangle$ , which therefore reveals  $\bar{k} \frac{\partial}{\partial x} u^{(a)}(1)$  as an affine function of  $a$ . Likewise, by multiplying [2.37]  $(1 - x)$  by, we obtain the expression

$$\bar{k} \frac{\partial}{\partial x} u^{(a)}(0) = \langle yp \rangle + (1 - a) \langle p \rangle.$$

Therefore,

$$R(a) = \frac{1}{2} \left( |\langle yp \rangle - a \langle p \rangle|^2 + |\langle yp \rangle + (1 - a) \langle p \rangle|^2 \right)$$

is a quadratic function of  $a$ , for which it is easy to determine the minimum. In the general case, we no longer have a simple formula of this kind. We do not have an explicit formula either, and it is not even clear that classic optimization results can be applied.

We will nevertheless implement extrema determination algorithms, focusing on the discrete version of the problem. We introduce the numerical unknown  $U^{(a)} = (u_1^{(a)}, \dots, u_N^{(a)}) \in \mathbb{R}^N$ , a vector whose coordinates are designed to be approximations of the evaluation of  $u^{(a)}$  at the points  $ih$ ,  $i \in \{1, \dots, N\}$  of a uniform grid with step  $h$ :

$$x_0 = 0 < x_1 = h < \dots < x_i = ih < \dots < x_N = Nh < x_{N+1} = (N+1)h = 1$$

(note the relation between  $N$  and  $h = 1/(N+1)$ ). This quantity is defined as a solution to the linear system

$$\frac{1}{h} \left( -k_{i+1/2} \frac{u_{i+1}^{(a)} - u_i^{(a)}}{h} - k_{i-1/2} \frac{u_i^{(a)} - u_{i-1}^{(a)}}{h} \right) = -p(ih - a)$$

for  $i \in \{1, \dots, N\}$ , using the convention  $u_0^{(a)} = 0 = u_{N+1}^{(a)}$  and  $k_{i+1/2} = k(h(i + 1/2))$ . In the matrix form, we have  $MU^{(a)} = F^{(a)}$ , with  $F_i^{(a)} = -p(ih - a)$ , and  $M$  is the symmetric matrix whose coefficients are given by

$$M_{ii} = \frac{k_{i+1/2} + k_{i-1/2}}{h^2}, \quad M_{ii+1} = -\frac{k_{i+1/2}}{h^2},$$

and  $M_{ij} = 0$  when  $|i - j| > 1$ . Note that

$$MU \cdot U = \sum_{i=1}^N \frac{k_{i+1/2}}{h^2} (U_{i+1} - U_i)^2,$$

which proves that  $M$  is indeed invertible. We associate the discrete unknown  $U^{(a)}$  with the rupture energy

$$\begin{aligned} R_h(a) &= \frac{1}{2} \left( k_{1/2} \frac{u_1^{(a)} - u_0^{(a)}}{h} \right)^2 + \frac{1}{2} \left( k_{N+1/2} \frac{u_{N+1}^{(a)} - u_N^{(a)}}{h} \right)^2 \\ &= \frac{1}{2} \left( k_{1/2} \frac{u_1^{(a)}}{h} \right)^2 + \frac{1}{2} \left( k_{N+1/2} \frac{u_N^{(a)}}{h} \right)^2 \end{aligned}$$

taking into account the boundary conditions. We will seek to numerically determine the optimal position using the gradient algorithm (even though, in this case too, it is far from clear that the conditions that guarantee convergence are satisfied). We therefore construct the sequence

$$a_{n+1} = a_n - \rho R'_h(a_n)$$

where  $\rho > 0$  is a parameter to be determined. In order to express the derivative of the rupture energy  $R_h$ , we introduce the vector  $V^{(a)} = \frac{d}{da} U^{(a)}$ , and the chain rule yields

$$R'_h(a) = k_{1/2} u_1^{(a)} v_1^{(a)} + k_{N+1/2} u_N^{(a)} v_N^{(a)}. \quad [2.38]$$

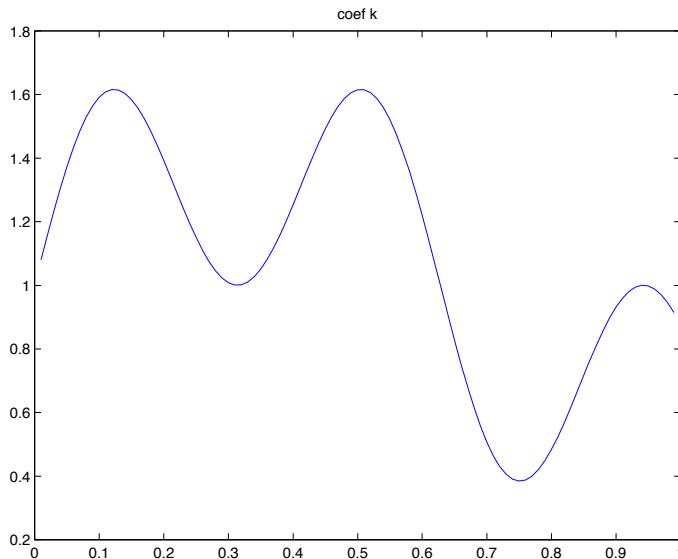
In order to put the algorithm in place, it is necessary to calculate the vector  $V^{(a)}$ . To this end, we take the derivative of the relation  $MU^{(a)} = F^{(a)}$  to obtain  $V^{(a)}$  as a solution of the linear system  $MV^{(a)} = G^{(a)}$ , where  $G^{(a)}$  has coordinates  $p'(x - a)$  (now assuming that  $p$  is  $C^1$ ).

We perform simulations in the case where the rigidity coefficient is given by

$$k(x) = 1 + 0.8 \sin(10x) \cos(5x),$$

for the graph of this function, see Figure 2.33. The applied force is obtained with the function

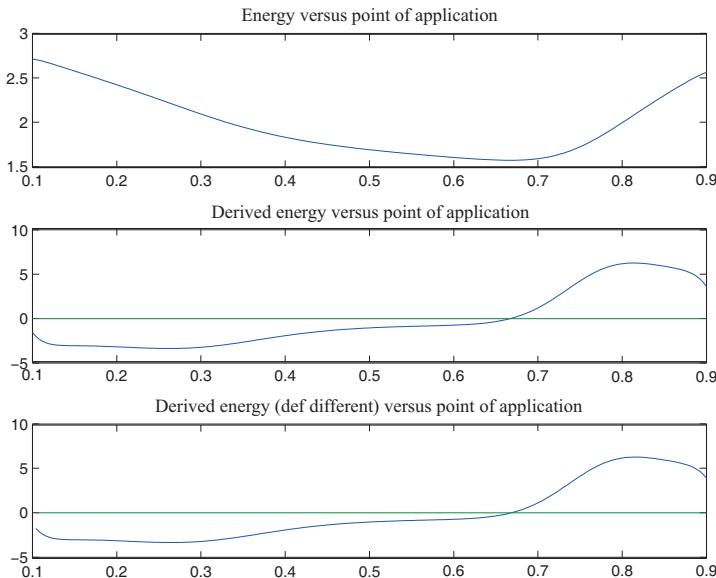
$$p(x) = 20 \exp(-(20x)^2).$$



**Figure 2.33.** Graph of the rigidity coefficient  $k$  as a function of position

It does not have compact support, but has significantly positive values only on a restricted domain (see the figures). We calculate the solution  $u^{(a)}$  for several values

of the parameter  $a$ , chosen between 0.1 and 0.9 using a uniform grid of  $2N$  points. Figure 2.34 shows the evolution of the rupture energy  $R_h(a)$  as a function of the charge position  $a$ . We also show the derivative  $\frac{d}{da}R_h(a)$ , which is calculated with either the formula [2.38] or the discrete derivation  $\frac{R_h(a_{k+1}) - R_h(a_k)}{a_{k+1} - a_k}$  for comparison. We observe that this derivative is zero for  $a$  that is approximately 0.66. Figure 2.36 shows the optimal position determined by the gradient algorithm, with  $\rho = 0.04$ . We stop the algorithm when the relative error  $\frac{|E(a_{k+1}) - E(a_k)|}{E(a_k)}$  is less than  $10^{-5}$ . We therefore find, in nine iterations starting from  $a_0 = 1/2$ , the optimal position  $a_{\text{opt}} = 0.666668$ , with  $R_h(a_{\text{opt}}) = 1.5708$  and  $\frac{d}{da}R_h(a_{\text{opt}}) = -3.5772 \times 10^{-5}$ . Figure 2.35 shows the configurations obtained for the consecutive iterations.



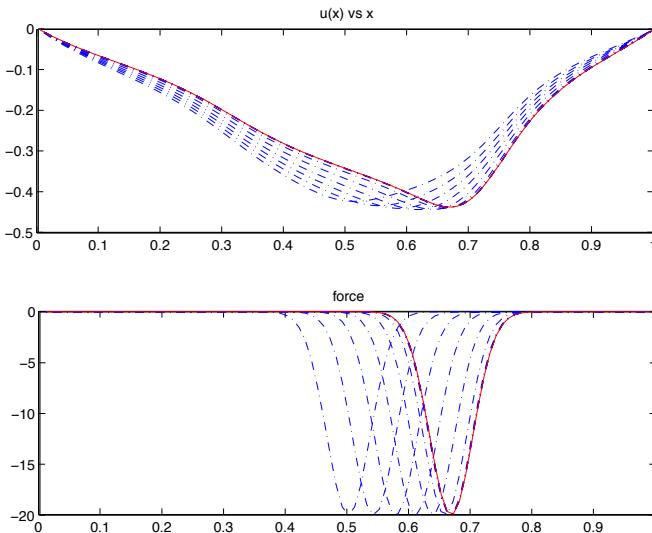
**Figure 2.34.** Evolution of the rupture energy and its derivative as a function of the point of the application force

## 2.10. Stokes problem

The Stokes equations describe the flow of an incompressible viscous fluid. We are going to study a stationary environment and first consider a two-dimensional context, where quantities depend only on the variables  $x$  and  $y$ . We let  $\nu > 0$  denote the viscosity of the fluid. External forces are described by a vector-valued function  $(x, y) \mapsto (F(x, y), G(x, y)) \in \mathbb{R}^2$ . The unknowns represent the velocity and pressure fields,  $(x, y) \mapsto (U(x, y), V(x, y)) \in \mathbb{R}^2$  and  $(x, y) \mapsto P(x, y) \in \mathbb{R}$ , respectively.

The conservation of momentum is reflected in the system of partial differential equations

$$\begin{aligned} \partial_x P - \nu(\partial_x^2 + \partial_y^2)U &= F, \\ \partial_y P - \nu(\partial_x^2 + \partial_y^2)V &= G. \end{aligned} \quad [2.39]$$



**Figure 2.35.** Successive configurations obtained with the gradient algorithm. For the color version of this figure, see [www.iste.co.uk/goudon/mathmodel.zip](http://www.iste.co.uk/goudon/mathmodel.zip)

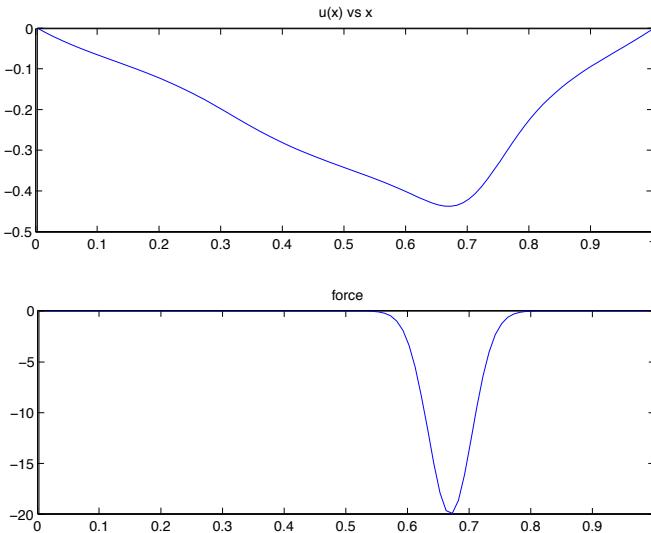
These equations show the equilibrium between the external forces on the one hand, and between pressure and viscous friction on the other hand. The fluid's incompressibility is reflected in the constraint

$$\partial_x U + \partial_y V = 0. \quad [2.40]$$

This means that a fluid domain may deform along the course of movement, but it retains a constant volume (see section 4.3 in [GOU 11]).

Indeed, consider a velocity field  $u : \mathbb{R} \times \mathbb{R}^N \rightarrow \mathbb{R}^N$ , which may also be dependent on time. We associate it with characteristic curves defined by the system of differential equations

$$\frac{d}{dt} X(t, x) = u(t, X(t, x)), \quad X(0, x) = x.$$



**Figure 2.36.** Optimal configuration determined by the gradient algorithm

In other words,  $X(t, x) \in \mathbb{R}^N$  is the position in time  $t \geq 0$  of a particle responding to the velocity field  $u$ , which starts at an instant  $t = 0$  at the position  $x \in \mathbb{R}^N$ . We fix a domain  $D_0$  and for  $t > 0$  we let  $D(t)$  denote its image at instant  $t$

$$D(t) = \{X(t, x_0), x_0 \in D_0\},$$

(see Figure 2.37). We consider the volume

$$\mathcal{V}(t) = \int_{D(t)} dx.$$

With the change of variable  $x = X(t, x_0)$ , we have

$$\frac{d}{dt} \mathcal{V}(t) = \frac{d}{dt} \left( \int_{D_0} |\det(A(t, x_0))| dx_0 \right) = \int_{D_0} \frac{d}{dt} |\det(A(t, x_0))| dx_0$$

where  $A(t, x_0)$  denotes the Jacobian matrix of that change of variables, that is,  $A_{ij}(t, x_0) = \frac{\partial}{\partial x_{0j}} X_i(t, x_0)$ . For  $t = 0$ , we have  $A(0, x_0) = \mathbb{I}$ , so  $\det(A(t, x_0))$  remains positive for all times. Now, we have

$$\begin{aligned} \det(A + H) &= \det(A(\mathbb{I} + A^{-1}H)) = \det(A)\det(\mathbb{I} + A^{-1}H) \\ &= \det(A)(1 + \text{Tr}(A^{-1}H) + \|H\|\epsilon(H)) \end{aligned}$$

where  $\lim_{\|H\| \rightarrow 0} \epsilon(H) = 0$ . We must therefore evaluate

$$\frac{d}{dt} \det A(t, x_0) = \det(A(t, x_0)) \operatorname{Tr} \left( A(t, x_0)^{-1} \frac{d}{dt} A(t, x_0) \right).$$

Returning to the differential system, we realize that

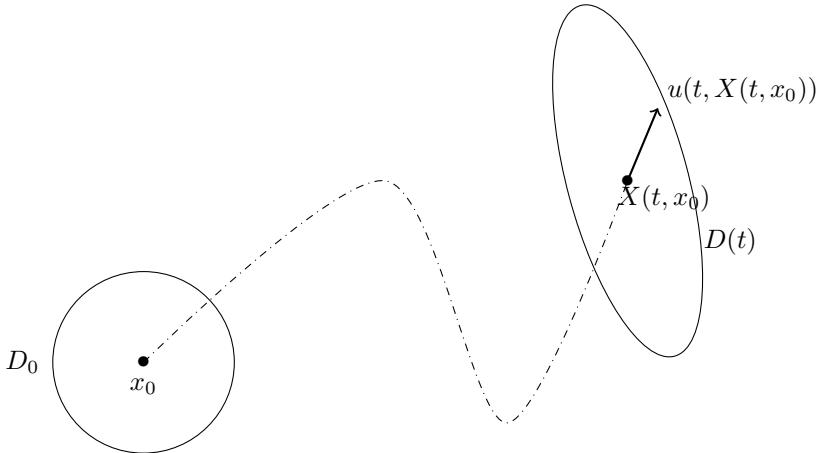
$$\frac{d}{dt} \frac{\partial}{\partial x_{0_j}} X_i(t, x_0) = \sum_{\ell=1}^N (\partial_\ell u_i)(t, X(t, x_0)) \frac{\partial}{\partial x_{0_j}} X_\ell(t, x_0),$$

with  $\frac{\partial}{\partial x_{0_j}} X_i(0, x_0) = \mathbb{I}$ . In the matrix form, this can be represented as

$$\frac{d}{dt} A(t, x_0) = \nabla u(t, X(t, x_0)) A(t, x_0),$$

with  $A(0, x_0) = \mathbb{I}$ , where  $\nabla u$  is the Jacobian matrix of  $u$ . It follows that

$$\begin{aligned} \frac{d}{dt} \det A(t, x_0) &= \det(A(t, x_0)) \operatorname{Tr} \left( A(t, x_0)^{-1} \nabla u(t, X(t, x_0)) A(t, x_0) \right) \\ &= \det(A(t, x_0)) \operatorname{Tr}(\nabla u(t, X(t, x_0))) \\ &= \det(A(t, x_0)) \operatorname{div}(u)((t, X(t, x_0))). \end{aligned}$$



**Figure 2.37.** Evolution of the domain  $D_0$

We conclude that

$$\frac{d}{dt} \mathcal{V}(t) = \int_{D_0} \det(A(t, x_0)) \operatorname{div}(u)(t, X(t, x_0)) dx_0 = \int_{D(t)} \operatorname{div}(u)(t, x) dx$$

using the change of variables  $x = X(t, x_0)$ ,  $\mathrm{d}x = \det(A(t, x_0)) \mathrm{d}x_0$ . Therefore, when the field  $u$  is such that  $\operatorname{div}(u) = 0$ , the volume  $\mathcal{V}(t)$  remains constant, although the shape of the domain may indeed change throughout time.

We can determine explicitly some specific solutions for this problem. We assume that the fluid is subject to the effects of gravity, which act in the direction of  $x$  (which is the “vertical” direction):  $F(x, y) = -g$ ,  $G(x, y) = 0$ , is constant. Moreover, the fluid flows between two plates with coordinates  $y = 0$  and  $y = L$ . We seek a solution in the form of functions independent of the variable  $x$ . We thus obtain

$$\begin{cases} 0 - \nu U''(y) = -g, \\ P'(y) - \nu V''(y) = 0, \\ V'(y) = 0. \end{cases}$$

These equations are completed by the boundary conditions

$$U(0) = U(L) = 0 = V(0) = V(L),$$

which express the fact that the fluid adheres to the walls of the domain in which it flows. It follows that  $V(y) = 0$ ,  $P(y) = \bar{P} \in \mathbb{R}$  is constant, and  $U$  is shaped like the parabola given by

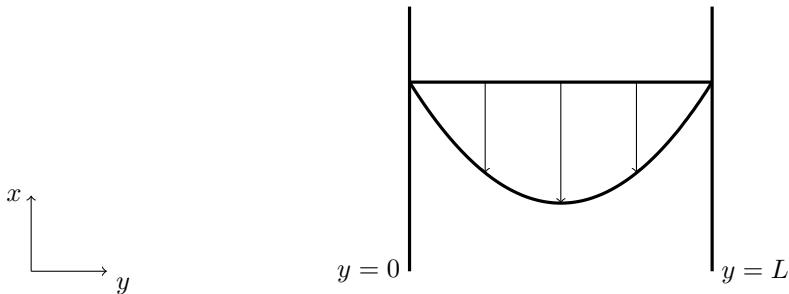
$$U(y) = \frac{g}{2\nu}y(y - L).$$

This situation is known as a *Poiseuille flow*. In practice, we can only seldom identify explicit situations of this kind and the use of numerical simulations is inevitable when trying to describe flows. We are therefore led to search for a relevant definition of approximated solutions. The analysis of problem [2.39]–[2.40] and its approximations calls for the use of fairly sophisticated functional tools. We will limit ourselves to studying a simplified version of the problem, inspired by [JOH 02], which deals with the most important challenges of the problem (the reader may refer to Part 1 of [TAR 82] for an introduction to the analysis methods for the Stokes and Navier–Stokes equations, or [BOY 06] for a more detailed and advanced treatment).

Henceforth, we will assume that the flow occurs within the domain  $[0, L] \times [0, 2\pi]$  and that boundary conditions are of the form

$$U(0, y) = U(L, y) = 0, \quad V(0, y) = V(L, y),$$

$$U(x, 0) = U(x, 2\pi) = 0, \quad \partial_y V(x, 0) = \partial_y V(x, 2\pi) = 0.$$



**Figure 2.38.** Velocity profile for a Poiseuille flow

We also assume that the force terms are of the form

$$F(x, y) = f(x) \sin(ky), \quad G(x, y) = g(x) \cos(ky), \quad k \in \mathbb{N} \setminus \{0\}.$$

We are looking for solutions that can be expressed in the same form

$$U(x, y) = u(x) \sin(ky), \quad V(x, y) = v(x) \cos(ky), \quad P(x, y) = p(x) \sin(ky).$$

Thus, equations [2.39]–[2.40] give

$$\begin{cases} \frac{d}{dx}p + \nu \left( k^2 - \frac{d^2}{dx^2} \right)u = f, \\ kp + \nu \left( k^2 - \frac{d^2}{dx^2} \right)v = g, \\ \frac{d}{dx}u - kv = 0, \end{cases} \quad [2.41]$$

with the boundary conditions  $u(0) = u(L) = 0 = v(0) = v(L)$ . Let  $\phi, \psi$  be regular functions, for example  $C_c^\infty([0, 1])$ . Using integration by parts, we obtain, with  $(u, v, p)$  being a solution of [2.41]:

$$\begin{aligned} & \nu \int_0^L \left( k^2(u\phi + v\psi) + \frac{d}{dx}u \frac{d}{dx}\phi + \frac{d}{dx}v \frac{d}{dx}\psi \right) dx \\ & + \int_0^L \left( \frac{d}{dx}p\phi + kp\psi \right) dx = \int_0^L (f\phi + g\psi) dx \end{aligned}$$

Note that if  $x \mapsto (\phi(x), \psi(x))$  satisfies the incompressibility constraint  $\frac{d}{dx}\phi - k\psi = 0$  and equals zero at  $x = 0$  and  $x = L$ , then we obtain

$$\int_0^L \left( \frac{d}{dx} p\phi + kp\psi \right) dx = \int_0^L p \left( -\frac{d}{dx}\phi + k\psi \right) dx = 0.$$

These remarks motivate us to introduce the following functional framework. We write

$$H_{0,k}^1 = \left\{ x \in [0, L] \mapsto (u(x), v(x)), u, v \in L^2([0, L]), \frac{d}{dx}u, \frac{d}{dx}v \in L^2([0, L]), \right. \\ \left. u(0) = u(L) = 0 = v(0) = v(L), \quad \frac{d}{dx}u - kv = 0 \right\}.$$

We equip this space with the usual norm  $H^1$ . It is indeed a closed subspace of  $H_0^1 \times H_0^1([0, L])$ . On this space, we define

$$a(u, v; \phi, \psi) = \nu \int_0^L \left( k^2(u\phi + v\psi) + \frac{d}{dx}u \frac{d}{dx}\phi + \frac{d}{dx}v \frac{d}{dx}\psi \right) dx$$

and

$$\ell(\phi, \psi) = \int_0^L (f\phi + g\psi) dx.$$

It is clear that  $a$  is a continuous and coercive bilinear form on  $H_{0,k}^1$  and  $\ell$  is a continuous linear form on  $H_{0,k}^1$  since  $f, g \in L^2([0, L])$ . From the Lax–Milgram theorem, we can infer that:

**THEOREM 2.16.**— for all  $f, g \in L^2([0, L])$ , there exists a unique element  $(u, v) \in H_{0,k}^1$ , such that  $a(u, v; \phi, \psi) = \ell(\phi, \psi)$  for all  $(\phi, \psi) \in H_{0,k}^1$ .

In this formulation, the pressure  $p$  disappears! In fact, it is constraint in the defining the functional framework. Allowing us to realize another integration by parts (which is, in fact, not correct because we cannot be sure that  $\frac{d^2}{dx^2}u$  and  $\frac{d^2}{dx^2}v$  are in  $L^2$ ), the relation  $a(u, v; \phi, \psi) = \ell(\phi, \psi)$  means that the vector

$$\begin{pmatrix} W \\ Z \end{pmatrix} = \begin{pmatrix} \nu k^2 u - \nu \frac{d^2}{dx^2} u - f \\ \nu k^2 v - \nu \frac{d^2}{dx^2} v - g \end{pmatrix} \quad [2.42]$$

satisfies

$$\int_0^L \begin{pmatrix} W \\ Z \end{pmatrix} \cdot \begin{pmatrix} \phi \\ \psi \end{pmatrix} dx = 0 \quad [2.43]$$

for every vector-valued test function  $(\phi, \psi)$  that *satisfies the constraint*  $\frac{d}{dx}\phi - k\psi = 0$ . The challenge is to identify the set of vectors  $(W, Z)$  that satisfy this orthogonality property. Therefore, if we have  $W(x) = \frac{d}{dx}p(x)$  and  $Z(x) = kp(x)$  for a certain scalar function  $x \mapsto p(x)$ , then, as seen above, [2.43] is satisfied by all  $(\phi, \psi) \in H_{0,k}^1$ . We will provide arguments to support the claim that it is the only possible solution, although the functional framework required for a full proof is quite elaborate.

**LEMMA 2.10** (Hodge decomposition).— Let  $U, V \in L^2([0, L[)$ . Then, there exist functions  $p \in H^1([0, L[)$ ,  $R, S \in L^2([0, L[)$ , such that  $U = \frac{d}{dx}p + R$ ,  $V = kp + S$  and for all  $\phi \in H^1([0, L[)$ , we have  $\int_0^L (R \frac{d}{dx}\phi + Sk\phi) dx = 0$ .

**PROOF.**— If all the functions are regular and the proposed decomposition holds, then we obtain the following equation defining  $p$ :

$$-\frac{d}{dx}U + kV = -\frac{d^2}{dx^2}p + k^2p - \frac{d}{dx}R + kS = -\frac{d^2}{dx^2}p + k^2p.$$

This suggests a definition for  $p$  as a solution to the variational problem  $\bar{a}(p, \phi) = \bar{\ell}(\phi)$ , with

$$\bar{a}(p, \phi) = \int_0^L \left( k^2p\phi + \frac{d}{dx}p \frac{d}{dx}\phi \right) dx, \quad \bar{\ell}(\phi) = \int_0^L \left( U \frac{d}{dx}\phi + kV\phi \right) dx,$$

for all  $\phi \in H^1([0, L[)$ . The Lax–Milgram theorem ensures the existence and uniqueness of a solution  $p \in H^1([0, L[)$  (if  $k = 0$ , uniqueness is guaranteed by further assuming that  $\int_0^L p dx = 0$ ). Implicitly, we therefore impose the Neumann condition for  $p$ . We then write  $R = U - \frac{d}{dx}p$  and  $S = V - kp$ , which are both elements of  $L^2([0, L[)$ . By definition, we have

$$\begin{aligned} \int_0^L \left( R \frac{d}{dx}\phi + kS\phi \right) dx &= \int_0^L \left( U \frac{d}{dx}\phi + kV\phi \right) dx \\ &\quad - \int_0^L \left( \frac{d}{dx}p \frac{d}{dx}\phi + k^2p\phi \right) dx = 0. \end{aligned}$$

This decomposition is, in fact, unique. Indeed, if there are two such decompositions  $U = \frac{d}{dx}p + R = \frac{d}{dx}q + R'$  and  $V = kp + S = kq + S'$ , then  $\pi = p - q$ ,  $\rho = R - R'$  and  $\sigma = S - S'$  satisfy  $\frac{d}{dx}\pi + \rho = 0 = k\pi + \sigma$ . Moreover, the relation  $\int_0^L (\rho \frac{d}{dx}\phi + \sigma k\phi) dx = 0$  is satisfied for all  $\phi \in H^1([0, L[)$ . However, this can be rewritten as  $\bar{a}(\pi, \phi) = 0$ . It follows that  $\pi = 0$  and then  $\rho = \sigma = 0$ .  $\square$

**NOTE 2.9.**— Throughout this discussion, the crucial point, which allows us to extend the argument to higher dimensions, is the duality between *gradient* and *divergence*

operators: for  $u : \Omega \subset \mathbb{R}^N \rightarrow \mathbb{R}^N$  a vector-valued function, such that  $u|_{\partial\Omega} = 0$ , and  $p : \Omega \rightarrow \mathbb{R}$  a scalar function, we have (with the appropriate regularity assumptions)

$$\int_{\Omega} \nabla p \cdot u \, dX = - \int_{\Omega} p \operatorname{div}(u) \, dX.$$

Here, in dimension  $N = 2$  ( $X = (x, y)$ ), the gradient operator

$$\nabla : p \mapsto \begin{pmatrix} \partial_x p \\ \partial_y p \end{pmatrix}$$

is replaced by the operator

$$\mathcal{D} : p \mapsto \begin{pmatrix} \partial_x p \\ kp \end{pmatrix}$$

and the duality relation becomes

$$\int_0^L \mathcal{D}p \cdot u \, dx = - \int_0^L p \mathcal{D}^* u \, dx,$$

with

$$\mathcal{D}^* : u = \begin{pmatrix} u_1 \\ u_2 \end{pmatrix} = \partial_x u_1 - ku_2,$$

which plays the role of the operator  $\operatorname{div}(u) = \partial_x u_1 + \partial_y u_2$ . It is necessary to understand lemma 2.10 as a decomposition  $u = \mathcal{D}p + \Phi$ , with  $\mathcal{D}^*\Phi = 0$ , which is of the form  $\operatorname{Ran}(\mathcal{D}) + \operatorname{Ker}(\mathcal{D}^*)$ . The multidimensional analogue is obtained by writing  $u = \nabla p + \Phi$ , with  $\Phi$  a vector function satisfying  $\operatorname{div}(\Phi) = 0$ . It is important to keep these remarks in mind to properly understand the arguments presented here.

This statement is important and has several applications in multidimensional versions. Here, it is not entirely sufficient for identifying the vector  $(W, Z)$  in [2.42]. Indeed, if we assume that this vector is an element of  $L^2([0, L[)$  (which is not the case in general), we can decompose  $(W, Z) = (\frac{d}{dx}p + R, kp + S)$  and obtain  $\int_0^L (W, Z) \cdot (\phi, \psi) \, dx = \int_0^L (R, S) \cdot (\phi, \psi) \, dx = 0$  for all  $(\phi, \psi) \in H_{0,k}^1$ , in addition to  $\int_0^L (R \frac{d}{dx}\phi + Sk\phi) \, dx = 0$  for all  $\phi \in H^1([0, L[)$ . However, this does not allow us to conclude that  $R = S = 0$ . We thus require more sophisticated functional tools.

*Step 1.* We begin by introducing  $H^{-1}([0, L[)$ , the dual space of  $H_0^1([0, L[)$ . It is the set of continuous linear forms on  $H_0^1([0, L[)$ : a linear form  $\ell$  on  $H_0^1([0, L[)$  is an

element of  $H^{-1}(]0, L[)$  when there exists a  $C > 0$ , such that for all  $u \in H_0^1(]0, L[)$ , we have

$$|\ell(u)| \leq C\|u\|_{H^1}, \quad \text{where } \|u\|_{H^1}^2 = \int_0^L \left( |u(x)|^2 + \left| \frac{d}{dx} u(x) \right|^2 \right) dx.$$

We can then write

$$\|\ell\|_{H^{-1}} = \sup_{u \in H_0^1 \setminus \{0\}} \frac{|\ell(u)|}{\|u\|_{H^1}}.$$

A key example is given in the following lemma, whose proof is a direct consequence of the Cauchy–Schwarz inequality.

LEMMA 2.11.– Let  $p \in L^2(]0, 1[)$ . We associate it with the linear form denoted as  $\frac{dp}{dx}$  and defined by

$$\text{for any } \varphi \in H_0^1(]0, L[), \frac{dp}{dx}(\varphi) = - \int_0^L p(x) \frac{d}{dx} \varphi(x) dx.$$

Then,  $\frac{dp}{dx} \in H^{-1}(]0, L[)$  with  $\|\frac{dp}{dx}\|_{H^{-1}} \leq \|p\|_{L^2}$ .

*Step 2.* Using the same reasoning, we can define the derivative  $\frac{d\ell}{dx}$  of a linear form  $\ell$  of  $H^{-1}(]0, L[)$  by writing

$$\text{for any } \varphi \in C_c^\infty(]0, L[), \frac{d\ell}{dx}(\varphi) = -\ell\left(\frac{d\varphi}{dx}\right).$$

We thus consider the following subspace:

$$\mathcal{X} = \left\{ \ell \in H^{-1}(]0, L[), \frac{d\ell}{dx} \in H^{-1}(]0, L[) \right\}.$$

If we let  $\ell \in \mathcal{X}$ , it therefore implies that there exists a  $C > 0$ , such that for all  $\varphi \in C_c^\infty(]0, L[)$ , we have

$$\left| \frac{d\ell}{dx}(\varphi) \right| \leq C\|\varphi\|_{H^1}. \quad [2.44]$$

(note that for the elements of  $H^{-1}$ , the estimate would additionally involve the norm of the second derivative of  $\varphi$  in  $L^2$ ).

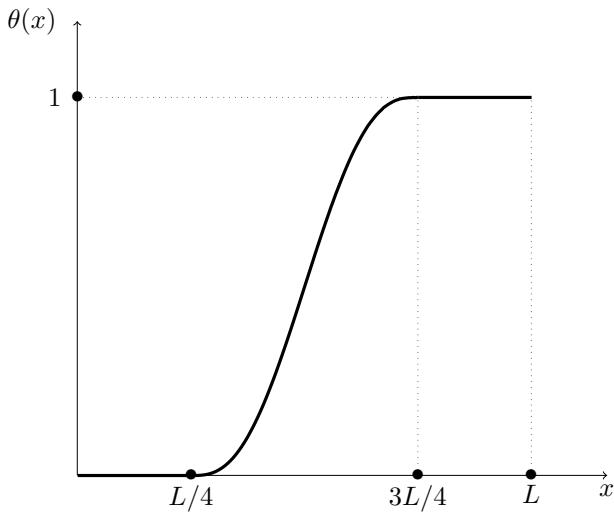
LEMMA 2.12.– We have  $\mathcal{X} = L^2([0, L])$ . Moreover, there exists two constants  $c, \bar{c} > 0$ , such that the equivalence of norms

$$c\|p\|_{L^2} \leq \|p\|_{\mathcal{X}} := \|p\|_{H^{-1}} + \left\| \frac{dp}{dx} \right\|_{H^{-1}} \leq \bar{c}\|p\|_{L^2}$$

is satisfied for all  $p \in L^2([0, L])$ .

PROOF.– Lemma 2.11 ensures the inclusion  $L^2([0, L]) \subset \mathcal{X}$  since it is clear that  $L^2([0, L]) \subset H^{-1}([0, L])$ . In particular, note that  $\|p\|_{\mathcal{X}} \leq 2\|p\|_{L^2}$ . The difficulty is in proving the inverse inclusion. Let  $\psi \in C_c^\infty([0, L])$ . We fix a  $C^\infty$  function  $\theta$  on  $[0, L]$ , such that

$$\begin{aligned} 0 &\leq \theta(x) \leq 1, \quad \text{for } x \in [0, L], \\ \theta(x) &= 1, \quad \text{for } x \in [3L/4, L], \\ \theta(x) &= 0, \quad \text{for } x \in [0, L/4]. \end{aligned}$$



**Figure 2.39.** Function  $x \mapsto \theta(x)$

Finally, we set

$$\varphi(x) = \int_0^x \psi(y) dy - \theta(x) \int_0^L \psi(y) dy.$$

This function is still  $C^\infty$  and its support is included in the open set  $]0, L[$ . Then, for  $\ell \in \mathcal{X}$ , we have

$$\begin{aligned}\ell(\psi) &= \ell\left(\frac{d}{dx}\varphi + \int_0^L \psi(y) dy \times \frac{d}{dx}\theta\right) \\ &= \ell\left(\frac{d}{dx}\varphi\right) + \int_0^L \psi(y) dy \times \ell\left(\frac{d}{dx}\theta\right).\end{aligned}$$

Since  $\ell \in \mathcal{X}$ , we have

$$\left|\frac{d\ell}{dx}(\varphi)\right| \leq \|\ell\|_{\mathcal{X}} \|\varphi\|_{H^1}.$$

Furthermore, the Cauchy–Schwarz inequality makes it possible to dominate  $\left|\int_0^L \psi(y) dy\right| \leq \sqrt{L} \|\psi\|_{L^2}$ . It follows that

$$\|\varphi\|_{H^1} \leq \|\psi\|_{L^2} \sqrt{2(1 + L^2 + L\|\theta\|_{H^1})}.$$

We let  $m_1$  denote the constant, which is dependent on  $L$   $\theta$ , involved in this estimate. Finally,  $\frac{d\theta}{dx} \in C_c^\infty(]0, L[)$ , so that  $\ell\left(\frac{d}{dx}\theta\right)$  is well estimated and dominated by  $\|\ell\|_{H^{-1}} \|\frac{d\theta}{dx}\|_{H^1} \leq \|\ell\|_{\mathcal{X}} \|\frac{d\theta}{dx}\|_{H^1}$ , we denote the right hand side as  $\|\ell\|_{\mathcal{X}} m_2$ . We thus obtain

$$|\ell(\psi)| = \left|\ell\left(\frac{d}{dx}\varphi\right) + \int_0^L \psi(y) dy \times \ell\left(\frac{d}{dx}\theta\right)\right| \leq \|\psi\|_{L^2} \|\ell\|_{\mathcal{X}} (m_1 + \sqrt{L} m_2).$$

Then,  $\ell$  is extended as a continuous linear form on  $L^2(]0, L[)$ , and by the Riesz theorem (theorem 4.37 in [GOU 11]), we can identify it to a function  $p \in L^2(]0, L[)$ :

$$\ell(\psi) = \int_0^L p(x) \psi(x) dx.$$

□

**NOTE 2.10.–** Lemma 2.12 can be adapted to higher dimensions where it is known as “Necas’s lemma” [NEC 96]. In the whole space (respectively in a periodic domain), it can be proved directly by using the Fourier transform (respectively Fourier series). Recall that  $\widehat{\nabla\phi}(\xi) = i\xi\widehat{\phi}(\xi)$  and therefore

$$\|\phi\|_{H^1}^2 = \frac{1}{(2\pi)^N} \int (1 + \xi^2) |\widehat{\phi}(\xi)|^2 d\xi.$$

This allows us to define the dual norm

$$\|\ell\|_{H^{-1}} = \frac{1}{(2\pi)^N} \int \frac{|\widehat{\ell}(\xi)|^2}{1 + \xi^2} d\xi.$$

We thus write

$$|\widehat{p}(\xi)|^2 = \frac{|\widehat{p}(\xi)|^2}{1+\xi^2} + \frac{\xi^2|\widehat{p}(\xi)|^2}{1+\xi^2} = \frac{|\widehat{p}(\xi)|^2}{1+\xi^2} + \frac{|\widehat{\nabla p}(\xi)|^2}{1+\xi^2}.$$

By integrating and using the Plancherel (theorem 6.11 in [GOU 11]), we find

$$\underbrace{\int |\widehat{p}(\xi)|^2 d\xi}_{(2\pi)^N \|p\|_{L^2}^2} = \underbrace{\int \frac{|\widehat{p}(\xi)|^2}{1+\xi^2} d\xi}_{(2\pi)^N \|p\|_{H^{-1}}^2} + \underbrace{\int \frac{|\widehat{\nabla p}(\xi)|^2}{1+\xi^2} d\xi}_{(2\pi)^N \|\nabla p\|_{H^{-1}}^2}.$$

The proof of this statement on a domain  $\Omega$  is somewhat technical (see, for example, section 1, Chapter IV in [BOY 06]).

*Step 3.* We now move on to vector-valued objects. We seek to identify

$$\mathcal{Y} = \left\{ \begin{pmatrix} \frac{dp}{dx} \\ kp \end{pmatrix}, p \in L^2([0, L]) \right\} \subset H^{-1}([0, L]),$$

with

$$(H_{0,k}^1)^\perp = \left\{ \begin{pmatrix} U \\ V \end{pmatrix} \in H^{-1}([0, L]), U(\phi) + V(\psi) = 0, \text{ for any } \begin{pmatrix} \phi \\ \psi \end{pmatrix} \in H_0^1([0, L]) \text{ such that } \frac{d\phi}{dx} - k\psi = 0 \right\}.$$

We know that

$$\mathcal{Y} \subset (H_{0,k}^1)^\perp,$$

since for all  $p \in L^2([0, L])$  and all  $(\phi, \psi) \in H_{0,k}^1$ , we have

$$\frac{dp}{dx}(\phi) + kp(\psi) = - \int_0^L p(x) \left( \frac{d\phi}{dx} - k\psi \right) dx = 0.$$

Suppose it is proved that

$$\mathcal{Y} \text{ is a closed subspace of } H^{-1}. \quad [2.45]$$

Then, we can conclude from the inclusion<sup>12</sup>

$$\mathcal{Y}^\perp \subset H_{0,k}^1. \quad [2.46]$$

Indeed, it implies<sup>13</sup>  $(H_{0,k}^1)^\perp \subset (\mathcal{Y}^\perp)^\perp = \overline{\mathcal{Y}}$ , with  $\overline{\mathcal{Y}} = \mathcal{Y}$ , by [2.45]. Now, if for all  $p \in L^2(]0, L[)$ ,  $(u, v) \in H_0^1$  satisfies

$$\int_0^L p(x) \left( -\frac{du}{dx}(x) + kv(x) \right) dx = 0$$

(which is equivalent to  $(u, v) \in \mathcal{Y}^\perp$ ), then we have  $-\frac{du}{dx}(x) + kv(x) = 0$  a. e.  $x \in ]0, L[$  so  $(u, v) \in H_{0,k}^1$ . To finish the proof, it remains to show [2.45], which is a product of the following statement.

LEMMA 2.13.– For  $k \neq 0$ , the (linear and continuous) operator

$$A : p \in L^2(]0, L[) \longmapsto \begin{pmatrix} \frac{dp}{dx} \\ kp \end{pmatrix} \in H^{-1}(]0, L[)$$

has a closed range.

PROOF.– Consider a sequence  $(p_n)_{n \in \mathbb{N}}$ , such that  $(Ap_n)_{n \in \mathbb{N}}$  converges in  $H^{-1}(]0, L[)$ : there exists  $p, q \in H^{-1}(]0, L[)$ , such that  $\lim_{n \rightarrow \infty} p_n = p$  and  $\lim_{n \rightarrow \infty} \frac{dp_n}{dx} = q$  in  $H^{-1}(]0, L[)$ . This implies that for all  $\varphi \in C_c^\infty(]0, L[)$ , we have

$$\int_0^L p_n \varphi dx \xrightarrow{n \rightarrow \infty} p(\varphi)$$

as well as

$$\frac{dp_n}{dx}(\varphi) = - \int_0^L p_n(x) \frac{d\varphi}{dx}(x) dx \xrightarrow{n \rightarrow \infty} q(\varphi) = -p\left(\frac{d\varphi}{dx}\right)$$

which indeed means that  $q = \frac{dp}{dx}$ . In fact, we can show that the operator  $\frac{d}{dx}$ , which is linear and continuous for  $L^2(]0, L[)$  in  $H^{-1}(]0, L[)$ , has a closed range and extend this property to all dimensions. This property is a consequence of a general statement, which is detailed in Appendix 3.  $\square$

<sup>12</sup> Here, for a subspace  $Z \subset H^{-1} \times H^{-1}$ , we denote as  $Z^\perp$  the set of  $(u, v) \in H_0^1$ , such that for all  $(\lambda, \mu) \in Z$ , we have  $\lambda(u) + \mu(v) = 0$ ; for more information on these orthogonality notions, the reader may refer to [BRÉ 05, section II.5].

<sup>13</sup> See [BRÉ 05, proposition II.12]

NOTE 2.11 (Periodic framework).— We can advance the same discussion by assuming that the functions are ( $L = 1$ )—periodic. The result is thus obtained in a more direct fashion, by arguing through Fourier series. We introduce the vector  $e(j) = (2i\pi j, -k)$ . The constraint defining  $H_{0,k}^1$  is thus represented in terms of Fourier coefficients

$$2i\pi j \hat{\phi}(j) - k \hat{\psi}(j) = 0 = e(j) \cdot \begin{pmatrix} \hat{\phi}(j) \\ \hat{\psi}(j) \end{pmatrix},$$

for all  $j \in \mathbb{Z}$ . We decompose any vector  $(\hat{\phi}, \hat{\psi})$  in the form

$$\begin{pmatrix} \hat{\phi} \\ \hat{\psi} \end{pmatrix} = \left( \mathbb{I} - \frac{e(j) \otimes e(j)}{|e(j)|^2} \right) \begin{pmatrix} \hat{\phi} \\ \hat{\psi} \end{pmatrix} + \frac{e(j) \otimes e(j)}{|e(j)|^2} \begin{pmatrix} \hat{\phi} \\ \hat{\psi} \end{pmatrix},$$

where

$$e(j) \cdot \left( \mathbb{I} - \frac{e(j) \otimes e(j)}{|e(j)|^2} \right) \begin{pmatrix} \hat{\phi} \\ \hat{\psi} \end{pmatrix} = 0.$$

For any vector-valued function  $x \mapsto (\phi, \psi)(x)$ , by [2.42]–[2.43],  $(W, Z)$  satisfies

$$\sum_{j \in \mathbb{Z}} \begin{pmatrix} \widehat{W}(j) \\ \widehat{Z}(j) \end{pmatrix} \cdot \left( \mathbb{I} - \frac{e(j) \otimes e(j)}{|e(j)|^2} \right) \begin{pmatrix} \widehat{\phi}(j) \\ \widehat{\psi}(j) \end{pmatrix} = 0$$

and therefore

$$\begin{aligned} \int_0^1 \begin{pmatrix} W \\ Z \end{pmatrix} \cdot \begin{pmatrix} \phi \\ \psi \end{pmatrix} dx &= \sum_{j \in \mathbb{Z}} \begin{pmatrix} \widehat{W}(j) \\ \widehat{Z}(j) \end{pmatrix} \cdot \begin{pmatrix} \widehat{\phi}(j) \\ \widehat{\psi}(j) \end{pmatrix} \\ &= \sum_{j \in \mathbb{Z}} \begin{pmatrix} \widehat{W}(j) \\ \widehat{Z}(j) \end{pmatrix} \cdot \frac{e(j) \otimes e(j)}{|e(j)|^2} \begin{pmatrix} \widehat{\phi}(j) \\ \widehat{\psi}(j) \end{pmatrix} \\ &= \sum_{j \in \mathbb{Z}} \frac{e(j) \otimes e(j)}{|e(j)|^2} \begin{pmatrix} \widehat{W}(j) \\ \widehat{Z}(j) \end{pmatrix} \cdot \begin{pmatrix} \widehat{\phi}(j) \\ \widehat{\psi}(j) \end{pmatrix}. \end{aligned}$$

Since this equality holds for all  $(\phi, \psi)$ , we conclude that

$$\begin{pmatrix} \widehat{W}(j) \\ \widehat{Z}(j) \end{pmatrix} = \frac{e(j) \otimes e(j)}{|e(j)|^2} \begin{pmatrix} \widehat{W}(j) \\ \widehat{Z}(j) \end{pmatrix} = e(j) \widehat{p}(j)$$

where  $p$  is the scalar function whose Fourier coefficients are given by

$$\widehat{p}(j) = \frac{e(j)}{|e(j)|^2} \cdot \begin{pmatrix} \widehat{W}(j) \\ \widehat{Z}(j) \end{pmatrix}.$$

In other words,  $W = \frac{d}{dx} p$  and  $Z = kp$ .

**NOTE 2.12.–** Stokes problem [2.39]–[2.40], presented within a domain  $\Omega$ , must be completed by boundary conditions. They affect the velocity  $U$  and do not involve the pressure. A simple model consists of assuming that the fluid adheres to the walls, which allows us to impose the constraint  $U|_{\partial\Omega} = 0$ . Note that the pressure  $P$  in [2.39]–[2.40] is defined only up to a constant: if  $x \mapsto (U(x), P(x))$  is a solution, then for all  $C \in \mathbb{R}$ ,  $x \mapsto (U(x), P(x) + C)$  is also a solution. For [2.41], we have chosen that constant: we assume the solution  $P(x, y)$  satisfies  $\int P dy dx = 0$ . From a mathematical point of view, the pressure is intimately associated with the incompressibility constraint, and we will see further below how to interpret it in terms of Lagrange multipliers. For [2.41], we can obtain an equation for  $p$  by combining the information as follows:

$$\begin{aligned} & \frac{d}{dx} \left( \frac{d}{dx} p + \nu k^2 u - \nu \frac{d^2}{dx^2} u \right) - k \left( kp + \nu k^2 v - \nu \frac{d^2}{dx^2} v \right) \\ &= \frac{d}{dx} f - kg \\ &= \frac{d^2}{dx^2} p - k^2 p - \nu \left( \frac{d^2}{dx^2} - k^2 \right) \left( \frac{d}{dx} u - kv \right) = \frac{d^2}{dx^2} p - k^2 p. \end{aligned} \quad [2.47]$$

Therefore,  $p$  satisfies an elliptic equation, whose right-hand side is determined by the data  $f$  and  $g$ . However, the difficulty is in the fact that we do not have boundary conditions for  $p$ .

A “naive” discretization of the system [2.41] with finite differences leads to the following discrete equations. We fix the constant step  $\Delta x > 0$  and use the following notation:  $x_0 = 0 < x_1 = \Delta x < \dots < x_j = j\Delta x < \dots < x_J < L = (J+1)\Delta x = x_{J+1}$ . Naturally, the numerical unknown should be constituted by the vectors  $U^{(J)} = (u_1, \dots, u_J)$ ,  $V^{(J)} = (v_1, \dots, v_J)$  and  $P^{(J)} = (p_1, \dots, p_J)$ . The first order derivatives are approximated using the centered method and the diffusion term using the usual formula. We obtain, for  $j \in \{1, \dots, J\}$ ,

$$\begin{aligned} & \frac{1}{2\Delta x} (p_{j+1} - p_{j-1}) - \nu \left( \frac{u_{j+1} - 2u_j + u_{j-1}}{\Delta x^2} - k^2 u_j \right) = f_j, \\ & kp_j - \nu \left( \frac{v_{j+1} - 2v_j + v_{j-1}}{\Delta x^2} - k^2 v_j \right) = g_j, \\ & \frac{1}{2\Delta x} (u_{j+1} - u_{j-1}) - kv_j = 0 \end{aligned} \quad [2.48]$$

The boundary conditions become

$$u_0 = u_{J+1} = 0 = v_0 = v_{J+1}.$$

This system has  $3 \times J$  equations and four boundary conditions, but has  $3 \times J + 6$  unknowns: the components of  $U^{(J)}, V^{(J)}, P^{(J)}$  and the quantities  $u_0, v_0, u_{J+1}, v_{J+1}$  as well as  $p_0, p_{J+1}$ , which are also involved in the formulas. The system is under-determined; there are more unknowns than equations. The difficulty arises because the discrete problem requires boundary conditions for the pressure, whereas the continuous problem makes no assumptions on that unknown. This difficulty reveals *parasite modes*, which are non-trivial solutions to the homogeneous discrete problem.

Indeed, when  $f = g = 0$ , the solution of the continuous problem is  $u = v = 0$ ,  $p = 0$ . Now, we can construct a non-zero solution of the discrete problem depending on the pressure boundary assumptions. To study this phenomenon, first note that, by subtracting the evaluations of the first equation at  $j + 1$  and  $j - 1$ ,

$$\begin{aligned} f_{j+1} - f_{j-1} &= \frac{(p_{j+2} - p_j) - (p_j - p_{j-2})}{2\Delta x} \\ &\quad - \nu \left( \frac{(u_{j+2} - 2u_{j+1} + u_j) - (u_j - 2u_{j-1} + u_{j-2})}{\Delta x^2} - k^2(u_{j+1} - u_{j-1}) \right) \\ &= \frac{p_{j+2} - 2p_j + p_{j-2}}{2\Delta x} \\ &\quad - \nu \left( \frac{(u_{j+2} - u_j) - 2(u_{j+1} - u_{j-1}) + (u_j - u_{j-2})}{\Delta x^2} - k^2(u_{j+1} - u_{j-1}) \right) \\ &= \frac{p_{j+2} - 2p_j + p_{j-2}}{2\Delta x} - 2\nu k \Delta x \left( \frac{v_{j+1} - 2v_j + v_{j-1}}{\Delta x^2} - k^2 v_j \right) \\ &= \frac{p_{j+2} - 2p_j + p_{j-2}}{2\Delta x} - 2k^2 \Delta x p_j - 2k \Delta x g_j. \end{aligned}$$

In other words,

$$k^2 p_j - \frac{p_{j+2} - 2p_j + p_{j-2}}{4\Delta x^2} = -kg_j - \frac{f_{j+1} - f_{j-1}}{2\Delta x}. \quad [2.49]$$

This equation is satisfied for  $j \in \{2, \dots, J - 1\}$ . We recognize a discrete version of the elliptic equation [2.47]. However, this discretization can uncouple the grids formed by “even” points ( $j = 2n$ ) and “odd” points ( $j = 2n + 1$ ). We thus construct

non-trivial solutions of the homogeneous problem, where  $f = g = 0$ . Suppose that  $J = 2N$  or  $2N - 1$ . We define  $Q = (q_1, \dots, q_{N-1})$  as a solution of

$$k^2 q_n - \frac{1}{(2\Delta x)^2} (q_{n+1} - 2q_n + q_{n-1}) = 0,$$

$$q_0 = 1, \quad q_N = 0.$$

In the matrix form, this becomes

$$\begin{aligned} & \left( k^2 \begin{pmatrix} 1 & 0 & \cdots & 0 \\ 0 & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & 1 \end{pmatrix} - \frac{1}{(2\Delta x)^2} \begin{pmatrix} -2 & 1 & 0 & \cdots & 0 \\ 1 & -2 & 1 & \cdots & \vdots \\ 0 & 1 & -2 & \cdots & 0 \\ \vdots & \ddots & \ddots & \ddots & 1 \\ 0 & \cdots & 0 & 1 & -2 \end{pmatrix} \right) Q \\ &= \frac{1}{(2\Delta x)^2} \begin{pmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix} \end{aligned}$$

(we can likewise define a vector  $Q$ , with  $q_0 = 0$  and  $q_N = 1$ ). We thus set, for  $n \in \{1, \dots, N\}$ ,  $p_{2n} = q_n$ ,  $u_{2n} = 0$ ,  $v_{2n+1} = 0$ , and we define  $u_{2n+1}, v_{2n}$  as

$$kp_{2n} + \nu \left( k^2 + \frac{2}{\Delta x^2} \right) v_{2n} = 0, \quad \frac{p_{2n+2} - p_{2n}}{2\Delta x} + \nu \left( k^2 + \frac{2}{\Delta x^2} \right) u_{2n+1} = 0$$

(for  $n \in \{1, \dots, N\}$  and  $n \in \{0, \dots, N-1\}$ , respectively). Finally, we write

$$p_{2n+1} = \frac{\nu}{k\Delta x^2} (v_{2n+2} + v_{2n}) = -\frac{k}{k^2\Delta x^2 + 2} (p_{2n+2} + p_{2n}).$$

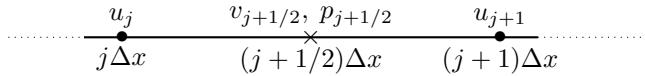
We confirm that, on the one hand,

$$\begin{aligned} \frac{p_{2n+1} - p_{2n-1}}{2\Delta x} &= -\frac{1}{(k^2\Delta x^2 + 2)2\Delta x} (p_{2n+2} - p_{2n-2}) \\ &= 0 - \frac{\nu}{\Delta x^2} (u_{2n+1} + u_{2n-1}), \end{aligned}$$

and, on the other hand,

$$\begin{aligned} \frac{1}{2\Delta x}(u_{2n+1} - u_{2n-1}) - kv_{2n} &= \nu \left( k^2 + \frac{2}{\Delta x^2} \right) \\ \left( k^2 p_{2n} - \frac{1}{(2\Delta x)^2} (p_{2n+2} - 2p_{2n} + p_{2n-2}) \right) &= 0. \end{aligned}$$

We thus obtain a non-trivial solution of [2.48] when  $f = g = 0$ .



**Figure 2.40.** Staggered grids: the horizontal velocity is evaluated at points  $j\Delta x$ , and the vertical velocity and pressure are evaluated at points  $(j + \frac{1}{2})\Delta x$

In order to overcome this difficulty, we use *staggered grids*, following a method introduced in [HAR 65]. The horizontal velocity is approximated at points  $x_j = j\Delta x$ , and the vertical velocity and pressure are approximated at points  $x_{j+1/2} = (j + 1/2)\Delta x$ . The unknowns are therefore  $U^{(J)} = (u_1, \dots, u_J) \in \mathbb{R}^J$ ,  $V^{(J)} = (v_{1/2}, \dots, v_{J+1/2}) \in \mathbb{R}^{J+1}$  and  $P^{(J)} = (p_{1/2}, \dots, p_{J+1/2}) \in \mathbb{R}^{J+1}$ . The discrete equations become

$$\begin{aligned} \frac{1}{\Delta x} (p_{j+1/2} - p_{j-1/2}) - \nu \left( \frac{u_{j+1} - 2u_j + u_{j-1}}{\Delta x^2} - k^2 u_j \right) &= f_j \quad \text{for } j \in \{1, \dots, J\}, \\ kp_{j+1/2} - \nu \left( \frac{v_{j+3/2} - 2v_{j+1/2} + v_{j-1/2}}{\Delta x^2} - k^2 v_{j+1/2} \right) &= g_{j+1/2} \quad \text{for } j \in \{0, \dots, J\}, \\ \frac{1}{\Delta x} (u_{j+1} - u_j) - kv_{j+1/2} &= 0 \quad \text{for } j \in \{0, \dots, J\}. \end{aligned}$$

This system is composed of  $J + (J + 1) + (J + 1) = 3J + 2$  equations. We complete them with the boundary conditions for the velocity

$$u_0 = u_{J+1} = 0 = v_{-1/2} = v_{J+3/2}.$$

These equations involve the  $3J + 2$  components of vectors  $U^{(J)}, V^{(J)}, P^{(J)}$  and the four terms imposed by the boundary conditions: there are as many equations as unknowns. In order to describe this system in the matrix form, we introduce the matrix with  $(J + 1)$  columns and  $J$  rows:

$$B = \frac{1}{\Delta x} \begin{pmatrix} -1 & 1 & 0 & \cdots & 0 & 0 \\ 0 & -1 & 1 & \cdots & 0 & 0 \\ \vdots & & \ddots & \ddots & & \vdots \\ 0 & 0 & 0 & \cdots & 0 & -1 \\ 0 & 0 & 0 & \cdots & 0 & 1 \end{pmatrix}$$

Its transpose is thus the matrix with  $J$  columns and  $(J + 1)$  rows:

$$B^\top = \frac{1}{\Delta x} \begin{pmatrix} -1 & 0 & \cdots & 0 \\ 1 & -1 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & & \ddots & \vdots \\ 0 & 0 & \cdots & 0 & 1 \\ 0 & 0 & \cdots & 0 & -1 \end{pmatrix}$$

In particular, we have

- for  $j \in \{1, \dots, J\}$ , the  $j$ th component of  $BP^{(J)}$  is  $p_{j+1/2} - p_{j-1/2}$ ;
- for  $j \in \{0, \dots, J\}$ , the  $(j + 1)$ th component of  $B^\top U^{(J)}$  is  $u_j - u_{j+1}$ .

(with the condition  $u_0 = 0 = u_{J+1}$ ). We will also use the matrix with  $J$  columns and  $J$  rows:

$$A = \frac{\nu}{\Delta x^2} \begin{pmatrix} 2 & -1 & 0 & \cdots & 0 \\ -1 & 2 & -1 & \ddots & \vdots \\ 0 & -1 & 2 & \ddots & 0 \\ \vdots & \ddots & 0 & -1 & 2 \\ 0 & 0 & -1 & 2 & 0 \end{pmatrix}$$

and we denote as  $\bar{A}$  the matrix with the same structure but with  $(J + 1)$  columns and  $(J + 1)$  rows. Similarly, we denote as  $\mathbb{I}$  the identity matrix on  $\mathbb{R}^J$  and as  $\bar{\mathbb{I}}$  the identity matrix on  $\mathbb{R}^{J+1}$ . We therefore obtain the following block expression:

$$\begin{pmatrix} \nu k^2 \mathbb{I} + A & 0 & B \\ 0 & \nu k^2 \bar{\mathbb{I}} + \bar{A} & k \bar{\mathbb{I}} \\ B^\top & k \bar{\mathbb{I}} & 0 \end{pmatrix} \begin{pmatrix} U^{(J)} \\ V^{(J)} \\ P^{(J)} \end{pmatrix} = \begin{pmatrix} F^{(J)} \\ G^{(J)} \\ 0 \end{pmatrix}$$

We can distinguish the velocity and pressure unknowns by writing

$$\mathcal{U}^{(J)} = \begin{pmatrix} U^{(J)} \\ V^{(J)} \end{pmatrix} \in \mathbb{R}^{2J+1}, \quad \mathcal{A}_k = \begin{pmatrix} \nu k^2 \mathbb{I} + A & 0 \\ 0 & \nu k^2 \bar{\mathbb{I}} + \bar{A} \end{pmatrix}, \quad \mathcal{B}_k = \begin{pmatrix} B \\ k \bar{\mathbb{I}} \end{pmatrix}.$$

Similarly, we concatenate the data  $F^{(J)}$  and  $G^{(J)}$  in a single vector  $\mathcal{F}^{(J)}$ . The problem becomes

$$\begin{pmatrix} \mathcal{A}_k & \mathcal{B}_k \\ \mathcal{B}_k^\top & 0 \end{pmatrix} \begin{pmatrix} \mathcal{U}^{(J)} \\ P^{(J)} \end{pmatrix} = \begin{pmatrix} \mathcal{F}^{(J)} \\ 0 \end{pmatrix}. \quad [2.50]$$

**THEOREM 2.17.**— The linear system [2.50] has a unique solution.

**PROOF.**— Two ingredients are crucial in analyzing the system [2.50]. On the one hand, studying the elliptic model problem indicates that the matrix  $\mathcal{A}_k$  is symmetric positive definite, and therefore invertible. On the other hand, note that the matrix  $\mathcal{B}_k$  is injective. The equation provided by the uppermost blocks of [2.50] can be rewritten in the form

$$\mathcal{U}^{(J)} = \mathcal{A}_k^{-1}(\mathcal{F}^{(J)} - \mathcal{B}_k P^{(J)}).$$

Since  $\mathcal{B}_k^\top \mathcal{U}^{(J)} = 0$ , it follows that  $P^{(J)}$  satisfies

$$\mathcal{B}_k^\top \mathcal{A}_k^{-1} \mathcal{B}_k P^{(J)} = \mathcal{B}_k^\top \mathcal{A}_k^{-1} \mathcal{F}^{(J)}.$$

Because  $\mathcal{B}_k$  is injective, the matrix  $\mathcal{B}_k^\top \mathcal{A}_k^{-1} \mathcal{B}_k$  is symmetric positive definite, so it is invertible, since for all  $Q \in \mathbb{R}^{J+1} \setminus \{0\}$ , we have

$$\mathcal{B}_k^\top \mathcal{A}_k^{-1} \mathcal{B}_k Q \cdot Q = \mathcal{A}_k^{-1} \mathcal{B}_k Q \cdot \mathcal{B}_k Q \geq \frac{1}{\lambda_M} |\mathcal{B}_k Q|^2 > 0$$

where  $\lambda_M > 0$  denotes the largest eigenvalue of  $\mathcal{A}_k$ . □

The pressure can be interpreted as the Lagrange multiplier associated with the incompressibility. Indeed, the constrained optimization theorem allows us to characterize the minimum of the functional

$$\mathcal{J} : \mathcal{V} \in \mathbb{R}^{2J+1} \longmapsto \frac{1}{2} \mathcal{A}_k \mathcal{V} \cdot \mathcal{V} - \mathcal{F}^{(J)} \cdot \mathcal{V}$$

on the domain

$$\mathcal{C} = \{\mathcal{V} \in \mathbb{R}^{2J+1}, \mathcal{B}_k^\top \mathcal{V} = 0\}.$$

Theorem 1.5 ensures that, with  $\mathcal{U}^{(J)}$  as the minimizer of  $\mathcal{J}$  on  $\mathcal{C}$ , there exists a vector  $(p_{1/2}, \dots, p_{J+1/2}) = P^{(J)} \in \mathbb{R}^{J+1}$ , such that

$$\nabla \mathcal{J}(\mathcal{U}^{(J)}) + \mathcal{B}_k P^{(J)} = 0.$$

We thus find [2.50]. This observation can be reformulated in terms of a *saddle point*.

**PROPOSITION 2.9.–** The pair  $(\mathcal{U}, P)$  is a solution of [2.50] if and only if  $(\mathcal{U}, P)$  is a saddle point of the functional

$$\mathcal{L}(\mathcal{V}, q) = \mathcal{J}(\mathcal{V}) + (q, \mathcal{B}_k^\top \mathcal{V})$$

which means that for all  $(\mathcal{V}, q)$ , we have  $\mathcal{L}(\mathcal{U}, q) \leq \mathcal{L}(\mathcal{U}, P) \leq \mathcal{L}(\mathcal{V}, P)$ .

**PROOF.–** Let  $(\mathcal{U}, P)$  be a solution of [2.50]. Then, we have

$$\mathcal{L}(\mathcal{U}, P) = \mathcal{J}(\mathcal{U})$$

because  $\mathcal{B}_k^\top \mathcal{U} = 0$ . Let  $\mathcal{V} = \mathcal{U} + h$ , so that

$$\begin{aligned} \mathcal{L}(\mathcal{V}, P) &= \mathcal{J}(\mathcal{U}) + (\mathcal{A}_k \mathcal{U} - \mathcal{F}^{(J)}) \cdot h + \frac{1}{2} \mathcal{A}_k h \cdot h + P \cdot \mathcal{B}_k^\top h \\ &\geq \mathcal{J}(\mathcal{U}) + (\mathcal{A}_k \mathcal{U} - \mathcal{F}^{(J)} + \mathcal{B}_k P) \cdot h = \mathcal{J}(\mathcal{U}) = \mathcal{L}(\mathcal{U}, P). \end{aligned}$$

Moreover, for all  $q \in \mathbb{R}^{J+1}$ , we have

$$\mathcal{L}(\mathcal{U}, P) - \mathcal{L}(\mathcal{U}, q) = (P - q, \mathcal{B}_k^\top \mathcal{U}) = 0.$$

This proves that  $(\mathcal{U}, P)$  is a saddle point of  $\mathcal{L}$ .

Reciprocally, let  $(\mathcal{U}, P)$  be a saddle point of  $\mathcal{L}$ . Then, for all  $\mathcal{V} = \mathcal{U} + th$ ,  $t > 0$ ,  $h \in \mathbb{R}^{2J+1}$ , we obtain

$$\mathcal{L}(\mathcal{V}, P) - \mathcal{L}(\mathcal{U}, P) = t(\mathcal{A}_k \mathcal{U} + \mathcal{B}_k P - \mathcal{F}^{(J)}) \cdot h + \frac{t^2}{2} \mathcal{A}_k h \cdot h \geq 0.$$

We divide this relation by  $t$  and let  $t$  tend to 0 to obtain

$$(\mathcal{A}_k \mathcal{U} + \mathcal{B}_k P - \mathcal{F}^{(J)}) \cdot h \geq 0.$$

This relation is satisfied for all  $h = \pm \eta \in \mathbb{R}^{2J+1}$ , so we have  $(\mathcal{A}_k \mathcal{U} + \mathcal{B}_k P - \mathcal{F}^{(J)}) \cdot \eta = 0$  for all vectors  $\eta$ . This means that  $(\mathcal{A}_k \mathcal{U} + \mathcal{B}_k P - \mathcal{F}^{(J)}) = 0$ . Finally, we also have  $\mathcal{L}(\mathcal{U}, P) - \mathcal{L}(\mathcal{U}, q) = (P - q, \mathcal{B}_k^\top \mathcal{U}) \geq 0$  for all vectors  $q = P \pm \zeta \in \mathbb{R}^{J+1}$ , which implies that  $\mathcal{B}_k^\top \mathcal{U} = 0$ . We have thus shown that  $(\mathcal{U}, P)$  is a solution of [2.50].  $\square$

In other words, we have  $\mathcal{L}(\mathcal{U}, p) = \max_q \min_{\mathcal{V}} \mathcal{L}(\mathcal{V}, q) = \min_{\mathcal{V}} \max_q \mathcal{L}(\mathcal{V}, q)$ . The objective is thus to minimize  $\mathcal{L}$  with respect to the state variable and maximize it with respect to the multiplier variable. One approach is to define approximations of  $(\mathcal{U}, P)$  by adapting the gradient method. We thus obtain

the *Arrow–Hurwicz algorithm*: given  $\mathcal{U}_0$  and  $p_0$ , we construct the sequences defined by the relations

$$\mathcal{U}_{n+1} = \mathcal{U}_n - \varrho \left[ \mathcal{A}_k \mathcal{U}_n - \mathcal{F}^{(J)} + \mathcal{B}_k p_n \right], \quad p_{n+1} = p_n + \alpha \varrho \mathcal{B}_k^\top \mathcal{U}_{n+1}.$$

Given  $\mathcal{U}_n, p_n$ , we find  $\mathcal{U}_{n+1}$  by “descending” in the direction  $\nabla_U \mathcal{L}(\mathcal{U}_n, p_n)$  with the step  $\varrho$ , and then we find  $p_{n+1}$  by “climbing up” following the direction  $\nabla_p \mathcal{L}(\mathcal{U}_{n+1}, p_n)$   $\alpha \varrho$  with the step. This algorithm depends on two parameters,  $\alpha, \varrho > 0$ , which must satisfy certain conditions to guarantee convergence.

**THEOREM 2.18.**— Let  $\lambda_m > 0$  be the smallest eigenvalue of  $\mathcal{A}_k$ . If

$$0 < \varrho < \frac{2\lambda_m}{\alpha \|\mathcal{B}_k^\top\|^2 + \|\mathcal{A}_k\|^2}, \quad [2.51]$$

then the sequence  $(\mathcal{U}_n, p_n)_{n \in \mathbb{N}}$  converges to  $(\mathcal{U}, P)$ , a solution of [2.50].

**PROOF.**— We denote  $e_n = \mathcal{U}_n - \mathcal{U}$  and  $r_n = p_n - P$ , which satisfy  $e_{n+1} = (\mathbb{I} - \varrho \mathcal{A}_k)e_n - \varrho \mathcal{B}_k r_n$  and  $r_{n+1} = r_n + \varrho \mathcal{B}_k^\top e_{n+1}$ . It follows that

$$\|e_{n+1}\|^2 = (\mathbb{I} - \varrho \mathcal{A}_k)e_n \cdot e_{n+1} - \varrho r_n \cdot \mathcal{B}_k^\top e_{n+1}.$$

We use this relation to calculate

$$\begin{aligned} \|r_{n+1}\|^2 &= \|r_n\|^2 + \alpha^2 \varrho^2 \|\mathcal{B}_k^\top e_{n+1}\|^2 + 2\alpha \varrho r_n \cdot \mathcal{B}_k^\top e_{n+1} \\ &= \|r_n\|^2 + \alpha^2 \varrho^2 \|\mathcal{B}_k^\top e_{n+1}\|^2 - 2\alpha \|e_{n+1}\|^2 + 2\alpha e_n \cdot (\mathbb{I} - \varrho \mathcal{A}_k)e_{n+1}. \end{aligned}$$

The Cauchy–Schwarz inequality, combined with  $2ab \leq a^2 + b^2$ , makes it possible to dominate

$$\begin{aligned} 2e_n \cdot (\mathbb{I} - \varrho \mathcal{A}_k)e_{n+1} &\leq 2\|e_n\| \|(\mathbb{I} - \varrho \mathcal{A}_k)e_{n+1}\| \leq \|e_n\|^2 + \|(\mathbb{I} - \varrho \mathcal{A}_k)e_{n+1}\|^2 \\ &\leq \|e_n\|^2 + \|e_{n+1}\|^2 - 2\varrho \mathcal{A}_k e_{n+1} \cdot e_{n+1} + \varrho^2 \|\mathcal{A}_k e_{n+1}\|^2 \\ &\leq \|e_n\|^2 + \|e_{n+1}\|^2 - 2\varrho \lambda_m \|e_{n+1}\|^2 + \varrho^2 \|\mathcal{A}_k\|^2 \|e_{n+1}\|^2 \end{aligned}$$

where we have used the fact that  $\mathcal{A}_k$  is symmetric positive definite and  $\mathcal{A}_k \xi \cdot \xi \geq \lambda_m \|\xi\|^2$  is satisfied for all  $\xi \in \mathbb{R}^{2J+1}$ . Therefore,

$$\|r_{n+1}\|^2 + \alpha \|e_{n+1}\|^2 \leq \|r_n\|^2 + \alpha \|e_n\|^2 + \alpha \varrho (\varrho (\alpha \|\mathcal{B}_k^\top\|^2 + \|\mathcal{A}_k\|^2) - 2\lambda_m) \|e_{n+1}\|^2. \quad [2.52]$$

When the condition [2.51] is satisfied, the last term contributes negatively in such a way that the sequence  $(\|r_n\|^2 + \alpha \|e_n\|^2)_{n \in \mathbb{N}}$  is decreasing. Evidently, this sequence is

bounded from below by 0. It follows that it converges. Passing to the limit in [2.52], we see that  $\|e_n\|$  tends to 0 when  $n \rightarrow \infty$ . Then, the relation  $e_{n+1} = (I - \varrho \mathbb{A})e_n - \varrho \mathcal{B}_k r_n$  implies that  $\mathcal{B}_k r_n$  also converges to 0, and, finally, since  $\mathcal{B}_k$  is injective,  $r_n$  converges to 0.  $\square$

The most attractive part of this algorithm is the fact that it does not require solving any equations: successive iterations are obtained using simple matrix products, without solving large linear systems. However, the condition [2.51] is constraining. In particular, it implies that  $0 < \varrho < \frac{2\lambda_m}{\lambda_M^2}$ , where  $\lambda_M$  denotes the largest eigenvalue of the symmetric matrix  $\mathcal{A}_k$ . This upper bound tends to 0 when the number of discretization points increases (it behaves like  $J^4$ ; see Figure 2.46): for a fine-grained discretization, we must take a small value of  $\varrho$  and thus perform a large number of iterations.

A rather similar approach is to define the sequences

$$\mathcal{A}_k \mathcal{U}_{n+1} = \mathcal{F}^{(J)} - \mathcal{B}_k p_n, \quad p_{n+1} = p_n + \rho \mathcal{B}_k^\top \mathcal{U}_{n+1}. \quad [2.53]$$

Here, given a known  $\mathcal{U}_n, p_n$ , we determine  $\mathcal{U}_{n+1}$  as the minimizer of  $\mathcal{U} \mapsto \mathcal{L}(\mathcal{U}, p_n)$ , and then we update  $p_{n+1}$  using a step from the gradient algorithm (following the direction of  $\nabla_p \mathcal{L}(\mathcal{U}_{n+1}, p_n)$  with step  $\varrho$ ). There is now only one parameter  $\varrho > 0$ . However, the method requires solving a linear system for each iteration. On the one hand, the size of these systems is somewhat reduced with respect to that of [2.50], and on the other hand, the structure of the matrix  $\mathcal{A}_k$  can be more favorable than that of the complete system [2.50], enabling the use of more efficient resolution methods. These remarks motivate the introduction of the *Uzawa iteration algorithm* [2.53].

**THEOREM 2.19.**— Let  $\lambda_m > 0$  be the smallest eigenvalue of  $\mathcal{A}_k$ . If

$$0 < \rho < \frac{2\lambda_m}{\|\mathcal{B}_k^\top\|^2}, \quad [2.54]$$

then the sequence  $(\mathcal{U}_n, p_n)_{n \in \mathbb{N}}$  defined by [2.53] converges to  $(\mathcal{U}, P)$ , a solution of [2.50].

**PROOF.**— Writing  $e_n = \mathcal{U}_n - \mathcal{U}$  and  $r_n = P_n - P$ , we have

$$\mathcal{A}_k e_{n+1} + \mathcal{B}_k r_n = 0, \quad r_{n+1} = r_n + \rho \mathcal{B}_k^\top e_{n+1}.$$

It follows that

$$\begin{aligned}\|r_{n+1}\|^2 &= \|r_n\|^2 + \rho^2 \|\mathcal{B}_k^\top e_{n+1}\|^2 + 2\rho \mathcal{B}_k^\top e_{n+1} \cdot r_n \\ &= \|r_n\|^2 + \rho^2 \|\mathcal{B}_k^\top e_{n+1}\|^2 + 2\rho e_{n+1} \cdot \mathcal{B}_k r_n \\ &= \|r_n\|^2 + \rho^2 \|\mathcal{B}_k^\top e_{n+1}\|^2 - 2\rho \mathcal{A}_k e_{n+1} \cdot e_{n+1}.\end{aligned}$$

We thus arrive at the following estimation:

$$\|r_{n+1}\|^2 \leq \|r_n\|^2 - \rho(2\lambda_m - \rho \|\mathcal{B}_k^\top\|^2) \|e_{n+1}\|^2. \quad [2.55]$$

The condition [2.54] ensures that the sequence  $\|r_n\|^2$  is decreasing and therefore has a positive or zero limit. Letting  $n \rightarrow \infty$  in [2.55], we obtain  $\lim_{n \rightarrow \infty} \|e_n\| = 0$ . Since the sequence  $(p_n)_{n \in \mathbb{N}}$  is bounded, it has at least one convergent subsequence with limit  $\bar{p}$ , and returning to [2.53], we have  $\mathcal{A}_k \mathcal{U} + \mathcal{B}_k \bar{p} = \mathcal{F}$ . We conclude by recalling the uniqueness of the solution of [2.50]:  $\bar{p} = P$  and the entire sequence converges to that limit.  $\square$

Another approach, making it possible to avoid having to introduce the vector  $P$ , deals with the constraint by penalization. Let  $0 < \varepsilon \ll 1$  be a given parameter, designed in practice to be “small”. We determine  $\mathcal{U}_\varepsilon$ , a minimizer on  $\mathbb{R}^{2J+1}$  of the functional

$$\mathcal{V} \mapsto \mathcal{J}(\mathcal{V}) + \frac{1}{2\varepsilon} \|\mathcal{B}_k^\top \mathcal{V}\|^2.$$

Note that this functional  $\mathcal{J}$  is strictly convex and  $\mathcal{U}_\varepsilon$  is a critical point of  $\mathcal{J}$ : it makes  $\nabla \mathcal{J}$  zero and is therefore a solution to the linear system

$$\mathcal{A}_k \mathcal{U}_\varepsilon + \frac{1}{\varepsilon} \mathcal{B}_k \mathcal{B}_k^\top \mathcal{U}_\varepsilon = \mathcal{F}^{(J)}. \quad [2.56]$$

The value of this method is provided by the following statement.

**THEOREM 2.20.–** When  $\varepsilon \rightarrow 0$ ,  $\mathcal{U}_\varepsilon$  converges to  $\mathcal{U}$  in  $\mathbb{R}^{2J+1}$  and there exists a  $P \in \mathbb{R}^{J+1}$ , such that  $(\mathcal{U}, P)$  is a solution of [2.50].

**PROOF.–** We calculate the scalar product of [2.56] with  $\mathcal{U}_\varepsilon$ . Using the Cauchy–Schwarz inequality, we obtain

$$\mathcal{A}_k \mathcal{U}_\varepsilon \cdot \mathcal{U}_\varepsilon + \frac{1}{\varepsilon} \mathcal{B}_k \mathcal{B}_k^\top \mathcal{U}_\varepsilon \cdot \mathcal{U}_\varepsilon = \mathcal{F}^{(J)} \cdot \mathcal{U}_\varepsilon \leq \|\mathcal{F}^{(J)}\| \|\mathcal{U}_\varepsilon\|.$$

As mentioned above,  $\mathcal{A}_k$  is symmetric positive definite, whereas  $\mathcal{B}_k \mathcal{B}_k^\top$  is symmetric and positive: for all  $\xi \in \mathbb{R}^{2J+1} \setminus \{0\}$ , we have  $\mathcal{A}_k \xi \cdot \xi \geq \lambda_m |\xi|^2$  and  $\mathcal{B}_k \mathcal{B}_k^\top \xi \cdot \xi \geq 0$ . It follows that the sequence  $(\mathcal{U}_\varepsilon)_{\varepsilon > 0}$  is bounded:  $\|\mathcal{U}_\varepsilon\| \leq \frac{\|\mathcal{F}^{(J)}\|}{\lambda_m}$ . We can extract a convergent subsequence:  $\lim_{\ell \rightarrow \infty} \mathcal{U}_{\varepsilon_\ell} = \mathcal{U}$ . Passing to the limit in [2.56], we obtain

$$\varepsilon_\ell (\mathcal{A}_k \mathcal{U}_{\varepsilon_\ell} - \mathcal{F}^{(J)}) + \mathcal{B}_k \mathcal{B}_k^\top \mathcal{U}_{\varepsilon_\ell} = 0 \xrightarrow[\ell \rightarrow \infty]{} \mathcal{B}_k \mathcal{B}_k^\top \mathcal{U} = 0$$

so  $\mathcal{U} \in \text{Ker}(\mathcal{B}_k \mathcal{B}_k^\top)$ . Now,  $\mathcal{B}_k \mathcal{B}_k^\top \mathcal{U} \cdot \mathcal{U} = \mathcal{B}_k^\top \mathcal{U} \cdot \mathcal{B}_k^\top \mathcal{U} = \|\mathcal{B}_k^\top \mathcal{U}\|^2$ , so  $\mathcal{U} \in \text{Ker}(\mathcal{B}_k^\top)$ : the limit satisfies the desired incompressibility constraint. Let  $\mathcal{V} \in \mathbb{R}^{2J+1}$ , such that  $\mathcal{B}_k^\top \mathcal{V} = 0$ . We have

$$\begin{aligned} 0 &= (\mathcal{A}_k \mathcal{U}_{\varepsilon_\ell} + \frac{1}{\varepsilon_\ell} \mathcal{B}_k \mathcal{B}_k^\top \mathcal{U}^{\varepsilon_\ell} - \mathcal{F}^{(J)}) \cdot \mathcal{V} = 0 + (\mathcal{A}_k \mathcal{U}_{\varepsilon_\ell} - \mathcal{F}^{(J)}) \cdot \mathcal{V} \\ &\xrightarrow[\ell \rightarrow \infty]{} (\mathcal{A}_k \mathcal{U} - \mathcal{F}^{(J)}) \cdot \mathcal{V} = 0. \end{aligned}$$

The limit  $\mathcal{U}$  is therefore such that  $(\mathcal{A}_k \mathcal{U} - \mathcal{F}^{(J)}) \in \text{Ker}(\mathcal{B}_k^\top)^\perp = \text{Ran}(\mathcal{B}_k)$ . Since we have shown that the system  $\mathcal{A}_k \mathcal{U} + \mathcal{B}_k P = \mathcal{F}^{(J)}$ ,  $\mathcal{B}_k^\top \mathcal{U} = 0$  has a unique solution, it follows that the entire sequence  $(\mathcal{U}_\varepsilon)_{\varepsilon > 0}$  converges.  $\square$

In practice, this kind of method, which can nevertheless be efficient, leads to two types of difficulties, both related to the fact that they depend on a parameter  $\varepsilon > 0$ . On the one hand, the role of the penalization parameter is not easy to analyze with respect to the discretization parameters. On the other hand, we can see that the matrix

$$\mathcal{A}_\varepsilon = \mathcal{A}_k + \frac{1}{\varepsilon} \mathcal{B}_k \mathcal{B}_k^\top$$

of the underlying system is ill conditioned. Indeed, the matrix  $\mathcal{A}_\varepsilon$  is symmetric positive definite and its conditioning is given by the ratio of the extreme eigenvalues. The largest eigenvalue of  $\mathcal{A}_\varepsilon$  is given by

$$\Lambda_\varepsilon = \max_{\mathcal{V} \neq 0} \frac{\mathcal{A}_\varepsilon \mathcal{V} \cdot \mathcal{V}}{\|\mathcal{V}\|^2} \geq \frac{1}{\varepsilon} \max_{\mathcal{V} \neq 0} \frac{\mathcal{B}_k \mathcal{B}_k^\top \mathcal{V} \cdot \mathcal{V}}{\|\mathcal{V}\|^2} = \frac{\mu}{\varepsilon}$$

where  $\mu$  is the largest eigenvalue of the symmetric matrix  $\mathcal{B}_k \mathcal{B}_k^\top$ . However, we note also that for all  $\mathcal{V} \neq 0$ , such that  $\mathcal{B}_k^\top \mathcal{V} = 0$ , we have

$$\frac{\mathcal{A}_\varepsilon \mathcal{V} \cdot \mathcal{V}}{\|\mathcal{V}\|^2} = \frac{\mathcal{A}_k \mathcal{V} \cdot \mathcal{V}}{\|\mathcal{V}\|^2} \leq \max_{\mathcal{V} \neq 0, \mathcal{B}_k^\top \mathcal{V} = 0} \frac{\mathcal{A}_k \mathcal{V} \cdot \mathcal{V}}{\|\mathcal{V}\|^2} = \sigma.$$

It follows that the smallest eigenvalue of  $\mathcal{A}_\varepsilon$  satisfies

$$\lambda_\varepsilon = \min_{\mathcal{V} \neq 0} \frac{\mathcal{A}_\varepsilon \mathcal{V} \cdot \mathcal{V}}{\|\mathcal{V}\|^2} \leq \min_{\mathcal{V} \neq 0, \mathcal{B}_k^\top \mathcal{V} = 0} \frac{\mathcal{A}_\varepsilon \mathcal{V} \cdot \mathcal{V}}{\|\mathcal{V}\|^2} \leq \sigma.$$

It follows that

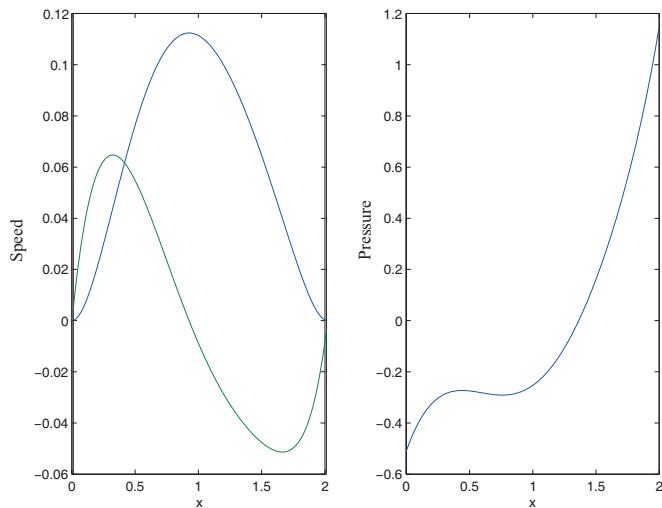
$$\text{cond}(A_\varepsilon) = \frac{\Lambda_\varepsilon}{\lambda_\varepsilon} \geq \frac{\mu}{\varepsilon\sigma}.$$

In order to appropriately approximate the constraint, we must take small values of  $\varepsilon$ ; however, the conditioning of the linear system that determines the approximate solution  $\mathcal{U}_\varepsilon$  is degraded. This difficulty is illustrated in Figure 2.48.

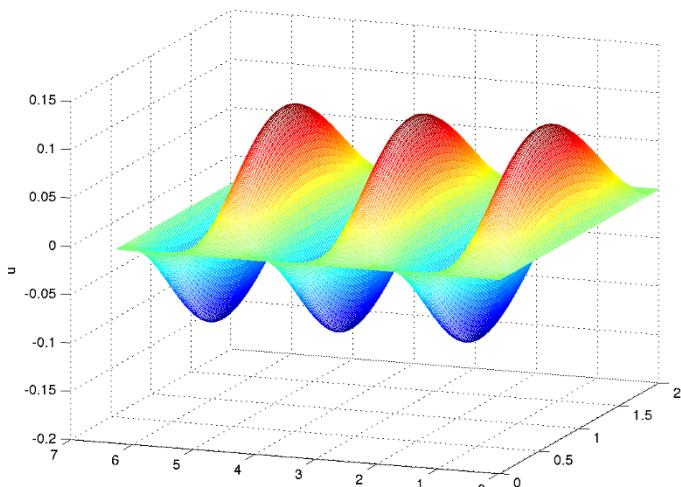
We test these methods numerically, in the case where  $0 < x < 2$  and  $k = 3$ . The applied force is given by

$$f(x) = 2 \sin(x), \quad g(x) = \cos(3x + .08).$$

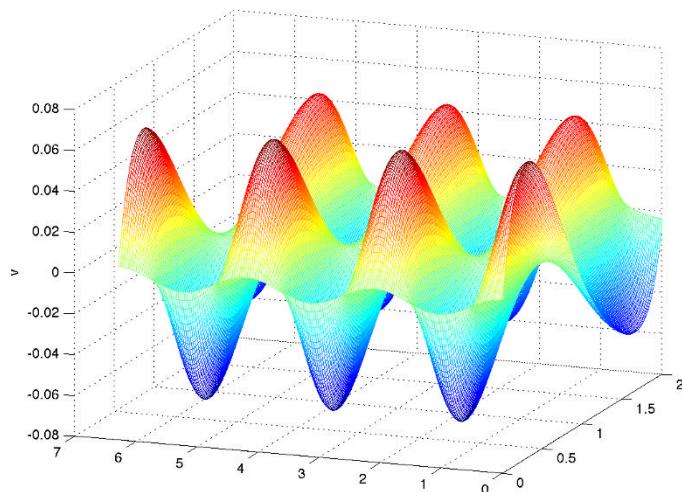
Figure 2.41 is obtained by directly solving the linear system. Given the conditions of this simulation (200 discretization points), conditioning is of the order of  $8.10^4$ . Figures 2.42–2.44 show the velocity  $(U(x, y), V(x, y))$  and pressure  $P(x, y)$  fields. Figure 2.45 shows the velocity field; we clearly see the formation of “vortices” that turn in opposite directions (for this figure,  $k = 3$ ; the number of structures clearly depends on the parameter  $k$ ). We obtain similar curves using the Arrow–Hurwicz algorithm, with, for example, the parameters  $\varrho = 3 \cdot 10^{-5}$  and  $\alpha = 1.1/\varrho$ . Note that  $\varrho$  must be chosen small enough to ensure convergence; if we seek to perform simulations with larger parameters, the calculation is unstable and can lead to outlying values. This phenomenon is explained by the fact that condition [2.51] becomes more constraining when the grid is finer, as is shown by the evolution of the norm of  $\mathcal{A}_k$  in Figure 2.46. With these parameters, we obtain a solution such that the relative error  $\frac{\|\mathcal{A}_k \mathcal{U}_{n+1} - \mathcal{F} + \mathcal{B}_k P_n\|}{\|\mathcal{F}\|}$  and the norm of the discrete divergence  $\|\mathcal{B}_k^\top \mathcal{U}_{n+1}\|$  are less than  $10^{-5}$ , within approximatively 33000 iterations. With the same stopping criteria, the Uzawa algorithm provides the result in 18 iterations for  $\varrho = 1.1$ . Therefore, even if it is necessary to solve linear systems at each iteration, this algorithm is faster. Figure 2.47 shows the velocity fields obtained using the penalization method with  $\varepsilon = (10^{-2})^k$ , and  $k$  varying from 1 to 6. Figure 2.48 shows the degradation of the conditioning when  $\varepsilon \rightarrow 0$ , while the incompressibility condition is better satisfied. This discussion clearly shows that the choice of the numerical method and the evaluation of its practical efficiency rely on subtle arguments.



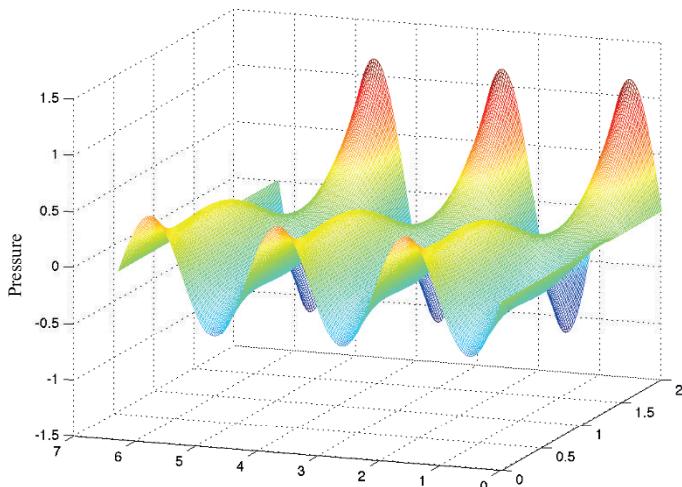
**Figure 2.41.** Simulation of Stokes problem (200 discretization points). For the color version of this figure, see [www.iste.co.uk/goudon/mathmodel.zip](http://www.iste.co.uk/goudon/mathmodel.zip)



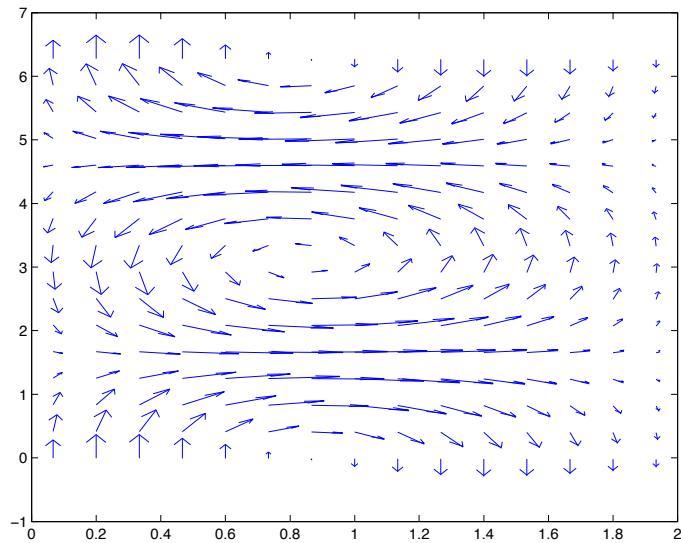
**Figure 2.42.** Simulation of Stokes problem (200 discretization points): horizontal component of the velocity field. For the color version of this figure, see [www.iste.co.uk/goudon/mathmodel.zip](http://www.iste.co.uk/goudon/mathmodel.zip)



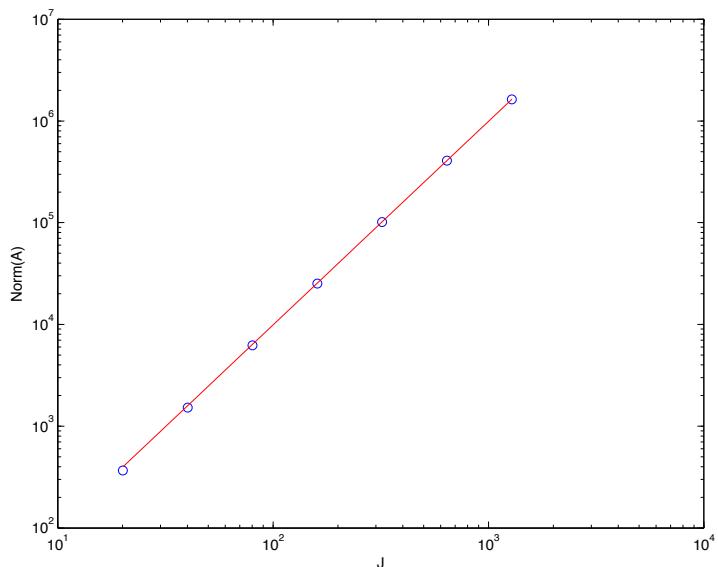
**Figure 2.43.** Simulation of Stokes problem (200 discretization points): vertical component of the velocity field. For the color version of this figure, see [www.iste.co.uk/goudon/mathmodel.zip](http://www.iste.co.uk/goudon/mathmodel.zip)



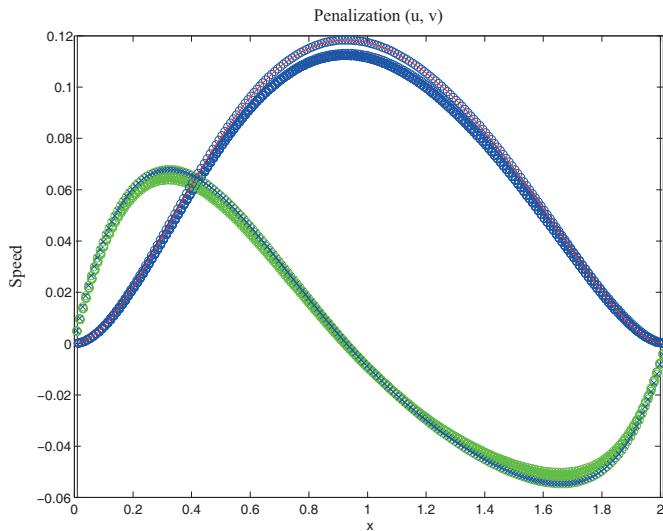
**Figure 2.44.** Simulation of Stokes problem (200 discretization points): pressure field. For the color version of this figure, see [www.iste.co.uk/goudon/mathmodel.zip](http://www.iste.co.uk/goudon/mathmodel.zip)



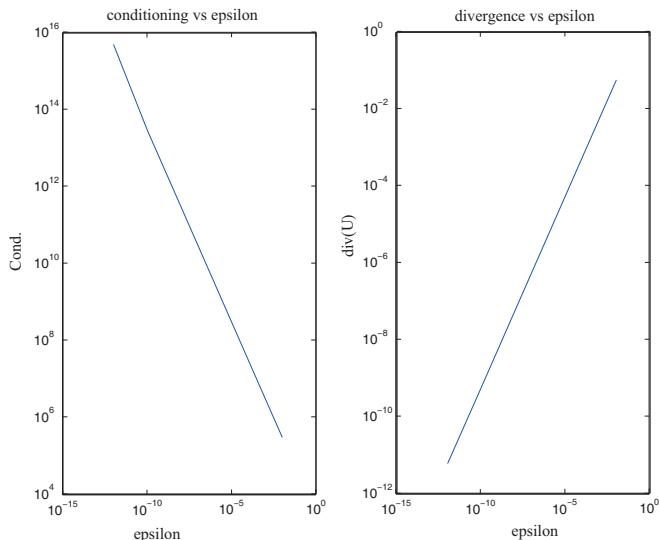
**Figure 2.45.** Simulation of Stokes problem: velocity field ( $k = 1$ )



**Figure 2.46.** Evolution of the norm of  $\mathcal{A}_k$  as a function of the number of discretization points  $J$ . For the color version of this figure, see [www.iste.co.uk/goudon/mathmodel.zip](http://www.iste.co.uk/goudon/mathmodel.zip)



**Figure 2.47.** Velocity fields using the penalization method for different values of  $\varepsilon$ . For the color version of this figure, see [www.iste.co.uk/goudon/mathmodel.zip](http://www.iste.co.uk/goudon/mathmodel.zip)



**Figure 2.48.** Conditioning of the penalized linear system and divergence of the velocity field as a function of  $\varepsilon$

---

# Numerical Simulations of Partial Differential Equations: Time-dependent Problems

---

*There are no impenetrable citadels, only badly attacked fortresses.*

Pierre Choderlos de Laclos  
(*Les liaisons dangereuses*)

## 3.1. Diffusion equations

We are interested in the numerical approximation of the heat equation in one spatial dimension,

$$\partial_t u = \kappa \partial_{xx}^2 u + f. \quad [3.1]$$

When the problem spans all space ( $x \in \mathbb{R}$ ), it can be simply resolved by the Fourier transform (see [GOU 11, section 6.4]); with  $u_{\text{Init}}$  being the initial value for [3.1], we obtain

$$u(t, x) = H(t)u_{\text{Init}}(x) + \int_0^t H(t-s)f(s, x) \, ds$$

where  $t \mapsto H(t)$  represents the family of operators defined by

$$H(t)v(x) = \int_{\mathbb{R}} \frac{1}{\sqrt{4\pi\kappa t}} \exp\left(-\frac{|x-y|^2}{4\kappa t}\right) v(y) \, dy.$$

In particular, we easily establish:

**PROPOSITION 3.1.–** If  $u_{\text{Init}} \geq 0$  and  $f \geq 0$  then  $u \geq 0$ .

If  $f = 0$ ,  $t \mapsto u(t, \cdot)$  is continuous over  $[0, \infty)$  with values within  $L^2(\mathbb{R})$  and we even obtain  $u \in C^\infty([\epsilon, T] \times \mathbb{R})$  for all  $0 < \epsilon < T < \infty$ .

**NOTE 3.1.–** The analysis of diffusion–reaction problems, of the type

$$\partial_t u = \partial_{xx}^2 u + F(u)$$

can prove to be surprising. An iterative scheme can show, with the help of Banach’s theorem, the existence and uniqueness of a solution defined over a sufficiently short interval of time. However, we could convince ourselves that in such problems, there is a competition between the regularizing effect of the heat equation  $\partial_t u = \partial_{xx}^2 u$  and the “explosive” effect of the ODE  $u'(t) = F(u(t))$ . More information on this fascinating topic can be found in [GIG 85, WEI 85] and, more recently, [MER 98].

The numerical analysis of this equation combines both the difficulties of the stationary problems (the choice of infinite-dimensional norms) and the stability issues related to temporal evolution. The simplest scheme that we can imagine depends upon an explicit Euler-type discretization for the time derivative and a finite difference discretization, with a uniform mesh of step size  $\Delta x > 0$ , for the diffusion operator  $\partial_{xx}^2$ . We thus define a sequence  $u_j^n$  by

$$\begin{aligned} u_j^0 &= u_{\text{Init}}(j\Delta x), \\ u_j^{n+1} &= u_j^n + \kappa \frac{\Delta t}{\Delta x^2} (u_{j+1}^n - 2u_j^n + u_{j-1}^n) + f_j^{n+1}, \end{aligned} \tag{3.2}$$

with  $f_j^{n+1} = f((n+1)\Delta t, j\Delta x)$ .

**LEMMA 3.1.–** We define the consistency error of scheme [3.2] as

$$\begin{aligned} \varepsilon_j^{n+1} &= \frac{u((n+1)\Delta t, j\Delta x) - u(n\Delta t, j\Delta x)}{\Delta t} \\ &\quad - \kappa \frac{u(n\Delta t, (j+1)\Delta x) - 2u(n\Delta t, j\Delta x) + u(n\Delta t, (j-1)\Delta x)}{\Delta x^2} \\ &\quad - f((n+1)\Delta t, j\Delta x) \end{aligned}$$

$(t, x) \mapsto u(t, x)$  being the solution of [3.1]. Hence, there exists a constant  $C > 0$ , such that

$$|\varepsilon_j^n| \leq C(\Delta t + \Delta x^2).$$

The constant  $C$  appearing in this statement depends upon the final time  $0 < T < \infty$  and the infinite norms of the derivatives of  $u$  with respect to  $x$  and  $t$ , to a relatively high order. We shall see that the stability analysis depends critically upon the norm used. The following example shows that, without sufficient precaution, nonsensical numerical approximations can be obtained. We take the initial value  $u_j^0 = 0$  if  $j \notin \{j_0, \dots, j_0 + k_0\}$  and  $u_j^0 = 1$  if  $j \in \{j_0, \dots, j_0 + k_0\}$ , and additionally  $f = 0$ . The numerical conditions are such that  $\kappa\Delta t = \Delta x^2$ . In other words, we have

$$u_j^{n+1} = u_{j+1}^n + u_{j-1}^n - u_j^n.$$

The numerical solution is described by the following table:

$n = 0 :$	0	0	0	0	1	1	1	1	1	1	1	1	0	0	0	0	0	
$n = 1 :$	0	0	0	1	0	1	1	1	1	1	1	1	0	1	0	0	0	
$n = 2 :$	0	0	1	-1	2	0	1	1	1	1	1	1	0	2	-1	1	0	0
$n = 3 :$	0	1	-2	4	-3	3	0	1	1	1	0	3	-3	4	-2	1	0	
$n = 4 :$	1	-3	7	-9	10	-6	4	0	1	1	0	4	-6	10	-9	7	-3	1

This result is obviously absurd, with the analysis having shown us that the solution  $u(t, x)$  takes values within  $[0, 1]$ . We shall see that the numerical parameters  $\Delta t$  and  $\Delta x$  must satisfy certain conditions for scheme [3.2] to give a meaningful result.

### 3.1.1. $L^2$ stability (von Neumann analysis) and $L^\infty$ stability: convergence

In order to study von Neumann stability, we must limit the problem to one with periodic condition: we consider equation [3.1] applied in  $\mathbb{R}$  with the condition  $u(t, x) = u(t, x + 2\pi)$ . Evidently,  $f$  is also subject to this periodic constraint. The segment  $[0, 2\pi]$  is discretized by uniform subdivision

$$0 < \Delta x < \dots < (J + 1)\Delta x = 2\pi.$$

The matrix of finite differences corresponding to the operator  $\frac{d^2}{dx^2}$  is written in the form

$$A_\# = \frac{1}{\Delta x^2} \begin{pmatrix} -2 & 1 & 0 & \cdots & 0 & 1 \\ 1 & 0 & & & & 0 \\ 0 & & \ddots & & & 0 \\ \vdots & & & \ddots & & \vdots \\ 0 & & & & 0 & 1 \\ 1 & 0 & \cdots & 0 & 1 & -2 \end{pmatrix}. \quad [3.3]$$

Scheme [3.2] takes the matrix form

$$\frac{1}{\Delta t}(U^{n+1} - U^n) = \kappa A_{\#} U^n + F^{n+1},$$

where  $F^{n+1}$  is the vector with coordinates  $f_1^{n+1}, \dots, f_J^{n+1}$ . We associate the sequence  $u_j^n$  defined by [3.2] with the sequence of functions defined over  $[0, 2\pi[$  by

$$x \mapsto u^n(x) = \sum_{j=0}^J u_j^n \mathbf{1}_{j\Delta x \leq x < (j+1)\Delta x}(x)$$

which we extend over  $\mathbb{R}$  with  $2\pi$ -periodicity. We label as  $L_{\#}^2$  the set of functions  $g : \mathbb{R} \rightarrow \mathbb{R}$  which are  $2\pi$ -periodic and satisfy

$$\|g\|_{L_{\#}^2}^2 = \int_0^{2\pi} |g(y)|^2 dy < \infty.$$

Thus, we note that

$$\|u^n\|_{L_{\#}^2}^2 = \Delta x \sum_{j=0}^J |u_j^n|^2.$$

**DEFINITION 3.1.–** We say that scheme [3.2] is  $L^2$ -stable, if we have

$$\|u^n\|_{L_{\#}^2} \leq \|u^0\|_{L_{\#}^2} + \Delta t \sum_{m=1}^n \|F^m\|_{L_{\#}^2}.$$

**NOTE 3.2.–** The quantity  $\Delta t \sum_{m=1}^n \|F^m\|_{L_{\#}^2}$  can be interpreted as the  $L^1(0, (n+1)\Delta t)$  norm of the step-wise function

$$\sum_{m=1}^n \|F^m\|_{L_{\#}^2} \mathbf{1}_{(m-1)\Delta t \leq t < m\Delta t}.$$

The definition of  $L^2$  stability is hence natural, since it appears as a discrete version of the energy estimate

$$\begin{aligned} \frac{1}{2} \frac{d}{dt} \int_0^{2\pi} |u(t, x)|^2 dx + \kappa \int_0^{2\pi} |\partial_x u(t, x)|^2 dx &= \int_0^{2\pi} f(t, x) u(t, x) dx \\ &\leq \left( \int_0^{2\pi} |f(t, x)|^2 dx \right)^{1/2} \left( \int_0^{2\pi} |u(t, x)|^2 dx \right)^{1/2}. \end{aligned}$$

We interpret this relation as a differential inequality for  $\sqrt{\int_0^{2\pi} |u(t, x)|^2 dx}$  and, upon integration, we are led to

$$\sqrt{\int_0^{2\pi} |u(t, x)|^2 dx} \leq \sqrt{\int_0^{2\pi} |u(0, x)|^2 dx} + \int_0^t \sqrt{\int_0^{2\pi} |f(t, x)|^2 dx} ds.$$

This criterion leads to a constraint relating the steps in time  $\Delta t$  and in space  $\Delta x$ . We first recall that

$$\|u^n\|_{L^2_\#}^2 = \sum_{j=0}^J \int_{j\Delta x}^{(j+1)\Delta x} |u_j^n|^2 dx = \Delta x \sum_{j=0}^J |u_j^n|^2.$$

Then, by using Plancherel's theorem, we can express the  $L^2$  norm with the Fourier coefficients: for  $g \in L^2_\#$ , we set

$$\widehat{g}(k) = \frac{1}{2\pi} \int_0^{2\pi} g(x) e^{-ikx} dx$$

and we have  $\|g\|_{L^2_\#}^2 = \sum_{k \in \mathbb{Z}} |\widehat{g}(k)|^2$ . Now, on the one hand, we have

$$2\pi \widehat{u^{n+1}}(0) = \Delta x \sum_{j=0}^J u_j^n + \Delta t \Delta x \sum_{j=0}^J F_j^{n+1} = 2\pi \left( \widehat{u^n}(0) + \Delta t \widehat{F^{n+1}}(0) \right)$$

and, on the other hand, for  $k \neq 0$

$$\begin{aligned} 2\pi \widehat{u^n}(k) &= \sum_{j=0}^J u_j^n \int_{j\Delta x}^{(j+1)\Delta x} e^{-ikx} dx = \sum_{j=0}^J u_j^n \frac{e^{-i(j+1)k\Delta x} - e^{-ijk\Delta x}}{-ik} \\ &= \sum_{j=0}^J u_j^n e^{-i(j+1/2)k\Delta x} \frac{2 \sin(k\Delta x/2)}{k}. \end{aligned}$$

Thus, scheme [3.2] leads to

$$\begin{aligned} \widehat{u^{n+1}}(k) &= \sum_{j=0}^J e^{-i(j+1/2)k\Delta x} \frac{2 \sin(k\Delta x/2)}{k} \\ &\times \left( u_j^n + \kappa \frac{\Delta t}{\Delta x^2} (u_{j+1}^n - 2u_j^n + u_{j-1}^n) + \Delta t F_j^{n+1} \right) \end{aligned}$$

$$\begin{aligned}
&= \sum_{j=0}^J e^{-i(j+1/2)k\Delta x} \frac{2 \sin(k\Delta x/2)}{k} u_j^n \left( 1 + \kappa \frac{\Delta t}{\Delta x^2} (e^{-ik\Delta x} - 2 + e^{ik\Delta x}) \right) \\
&\quad + \Delta t \sum_{j=0}^J e^{-i(j+1/2)k\Delta x} \frac{2 \sin(k\Delta x/2)}{k} F_j^{n+1}.
\end{aligned}$$

We set

$$M(k) = \left( 1 + \kappa \frac{\Delta t}{\Delta x^2} (e^{-ik\Delta x} - 2 + e^{ik\Delta x}) \right)$$

a quantity called the *amplification factor* of scheme [3.2], such that

$$\begin{aligned}
\widehat{u^{n+1}}(k) &= \frac{M(k)}{2\pi} \sum_{j=0}^J e^{-i(j+1/2)k\Delta x} \frac{2 \sin(k\Delta x/2)}{k} u_j^n \\
&\quad + \frac{\Delta t}{2\pi} \sum_{j=0}^J e^{-i(j+1/2)k\Delta x} \frac{2 \sin(k\Delta x/2)}{k} F_j^{n+1} \\
&= M(k) \widehat{u^n}(k) + \Delta t \widehat{F^{n+1}}(k).
\end{aligned}$$

However, we note that

$$M(k) = 1 - 2\kappa \frac{\Delta t}{\Delta x^2} + 2\kappa \frac{\Delta t}{\Delta x^2} \cos(k\Delta x) = 1 - 4 \frac{\Delta t}{\Delta x^2} \sin^2(k\Delta x/2).$$

Thus, we have

$$1 - 4\kappa \frac{\Delta t}{\Delta x^2} \leq M(k) \leq 1$$

and the condition  $|M(k)| \leq 1$  is satisfied for all  $k$  when  $2\kappa \frac{\Delta t}{\Delta x^2} \leq 1$ . With this condition satisfied, it follows that

$$\|\widehat{u^{n+1}}\|_{\ell^2} \leq \|M \widehat{u^n}\|_{\ell^2} + \Delta t \|\widehat{F^{n+1}}\|_{\ell^2} \leq \|\widehat{u^n}\|_{\ell^2} + \Delta t \|\widehat{F^{n+1}}\|_{\ell^2}$$

from which we conclude

$$\|\widehat{u^n}\|_{\ell^2} \leq \|\widehat{u^0}\|_{\ell^2} + \Delta t \sum_{m=1}^n \|\widehat{F^m}\|_{\ell^2}.$$

We have thus established the following result.

**PROPOSITION 3.2.**— Scheme [3.2] is  $L^2$ —stable for the condition of stability

$$2\kappa \frac{\Delta t}{\Delta x^2} \leq 1. \quad [3.4]$$

As for ODEs, we will make use of the principle that the combination of consistency and stability properties allows the proof of the convergence of the numerical scheme, i.e. that the numerical solution is indeed an approximation of the solution to the continuous problem as the time and space steps tend towards 0.

**THEOREM 3.1.**— Suppose that [3.4] is satisfied. We set  $0 < T = (N + 1)\Delta t < \infty$ . We then obtain

$$\lim_{\Delta t \rightarrow 0, \Delta x \rightarrow 0} \left\{ \sup_{n \in \{0, \dots, N\}} \left( \Delta x \sum_{j=0}^J |u_j^n - u(n\Delta t, j\Delta x)|^2 \right) \right\} = 0.$$

**PROOF.**— We note that [3.4] implies that  $\Delta t \rightarrow 0$  as  $\Delta x \rightarrow 0$ . We also recall that the number of iterations in time,  $N$ , is related to the final time,  $T$ , and the time step,  $\Delta t$  (just as the number of discrete points in space,  $J$ , is related to the step size,  $\Delta x$ ). We denote  $e_j^n = u_j^n - u(n\Delta t, j\Delta x)$  and  $e^n = (e_0^n, \dots, e_J^n)$ , so that

$$e^{n+1} = e^n + \Delta t(\kappa A_\# e^n + \epsilon^{n+1})$$

with  $\epsilon_j^n$  being the local truncation error. The stability implies that

$$\|e^n\|_{L^2_\#} \leq \|e^0\|_{L^2_\#} + \Delta t \sum_{m=1}^n \sqrt{\Delta x \sum_{j=0}^J |\epsilon_j^m|^2}.$$

Since  $e^0 = 0$ , the consistency estimate leads to the inequality

$$\|e^n\|_{L^2_\#}^2 \leq n\Delta t C (\Delta t + \Delta x^2) \leq CT (\Delta t + \Delta x^2). \quad \square$$

In this statement, and those that follow, we must always take care that the final time,  $T$ , and the size of the domain, here equal to 1, are fixed and the numerical parameters (step size  $\Delta t, \Delta x$  vs. the number of iterations  $N$  and discretization points  $J$ ) are related to each other by the relationships  $N\Delta t = T$  and  $(J + 1)\Delta x = 1$ .

We have seen that the continuous problem [3.1] satisfies a maximum principle: if  $f = 0$  and the initial value satisfies  $0 \leq u_{\text{Init}} \leq M < \infty$ , then the same is true for the solution, since  $0 \leq u(t, x) \leq M$  is satisfied for all  $(t, x)$ . We would like to identify the conditions on  $\Delta t$  and  $\Delta x$  that would preserve this property. We shall deduce from this an estimate of the convergence in the  $L^\infty$  norm.

**DEFINITION 3.2.**— We say that scheme [3.2] is  $L^\infty$ –stable if, given  $0 \leq u_j^0 \leq M < \infty$  and  $f_j^n = 0$  for all  $j$ , we have  $0 \leq u_j^n \leq M$  for all  $n, j \in \mathbb{N}$ .

We re-write [3.2] in the form

$$u_j^{n+1} = \left(1 - 2\kappa \frac{\Delta t}{\Delta x^2}\right) u_j^n + \kappa \frac{\Delta t}{\Delta x^2} (u_{j+1}^n + u_{j-1}^n),$$

such that the condition  $2\kappa \frac{\Delta t}{\Delta x^2} \leq 1$  reveals  $u_j^{n+1}$  as a convex combination of  $u_j^n$ ,  $u_{j+1}^n$  and  $u_{j-1}^n$ .

**PROPOSITION 3.3.**— Scheme [3.2] is  $L^\infty$ –stable given the condition [3.4].

**THEOREM 3.2.**— Suppose [3.4]. Then, scheme [3.2] is convergent in that, with fixed  $0 < T = N\Delta t < \infty$ , we have

$$\lim_{\Delta t \rightarrow 0, \Delta x \rightarrow 0} \left( \sup_{n \in \{0, \dots, N\}, j} |u_j^n - u(n\Delta t, j\Delta x)| \right) = 0.$$

**PROOF.**— We again use  $e_j^n = u_j^n - u(n\Delta t, j\Delta x)$  and the relation

$$\frac{e_j^{n+1} - e_j^n}{\Delta t} - \kappa \frac{e_{j+1}^n - 2e_j^n + e_{j-1}^n}{\Delta x^2} = -\epsilon_j^{n+1}$$

the local truncation error, or

$$e_j^{n+1} = \left(1 - 2\kappa \frac{\Delta t}{\Delta x^2}\right) e_j^n + \kappa \frac{\Delta t}{\Delta x^2} (e_{j+1}^n + e_{j-1}^n) + \Delta t \epsilon_j^{n+1}.$$

By [3.4], we have  $1 - 2\kappa \frac{\Delta t}{\Delta x^2} \geq 0$ , which gives the inequality

$$\begin{aligned} |e_j^{n+1}| &\leq \left(1 - 2\kappa \frac{\Delta t}{\Delta x^2}\right) \sup_\ell |e_\ell^n| + 2\kappa \frac{\Delta t}{\Delta x^2} \sup_k |e_k^n| + \Delta t |\epsilon_j^{n+1}| \\ &\leq \sup_\ell |e_\ell^n| + C \Delta t (\Delta t + \Delta x^2) \end{aligned}$$

by making use of the consistency estimate. It follows that

$$|e_j^n| \leq \sup_\ell |e_\ell^0| + C n \Delta t (\Delta t + \Delta x^2) \leq \sup_\ell |e_\ell^0| + CT (\Delta t + \Delta x^2). \quad \square$$

More generally, the stability condition [3.4] can again be interpreted with the help of the spectral properties of the matrix of the linear system corresponding to the discrete problem. For this, we recall the following statement.

LEMMA 3.2.– The sequence  $(A^k)_{k \in \mathbb{N}}$  is bounded if the modulus of the eigenvalues of  $A$  are less than or equal to 1 and the eigenvalues of modulus 1 are semisimple.

PROOF.– The demonstration uses the Jordan decomposition of matrix  $A$ :  $A = PJP^{-1}$ , where the matrix  $J$  is made up of  $r$  diagonal blocks of the form

$$\left( \begin{array}{cccccc} \lambda & 1 & 0 & \cdots & 0 \\ 0 & & & & \\ \vdots & & & & 0 \\ 0 & & & & \\ 0 & \cdots & 0 & & \lambda \end{array} \right) \in \mathcal{M}_q.$$

For a given eigenvalue  $\lambda$ , the number of these blocks in the decomposition corresponds to the geometric multiplicity of  $\lambda$ , while the sum of the dimensions  $q$  of these blocks gives the algebraic multiplicity. When  $\lambda$  is semisimple, there is no nilpotent part in the Jordan decomposition of this eigenvalue and we can re-write the block corresponding to  $\lambda$  in a purely diagonal form. We therefore have  $A^k = P J^k P^{-1}$  where, for sufficiently large  $k$ ,  $J^k$  is composed of diagonal blocks of the form

$$\left( \begin{array}{ccccc} \lambda^k & C_k^1 \lambda^{k-1} & C_k^2 \lambda^{k-2} & \cdots & C_k^{q-1} \lambda^{k-q+1} \\ 0 & & & & \\ \vdots & & & & C_k^2 \lambda^{k-2} \\ 0 & & & & C_k^1 \lambda^{k-1} \\ 0 & \cdots & 0 & & \lambda^k \end{array} \right)$$

with  $C_k^p = \frac{k!}{p!(k-p)!}$ , where  $p$  is smaller than  $N$ , the dimension of  $A$ . If  $|\lambda| < 1$ , the elements of this matrix tend towards 0 as  $k \rightarrow \infty$ . For semisimple  $|\lambda|$  with  $\lambda = 1$ , the corresponding Jordan block is also bounded, since it has no nilpotent component.  $\square$

This statement also allows us to establish the stability of [3.2] under condition [3.4]. Under Dirichlet conditions, the problem is written in the matrix form  $U_{k+1} = (\mathbb{I} + \kappa\Delta t A)U_k$  with

$$A = \frac{1}{\Delta x^2} \begin{pmatrix} -2 & 1 & 0 & \cdots & 0 & 0 \\ 1 & -2 & 1 & \cdots & 0 & 0 \\ 0 & 1 & -2 & 1 & \cdots & 0 \\ \vdots & 0 & \ddots & \ddots & \ddots & 0 \\ 0 & 0 & \cdots & 0 & 1 & -2 \end{pmatrix}. \quad [3.5]$$

One method of defining the stability is to require that  $(\mathbb{I} + \kappa\Delta t A)^k$  remains uniformly bounded for  $k \in \mathbb{N}$ . We have seen that the eigenvalues of  $(\mathbb{I} + \kappa\Delta t A)$  (with  $\Delta x = 1/(J+1)$ ) are

$$1 - 4\kappa\Delta t(J+1)^2 \sin^2 \left( \frac{k\pi}{2(J+1)} \right)$$

for  $k \in \{1, \dots, J\}$ . The spectrum of  $(\mathbb{I} + \Delta t A)$  lies within  $] -1, +1[$  when

$$\kappa\Delta t \times (J+1)^2 < \frac{1}{2},$$

a condition that ensures the stability of the method.

### 3.1.2. Implicit schemes

Condition [3.4], which ensures the stability and convergence of scheme [3.2], is very restrictive: for a given spatial precision, set by  $\Delta x$ , the time step must be of the order of  $\Delta x^2$ , hence placing a significant penalty on the computation time required to reach a fixed final simulation time  $T$ . Learning from ODE analysis, we can aim to relax this constraint by using an implicit scheme. To this end, we now construct numerical approximations  $u_j^n$  with the scheme

$$\begin{aligned} u_j^0 &= u_{\text{Init}}(j\Delta x), \\ u_j^{n+1} &= u_j^n + \kappa \frac{\Delta t}{\Delta x^2} (u_{j+1}^{n+1} - 2u_j^{n+1} + u_{j-1}^{n+1}) + \Delta t f((n+1)\Delta t, j\Delta x). \end{aligned} \quad [3.6]$$

If the problem is considered to be posed over the segment  $x \in [0, 1]$ , given the Dirichlet conditions  $u(0) = 0 = u(1)$ , this signifies that the vector  $U^{n+1} = (u_1^{n+1}, \dots, u_J^{n+1})$ , where  $(J + 1)\Delta x = 1$ , is obtained by solving the linear system

$$(\mathbb{I} - \kappa\Delta t A)U^{n+1} = U^n + \Delta t F^{n+1},$$

the matrix  $A$  being defined in [3.5]. If the problem is considered with periodic conditions, the linear system is written as

$$(\mathbb{I} - \kappa\Delta t A_{\#})U^{n+1} = U^n.$$

with  $A_{\#}$  defined in [3.3].

The consistency error here is defined by

$$\varepsilon_j^{n+1} = \frac{u((n+1)\Delta t, j\Delta x) - u(n\Delta t, j\Delta x)}{\Delta t}$$

$$-\kappa \frac{u((n+1)\Delta t, (j+1)\Delta x) - 2u((n+1)\Delta t, j\Delta x) + u((n+1)\Delta t, (j-1)\Delta x)}{\Delta x^2} \\ - \Delta t f((n+1)\Delta t, j\Delta x).$$

The scheme is of order 1 with respect to time, and of order 2 with respect to the spatial variable, since  $\varepsilon_j^n$  satisfies

$$|\varepsilon_j^n| \leq C(\Delta t + \Delta x^2)$$

(where  $C > 0$  depends upon  $\|u\|_{C^4([0,T] \times [0,2\pi])}$ ). Let us start by studying the von Neumann stability.

**PROPOSITION 3.4.–** Scheme [3.6] is unconditionally  $L^2$ –stable.

**PROOF.–** This time we have

$$\widehat{u^n}(k) + \Delta t \widehat{F^{n+1}}(k) \\ = \sum_{j=0}^J e^{-i(j+1/2)k\Delta x} \frac{2 \sin(k\Delta x/2)}{k} \left( u_j^{n+1} - \kappa \frac{\Delta t}{\Delta x^2} (u_{j+1}^{n+1} - 2u_j^{n+1} + u_{j-1}^{n+1}) \right)$$

$$\begin{aligned}
&= \sum_{j=0}^J e^{-i(j+1/2)k\Delta x} \frac{2\sin(k\Delta x/2)}{k} u_j^{n+1} \left( 1 - \kappa \frac{\Delta t}{\Delta x^2} (e^{-ik\Delta x} - 2 + e^{ik\Delta x}) \right) \\
&= \sum_{j=0}^J e^{-i(j+1/2)k\Delta x} \frac{2\sin(k\Delta x/2)}{k} u_j^{n+1} \left( 1 + 4\kappa \frac{\Delta t}{\Delta x^2} \sin^2(k\Delta x/2) \right).
\end{aligned}$$

The amplification factor is now expressed as

$$0 \leq M(k) = \frac{1}{1 + 4\kappa \frac{\Delta t}{\Delta x^2} \sin^2(k\Delta x/2)} \leq 1$$

and we have

$$\widehat{u^{n+1}}(k) = M(k)(\widehat{u^n}(k) + \Delta t \widehat{F^{n+1}}(k)).$$

We immediately deduce from this the stability estimate.  $\square$

As for the explicit scheme, we can combine the stability and consistency properties and use them to find the convergence of the scheme.

**COROLLARY 3.1.**— Scheme [3.6] is convergent in the sense that, with fixed  $0 < T = N\Delta t < \infty$ , we have

$$\lim_{\Delta t \rightarrow 0, \Delta x \rightarrow 0} \left\{ \sup_{n \in \{0, \dots, N\}} \left( \Delta x \sum_{j=0}^J |u_j^n - u(n\Delta t, j\Delta x)|^2 \right) \right\} = 0.$$

In order to establish the same result for the  $L^\infty$  norm, we will exploit the fact that  $\mathbb{I} - \kappa\Delta t A$  is an  $M$ –matrix (see definition 2.3 and theorem 2.7). We will use this to deduce that scheme [3.6] preserves the maximum principle, is  $L^\infty$ –stable and, finally, is convergent.

**PROPOSITION 3.5.**— For all  $\Delta t, \Delta x > 0$ , scheme [3.6] exhibits the following properties:

- i) if  $u_{\text{Init}} \geq 0$  and  $f \geq 0$  then  $u_j^n \geq 0$  for all  $n, j$ ;
- ii) if  $f = 0$  and  $0 \leq u_{\text{Init}} \leq M$ , then  $0 \leq u_j^n \leq M$  for all  $n, j$ ;
- iii) the scheme is unconditionally  $L^\infty$ –stable :  $\sup_j |u_j^n| \leq \sup_j |u_j^0| + n\Delta t \|f\|_\infty$ .

PROOF.– Point i) is a direct consequence of the fact that  $\mathbb{I} - \kappa\Delta t A$  is an  $M$ -matrix. We then write the scheme in the form

$$\left(1 + 2\kappa \frac{\Delta t}{\Delta x^2}\right)u_j^{n+1} - \kappa \frac{\Delta t}{\Delta x^2} u_{j+1}^{n+1} - \kappa \frac{\Delta t}{\Delta x^2} u_{j-1}^{n+1} = u_j^n + \Delta t f((n+1)\Delta t, j\Delta x)$$

which leads to

$$\left(1 + 2\kappa \frac{\Delta t}{\Delta x^2}\right) \sup_j |u_j^{n+1}| \leq \sup_k |u_k^n| + \Delta t \|f\|_\infty + 2\kappa \frac{\Delta t}{\Delta x^2} \sup_k |u_k^{n+1}|,$$

from which we deduce iii) and hence ii).  $\square$

COROLLARY 3.2.– Scheme [3.6] is convergent in the sense that, with fixed  $0 < T = N\Delta t < \infty$ , we have

$$\lim_{\Delta t \rightarrow 0, \Delta x \rightarrow 0} \left( \sup_{n \in \{0, \dots, N\}, j} |u_j^n - u(n\Delta t, j\Delta x)| \right) = 0.$$

We can try to improve the order of consistency in time, to obtain a scheme of order 2 in both variables. One idea could be to use a centered-time scheme, approximating  $\partial_t u$  by  $\frac{u_j^{n+1} - u_j^{n-1}}{2\Delta t}$ , but, besides the issue of defining the first iteration, the scheme thus obtained is unstable (see Figures 3.5). Another approach could be to reproduce the method adopted for ODEs and which leads to the Crank–Nicolson scheme. More generally, we construct, for  $0 \leq \theta \leq 1$ , the following scheme

$$\begin{aligned} u_j^0 &= u_{\text{Init}}(j\Delta x), \\ u_j^{n+1} &= u_j^n + \kappa \frac{\Delta t}{\Delta x^2} \left( \theta(u_{j+1}^{n+1} - 2u_j^{n+1} + u_{j-1}^{n+1}) \right. \\ &\quad \left. + (1-\theta)(u_{j+1}^n - 2u_j^n + u_{j-1}^n) \right). \end{aligned} \tag{3.7}$$

The numerical solution is updated by solving the linear system

$$(\mathbb{I} - \theta\kappa\Delta t A)U^{n+1} = (\mathbb{I} + (1-\theta)\kappa\Delta t A)U^n.$$

LEMMA 3.3.– The consistency error of scheme [3.7] is of order 1 in time and of order 2 in space for  $0 \leq \theta \leq 1$  and  $\theta \neq 1/2$ , and is of order 2 in both variables for  $\theta = 1/2$ :

$$\begin{aligned} \varepsilon_j^n &= \frac{u((n+1)\Delta t, j\Delta x) - u(n\Delta t, j\Delta x)}{\Delta t} \\ &\quad - \kappa\theta \frac{u((n+1)\Delta t, (j+1)\Delta x) - 2u((n+1)\Delta t, j\Delta x) + u((n+1)\Delta t, (j-1)\Delta x)}{\Delta x^2} \\ &\quad - \kappa(1-\theta) \frac{u(n\Delta t, (j+1)\Delta x) - 2u(n\Delta t, j\Delta x) + u(n\Delta t, (j-1)\Delta x)}{\Delta x^2} \end{aligned}$$

satisfying

$$|\varepsilon_j^n| \leq C(\Delta t + \Delta x^2) \quad \text{if } 0 \leq \theta \leq 1 \text{ and } \theta \neq 1/2,$$

$$|\varepsilon_j^n| \leq C(\Delta t^2 + \Delta x^2) \quad \text{if } \theta = 1/2.$$

In this statement, the constant  $C$  depends upon the  $L^\infty$  norms of  $\partial_{t,x}^k u$  up until a sufficiently large index  $k \geq 4$ . The interest in this family of schemes, in particular for the case  $\theta = 1/2$ , is bolstered by the following stability analysis.

**PROPOSITION 3.6.–** Scheme [3.7] is unconditionally  $L^2$ –stable for all  $1/2 \leq \theta \leq 1$ ; for  $0 \leq \theta < 1/2$ , it is stable under the condition  $\kappa \frac{\Delta t}{\Delta x^2} \leq \frac{1}{2(1-2\theta)}$ .

**PROOF.–** For this scheme, we obtain the relation

$$\begin{aligned} & \sum_{j=0}^J e^{-i(j+1/2)k\Delta x} \frac{2 \sin(k\Delta x/2)}{k} u_j^{n+1} \left( 1 + 4\theta\kappa \frac{\Delta t}{\Delta x^2} \sin^2(k\Delta x/2) \right) \\ &= \sum_{j=0}^J e^{-i(j+1/2)k\Delta x} \frac{2 \sin(k\Delta x/2)}{k} u_j^n \left( 1 - 4(1-\theta)\kappa \frac{\Delta t}{\Delta x^2} \sin^2(k\Delta x/2) \right) \\ & \quad + \Delta t \widehat{F^{n+1}}(k). \end{aligned}$$

The amplification factor then becomes

$$M(k) = \frac{1 - 4(1-\theta)\kappa \frac{\Delta t}{\Delta x^2} \sin^2(k\Delta x/2)}{1 + 4\theta\kappa \frac{\Delta t}{\Delta x^2} \sin^2(k\Delta x/2)}$$

and we have

$$|\widehat{u^{n+1}}(k)| \leq |M(k) \widehat{u^n}(k)| + \Delta t |\widehat{F^{n+1}}(k)|,$$

since  $1 + 4\theta\kappa \frac{\Delta t}{\Delta x^2} \sin^2(k\Delta x/2) \geq 1$ . We always have  $M(k) \leq 1$ , but the estimate  $M(k) \geq -1$  sets

$$4\kappa \frac{\Delta t}{\Delta x^2} \sin^2(k\Delta x/2) (1 - 2\theta) \leq 2.$$

This relationship is always satisfied when  $\theta \geq 1/2$ ; otherwise, we need to set  $4\kappa \frac{\Delta t}{\Delta x^2} \leq \frac{2}{1-2\theta}$ .  $\square$

By repeating the arguments discussed above, we can deduce the convergence of the scheme in terms of the  $L^2$  norm.

COROLLARY 3.3.– Let  $0 \leq \theta \leq 1$ ; if  $0 \leq \theta < 1/2$ , we also assume that  $\kappa \frac{\Delta t}{\Delta x^2} \leq \frac{1}{2(1-2\theta)}$ . Then, scheme [3.7] is convergent in the sense that, with fixed  $0 < T = N\Delta t < \infty$ , we have

$$\lim_{\Delta t \rightarrow 0, \Delta x \rightarrow 0} \left\{ \sup_{n \in \{0, \dots, N\}} \left( \Delta x \sum_{j=0}^J |u_j^n - u(n\Delta t, j\Delta x)|^2 \right) \right\} = 0.$$

In the case where  $\theta = 1/2$ , the error is of order 2: it is controlled by  $\Delta t^2 + \Delta x^2$ . However, this result applies for the  $L^2$  norm only: the scheme works with respect to the  $L^\infty$  norm at the price of new constraints, again requiring a choice of time step of order  $\Delta x^2$ , as demonstrated by the following stability analysis.

PROPOSITION 3.7.– Suppose that  $\kappa \frac{\Delta t}{\Delta x^2} \leq \frac{1}{2(1-\theta)}$ . Then, scheme [3.7] exhibits the following properties:

- i) if  $u_{\text{Init}} \geq 0$  and  $f \geq 0$  then  $u_j^n \geq 0$  for all  $n, j$ ;
- ii) if  $f = 0$  and  $0 \leq u_{\text{Init}} \leq M$ , then  $0 \leq u_j^n \leq M$  for all  $n, j$ ;
- iii) the scheme is  $L^\infty$ -stable :  $\sup_j |u_j^n| \leq \sup_j |u_j^0| + n\Delta t \|f\|_\infty$ .

In this case, it is convergent in the sense that, with fixed  $0 < T = N\Delta t < \infty$ , we have

$$\lim_{\Delta t \rightarrow 0, \Delta x \rightarrow 0} \left( \sup_{n \in \{0, \dots, N\}, j} |u_j^n - u(n\Delta t, j\Delta x)| \right) = 0.$$

PROOF.– In matrix form, the scheme is written as

$$(\mathbb{I} - \theta\kappa\Delta t A)U^{n+1} = (\mathbb{I} + (1-\theta)\kappa\Delta t A)U^n + \Delta t F^{n+1/2}. \quad [3.8]$$

We have seen that  $(\mathbb{I} - \theta\kappa\Delta t A)$  is an  $M$ -matrix; hence, the components of  $U^{n+1}$  are positive, provided that the components of the right-hand side of [3.8] are also positive. Suppose  $f \geq 0$  and the components of  $U^n$  are positive, this is satisfied when all components of  $(\mathbb{I} + (1-\theta)\kappa\Delta t A)$  are positive. Thus, this requires  $1 - 2(1-\theta)\kappa \frac{\Delta t}{\Delta x^2} \geq 0$ .  $\square$

### 3.1.3. Finite element discretization

We construct a numerical approximation

$$u_h(t, x) = \sum_{n=0}^N \sum_{j=1}^M u_j^n \mathbf{1}_{[n\Delta t, (n+1)\Delta t]} \phi_j(x)$$

where the functions  $\phi_1, \dots, \phi_J$  form a basis of the space of approximation  $V_h$ . (Here, the index  $h$  and the number of basis functions  $J$  are related and dependent upon both the refinement of the discretization and the degree of the polynomials used to define the approximation over each element.) We define a linear system satisfied by  $u_j^n$  by making use of the variational structure, adapted to the time-dependent problem. For  $\varphi \in C_c^\infty([0, 1[)$ , we have

$$\frac{d}{dt} \int_0^1 u(t, x) \varphi(x) dx + \int_0^1 a(x) \frac{d}{dx} u(t, x) \frac{d}{dx} \varphi(x) dx = \int_0^1 f(t, x) \varphi(x) dx.$$

The discrete problem is obtained by using the explicit Euler scheme to describe the evolution of  $t \mapsto \int_0^1 u(t, x) \varphi_j(x) dx$ . We obtain the following linear system

$$\frac{1}{\Delta t} (M U^{n+1} - M U^n) = A U^n$$

with

$$A_{i,j} = \int \frac{d}{dx} \phi_i(x) \frac{d}{dx} \phi_j(x) dx$$

and

$$M_{i,j} = \int \phi_i(x) \phi_j(x) dx,$$

which we call the *stiffness matrix* and the *mass matrix*, respectively. Thus, even with an explicit scheme,  $U^{n+1}$  is not directly determined and we must resolve a linear system to update the unknown values. For example, for the  $\mathbb{P}_1$  elements on a uniform mesh with step size  $h > 0$ , we have

$$M_{i,j} = 0 \quad \text{if } |i - j| > 1,$$

and

$$\begin{aligned} M_{i,i+1} &= \int_{ih}^{(i+1)h} \frac{(i+1)h - x}{h} \frac{x - ih}{h} dx \\ &= - \int_{ih}^{(i+1)h} \frac{(x - ih)^2}{h^2} dx + \int_{ih}^{(i+1)h} \frac{x - ih}{h} dx \\ &= h \left( -\frac{1}{3} + \frac{1}{2} \right) = \frac{h}{6}, \\ M_{i,i} &= \int_{ih}^{(i+1)h} \left| \frac{(i+1)h - x}{h} \right|^2 dx + \int_{(i-1)h}^{ih} \left| \frac{x - (i-1)h}{h} \right|^2 dx = \frac{2h}{3}. \end{aligned}$$

In order to avoid the resolution of a linear system like this, we sometimes use an approximate expression of the matrix  $M$  which allows direct determination of  $U^{n+1}$ . The trapeze formula, which is explained further in Appendix 2, allows the approximation of  $\int_0^1 g(y) dy$  by  $h \sum_{j=0}^{J-1} (g(jh) + \frac{g((j+1)h) - g(jh)}{2}) = \frac{h}{2} \sum_{j=0}^{J-1} (g(jh) + g((j+1)h))$ . By using this approximation to evaluate the coefficients of the mass matrix, we define

$$\begin{aligned}\widetilde{M}_{i,i} &= \frac{h}{2} \sum_{j=0}^{J-1} (|\phi_i(jh)|^2 + |\phi_i((j+1)h)|^2) = \frac{h}{2} 2|\phi_i(ih)|^2 = h, \\ \widetilde{M}_{i,j} &= \frac{h}{2} \sum_{k=0}^{J-1} (\phi_i(kh)\phi_j(kh) + \phi_i((k+1)h)\phi_j((k+1)h)) = 0 \quad \text{for } i \neq j.\end{aligned}$$

We therefore settle with setting

$$\frac{1}{\Delta t}(\widetilde{M}U^{n+1} - \widetilde{M}U^n) = \frac{h}{\Delta t}(U^{n+1} - U^n) = AU^n.$$

Such techniques, consisting of simplifying the expression for the mass matrix, are referred to as *mass condensation methods* (mass-lumping). Now, resolution of the linear system is trivial since  $\widetilde{M}$  is a diagonal matrix (and even scalar in this example where we actually retrieve the finite difference scheme). Beyond a simplification of the linear system determining  $U^{n+1}$ , another advantage of this method is that it restores the maximum principle (see [THO 06, theorem 15.5]). However, it can also induce degradation of the convergence order [CHA 12].

### 3.1.4. Numerical illustrations

We start by studying the case where the domain of calculation is the interval  $[0, 2\pi]$  with periodic conditions and the initial data is

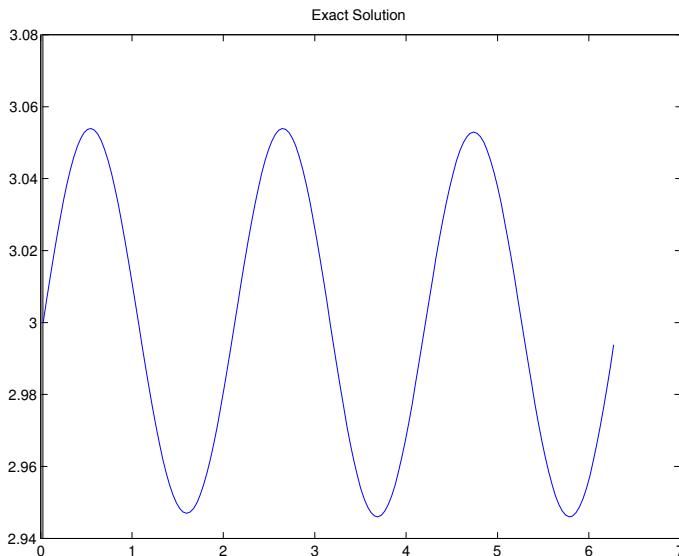
$$u_{\text{Init}}(x) = 3 + 0.8 \sin(3x) + 1.2 \sin(5x). \quad [3.9]$$

In this case, Fourier analysis explicitly yields the solution

$$u(t, x) = 3 + 0.8e^{-3^2 \kappa t} \sin(3x) + 1.2e^{-5^2 \kappa t} \sin(5x)$$

where  $\kappa > 0$  is the diffusion coefficient (see Figure 3.1 for the graph of this solution at final time  $T = 0.1$ ). Here, we have fixed  $\kappa = 3$ . First, we compare the explicit and implicit Euler methods and the Crank–Nicolson method, with a finite difference discretization over a uniform mesh of step size  $\Delta x > 0$ , under the condition  $2\kappa \frac{\Delta t}{\Delta x^2} < 1$ , which then ensures the stability of the  $L^2$  norm as well as the  $L^\infty$  norm

for these three finite difference schemes. Specifically, here, we set  $2\kappa \frac{\Delta t}{\Delta x^2} = 0.9$ . We also compare these numerical solutions with a discrete Fourier transform method  $\text{ifft}(\exp(-\kappa \ell^2 t) \text{fft}(u_{\text{Init}}))$ , where  $\ell$  spans a discrete set of frequencies (see Figure 3.2). The final time is  $T = 0.1$  and we test uniform meshes with  $N = n \times N_0$  points,  $n \in \{1, \dots, 7\}$  and  $N_0 = 24$ . We evaluate the evolution of the  $L^2$  and  $L^\infty$  norms of the difference between the numerical solution and the exact solution as a function of the discretization (see Figure 3.3). We indeed observe convergence of order 2 as a function of the spatial discretization length.

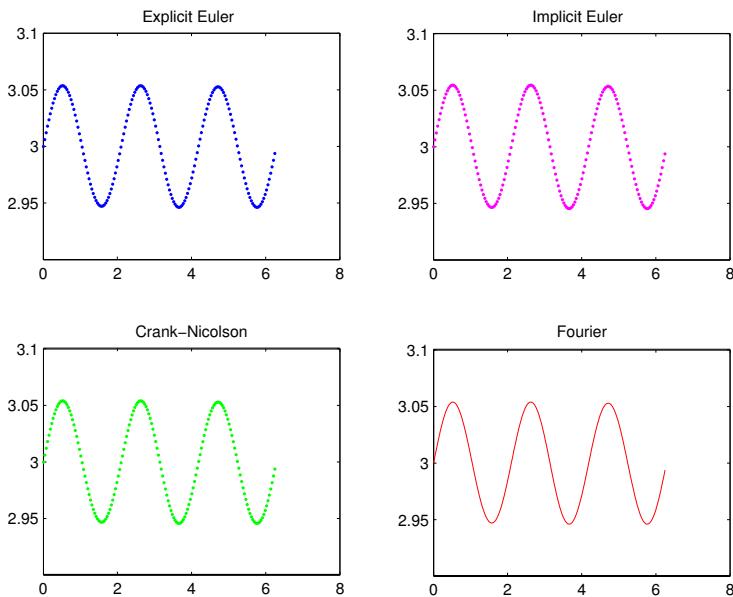


**Figure 3.1.** Exact solution for the heat equation

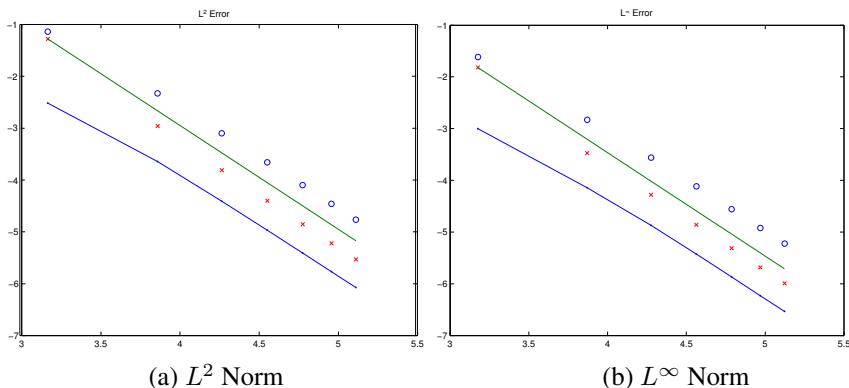
If we free ourselves of the CFL condition, the explicit Euler scheme becomes unstable (another illustration is given later on). Hence, we only continue comparison of the implicit Euler and Crank–Nicolson methods, for the same meshes, but this time fixing the time interval  $\Delta t = 2 \cdot 10^{-3}$ . This time we observe, as shown in Figure 3.4, that the error curves as a function of spatial step size for the implicit Euler method stop decreasing; it is the time error that starts to dominate. For the Crank–Nicolson method, over the meshes considered, we again observe a decrease in the error of order 2.

Figures 3.5 show the results obtained with the scheme

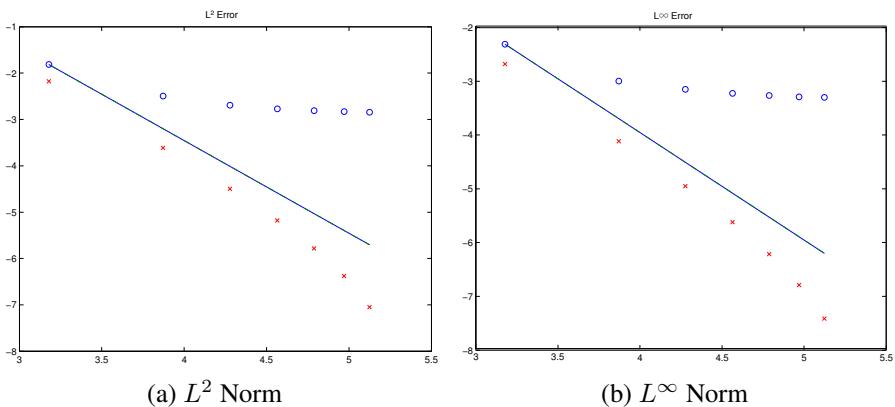
$$\frac{u_j^{n+1} - u_j^{n-1}}{2\Delta t} = \kappa \frac{u_{j+1}^n - 2u_j^n + u_{j-1}^n}{\Delta x^2}.$$



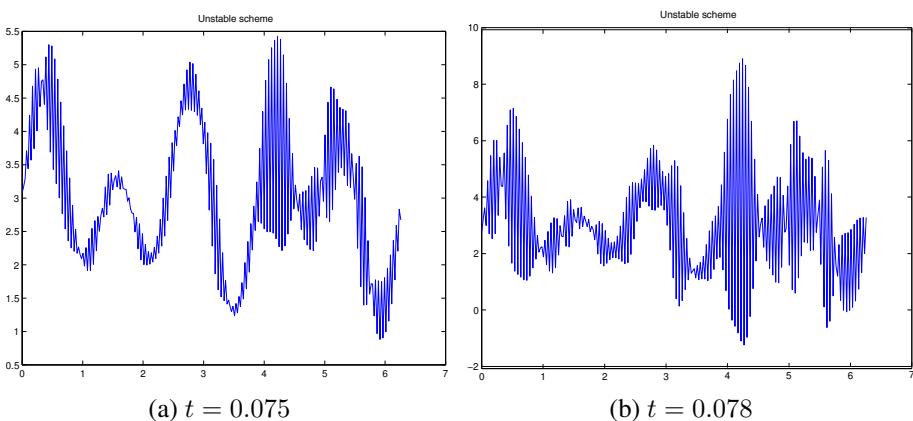
**Figure 3.2.** Numerical solutions (finite difference) for the heat equation under CFL condition. For the color version of this figure, see [www.iste.co.uk/goudon/mathmodel.zip](http://www.iste.co.uk/goudon/mathmodel.zip)



**Figure 3.3.** Evolution of the error as a function of mesh under CFL: the explicit Euler ( $\times$ ), implicit Euler ( $\circ$ ), and Crank–Nicolson (blue) finite difference schemes and straight line of slope 2 (green), log–log scale. For the color version of this figure, see [www.iste.co.uk/goudon/mathmodel.zip](http://www.iste.co.uk/goudon/mathmodel.zip)

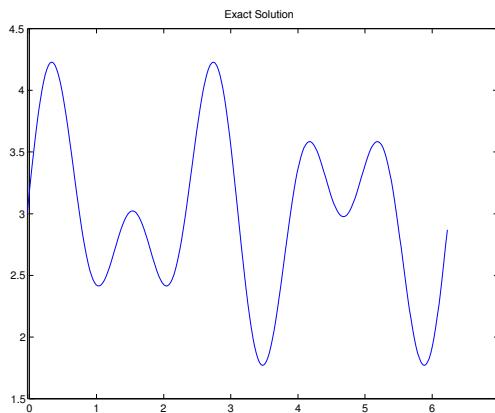


**Figure 3.4.** Evolution of the error as a function of mesh without obeying the CFL: the implicit Euler ( $\circ$ ) and Crank–Nicolson ( $\times$ ) finite difference schemes and straight line of slope 2, log–log scale. For the color version of this figure, see [www.iste.co.uk/goudon/mathmodel.zip](http://www.iste.co.uk/goudon/mathmodel.zip)



**Figure 3.5.** Simulation with an unstable scheme

The exact solution is shown in Figure 3.6. This scheme is reliable of order 2 in time and space, but it is unstable, as the figures show, even though the time interval has been chosen such that  $2\kappa \frac{\Delta t}{\Delta x^2} < 1$ .



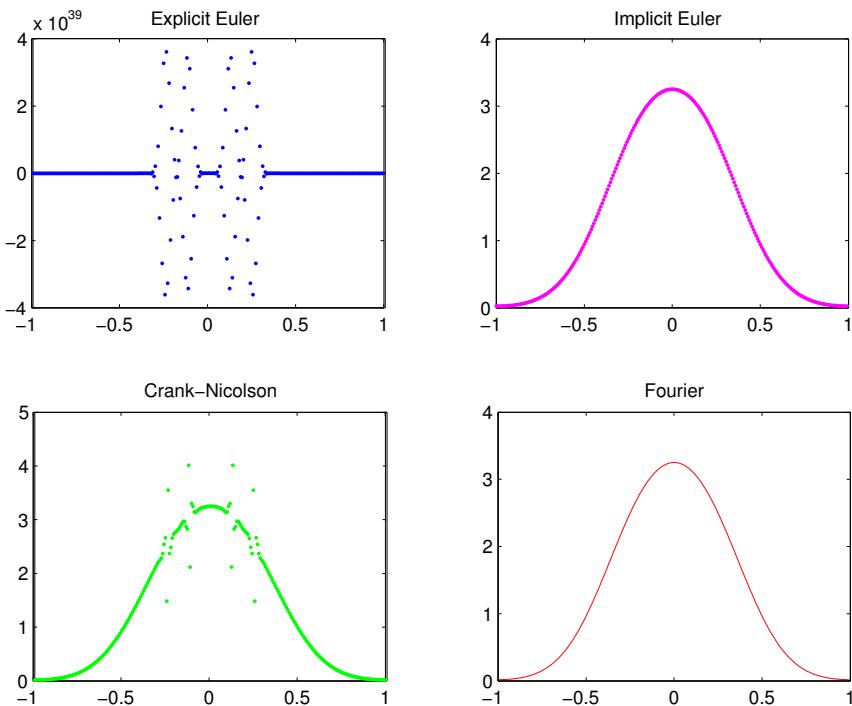
**Figure 3.6.** Exact solution at time  $t = 0.0078$

Finally, we test the behavior of these finite difference schemes for the initial value

$$u_{\text{Init}}(x) = 10 \times (\mathbf{1}_{[-1/4, -1/8]}(x) + (\mathbf{1}_{[1/8, 1/4]}(x)) + \frac{1}{2} \times \mathbf{1}_{[-1/8, 1/8]}(x) \quad [3.10]$$

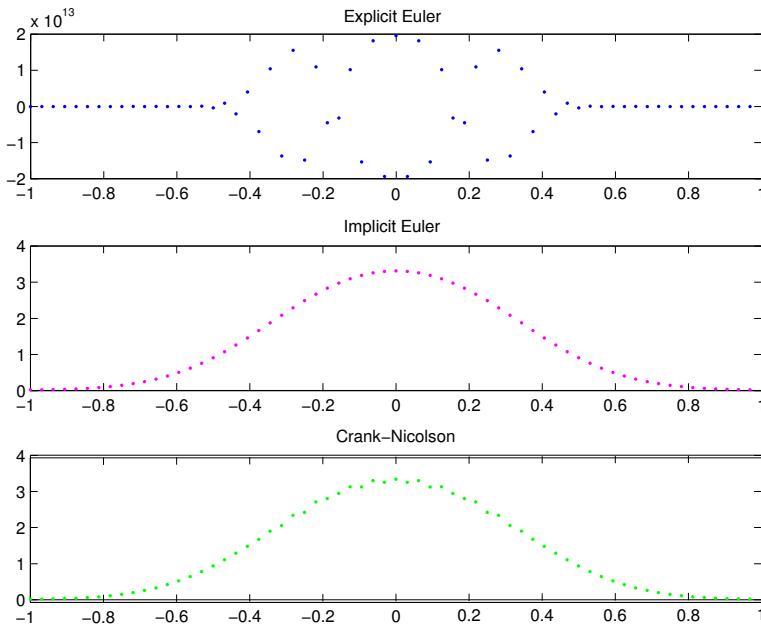
over the interval  $[-1, +1]$ . We are still working with periodic boundaries and  $\kappa = 3$ . The spatial discretization is set by 256 equally separated points (i.e.  $\Delta x = 0.0078$ ). The time step is  $5 \cdot 10^{-4}$  and we stop the simulation at time  $T = 0.01$  (20 time iterations). The result is presented in Figure 3.7. We clearly observe the instability of the explicit Euler scheme under these conditions. We also note parasitic oscillations for the Crank–Nicolson scheme, revealing the lack of stability in the  $L^\infty$  norm for the numerical conditions adopted. However, these oscillations fade as time progresses.

We meet the same problem with a finite element discretization strategy. Figures 3.8 and 3.9 show the solutions obtained at the final time  $T = 0.01$  for the initial value [3.10] with the explicit, implicit and Crank–Nicolson  $\mathbb{P}_1$  methods or the explicit, explicit with mass condensation, implicit and Crank–Nicolson  $\mathbb{P}_2$  methods. The boundary conditions are the homogeneous Dirichlet conditions and we have taken for this simulation a time step  $\Delta t = 4 \frac{\Delta x^2}{2\kappa}$ , which violates the stability condition obtained by finite difference methods of analysis (we recall that finite difference and finite element schemes in one dimension over a uniform mesh are equivalent). The explicit schemes are unstable, even with mass condensation; the Crank–Nicolson scheme produces oscillations (which are even more sensitive for shorter simulation times). It is interesting to test shorter time intervals: with  $\Delta t = \epsilon \frac{\Delta x^2}{2\kappa}$  with  $0 < \epsilon < 1$  in these conditions the explicit  $\mathbb{P}_1$  scheme is stable (it is in fact the finite difference scheme), but the explicit  $\mathbb{P}_2$  schemes can continue to be unstable.



**Figure 3.7.** Numerical solutions (finite difference) for the heat equation for discontinuous data, not obeying the stability condition. For the color version of this figure, see [www.iste.co.uk/goudon/mathmodel.zip](http://www.iste.co.uk/goudon/mathmodel.zip)

Finally, we proceed to an analysis of convergence by numerical experimentation by returning to the periodic framework with the initial condition [3.9] for which we explicitly know the solution. Evidently with  $\mathbb{P}_1$  methods, we retrieve the results of Figures 3.3 and 3.4. The error curves for the implicit and Crank–Nicolson  $\mathbb{P}_2$  schemes are reproduced in Figure 3.10. We clearly observe an improvement of one order of convergence with the Crank–Nicolson scheme. We impose here a time step  $\Delta t = 4 \frac{\Delta x^2}{2k}$ . With the implicit scheme, the error behavior is dominated by the consistency error in time and is therefore proportional to  $\Delta x^2$ , while with the Crank–Nicolson scheme it remains dominated by the spatial error which behaves as  $\Delta x^3$ . However, for short times, the Crank–Nicolson scheme can produce oscillations and losses of positivity of the solution, a fault which can become intolerable with regard to the physical interpretation of the unknown quantity and its eventual use in a coupled problem.



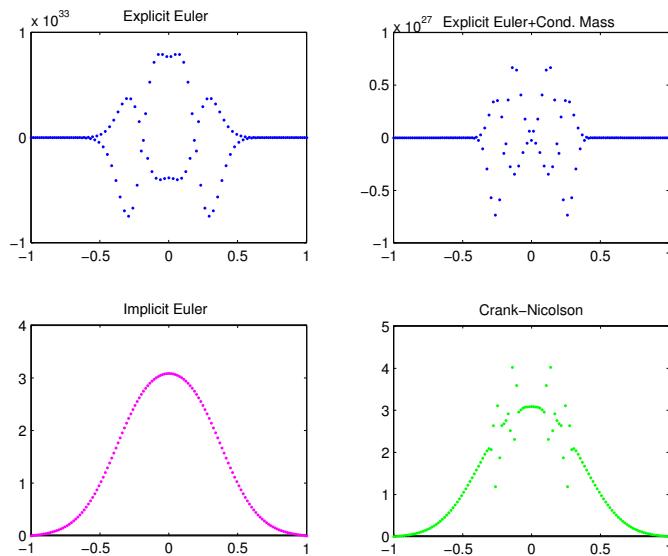
**Figure 3.8.** Numerical solutions ( $\mathbb{P}_1$  finite element) for the heat equation with discontinuous data, not obeying the DF stability condition. For the color version of this figure, see [www.iste.co.uk/goudon/mathmodel.zip](http://www.iste.co.uk/goudon/mathmodel.zip)

Let us finish with a few practical remarks about the implementation of these algorithms. To construct the stiffness and mass matrices, we follow the procedure explained in section 2.4.2, bearing in mind that the elemental mass matrix whose coefficients are defined by the integrals

$$\begin{aligned} \int_0^1 \psi_0(x)\psi_0(x) dx, & \quad \int_0^1 \psi_0(x)\psi_{1/2}(x) dx, & \quad \int_0^1 \psi_0(x)\psi_1(x) dx, \\ \int_0^1 \psi_{1/2}(x)\psi_0(x) dx, & \quad \int_0^1 \psi_{1/2}(x)\psi_{1/2}(x) dx, & \quad \int_0^1 \psi_{1/2}(x)\psi_1(x) dx, \\ \int_0^1 \psi_1(x)\psi_0(x) dx, & \quad \int_0^1 \psi_1(x)\psi_{1/2}(x) dx, & \quad \int_0^1 \psi_1(x)\psi_1(x) dx, \end{aligned}$$

is expressed as

$$N_{\text{loc}} = \frac{1}{30} \begin{pmatrix} 4 & 2 & -1 \\ 2 & 16 & 2 \\ -1 & 2 & 4 \end{pmatrix}.$$

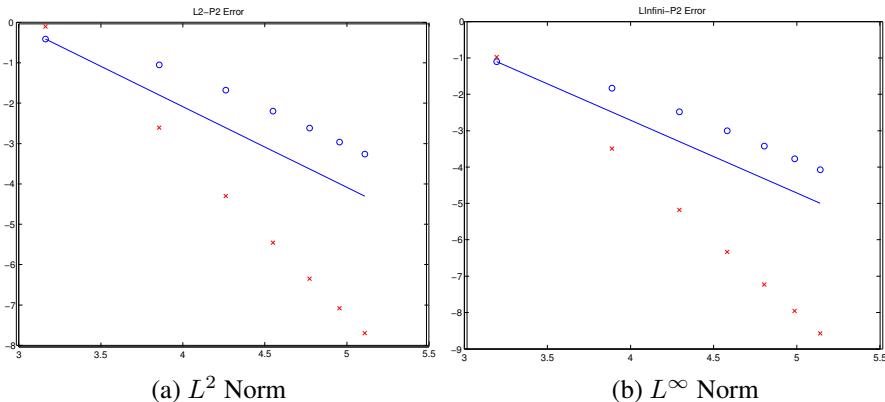


**Figure 3.9.** Numerical solutions ( $\mathbb{P}_2$  finite element) for the heat equation with discontinuous data, not obeying the DF stability condition. For the color version of this figure, see [www.iste.co.uk/goudon/mathmodel.zip](http://www.iste.co.uk/goudon/mathmodel.zip)

The scaling to obtain the mass matrix associated with the mesh of uniform step size  $\Delta x$  is achieved by multiplying by  $\Delta x$  (while the stiffness matrix is divided by  $\Delta x$ ). Although satisfying the Dirichlet conditions does not pose any difficulties, the construction of the matrices for the periodic problem is a bit more delicate. The (symmetric) structure of the  $\mathbb{P}_2$  matrices reproduces these periodic conditions, and we obtain matrices of the form

$$\left( \begin{array}{ccccccc} * & * & * & 0 & \dots & 0 & * \\ * & * & * & 0 & \dots & \dots & 0 \\ * & * & * & * & * & 0 & \dots \\ 0 & 0 & * & * & * & 0 & \dots \\ \vdots & \vdots & & & & & \vdots \\ 0 & 0 & \dots & 0 & * & * & * \\ 0 & 0 & \dots & 0 & * & * & * \\ * & 0 & \dots & 0 & * & * & * \\ * & 0 & \dots & 0 & * & * & * \end{array} \right)$$

(There are either 3 or 5 non-zero coefficients per row and per column; for the stiffness matrix, the sum of the coefficients for each row is zero. In this expression, the first coordinate corresponds to a node at abscissa 0, and the last coordinate to a node at abscissa  $2\pi - \Delta x/2$ .)



**Figure 3.10.** Error curves for the simulation of the heat equation by  $\mathbb{P}_2$  finite element schemes (implicit  $\circ$ , Crank–Nicolson  $\times$ , straight line of gradient 2)

### 3.2. From transport equations towards conservation laws

### 3.2.1. *Introduction*

Finite volume methods are especially suited for dealing with conservation laws because these equations correspond exactly to balances (of mass, energy, etc.), where the gains and losses in a domain arise from exchanges at the interfaces of the domain. Let us consider the example of passive particles immersed in a fluid: the particles (for example, a pollutant dispersed in a river) are described by their mass density  $u(t, x)$ . Thus, for all domain  $\Omega \subset \mathbb{R}^N$ , the integral  $\int_{\Omega} u(t, x) dx$  represents the mass contained at time  $t$  in the domain  $\Omega$ . The fluid is described by the velocity field  $(t, x) \mapsto a(t, x) \in \mathbb{R}^N$ , which we here take to be known. Any variations in mass are due to gains and losses across the boundary  $\partial\Omega$  and we obtain the following balance:

$$\frac{d}{dt} \int_{\Omega} u(t, x) dx = - \int_{\partial\Omega} u(t, x) a(t, x) \cdot \nu(x) d\sigma(x)$$

where  $\nu(x)$  represents the unit vector pointing outward  $\Omega$  at a point  $x \in \partial\Omega$  (such that  $a(t, x) \cdot \nu(x) > 0$  corresponds to a flux flowing from the inside to the outside of  $\Omega$ ,

hence a mass loss, which explains the negative sign in front of the right-hand side of the balance, see also note 3.3). We have used  $d\sigma$  to designate the Lebesgue measure over  $\partial\Omega$ . By integrating by parts, we deduce that

$$\int_{\Omega} (\partial_t u + \operatorname{div}_x(au))(t, x) dx = 0,$$

using  $\operatorname{div}_x$  to represent the differential operator defined over vector-valued functions

$$U : x \in \mathbb{R}^N \longmapsto U(x) = (U_1(x), \dots, U_N(x_N)) \in \mathbb{R}^N$$

by

$$\operatorname{div}_x(U)(x) = \sum_{j=1}^N \partial_{x_j} U_j(x).$$

This relation being satisfied for all domains  $\Omega$ , we thus obtain

$$\partial_t u + \operatorname{div}_x(au) = 0.$$

Provided that the velocity  $a$  is a “reasonably regular” given function, the analysis of this equation poses little difficulty. As we shall see later, it can be solved by the method of characteristics which shows that there exists a unique regular solution (let us say of class  $C^1$ , for  $C^1$  initial data). Considerable difficulties arise when the velocity depends upon the unknown  $u$  itself. We will study a variety of numerical approximation strategies for this equation. In particular, finite volume approaches correspond very closely to the principles that led to the equation, as a consequence of the balance of exchanges across the interfaces.

NOTE 3.3.– In dimension one, the domain  $\Omega$  is a segment  $[\alpha, \beta]$ , and the balance relation simply takes the form

$$\frac{d}{dt} \int_{\alpha}^{\beta} u(t, x) dx = -(a(t, \beta)u(t, \beta) - a(t, \alpha)u(t, \alpha)).$$

A velocity which is positive at  $\beta$  or negative at  $\alpha$  leads to mass loss.

Generally, we are from now on interested in a scalar quantity  $u : [0, T] \times \mathbb{R} \rightarrow \mathbb{R}$ , which satisfies the equation

$$\partial_t u + \partial_x f(u) = 0. \quad [3.11]$$

We will discuss later the properties of solutions following the nature of the flux  $u \mapsto f(u)$ , the linear case corresponding to  $f(u) = a \times u$ , for a given function  $a : [0, T] \times \mathbb{R} \rightarrow \mathbb{R}$ . Two classic nonlinear examples are:

- Burgers' equation where  $f(u) = u^2$ . This relates to a very simplified model of the equations of gas dynamics (but which preserve certain fundamental difficulties) introduced by [BAT 15, BUR 40].
- The Lighthill–Whitham–Richards equation for road traffic [LIG 55, RIC 56], where  $f(u) = u \times V$ , with

$$V = V_M(1 - u/u_M).$$

Here,  $u(t, x)$  represents a density of vehicles at time  $t$  and position  $x$  on a highway: the integral  $\int_a^b u(t, x) dx$  gives the number of vehicles present at time  $t$  over the section  $[a, b]$ . The function  $V$  describes the velocity of these vehicles: if the density  $u$  is low, the vehicles drive at a velocity close to the maximum speed  $V_M > 0$ , if the density  $u$  approaches the maximum density  $u_M > 0$ , the velocity tends to become zero<sup>1</sup>.

In these examples, we must not confuse  $V(u)$ , such as  $f(u) = u \times V(u)$ , and the derivative of the flux  $f'(u)$ . The formalism can also be adapted to apply to systems where  $u$  becomes an unknown vector quantity  $u : [0, T] \times \mathbb{R} \rightarrow \mathbb{R}^D$ . We will focus in particular on Euler equations for gas dynamics:

- Isentropic Euler equations  $u = (\rho, J) \in \mathbb{R}^2$ , with

$$f(u) = \begin{pmatrix} J \\ J^2/\rho + \kappa\rho^\gamma \end{pmatrix}$$

where  $\kappa > 0$  and  $\gamma > 1$  (or isothermal Euler equations if  $\gamma = 1$ ).

- Compressible Euler equations  $u = (\rho, J, \mathcal{E}) \in \mathbb{R}^3$ , with

$$f(u) = \begin{pmatrix} J \\ J^2/\rho + p(\rho, e) \\ (\mathcal{E} + p(\rho, e))J/\rho \end{pmatrix}$$

where  $p$  is a function with positive values, satisfying certain properties imposed by the physics of the situation, and  $e = \frac{\mathcal{E}}{\rho} - \frac{J^2}{2\rho^2}$  (the internal energy).

---

<sup>1</sup> This is a very simple model; a richer description in the form of a *system* of conservation laws has recently been proposed [AW 00] and has since become a very active area of research.

As mentioned above, the analysis of such nonlinear systems is extremely delicate and is still a very active area of research. We shall present slightly later some of the technical difficulties these problems present, and more on the subject can be found in [BEN 07, DAF 10, GOD 91, SER 96a, SER 96b]. In order to treat such equations numerically, it is important to properly understand the principles of discretization over linear models for which we can make use of explicit solutions that allow meaningful comparisons.

Finite volume discretization of these equations is based upon the following principles. We will use the following notations (refer back to Figure 2.6):

- $x_j$  represents a mesh node,
- $\mathcal{C}_j = [x_{j-1/2}, x_{j+1/2}]$  represents the control volume centered upon point  $x_j$ ,
- $h_j = x_{j+1/2} - x_{j-1/2}$  is the size of the control volume  $\mathcal{C}_j$ ,
- $h = \max\{h_j, j \in \mathbb{Z}\}$ ,  $\underline{h} = \min\{h_j | j \in \mathbb{Z}\}$ ,
- $\Delta t$  signifies the time step, which we will assume to be constant and denote  $t^n = n\Delta t$ .

We will focus on families of such meshes, parameterized by the step size  $h$ . In order to study the convergence properties of the approximations, we will assume that the elements of these families are such that  $\underline{h} = kh$ , for a given  $0 < k \leq 1$ , independent of  $h$  (such a mesh family is said to be *quasi-uniform*). The unknown value  $U_j^n$  is seen as an approximation of the average over the cell  $\mathcal{C}_j$  that is  $\frac{1}{h_j} \int_{\mathcal{C}_j} u(t^n, y) dy$ . By integrating equation [3.11] over the domain  $[t^n, t^{n+1}] \times \mathcal{C}_j$ , we obtain

$$\begin{aligned} \frac{1}{h_j} \int_{\mathcal{C}_j} u(t^{n+1}, y) dy - \frac{1}{h_j} \int_{\mathcal{C}_j} u(t^n, y) dy \\ + \frac{1}{h_j} \int_{t^n}^{t^{n+1}} (f(u(s, x_{j+1/2})) - f(u(s, x_{j-1/2}))) ds = 0. \end{aligned}$$

We draw inspiration from this formula to update the unknown values

$$\frac{1}{\Delta t} (U_j^{n+1} - U_j^n) + \frac{1}{h_j} (F_{j+1/2}^n - F_{j-1/2}^n) = 0 \quad [3.12]$$

where the *numerical flux*  $F_{j+1/2}^n$  is thus interpreted as an approximation of the physical flux between times  $t^n$  and  $t^{n+1}$  over the interface  $x_{j+1/2}$ , namely

$$\frac{1}{\Delta t} \int_{t^n}^{t^{n+1}} f(u(s, x_{j+1/2})) ds.$$

The design of the approximation schemes for [3.11] thus depends upon an adequate definition of the numerical fluxes  $F_{j+1/2}^n$ . To determine the flux at  $x_{j+1/2}$ , the most simple constructions involve only the adjacent cells, centered at  $x_j$  and  $x_{j+1}$ ; we thus obtain a three-point scheme to find the numerical flux in the form,

$$F_{j+1/2}^n = \mathbb{F}(U_{j+1}^n, U_j^n).$$

For example, the centered flux

$$F_{j+1/2}^n = \frac{1}{2}(f(U_{j+1}^n) + f(U_j^n))$$

at first seems to be applicable, since, over a uniform mesh of step size  $h$ , it leads to an approximation of the spatial derivative by

$$\frac{1}{h}(F_{j+1/2}^n - F_{j-1/2}^n) = \frac{f(U_{j+1}^n) - f(U_{j-1}^n)}{2h},$$

with a consistency error of order 2. However, we shall see that this scheme has a bad behavior and cannot be used.

### 3.2.2. Transport equation: method of characteristics

We start by addressing the linear case  $f(u) = a \times u$ , i.e.:

$$\partial_t u + \partial_x(au) = 0, \quad u(t=0, x) = u_{\text{Init}}(x). \quad [3.13]$$

with a given function  $a : \mathbb{R}^+ \times \mathbb{R} \rightarrow \mathbb{R}$ . We can explicitly solve this partial derivative equation by using arguments from the theory of ordinary differential equations. We introduce characteristic curves  $X(s; t, x)$ , solutions of the differential equation

$$\frac{d}{ds}X(s; t, x) = a(s, X(s; t, x)), \quad X(t; t, x) = x. \quad [3.14]$$

We interpret  $X(s; t, x)$  as the position occupied at time  $s$  by a particle having left position  $x$  at time  $t$ . It is important to bear this interpretation in mind to build an intuition on the behavior of the solutions. In particular, this makes clear that

$$X(t; s, X(s; r, x)) = X(t; r, x).$$

If  $u$  is a solution of [3.13], by applying the chain rule, we obtain

$$\begin{aligned}\frac{d}{ds} \left[ u(s, X(s; t, x)) \right] &= (\partial_t u)(s, X(s; t, x)) + \frac{d}{ds} X(s; t, x) \times (\partial_x u)(s, X(s; t, x)) \\ &= ((\partial_t + a \partial_x)u)(s, X(s; t, x)) = -(u \times \partial_x a)(s, X(s; t, x)).\end{aligned}$$

It then remains to integrate this simple linear equation, satisfied by  $s \mapsto u(s, X(s; t, x))$ , to obtain

$$u(t, x) = u_{\text{Init}}(X(0; t, x)) J(0; t, x), \quad [3.15]$$

with

$$J(s; t, x) = \exp \left( \int_t^s (\partial_x a)(\sigma, X(\sigma; t, x)) d\sigma \right). \quad [3.16]$$

In fact,  $J(s; t, x)$  is nothing other than the Jacobian of the change in variable  $y = X(s; t, x)$  (or equivalently  $x = X(t; s, y)$ ). By differentiating [3.14], we obtain that  $Y(s; t, x) = \partial_x X(s; t, x)$  satisfies

$$\frac{d}{ds} Y(s; t, x) = (\partial_x a)(s, X(s; t, x)) \times Y(s; t, x), \quad Y(t; t, x) = 1.$$

Again, we have a simple linear differential equation whose solution is

$$Y(s; t, x) = \exp \left( \int_t^s (\partial_x a)(\sigma, X(\sigma; t, x)) d\sigma \right).$$

We thus have  $J(s; t, x) = \partial_x X(s; t, x)$  and  $dy = J(s; t, x) dx$ . In particular, we deduce from this that [3.13] is indeed a conservation equation since the change in variable  $y = X(0; t, x)$ ,  $dy = J(0; t, x) dx$  gives

$$\int_{\mathbb{R}} u(t, x) dx = \int_{\mathbb{R}} u_{\text{Init}}(X(0; t, x)) J(0; t, x) dx = \int_{\mathbb{R}} u_{\text{Init}}(y) dy.$$

This treatment assumes that the characteristic curves  $X(s; t, x)$  are well-defined, i.e. that we can apply the Cauchy–Lipschitz theorem to solve [3.14]. For example, we assume that

h1)  $a$  is of class  $C^1$ ,

h2) there exists  $C > 0$ , such that for all  $t, x \in \mathbb{R}$ , we have  $|a(t, x)| \leq C(1 + |x|)$ .

In the particular case where  $a(t, x) = c \in \mathbb{R}$  is constant, we simply have  $X(s; t, x) = x + c(s - t)$  and  $J(s; t, x) = 1$ .

**THEOREM 3.3.**— Under the hypotheses h1)-h2), for any initial data  $u_{\text{Init}} \in C^1$ , problem [3.13] has a unique solution  $u$  of class  $C^1$  defined by [3.15]–[3.16]. If  $u_{\text{Init}} \in L^1(\mathbb{R})$ , then  $u$  is integrable and has the same integral as  $u_{\text{Init}}$ . If  $\partial_x a \in L^\infty$ , and  $0 \leq u_{\text{Init}}(x) \leq M$  for all  $x \in \mathbb{R}$  then, for all  $0 \leq t \leq T < \infty$ ,  $x \in \mathbb{R}$ , we have  $0 \leq u(t, x) \leq M e^{\bar{T} \|\partial_x a\|_\infty}$ .

It is interesting to note that formula [3.15]–[3.16] retains its meaning when  $u_{\text{Init}}$  is a function of  $L^p(\mathbb{R})$ , not necessarily continuous. Hence, we can naturally extend the notion of solutions to this context.

**THEOREM 3.4.**— Let  $u_{\text{Init}} \in L^1 \cap L^\infty(\mathbb{R})$ . Then,  $(t, x) \mapsto u(t, x)$  defined by [3.15]–[3.16] is the unique function satisfying

$$\lim_{t \rightarrow t_0} \int_{\mathbb{R}} |u(t, x) - u(t_0, x)| dx = 0 \quad [3.17]$$

and  $u$  is a weak solution of [3.13] in the sense that for all functions  $\phi \in C_c^1([0, \infty[\times\mathbb{R})$ , we have

$$-\int_0^\infty \int_{\mathbb{R}} u(t, x) (\partial_t \phi(t, x) + a(t, x) \partial_x \phi(t, x)) dx dt - \int_{\mathbb{R}} \rho_{\text{Init}}(x) \phi(0, x) dx = 0.$$

**PROOF.**— We start by establishing a continuity property of [3.13] with respect to the initial data. By noting  $u^i(t, x) = u_{\text{Init}}^i(X(0; t, x)) J(0; t, x)$  for  $i \in \{1, 2\}$ , we have

$$\begin{aligned} \int_{\mathbb{R}} |u^1(t, x) - u^2(t, x)| dx &= \int_{\mathbb{R}} |u_{\text{Init}}^1(X(0; t, x)) - u_{\text{Init}}^2(X(0; t, x))| J(0; t, x) dx \\ &= \int_{\mathbb{R}} |u_{\text{Init}}^1(y) - u_{\text{Init}}^2(y)| dy. \end{aligned}$$

As a consequence, it will be sufficient to prove [3.17] for continuous data with compact support,  $u_{\text{Init}}$ , to then extend it to integrable functions by a density argument (see [GOU 11, theorem 4.13]).

The demonstration then makes use of the continuity properties of the characteristic curves. The general theory of differential equations ensures that the mapping  $(t, s, x) \mapsto X(t; s, x)$  is of class  $C^1$ . We also note that when the variables  $t, s, x$  describe a compact set, the trajectories  $X(t; s, x)$  remain in a compact set of  $\mathbb{R}$ .

The continuity in time for values in  $L^1$  is demonstrated by a change in variables and by invoking the Lebesgue theorem. Indeed, we write the following inequalities:

$$\begin{aligned}
 & \int_{\mathbb{R}} |u(t, x) - u(t_0, x)| dx \\
 &= \int_{\mathbb{R}} |u_{\text{Init}}(X(0; t, x))J(0; t, x) - u_{\text{Init}}(X(0; t_0, x))J(0; t_0, x)| dx \\
 &\leq \int_{\mathbb{R}} |u_{\text{Init}}(X(0; t, x)) - u_{\text{Init}}(X(0; t_0, x))| |J(0; t, x)| dx \\
 &\quad + \int_{\mathbb{R}} |u_{\text{Init}}(X(0; t_0, x))| \left| \frac{J(0; t, x)}{J(0; t_0, x)} - 1 \right| |J(0; t_0, x)| dx \tag{3.18} \\
 &\leq \int_{\mathbb{R}} |u_{\text{Init}}(y) - u_{\text{Init}}(X(0; t_0, X(t; 0, y)))| dy \\
 &\quad + \int_{\mathbb{R}} |u_{\text{Init}}(y)| \left| \frac{J(0; t, X(t_0; 0, y))}{J(0; t_0, X(t_0; 0, y))} - 1 \right| dy.
 \end{aligned}$$

We thus recall that, for fixed  $t_0$  and  $y$ , on the one hand, we have

$$\lim_{t \rightarrow t_0} X(0; t_0, X(t; 0, y)) = X(0; t_0, X(t_0; 0, y)) = X(0; 0, y) = y$$

and, on the other hand, we can establish that

$$\begin{aligned}
 \lim_{t \rightarrow t_0} J(0; t, X(t_0; 0, y)) &= \lim_{t \rightarrow t_0} \exp \left( \int_t^0 (\partial_x a)(\sigma, X(\sigma; t, X(t_0; 0, y))) d\sigma \right) \\
 &= \exp \left( \int_{t_0}^0 (\partial_x a)(\sigma, X(\sigma; t_0, X(t_0; 0, y))) d\sigma \right) \\
 &= J(0; t_0, X(t_0; 0, y)) \\
 &= \exp \left( \int_{t_0}^0 (\partial_x a)(\sigma, X(\sigma; 0, y)) d\sigma \right) = \frac{1}{J(t_0; 0, y)}.
 \end{aligned}$$

Thus, the integrands in [3.18] do indeed tend towards 0 as  $t \rightarrow t_0$ . Furthermore, while  $t_0 \in \mathbb{R}$  is fixed, when  $t$  varies within, for example,  $[t_0 - 1, t_0 + 1]$  and  $y$  within a compact  $K$  of  $\mathbb{R}$  (as it happens, the support of  $u_{\text{Init}}$ ), the quantity  $\frac{J(0; t, X(t_0; 0, y))}{J(0; t_0, X(t_0; 0, y))} = J(0; t, X(t_0; 0, y))J(t_0; 0, y)$  is bounded by a constant (which depends upon the compact  $K$ ). Finally,  $\{y \in \mathbb{R}, X(0; t_0, X(t; 0, y)) \in K, t_0 - 1 \leq t \leq t_0 + 1\} = \{y \in \mathbb{R}, X(t; 0, y) \in X(t_0; 0, K), t_0 - 1 \leq t \leq t_0 + 1\} = \{X(0; t, X(t_0, 0, K)), t_0 - 1 \leq t \leq t_0 + 1\}$  is again a compact set of  $\mathbb{R}$ . These remarks ensure the domination required to be able to conclude with the Lebesgue

theorem and we deduce from this that [3.17] is satisfied for continuous  $u_{\text{Init}}$  with compact support, and hence finally for  $u_{\text{Init}} \in L^1(\mathbb{R})$ .

We then show that  $u$  is a weak solution by calculating

$$\begin{aligned} & \int_0^\infty \int_{\mathbb{R}} u(t, x) (\partial_t \phi(t, x) + a(t, x) \partial_x \phi(t, x)) dx dt \\ &= \int_0^\infty \int_{\mathbb{R}} u_{\text{Init}}(X(0; t, x)) (\partial_t \phi(t, x) + a(t, x) \partial_x \phi(t, x)) J(0; t, x) dx dt \\ &= \int_0^\infty \int_{\mathbb{R}} u_{\text{Init}}(y) (\partial_t \phi(t, X(t; 0, y)) + a(t, X(t; 0, y)) \partial_x \phi(t, X(t; 0, y))) dy dt \\ &= \int_{\mathbb{R}} u_{\text{Init}}(y) \left( \int_0^\infty \frac{d}{dt} [\phi(t, X(t; 0, y))] dt \right) dy \\ &= \int_{\mathbb{R}} u_{\text{Init}}(y) \phi(0, y) dy. \end{aligned}$$

To demonstrate the uniqueness, we must prove that if  $u$  satisfies the weak formulation with  $u_{\text{Init}} = 0$  then  $u$  is null almost everywhere. This conclusion is based upon a duality argument, sometimes referred to as the H\"olmgren method. Let  $\psi \in C_c^\infty((0, +\infty) \times \mathbb{R})$ , such that  $\psi(t, \cdot) = 0$  for  $t \geq T$ . We solve  $\partial_t \phi + a \partial_x \phi = \psi$  with the final data  $\phi|_{t=T} = 0$ . More precisely, by integrating along the characteristics, we obtain

$$\phi(t, x) = \int_T^t \psi(\sigma, X(\sigma; t, x)) d\sigma \in C_c^1([0, +\infty) \times \mathbb{R}).$$

We can then use this test function in the weak form, which leads to

$$\int_0^\infty \int_{\mathbb{R}} u(t, x) \psi(t, x) dx dt = 0.$$

With this having been satisfied for all  $\psi \in C_c^\infty((0, +\infty) \times \mathbb{R})$ , we conclude from this that  $u(t, x) = 0$  almost everywhere.  $\square$

### 3.2.3. Upwinding principles: upwind scheme

As we had mentioned previously, the construction of a numerical scheme for [3.13] seeks to “copy” the mass balance over the control volumes. The difficulty therefore lies in defining an appropriate approximation of the flux over the interface  $x_{j+1/2}$ . The upwind flux will look for the information where it comes from, with regard to the

equation: from the left if  $a(t, x) \geq 0$ , and from the right if  $a(t, x) \leq 0$ . The scheme is thus adapted to the direction of information propagation by setting

$$F_{j+1/2}^n = \begin{cases} a(t^n, x_{j+1/2}) U_{j+1}^n & \text{if } a(t^n, x_{j+1/2}) \leq 0, \\ a(t^n, x_{j+1/2}) U_j^n & \text{if } a(t^n, x_{j+1/2}) \geq 0, \end{cases} \quad [3.19]$$

LEMMA 3.4.– The upwind scheme [3.19] preserves the positivity under the CFL condition  $\|a\|_\infty \frac{\Delta t}{h} \leq 1/2$ : if  $u_j^0 \geq 0$  for all  $j$ , then  $u_j^n \geq 0$  for all  $n, j$ .

PROOF.– By linearity, it suffices to prove that  $u_j^{n+1}$  remains positive when  $u_{j-1}^n, u_j^n, u_{j+1}^n$  are positive. We set  $a_{j+1/2}^n = a(t^n, x_{j+1/2})$ . We recall the notation

$$[a]_+ = \max(a, 0) = \frac{a + |a|}{2}, \quad [a]_- = \min(a, 0) = \frac{a - |a|}{2},$$

such that  $a = [a]_+ + [a]_-$  and  $|a| = [a]_+ - [a]_-$ . The scheme is expressed as

$$\begin{aligned} u_j^{n+1} &= u_j^n - \frac{\Delta t}{h_j} \left( [a_{j+1/2}^n]_+ u_j^n + [a_{j+1/2}^n]_- u_{j+1}^n - [a_{j-1/2}^n]_+ u_{j-1}^n - [a_{j-1/2}^n]_- u_j^n \right) \\ &= u_j^n \left( 1 - \frac{\Delta t}{h_j} [a_{j+1/2}^n]_+ + \frac{\Delta t}{h_j} [a_{j-1/2}^n]_- \right) \\ &\quad - \frac{h_j}{h_j} [a_{j+1/2}^n]_- u_{j+1}^n + \frac{\Delta t}{h_j} [a_{j-1/2}^n]_+ u_{j-1}^n. \end{aligned}$$

The terms involving  $u_{j-1}^n$  and  $u_{j+1}^n$  contribute positively. Moreover, we have

$$0 \leq [a_{j+1/2}^n]_+ - [a_{j-1/2}^n]_- \leq 2\|a\|_\infty.$$

Thus, the CFL constraint ensures that

$$\left( 1 - \frac{\Delta t}{h_j} [a_{j+1/2}^n]_+ + \frac{\Delta t}{h_j} [a_{j-1/2}^n]_- \right) \geq 0,$$

which thus implies that  $u_j^{n+1} \geq 0$ . □

NOTE 3.4.– The proof shows that, *in the specific case where the velocity  $a$  has constant sign*, the upwind scheme is  $L^\infty$ -stable under CFL  $\|a\|_\infty \frac{\Delta t}{h} \leq 1$ . In this case, we show in fact that if  $m \leq u_j^0 \leq M$  for all  $j$ , then  $m \leq u_j^n \leq M$  for all  $n, j$ , as for the continuous problem. When the sign of  $a$  is variable, the condition that guarantees the preservation of positivity is slightly more restrictive.

### 3.2.4. Linear transport at constant speed: analysis of FD and FV schemes

In this section, we describe the particular case where the velocity  $a(t, x) = c \in \mathbb{R}$  is constant. The solution of [3.13] is thus simply

$$u(t, x) = u_{\text{Init}}(x - ct).$$

Of course, with the solution being explicitly known, the numerical simulation of this problem is of little interest in itself. However, the detailed analysis of this simple case and the direct comparison between numerical solutions and the exact solution uncovers certain guiding principles which will govern the design of numerical methods for more complex problems.

From a finite difference perspective, the unknown value  $\bar{u}_j^n$  is interpreted as an approximation of  $u(n\Delta t, x_j)$ . It seems quite natural, to start with, to use an Euler scheme with centered discretization for the spatial variable:

$$\frac{\bar{u}_j^{n+1} - \bar{u}_j^n}{\Delta t} = -c \frac{\bar{u}_{j+1}^n - \bar{u}_{j-1}^n}{x_{j+1} - x_{j-1}}. \quad [3.20]$$

In this context, the FD-upwind scheme takes the form

$$\frac{\bar{u}_j^{n+1} - \bar{u}_j^n}{\Delta t} = \begin{cases} -c \frac{\bar{u}_j^n - \bar{u}_{j-1}^n}{h_{j-1/2}} & \text{if } c > 0, \\ -c \frac{\bar{u}_{j+1}^n - \bar{u}_j^n}{h_{j+1/2}} & \text{if } c < 0. \end{cases} \quad [3.21]$$

We will note the difference with the FV-upwind scheme, which uses the same formulae but where  $h_j = x_{j+1/2} - x_{j-1/2}$  replaces  $h_{j+1/2} = x_{j+1} - x_j$  (see [3.12]). The two schemes match, however, in the case of a uniform mesh of constant mesh size  $h$ . The first element of analysis in these finite difference schemes lies in the concept of consistency, which generalizes the definition introduced in lemma 3.1.

**DEFINITION 3.3.–** We say that a scheme

$$u_j^{n+1} = \Phi(\Delta t, u_{j-k}^n, \dots, u_{j+k}^n, h_{j-k+1/2}, \dots, h_{j+k+1/2})$$

is *consistent (in the sense of finite differences)* with an equation (E), if, given  $u: (t, x) \mapsto u(t, x)$  a regular solution of (E), then the sequence defined by

$$\epsilon_j^n = \frac{u(t^{n+1}, x_j) - \Phi(\Delta t, u(t^n, x_{j-k}), \dots, u(t^n, x_{j+k}), h_{j-k+1/2}, \dots, h_{j+k+1/2})}{\Delta t}$$

satisfies

$$\lim_{\Delta t, h \rightarrow 0} \sup_{0 \leq t^n \leq T} \sup_j |\epsilon_j^n| = 0.$$

NOTE 3.5.– We define the concept of consistency in the  $L^p$  norm in the same manner by replacing  $\sup_j |\epsilon_j^n|$  by  $\sum_j h_j |\epsilon_j^n|^p$ .

In what follows, we should be aware that the consistency analysis to be carried out makes use of the regularity of the exact solution  $u$  of the problem (which we do not always know). The centered scheme thus seems attractive, since, over a uniform mesh, it is consistent of order 2 in space and of order 1 in time. Indeed, with  $h_{j+1/2} = x_{j+1} - x_j$ , we obtain

$$\begin{aligned} \epsilon_j^n &= \frac{u(t^{n+1}, x_j) - u(t^n, x_j)}{\Delta t} + c \frac{1}{x_{j+1} - x_{j-1}} (u(t^n, x_{j+1}) - u(t^n, x_{j-1})) \\ &= \frac{1}{\Delta t} \left( \partial_t u(t^n, x_j) \Delta t + \partial_{tt}^2 u(t^n, x_j) \frac{\Delta t^2}{2} \right) \\ &\quad + c \frac{1}{x_{j+1} - x_{j-1}} \left( \partial_x u(t^n, x_j) (h_{j+1/2} + h_{j-1/2}) \right. \\ &\quad \left. + \partial_{xx}^2 u(t^n, x_j) \frac{h_{j+1/2}^2 - h_{j-1/2}^2}{2} + \partial_{xxx}^3 u(t^n, x_j) \frac{h_{j+1/2}^3 + h_{j-1/2}^3}{6} \right) + R_j^n \quad [3.22] \\ &= \partial_{tt}^2 u(t^n, x_j) \frac{\Delta t}{2} + c \partial_{xx}^2 u(t^n, x_j) \frac{h_{j+1/2}^2 - h_{j-1/2}^2}{2(h_{j+1/2} + h_{j-1/2})} \\ &\quad + c \partial_{xxx}^3 u(t^n, x_j) \frac{h_{j+1/2}^3 + h_{j-1/2}^3}{6(h_{j+1/2} + h_{j-1/2})} + R_j^n \end{aligned}$$

where the final term can be estimated by making use of the regularity of  $u$ : for example, we have

$$|R_j^n| \leq \|c \partial_{xxx}^4 u\|_\infty h^3 + \|\partial_{ttt}^3 u\|_\infty \Delta t^2.$$

When the mesh is uniform  $h_{j+1/2}^2 - h_{j-1/2}^2 = 0$  and, in this case, we deduce from this that

for the centered scheme [3.20]  $\sup_j |\epsilon_j^n| \leq C(h^2 + \Delta t)$ .

We conduct a similar analysis for the upwind scheme, for example, for  $c > 0$

$$\begin{aligned}\epsilon_j^n &= \frac{u(t^{n+1}, x_j) - u(t^n, x_j)}{\Delta t} + c \frac{1}{x_j - x_{j-1}} (u(t^n, x_j) - u(t^n, x_{j-1})) \\ &= \frac{1}{\Delta t} \left( \partial_t u(t^n, x_j) \Delta t + \partial_{tt}^2 u(t^n, x_j) \frac{\Delta t^2}{2} \right) \\ &\quad + c \frac{1}{h_{j-1/2}} \left( \partial_x u(t^n, x_j) h_{j-1/2} + \partial_{xx}^2 u(t^n, x_j) \frac{h_{j-1/2}^2}{2} \right) + R_j^n \\ &= \partial_{tt}^2 u(t^n, x_j) \frac{\Delta t}{2} + c \partial_{xx}^2 u(t^n, x_j) \frac{h_{j-1/2}}{2} + R_j^n.\end{aligned}\tag{3.23}$$

We deduce from this the following estimate of consistency of order 1:

$$\text{for the upwind scheme [3.21]} \sup_j |\epsilon_j^n| \leq C(h + \Delta t).$$

In these estimates, the constant  $C$  depends upon the norm of  $u$  in  $C^k$  for a suitable  $k$ . However, the centered scheme is hindered by a stability fault, which is immediately noticeable in the simulations.

**THEOREM 3.5.**— The centered scheme [3.20] is neither  $L^\infty$ -stable, nor  $L^2$ -stable.

**PROOF.**— We only consider the situation where the mesh size is constant  $h_j = h$  and we study the case of periodic boundary conditions. We use the same reasoning as for the heat equation to define the  $L^2$  stability: the index  $j$  describes  $\{0, \dots, J\}$  and the scheme is completed by the periodicity relations. The sequence  $u_j^n$  defined by the scheme is then associated with the sequence of functions defined over  $[0, 2\pi]$  by

$$x \mapsto u^n(x) = \sum_{j=0}^J u_j^n \mathbf{1}_{jh \leq x < (j+1)h}(x)$$

and that we extend over  $\mathbb{R}$  by  $2\pi$ -periodicity. We use  $L^2_\#$  to represent the set of functions  $g : \mathbb{R} \rightarrow \mathbb{R}$  which are  $2\pi$ -periodic, and such that

$$\|g\|_{L^2_\#}^2 = \int_0^{2\pi} |g(y)|^2 dy < \infty.$$

The  $L^2$ -stability means that for all  $n \in \mathbb{N}$ , we have

$$\|u^n\|_{L^2_\#} \leq \|u^0\|_{L^2_\#}.$$

For scheme [3.20], we have

$$\begin{aligned}
 2\pi\widehat{\bar{u}^{n+1}}(k) &= \sum_{j=0}^J \bar{u}_j^{n+1} e^{-i(j+1/2)kh} \frac{2\sin(kh/2)}{k} \\
 &= \sum_{j=0}^J \left( \bar{u}_j^n - c \frac{\Delta t}{2h} (\bar{u}_{j+1}^n - \bar{u}_{j-1}^n) \right) e^{-i(j+1/2)kh} \frac{2\sin(kh/2)}{k} \\
 &= \sum_{j=0}^J \bar{u}_j^n \left( 1 - c \frac{\Delta t}{2h} (e^{-ikh} - e^{-ikh}) \right) e^{-i(j+1/2)kh} \frac{2\sin(kh/2)}{k} \\
 &= \sum_{j=0}^J \bar{u}_j^n \left( 1 - 2ic \frac{\Delta t}{2h} \sin(kh) \right) e^{-i(j+1/2)kh} \frac{2\sin(kh/2)}{k}.
 \end{aligned}$$

This calculation brings out the amplification factor

$$M(k) = \left( 1 - ic \frac{\Delta t}{h} \sin(kh) \right).$$

Its modulus satisfies

$$|M(k)| = 1 + c^2 \frac{\Delta t^2}{h^2} \sin^2(kh) \geq 1.$$

The scheme is therefore always unstable in the  $L^2$  norm. In the  $L^\infty$  norm, it is enough to observe that, for a step function which jumps from 1 to 0 at the node  $j_0$  as initial data, the scheme yields values strictly greater than 1 from the first step in time. Actually, we find  $u_{j_0}^1 = u_{j_0}^0 - \frac{\Delta t}{2\Delta x} (u_{j_0+1}^0 - u_{j_0-1}^0) = 1 - \frac{\Delta t}{2\Delta x} (0 - 1) = 1 + \frac{\Delta t}{2\Delta x} > 1$ . Figure 3.12 illustrates how this stability fault materializes.  $\square$

**THEOREM 3.6.–** Over a uniform mesh of step size  $h > 0$ , the upwind scheme [3.21] is  $L^\infty$ -stable and  $L^2$ -stable under the condition  $|c|\Delta t/h \leq 1$ .

**PROOF.–** The amplification factor of scheme [3.21] is expressed as

$$\begin{aligned}
 \text{for } c > 0, \quad M(k) &= 1 + 2ic \frac{\Delta t}{h} e^{ikh/2} \sin(kh/2) \\
 &= 1 - 2c \frac{\Delta t}{h} \sin^2(kh/2) + 2ic \frac{\Delta t}{h} \cos(kh/2) \sin(kh/2), \\
 \text{for } c < 0, \quad M(k) &= 1 - 2i|c| \frac{\Delta t}{h} e^{-ikh/2} \sin(kh/2) \\
 &= 1 - 2|c| \frac{\Delta t}{h} \sin^2(kh/2) - 2i|c| \frac{\Delta t}{h} \cos(kh/2) \sin(kh/2),
 \end{aligned}$$

It follows that

$$\begin{aligned} |M(k)|^2 &= 1 + 4c^2 \frac{\Delta t^2}{h^2} \sin^4(kh/2) \\ &\quad - 4|c| \frac{\Delta t}{h} \sin^2(kh/2) + 4c^2 \frac{\Delta t^2}{h^2} \sin^2(kh/2)(1 - \sin^2(kh/2)) \\ &= 1 + 4|c|\Delta t h \sin^2(kh/2) \left( |c| \frac{\Delta t}{h} - 1 \right). \end{aligned}$$

The CFL condition thus ensures  $|M(k)| \leq 1$ .

The  $L^\infty$  stability follows on from the fact that the iterated  $\bar{u}_j^{n+1}$  is written, when the CFL condition is satisfied, as a convex combination of  $\bar{u}_j^n, \bar{u}_{j-1}^n, \bar{u}_{j+1}^n$ . (If the meshes are quasi-uniform but not uniform, the condition for  $L^\infty$ -stability becomes  $\frac{|c|\Delta t}{kh} \leq 1$ .)  $\square$

As a consequence of these properties, we can establish the convergence of the upwind scheme for linear transport at a constant speed.

**THEOREM 3.7.**— Let  $u_{\text{Init}} \in C_\#^2([0, 2\pi])$ . Let  $u$  be the associated solution of [3.13], with a constant velocity  $a(t, x) = c$ . Finally, let  $0 < T < \infty$ . Then, the sequence  $\bar{u}_j^n$  defined by [3.21], for a family of quasi-uniform meshes such that the condition  $\frac{|c|\Delta t}{kh} \leq 1$  holds, satisfies

$$\max_{0 \leq t^n \leq T} \sup_j |\bar{u}_j^n - u(t^n, x_j)| \leq CT h$$

where  $C$  depends on  $\|u\|_{C^2}$ .

**PROOF.**— We set  $e_j^n = \bar{u}_j^n - u(t^n, x_j)$ . We only treat the case  $c > 0$ , the demonstration being identical for negative velocities. We deduce from the recurrence relation

$$\frac{1}{\Delta t}(e_j^{n+1} - e_j^n) + \frac{c}{h_{j-1/2}}(e_j^n - e_{j-1}^n) = -\epsilon_j^{n+1}$$

that

$$|e_j^{n+1}| \leq |e_j^n| \left( 1 - \frac{c\Delta t}{h_{j-1/2}} \right) + \frac{c\Delta t}{h_{j-1/2}} |e_{j-1}^n| + \Delta t |\epsilon_j^{n+1}| \leq \sup_k |e_k^n| + C\Delta t h.$$

To obtain this inequality, we have used both the CFL condition and the consistency estimate. An argument by recurrence thus leads to the stated assertion.  $\square$

The consistency analysis [3.22] of the centered scheme reveals the origin of the instability and offers ideas for how to remedy it. Indeed, we can rewrite [3.22] in the following manner:

$$\begin{aligned}\epsilon_j^n &= \frac{u(t^{n+1}, x_j) - u(t^n, x_j)}{\Delta t} - c \frac{1}{2h} (u(t^n, x_j + h) - u(t^n, x_j - h)) \\ &= \left( \partial_t u + c \partial_x u + \frac{\Delta t}{2} \partial_{tt}^2 u \right)(t^n, x_j) + \tilde{R}_j^n(\Delta t, h) \\ &= \left( \partial_t u + c \partial_x u + c^2 \frac{\Delta t}{2} \partial_{xx}^2 u \right)(t^n, x_j) + \tilde{R}_j^n(\Delta t, h)\end{aligned}\quad [3.24]$$

since  $\partial_t u = -c \partial_x u$  and thus  $\partial_{tt}^2 u = -c \partial_x \partial_t u = c^2 \partial_{xx}^2 u$ . In this relationship, the remaining term is evaluated as  $|\tilde{R}_j^n(\Delta t, h)| \leq C(\Delta t^2 + h^2)$ , where  $C$  depends upon the norm of  $u$  in  $C^4$ . This demonstrates that scheme [3.20] defines a consistent approximation of order 2 in time and space of the PDE

$$\partial_t v + c \partial_x v + c^2 \frac{\Delta t}{2} \partial_{xx}^2 v = 0.$$

This is an “anti-diffusion” equation which is inherently unstable<sup>2</sup>. For the upwind scheme, the same analysis leads to

$$\begin{aligned}\epsilon_j^n &= \left( \partial_t u + c \partial_x u + \frac{\Delta t}{2} \underbrace{\partial_{tt}^2 u}_{=c^2 \partial_{xx}^2 u} - \frac{ch}{2} \partial_{xx}^2 u \right)(t^n, x_j) + \tilde{R}_j^n(\Delta t, h) \\ &= \left( \partial_t u + c \partial_x u + \frac{c}{2}(c\Delta t - h) \partial_{xx}^2 u \right)(t^n, x_j) + \tilde{R}_j^n(\Delta t, h).\end{aligned}$$

Under the CFL condition, the coefficient  $\delta = \frac{c}{2}(h - c\Delta t)$  is indeed positive and the scheme is consistent of order 2 in time and space with the diffusion equation

$$\partial_t v + c \partial_x v - \delta \partial_{xx}^2 v = 0.$$

In this type of argument, we can look to introduce an artificial diffusion term (in this context, we are talking about “artificial viscosity”) so as to compensate the anti-diffusive character of scheme [3.20]. We assume that the condition  $|c|\Delta t/h < 1$  is achieved. We thus choose  $\epsilon > 0$  depending on the discretization parameters, so that  $c^2 \frac{\Delta t}{2} = (c\Delta t/h)^2 \frac{h^2}{2\Delta t} \leq \frac{h^2}{2\Delta t} \leq \epsilon$  and we reach the modified equation

$$\partial_t \tilde{v} + c \partial_x \tilde{v} - \left( \epsilon - c^2 \frac{\Delta t}{2} \right) \partial_{xx}^2 \tilde{v} = 0.$$

---

<sup>2</sup> We can convince ourselves of this by examining the Fourier transform of the solution that satisfies  $\partial_t \hat{v}(t, \xi) = (ic\xi + k\xi^2) \hat{v}(t, \xi)$ , with  $k = c^2 \frac{\Delta t}{2}$ . It follows that  $\hat{v}(t, \xi) = \exp(t(ic\xi + k\xi^2)) \hat{v}(0, \xi)$ . As  $k > 0$ , we deduce from this that  $\|\hat{v}(t, \cdot)\|_{L^2} = \|u(t, \cdot)\|_{L^2} \rightarrow \infty$  as  $t \rightarrow \infty$ , with an exponentially fast growth.

With  $\epsilon = \frac{h^2}{2\Delta t}$ , we obtain the following scheme

$$\frac{\bar{u}_j^{n+1} - \bar{u}_j^n}{\Delta t} = -\frac{c}{2h}(\bar{u}_{j+1}^n - \bar{u}_{j-1}^n) + \frac{1}{2}(\bar{u}_{j+1}^n - 2\bar{u}_j^n + \bar{u}_{j-1}^n).$$

This is the Lax–Friedrichs scheme. We can verify that the scheme is stable and consistent of order 1.

In order to properly understand the influence of the “numerical diffusion” term, we can focus on the convection–diffusion equation at a constant velocity  $c > 0$ :

$$\partial_t u_\delta + c \partial_x u_\delta = \delta \partial_{xx}^2 u_\delta, \quad [3.25]$$

where  $0 < \delta \ll 1$ . In fact, we note that  $v(t, x) = u_\delta(t/\delta, x + ct/\delta)$  simply satisfies  $\partial_t v = \partial_{xx}^2 v$ , an equation for which we know the expression of the solution for fixed initial data. Thus, if  $u_\delta|_{t=0} = u_{\text{Init}}$ , we obtain

$$u_\delta(t, x) = \int_{\mathbb{R}} u_{\text{Init}}(y) \exp\left(-\frac{|x - ct - y|^2}{4\delta t}\right) \frac{dy}{\sqrt{4\pi\delta t}}.$$

Notably, even though the initial data are discontinuous, the solution  $u_\delta$  is  $C^\infty$  for all times  $t > 0$ . We recover the regularizing effect of the heat equation. Over numerical simulations, we shall indeed see the profile of the approximate solution regularizing for schemes that induce such a diffusion effect (see Figures 3.13 and 3.14). Nevertheless, we observe that

$$\begin{aligned} & \int_{\mathbb{R}} |u_\delta(t, x) - u_{\text{Init}}(x - ct)| dx \\ &= \int_{\mathbb{R}} \left| \int_{\mathbb{R}} e^{-|x-ct-y|^2/(4\delta t)} (u_{\text{Init}}(y) - u_{\text{Init}}(x - ct)) \frac{dy}{\sqrt{4\pi\delta t}} \right| dx \\ &\leq \int_{\mathbb{R}} \int_{\mathbb{R}} |u_{\text{Init}}(y) - u_{\text{Init}}(x - ct)| e^{-|x-ct-y|^2/(4\delta t)} \frac{dy}{\sqrt{4\pi\delta t}} dx \\ &\leq \int_{\mathbb{R}} \int_{\mathbb{R}} |u_{\text{Init}}(X - \sqrt{4\delta t}Y) - u_{\text{Init}}(X)| e^{-Y^2} \frac{dY}{\sqrt{\pi}} dX \end{aligned}$$

with the change in variables  $X = x - ct$  and  $Y = (x - ct - y)/\sqrt{4\delta t}$ . The continuity of translations in  $L^1$  and the Lebesgue theorem allow us to conclude that, for fixed  $t$  when  $\delta$  tends towards 0,  $u_\delta(t, \cdot)$  converges in  $L^1(\mathbb{R})$  towards  $u(t, \cdot)$ , the solution for the transport equation at velocity  $c$  and for the initial data  $u_{\text{Init}}$ .

We can go slightly further by choosing  $\epsilon$  so as to cancel the diffusion term. As a consequence, we cancel the terms of order 1 in the consistency analysis. The resulting scheme is written as

$$\frac{\bar{u}_j^{n+1} - \bar{u}_j^n}{\Delta t} = -\frac{c}{2h}(\bar{u}_{j+1}^n - \bar{u}_{j-1}^n) + \frac{\Delta t}{2} \left(\frac{c}{h}\right)^2 (\bar{u}_{j+1}^n - 2\bar{u}_j^n + \bar{u}_{j-1}^n).$$

This is the Lax–Wendroff scheme [LAX 60]. Accordingly, it is consistent of order 2 in time and space.

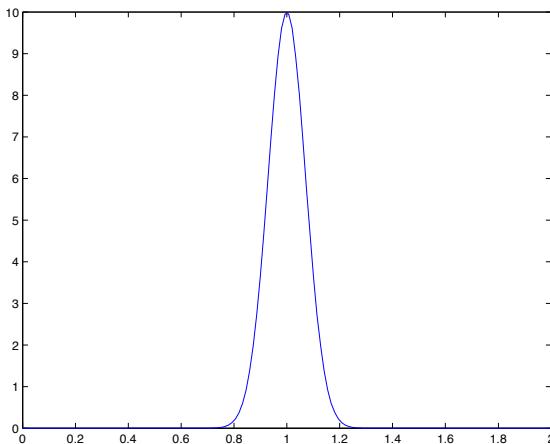
We compare these different schemes, given the following details:

- the simulation is conducted over the interval  $[0, 2]$ , with periodic conditions;
- the velocity is constant,  $c = 1.37$ ;
- the time and space steps are constant, and related by the condition  $c\Delta t = 0.9h$ .

We conduct a first simulation with the initial data (see Figure 3.11)

$$u_{\text{Init}}(x) = 10 \exp(-100|x - 1|^2). \quad [3.26]$$

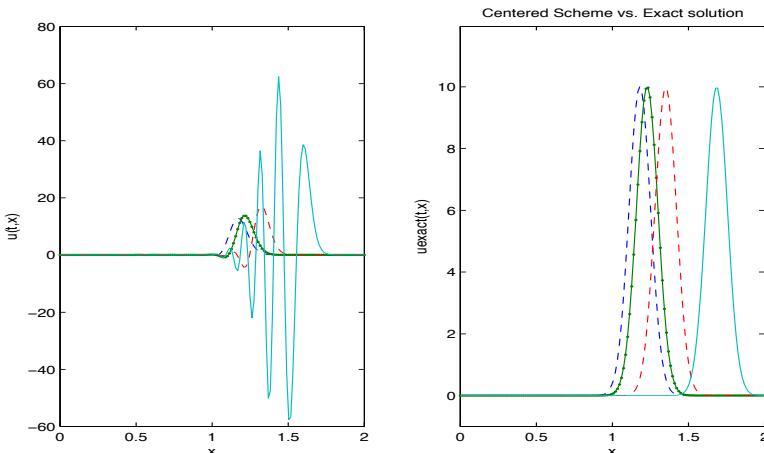
We recall that the exact solution is known and is expressed as  $u_{\text{Init}}(x - ct)$ .



**Figure 3.11.** Initial data [3.26]

Figure 3.12 shows the solution obtained with scheme [3.20]. The solution respects neither the positivity nor the upper bound of the exact solution, with

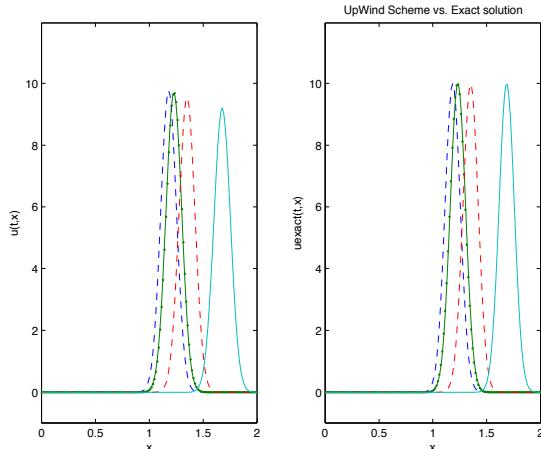
oscillations appearing and growing over time, which illustrates the numerical instability phenomenon. Figure 3.14 shows the solution obtained with different schemes in comparison with the exact solution (upwind, Lax–Friedrichs, Lax–Wendroff and the Després–Lagoutière schemes, the latter being a recent innovation described in [DES 01]). We observe that the maximum of the solution is poorly respected by the upwind and Lax–Friedrichs schemes, which also have the tendency of spreading the solution (the numerical solution takes significantly positive values over a larger interval than the exact solution). The solution of the Després–Lagoutière scheme presents a more irregular profile, with plateaus forming. The solution produced by Lax–Wendroff is better. These differences diminish as the discretization is refined. Figure 3.15 allows verification of the stated orders of convergence, confirming the visually observed behavior.



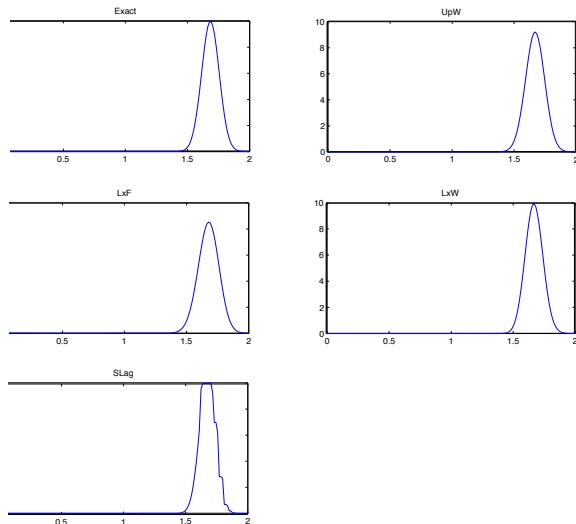
**Figure 3.12.** Simulation of the transport equation with the centered scheme ( $h = 1.34 \cdot 10^{-2}$ : numerical solution for different times  $0 \leq t \leq .5$ )

NOTE 3.6.– It is tempting, on the strength of experience gained with ODEs and diffusion problems, to look to relax the stability constraint  $|c|\frac{\Delta t}{h} < 1$  and to construct a method that works under less restrictive conditions in the form of an implicit scheme. Thus, in the case of constant velocity  $c > 0$ , we would write

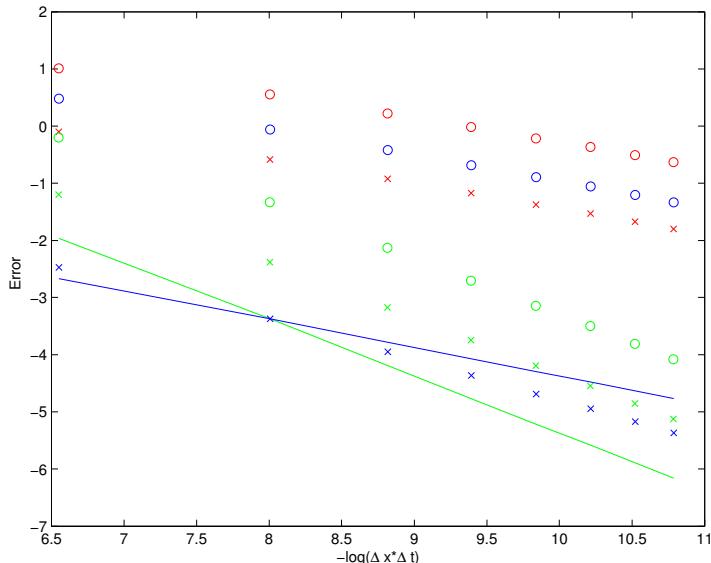
$$\frac{u_j^{n+1} - u_j^n}{\Delta t} + c \frac{u_j^{n+1} - u_{j-1}^{n+1}}{h} = 0.$$



**Figure 3.13.** Simulation of the transport equation with the upwind scheme ( $h = 1.34 \cdot 10^{-2}$ : numerical solution for different times  $0 \leq t \leq .5$ ). For the color version of this figure, see [www.iste.co.uk/goudon/mathmodel.zip](http://www.iste.co.uk/goudon/mathmodel.zip)



**Figure 3.14.** Simulation of the transport equation, numerical solutions at time  $T = .5$  ( $h = 1.34 \cdot 10^{-2}$ ): exact solution (Exact), upwind scheme (UpW), Lax–Friedrichs scheme (LxF), Lax–Wendroff scheme (LxW), Després–Lagoutière scheme (SLag)



**Figure 3.15.** Simulation of the transport equation, evolution of the error in  $L^2$  and  $L^\infty$  norms for upwind scheme (UpW, in blue), Lax–Friedrichs scheme (LxF, in red) and Lax–Wendroff scheme (LxW, in green). The  $L^2$  norms are represented by  $\times$ , and the  $L^\infty$  norms are represented by  $\circ$ ; for comparison with reference to straight lines of slopes 1 and 2, respectively. For the color version of this figure, see [www.iste.co.uk/goudon/mathmodel.zip](http://www.iste.co.uk/goudon/mathmodel.zip)

This scheme does indeed preserve the positivity of the solutions, as well as the  $L^\infty$  bounds. Indeed, assuming that  $0 \leq u_j^n \leq M$  for all  $j$  and that  $u_{j_0}^{n+1} = \min_j u_j^{n+1}$  (respectively  $u_{j_0}^{n+1} = \max_j u_j^{n+1}$ ), then we have

$$0 \leq u_{j_0}^n = u_{j_0}^{n+1} \left( 1 + c \frac{\Delta t}{h} \right) - c \frac{\Delta t}{h} u_{j_0-1}^{n+1} \leq u_{j_0}^{n+1} \left( 1 + c \frac{\Delta t}{h} \right) - c \frac{\Delta t}{h} u_{j_0}^{n+1} = u_{j_0}^{n+1}$$

(respectively  $M \geq u_{j_0}^n \geq u_{j_0}^{n+1}$ ). This scheme has been analyzed under much more general conditions in [BOY 12], and the fact that it allows, a priori, for the use of larger time intervals is an advantage that makes it useful for certain applications. However, its use is often regarded with caution for at least two reasons:

- Updating the unknown values is done by solving a linear system

$$(\mathbb{I} + \frac{\Delta t}{h} C) U^{n+1} = U^n$$

(which is invertible since the matrix is strictly diagonally dominant). If the matrix  $(\mathbb{I} + \frac{\Delta t}{h} C)$  has a very sparse structure, its inverse  $(\mathbb{I} + \frac{\Delta t}{h} C)^{-1}$  is a full matrix. Thus, the approximation  $u_j^{n+1}$  at the point on the grid  $x_j$  involves the solution at the preceding

discrete time over *all* points in the domain of calculation. This contradicts the principle of propagation at a finite speed of the continuous problem (the solution  $u(t^{n+1}, x_j) = u(t^n, x - c\Delta t)$  depends only on  $u(t^n, y)$  for the points  $y \in [x - c\Delta t, x]$ ). More precisely, the matrix  $C$  is expressed as

$$C = c \begin{pmatrix} 1 & 0 & \cdots & 0 \\ -1 & 1 & \cdots & 0 \\ 0 & -1 & 1 & \cdots \\ \vdots & \vdots & \ddots & 0 \\ 0 & 0 & \cdots & -1 & 1 \end{pmatrix}.$$

In fact, it is convenient to express the matrix  $(\mathbb{I} + \frac{\Delta t}{h} C)$  of the linear system in the form  $(1 + \frac{c\Delta t}{h})(\mathbb{I} - \lambda N)$ , where  $N$  is the nilpotent matrix

$$N = \begin{pmatrix} 0 & \cdots & 0 \\ 1 & \cdots & 0 \\ 0 & \cdots & 0 \\ \vdots & \ddots & 0 \\ 0 & 0 & \cdots & 1 & 0 \end{pmatrix},$$

and  $\lambda = \frac{c\Delta t}{1 + \frac{c\Delta t}{h}}$ . The calculation of the inverse thus becomes very simple; we obtain

$$\begin{aligned} (\mathbb{I} + \frac{\Delta t}{h} C)^{-1} &= \frac{1}{1 + \frac{c\Delta t}{h}} \sum_{k=0}^{\infty} \lambda^k N^k = \frac{1}{1 + \frac{c\Delta t}{h}} \sum_{k=0}^{J-1} \lambda^k N^k \\ &= \frac{1}{1 + \frac{c\Delta t}{h}} \begin{pmatrix} 1 & 0 & \cdots & 0 \\ \lambda & 1 & \cdots & 0 \\ \lambda^2 & \lambda & 1 & \cdots \\ \vdots & \vdots & \ddots & 0 \\ \lambda^{J-1} & \cdots & \lambda^2 & \lambda & 1 \end{pmatrix}. \end{aligned}$$

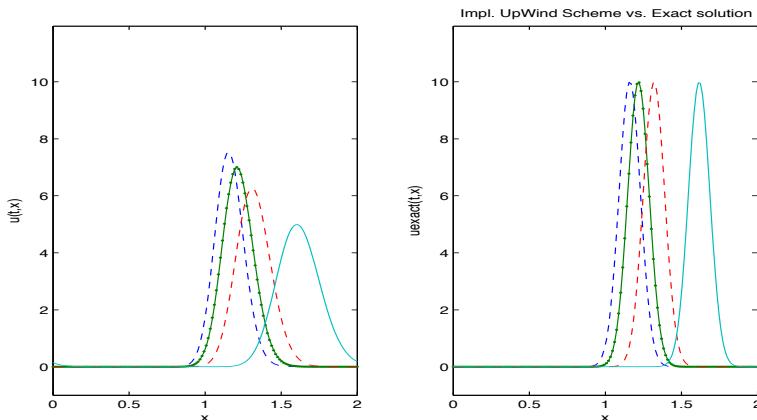
The coefficients of this matrix are indeed positive, which leads back to an argument that the maximum principle is satisfied. We remark that the evaluation of the solution at a point  $x_0$  depends on the values over all points situated to the left of  $x_0$ .

– As noted in [BRU 97], explicit and implicit schemes result in an approximation which is consistent up to  $\mathcal{O}(\Delta t^2 + \Delta x^2)$  of the equation

$$\partial_t u_\delta + c \partial_x u_\delta = \delta \partial_{xx}^2 u_\delta$$

where  $\delta$  is a function of the numerical parameters which depend on the scheme used

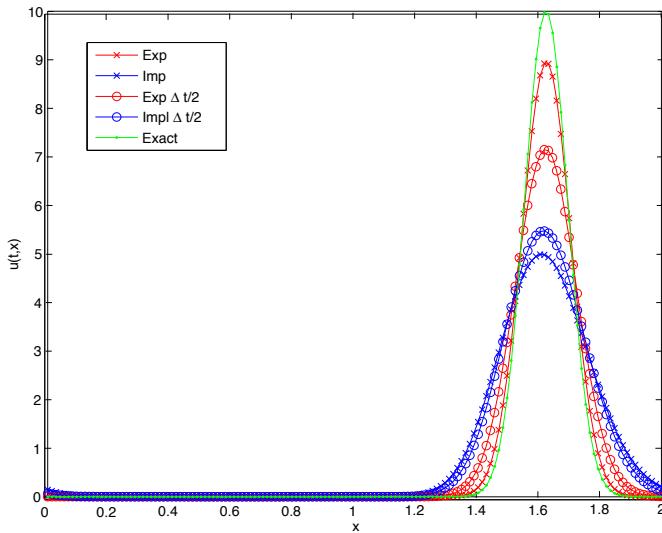
$$\delta_{\text{Exp}} = \frac{c}{2}(h - c\Delta t), \quad \delta_{\text{Imp}} = \frac{c}{2}(h + c\Delta t).$$



**Figure 3.16.** Simulation of the transport equation by the implicit upwind scheme and comparison with the exact solution at different times. For the color version of this figure, see [www.iste.co.uk/goudon/mathmodel.zip](http://www.iste.co.uk/goudon/mathmodel.zip)

In the explicit case, we recover the condition  $\Delta t \leq \frac{h}{c}$ , without which this convection-diffusion equation is unstable (examine the behavior of the Fourier transform), but by increasing  $\Delta t$ , we reduce the diffusion. In the implicit case, the diffusion term does not cancel and grows with  $\Delta t$ . The implicit scheme inherently introduces a stronger numerical diffusion which further distorts the result. Figure 3.16 compares the exact solution with the numerical solution given by the implicit upwind scheme: this result is to be compared with Figure 3.13, which shows the solution from the explicit upwind scheme, in the same numerical conditions ( $h = 0.0134$ ,  $\Delta t = 0.0083$ ). We clearly see that the implicit solution is subjected to stronger numerical diffusion. This analysis is completed by Figure 3.17, where we examine the influence of the time step on the performances of the explicit and implicit schemes, for

a fixed mesh size. We recover the stated behavior: reducing the time step  $\Delta t$  increases the numerical diffusion for the explicit scheme and reduces it for the implicit scheme.



**Figure 3.17.** Behavior of the explicit and implicit upwind schemes for different time steps. For the color version of this figure, see [www.iste.co.uk/goudon/mathmodel.zip](http://www.iste.co.uk/goudon/mathmodel.zip)

We have seen that we could attribute meaning to equation [3.13] when the initial data are discontinuous. It is interesting to study the behavior of numerical schemes in this situation, because, as we shall see later, nonlinear problems inevitably lead to the appearance of such discontinuities. The data are now (see Figure 3.18)

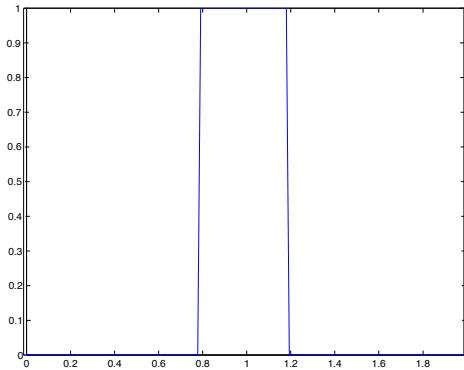
$$u_{\text{Init}} = \mathbf{1}_{1.8 \leq x \leq 1.2}, \quad [3.27]$$

for the simulations shown in Figure 3.19.

The upwind and Lax–Friedrichs schemes exhibit smooth and softened solutions, with spread support and reduced extrema: the effect of the numerical diffusion is recognizable. The Lax–Wendroff scheme presents strong parasitic oscillations at the points of discontinuity. The Després–Lagoutière [DES 01] scheme performs spectacularly in this particular case. From the error curve in Figure 3.20, it is apparent that the orders of convergence are altered by the presence of discontinuities. We also note that theorem 3.7 specifically does not apply to this situation, since the solution is not even continuous! However, we can justify the estimate of the observed error, by introducing slightly more elaborate analysis tools. The “natural” space for

the study of such solutions is the BV set of bounded functions, whose derivative (in the sense of distributions) is a measure. Specifically, we say that  $u : \mathcal{I} \subset \mathbb{R} \rightarrow \mathbb{R}$  is an element of  $\text{BV}(\mathcal{I})$  if  $u \in L^\infty(\mathcal{I})$  and there exists  $C > 0$ , such that for all functions  $\psi \in C_c^1(\mathcal{I})$ , we have

$$\left| \int_{\mathcal{I}} u\psi'(x) dx \right| \leq C\|\psi\|_{L^\infty}.$$



**Figure 3.18.** Initial condition [3.27]

We thus represent with  $\|u'\|_{\mathcal{M}^1}$  the infimum of such constants  $C$  and  $\|u\|_{\text{BV}} = \|u\|_{L^\infty} + \|u'\|_{\mathcal{M}^1}$ . Theorem 3.7 extends to data of this type, but in the  $L^1$  norm and with a fractional order of convergence.

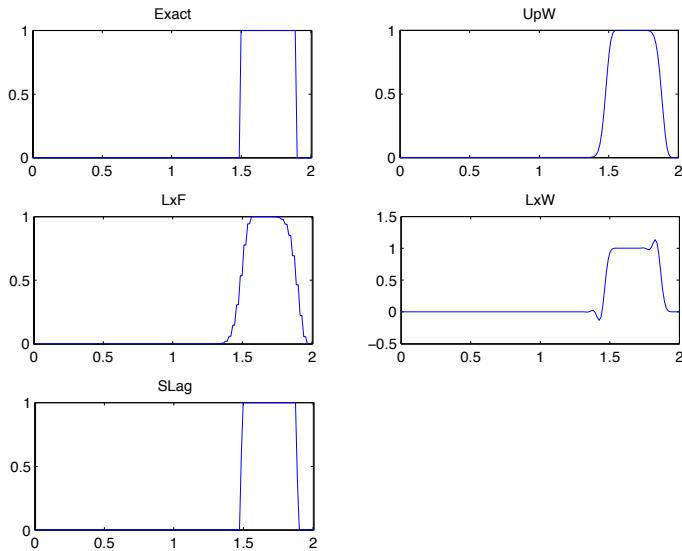
**THEOREM 3.8.–** Let  $u_{\text{Init}} \in \text{BV}_\#([0, 2\pi])$ . Let  $u$  be the associated solution of [3.13], with a constant velocity  $a(t, x) = c$ . Finally, let  $0 < T < \infty$ . Then, the sequence  $\bar{u}_j^n$  defined by [3.21] with a uniform mesh size  $h$  and under the condition  $|c|\Delta t/h \leq 1$  satisfies

$$\max_{0 \leq t^n \leq T} h \sum_j |\bar{u}_j^n - u(t^n, x_j)| \leq C\|u_{\text{Init}}\|_{\text{BV}}(h + \sqrt{Th}),$$

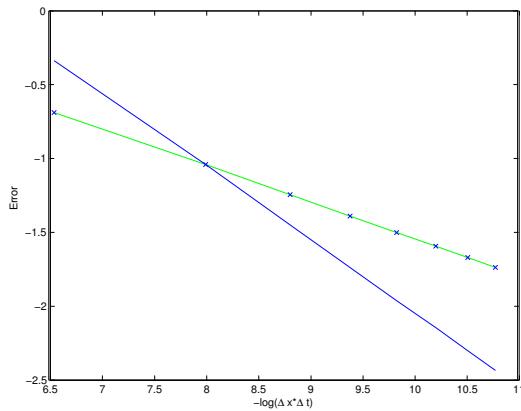
where  $C > 0$  does not depend on  $h$ , nor  $T$ , nor the data  $u_{\text{Init}}$ .

**PROOF.–** We introduce a sequence of mollifiers  $\zeta_\epsilon(x) = \frac{1}{\epsilon}\zeta(x/\epsilon)$ , where  $\zeta \in C_c^\infty(\mathbb{R})$  is supported in  $[-1, +1]$ , positive and of integral equal to 1. We set

$$u_{\text{Init}, \epsilon} = \zeta_\epsilon * u_{\text{Init}}(x).$$



**Figure 3.19.** Simulation of the transport equation, numerical solutions at time  $T = .5$  ( $h = 1.34 \cdot 10^{-2}$ ): exact solution (Exact), upwind scheme (UpW), Lax–Friedrichs scheme (LxF), Lax–Wendroff scheme (LxW) and Després–Lagoutière scheme (SLag)



**Figure 3.20.** Simulation of the transport equation: Order of convergence in the  $L^2$  norm with the upwind scheme and comparison with straight lines of slopes 1 and  $1/2$

We have  $u_{\text{Init},\epsilon} \in C^2$  and we prove without difficulty that  $\|u_{\text{Init},\epsilon}\|_{\text{BV}} \leq \|u_{\text{Init}}\|_{\text{BV}}$ , while  $\|\partial_{xx}^2 u_{\text{Init},\epsilon}\|_{L^\infty} = \|\partial_x \zeta_\epsilon \star \partial_x u_{\text{Init}}\|_{L^\infty} \leq \frac{1}{\epsilon} \|\partial_x \zeta\|_{L^\infty} \|u_{\text{Init}}\|_{\text{BV}}$ . Furthermore, we have

$$\begin{aligned} |u_{\text{Init},\epsilon}(x) - u_{\text{Init}}(x)| &= \left| \int \frac{1}{\epsilon} \zeta\left(\frac{x-y}{\epsilon}\right) (u_{\text{Init}}(y) - u_{\text{Init}}(x)) dy \right| \\ &\leq \int \int_0^1 \zeta(z) |\partial_x u_{\text{Init}}(x + \epsilon \theta z)| \epsilon \theta z d\theta dz \end{aligned}$$

which leads to

$$\|u_{\text{Init},\epsilon} - u_{\text{Init}}\|_{L^1} \leq \epsilon \|z \zeta(z)\|_{L^1} \|u_{\text{Init}}\|_{\text{BV}}.$$

Next, we exploit the following property of the upwind scheme: if the sequence  $v_j^n$  satisfies

$$v_j^{n+1} = v_j^n - c \frac{\Delta t}{h} (v_j^n - v_{j-1}^n) + \eta_j^n,$$

then, under the condition  $0 < c\Delta t/h \leq 1$ ,

$$\begin{aligned} h \sum_j |v_j^{n+1}| &\leq \left(1 - c \frac{\Delta t}{h}\right) h \sum_j |v_j^{n+1}| + c \frac{\Delta t}{h} h \sum_j |v_{j-1}^{n+1}| + h \sum_j |\eta_j^n| \\ &\leq h \sum_j |v_j^n| + h \sum_j |\eta_j^n| \leq h \sum_j |v_j^n| + h \sum_{k=0}^n \sum_j |\eta_j^k|. \end{aligned}$$

We represent with  $(t, x) \mapsto u(t, x)$  and  $(t, x) \mapsto u_\epsilon(t, x)$  the solutions of [3.13], with  $a(t, x) = c > 0$  and for initial data  $u(0, x) = u_{\text{Init}}(x)$ ,  $u_\epsilon(0, x) = \zeta_\epsilon \star u_{\text{Init}}(x)$ , respectively. Similarly, we use  $u_j^n$  and  $u_{\epsilon,j}^n$  to represent the numerical solutions of the upwind scheme with data  $u_j^0 = u_{\text{Init}}(x_j)$  and  $u_{j,\epsilon}^0 = \zeta_\epsilon \star u_{\text{Init}}(x_j)$ , respectively. The linearity of the equation and of the scheme allow us to write

$$\begin{aligned} h \sum_j |u_j^n - u(t^n, x_j)| &\leq h \sum_j (|u_j^n - u_{\epsilon,j}^n| + |u_{\epsilon,j}^n - u_\epsilon(t^n, x_j)| + |u_\epsilon(t^n, x_j) - u(t^n, x_j)|) \\ &\leq h \sum_j |u_j^0 - u_{\epsilon,j}^0| + h \sum_j |u_{\epsilon,j}^n - u_\epsilon(t^n, x_j)| + h \sum_j |u_\epsilon(0, x_j) - u(0, x_j)| \\ &\leq 2h \sum_j |u_{\text{Init}}(x_j) - \zeta_\epsilon \star u_{\text{Init}}(x_j)| + h \sum_j |u_{\epsilon,j}^n - u_\epsilon(t^n, x_j)|. \end{aligned}$$

We set  $U_\epsilon(x) = u_{\text{Init}}(x) - \zeta_\epsilon \star u_{\text{Init}}(x)$ . We can evaluate

$$\begin{aligned} h \sum_j |U_\epsilon(x_j)| &\leq \sum_j \left| h U_\epsilon(x_j) - \int_{x_j}^{x_{j+1}} U_\epsilon(y) dy \right| + \sum_j \int_{x_j}^{x_{j+1}} |U_\epsilon(y)| dy \\ &\leq \sum_j \int_{x_j}^{x_{j+1}} |U_\epsilon(x_j) - U_\epsilon(y)| dy + \int |U_\epsilon(y)| dy \\ &\leq \sum_j \int_{x_j}^{x_{j+1}} \int_0^1 |U'_\epsilon(x_j + \theta(y - x_j))| (y - x_j) d\theta dy \\ &\quad + \epsilon \|z\zeta(z)\|_{L^1} \|u_{\text{Init}}\|_{\text{BV}} \\ &\leq \sum_j \int_0^1 \int_{x_j}^{x_j + \theta(x_{j+1} - x_j)} |U'_\epsilon(z)| dz d\theta h dz d\theta + \epsilon \|z\zeta(z)\|_{L^1} \|u_{\text{Init}}\|_{\text{BV}} \\ &\leq h \|U_\epsilon\|_{\text{BV}} + \epsilon \|z\zeta(z)\|_{L^1} \|u_{\text{Init}}\|_{\text{BV}} \leq C(h + \epsilon) \|u_{\text{Init}}\|_{\text{BV}}. \end{aligned}$$

Furthermore, the convergence analysis conducted for the case of regular data ensures that

$$h \sum_j |u_{\epsilon,j}^n - u_\epsilon(t^n, x_j)| \leq CT h \|\zeta_\epsilon \star u_{\text{Init}}\|_{C^2} \leq CT h \frac{\|u_{\text{Init}}\|_{\text{BV}}}{\epsilon}.$$

We thus obtain

$$h \sum_j |u_j^n - u(t^n, x_j)| \leq C \|u_{\text{Init}}\|_{\text{BV}} \left( h + \epsilon + \frac{Th}{\epsilon} \right).$$

By minimizing the right-hand term with respect to  $\epsilon$ , we conclude from this that

$$h \sum_j |u_j^n - u(t^n, x_j)| \leq C \|u_{\text{Init}}\|_{\text{BV}} (h + \sqrt{Th}). \quad \square$$

Let us return to the case of a non-uniform mesh and a finite volume discretization, i.e. in the present case, assuming that the velocity  $a$  has positive values,

$$u_j^{n+1} = u_j^n - \frac{\Delta t}{h_j} (a_{j+1/2} u_j - a_{j-1/2} u_{j-1}).$$

We remark that the upwind scheme is not consistent (in the FD sense). Indeed, by using the Taylor expansion of  $u$ , we obtain

$$\begin{aligned}
& \frac{u(t^{n+1}, x_j) - u(t^n, x_j)}{\Delta t} + \frac{1}{h_j} (a(x_{j+1/2})u(t^n, x_j) - a(x_{j-1/2})u(t^n, x_{j-1})) \\
&= \frac{u(t^{n+1}, x_j) - u(t^n, x_j)}{\Delta t} + \frac{1}{h_j} (a(x_{j+1/2})u(t^n, x_{j+1/2}) \\
&\quad - a(x_{j-1/2})u(t^n, x_{j-1/2})) + \frac{a(x_{j+1/2})}{h_j} (u(t^n, x_j) - u(t^n, x_{j+1/2})) \\
&\quad - \frac{a(x_{j-1/2})}{h_j} (u(t^n, x_{j-1}) - u(t^n, x_{j-1/2})) \\
&= (\partial_t u + \partial_x(au))(t^n, x_j) \\
&\quad + a(x_{j+1/2})\partial_x u(t^n, x_{j+1/2})\frac{x_j - x_{j+1/2}}{h_j} \\
&\quad - a(x_{j-1/2})\partial_x u(t^n, x_{j-1/2})\frac{x_{j-1} - x_{j-1/2}}{h_j} + R_j^n.
\end{aligned}$$

The error term  $R_j^n$  tends to 0 as  $\Delta t, h \rightarrow 0$ . However, the leading term in the expansion of

$$a(x_{j+1/2})\partial_x u(t^n, x_{j+1/2})\frac{x_j - x_{j+1/2}}{h_j} - a(x_{j-1/2})\partial_x u(t^n, x_{j-1/2})\frac{x_{j-1} - x_{j-1/2}}{h_j}$$

is

$$-a\partial_x u(t^n, x_j)\frac{1}{h_j}\left(\frac{h_j}{2} - \frac{h_{j-1}}{2}\right) = -a\partial_x u(t^n, x_j)\frac{1}{2}\left(1 - \frac{h_{j-1}}{h_j}\right).$$

For general meshes, this term does not cancel and does not tend to 0.

Despite the non-consistency in the finite difference sense of the finite volume scheme over any grid, we shall nevertheless be able to establish the convergence of the FV scheme, in the  $L^2$  norm, by using a very different outlook (here for the case  $a(t, x) = c > 0$ ). We introduce the operator  $A_h$ , which associates with  $u = (u_1, \dots, u_{N_h}) \in \mathbb{R}^{N_h}$  the vector of  $\mathbb{R}^{N_h}$  whose coordinates are given by

$$[A_h u]_j = -c\frac{1}{h_j}(u_j - u_{j-1}),$$

with the convention  $u_0 = u_{N_h}$  (periodicity condition). For  $\tau > 0$ , we have

$$[(\mathbb{I} + \tau A_h)u]_j = \left(1 - c\frac{\tau}{h_j}\right)u_j + c\frac{\tau}{h_j}u_{j-1}$$

which appears as a convex combination of  $u_j$  and  $u_{j-1}$  as soon as  $0 < \tau \leq \frac{kh}{c}$ . In this case, we thus have

$$\|\mathbb{I} + \tau A_h\| \leq 1.$$

As a consequence, we also have

$$e^{tA_h} = e^{-\tau} e^{\tau(\mathbb{I} + t/\tau A_h)} = e^{-\tau} \sum_{n=0}^{\infty} \frac{\tau^n}{n!} (\mathbb{I} + t/\tau A_h)^n$$

which satisfies, by taking  $\tau \geq tc/kh$ ,

$$\|e^{tA_h}\| \leq e^{-\tau} \sum_{n=0}^{\infty} \frac{\tau^n}{n!} = 1.$$

We start by focusing on the semi-discrete problem by considering  $t \mapsto \tilde{u}(t) \in \mathbb{R}^{N_h}$  the solution of

$$\frac{d}{dt} \tilde{u}(t) = A_h \tilde{u}(t), \quad \tilde{u}_j(0) = \int_{x_{j-1/2}}^{x_{j+1/2}} u_{\text{Init}}(y) dy.$$

We thus have  $\tilde{u}(t) = e^{tA_h} \tilde{u}_j(0)$  and we set

$$\tilde{u}_h(t, x) = \sum_{j=1}^{N_h} \tilde{u}_j(t) \mathbf{1}_{x_{j-1/2} \leq x < x_{j+1/2}}.$$

We calculate

$$\begin{aligned} \frac{1}{2} \frac{d}{dt} \int |\tilde{u}_h - u|^2 dx &= \frac{1}{2} \frac{d}{dt} \left\{ \int |\tilde{u}_h|^2 dx + \int |u|^2 dx - 2 \int u \tilde{u}_h dx \right\} \\ &= \frac{1}{2} \frac{d}{dt} \sum_{j=1}^{N_h} h_j |\tilde{u}_j|^2 + 0 - \frac{d}{dt} \left( \sum_{j=1}^{N_h} \tilde{u}_j \int_{x_{j-1/2}}^{x_{j+1/2}} u dx \right) \\ &= - \sum_{j=1}^{N_h} c \tilde{u}_j (\tilde{u}_j - \tilde{u}_{j-1}) + \sum_{j=1}^{N_h} \frac{c}{h_j} (\tilde{u}_j - \tilde{u}_{j-1}) \int_{x_{j-1/2}}^{x_{j+1/2}} u dx + \sum_{j=1}^{N_h} \tilde{u}_j \int_{x_{j-1/2}}^{x_{j+1/2}} c \partial_x u dx \end{aligned}$$

$$\begin{aligned}
&= -\frac{c}{2} \sum_{j=1}^{N_h} (\tilde{u}_j - \tilde{u}_{j-1})^2 \\
&\quad + c \underbrace{\sum_{j=1}^{N_h} (\tilde{u}_j - \tilde{u}_{j-1}) u(t, x_{j-1/2}) + c \sum_{j=1}^{N_h} \tilde{u}_j (u(t, x_{j+1/2}) - u(t, x_{j-1/2}))}_{=0} \\
&\quad + c \sum_{j=1}^{N_h} \frac{\tilde{u}_j - \tilde{u}_{j-1}}{h_j} \left( \int_{x_{j-1/2}}^{x_{j+1/2}} u \, dx - h_j u(t, x_{j-1/2}) \right).
\end{aligned}$$

We hence use the Cauchy–Schwarz inequality and the Young inequality  $ab \leq \frac{1}{4}a^2 + b^2$  to evaluate

$$\begin{aligned}
&\frac{1}{2} \frac{d}{dt} \int |\tilde{u}_h - u|^2 \, dx + \frac{c}{4} \sum_{j=1}^{N_h} (\tilde{u}_j - \tilde{u}_{j-1})^2 \\
&\leq c \sum_{j=1}^{N_h} \left( \frac{1}{h_j} \int_{x_{j-1/2}}^{x_{j+1/2}} u \, dx - u(t, x_{j-1/2}) \right)^2 \\
&\leq c \sum_{j=1}^{N_h} \left( \frac{1}{h_j} \int_{x_{j-1/2}}^{x_{j+1/2}} (u(t, x) - u(t, x_{j-1/2})) \, dx \right)^2 \\
&\leq \sum_{j=1}^{N_h} \frac{c}{h_j^2} \left( \int_{x_{j-1/2}}^{x_{j+1/2}} \int_0^1 \partial_x u(t, x_{j-1/2} + \theta(x - x_{j-1/2})) (x - x_{j-1/2}) \, d\theta \, dx \right)^2 \\
&\leq \sum_{j=1}^{N_h} \frac{c}{h_j^2} \int_0^1 \int_{x_{j-1/2}}^{x_{j+1/2}} |\partial_x u(t, x_{j-1/2} + \theta(x - x_{j-1/2}))|^2 \, dx \, d\theta \\
&\quad \times \int_0^1 \int_{x_{j-1/2}}^{x_{j+1/2}} (x - x_{j-1/2})^2 \, dx \, d\theta \\
&\leq \frac{c}{3} \sum_{j=1}^{N_h} h_j \int_0^1 \int_{x_{j-1/2}}^{x_{j+1/2}} |\partial_x u(t, x_{j-1/2} + \theta(x - x_{j-1/2}))|^2 \, dx \, d\theta \\
&\leq \frac{c}{3} \sum_{j=1}^{N_h} h_j \int_0^1 \int_{x_{j-1/2}}^{x_{j+1/2}} |\partial_x u(t, z)|^2 \, dz \, d\theta \leq \frac{c}{3} \|u\|_{H^1}^2 h.
\end{aligned}$$

We have thus established that there exists  $C > 0$ , such that for  $0 \leq t \leq T$

$$\|(\tilde{u}_h - u)(t)\|_{L^2} \leq \|(\tilde{u}_h - u)(0)\|_{L^2} + C\|u\|_{H^1}^2 \sqrt{Th}.$$

To conclude, there remains the evaluation of the error between the semi-discrete model and the numerical model. Now, we have

$$\begin{aligned}\tilde{u}_j(t^{n+1}) &= \tilde{u}_j(t^n) + \int_{t^n}^{t^{n+1}} [A_h \tilde{u}]_j(s) \, ds \\ &= \tilde{u}_j(t^n) + \Delta t [A_h \tilde{u}]_j(t^n) + \left[ A_h \int_{t^n}^{t^{n+1}} (\tilde{u}(s) - \tilde{u}(t^n)) \, ds \right]_j.\end{aligned}$$

It follows that  $e_j^n = u_j^n - \tilde{u}_j(t^n)$  satisfies

$$e_j^{n+1} = e_j^n + \Delta t [A_h e^n]_j + \eta_j^n$$

with

$$\eta_j^n = [A_h r^n]_j, \quad r_j^n = \int_{t^n}^{t^{n+1}} (\tilde{u}_j(s) - \tilde{u}_j(t^n)) \, ds.$$

An argument by recurrence immediately leads to

$$e^n = (\mathbb{I} + \Delta t A_h)^n e^0 + \Delta t \sum_{k=0}^{n-1} (\mathbb{I} + \Delta t A_h)^{n-1-k} A_h r^k.$$

We are therefore led to evaluate the norm of the operators

$$B_m = (\mathbb{I} + \Delta t A_h)^m A_h.$$

The stability constraint leads to the requirement that the parameter

$$\lambda = \frac{c \Delta t}{kh}$$

is an element of  $]0, 1[$ . We thus have

$$\begin{aligned}B_m &= \left( (1 - \lambda) \mathbb{I} + \lambda \left( \mathbb{I} + \frac{kh}{c} A_h \right) \right)^m A_h \\ &= \sum_{\ell=0}^m C_m^\ell (1 - \lambda)^{m-\ell} \lambda^\ell \left( \mathbb{I} + \frac{kh}{c} A_h \right)^\ell \left( \mathbb{I} + \frac{kh}{c} A_h - \mathbb{I} \right) \frac{c}{kh}.\end{aligned}$$

We set  $a_{m,\ell} = C_m^\ell (1 - \lambda)^{m-\ell} \lambda^\ell$  for  $\ell \in \{0, \dots, m\}$ , and  $a_{m,\ell} = 0$  otherwise, in order that

$$\begin{aligned} B_m &= \frac{c}{kh} \sum_\ell a_{m,\ell} \left\{ \left( \mathbb{I} + \frac{kh}{c} A_h \right)^{\ell+1} - \left( \mathbb{I} + \frac{kh}{c} A_h \right)^\ell \right\} \\ &= \frac{\lambda}{\Delta t} \sum_\ell (a_{m,\ell-1} - a_{m,\ell}) \left( \mathbb{I} + \frac{kh}{c} A_h \right)^\ell. \end{aligned}$$

We calculate directly

$$\begin{aligned} a_{m,\ell-1} - a_{m,\ell} &= (1 - \lambda)^{m-\ell} \lambda^{\ell-1} \frac{m!}{\ell!(m-\ell)!} (\ell(1 - \lambda) - (m - \ell)\lambda) \\ &= (1 - \lambda)^{m-\ell} \lambda^{\ell-1} \frac{m!}{\ell!(m-\ell)!} (\ell - m\lambda). \end{aligned}$$

Since  $\|\mathbb{I} + \frac{kh}{c} A_h\| \leq 1$ , this comes to

$$\|B_m\| \leq \frac{\lambda}{\Delta t} \left( \sum_{0 \leq \ell \leq m\lambda} (a_{m,\ell} - a_{m,\ell-1}) + \sum_{\ell > m\lambda} (a_{m,\ell-1} - a_{m,\ell}) \right) = 2 \frac{\lambda}{\Delta t} a_{m,\ell_0}$$

where  $\ell_0$  is the floor function of  $m\lambda$ . To conclude, we will use the estimate

$$0 \leq a_{m,\ell} \leq \min \left( 1, \frac{2}{\sqrt{\lambda(1 - \lambda)m}} \right). \quad [3.28]$$

Assuming that this inequality is satisfied, we obtain

$$\begin{aligned} \left\| \Delta t \sum_{k=0}^{n-1} (\mathbb{I} + \Delta t A_h)^{n-1-k} A_h r^k \right\| &\leq \Delta t \sum_{k=0}^{n-1} \|\mathbb{I} + \Delta t A_h\|^{n-1-k} \|r^k\| \\ &\leq \Delta t \sum_{k=0}^{n-1} 2 \frac{\lambda}{\Delta t} \min \left( 1, \frac{2}{\sqrt{\lambda(1 - \lambda)(n - 1 - k)}} \right) \|r^k\| \\ &\leq 2\lambda \sup_{k \in \{0, \dots, n-1\}} \|r^k\| \left( 2 + \sum_{m=2}^{n-1} \frac{2}{\sqrt{\lambda(1 - \lambda)m}} \right) \\ &\leq 2\lambda \sup_{k \in \{0, \dots, n-1\}} \|r^k\| \left( 2 + \frac{2}{\sqrt{\lambda(1 - \lambda)}} \sum_{m=1}^{n-1} \int_m^{m+1} \frac{dy}{\sqrt{y}} \right) \\ &\leq 2\lambda \sup_{k \in \{0, \dots, n-1\}} \|r^k\| \left( 2 + \frac{2}{\sqrt{\lambda(1 - \lambda)}} 2(\sqrt{n} - 1) \right). \end{aligned}$$

Now  $\lambda(1 - \lambda) \leq 1/2$  so  $2 - \frac{4}{\sqrt{\lambda(1-\lambda)}} \leq 0$  and we thus arrive at

$$\left\| \Delta t \sum_{k=0}^{n-1} (\mathbb{I} + \Delta t A_h)^{n-1-k} A_h r^k \right\| \leq \lambda \sup_{k \in \{0, \dots, n-1\}} \|r^k\| \frac{8\sqrt{n}}{\sqrt{\lambda(1-\lambda)}}.$$

We now recall that

$$r_j^n = \int_{t^n}^{t^{n+1}} \int_{t^n}^s \frac{d}{d\sigma} \tilde{u}_j(\sigma) d\sigma ds = \int_{t^n}^{t^{n+1}} \int_{t^n}^s [A_h \tilde{u}]_j(\sigma) d\sigma ds$$

so

$$\begin{aligned} \|r^n\| &\leq \int_{t^n}^{t^{n+1}} \int_{t^n}^{t^{n+1}} \|A_h\| \|\tilde{u}(\sigma)\| d\sigma ds = \int_{t^n}^{t^{n+1}} \int_{t^n}^{t^{n+1}} \|A_h\| \|e^{\sigma A_h} \tilde{u}(0)\| d\sigma ds \\ &\leq \Delta t^2 \|A_h\| \sup_{\sigma \geq 0} \|e^{\sigma A_h}\| \|\tilde{u}(0)\| \end{aligned}$$

Now we can set the inequality

$$\|A_h\| \leq \frac{c}{kh} \left( \|\mathbb{I} + \frac{kh}{c} A_h\| + \|\mathbb{I}\| \right) \leq \frac{2c}{kh},$$

and we recall that  $\|e^{\sigma A_h}\| \leq 1$ . We thus arrive at

$$\|r^n\| \leq \frac{2c\Delta t^2}{kh} \|\tilde{u}(0)\| = 2\lambda\Delta t \|\tilde{u}(0)\|.$$

We deduce from this that

$$\begin{aligned} \left\| \Delta t \sum_{k=0}^{n-1} (\mathbb{I} + \Delta t A_h)^{n-1-k} A_h r^k \right\| &\leq \frac{16\lambda^2 \Delta t \sqrt{n}}{\sqrt{\lambda(1-\lambda)}} \|\tilde{u}(0)\| \\ &= \frac{16\lambda^2}{\sqrt{1-\lambda}} \sqrt{\frac{n\Delta t}{c}} \sqrt{kh} \|\tilde{u}(0)\|. \end{aligned}$$

Thus, we can find a constant  $C > 0$ , which depends on the velocity  $c$  and on  $\lambda \in ]0, 1[$ , such that

$$\|u^n - \tilde{u}(t^n)\| \leq \underbrace{\|u^0 - \tilde{u}(0)\|}_{=0} + C\sqrt{Th}.$$

We introduce the stepwise constant function

$$u_h^n(x) = \sum_{j=1}^{N_h} u_j^n \mathbf{1}_{x_{j-1/2} \leq x < x_{j+1/2}}.$$

By using the triangle inequality, we conclude that

$$\begin{aligned} \|u_h^n - u(t^n)\|_{L^2} &\leq \|u_h^n - \tilde{u}(t^n)\|_{L^2} + \|\tilde{u}(t^n) - u(t^n)\|_{L^2} \\ &\leq \|\tilde{u}(0) - u(0)\|_{L^2} + M\sqrt{Th}, \end{aligned}$$

for a certain constant  $M > 0$  dependent on  $c$  and  $\lambda$ .

In order to demonstrate [3.28], the starting point is to note that  $C_m^\ell \alpha^\ell \beta^{m-\ell}$  can be interpreted as the  $\ell^{\text{th}}$  Fourier coefficient of the  $2\pi$ -periodic function  $\theta \mapsto (\alpha + \beta e^{i\theta})^m$ , since

$$\begin{aligned} \frac{1}{2\pi} \int_0^{2\pi} (\alpha + \beta e^{i\theta})^m e^{-i\ell\theta} d\theta &= \frac{1}{2\pi} \sum_{k=0}^m C_m^k \alpha^{m-k} \beta^k \int_0^{2\pi} e^{i\theta(k-\ell)} d\theta \\ &= \sum_{k=0}^m C_m^k \alpha^{m-k} \beta^k \delta_{k,\ell} = C_m^\ell \alpha^{m-\ell} \beta^\ell. \end{aligned}$$

Now, the classic formulae of trigonometry allow us to rewrite

$$(\alpha + \beta e^{i\theta})^m = (\alpha^2 + \beta^2 + 2\alpha\beta \cos(\theta))^{m/2} = ((\alpha + \beta)^2 - 4\alpha\beta \sin^2(\theta/2))^{m/2}$$

By also using the elementary inequalities<sup>3</sup>  $\sin^2(\theta/2) \geq \frac{\theta^2}{\pi^2}$  for all  $0 \leq \theta \leq \pi$  and  $1 - x \leq e^{-x}$  for all  $x \geq 0$ , we obtain

$$\begin{aligned} \frac{1}{2\pi} \int_0^{2\pi} (\alpha + \beta e^{i\theta})^m e^{-i\ell\theta} d\theta &= \frac{1}{\pi} \int_0^\pi ((\alpha + \beta)^2 - 4\alpha\beta \sin^2(\theta/2))^{m/2} d\theta \\ &\leq \frac{(\alpha + \beta)^2}{\pi} \int_0^\pi \left(1 - 4 \frac{\alpha\beta}{(\alpha + \beta)^2} \frac{\theta^2}{\pi^2}\right)^{m/2} d\theta \\ &\leq \frac{(\alpha + \beta)^2}{\pi} \int_0^\pi \exp\left(m \frac{2\alpha\beta}{\pi^2(\alpha + \beta)^2} \theta^2\right) d\theta \\ &\leq \frac{(\alpha + \beta)^2}{\pi} \int_0^\infty \exp\left(m \frac{2\alpha\beta}{\pi^2(\alpha + \beta)^2} \theta^2\right) d\theta = \frac{(\alpha + \beta)^3}{\sqrt{2m\alpha\beta}}. \end{aligned}$$

---

<sup>3</sup> The second inequality can be obtained by a simple study of the function, the first coming from the fact that the function  $z \mapsto \sin(z)$  is positively valued and concave over  $[0, \pi/2]$ ; it follows that  $\sin(tz + (1-t)z') \geq t \sin(z) + (1-t) \sin(z')$  for all  $0 \leq t \leq 1$ ,  $z, z' \in [0, \pi/2]$ , which we apply with  $z = 0$ ,  $z' = \pi/2$  and  $\theta = tz + (1-t)z'$ .

Although the problem seems to be very simple, the analysis of the finite volume scheme calls upon quite sophisticated arguments. This analysis is of real value, motivated by extensions to multi-dimensional frameworks, with discretization grids of general geometric constructions, and by nonlinear problems. It is thus a current theme of research, opened in particular by [KUZ 76]. Proofs are available giving the convergence in  $\mathcal{O}(h^{1/2})$  in the  $L^\infty(0, T; L^1(\mathbb{R}^N))$  norm for integrable data  $L^1 \cap BV$  or in  $\mathcal{O}(h^{1/2-\epsilon})$  for all  $\epsilon > 0$  in the  $L^\infty((0, T) \times \mathbb{R}^N)$  norm for Lipschitz data. The details of this analysis are found in [MER 07, MER 08], while a very original outlook, interpreting the numerical scheme in terms of Markov chains, is developed in [DEL 11]. Here, we have followed the arguments presented in [DES 04a, DES 04b].

### 3.2.5. Two-dimensional simulations

Once we have an effective scheme for dealing with transport in dimension one, it is relatively simple to obtain a multi-dimensional scheme by a directional decomposition approach. The idea is inspired by the solution of a differential system

$$\frac{d}{dt}X = (A + B)X$$

by a method where we successively solve a system involving only matrix  $A$  followed by a system involving only matrix  $B$ . More specifically, knowing  $X_n$ , we start by solving over  $[n\Delta t, (n+1)\Delta t]$

$$\frac{d}{dt}Y = AY, \quad Y(n\Delta t) = X_n,$$

then we solve, still over  $[n\Delta t, (n+1)\Delta t]$

$$\frac{d}{dt}Z = BZ, \quad Z(n\Delta t) = Y((n+1)\Delta t)$$

and we set

$$X_{n+1} = Z((n+1)\Delta t).$$

In other words, given that we know how to solve these two equations exactly, we have

$$Y((n+1)\Delta t) = e^{A\Delta t}X_n, \quad X_{n+1} = e^{B\Delta t}e^{A\Delta t}X_n.$$

We must compare this expression with the exact solution

$$X((n+1)\Delta t) = e^{(A+B)\Delta t}X(n\Delta t).$$

By writing  $e_n = X(n\Delta t) - X_n$ , we thus obtain

$$e_{n+1} = e^{B\Delta t} e^{A\Delta t} e_n + [e^{(A+B)\Delta t} - e^{B\Delta t} e^{A\Delta t}] X(t_n).$$

Now, we have

$$e^{(A+B)\Delta t} = \mathbb{I} + (A + B)\Delta t + (A^2 + AB + BA + B^2) \frac{\Delta t^2}{2} + H_1(\Delta t)\Delta t^2$$

and

$$e^{B\Delta t} e^{A\Delta t} = \mathbb{I} + A\Delta t + B\Delta t + A^2 \frac{\Delta t^2}{2} + B^2 \frac{\Delta t^2}{2} + BA\Delta t^2 + H_2(\Delta t)\Delta t^2$$

where  $\lim_{\Delta t \rightarrow 0} H_j(\Delta t) = 0$ . It follows that

$$e^{(A+B)\Delta t} - e^{B\Delta t} e^{A\Delta t} = (AB - BA) \frac{\Delta t^2}{2} + H(\Delta t)\Delta t^2$$

with  $\lim_{\Delta t \rightarrow 0} H(\Delta t) = 0$ . We set  $0 < T < \infty$  and set  $\Delta t = T/N$ , for a non-null integer  $N$ . We can thus find a constant  $C_T > 0$ , such that for all  $n \in \{0, \dots, N\}$ , we have

$$|e_{n+1}| \leq C_T(|e_n| + \Delta t^2).$$

Then, by recurrence, we obtain (with  $e_0 = X(0) - X_0 = 0$ )

$$|e_n| \leq C_T n \Delta t^2 \leq T C_T \Delta t.$$

The decomposition thus induces an error of order one in  $\Delta t$ . We can improve this decomposition error with the following strategy:

$$\text{solve over } [n\Delta t, (n + 1/2)\Delta t] \quad \frac{d}{dt} Y = AY, \quad Y(n\Delta t) = X_n,$$

$$\text{solve over } [n\Delta t, (n + 1)\Delta t] \quad \frac{d}{dt} Z = BZ, \quad Z(n\Delta t) = Y((n + 1/2)\Delta t),$$

$$\text{solve over } [n\Delta t, (n + 1/2)\Delta t] \quad \frac{d}{dt} \tilde{Y} = A\tilde{Y}, \quad \tilde{Y}(n\Delta t) = Z((n + 1)\Delta t).$$

We thus set  $X_{n+1} = \tilde{Y}((n + 1/2)\Delta t)$ . In other words, we have

$$X_{n+1} = e^{A\Delta t/2} e^{B\Delta t} e^{A\Delta t/2} X_n.$$

This decomposition, called the *Strang decomposition*, induces an error of order two in  $\Delta t$ .

We can derive inspiration from this reasoning to numerically solve the equation

$$\partial_t \rho + \partial_x(u\rho) + \partial_y(v\rho) = 0,$$

over the domain  $\Omega = [0, \ell] \times [0, L]$ , with given  $u, v : \Omega \mapsto \mathbb{R}$ , letting the operator  $\partial_x(u\bullet)$  play the role of  $A$  and the operator  $\partial_y(v\bullet)$  play that of  $B$ . Thus, we consider a mesh of  $\Omega$  of step size  $\Delta x$ ,  $\Delta y$  and, given  $\rho_{i,j}^n$ , an approximation of  $\rho(n\Delta t, i\Delta x, j\Delta y)$ , we first define

$$\begin{aligned}\rho_{i,j}^* &= \rho_{i,j}^n - \frac{\Delta t}{\Delta x}(F_{i+1/2,j} - F_{i-1/2,j}), \\ F_{i+1/2,j} &= [u((i+1/2)\Delta x, j\Delta y)]_+ \rho_{i,j}^n + [u((i+1/2)\Delta x, j\Delta y)]_- \rho_{i+1,j}^n\end{aligned}$$

with  $[z]_+ = \max(0, z) \geq 0$ ,  $[z]_- = \min(0, z) \leq 0$ , then we set

$$\begin{aligned}\rho_{i,j}^{n+1} &= \rho_{i,j}^* - \frac{\Delta t}{\Delta y}(G_{i,j+1/2} - G_{i,j-1/2}), \\ G_{i,j+1/2} &= [v(i\Delta x, (j+1/2)\Delta y)]_+ \rho_{i,j}^* + [v(i\Delta x, (j+1/2)\Delta y)]_- \rho_{i,j+1}^*\end{aligned}$$

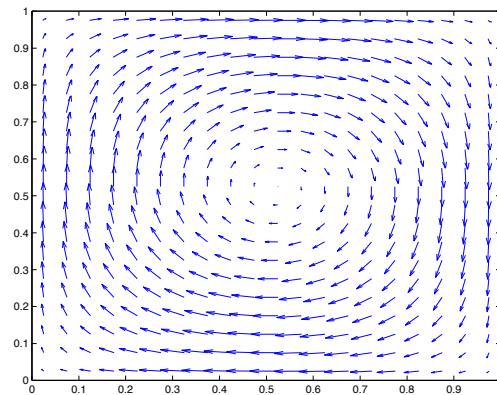
We use this scheme to conduct the transport equation simulation with  $\ell = 1 = L$  and the velocity field

$$u(x, y) = -\sin(\pi x) \cos(\pi y), \quad v(x, y) = \cos(\pi x) \sin(\pi y).$$

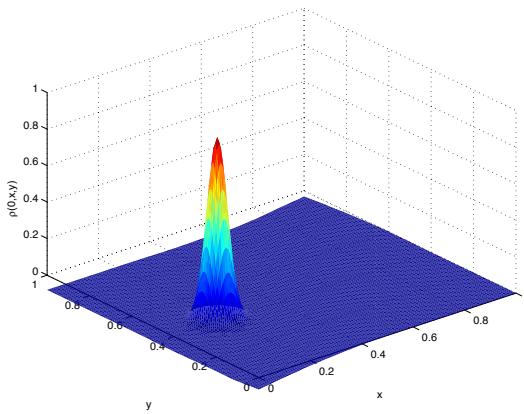
This vector field is represented in Figure 3.21; we see that it acts to rotate the density  $\rho$  in the clockwise direction. The initial data

$$\rho_{\text{Init}}(x, y) = \exp(-500((x - 0.25)^2 + (y - 0.5)^2))$$

are represented in Figure 3.22. The evolution of the solution is presented in Figure 3.23 ( $\Delta x = 1/100$ ,  $\Delta y = 1/120$ , CFL number = 0.4): we clearly see the rotation effect imposed by the velocity field. However, this result is not very satisfactory, since the numerical diffusion is significant: the support of the numerical solution broadens and it is clear from Figure 3.24 that, at the final time  $T = 1.5$ , the maximum of the solution is very distinctly underestimated. The final thing to be said about this approach is that it inherently imposes the use of Cartesian grids. However, in practice, for equations involving variable coefficients and complex geometries, it can be attractive to work with more general meshes. This is one of the reasons that motivates the development of finite volume methods.



**Figure 3.21.** 2D velocity field



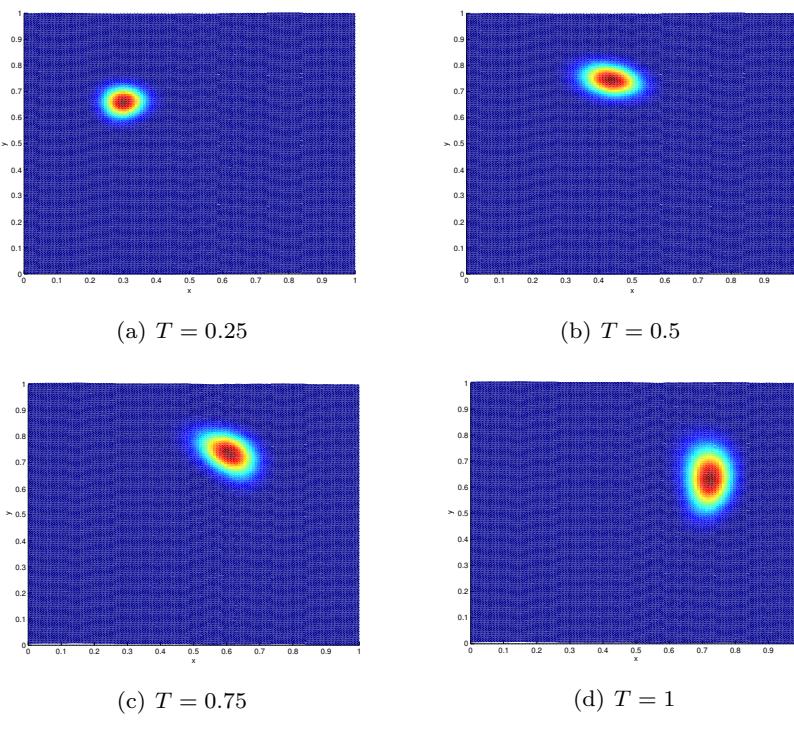
**Figure 3.22.** Initial 2D condition. For the color version of this figure, see [www.iste.co.uk/goudon/mathmodel.zip](http://www.iste.co.uk/goudon/mathmodel.zip)

### 3.2.6. The dynamics of prion proliferation

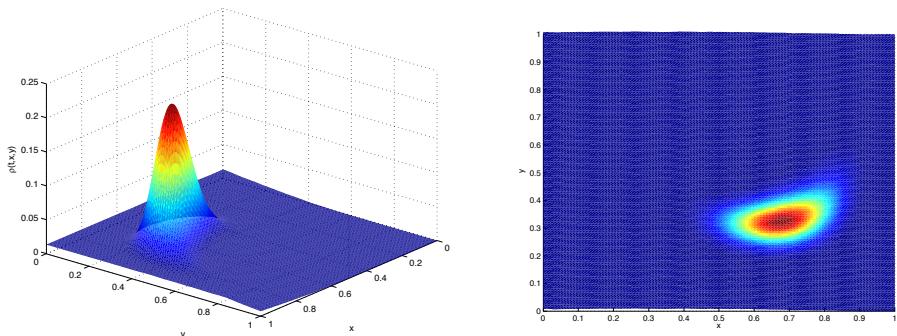
We are interested here in the mathematical modeling of the development of spongiform encephalopathy, a disease whose human form is known by the name “Creutzfeldt–Jakob disease”, which is caused by the action of an infectious agent called a prion. The same arguments apply to descriptions of other pathologies, such as Alzheimer’s disease, for example. The prion consists of a modified form of the

protein  $\text{PrP}^C$ , designated  $\text{PrP}^{Sc}$ . Molecules of  $\text{PrP}^C$  are present in the body and play a protective role in metabolism. The disease corresponds to the formation of aggregates of  $\text{PrP}^C$  molecules, which are more stable than the free form and which resist treatment. The presence of aggregated  $\text{PrP}^{Sc}$  molecules promotes the formation of new aggregates; this results in a reduction of the number of free  $\text{PrP}^C$  molecules and thus the disappearance of natural defenses which leads to irreversible neuron damage. The description of the mechanism of the disease is based on the following hypotheses. We adopt a “continuous” description, where the aggregated molecules are characterized by their size  $x \geq 0$ , while the  $\text{PrP}^C$  are seen as monomers of infinitely small size with respect to the aggregates. The  $\text{PrP}^C$  and  $\text{PrP}^{Sc}$  molecules are subject to:

- a natural metabolic degradation characterized by distinct rates for monomers and aggregates,  $\gamma$  and  $\mu(x)$ , respectively;
- a polymerization mechanism, described by a rate  $\tau(x)$ , where  $\text{PrP}^C$  monomers attach to  $\text{PrP}^{Sc}$  molecules to form a larger chain;
- a fragmentation mechanism where aggregates break up into smaller chains.



**Figure 3.23.** Solution of the transport equation at different times. For the color version of this figure, see [www.iste.co.uk/goudon/mathmodel.zip](http://www.iste.co.uk/goudon/mathmodel.zip)



**Figure 3.24.** Solution of the transport equation at time  $T = 1.5$ . For the color version of this figure, see [www.iste.co.uk/goudon/mathmodel.zip](http://www.iste.co.uk/goudon/mathmodel.zip)

The dynamics of prion proliferation can thus be described as the evolution of the population, represented as  $V(t)$ , of  $\text{PrP}^C$  at time  $t \geq 0$ , and the density  $u(t, x)$  of polymers of size  $x \geq 0$  at time  $t \geq 0$ . More precisely,  $\int_a^b u(t, x) dx$  gives the number of polymers whose size lies within  $a$  and  $b$  at time  $t \geq 0$ . The fragmentation phenomenon is subject to a threshold effect: there exists a critical size  $x_0 \geq 0$  beneath which the aggregates are not stable and are immediately dissolved into the monomer form. The fragmentation is described by two parameters: a fragmentation rate  $\beta(x)$  and the probability density, written as  $x \mapsto \kappa(x, y)$ , that a polymer of size  $y$  fragments into a chain of size  $x$  and a chain of size  $y - x$ . Evidently, this only has meaning for  $y \geq x$ . The nucleus of fragmentation hence satisfies a natural symmetry property

$$\kappa(y - x, y) = \kappa(x, y)$$

and the normalization constraint

$$\int_0^\infty \kappa(x, y) dx = \begin{cases} 0 & \text{if } y \leq x_0, \\ 1 & \text{if } y > x_0. \end{cases}$$

We thus arrive at the following system, which couples an ordinary differential equation with an equation of integro-differential nature,

$$\left\{ \begin{array}{l} \frac{d}{dt} V(t) = \lambda - \gamma V(t) - V(t) \int_{x_0}^\infty \tau(x) u(t, x) dx \\ \quad + 2 \int_0^{x_0} x \left( \int_{x_0}^\infty \beta(y) \kappa(x, y) u(t, y) dy \right) dx, \\ \frac{\partial}{\partial t} u(t, x) + V(t) \frac{\partial}{\partial x} (\tau(x) u(t, x)) \\ \quad = -(\mu(x) + \beta(x)) u(t, x) + 2 \int_x^\infty \beta(y) \kappa(x, y) u(t, y) dy, \end{array} \right. [3.29]$$

for  $t \geq 0$ ,  $x \geq x_0$ , where  $\lambda > 0$  is a natural source of monomers. The factors of 2 take into account the symmetry of the fragmentations  $y \rightarrow (y - x, x)$  and  $y \rightarrow (x, y - x)$ , which contribute to the formation of polymers of size  $x$ . The problem is completed by data of initial conditions  $V(0) = V_0$ ,  $u(0, x) = u_0(x)$  and the boundary condition  $u(t, x_0) = 0$ , which accounts for the fact that monomers of size less than or equal to  $x_0$  fragment instantaneously and are not observable.

A crucial consequence of the hypotheses regarding  $\kappa$  is the equality  $\int_0^y x\kappa(x, y) dx = y/2$ . This results in the following calculation:

$$\begin{aligned} \int_0^y x\kappa(x, y) dx &= \int_0^y x\kappa(y - x, y) dx = \int_0^y (y - z)\kappa(z, y) dz \\ &= \frac{y}{2} \int_0^y \kappa(x, y) dx = \frac{y}{2}. \end{aligned}$$

This relation expresses a mass conservation property: the mass (proportional to  $y$ ) of the initial fragment is contained in the fragmentation products of size  $x$  and  $y - x$ . We thus observe that

$$\frac{d}{dt} \left( V(t) + \int_{x_0}^{\infty} xu(t, x) dx \right) = \lambda - \gamma V(t) - \int_{x_0}^{\infty} \mu(x)xu(t, x) dx.$$

This calculation is based on integration by parts and exploits the observation:

$$\begin{aligned} &2 \int_{x_0}^{\infty} x \int_x^{\infty} \beta(y)\kappa(x, y)u(t, y) dy dx + 2 \int_0^{x_0} x \left( \int_{x_0}^{\infty} \beta(y)\kappa(x, y)u(t, y) dy \right) dx \\ &= 2 \int_{x_0}^{\infty} \left( \int_{x_0}^y x\kappa(x, y) dx \right) \beta(y)u(t, y) dy + 2 \int_{x_0}^{\infty} \left( \int_0^{x_0} x\kappa(x, y) dx \right) \beta(y)u(t, y) dy \\ &= 2 \int_{x_0}^{\infty} \left( \int_0^y x\kappa(x, y) dx \right) \beta(y)u(t, y) dy = 2 \int_{x_0}^{\infty} \frac{y}{2} \beta(y)u(t, y) dy. \end{aligned}$$

### *A simplified model*

First, we can assume that

- the rates of degradation  $\mu$  and polymerization  $\tau$  do not depend on the size variable  $x$ ;
- the rate of fragmentation is proportional to the size of the aggregate:  $\beta(x) = \bar{\beta}x$  for a positive constant  $\bar{\beta}$ ;
- the probability of fragmentation is uniform over the length of the aggregate  $\kappa(x, y) = 0$  if  $y \leq x_0$  or  $y \leq x$ ,  $\kappa(x, y) = 1/y$  if  $y > x_0$  and  $y > x > 0$ .

With these definitions, when it is not null, the product  $\beta(y)\kappa(x,y)$  simply takes the value  $\bar{\beta}$ . These hypotheses allow us to obtain a closed ODE system for  $V(t)$ , the total number of polymers  $U(t) = \int_{x_0}^{\infty} u(t,x) dx$  and  $P(t) = \int_{x_0}^{\infty} x u(t,x) dx$ , which gives the number of monomers in the agglomerated form (or, up to a proportionality factor, the mass of polymers). In fact, we obtain

$$\begin{cases} \frac{d}{dt}U = \bar{\beta}P - (\mu + 2\bar{\beta}x_0)U, \\ \frac{d}{dt}V = \lambda - \gamma V - \tau VU + \bar{\beta}x_0^2U, \\ \frac{d}{dt}P = \tau VU - \mu P - \bar{\beta}x_0^2U. \end{cases} \quad [3.30]$$

(We might note that this system has a structure very similar to the one studied in section 1.2.7.) The analysis starts with

**THEOREM 3.9.** – For all initial values  $(U_0, V_0, P_0) \in \mathcal{X} = \{(a, b, c) \in \mathbb{R}^3, a \geq 0, b \geq 0, c \geq x_0a\}$ , problem [3.30] admits a unique solution  $t \mapsto (U(t), V(t), P(t)) \in \mathcal{X}$ , defined over  $[0, +\infty[$ .

By setting

$$F : (a, b, c) \in \mathcal{X} \mapsto \begin{pmatrix} \bar{\beta}c - (\mu + 2\bar{\beta}x_0)a, \\ \lambda - \gamma b - \tau ab + \bar{\beta}x_0^2a, \\ \tau ab - \mu c - \bar{\beta}x_0^2a \end{pmatrix}$$

we write the differential system in the (autonomous) form

$$\frac{d}{dt}Y = F(Y).$$

Since the function  $F$  is polynomial, for all initial data  $Y_{\text{Init}} \in \mathcal{X}$ , the Cauchy–Lipschitz theorem ensures the existence and uniqueness of a solution  $t \mapsto Y(t)$  defined over an interval  $[0, T[$  and such that  $Y(0) = Y_{\text{Init}}$ . We then observe that, if  $U_{\text{Init}} = P_{\text{Init}} = 0$ , then the corresponding solution is written as  $U(t) = 0$ ,  $P(t) = 0$  and

$$V(t) = e^{-\gamma t}V_{\text{Init}} + \frac{\lambda}{\gamma}(1 - e^{-\gamma t}). \quad [3.31]$$

By uniqueness, and the system being autonomous, this expression determines all the solutions for which  $U$  and  $P$  simultaneously cancel with each other (in at least a time  $t_0$ ). We now consider the solution associated with initial values  $U_{\text{Init}} > 0$ ,  $V_{\text{Init}} >$

0 and  $P_{\text{Init}} > x_0 U_{\text{Init}}$ . By continuity, we can assume that the solution still satisfies such inequalities over a given interval  $[0, t_0[$ . If  $V(t_0) = 0$ , then  $\frac{d}{dt}V(t_0) \geq \lambda > 0$  implies that  $t \mapsto V(t)$  is strictly increasing over a neighborhood  $[t_0 - \eta, t_0[$ , which leads to a contradiction. Similarly, if  $U(t_0) = 0$ , then  $\frac{d}{dt}U(t_0) = \bar{\beta}P(t_0) > 0$  (since the initial datas and the uniqueness of the solution to the Cauchy problem exclude the possibility that  $U$  and  $P$  cancel at the same time), which again leads to a contradiction. Finally, we calculate  $\frac{d}{dt}(P - x_0U) = \tau VU - (\mu + \bar{\beta}x_0)(P - x_0U)$ . Thus, if  $(P - x_0U)(t_0) = 0$ , we obtain  $\frac{d}{dt}(P - x_0U)(t_0) = \tau VU(t_0) > 0$ , and we arrive once more at a contradiction. We conclude from this that the solution satisfies  $U(t) > 0$ ,  $V(t) > 0$  and  $P(t) > x_0U(t)$  for all  $t \in [0, T[$ . We will exploit the positivity of the solutions to demonstrate that the solution is globally defined ( $T = +\infty$ ). Indeed, we note that there exists  $C > 0$ , which depends on the coefficients  $\bar{\beta}, \mu$ , such that

$$\frac{d}{dt}(U + V + P) \leq \lambda + C(U + V + P).$$

(Note that taking the sum  $\frac{d}{dt}(U + V + P)$  allows us to offset the quadratic terms  $\tau UV$  of [3.30].) Grönwall's lemma thus implies that  $(U + V + P)(t) \leq e^{Ct}((U + V + P)(0) + \lambda t)$ , an estimate which guarantees that the solutions, which we recall are positive, are defined for all times  $t \geq 0$ .

We shall focus on stationary solutions of [3.30] and on their stability. We immediately note that  $(0, \lambda/\gamma, 0)$  is a stationary solution of [3.30]. This solution describes a healthy state because there are no agglomerated molecules. This state also corresponds to the asymptotic behavior of solution [3.31] as  $t \rightarrow \infty$ . However, we note that

$$\bar{U} = \frac{\bar{\beta}\lambda\tau - \gamma(x_0\bar{\beta} + \mu)^2}{\mu\tau(2x_0\bar{\beta} + \mu)}, \quad \bar{V} = \frac{(x_0\bar{\beta} + \mu)^2}{\bar{\beta}\tau}, \quad \bar{P} = \frac{\bar{\beta}\lambda\tau - \gamma(x_0\bar{\beta} + \mu)^2}{\bar{\beta}\tau\mu}$$

is also a stationary solution. This solution is physically admissible when  $\sqrt{\bar{\beta}\lambda\tau/\gamma} > x_0\bar{\beta} + \mu$  and thus corresponds to an infected state. From a biological point of view, the stability of these states is a critical question. It would be natural to study the linear stability of these solutions, but in fact we can establish a more precise result, [GRE 06, PRÜ 06].

**THEOREM 3.10.– i)** If  $\sqrt{\bar{\beta}\lambda\tau/\gamma} < x_0\bar{\beta} + \mu$ , then the healthy state  $(0, \lambda/\gamma, 0)$  is globally stable and  $(U, V, P) \rightarrow (0, \lambda/\gamma, 0)$  as  $t \rightarrow \infty$ .

ii) If  $\sqrt{\bar{\beta}\lambda\tau/\gamma} > x_0\bar{\beta} + \mu$ , then the infected state  $(\bar{U}, \bar{V}, \bar{P})$  is globally stable.

In order to demonstrate *i*), the starting point is the calculation, for  $A > 0$ ,

$$\frac{d}{dt} \left( \frac{1}{2} \left( V - \frac{\lambda}{\gamma} \right)^2 + AP + A \frac{\mu}{\bar{\beta}} U \right) = -\gamma \left( V - \frac{\lambda}{\gamma} \right)^2 - U \Pi_A(V),$$

where

$$\begin{aligned}\Pi_A(V) &= \left(V - \frac{\lambda}{\gamma}\right)(\tau V - \bar{\beta}x_0^2) - A \left(\tau V - \bar{\beta}x_0^2 - \frac{\mu^2}{\bar{\beta}} - 2\mu x_0\right) \\ &= \tau V^2 - \left(\bar{\beta}x_0^2 + \tau \frac{\lambda}{\gamma} + A\tau\right)V + A \left(\bar{\beta}x_0^2 + \frac{\mu^2}{\bar{\beta}} + 2\mu x_0\right) + \frac{\lambda}{\gamma} \bar{\beta}x_0^2\end{aligned}$$

is a polynomial of order 2 in  $V$ . The condition  $\sqrt{\bar{\beta}\lambda\tau/\gamma} < x_0\bar{\beta} + \mu$  allows us to show that  $A > 0$  and  $\underline{\pi} > 0$ , such that for all real  $z$ ,  $\Pi_A(z) \geq \underline{\pi} > 0$ . Specifically, we note that  $\lim_{|V| \rightarrow \infty} \Pi_A(V) = +\infty$ , thus we only need to find  $A$ , such that  $V \mapsto \Pi_A(V)$  has no real roots. This leads to the requirement that the discriminant

$$\begin{aligned}d(A) &= \left(\bar{\beta}x_0^2 + \tau \frac{\lambda}{\gamma} + A\tau\right)^2 - 4\tau A \left(\bar{\beta}x_0^2 + \frac{\mu^2}{\bar{\beta}} + 2\mu x_0\right) - 4\tau \frac{\lambda}{\gamma} \bar{\beta}x_0^2 \\ &= \tau^2 A^2 + 2\tau \left(\tau \frac{\lambda}{\gamma} - \bar{\beta}x_0^2 - 2\frac{\mu^2}{\bar{\beta}} - 4\mu x_0\right) A + \left(\bar{\beta}x_0^2 + \tau \frac{\lambda}{\gamma}\right)^2 - 4\tau \frac{\lambda}{\gamma} \bar{\beta}x_0^2\end{aligned}$$

is strictly negative for certain values of  $A$ . However,  $A \mapsto d(A)$  is again a polynomial of order 2 which satisfies  $\lim_{A \rightarrow \infty} d(A) = +\infty$  and is thus sufficient to ensure that this polynomial has real roots. We are therefore led to prove that

$$\begin{aligned}&\left(\tau \frac{\lambda}{\gamma} - \bar{\beta}x_0^2 - 2\frac{\mu^2}{\bar{\beta}} - 4\mu x_0\right)^2 - \left(\bar{\beta}x_0^2 + \tau \frac{\lambda}{\gamma}\right)^2 + 4\tau \frac{\lambda}{\gamma} \bar{\beta}x_0^2 \\ &= \left(2\frac{\mu^2}{\bar{\beta}} + 4\mu x_0\right)^2 - 2 \left(2\frac{\mu^2}{\bar{\beta}} + 4\mu x_0\right) \left(\tau \frac{\lambda}{\gamma} - \bar{\beta}x_0^2\right) \\ &= 2 \left(2\frac{\mu^2}{\bar{\beta}} + 4\mu x_0\right) \left(\frac{\mu^2}{\bar{\beta}} + 2\mu x_0 - \tau \frac{\lambda}{\gamma} + \bar{\beta}x_0^2\right) \\ &= \frac{2}{\bar{\beta}} \left(2\frac{\mu^2}{\bar{\beta}} + 4\mu x_0\right) \left((x_0\bar{\beta} + \mu)^2 - \bar{\beta}\tau \frac{\lambda}{\gamma}\right) > 0.\end{aligned}$$

Thus, by assuming  $\sqrt{\bar{\beta}\lambda\tau/\gamma} < x_0\bar{\beta} + \mu$ , we must take  $A \in ]A_-, A_+[$ , with

$$\begin{aligned}A_{\pm} &= \frac{1}{\tau^2} \left\{ \tau \left(-\tau \frac{\lambda}{\gamma} + \bar{\beta}x_0^2 + 2\frac{\mu^2}{\bar{\beta}} + 4\mu x_0\right) \right. \\ &\quad \left. \pm \tau \sqrt{\frac{2}{\bar{\beta}} \left(2\frac{\mu^2}{\bar{\beta}} + 4\mu x_0\right) \left((x_0\bar{\beta} + \mu)^2 - \bar{\beta}\tau \frac{\lambda}{\gamma}\right)} \right\}.\end{aligned}$$

As

$$\begin{aligned}
& \left( -\tau \frac{\lambda}{\gamma} + \bar{\beta} x_0^2 + 2 \frac{\mu^2}{\bar{\beta}} + 4\mu x_0 \right)^2 \\
&= \left( -\tau \frac{\lambda}{\gamma} + \bar{\beta} x_0^2 \right)^2 + \left( 2 \frac{\mu^2}{\bar{\beta}} + 4\mu x_0 \right)^2 + \frac{2}{\bar{\beta}} \left( -\tau \bar{\beta} \frac{\lambda}{\gamma} + \bar{\beta}^2 x_0^2 \right) \left( 2 \frac{\mu^2}{\bar{\beta}} + 4\mu x_0 \right) \\
&= \left( -\tau \frac{\lambda}{\gamma} + \bar{\beta} x_0^2 \right)^2 + \left( 2 \frac{\mu^2}{\bar{\beta}} + 4\mu x_0 \right)^2 \\
&\quad + \frac{2}{\bar{\beta}} \left( -\tau \bar{\beta} \frac{\lambda}{\gamma} + (\bar{\beta} x_0 + \mu)^2 \right) \left( 2 \frac{\mu^2}{\bar{\beta}} + 4\mu x_0 \right) - \frac{2}{\bar{\beta}} (\mu^2 + 2\mu \bar{\beta} x_0) \left( 2 \frac{\mu^2}{\bar{\beta}} + 4\mu x_0 \right) \\
&= \left( -\tau \frac{\lambda}{\gamma} + \bar{\beta} x_0^2 \right)^2 + \frac{2}{\bar{\beta}} \left( -\tau \bar{\beta} \frac{\lambda}{\gamma} + (\bar{\beta} x_0 + \mu)^2 \right) \left( 2 \frac{\mu^2}{\bar{\beta}} + 4\mu x_0 \right) \\
&\geq \frac{2}{\bar{\beta}} \left( -\tau \bar{\beta} \frac{\lambda}{\gamma} + (\bar{\beta} x_0 + \mu)^2 \right) \left( 2 \frac{\mu^2}{\bar{\beta}} + 4\mu x_0 \right)
\end{aligned}$$

we observe that  $A_- > 0$ , which allows us to find  $A > 0$ , such that  $V \mapsto \Pi_A(V)$  is greater than a strictly positive constant  $\underline{\pi}$ . Finally, for an appropriate choice of  $A > 0$  following this, we thus obtain

$$\frac{d}{dt} \left( \frac{1}{2} \left( V - \frac{\lambda}{\gamma} \right)^2 + AP + A \frac{\mu}{\bar{\beta}} U \right) \leq -\gamma \left( V - \frac{\lambda}{\gamma} \right)^2 - \underline{\pi} U,$$

By integrating and recalling that  $AP(t) \geq 0$ , we arrive at

$$\begin{aligned}
& \frac{1}{2} \left( V(t) - \frac{\lambda}{\gamma} \right)^2 + A \frac{\mu}{\bar{\beta}} U(t) \leq \frac{1}{2} \left( V(0) - \frac{\lambda}{\gamma} \right)^2 + AP(0) + A \frac{\mu}{\bar{\beta}} U(0) \\
&\leq \frac{1}{2} \left( V_{\text{Init}} - \frac{\lambda}{\gamma} \right)^2 + AP_{\text{Init}} + A \frac{\mu}{\bar{\beta}} U_{\text{Init}} - \underline{\pi} \int_0^t \left\{ \left( V(s) - \frac{\lambda}{\gamma} \right)^2 + U(s) \right\} ds
\end{aligned}$$

where we have written  $\underline{\gamma} = \min(\gamma, \underline{\pi})$ . Hence, we can apply the techniques of section 1.2.7 to conclude that  $\lim_{t \rightarrow \infty} V(t) = \lambda/\gamma$ ,  $\lim_{t \rightarrow \infty} U(t) = 0$ , then  $\lim_{t \rightarrow \infty} P(t) = 0$ . We could also study the linear stability of equilibrium states. The proof of ii) is much more delicate (see [PRÜ 06]) but we can numerically check the validity of this statement.

### *Distribution of sizes; stationary states*

The preceding statements only give information, in the infected case, about the asymptotic macroscopic quantities  $\bar{U}$  and  $\bar{P}$ , but do not tell us anything with regard

to the size distribution of the equilibrium solutions. We notice that the stationary solutions  $(V_\infty, u_\infty)$  of problem [3.29] satisfy

$$\begin{cases} (a) V_\infty \left( \gamma + \int_{x_0}^{\infty} \tau u_\infty(x) dx \right) = \lambda + 2 \int_0^{x_0} x \left( \int_{x_0}^{\infty} \beta(y) \kappa(x, y) u_\infty(y) dy \right) dx, \\ (b) V_\infty \frac{\partial}{\partial x} (\tau u_\infty(x)) = -(\mu + \beta(x)) u_\infty(x) + 2 \int_x^{\infty} \beta(y) \kappa(x, y) u_\infty(y) dy, \end{cases} \quad [3.32]$$

with the boundary condition  $u_\infty(x_0) = 0$ . Figure 3.25 shows the solution in the case where  $\mu$  and  $\tau$  are constant,  $\beta(x) = \bar{\beta}x$  and  $x_0 = 0$ .

We shall see later how to obtain this solution, but this profile does not correspond to recorded observations in which we see several maxima present in the size distribution of the aggregates (multi-modal profile). The idea is to look for such complex profiles by considering size-variable rates  $\tau$ , as has been demonstrated in [CAL 09].

From now on, we adopt the following simplifying hypotheses:

- $\mu$  is constant,
- $\beta(x) = \bar{\beta}x$ ,
- $\kappa(x, y) = \frac{1}{y}$  when  $y > x_0$  and  $0 < x < y$ ,

–  $x_0 = 0$ , this final hypothesis signifying that the unstable aggregates are of negligible size (in the observations we do not see chains containing less than a dozen monomers), and we are interested in the effect of possible variations of  $\tau$  with respect to the size variable. Under these hypotheses, system [3.32] becomes

$$\begin{cases} (a) V_\infty \left( \gamma + \int_0^{\infty} \tau u_\infty(x) dx \right) = \lambda, \\ (b) V_\infty \frac{\partial}{\partial x} (\tau u_\infty(x)) = -(\mu + \bar{\beta}x) u_\infty(x) + 2\bar{\beta} \int_x^{\infty} u_\infty(y) dy, \end{cases} \quad [3.33]$$

with  $u_\infty(0) = 0$ . We thus make the following crucial observations:

- The solutions of [3.33b] are not unique: if, given  $V_\infty, u_\infty$  is a solution, then for all  $z \in \mathbb{R}$ ,  $zu_\infty$  also satisfies [3.33b].

– Equation [3.33b] is sufficient to determine the value of  $V_\infty$ , knowing  $u_\infty$ . Specifically, we multiply [3.33b] by  $x$ , then integrate, using integration by parts,

$$2 \int_0^\infty x \left( \int_x^\infty u_\infty(y) dy \right) dx = \int_0^\infty x^2 u_\infty(x) dx,$$

to finally obtain

$$V_\infty = \frac{\mu \int_0^\infty x u_\infty(x) dx}{\int_0^\infty \tau u_\infty(x) dx}. \quad [3.34]$$

(Note that this value is invariant under exchange of  $u_\infty$  for  $z u_\infty$ .)

The quantity determined by [3.34] does not depend on  $\lambda, \gamma$ ; but, knowing a solution  $u_\infty$  of [3.33b], we choose the solution  $z_\infty u_\infty$  of [3.33b] by imposing

$$z_\infty = \frac{\lambda/V_\infty - \gamma}{\int_0^\infty \tau u_\infty(x) dx}.$$

Thus, if  $u_\infty$  is a solution of [3.33b] normalized by  $\int_0^\infty u_\infty dx = 1$ , then  $z_\infty$  is nothing other than the total number of polymers in the infected state. Such a solution does not exist for all coefficient values. Notably, we must impose the condition  $V_\infty < \lambda/\gamma$  for a meaningful equilibrium to exist.

In order to return to a more familiar framework, we differentiate [3.33b] and we obtain

$$V_\infty \frac{\partial^2}{\partial x^2} (\tau u_\infty(x)) + \frac{\partial}{\partial x} ((\mu + \bar{\beta}x) u_\infty(x)) + 2\bar{\beta} u_\infty(x) = 0, \quad [3.35]$$

with the boundary conditions

$$u_\infty(0) = 0, \quad V_\infty \frac{\partial}{\partial x} (\tau u_\infty)(0) = 2\bar{\beta} \int_0^\infty u_\infty(y) dy. \quad [3.36]$$

In the case that  $\tau$  is constant, we can find an explicit formula for the solution  $u_\infty$ , a solution of

$$V_\infty \tau \partial_{xx}^2 u_\infty(x) + (\mu + \bar{\beta}x) \partial_x u_\infty(x) + 3\bar{\beta} u_\infty = 0.$$

We introduce the change of variable

$$X = x \sqrt{\frac{\bar{\beta}}{\tau V_\infty}} + \frac{\mu}{\sqrt{\bar{\beta} \tau V_\infty}} - 1$$

that is to say

$$x = X \sqrt{\frac{\tau V_\infty}{\bar{\beta}}} - \frac{\mu}{\bar{\beta}} + \sqrt{\frac{\tau V_\infty}{\bar{\beta}}}.$$

We thus set  $\mathcal{U}(X) = u_\infty(x)$ , which leads to  $\partial_x u_\infty(x) = \sqrt{\frac{\bar{\beta}}{\tau V_\infty}} \partial_X \mathcal{U}(X)$ . Thus,  $\mathcal{U}$  satisfies

$$\partial_{XX}^2 \mathcal{U} + (X + 1) \partial_X \mathcal{U} + 3\mathcal{U} = 0.$$

We introduce the function

$$\Phi(X) = (X + X^2/2)$$

and, since  $\partial_{XX}^2 \Phi = 1$ , we rewrite the equation in the form

$$\begin{aligned} -\partial_X(\partial_X \mathcal{U} + \partial_X \Phi \mathcal{U}) &= -\partial_X \left( e^{-\Phi} \partial_X \left( \frac{\mathcal{U}}{e^{-\Phi}} \right) \right) \\ &= -(\partial_{XX}^2 \mathcal{U} + \partial_X \Phi \partial_X \mathcal{U} + \partial_{XX}^2 \Phi \mathcal{U}) = 2\mathcal{U}. \end{aligned}$$

where  $\mathcal{U}$  appears as an eigenfunction, with the eigenvalue 2, of the Fokker–Planck operator  $-\partial_X(e^{-\Phi} \partial_X(\frac{\cdot}{e^{-\Phi}}))$ . We prove, since  $\partial_{XX}^2 \Phi - (\partial_X \Phi)^2 = -2\Phi$ , that  $\mathcal{U} = \Phi e^{-\Phi}$  is a solution of this equation with

$$\mathcal{U}(0) = 0, \quad \partial_X \mathcal{U}(0) = 1.$$

In this particular case, we thus have

$$u_\infty(x) = \mathcal{U} \left( x \sqrt{\frac{\bar{\beta}}{\tau V_\infty}} + \frac{\mu}{\sqrt{\bar{\beta} \tau V_\infty}} - 1 \right).$$

We obtain  $V_\infty$  by exploiting [3.34]. In fact, by directly integrating [3.32b] and taking account of the boundary conditions, we arrive at

$$\mu \int_0^\infty u_\infty dx = \bar{\beta} \int_0^\infty x u_\infty dx.$$

Returning to [3.34], we deduce from this that

$$V_\infty = \frac{\mu^2}{\bar{\beta}\tau}.$$

We have thus completely determined the stationary solution, which is obtained as a dilatation of the function  $\mathcal{U} = \Phi e^{-\Phi}$ .

In the case where  $\tau$  varies as a function of the size  $x$ , we will numerically evaluate the solutions of [3.32] and demonstrate the formation of bimodal profiles. The numerical scheme works in the following manner. As we observed, there is not uniqueness in the solutions of [3.33b], or of [3.35]–[3.36]; we shall thus parameterize the solutions by their integral. We first seek an approximation of the normalized solution, i.e. satisfying  $\int_0^\infty u_\infty(x) dx = 1$ . In particular, the differential in  $x = 0$  in [3.36] is precisely determined by this constraint which prescribes  $\int u_\infty dx$ . Then, the idea is to reinterpret  $x$  as a variable “of time” and to see the equation, with given  $V$ ,

$$\begin{aligned} V \frac{\partial^2}{\partial x^2} (\tau w(x)) + \frac{\partial}{\partial x} ((\mu + \bar{\beta}x)w(x)) + 2\bar{\beta}w(x) &= 0, \\ u_\infty(0) = 0, \quad V \frac{\partial}{\partial x}(\tau w)(0) &= 2\bar{\beta}, \end{aligned} \tag{3.37}$$

as a Cauchy problem of order 2. Finally, to obtain  $V$ , the numerical scheme corresponds to iterations of the mapping

$$\Psi : V \mapsto \Psi(V) = \frac{\int_0^\infty \mu x w dx}{\int_0^\infty \tau w dx},$$

which acts on formula [3.34], where  $w$  is the solution of the stationary problem [3.37] with coefficient  $V$ . The algorithm used corresponds to a discrete version of this process. Let  $h$  be a fixed discretization step size. For fixed  $V^n > 0$ , we define  $(w_i^{n+1})_{i \in \mathbb{N}}$  as a solution of

$$\begin{aligned} \frac{1}{h^2} (\tau_{i+1} w_{i+1}^{n+1} - 2\tau_i w_i^{n+1} + \tau_{i-1} w_{i-1}^{n+1}) \\ + \frac{1}{2h} \frac{1}{V^n} ((\mu + \bar{\beta}x_{i+1})w_{i+1}^{n+1} - (\mu + \bar{\beta}x_{i-1})w_{i-1}^{n+1}) + 2\frac{\bar{\beta}}{V^n} w_i^{n+1} &= 0 \end{aligned}$$

for  $i \in \mathbb{N}$  with

$$w_0^{n+1} = 0, \quad \frac{\tau_1 w_1^{n+1} - \tau_0 w_0^{n+1}}{h} = \frac{2\bar{\beta}}{V^n}.$$

This discrete relation is consistent at order 2 with equation [3.37]. In practice, we must stop the calculation for a certain maximum value of the index  $i$ . As the quantity we are looking for is integrable, we expect (although it is evidently not true for all integrable functions) that it tends towards 0 at infinity and we will work with  $i \in \{0, 1, \dots, I_{\max}\}$ ,  $I_{\max}$  chosen to be “sufficiently large” and we numerically demonstrate that the last iterations do indeed take small values. Thus, having determined  $w^{n+1}$ , we set

$$V^{n+1} = \frac{h\mu \sum_i x_i w_i^{n+1}}{h \sum_i \tau_i w_i^{n+1}},$$

which is a discrete version of [3.34] (in fact, we should rather use the trapeze formula to approximate the integrals with second order accuracy). Even if we do not have proof of convergence, we hope that the scheme converges to the fixed point of  $\Psi$ , which is indeed the solution we are looking for, and this algorithm is designed to give in several iterations an approximation of the normalized stationary solution  $(V_\infty, u_\infty)$ . In order to validate the algorithm, we test it for finding the exact solution in the simple case where  $x_0 = 0$ ,  $\bar{\beta} = 1$ ,  $\mu = 1$ , and  $\tau = 1$ . We start from  $V^0 = 1.8$  and we stop the algorithm when the relative error between two iterations is less than a fixed threshold (here  $\frac{|V^{n+1} - V^n|}{|V^n|} < 10^{-6}$ ). The maximum size is  $X_{\max} = 10$ , and we discretize  $[0, 10]$  with 5,000 points. The approximated solution is obtained in nine iterations. Figure 3.25 compares this numerical solution with the exact solution: the curves are indistinguishable. Figure 3.26 shows the different profiles of the evolution of  $V$  with successive iterations. The algorithm does indeed find the theoretical value of  $V_\infty$ .

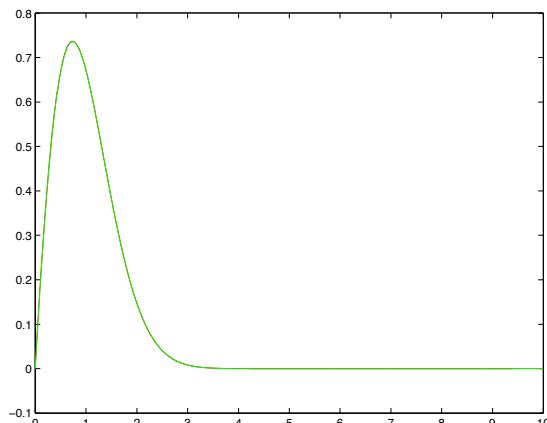
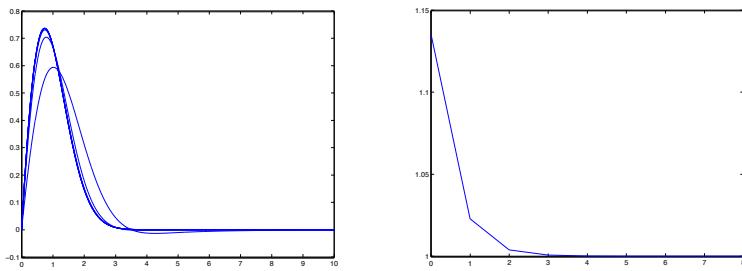


Figure 3.25. Stationary solution for constant  $\tau$

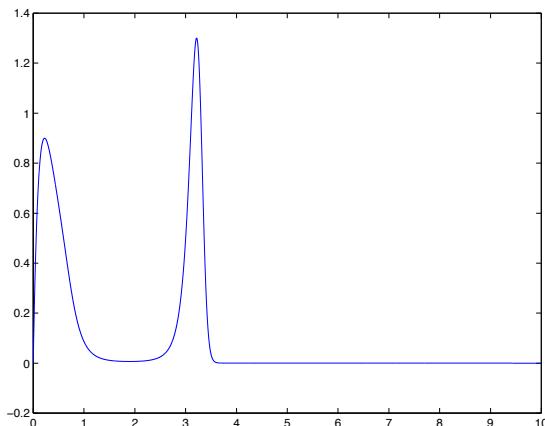


**Figure 3.26.** Stationary solution for constant  $\tau$ : profiles for  $w^n$  (on the left) and the evolution of  $V^n$  (on the right) as a function of the number of iterations

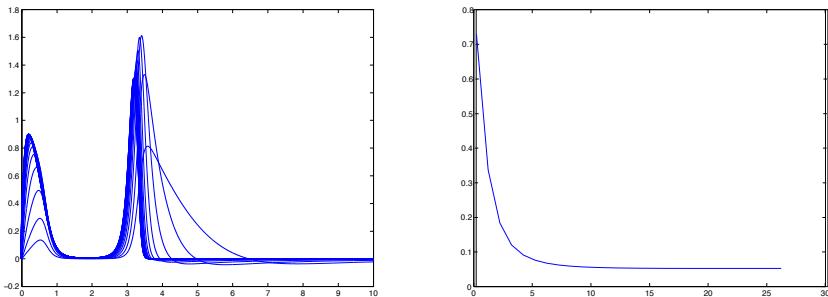
Figure 3.27 displays the solution obtained with the function

$$\tau(x) = \tau_0 + A \exp\left(-\frac{(x-m)^2}{\sigma^2}\right).$$

The values are as follows:  $\bar{\beta} = 0.03$ ,  $\mu = 0.05$ ,  $\tau_0 = 0.1$ ,  $A = 100$ ,  $m = 2$  and  $\sigma = 1/2$ . The algorithm stops after 27 iterations for a relative error on the iterations of  $V$  beneath the threshold  $10^{-4}$  (see Figure 3.28 for the evolution of  $V$  and its profiles). We find  $V_\infty = 0.0527$ . The profile, which we can demonstrate is normalized, does indeed display two “peaks”, more in agreement with the experimental observations.



**Figure 3.27.** Stationary solution for variable  $\tau$



**Figure 3.28.** Stationary solution for variable  $\tau$ : profiles for  $w^n$  (on the left) and the evolution of  $V^n$  (on the right) as a function of the number of iterations

The question of stability of the stationary solutions, in healthy or infected states, is delicate; it is an open problem which is the subject of active research [CAL 09]. We can nevertheless numerically test the evolution of solutions of the simplified system

$$\begin{cases} \frac{d}{dt}V(t) = \lambda - \gamma V(t) - V(t) \int_{x_0}^{\infty} \tau(x)u(t,x) dx + \bar{\beta}x_0^2 \int_{x_0}^{\infty} u(t,y) dy, \\ \frac{\partial}{\partial t}u(t,x) + V(t)\frac{\partial}{\partial x}(\tau(x)u(t,x)) = -(\mu(x) + \beta(x))u(t,x) + 2\bar{\beta} \int_x^{\infty} u(t,y) dy. \end{cases} \quad [3.38]$$

The rate of growth being positive, the upwinding principles lead to the following discretized version

$$\begin{aligned} \frac{V^{n+1} - V^n}{\Delta t} &= \lambda - \gamma V^n - V^n h \sum_{i \geq i_0} \tau_i U_i^n + \bar{\beta} x_0^2 h \sum_{i \geq i_0} U_i^n, \\ \frac{U_i^{n+1} - U_i^n}{\Delta t} + V^n \frac{\tau_i U_i^n - \tau_{i-1} U_{i-1}^n}{\Delta x} &= -(\mu + \beta x_i) U_i^n + 2\bar{\beta} h \sum_{j \geq i} U_j^n. \end{aligned}$$

We carry out simulations with the initial conditions

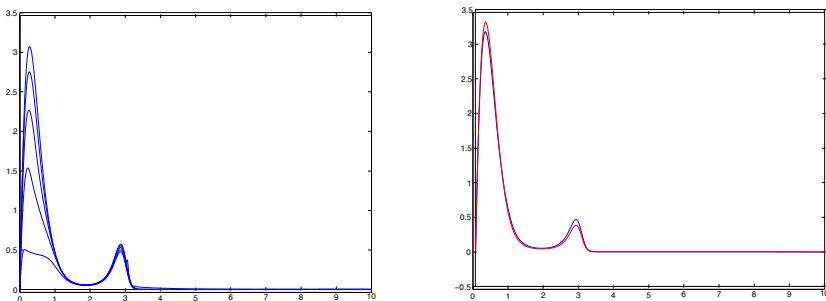
$$U_i^1 = \frac{x_i^2}{2(1+x_i^4)},$$

and setting  $U_1^n = 0$ . The coefficients are  $\bar{\beta} = 0.03$ ,  $\mu = 0.02$ ,  $\lambda = 0.06$  and  $\gamma = 1$ . We test the case where  $\tau$  is variable, with  $\tau_0 = 0.2$  and  $A = 10$ ,  $m = 2$  and  $\sigma = 1/2$ . We work over an interval of maximum size  $X_{\max} = 10$ , until final time  $T_{\max} = 2$ . The step size in space has a value of  $\Delta x = \frac{1}{200}$  and the step size in time is calculated with a CFL number of 0.8. Figure 3.29 shows the profiles of the size distributions at times  $T = 30, 60, 90, 120$  and  $150$ , as well as a comparison with the stationary profile of the same mass (here around 2.24). We find  $V_{\infty} = 0.0268$  both from the

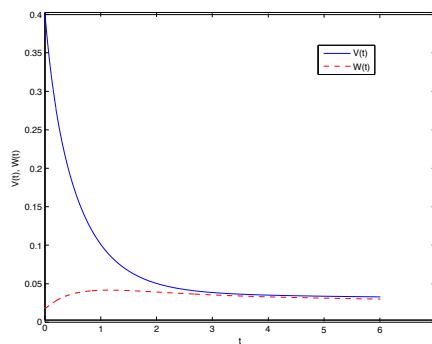
algorithm that looks for a stationary solution, and as the value of  $V$  at final time for the evolution problem. We obtain the same value by evaluating [3.34] at final time. Figure 3.29 shows the comparative evolutions of  $V(t)$  and of

$$W(t) = \frac{\mu \int_0^\infty xu(t, x) dx}{\int_0^\infty \tau u(t, x) dx}$$

which is the quantity given in [3.34]. These two functions do indeed converge towards the same value, which coincides with  $V_\infty$  obtained by looking for stationary solutions. We note that the profile takes more time to develop than  $V$  and  $W$  need to approach their asymptotic value.



**Figure 3.29.** Evolution of the size distribution profile as a function of time for variable  $\tau$  and comparison with the stationary state of the same mass (solution at time  $T = 150$ ). For the color version of this figure, see [www.iste.co.uk/goudon/mathmodel.zip](http://www.iste.co.uk/goudon/mathmodel.zip)



**Figure 3.30.** Evolution of  $V(t)$  (solid line) and  $W(t)$  (dashes). For the color version of this figure, see [www.iste.co.uk/goudon/mathmodel.zip](http://www.iste.co.uk/goudon/mathmodel.zip)

### 3.3. Wave equation

In this section, we are interested in solving the wave equation

$$(\partial_{tt}^2 - c^2 \partial_{xx}^2)u = 0.$$

We will see later how this equation can be interpreted as a linear system of conservation laws. To start, we recall the well-posed nature of this equation.

**THEOREM 3.11.**— Let  $c > 0$ ,  $g \in C^1(\mathbb{R})$  and  $h \in C^0(\mathbb{R})$ . The Cauchy problem

$$(\partial_{tt}^2 - c^2 \partial_{xx}^2)u = 0, \quad u(0, x) = g(x), \quad \partial_t u(0, x) = h(x), \quad [3.39]$$

has a unique solution  $u \in C^1(\mathbb{R} \times \mathbb{R})$ , determined by *D'Alembert's formula*

$$u(t, x) = \frac{1}{2} \left( g(x - ct) + g(x + ct) + \frac{1}{c} \int_{x-ct}^{x+ct} h(y) dy \right). \quad [3.40]$$

**PROOF.**— With  $g$  of class  $C^1$  and continuous  $h$ , the function  $u$  given by [3.40] is of class  $C^1$  over  $\mathbb{R} \times \mathbb{R}$  and its derivatives satisfy

$$\partial_t u(t, x) = \frac{c}{2} (g'(x + ct) - g'(x - ct)) + \frac{1}{2} (h(x + ct) + h(x - ct)),$$

$$\partial_x u(t, x) = \frac{1}{2} (g'(x + ct) + g'(x - ct)) + \frac{1}{2c} (h(x + ct) - h(x - ct)).$$

In particular, we have  $u(0, x) = g(x)$  and  $\partial_t u(0, x) = h(x)$ . It thus follows that

$$(\partial_t + c\partial_x)u(t, x) = cg'(x + ct) + h(x + ct),$$

$$(\partial_t - c\partial_x)u(t, x) = -cg'(x - ct) + h(x - ct).$$

If  $g$  and  $h$  are more regular, for example,  $C^2$  and  $C^1$ , respectively, we can differentiate once more and demonstrate that  $(\partial_t - c\partial_x)(\partial_t + c\partial_x)u(t, x) = 0 = (\partial_t + c\partial_x)(\partial_t - c\partial_x)u(t, x) = (\partial_{tt}^2 - c^2 \partial_{xx}^2)u(t, x)$ . In order to work in the  $C^1$  framework, we must interpret the derivatives in a weak sense. Let  $\varphi \in C_c^\infty(\mathbb{R} \times \mathbb{R})$ . We calculate

$$\begin{aligned} \langle (\partial_{tt}^2 - c^2 \partial_{xx}^2)u, \varphi \rangle &= -\langle (\partial_t + c\partial_x)u, (\partial_t - c\partial_x)\varphi \rangle \\ &= -\iint_{\mathbb{R} \times \mathbb{R}} [cg'(x + ct) + h(x + ct)](\partial_t - c\partial_x)\varphi(t, x) dx dt. \end{aligned}$$

We use the change of variables  $X = x + ct$ ,  $T = x - ct$  whose Jacobian is

$$\begin{pmatrix} \partial_x X & \partial_t X \\ \partial_x T & \partial_t T \end{pmatrix} = \begin{pmatrix} 1 & c \\ 1 & -c \end{pmatrix}$$

such that  $dX dT = 2c dx dt$ . We also introduce  $\psi(T, X) = \varphi(t, x)$ , which satisfies

$$\begin{aligned} \partial_t \varphi(t, x) &= \partial_t X \partial_X \psi(T, X) + \partial_t T \partial_T \psi(T, X) \\ &= c \partial_X \psi(T, X) - c \partial_T \psi(T, X), \\ \partial_x \varphi(t, x) &= \partial_x X \partial_X \psi(T, X) + \partial_x T \partial_T \psi(T, X) \\ &= \partial_X \psi(T, X) + \partial_T \psi(T, X), \end{aligned}$$

and  $(\partial_t - c \partial_x) \varphi(t, x) = -2c \partial_T \psi(T, X)$ . We thus arrive at

$$\begin{aligned} \langle (\partial_{tt}^2 - c^2 \partial_{xx}^2) u, \varphi \rangle &= 4c^2 \iint_{\mathbb{R} \times \mathbb{R}} [cg'(X) + h(X)] \partial_T \psi(T, X) dX dT \\ &= 4c^2 \int_{\mathbb{R}} [cg'(X) + h(X)] \left( \int_{\mathbb{R}} \partial_T \psi(T, X) dT \right) dX = 0, \end{aligned}$$

using the fact that  $\psi$  also has compact support. We have shown that  $u$  is a solution of [3.39]. Finally, we remark that

if  $\text{supp}(g), \text{supp}(h) \subset [-R, R]$ , then  $\text{supp}(u(t, \cdot)) \subset [-R - c|t|, R + c|t|]$ . [3.41]

This fundamental property indicates that the wave equation propagates information at a finite speed (like the transport equation and unlike the heat equation).

Now assume that we have two solutions of [3.39], in the weak sense studied previously. Then, their difference  $v(t, x) \in C^1(\mathbb{R} \times \mathbb{R})$  satisfies [3.39] with  $g = 0 = h$ . We shall show that  $v$  is identically zero. Let us have fixed  $T \in \mathbb{R}$  and  $\zeta \in C_c^\infty(\mathbb{R})$ . We set

$$\psi(t, x) = \frac{1}{2c} \int_{x-c(t-T)}^{x+c(t-T)} \zeta(y) dy.$$

In other words,  $\psi \in C^\infty(\mathbb{R} \times \mathbb{R})$  is a solution of  $(\partial_{tt}^2 - c^2 \partial_{xx}^2)\psi = 0$ ,  $\psi(T, x) = 0$  and  $\partial_t \psi(T, x) = \zeta(x)$ . We set

$$I(t) = \int_{\mathbb{R}} \partial_t v(t, x) \psi(t, x) dx - \int_{\mathbb{R}} v(t, x) \partial_t \psi(t, x) dx.$$

This quantity is well defined since, by assuming  $\text{supp}(\zeta) \subset [-R, +R]$ , for all fixed  $t$ ,  $x \mapsto \psi(t, x)$  is supported in  $[-R - c|t - T|, R + c|t - T|]$ . We even have  $I \in C^0(\mathbb{R})$  with  $I(0) = 0$  and  $I(T) = -\int v(T, x)\zeta(x) dx$ . We shall finish by showing that  $I$  is identically null. Let  $\theta \in C_c^\infty(\mathbb{R})$ . By using integrations by parts, we evaluate

$$\begin{aligned} \int_{\mathbb{R}} I(t)\theta'(t) dt &= \iint_{\mathbb{R} \times \mathbb{R}} \partial_t v(t, x)\psi(t, x)\theta'(t) dx dt - \iint_{\mathbb{R} \times \mathbb{R}} v(t, x)\partial_t \psi(t, x)\theta'(t) dx dt \\ &= \iint_{\mathbb{R} \times \mathbb{R}} \partial_t v(t, x)\partial_t(\psi(t, x)\theta(t)) dx dt - \iint_{\mathbb{R} \times \mathbb{R}} \partial_t v(t, x)\partial_t \psi(t, x)\theta(t) dx dt \\ &\quad - \iint_{\mathbb{R} \times \mathbb{R}} v(t, x)\partial_t \psi(t, x)\theta'(t) dx dt \\ &= \iint_{\mathbb{R} \times \mathbb{R}} \partial_t v(t, x)\partial_t(\psi(t, x)\theta(t)) dx dt + \iint_{\mathbb{R} \times \mathbb{R}} v(t, x)\partial_t(\partial_t \psi(t, x)\theta(t)) dx dt \\ &\quad - \iint_{\mathbb{R} \times \mathbb{R}} v(t, x)\partial_t \psi(t, x)\theta'(t) dx dt \\ &= \iint_{\mathbb{R} \times \mathbb{R}} \partial_t v(t, x)\partial_t(\psi(t, x)\theta(t)) dx dt + \iint_{\mathbb{R}} v(t, x)\partial_{tt}^2 \psi(t, x)\theta(t) dx dt \\ &= \iint_{\mathbb{R} \times \mathbb{R}} \partial_t v(t, x)\partial_t(\psi(t, x)\theta(t)) dx dt + \iint_{\mathbb{R}} v(t, x)c^2 \partial_{xx}^2 \psi(t, x)\theta(t) dx dt \\ &= \iint_{\mathbb{R} \times \mathbb{R}} (\partial_t - c\partial_x)v(t, x)(\partial_t + c\partial_x)(\psi(t, x)\theta(t)) dx dt = 0. \end{aligned}$$

We can apply this reasoning again to

$$\theta(t) = \int_0^t \left( h(\tau) - \kappa(\tau) \int_0^T h(s) ds \right) d\tau$$

where  $h$  and  $\kappa$  are continuous functions, supported in  $[0, T]$ , additionally with  $\int_0^T \kappa(s) ds = 1$ . In particular,  $\theta$  is  $C^1$  over  $[0, T]$  (without loss of generality, we assume here that  $T > 0$ ) and satisfies  $\theta(0) = 0$  as well as  $\theta(T) = 0$  (thanks to the normalization hypothesis over  $\kappa$ ), which justifies the integrations by parts carried out above. As  $\theta'(t) = h(t) - \kappa(t) \int_0^T h(s) ds$ , we thus obtain

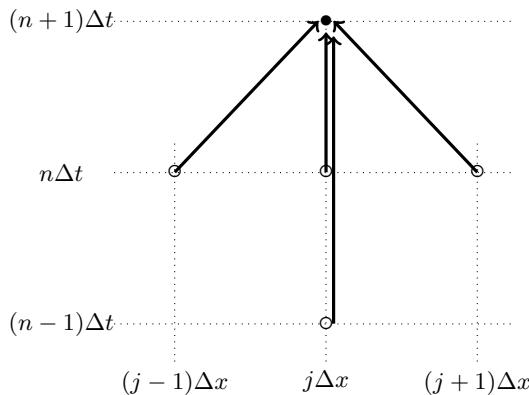
$$\int_{\mathbb{R}} I(t)\theta'(t) dt = \int_0^T I(t)\theta'(t) dt = 0 = \int_0^T \left( I(t) - \int_0^T I(s)\kappa(s) ds \right) h(t) dt.$$

With this being satisfied for all test functions  $h$ , we deduce that  $I(t)$  is constant over  $[0, T]$  with  $I(0) = 0$ , thus  $I(t) = 0$  over  $[0, T]$ , and finally  $v(t, x) = 0$ .  $\square$

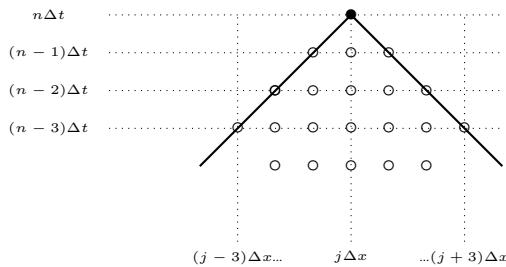
The property of finite speed of propagation, see [3.41], also plays a fundamental role in the stability analysis of numerical schemes for the wave equation. Formula [3.40] shows that the evaluation of the solution  $u$  at time  $t$  and position  $x$  depends only on the values taken by the data in the interval, called the *light cone*,  $\mathcal{C}(t, x) = [x - c|t|, x + c|t|]$ . A natural finite difference scheme for the wave equation is expressed as

$$\frac{1}{\Delta t^2} (u_j^{n+1} - 2u_j^n + u_j^{n-1}) = \frac{c^2}{\Delta x^2} (u_{j+1}^n - 2u_j^n + u_{j-1}^n). \quad [3.42]$$

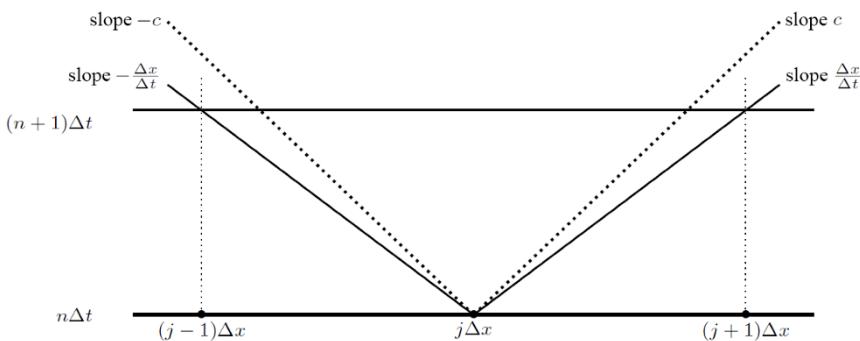
To initiate the scheme, we use  $u_j^0 = g(j\Delta x)$  and  $\frac{u_j^0 - u_j^{-1}}{\Delta t} = h(j\Delta x)$ . Thus, the approximated solution at time  $(n+1)\Delta t$  and position  $j\Delta x$  depends on the approximated solution at times  $n\Delta t$  and  $(n-1)\Delta t$  and neighboring positions  $(j \pm 1)\Delta x$  and  $j\Delta x$  (see Figure 3.31). By recurrence, we thus establish that  $u_j^n$  depends only on the initial state  $u_k^0$  and  $u_k^{-1}$  for indices  $k$ , such that  $|j - k| \leq n$ . In other words,  $u_j^n$  depends only on the values taken by data in the interval  $C_j^n = [(j-n)\Delta x, (j+n)\Delta x]$  (see Figure 3.32). Thus, to obtain results consistent with the continuous problem, we must impose  $\mathcal{C}(n\Delta t, j\Delta x) \subset C_j^n$ : if not, we could affect the exact solution at  $(n\Delta t, j\Delta x)$  with a modification of the initial data at the boundaries of  $\mathcal{C}(n\Delta t, j\Delta x)$  without the approximated solution being changed. This constraint leads to the condition  $c\Delta t \leq \Delta x$ . Geometrically, this means that the wave starting from  $j\Delta x$  does not reach, in a time  $\Delta t$ , the boundaries of the discretization interval  $[(j-1)\Delta x, (j+1)\Delta x]$  (see Figure 3.33).



**Figure 3.31.** Determination of the numerical solution at  $(n+1)\Delta t$  and  $j\Delta x$



**Figure 3.32.** Light cone of the numerical solution

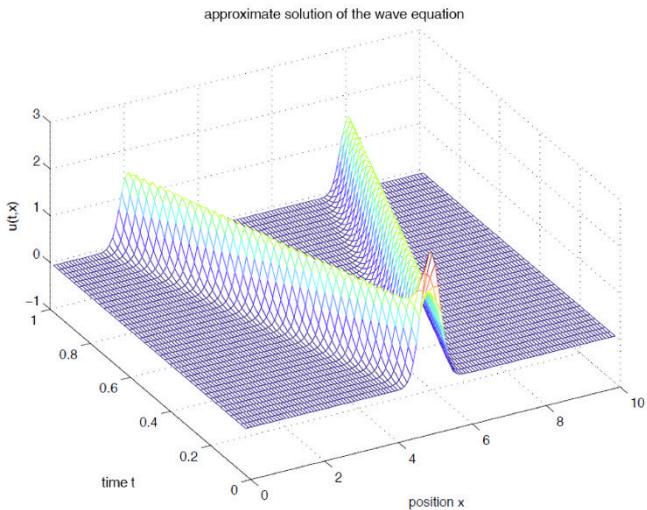


**Figure 3.33.** CFL condition for the wave equation: the “numerical speed”  $\frac{\Delta x}{\Delta t}$  must be greater than the speed of propagation  $c$ , and the numerical cone contains the physical cone

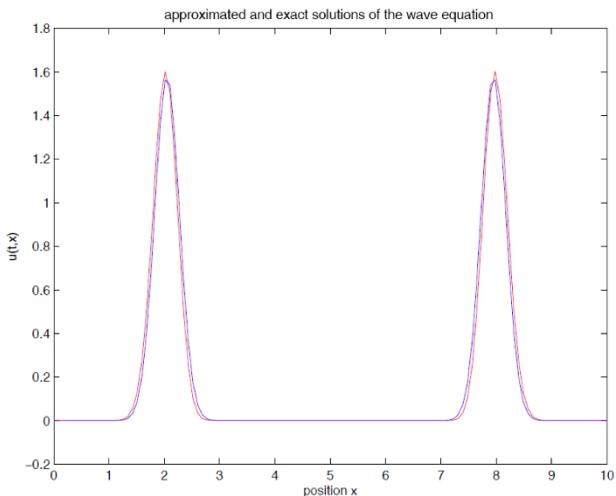
These concepts, demonstrated by R. Courant, K. Friedrichs and H. Lewy in 1928 [COU 28], are illustrated in Figures 3.34–3.39, where we present the numerical results obtained with and without the condition  $c\Delta t \leq \Delta x$  being satisfied. For these numerical tests,  $c = 3$  and the initial data are

$$g(x) = \sqrt{10} \exp(-10|x - 5|^2), h(x) = 0.$$

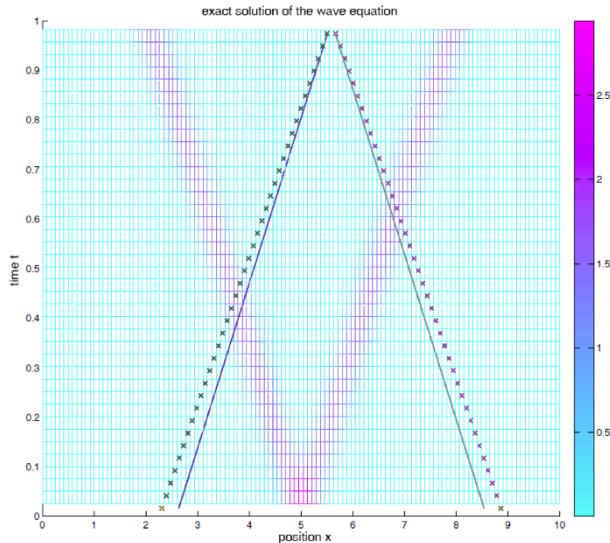
We carry out the simulation over the domain  $[0, 10]$ , supplementing the wave equation with homogeneous Dirichlet conditions  $u(t, 0) = u(t, 10) = 0$ . The final time here is  $T = 0.7$ . In particular, this time is short enough that the core initial information, concentrated at the position  $x = 5$ , interacts only negligibly with the boundary and the solution obtained corresponds to the solution of the problem posed over the whole of  $\mathbb{R}$ . When the stability condition is not satisfied, we see parasitic oscillations appear, which grow with time and rapidly render the numerical solution incoherent.



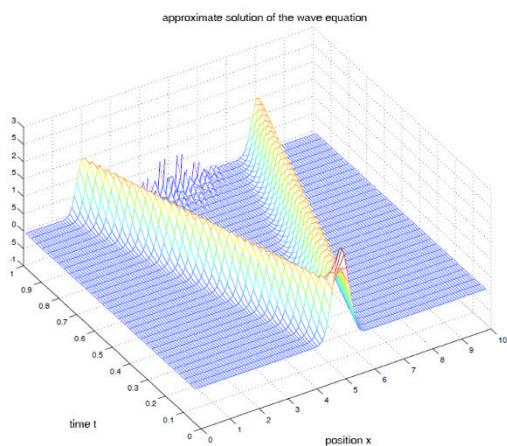
**Figure 3.34.** Simulation of the wave equation under the stability constraint ( $c\Delta t/\Delta x = .9$ ). For the color version of this figure, see [www.iste.co.uk/goudon/mathmodel.zip](http://www.iste.co.uk/goudon/mathmodel.zip)



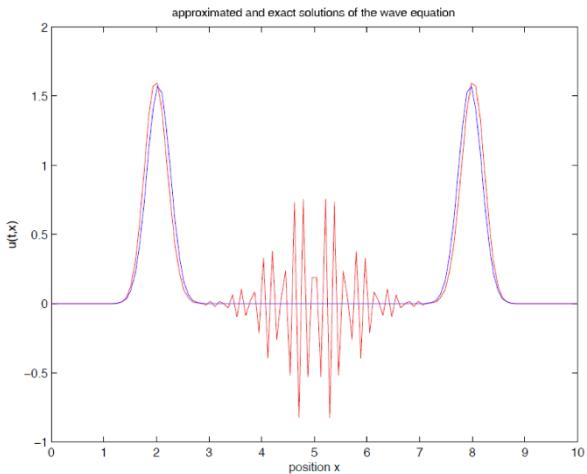
**Figure 3.35.** Simulation of the wave equation under the stability constraint ( $c\Delta t/\Delta x = .9$ ). Comparison with the exact solution at final time  $T = .7$ . For the color version of this figure, see [www.iste.co.uk/goudon/mathmodel.zip](http://www.iste.co.uk/goudon/mathmodel.zip)



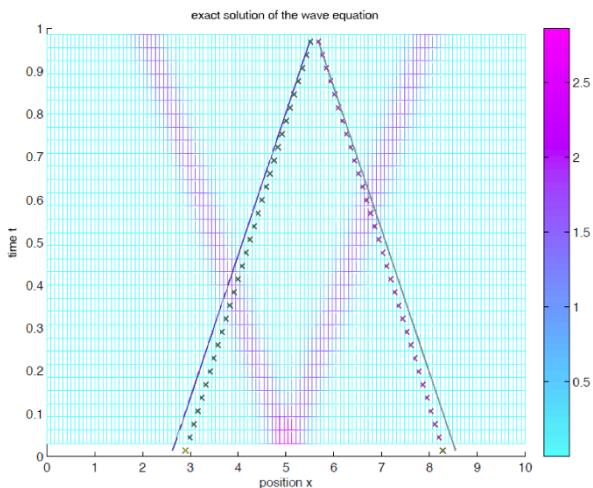
**Figure 3.36.** Simulation of the wave equation under the stability constraint ( $c\Delta t/\Delta x = .9$ ). Representation of the theoretical light cone (solid line) and the numerical light cone (crosses). For the color version of this figure, see [www.iste.co.uk/goudon/mathmodel.zip](http://www.iste.co.uk/goudon/mathmodel.zip)



**Figure 3.37.** Simulation of the wave equation beyond the stability constraint ( $c\Delta t/\Delta x = 1.1$ ). For the color version of this figure, see [www.iste.co.uk/goudon/mathmodel.zip](http://www.iste.co.uk/goudon/mathmodel.zip)



**Figure 3.38.** Simulation of the wave equation beyond the stability constraint ( $c\Delta t/\Delta x = 1.1$ ). Comparison with the exact solution at final time  $T = .7$ . For the color version of this figure, see [www.iste.co.uk/goudon/mathmodel.zip](http://www.iste.co.uk/goudon/mathmodel.zip)



**Figure 3.39.** Simulation of the wave equation beyond the stability constraint ( $c\Delta t/\Delta x = 1.1$ ). Representation of the theoretical light cone (solid line) and the numerical light cone (crosses). For the color version of this figure, see [www.iste.co.uk/goudon/mathmodel.zip](http://www.iste.co.uk/goudon/mathmodel.zip)

Another approach to the wave equation consists of writing it in the form of a system

$$\partial_t u + c \partial_x v = 0, \quad \partial_x v + c \partial_x u = 0.$$

Indeed, by taking the derivative in time of the first equation and the derivative in space of the second equation, we recover that  $u$  is a solution of  $(\partial_{tt}^2 - c^2 \partial_{xx}^2)u = 0$ . We set  $U(t, x) = (u(t, x), v(t, x))$  and the wave equation thus appears as a linear system of conservation laws

$$\partial_t U + W \partial_x U = 0, \quad W = \begin{pmatrix} 0 & c \\ c & 0 \end{pmatrix}.$$

We note that the Fourier transform (in space) of  $U(t, \cdot)$  satisfies

$$\partial_t \widehat{U}(t, \xi) + A(\xi) \widehat{U}(t, \xi) = 0, \quad A(\xi) = \begin{pmatrix} 0 & ic\xi \\ ic\xi & 0 \end{pmatrix}.$$

The eigenvalues of  $A(\xi)$  are  $\pm ic\xi$  and the eigenspaces are spanned by  $(1, \pm 1)$ . As

$$P^{-1} A(\xi) P = \begin{pmatrix} ic\xi & 0 \\ 0 & -ic\xi \end{pmatrix}, \quad P = \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix},$$

we introduce  $V(t, \xi) = P \widehat{U}(t, \xi)$ , which satisfies the diagonal system

$$\partial_t V(t, \xi) = \begin{pmatrix} ic\xi & 0 \\ 0 & -ic\xi \end{pmatrix} V(t, \xi).$$

We represent the components of the inverse Fourier transform  $V(t, \xi)$  with  $w_+(t, x)$  and  $w_-(t, x)$ ; they are thus solutions of the transport equations with velocities  $\pm c$ :

$$\partial_t w_{\pm} \pm c \partial_x w_{\pm} = 0$$

and we recover  $U(t, x)$  by applying  $P^{-1}$  to the vector  $(w_+(t, x), w_-(t, x))$ . We thus retrieve  $u = (w_+ + w_-)/2$  and  $v = (w_+ - w_-)/2$  as a combination of waves propagating at speeds  $\pm c$ . In an equivalent manner, we relate the equation for  $(w_+, w_-)$  to that satisfied by  $U = (u, v)$  by diagonalizing the matrix  $W$ .

This approach motivates the design of a scheme based on the upwind discretization of the transport equations satisfied by  $w_{\pm}$ , at positive/negative speeds. Specifically, we set

$$\begin{aligned} w_{+,j}^{n+1} &= w_{+,j}^n - \frac{c\Delta t}{\Delta x}(w_{+,j}^n - w_{+,j-1}^n), \\ w_{-,j}^{n+1} &= w_{-,j}^n + \frac{c\Delta t}{\Delta x}(w_{-,j+1}^n - w_{-,j}^n). \end{aligned}$$

Thus, the corresponding scheme for  $u, v$  takes the following form

$$\begin{aligned} u_j^{n+1} &= (w_{+,j}^{n+1} + w_{-,j}^{n+1})/2 \\ &= u_j^n - \frac{c\Delta t}{2\Delta x}(v_{j+1}^n - v_{j-1}^n) + \frac{c\Delta t}{2\Delta x}(u_{j+1}^n - 2u_j^n + u_{j-1}^n) \\ v_j^{n+1} &= (w_{+,j}^{n+1} - w_{-,j}^{n+1})/2 \\ &= v_j^n - \frac{c\Delta t}{2\Delta x}(u_{j+1}^n - u_{j-1}^n) + \frac{c\Delta t}{2\Delta x}(v_{j+1}^n - 2v_j^n + v_{j-1}^n). \end{aligned}$$

We thus recover a scheme of the Lax–Friedrichs type, with a centered finite difference to approximate the spatial derivative  $\partial_x$  and the addition of a numerical diffusion of order  $\Delta x$ . A comparison with the finite difference scheme is given in Figure 3.40: exact and FD solutions are practically indistinguishable, but the Lax–Friedrichs scheme displays a stronger numerical diffusion (underestimation of the extrema and spreading of the solution).

## 3.4. Nonlinear problems: conservation laws

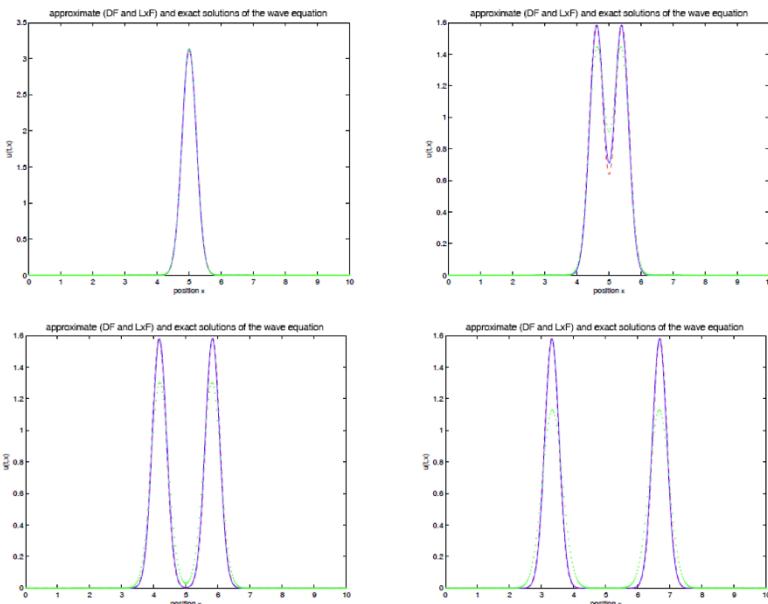
### 3.4.1. Scalar conservation laws

Evidently, the numerical simulation of the linear transport equation with constant coefficients is of limited interest in itself, since we know the exact solution! The motivation for the discussion put forward in section 3.2.4 is, precisely because we have a simple explicit expression of the solution, on the one hand, to build up an intuition for constructing reliable methods, and, on the other hand, to lay down a theoretical basis and put in place tools for analysis of the numerical schemes. We can then extend this outlook for nonlinear situations, which are more relevant. We start by focusing on the case where the unknown  $u$  is a function with *scalar* values.

#### 3.4.1.1. Elements of analysis

The remarkable fact arising from the analysis of solutions of conservation laws is the potential for singularities to appear in finite time: even if the initial data are infinitely smooth, the solution can become discontinuous in finite time. This

phenomenon is the exact opposite of the situation demonstrated for the diffusion equations. Indeed, we saw for the heat equation that, although the initial data can be discontinuous, the solution is instantaneously  $C^\infty$ . For the linear transport equation, the regularity is conserved: initial  $C^k$  data produce a solution of class  $C^k$ . For scalar conservation laws, in general,  $C^\infty$  data produce a discontinuous solution! Evidently, this phenomenon makes it difficult to give meaning to the equation when the solution is not even continuous: we must work within a framework of *weak solutions*. It is also important to highlight another fundamental difference with the heat equation. The heat equation *propagates information at infinite speed*. Indeed, the convolution relation shows that at all points  $x$  and at each point in time, the solution depends on the whole of the initial data, independently of its support. For the conservation laws, as for the transport equation and the wave equation, the information is propagated at a finite speed and the solution at a position  $x$  and a time  $t$  depends only on the values of the initial data within a limited domain.



**Figure 3.40.** Simulation of the wave equation: comparison of the Lax–Friedrichs-type scheme and the finite difference scheme. For the color version of this figure, see [www.iste.co.uk/goudon/mathmodel.zip](http://www.iste.co.uk/goudon/mathmodel.zip)

We can try to reproduce the argument held for the linear transport equation by defining the solution of

$$\partial_t \rho + \partial_x f(\rho) = 0 \quad [3.43]$$

by integration along the characteristic curves. These are defined by the differential equation

$$\frac{d}{ds} X(s; t, x) = f'(\rho(s, X(s; t, x))), \quad X(t; t, x) = x. \quad [3.44]$$

Thus,  $X(s; t, x)$  is the position at time  $s$  of a particle occupying the position  $x$  at time  $t$  and subject to the velocity field  $(t, x) \mapsto f'(\rho(t, x))$ . The chain rule again leads to

$$\begin{aligned} \frac{d}{ds} [\rho(s, X(s; t, x))] &= \nabla_{t,x} \rho(s, X(s; t, x)) \cdot \frac{d}{dt} \left( \begin{matrix} s \\ X(s; t, x) \end{matrix} \right) \\ &= (\partial_t \rho)(s, X(s; t, x)) + \partial_s X(s; t, x) (\partial_x \rho)(s, X(s; t, x)) \\ &= (\partial_t \rho + f'(\rho) \partial_x \rho)(s, X(s; t, x)) = 0. \end{aligned}$$

We thus arrive at

$$\rho(t, x) = \rho_{\text{Init}}(X(0; t, x)), \quad [3.45]$$

where  $\rho_{\text{Init}}$  is the initial data, for  $t = 0$ , of problem [3.43]. However, in contrast to the linear situation, expression [3.45] is now of *implicit* nature, since the characteristics themselves depend on the solution  $\rho$  via formula [3.44]. We can nevertheless exploit the fact that the solution is conserved along the characteristic curves  $\rho(t, X(t; 0, x)) = \rho_{\text{Init}}(x)$  to rewrite [3.44] in the form

$$\frac{d}{dt} X(t; 0, x) = f'(\rho(t, X(t; 0, x))) = f'(\rho_{\text{Init}}(x)), \quad X(0; 0, x) = x.$$

It follows that

$$X(t; 0, x) = x + t f'(\rho_{\text{Init}}(x)).$$

For fixed  $t > 0$ , we use  $\phi_t$  to signify the function  $x \mapsto \phi_t(x) = x + t f'(\rho_{\text{Init}}(x))$ . To express  $\rho(t, x)$  as a function of the initial data  $\rho_{\text{Init}}$ , we must determine the inverse of the function  $\phi_t$ , which gives  $X(0; t, x)$ , the position at time 0 of a particle leaving  $x$  at time  $t$ . We note that  $\phi'_t(x) = 1 + t f''(\rho_{\text{Init}}(x)) \rho'_{\text{Init}}(x)$  does not necessarily have a constant sign. Thus,  $\phi_t$  has an inverse defined over  $\mathbb{R}$  when  $f$  is convex and  $\rho_{\text{Init}}$  is monotonically increasing (since, in this case, we have  $\phi'_t(x) > 0$ , so  $x \mapsto \phi_t(x)$  is strictly monotonically increasing), but as a generality there can exist times  $t > 0$ , for which  $\phi_t$  is not invertible. Formulae [3.44] and [3.45] then no longer have meaning.

Let us now consider another approach that reveals this phenomenon of loss of regularity. We set  $v(t, x) = \partial_x \rho(t, x)$ . By differentiating [3.43], we obtain

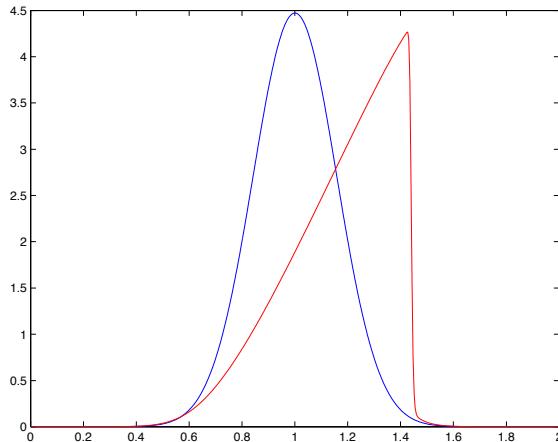
$$\partial_t v + f'(\rho) \partial_x v = -f''(\rho) v^2.$$

We restrict ourselves to the case of Burgers' equation, where  $f(\rho) = \rho^2/2$ . Thus,  $f''(\rho) = 1$  and, along the characteristics,  $v$  satisfies the Riccati equation:  $V(t) = v(t, X(t; 0, x))$  satisfying

$$\frac{d}{dt} V = -V^2.$$

The solutions of this equation are of the form  $V(t) = \frac{1}{t+1/V(0)}$ . In other words, we have

$$v(t, x) = \left( t + \frac{1}{\partial_x \rho_{\text{Init}}(X(0; t, x))} \right)^{-1}. \quad [3.46]$$



**Figure 3.41.** Loss of regularity of solutions of Burgers' equation. For the color version of this figure, see [www.iste.co.uk/goudon/mathmodel.zip](http://www.iste.co.uk/goudon/mathmodel.zip)

However, the solutions of the Riccati equation are not globally defined (the function  $x^2$  not being uniformly Lipschitzian): if  $V(0) < 0$ , then there exists  $0 < T < \infty$ , such that  $V(t) = \frac{1}{t+1/V(0)} \rightarrow \infty$  as  $t \rightarrow T$ . Thus, when  $\rho_{\text{init}}$  is decreasing, formula [3.46] does not have any meaning for all times. Figure 3.41 illustrates this phenomenon. It shows the solution of Burgers' equation at a positive time, starting from regular, but not monotonic, data (specifically with a Gaussian

profile). We note that the solution exhibits a discontinuity at position  $x \simeq 1.43$ . In particular, we note that the solution  $\rho(t, x)$  remains bounded; by contrast, its derivative in space  $\partial_x \rho(t, x)$  tends to become infinite at that position and time. We thus understand the fault of the analysis by characteristics: with the term on the right-hand side of [3.44] not being Lipschitzian, the hypotheses of the Cauchy–Lipschitz theorem fail and the characteristic curves cannot be defined for all  $(t, x)$ .

Even if we consider regular initial data, it is thus absolutely unrealistic to expect the solution of [3.43] to remain regular for all times. We must therefore work in a framework of weak solutions.

**DEFINITION 3.4.–** We say that a function  $\rho \in L^1_{\text{loc}}([0, \infty] \times \mathbb{R})$  is a *weak solution* of [3.43] associated with the initial data  $\rho_{\text{Init}}$  if  $f(\rho) \in L^1_{\text{loc}}([0, \infty] \times \mathbb{R})$  and, for all test functions  $\varphi \in C_c^1([0, \infty] \times \mathbb{R})$ , we have

$$-\int_0^\infty \int_{\mathbb{R}} (\rho \partial_t \phi + f(\rho) \partial_x \phi)(t, x) dx dt - \int_{\mathbb{R}} \rho_{\text{Init}}(x) \phi(0, x) dx = 0. \quad [3.47]$$

We note that the integrals that appear in [3.47] themselves have meaning even though  $\rho$  is potentially discontinuous. This formula comes from multiplying [3.43] by  $\varphi$  and proceeding by integration by parts. This concept thus allows us to consider discontinuous solutions of [3.43]. Actually, we can characterize the discontinuities of the solutions of [3.43]. Indeed, let us assume that  $(t, x) \mapsto \rho(t, x)$  is a weak solution of [3.43], of class  $C^1$  on both sides of a curve of discontinuity  $\Gamma = \{(x, t) \in \mathbb{R} \times [0, \infty[, x = s(t)\}$ . The function  $t \mapsto s(t)$  characterizes the propagation of the singularity. We write  $\Omega^- = \{(x, t) \in \mathbb{R} \times [0, \infty[, x < s(t)\}$  and  $\Omega^+ = \{(x, t) \in \mathbb{R} \times [0, \infty[, x > s(t)\}$ . We represent with  $\nu^\pm = (\nu_x^\pm, \nu_t^\pm)$  the vector normal to  $\Gamma$ , pointing outward of  $\Omega^\pm$ . More specifically, we have

$$\nu^- = \frac{1}{\sqrt{1 + |\dot{s}(t)|^2}} \begin{pmatrix} 1 \\ -\dot{s}(t) \end{pmatrix}, \quad \nu^+ = -\nu^-.$$

Let  $\varphi$  be a test function whose support is strictly enclosed within  $]0, \infty[ \times \mathbb{R}$ . As  $\rho$  satisfies [3.47], we have

$$\begin{aligned} 0 &= -\iint (\rho \partial_t \phi + f(\rho) \partial_x \phi) dx dt = -\iint_{\Omega^-} \dots dx dt - \iint_{\Omega^+} \dots dx dt \\ &= \iint_{\Omega^-} (\partial_t \rho + \partial_x f(\rho)) \phi dx dt + \iint_{\Omega^+} (\partial_t \rho + \partial_x f(\rho)) \phi dx dt \\ &\quad - \int_{\Gamma^-} (\rho \nu_t^- + f(\rho) \nu_x^-) \phi d\gamma - \int_{\Gamma^+} (\rho \nu_t^+ + f(\rho) \nu_x^+) \phi d\gamma. \end{aligned}$$

We have used  $\Gamma^\pm$  here to represent the boundaries of  $\Omega^\pm$ . As  $\rho$  is a *regular* solution of [3.43] in the domains  $\Omega^\pm$ , we have

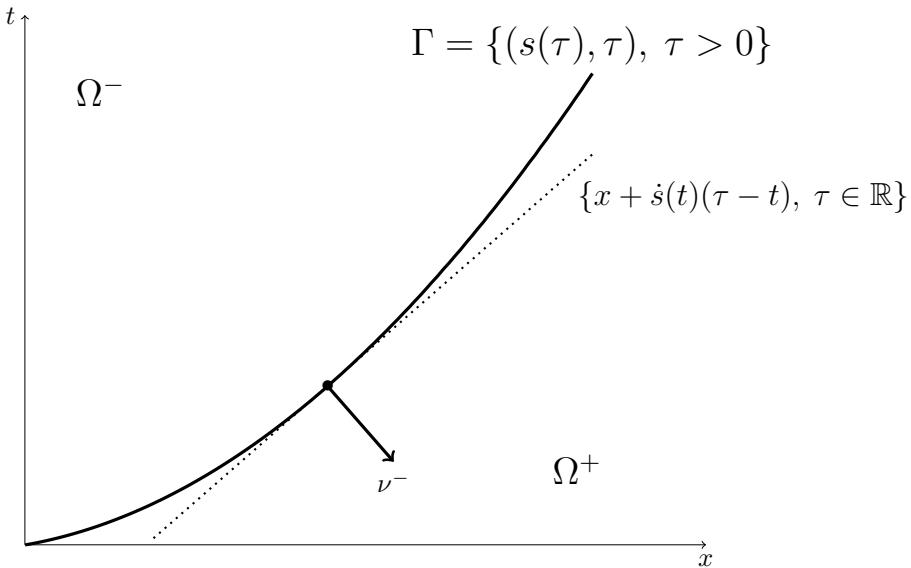
$$\iint_{\Omega^\pm} (\partial_t \rho + \partial_x f(\rho)) \phi \, dx \, dt = 0.$$

Taking account of the fact that  $\Gamma^\pm = \Gamma$  with  $\nu^- = -\nu^+$ , we arrive at

$$\begin{aligned} 0 &= - \iint (\rho \partial_t \phi + f(\rho) \partial_x \phi) \, dx \, dt \\ &= \int_{\Gamma} [(\rho^+ - \rho^-) \nu_t^- + (f(\rho^+) - f(\rho^-)) \nu_x^-] \phi \, d\gamma \end{aligned}$$

where  $\rho^\pm$  signifies the limit of  $\rho$  over  $\Gamma$  coming from  $\Omega^\pm$ . We deduce from this that the following relation – called the *Rankine–Hugoniot relation* – between jumps in  $\rho$  and in  $f(\rho)$ :

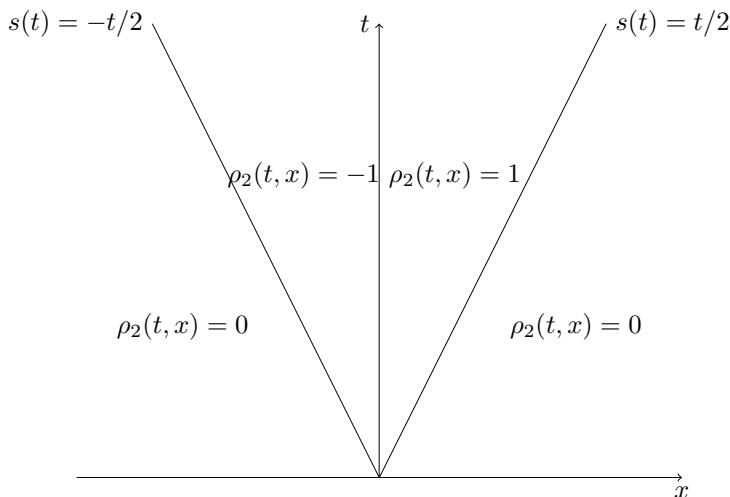
$$(\rho^+ - \rho^-) \dot{s} - (f(\rho^+) - f(\rho^-)) = [\![\rho]\!] \dot{s} - [\![f(\rho)]!] = 0.$$



**Figure 3.42.** Curve of discontinuity and Rankine–Hugoniot relations

The identification of this phenomenon of loss of regularity dates back to the middle of the 19th century when Challis demonstrated the paradox that the solutions

of equations of the dynamics of isentropic gases (which we will meet later) can take different values at one point at certain times [CHA 48]. Stokes then established the jump relations [STO 48], before Riemann [RIE 58/59], Rankine [RAN 70] and Hugoniot [HUG 87, HUG 89] pushed the analysis further. However, considering discontinuous solutions brings new difficulties: we realize that weak solutions of such nonlinear problems are not unique. For example, for Burgers' equation  $f(\rho) = \rho^2/2$ , with identically null initial data  $\rho_{\text{Init}}(x) = 0$ , we can prove that  $\rho_1(t, x) = 0$  and  $\rho_2(t, x) = \mathbf{1}_{-t/2 \leq x < 0} - \mathbf{1}_{0 < x \leq t/2}$  are both solutions in the sense of definition 3.4 (they both satisfy the Rankine–Hugoniot relation, by setting for  $\rho_2$ :  $s(t) = -t/2$ ,  $s(t) = 0$  and  $s(t) = +t/2$  in order to characterize the three curves of discontinuity of this function, see Figure 3.43). The difficulty thus consists of finding a criterion which allows us to choose from all the weak solutions one (and only one) specific solution.



**Figure 3.43.** Non-entropic solution of the Burgers' equation for identically null initial data.

**DEFINITION 3.5.–** A weak solution  $\rho$  of [3.43] is said to be *entropic*, if, for all convex functions  $\eta$ , by representing an antiderivative of  $\eta' f'$  with  $q$ , we have

$$-\int_0^\infty \int_{\mathbb{R}} (\eta(\rho) \partial_t \phi + q(\rho) \partial_x \phi)(t, x) dx dt - \int_{\mathbb{R}} \eta(\rho_{\text{Init}})(x) \phi(0, x) dx \leq 0 \quad [3.48]$$

for all positive test functions  $\phi \geq 0$ . (We sometimes write this in the abridged form “ $\partial_t \eta(\rho) + \partial_x q(\rho) \leq 0$ ” and we say that the functions  $(\eta, q)$  form an entropy–entropy flux pair.)

Evidently, if  $\rho$  is a regular solution of [3.43], we can develop the derivatives in [3.43] and write the equation in the form

$$\partial_t \rho + f'(\rho) \partial_x \rho = 0.$$

By applying the chain rule, we thus obtain

$$\partial_t \eta(\rho) = -\eta'(\rho) f'(\rho) \partial_x \rho = -q'(\rho) \partial_x \rho = -\partial_x q(\rho).$$

However, if  $\rho$  is discontinuous, this non-conservative form has no meaning (the product  $f'(\rho) \partial_x \rho$  being, at the points of discontinuity, the product of a discontinuous function with a Dirac measure). In fact, a discontinuous solution of [3.43] cannot, in general, also be a weak solution of  $\partial_t \eta(\rho) + \partial_x q(\rho) = 0$ . Indeed, returning to the calculations that led to the Rankine–Hugoniot relation, we obtain

$$\begin{aligned} \llbracket \eta(\rho) \rrbracket \dot{s} - \llbracket q(\rho) \rrbracket &= (\eta(\rho^+) - \eta(\rho^-)) \dot{s} - (q(\rho^+) - q(\rho^-)) \\ &= \int_{\rho^-}^{\rho^+} \eta'(z) \dot{s} dz - \int_{\rho^-}^{\rho^+} q'(z) dz \\ &= \int_{\rho^-}^{\rho^+} \eta'(z) \dot{s} dz - \int_{\rho^-}^{\rho^+} \eta' f'(z) dz \quad \text{which becomes by integrating by parts} \\ &= - \int_{\rho^-}^{\rho^+} \eta''(z) \left( \frac{f(\rho^+) - f(\rho^-)}{\rho^+ - \rho^-} (z - \rho^-) - (f(z) - f(\rho^-)) \right) dz \\ &= - \int_{\rho^-}^{\rho^+} \eta''(z) (z - \rho^-) \left( \frac{f(\rho^+) - f(\rho^-)}{\rho^+ - \rho^-} - \frac{f(z) - f(\rho^-)}{z - \rho^-} \right) dz. \end{aligned}$$

As  $\eta$  is convex,  $z \mapsto \eta''(z)(z - \rho^-)$  is of constant sign over the interval  $I$  defined by  $\rho^+$  and  $\rho^-$ . If we assume that  $f$  is convex or concave over  $I$ , the integrand is thus of constant sign and  $\llbracket \eta(\rho) \rrbracket \dot{s} - \llbracket q(\rho) \rrbracket$  cancels if and only if  $f(z) - f(\rho^-) = \frac{f(\rho^+) - f(\rho^-)}{\rho^+ - \rho^-}(z - \rho^-)$  for all  $z \in I$ . This is only possible if  $f$  is a linear function over  $I$ . The entropic criterion thus aims to select the admissible discontinuities. By returning to the calculations which led to the Rankine–Hugoniot relation, we find that the selection criterion of definition 3.5 is formulated as

$$\llbracket \eta(\rho) \rrbracket \dot{s} = (\eta(\rho^+) - \eta(\rho^-)) \dot{s} \geq (q(\rho^+) - q(\rho^-)) = \llbracket q(\rho) \rrbracket. \quad [3.49]$$

We can immediately demonstrate that it is not satisfied, over the three curves of discontinuity, by the weak solution of Burgers' equation shown in Figure 3.43, for example, with the function  $\eta(z) = z^2$ . Indeed, this criterion restores the uniqueness as stated by the following theorem [KRU 70].

**THEOREM 3.12** (Kruzkov's theorem).— The scalar conservation law [3.43] has a unique weak entropic solution satisfying for all  $0 < R < \infty$  the continuity property

$$\lim_{T \rightarrow 0} \left( \int_0^T \int_{|x| \leq R} |\rho(t, x) - \rho(0, x)| \, dx \, dt \right) = 0. \quad [3.50]$$

The demonstration of this result is technical and lies beyond the scope of this book. The basic idea is to use relations [3.48] for a set of entropy functions  $\eta$  chosen so as to evaluate the distance between two solutions. With a radically different approach, we can show, in the particular case where the flux  $f$  is a convex function, that it is sufficient to satisfy relation [3.48] for a *single* convex entropy function  $\eta$ , see [PAN 94] and [DE 04]. In certain references, criterion [3.48] is stated for test functions in  $C_c^\infty([0, \infty[ \times \mathbb{R})$ , so that the initial data term disappears. If we impose [3.48] for test functions which do not necessarily cancel at  $t = 0$ , the solutions obtained satisfy property [3.50]. This technical point is delicate and subtle, continuity in time being an essential property to prove uniqueness. Under certain hypotheses for the flux  $f$ , we can relax condition [3.50] and only impose [3.48] for  $\phi \in C_c^\infty([0, \infty[ \times \mathbb{R})$ , see [CHE 00, VAS 01]. Here, we will retain the following practical criterion.

**LEMMA 3.5** (Lax's criterion).— The discontinuities of a weak entropic solution of [3.43] satisfy

$$f'(u_+) \leq \dot{s} \leq f'(u_-).$$

In particular, when the flux  $f$  is convex, the admissible discontinuities are such that  $u_+ \leq u_-$ .

**PROOF.**— We consider the Kruzhkov entropy/flux entropy pairs

$$\eta_k(u) = |u - k|, \quad q_k(u) = (f(u) - f(k))\operatorname{sgn}(u - k),$$

which define a family of convex functions, and their associated fluxes, parameterized by  $k \in \mathbb{R}$ . The entropic criterion

$$\llbracket |u - k| \rrbracket \dot{s} \geq \llbracket (f(u) - f(k))\operatorname{sgn}(u - k) \rrbracket$$

leads to the following conclusions:

– with  $k < \min(u_-, u_+)$ , and  $k > \max(u_-, u_+)$ , we recover the Rankine–Hugoniot relation, since in these cases  $\operatorname{sgn}(u_+ - k) = \operatorname{sgn}(u_- - k)$  and we obtain, on the one hand, ( $k < \min(u_-, u_+)$ )

$$\begin{aligned} f(u_+) - f(k) - f(u_-) + f(k) &= f(u_+) - f(u_-) \\ &\leq \dot{s}(u_+ - k - u_- - k) = \dot{s}(u_+ - u_-) \end{aligned}$$

and, on the other hand, ( $k > \max(u_-, u_+)$ )

$$\begin{aligned} f(u_-) - f(k) - f(u_+) + f(k) &= f(u_-) - f(u_+ -) \\ &\leq \dot{s}(k - u_+ - k + u_-) = \dot{s}(u_- - u_+). \end{aligned}$$

- with  $k = \theta u_- + (1 - \theta)u_+$ ,  $0 \leq \theta \leq 1$ , we arrive at

$$\begin{aligned} \dot{s}(|u_+ - \theta u_- - (1 - \theta)u_+| - |u_- - \theta u_- - (1 - \theta)u_+|) \\ &= -(1 - 2\theta)\dot{s}|u_+ - u_-| \\ &\geq (f(u_+) - f(k))\operatorname{sgn}(u_+ - \theta u_- - (1 - \theta)u_+) \\ &\quad - (f(u_-) - f(k))\operatorname{sgn}(u_- - \theta u_- - (1 - \theta)u_+) \\ &\geq \operatorname{sgn}(u_+ - u_-)(f(u_+) - 2f(k) + f(u_-)). \end{aligned}$$

By using the expression  $\dot{s} = \frac{f(u_+) - f(u_-)}{u_+ - u_-}$ , we can write

$$\dot{s}|u_+ - u_-| = \operatorname{sgn}(u_+ - u_-)(f(u_+) - f(u_-)).$$

We are thus led to

$$\operatorname{sgn}(u_+ - u_-)\left(\theta f(u_-) + (1 - \theta)f(u_+) - f(\theta u_- + (1 - \theta)u_+)\right) \leq 0$$

or again, for  $0 < \theta < 1$ ,

$$\operatorname{sgn}(u_+ - u_-)\left(\frac{f(u_+ + \theta(u_- - u_+)) - f(u_+)}{\theta} + f(u_+) - f(u_-)\right) \geq 0.$$

By taking the limit  $\theta \rightarrow 0$ , we arrive at

$$\underbrace{(u_+ - u_-)\operatorname{sgn}(u_+ - u_-)}_{=|u_+ - u_-|} \times \left(-f'(u_+) + \underbrace{\frac{f(u_+) - f(u_-)}{u_+ - u_-}}_{=\dot{s}}\right) \geq 0.$$

We obtain the other of Lax's inequality criteria by taking the limit  $\theta \rightarrow 1$ .  $\square$

It is also important to understand, at least intuitively, the motivation behind the entropic criterion. The idea, notably developed in [LAD 57, LAX 54, LAX 57,

LAX 60, OLE 57], consists of seeing problem [3.43] as the limit when  $\epsilon \rightarrow 0$  of the equations

$$\partial_t \rho_\epsilon + \partial_x f(\rho_\epsilon) = \epsilon \partial_{xx}^2 \rho_\epsilon. \quad [3.51]$$

The introduction of the “viscous” term  $\epsilon \partial_{xx}^2 \rho_\epsilon$  can be motivated either with analogy to the derivation, at least formally, of the Euler equations starting from the Navier–Stokes equations when the viscosity becomes small (this is the concept which motivates the name “approximation by vanishing viscosity”), or from considerations of the numerical order: as we saw in the linear case, the “good” numerical schemes inherently introduce such a diffusion term, with a parameter  $\epsilon$  dependent on the discretization parameters (see [3.25]). The approach for the analysis of [3.43] can thus be outlined as follows:

- Solve the parabolic nonlinear problem [3.51] with  $\epsilon > 0$ . The equation remains nonlinear, we expect that the diffusion term will allow us to work with more regular solutions (but and, thus, if we think in terms of sequences of solutions, to work with better compactness properties), which will lead to the establishment of a fixed-point strategy<sup>4</sup>.
- Establish *uniform* estimates with respect to the parameter  $\epsilon > 0$ .
- Deduce from these the compactness properties.
- Take the limit  $\epsilon \rightarrow 0$ . At this stage, the difficulty lies in dealing with the nonlinear term  $f(\rho_\epsilon)$ , which necessitates strong convergence properties of  $\rho_\epsilon$ .
- Establish the entropic criterion and the uniqueness of entropic solutions.

These different points are based on extremely subtle analysis techniques which will not be explained in detail here; for more information, see [GOD 91, SER 96a]. However, we can explain from this point of view the origin of the entropic criterion. Indeed, the solutions  $\rho_\epsilon$  of [3.51] being regular, we have

$$\begin{aligned} \partial_t \eta(\rho_\epsilon) + \partial_x q(\rho_\epsilon) &= \epsilon \eta'(\rho_\epsilon) \partial_{xx}^2 \rho_\epsilon \\ &= \epsilon \partial_x (\eta'(\rho_\epsilon) \partial_x \rho_\epsilon) - \epsilon \eta''(\rho_\epsilon) |\partial_x \rho_\epsilon|^2. \end{aligned}$$

---

<sup>4</sup> In certain cases, we can in fact explicitly solve the nonlinear problem. Thus, for the viscous Burgers’ equation  $\partial_t \rho_\epsilon + \partial_x (\rho_\epsilon^2/2) = \epsilon \partial_{xx}^2 \rho_\epsilon$ , we introduce the *Hopf–Cole transformation*:  $v_\epsilon(t, x) = \exp(-\frac{1}{2\epsilon} \int_{-\infty}^x \rho_\epsilon(t, y) dy)$ , so that  $\rho_\epsilon(t, x) = -2\epsilon \frac{\partial_x v_\epsilon(t, x)}{v_\epsilon(t, x)} = -2\epsilon \partial_x \ln(v_\epsilon(t, x))$ , see [HOP 50]. We demonstrate that  $v_\epsilon(t, x)$  satisfies the heat equation  $\partial_t v_\epsilon = \epsilon \partial_{xx}^2 v_\epsilon$ ; we thus have an explicit formula giving  $\rho_\epsilon(t, x)$  as a function of  $\rho_{\text{Init}}$ .

By integrating, we deduce from this

$$\frac{d}{dt} \int \eta(\rho_\epsilon) dx + \epsilon \int \eta''(\rho_\epsilon) |\partial_x \rho_\epsilon|^2 dx = 0.$$

With  $\eta(\rho) = \rho^2/2$ , we deduce from this that, if  $\rho_{\text{init}} \in L^2(\mathbb{R})$ ,

$\rho_\epsilon$  is bounded in  $L^\infty(0, T; L^2(\mathbb{R}))$ .

and

$\sqrt{\epsilon} \partial_x \rho_\epsilon$  is bounded in  $L^2([0, T] \times \mathbb{R})$ .

More generally, with  $\eta(\rho) = \rho^p/p$  or  $\eta(\rho) = [\rho - \|\rho_{\text{Init}}\|_\infty]_+$ , we show that  $\rho_\epsilon$  is bounded in  $L^\infty(0, T; L^p(\mathbb{R}))$  if  $\rho_{\text{init}} \in L^p(\mathbb{R})$  for all  $1 \leq p \leq \infty$  (notably,  $|\rho_\epsilon(t, x)| \leq \|\rho_{\text{Init}}\|_\infty$ ). Thus, we write

$$\partial_t \rho_\epsilon + \partial_x f(\rho_\epsilon) = \sqrt{\epsilon} \partial_x (\sqrt{\epsilon} \partial_x \rho_\epsilon) \xrightarrow[\epsilon \rightarrow 0]{} 0$$

and the only difficulty lies in taking the limit in  $f(\rho_\epsilon)$ . Finally, if  $\eta$  is convex, we have

$$\partial_t \eta(\rho_\epsilon) + \partial_x q(\rho_\epsilon) = \underbrace{\sqrt{\epsilon} \partial_x (\sqrt{\epsilon} \eta'(\rho_\epsilon) \partial_x \rho_\epsilon)}_{\xrightarrow[\epsilon \rightarrow 0]{} 0} - \underbrace{\epsilon \eta''(\rho_\epsilon) |\partial_x \rho_\epsilon|^2}_{\leq 0},$$

a relation which motivates the entropic criterion: the term  $-\epsilon \eta''(\rho_\epsilon) |\partial_x \rho_\epsilon|^2$  is bounded, a priori its limit as  $\epsilon \rightarrow 0$  is not null, but has a sign.

To resume, we must keep in mind:

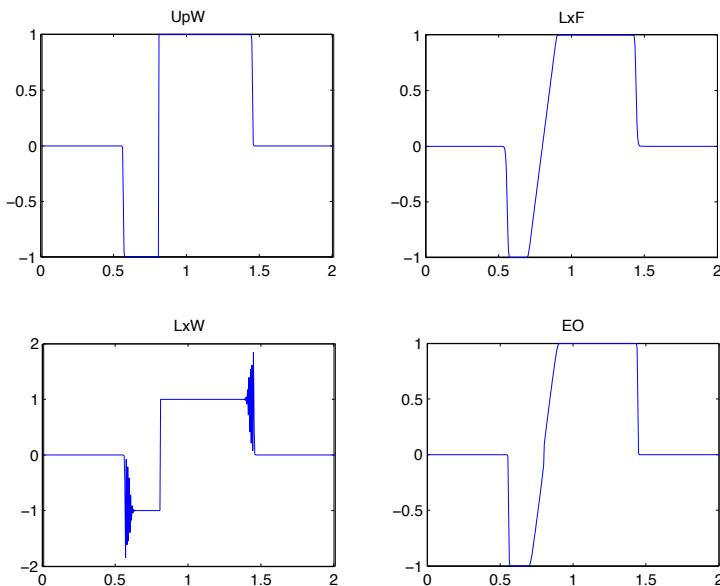
- The solutions of [3.43] generally become discontinuous in finite time, even if the initial data are regular.
- This phenomenon requires us to work in a framework of weak solutions.
- The weak solutions are not unique; the uniqueness is re-established by also imposing an entropic criterion.

In terms of numerical simulations, this discussion leads to the following difficulties:

- The schemes must be capable of reproducing these phenomena concerning the appearance of singularities,
- and of following their propagation as time develops.

- The solutions of [3.43] satisfy *a priori* natural conditions that the numerical solutions must conserve (preservation of the integral  $\int \rho(t, x) dx = \int \rho_{\text{Init}}(x) dx$ ,  $L^\infty$  estimates, etc.).
- The scheme must find the entropic solution and thus the numerical solutions must satisfy a discrete version of the entropic criterion.

Following all this, Figure 3.44 shows the numerical solutions obtained with four different schemes for the *same* Burgers' equation and starting with the same initial state. These solutions display discontinuities. However, the scheme labeled LxW exhibits strong oscillations and the solution produced by the UpW scheme does not satisfy Lax's criterion, in contrast to the LxF and EO schemes, which find an admissible solution.



**Figure 3.44.** Simulations of Burgers' equation with four different numerical schemes

### 3.4.1.2. Lax–Friedrichs and Rusjanov schemes for scalar conservation laws

The idea here is to draw inspiration from the study of the linear case to decentre the flux in an adequate manner. To this end, we set  $c = \sup_{|u| \leq M} |f'(u)|$  and we rewrite the equation in the form

$$\partial_t u + \partial_x f(u) = \partial_t u + \frac{1}{2} \partial_x (f(u) - cu) + \frac{1}{2} \partial_x (f(u) + cu) = 0.$$

We note that for  $f(u) - cu$ , the “velocity”  $f'(u) - c$  is, by construction, negative, while for  $f(u) + cu$ , the “velocity”  $f'(u) + c$  is, by construction, positive. This yields the upwinding strategy: to the right for the first term, and to the left for the second term. Thus, the numerical flux is defined by

$$\mathbb{F}(u_{j+1}, u_j) = \frac{1}{2}(f(u_{j+1}) - cu_{j+1}) + \frac{1}{2}(f(u_j) + cu_j) \quad [3.52]$$

which thus leads to the scheme

$$\begin{aligned} u_j^{n+1} &= u_j^n - \frac{\Delta t}{h_j} (\mathbb{F}(u_{j+1}^n, u_j^n) - \mathbb{F}(u_j^n, u_{j-1}^n)) \\ &= u_j^n - \frac{\Delta t}{2h_j} (f(u_{j+1}) - f(u_{j-1})) + c \frac{\Delta t}{2h_j} (u_{j+1}^n - 2u_j^n + u_{j-1}^n). \end{aligned} \quad [3.53]$$

This scheme can be interpreted as a consistent approximation in the finite difference sense, of order 1 in time, 2 in space, of the equation

$$\partial_t u + \partial_x f(u) = \frac{ch}{2} \partial_{xx}^2 u.$$

However, as we discussed in detail in the previous section for the linear transport equation, the concept of consistency in the finite difference sense is not really appropriate for the analysis of these schemes. It suits to introduce a new definition of consistency, adapted to the finite volume formalism.

**DEFINITION 3.6.–** A FV scheme [3.12] with a numerical flux  $F_{j+1/2} = \mathbb{F}(u_{j+1}, u_j)$  is said to be *consistent in the FV sense* with the equation  $\partial_t u + \partial_x f(u)$  if the numerical flux function  $\mathbb{F}$  is such that

$$\mathbb{F}(u, u) = f(u).$$

Scheme [3.52]–[3.53] is evidently consistent in the sense of definition 3.6. Furthermore, we can identify a condition over the numerical parameters, which guarantees the preservation of natural estimates on the solution of the problem.

**PROPOSITION 3.8.–** We assume that  $f$  is a convex function and that the numerical parameters are such that

$$\sup_{|u| \leq M} |f'(u)| \frac{\Delta t}{kh} \leq 1.$$

Thus, scheme [3.52]–[3.53] is  $L^\infty$ -stable: if  $|u_j^0| \leq M$  for all  $j$ , then  $|u_j^n| \leq M$  for all  $n, j$ .

PROOF.– As  $f$  is convex, for all  $x, y$ , we have

$$f(y) \geq f(x) + f'(x)(y - x).$$

We deduce from this that

$$f(u_{j+1}^n) \geq f(u_{j+1}^n) + f'(u_{j+1}^n)(u_{j+1}^n - u_{j+1}^n)$$

and

$$f(u_{j+1}^n) \geq f(u_{j+1}^n) + f'(u_{j+1}^n)(u_{j+1}^n - u_{j+1}^n).$$

It follows that

$$\begin{aligned} u_j^{n+1} &= \frac{c}{2} \frac{\Delta t}{h_j} (u_{j+1}^n + u_{j-1}^n) + u_j^n \left(1 - c \frac{\Delta t}{h_j}\right) - \frac{f(u_{j+1}^n) - f(u_{j-1}^n)}{2} \frac{\Delta t}{h_j} \\ &\leq \frac{c}{2} \frac{\Delta t}{h_j} (u_{j+1}^n + u_{j-1}^n) + u_j^n \left(1 - c \frac{\Delta t}{h_j}\right) - \frac{\Delta t}{2h_j} f'(u_{j-1}^n) (u_{j+1}^n - u_{j-1}^n) \\ &\leq \left(1 - c \frac{\Delta t}{h_j}\right) u_j^n + \frac{\Delta t}{2h_j} \underbrace{(c - f'(u_{j-1}^n))}_{\geq 0} u_{j+1}^n + \frac{\Delta t}{2h_j} \underbrace{(c + f'(u_{j-1}^n))}_{\geq 0} u_{j-1}^n \\ &\leq \left(1 - c \frac{\Delta t}{h_j} + \frac{\Delta t}{2h_j} (c - f'(u_{j-1}^n) + c + f'(u_{j-1}^n))\right) \underbrace{\sup_k |u_k^n|}_{=M} \leq M. \end{aligned}$$

Likewise, we obtain

$$\begin{aligned} u_j^{n+1} &\geq \left(1 - c \frac{\Delta t}{h_j}\right) u_j^n + \frac{\Delta t}{2h_j} \underbrace{(c - f'(u_{j+1}^n))}_{\geq 0} u_{j+1}^n + \frac{\Delta t}{2h_j} \underbrace{(c + f'(u_{j+1}^n))}_{\geq 0} u_{j-1}^n \\ &\geq \left(1 - c \frac{\Delta t}{h_j} + \frac{\Delta t}{2h_j} (c - f'(u_{j-1}^n) + c + f'(u_{j-1}^n))\right) \underbrace{\inf_k |u_k^n|}_{=m} \geq m. \end{aligned}$$

The argument also demonstrates that the CFL constraint is still satisfied:  $f$  is convex, thus  $f'$  is monotonically increasing and we always have  $|f'(u_j^n)| \frac{\Delta t}{h_j} \leq 1$ .  $\square$

More generally, it is interesting to introduce the following concept.

DEFINITION 3.7.– We say that the numerical flux  $F_{j+1/2} = \mathbb{F}(u_{j+1}, u_j)$  for [3.12] is *monotone* if the function  $(u, v) \mapsto \mathbb{F}(u, v)$  satisfies

$$\partial_u \mathbb{F}(u, v) \leq 0, \quad \partial_v \mathbb{F}(u, v) \geq 0.$$

This definition gives a simple, practical criterion to satisfy the maximum principle, that is to say the preservation of  $L^\infty$  bounds by the numerical solution. The idea is then to identify a convex combination by writing

$$\begin{aligned} u_j^{n+1} &= u_j^n - \frac{\Delta t}{h_j} (\mathbb{F}(u_{j+1}^n, u_j^n) - \mathbb{F}(u_j^n, u_{j-1}^n)) \\ &= u_j^n - \frac{\Delta t}{h_j} (\mathbb{F}(u_{j+1}^n, u_j^n) - \mathbb{F}(u_j^n, u_j^n)) - \frac{\Delta t}{h_j} (\mathbb{F}(u_j^n, u_j^n) - \mathbb{F}(u_j^n, u_{j-1}^n)) \\ &= u_j^n + \lambda_j^n (u_{j+1}^n - u_j^n) - \mu_j^n (u_j^n - u_{j-1}^n), \end{aligned}$$

where we have set

$$\begin{aligned} \lambda_j^n &= -\frac{\Delta t}{h_j} \frac{\mathbb{F}(u_{j+1}^n, u_j^n) - \mathbb{F}(u_j^n, u_j^n)}{u_{j+1}^n - u_j^n} = -\int_0^1 \partial_1 \mathbb{F}(u_j^n + \theta(u_{j+1}^n - u_j^n), u_j^n) d\theta, \\ \mu_j^n &= \frac{\Delta t}{h_j} \frac{\mathbb{F}(u_j^n, u_j^n) - \mathbb{F}(u_j^n, u_{j-1}^n)}{u_j^n - u_{j-1}^n} = \int_0^1 \partial_2 \mathbb{F}(u_j^n, u_{j-1}^n + \theta(u_j^n - u_{j-1}^n)) d\theta. \end{aligned}$$

We thus obtain

$$u_j^{n+1} = (1 - \lambda_j^n - \mu_j^n) u_j^n + \lambda_j^n u_{j+1}^n + \mu_j^n u_{j-1}^n.$$

Assume that the coefficients  $\lambda_j^n$  and  $\mu_j^n$  are positive. Let us assume then that  $u_j^n$  are in  $[-M, M]$  and that the flux function  $\mathbb{F}$  is  $\Lambda$ -Lipschitzian over  $[-M, M] \times [-M, M]$ . Thus, we observe that  $\lambda_j^n \leq \frac{\Delta t}{h_j} \Lambda$ ,  $\mu_j^n \leq \frac{\Delta t}{h_j} \Lambda$ . It follows that  $u_j^{n+1}$  remains in the interval  $[-M, M]$  under the condition  $2 \frac{\Delta t}{h_j} \Lambda \leq 1$ .

**LEMMA 3.6.**— We assume that the flux  $\mathbb{F}$  is monotone. We also assume that  $\mathbb{F}$  is locally Lipschitzian. If we have  $|u_j^n| \leq M$  for all  $j$  and the time step satisfies  $2 \frac{\Delta t}{h_j} \Lambda \leq 1$ , where  $\Lambda$  represents the Lipschitz constant of  $\mathbb{F}$  over  $[-M, M] \times [-M, M]$ , then for all  $j$ , we again have  $|u_j^{n+1}| \leq M$ .

**THEOREM 3.13 (Lax–Wendroff theorem).**— We assume that

- i) The scheme is consistent (in the FV sense),
- ii) The scheme is  $L^\infty$ -stable so that for all  $n, j$ ,  $|u_j^n| \leq M$ .

We set

$$u_\Delta(t, x) = \sum_{n,j} u_j^n \mathbf{1}_{n\Delta t \leq t < (n+1)\Delta t} \mathbf{1}_{x_{j-1/2} \leq x < x_{j+1/2}}.$$

We assume that  $u_\Delta$  has a limit  $u$  in  $L^1_{\text{loc}}([0, T] \times \mathbb{R})$  as  $\Delta t$  and  $h$  go to 0. Thus,  $u$  is a weak solution of [3.43].

This statement [LAX 60] still raises two questions:

- a) how can we guarantee the *strong* convergence of  $u_\Delta$  towards  $u$ ?
- b) the weak solutions of conservation law [3.43] are not unique; to guarantee the uniqueness, the entropy inequalities must also be satisfied. Hence, is the scheme used *entropic*?

To answer a), a compactness argument must be used. The idea is to check a discrete form of the derivatives of  $u$ . We explain the strategy by returning to the viscous approximation of [3.43]

$$\partial_t u_\varepsilon + \partial_x f(u_\varepsilon) = \varepsilon \partial_{xx}^2 u_\varepsilon.$$

By differentiating with respect to the variable  $x$ , we again obtain

$$\partial_t (\partial_x u_\varepsilon) + \partial_x (f'(u_\varepsilon) \partial_x u_\varepsilon) = \varepsilon \partial_{xx}^2 (\partial_x u_\varepsilon).$$

Let  $Z : \mathbb{R} \rightarrow \mathbb{R}$  be a given convex function. By multiplying this relation by  $Z'(\partial_x u_\varepsilon)$  and then integrating

$$\begin{aligned} & \frac{d}{dt} \int Z(\partial_x u_\varepsilon) dx + \varepsilon \int Z''(\partial_x u_\varepsilon) |\partial_{xx}^2 u_\varepsilon|^2 dx \\ &= - \int \partial_x (f'(u_\varepsilon) \partial_x u_\varepsilon) Z'(\partial_x u_\varepsilon) dx \\ &= - \int \left[ \partial_x (f'(u_\varepsilon)) \partial_x u_\varepsilon Z'(\partial_x u_\varepsilon) + f'(u_\varepsilon) \partial_x (\partial_x u_\varepsilon) Z'(\partial_x u_\varepsilon) \right] dx \\ &= - \int \left[ \partial_x (f'(u_\varepsilon)) \partial_x u_\varepsilon Z'(\partial_x u_\varepsilon) + f'(u_\varepsilon) \partial_x (Z(\partial_x u_\varepsilon)) \right] dx. \end{aligned}$$

As  $Z$  is convex, the second term on the left-hand side is positive. We now seek to remove the term on the right-hand side by a judicious choice of function  $Z$ . Specifically, with  $Z(s) = |s|$ , we have  $Z'(s) = \text{sgn}(s)$  and  $\partial_x u_\varepsilon Z'(\partial_x u_\varepsilon) = |\partial_x u_\varepsilon| = Z(\partial_x u_\varepsilon)$ , such that the term on the right-hand side becomes

$$- \int \partial_x (f'(u_\varepsilon) Z(\partial_x u_\varepsilon)) dx = 0.$$

Finally, we deduce from this that

$$\int |\partial_x u_\varepsilon|(t, x) dx \leq \int |\partial_x u_\varepsilon|(0, x) dx.$$

The calculation remains formal because, for this choice of function  $Z''(s) = \delta(s = 0)$ , the integrations by parts are not justified. To make the argument rigorous, we must work with functions  $Z_\eta(s)$  that are regular, convex and convergent towards  $|s|$  (for example  $Z_\eta(s) = \frac{s}{\sqrt{\eta+s^2}}$ ). We thus obtain a “BV bound” on the solution of [3.43], when the initial data itself satisfy  $\partial_x u_{\text{Init}} \in L^1(\mathbb{R})$ . [EVA 92, chapter 5] can be consulted for the construction and properties of the space

$$\text{BV}(\mathbb{R}) = \{u \in L^\infty(\mathbb{R}), \partial_x u \in L^1(\mathbb{R})\}.$$

In particular, for all bounded domains  $\Omega$ , a set bounded in  $\text{BV}(\Omega)$  embeds compactly in  $L^1(\Omega)$  (see [EVA 92, theorem 4, Sec. 5.2.3]). We will take note that the function  $u_\varepsilon$  depends on the two variables of time and space: to establish the compactness in  $L^1_{\text{loc}}([0, T] \times \mathbb{R})$ , it must thus be shown (see [GOU 11, theorem 7.56]) that for all functions  $\varphi \in C_c^\infty([0, T] \times \mathbb{R})$ , we have

$$\lim_{h,k \rightarrow 0} \sup_{\varepsilon > 0} \int_0^t \int_{\mathbb{R}} |u_\varepsilon(t+k, x+h) - u_\varepsilon(t, x)| |\varphi(t, x)| dx dt = 0.$$

The *BV* estimate provides the necessary control in terms of the spatial variable, but we still need to achieve the desired behavior in terms of the temporal variable. To this end, we return to the equation where  $\partial_t u_\varepsilon = \partial_x(-f(u_\varepsilon) + \varepsilon \partial_x u_\varepsilon)$  is expressed as the derivative in space of functions uniformly bounded in  $L^1((0, T) \times (-R, R))$  for all  $0 < R < \infty$ . This estimate allows us to obtain the compactness in the  $L^1$  norm. The adaptation of these arguments to the discrete problem is based on the following concept, which gives a discrete analog of the *BV* framework [LAX 77, HAR 84].

**DEFINITION 3.8.–** A FV scheme is said to be *TVD* if we have

$$\sum_j |u_{j+1}^{n+1} - u_j^{n+1}| \leq \sum_j |u_{j+1}^n - u_j^n|.$$

**LEMMA 3.7 (Le Roux–Harten lemma).–** A scheme of the form

$$u_j^{n+1} = u_j^n + C_{j+1/2}^n (u_{j+1}^n - u_j^n) - D_{j-1/2}^n (u_j^n - u_{j-1}^n)$$

with

$$C_{j+1/2}^n \geq 0 \quad D_{j+1/2}^n \geq 0, \quad C_{j+1/2}^n + D_{j-1/2}^n \leq 1$$

is  $L^\infty$ -stable. If additionally

$$C_{j+1/2}^n + D_{j+1/2}^n \leq 1$$

then the scheme is TVD.

PROOF.– We identify a convex combination by writing

$$u_j^{n+1} = (1 - C_{j+1/2}^n - D_{j-1/2}^n)u_j^n + C_{j+1/2}^n u_{j+1}^n + D_{j-1/2}^n u_{j-1}^n.$$

This guarantees the preservation of the  $L^\infty$  bounds. Next, we must evaluate

$$\begin{aligned} u_{j+1}^{n+1} - u_j^{n+1} &= u_{j+1}^n + C_{j+3/2}^n(u_{j+2}^n - u_{j+1}^n) - D_{j+1/2}^n(u_{j+1}^n - u_j^n) \\ &\quad - u_j^n - C_{j+1/2}^n(u_{j+1}^n - u_j^n) + D_{j-1/2}^n(u_j^n - u_{j-1}^n) \\ &= (u_{j+1}^n - u_j^n)(1 - C_{j+1/2}^n - D_{j+1/2}^n) \\ &\quad + C_{j+3/2}^n(u_{j+2}^n - u_{j+1}^n) + D_{j-1/2}^n(u_j^n - u_{j-1}^n). \end{aligned}$$

In the term on the right-hand side, all the coefficients are assumed to be positive, and we thus obtain, after summation by parts,

$$\begin{aligned} \sum_j |u_{j+1}^{n+1} - u_j^{n+1}| &\leq \sum_j |u_{j+1}^n - u_j^n| (1 - C_{j+1/2}^n - D_{j+1/2}^n + C_{j+1/2}^n + D_{j+1/2}^n) \\ &\leq \sum_j |u_{j+1}^n - u_j^n|. \end{aligned}$$

The positivity of  $C_{j+1/2}^n, D_{j-1/2}^n$  can be interpreted as a condition of upwinding the scheme, the other condition is in fact a stability constraint on the step size in time and space.  $\square$

LEMMA 3.8.– Under the CFL constraint, scheme [3.52]–[3.53] is TVD.

PROOF.– We seek to write the scheme in the form said to be incremental, which appears in the statement of the Le Roux–Harten lemma. We have

$$\begin{aligned} u_j^{n+1} &= u_j^n + (u_{j+1}^n - u_j^n) \frac{c\Delta t}{2h_j} - (f(u_{j+1}^n) - f(u_j^n)) \frac{c\Delta t}{2h_j} \\ &\quad - (u_j^n - u_{j-1}^n) \frac{c\Delta t}{2h_j} - (f(u_j^n) - f(u_{j-1}^n)) \frac{c\Delta t}{2h_j}. \end{aligned}$$

We thus set

$$C_{j+1/2}^n = \frac{\Delta t}{2h_j}(c - a_{j+1/2}^n), \quad C_{j+1/2}^n = \frac{\Delta t}{2h_j}(c + a_{j+1/2}^n)$$

where

$$f(u_{j+1}^n) - f(u_j^n) = \underbrace{\int_0^1 f'(u_j^n + \tau(u_{j+1}^n - u_j^n)) d\tau}_{=a_{j+1/2}^n} (u_{j+1}^n - u_j^n).$$

As a result of the CFL condition, we indeed have  $C_{j+1/2}^n + D_{j+1/2}^n = \frac{c\Delta t}{h_j} \leq 1$ , together with the positivity.  $\square$

### 3.4.1.3. Numerical illustrations

The implementation of the methods is made up of two stages: the calculation of the fluxes  $F_{j\pm 1/2}$ , and the updating of the unknown values which simply take the form  $u \leftarrow u - \frac{\Delta t}{\Delta x}(Fp - Fm)$  (where  $u$  is a vector containing the  $J$  coordinates  $u_1, \dots, u_J$ , while  $Fp$  and  $Fm$  have the fluxes as their coordinates). The second stage is thus common, only the evaluation of the fluxes is specific to the method used. We shall compare, for Burgers' equation ( $f(\rho) = \rho^2/2$ ) and the Lighthill–Whitham–Richards (LWR) equation ( $f(\rho) = V\rho(1-\rho)$ ), the behaviors of several schemes:

– Lax–Friedrichs:

$$F_{j+1/2} = \mathbb{F}(u_{j+1}, u_j) = \frac{f(u_j) + f(u_{j+1})}{2} - \frac{\Delta x}{2\Delta t}(u_{j+1} - u_j).$$

The centered scheme is stabilized by the numerical viscosity  $\frac{\Delta x^2}{2\Delta t}$ .

– Rusjanov:

$$F_{j+1/2} = \mathbb{F}(u_{j+1}, u_j) = \frac{f(u_j) + f(u_{j+1})}{2} - M_{j+1/2} \frac{u_{j+1} - u_j}{2},$$

where  $M_{j+1/2} = \max(|f'(u_{j+1})|, |f'(u_j)|)$ . This is a variant of the Lax–Friedrichs scheme where the numerical viscosity is locally defined.

– Lax–Wendroff:

$$F_{j+1/2} = \mathbb{F}(u_{j+1}, u_j) = \frac{f(u_j) + f(u_{j+1})}{2} - \frac{\Delta t}{2\Delta x} f'\left(\frac{u_{j+1} + u_j}{2}\right)(f(u_{j+1}) - f(u_j)).$$

This scheme is consistent of order 2 in time and space [LAX 60, RIC 63].

– Enquist–Osher:

$$F_{j+1/2} = \mathbb{F}(u_{j+1}, u_j) = \int_0^{u_{j+1}} [f'(z)]_- dz - \int_0^{u_j} [f'(z)]_+ dz$$

where  $[X]_\pm = \max(0, \pm X) = \pm \frac{X \pm |X|}{2} \geq 0$ . To obtain this scheme, we write

$$f(\rho) = \int_0^\rho f'(z) dz = \int_0^\rho [f'(z)]_+ dz - \int_0^\rho [f'(z)]_- dz,$$

then, to evaluate this quantity at the interface  $x_{j+1/2}$ , we upwind according to the sign, i.e. by using the unknown  $u_{j+1}$  for the negative part, and by using the unknown

$u_j$  for the positive part. We will see a bit later another interpretation of this scheme. For Burgers' equation, we have

$$F_{j+1/2}^{\text{Burg}} = \rho_j \frac{\rho_j + |\rho_j|}{4} + \rho_{j+1} \frac{\rho_{j+1} - |\rho_{j+1}|}{4}$$

and for the LWR equation, we have

$$\begin{aligned} F_{j+1/2}^{\text{LWR}} = & (3u_j(1-u_j)) \times \mathbf{1}_{(3-6u_j) \geq 0} + 3/4 \times \mathbf{1}_{(3-6u_j) < 0} \\ & + (3u_{j+1}(1-u_{j+1}) - 3/4) \times \mathbf{1}_{(3-6u_{j+1}) < 0}. \end{aligned}$$

– Upwind:

$$F_{j+1/2} = \mathbb{F}(u_{j+1}, u_j) = \begin{cases} f(u_j) & \text{if } f'\left(\frac{u_{j+1} + u_j}{2}\right) \geq 0, \\ f(u_{j+1}) & \text{if } f'\left(\frac{u_{j+1} + u_j}{2}\right) < 0. \end{cases}$$

– Non-conservative upwind:

$$F_{j+1/2} = \mathbb{F}(u_{j+1}, u_j) = (u_{j+1} - u_j) \frac{f'(u_j) - |f'(u_j)|}{2}.$$

This scheme is based on the non-conservative formulation of the equation  $\partial_t \rho + f'(\rho) \partial_x \rho = 0$  and the approximation of the spatial derivative  $\partial_x \rho$  is decentered according to the sign of  $f'(\rho_j)$ .

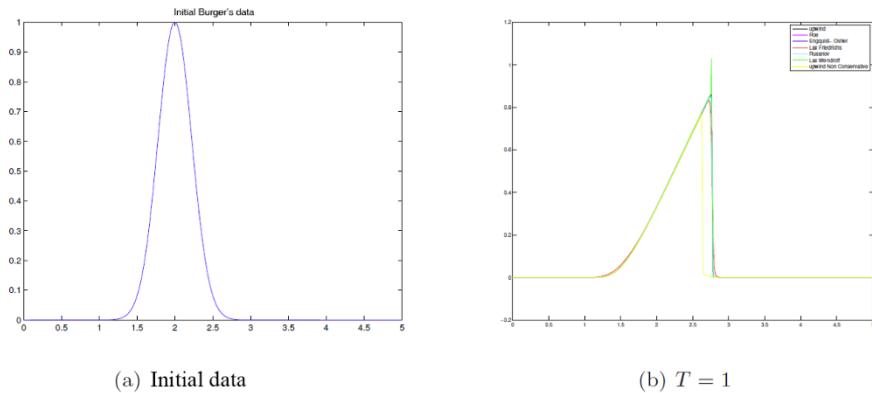
– Roe:

$$F_{j+1/2} = \mathbb{F}(u_{j+1}, u_j) = \begin{cases} f(u_j) & \text{if } \frac{f(u_{j+1}) - f(u_j)}{u_{j+1} - u_j} \geq 0, \\ f(u_{j+1}) & \text{if } \frac{f(u_{j+1}) - f(u_j)}{u_{j+1} - u_j} < 0. \end{cases}$$

The idea here is the same as for the upwind scheme, but with  $f'$  replaced by a discrete evaluation.

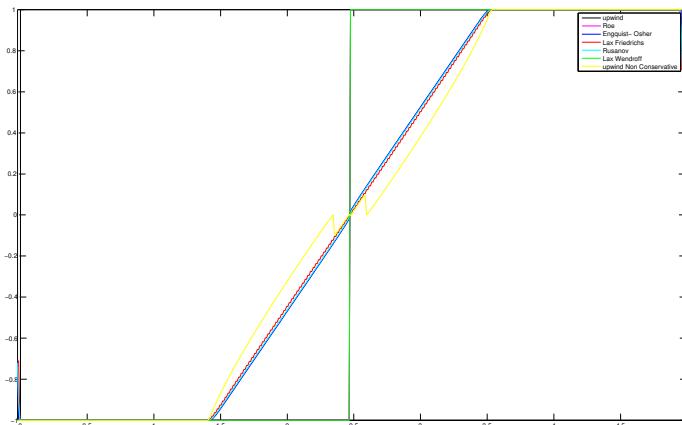
We start by conducting the simulations for Burgers' equation: the higher the  $\rho$ , the higher the velocity  $f'(\rho) = \rho$ . Here, the domain of calculation is the interval  $[0, 5]$ , discretized with a uniform step size  $\Delta x = 1/100$ . The time interval is fixed by  $\Delta t = 0.95 \times \Delta x$ . The results obtained for Gaussian data, centered on  $x = 2$ , are given in Figure 3.45. We do indeed observe the formation of a singularity: the solution becomes discontinuous and this discontinuity propagates from the left to the right. The Enquist–Osher, upwind and Roe schemes are identical in this particular case. The effect of numerical diffusion is more sensitive with the Lax–Friedrichs scheme.

The Lax–Wendroff scheme overestimates the value of  $\rho$  at the point of discontinuity (it exceeds the limit  $\rho_{\text{Max}} = 1$ ). The non-conservative upwind scheme produces a solution whose speed of propagation of the singularity is incorrect.



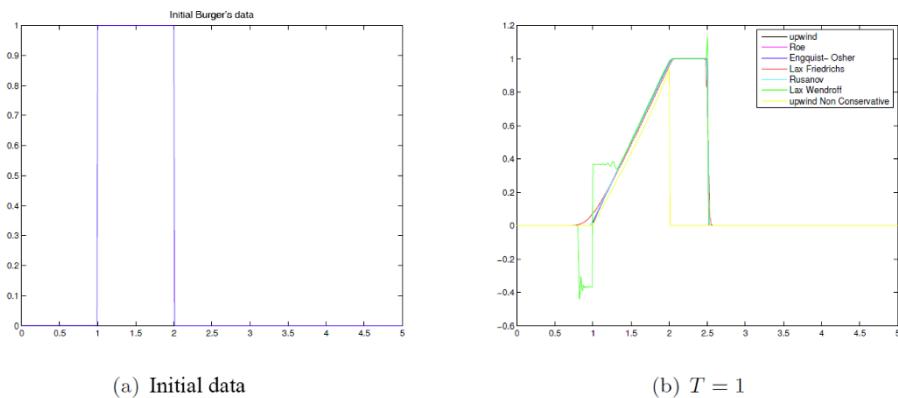
**Figure 3.45.** Simulation of Burgers' equation by several schemes. For the color version of this figure, see [www.iste.co.uk/goudon/mathmodel.zip](http://www.iste.co.uk/goudon/mathmodel.zip)

Figure 3.46 displays the solutions obtained with the data  $1_{x \geq 2.5} - 1_{x \leq 2.5}$ . The data are discontinuous, but the discontinuity does not satisfy the Lax criterion: the stationary solution is not admissible. However, it is the solution produced by the Lax–Wendroff, Roe and upwind schemes which are thus not entropic. The non-conservative scheme also gives an erroneous solution. The other schemes yield the anticipated solution.



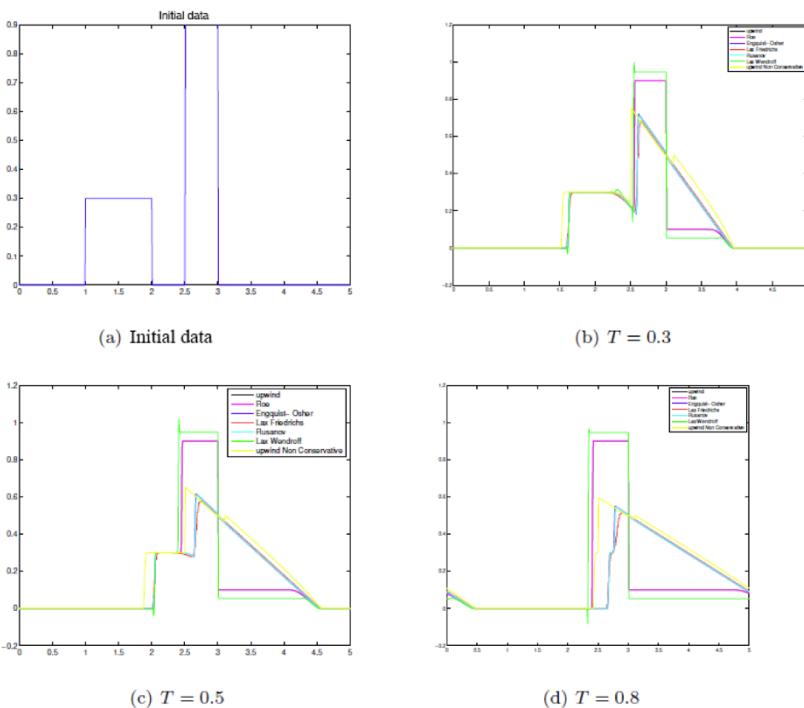
**Figure 3.46.** Simulation of Burgers' equation by several schemes (discontinuous data). For the color version of this figure, see [www.iste.co.uk/goudon/mathmodel.zip](http://www.iste.co.uk/goudon/mathmodel.zip)

Figure 3.47 displays the solutions obtained with a step function as initial data. We recover the same conclusions. The non-conservative scheme does not propagate the singularity at the correct speed. The Lax–Wendroff scheme overestimates the value of the solution at the discontinuity. It also produces oscillations and constructs an artificial discontinuity, with a negative value in the solution. The Lax–Friedrichs scheme entails stronger numerical diffusion (spreading of the discontinuity, smoother solutions). For these data, the upwind and Roe schemes overlap with the Enquist–Osher scheme (but we bear in mind that this is a specific case, the previous example demonstrating that these schemes can produce an inadmissible solution).



**Figure 3.47.** Simulation of Burgers' equation by several schemes. For the color version of this figure, see [www.iste.co.uk/goudon/mathmodel.zip](http://www.iste.co.uk/goudon/mathmodel.zip)

We next test the LWR model over the domain  $[0, 5]$  with a maximum speed  $V = 3$ . The (uniform) step size in space is  $\Delta x = 1/100$ . We set  $\Delta t = 0.95 \times \frac{\Delta x}{3}$ . The numerical results are shown in Figure 3.48. The initial data are formed of a double step: the one on the left, of density 0.3, moves quicker than the one on the right, held up with a density of 0.9. The behaviors of different schemes, as already showcased with Burgers' equation, stand out even more. The non-conservative scheme does not yield the correct speeds of propagation. The Lax–Wendroff scheme produces incorrect profiles (inadmissible discontinuities, negative values of density, overestimation of the extrema). The solutions of the Roe and upwind schemes are practically indistinguishable; they also display inadmissible discontinuities. Finally, the profile of the Lax–Friedrichs scheme is affected by a stronger numerical diffusion. As for the behavior of the solutions, we can present the following elements. The fronts, in front of the discontinuities, relax and the block on the left-hand side catches that on the right-hand side. At time  $T = 0.8$ , the two blocks have merged. We also see that the blocks leave the domain of calculation. As we have set periodic conditions, we thus see these blocks re-appearing at  $x = 0$ .



**Figure 3.48.** Simulation of the LWR equation by several schemes. For the color version of this figure, see [www.iste.co.uk/goudon/mathmodel.zip](http://www.iste.co.uk/goudon/mathmodel.zip)

#### **3.4.1.4. Propagation of a forest fire**

The example which will now be studied is inspired by [LE 99]. We seek to describe the evolution of a forest fire. The forest is represented by the entire  $\mathbb{R}^2$  plane, in order to overlook any difficulties related to the boundary conditions. We describe the phenomenon of combustion by the evolution of the flame front which separates the already-burnt part from the part not yet consumed. This front is described by a curve

$$y = p(t, x).$$

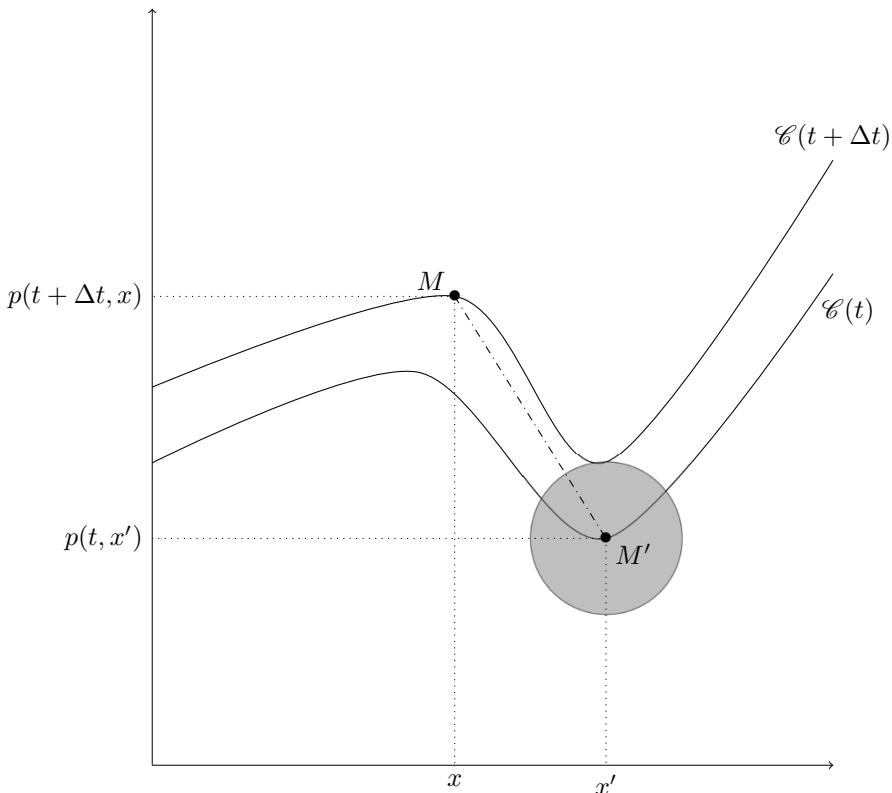
We shall exploit a certain number of simple modeling rules to establish an equation of evolution satisfied by  $p$ . Specifically, at time  $t \geq 0$ , the front of the fire is given by

$$\mathcal{C}(t) = \{(x, y) \in \mathbb{R}^2, y = p(t, x)\},$$

and the domain

$$\{(x, y) \in \mathbb{R}^2, y < p(t, x)\}$$

represents the part of the forest already consumed at time  $t > 0$ . The combustion phenomenon is characterized by a speed of propagation  $c$ , which we will assume to be constant (even though this hypothesis is not very realistic: with the distribution of different types of vegetation, the soil humidity, etc., not being homogeneous, the speed of propagation should be locally defined). Thus, over the time interval  $[t, t + \Delta t]$ , all points of  $\mathbb{R}^2$  situated beyond  $\mathcal{C}(t)$  and at a distance less than  $c\Delta t$  from a point of  $\mathcal{C}(t)$  are burnt. If we consider any two points  $M$  and  $M'$  situated, respectively, on the flame fronts  $\mathcal{C}(t + \Delta t)$  and  $\mathcal{C}(t)$  (see Figure 3.49), then the distance between these two points must be greater than or equal to  $c\Delta t$ , otherwise  $M$  would be situated in the burnt zone surrounding  $M'$  in time  $\Delta t$ .



**Figure 3.49.** Evolution of a flame front. The circle centered on  $M'$  and of radius  $c\Delta t$  represents the zone around this point burnt during the time  $\Delta t$ . The length of the segment  $MM'$  is  $\sqrt{(x - x')^2 + (p(t + \Delta t, x) - p(t, x'))^2}$ . This length thus must be  $\geq c\Delta t$

From a mathematical point of view, this is translated into the following property:

$$\forall (x, x') \in \mathbb{R}^2, \quad |x - x'|^2 + |p(t + \Delta t, x) - p(t, x')|^2 \geq c^2 \Delta t^2. \quad [3.54]$$

Furthermore, for all  $x \in \mathbb{R}$ , we show that

$$\inf_{x' \in \mathbb{R}} \{ |x - x'|^2 + |p(t + \Delta t, x) - p(t, x')|^2 \} = c^2 \Delta t^2. \quad [3.55]$$

Indeed, we deduce immediately from [3.54] that the infimum over  $x'$  present in [3.55] is greater than or equal to  $c\Delta t$ . Assuming that the infimum is strictly greater than  $c\Delta t$  implies that, for all  $x'$ , the points  $(x', p(t, x'))$  of the front  $\mathcal{C}(t)$  are such that  $|x - x'|^2 + |p(t + \Delta t, x) - p(t, x')|^2 > c^2 \Delta t^2$ . We can hence find points of coordinates  $(x, y)$  with  $y < p(t + \Delta t, x)$  and such that the distance to  $(x, p(t, x))$  – that is to say  $(y - p(t, x))$  – remains strictly greater than  $c\Delta t$ . These points, being under the curve  $\mathcal{C}(t + \Delta t)$ , should be burnt. Now, these are points of abscissa  $x$ , which cannot be affected from  $(x, p(t, x))$  in time  $\Delta t$  by moving at speed  $c$  (in Figure 3.49, these points are outside the gray circle). We thus arrive at a contradiction.

We seek to make use of [3.55] to obtain a partial differential equation describing the evolution of the front. We can again write [3.55] in the form

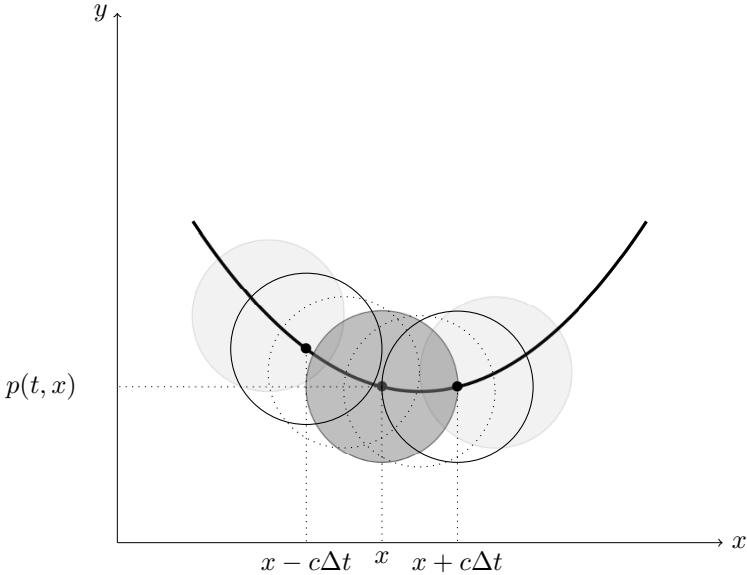
$$\sup_{x' \in \mathbb{R}} \{ c^2 \Delta t^2 - |x - x'|^2 - |p(t + \Delta t, x) - p(t, x')|^2 \} = 0.$$

In fact, we can restrict the set over which we take this supremum

$$\sup_{|x - x'| \leq c\Delta t} \{ c^2 \Delta t^2 - |x - x'|^2 - |p(t + \Delta t, x) - p(t, x')|^2 \} = 0. \quad [3.56]$$

(since when  $|x - x'| > c\Delta t$ , the evaluated quantity  $\{ \dots \}$  is strictly negative). We can continue to make use of this observation. A point with coordinates  $(x, y) = (x, p(t + \Delta t, x))$  is situated on the front  $\mathcal{C}(t + \Delta t)$  if it originates from a point on  $\mathcal{C}(t)$ . As the discs of center  $(x', p(t, x'))$  and radius  $c\Delta t$  are burnt between  $t$  and  $t + \Delta t$ , it follows that  $(x, p(t + \Delta t, x)) \in \bigcup_{x'} C((x', p(t, x')), c\Delta t)$ , the union of the circles centered on  $(x', p(t, x'))$ , of radius  $c\Delta t$ . In this discussion, the property of propagation at a finite speed is fundamental: the position of the front at abscissa  $x$  at time  $t + \Delta t$  depends only on the state of the front at time  $t$  for abscissae  $x' \in [x - c\Delta t, x + c\Delta t]$  (see Figure 3.50). The points whose abscissa  $x'$  is further from  $x$  than  $c\Delta t$  cannot influence the front at  $x$  at time  $t + \Delta t$ . We thus have  $(x, p(t + \Delta t, x)) \in \bigcup_{|x - x'| \leq c\Delta t} C((x', p(t, x')), c\Delta t)$ ; furthermore, since the front propagates in the direction of increasing ordinates, we have

$$p(t + \Delta t, x) \geq p(t, x') \quad \text{for all } |x - x'| \leq c\Delta t.$$



**Figure 3.50.** Evolution of a flame front

Thus, the quantity at stake in relation [3.56] appears now as the difference of the squares of two positive quantities. We thus deduce from [3.56] that we also have

$$\begin{aligned} \sup_{|x-x'| \leq c\Delta t} \{ \sqrt{c^2 \Delta t^2 - |x - x'|^2} - (p(t + \Delta t, x) - p(t, x')) \} &= 0. \\ &= \sup_{|x-x'| \leq c\Delta t} \{ \sqrt{c^2 \Delta t^2 - |x - x'|^2} + p(t, x') \} - p(t + \Delta t, x). \end{aligned}$$

In other words, for all  $x \in \mathbb{R}$ , we have

$$p(t + \Delta t, x) = \sup_{|x-x'| \leq c\Delta t} \{ p(t, x') + \sqrt{c^2 \Delta t^2 - |x - x'|^2} \}.$$

It follows that

$$\begin{aligned} \frac{p(t + \Delta t, x) - p(t, x)}{\Delta t} &= \sup_{|x-x'|/\Delta t \leq c} \left\{ \frac{p(t, x') - p(t, x)}{\Delta t} + \sqrt{c^2 - \frac{|x - x'|^2}{\Delta t^2}} \right\} \\ &= \sup_{|\xi| \leq c} \left\{ \frac{p(t, x + \xi \Delta t) - p(t, x)}{\Delta t} + \sqrt{c^2 - \xi^2} \right\}. \end{aligned}$$

By assuming that  $p$  is a sufficiently regular function and by taking  $\Delta t$  tending towards 0, we are thus led to

$$\frac{\partial}{\partial t} p(t, x) = \sup_{|\xi| \leq c} \left\{ \xi \frac{\partial}{\partial x} p(t, x) + \sqrt{c^2 - \xi^2} \right\}.$$

In order to obtain a more explicit expression, we introduce the function

$$F_h(\xi) = \xi h + \sqrt{c^2 - \xi^2}.$$

We note that  $F'_h(\xi) = h - \frac{\xi}{\sqrt{c^2 - \xi^2}}$  cancels when  $\xi$  takes the value

$$\xi_{\text{opt}} = \frac{c h}{\sqrt{1 + h^2}}, \quad [3.57]$$

while  $F''_h(\xi) = -\frac{1}{\sqrt{c^2 - \xi^2}} - \frac{\xi^2}{(c^2 - \xi^2)^{3/2}} < 0$ . Formula [3.57] thus determines, for fixed  $h$ , the maximum of the function  $F_h$ :

$$\sup_{\xi} F_h(\xi) = F_h(\xi_{\text{opt}}) = \frac{ch^2}{\sqrt{1 + h^2}} + \sqrt{c^2 - \frac{c^2 h^2}{1 + h^2}} = c\sqrt{1 + h^2}.$$

We deduce from this that  $(t, x) \mapsto p(t, x)$  satisfies the partial differential equation

$$\frac{\partial}{\partial t} p + f \left( \frac{\partial}{\partial x} p \right) = 0, \quad f(s) = -c\sqrt{1 + s^2}. \quad [3.58]$$

Hence, we set

$$u = \frac{\partial}{\partial x} p$$

which thus satisfies the conservation law

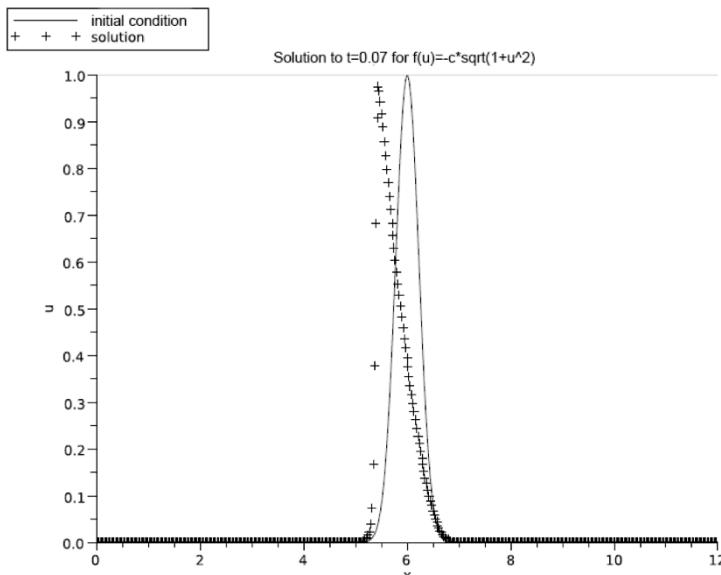
$$\frac{\partial}{\partial t} u + \frac{\partial}{\partial x} f(u) = 0. \quad [3.59]$$

The flux  $u \mapsto f(u)$  is determined by the solution of the optimization problem, which consists of maximizing, for fixed  $u$ ,  $\xi \mapsto u\xi + \sqrt{c^2 - \xi^2}$ :  $f$  thus corresponds to the *Legendre transformation* of the function  $\xi \mapsto \sqrt{c^2 - \xi^2}$ . We can now adopt two outlooks: either focus directly on the *Hamilton–Jacobi equation* [3.58] or focus on [3.59] and reconstruct  $p(t, x)$  by integration in space.

Before considering the numerical approximation of these equations, we can look back at the general arguments leading from equation [3.45]: here, we arrive at

$$u(t, x) = u_{\text{Init}}(X(0; t, x)), \quad X(t; 0, x) = x + t f'(u_{\text{Init}}(x)).$$

Obtaining  $X(0; t, x)$  comes back to taking the inverse of the function  $x \mapsto \phi_t(x) = x + t f'(u_{\text{Init}}(x))$ , at fixed  $t$ . Here, the function  $f$  defined in [3.58] is concave; thus, when the initial data  $u_{\text{Init}}$  decreases, the function  $\varphi_t$  increases for all  $t > 0$  and we can determine  $x$  as a function of  $X$ . This is no longer the case when  $u_{\text{Init}}$  is increasing and this translates to the appearance of singularities in finite time, as shown in Figure 3.51.



**Figure 3.51.** Formation of a singularity for equation [3.59]

The first numerical strategy directly follows the principles which guided the formulation of the equations. Indeed, we can express the solution in  $(t, x)$  in the form

$$p(t, x) = \sup_{-c \leq \xi \leq c} \{ p_{\text{Init}}(x + \xi t) + \sqrt{c^2 t^2 - \xi^2 t^2} \}. \quad [3.60]$$

Having discretized the domain of study  $[0, L]$  with a constant step size  $\Delta x > 0$ , we define the functions

$$p_j(t) = \sup_{|k-j| \leq ct/\Delta x} \{ p_{\text{Init}}(k\Delta x) + \sqrt{c^2 t^2 - (k-j)^2 \Delta x^2} \}, \quad [3.61]$$

for  $t > 0$  and  $j \in \{1, \dots, J\}$ , where  $J$  is the floor function of  $L/\Delta x$ . The function  $p_j(t)$  is used to approximate the exact value  $p(t, x_j)$  of the solution at the point of abscissa  $x_j = j\Delta x$ . We note that this relation directly supplies the solution approximated in an arbitrary time  $t > 0$  and is not based on an iterative process. The practical difficulty comes from the fact that this definition involves, for a given time  $t$  in  $[0, T]$ , points  $x_k$  which are outside the domain of study (and even more points outside  $[0, L]$  are needed as  $t$  increases). We must thus use data  $p_{\text{Init}}$  defined over a sufficiently large interval, containing  $[0, L]$ . Figure 3.53 shows the evolution of the front obtained by this method, with  $L = 12$ ,  $c = 2$ ,  $\Delta x = 0.02$  and initial data

$$p_{\text{Init}}(x) = \mathbf{1}_{[1.8, 6.6]}(x)(2.4 - |x - 4.2|) + \mathbf{1}_{[6.6, 9]}(x)(0.6 - b|x - 7.8|), \quad [3.62]$$

where  $b = 0.5$  and  $\mathbf{1}_A$  signifies the characteristic function of a set  $A$ . This function is extended beyond the domain  $[0, L]$  in a constant manner.

The second approach consists of solving the conservation law [3.59] instead, for which we have schemes that were studied earlier. For example, we can put in place the Lax–Friedrichs scheme for [3.59]

$$u_j^{n+1} - u_j^n = -\frac{\Delta t}{2\Delta x}(f(u_{j+1}^n) - f(u_{j-1}^n)) + \frac{u_{j+1}^n - 2u_j^n + u_{j-1}^n}{2}.$$

The unknown values correspond to the indices  $j \in \{1, \dots, J = L/\Delta x\}$ . The scheme involves supplementary values which we fix by setting

$$u_0 = u_1 \quad \text{and} \quad u_{J+1} = u_J.$$

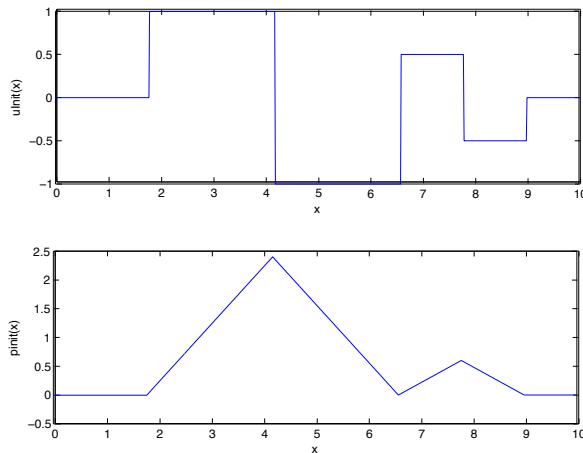
As initial data, we take<sup>5</sup>

$$u_{\text{Init}}(x) = \mathbf{1}_{[1.8, 4.2]}(x) + b\mathbf{1}_{[6.6, 7.8]}(x) - \mathbf{1}_{[4.2, 6.6]}(x) - b\mathbf{1}_{[7.8, 9]}(x), \quad [3.63]$$

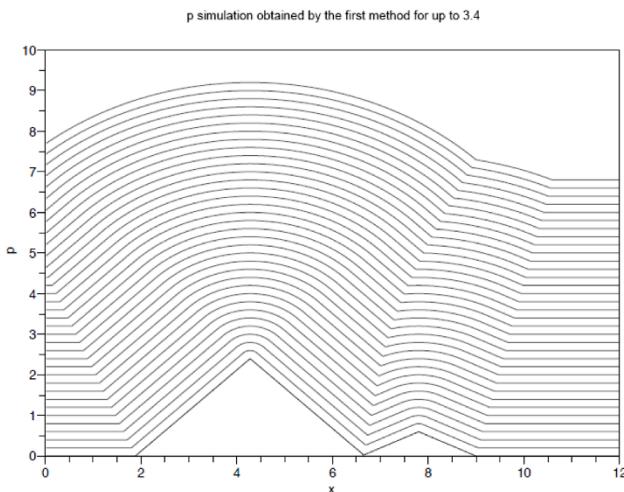
with  $b = 0.5$ . Of course [3.63] is nothing other than the derivative of the function  $p_{\text{Init}}$  shown in Figure 3.62 (see Figure 3.52). Figure 3.54 shows the evolution of the solution  $u$  of the conservation law [3.59] obtained by this scheme. Figure 3.55 gives the evolution of  $p$  satisfying  $\partial_t p = -f(u)$  corresponding to these data, that is to say that we determine  $p^{n+1}$  from  $u^{n+1}$  with the formula  $p_j^{n+1} = p_j^n - \Delta t f(u_j^{n+1})$  where  $p_j^0 = \Delta x \sum_{k=1}^j u_k^0$ . Here, the flux function  $f$  given by [3.58] is *concave*. In agreement with Lax's criterion of lemma 3.5, the admissible discontinuities for [3.59] must satisfy  $u_+ > u_-$ . This distinction between discontinuities is clearly visible in Figures 3.54 and 3.52: the initial discontinuities where  $u$  is decreasing are regularized, the only discontinuities where  $u$  is growing persist. For the front, this

<sup>5</sup> These data are “well chosen”; by considering other initial states, we obtain results that are qualitatively of the same nature but with possible numerical anomalies localized over certain points.

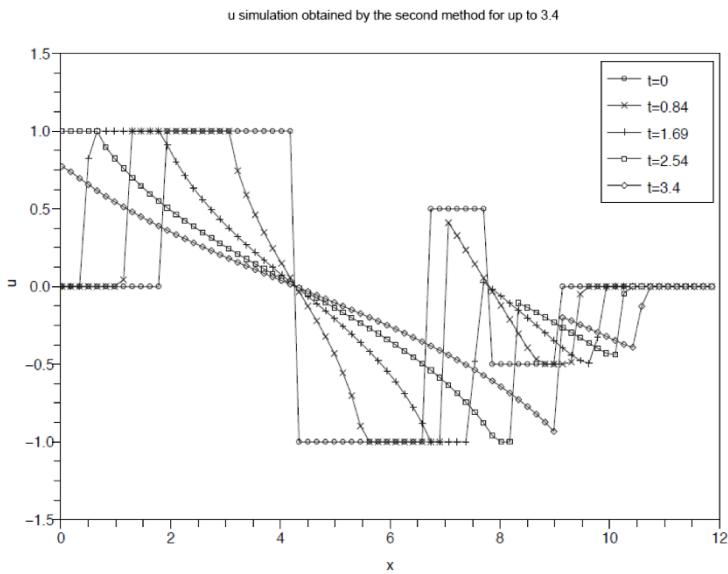
property is translated by the fact that the reflex angles  $\vee$  are maintained, while the salient angles  $\wedge$  are regularized, as we can see in Figures 3.53 and 3.56. We also clearly see the movement of these reflex angles in Figure 3.56. To illustrate this important difference in behavior, we conduct a test by inverting the orientation of triangles (see Figure 3.57). The evolution of these data is shown in Figure 3.58: the outward angles are well regularized, and the inward angles persisting, producing a clearly visible singularity.



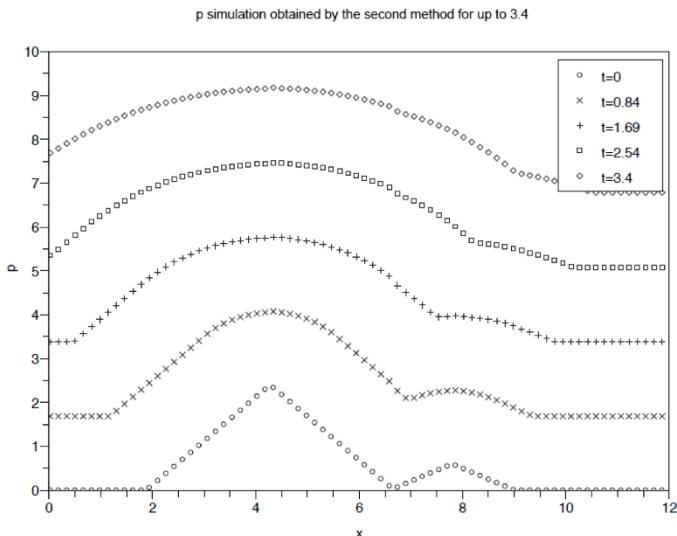
**Figure 3.52.** Initial data for the combustion problem



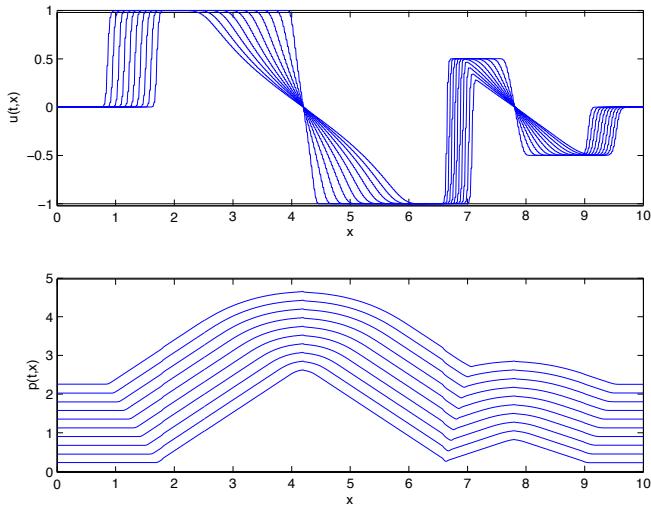
**Figure 3.53.** Evolution of the flame front for the data in Figure 3.52



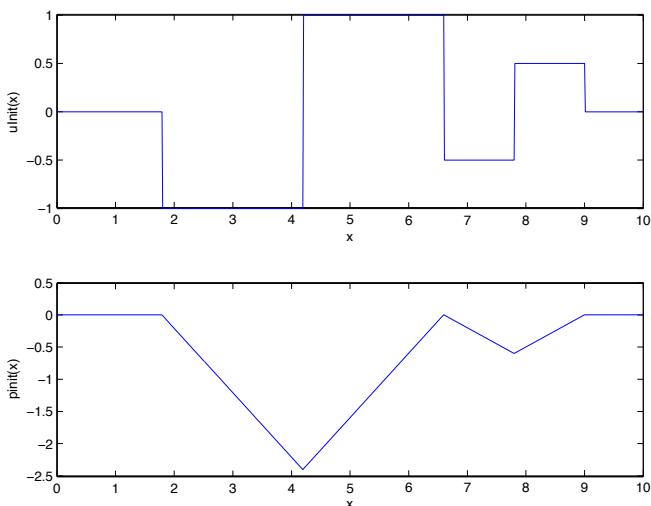
**Figure 3.54.** Evolution of the solution  $u$  of the nonlinear problem for the data in Figure 3.52



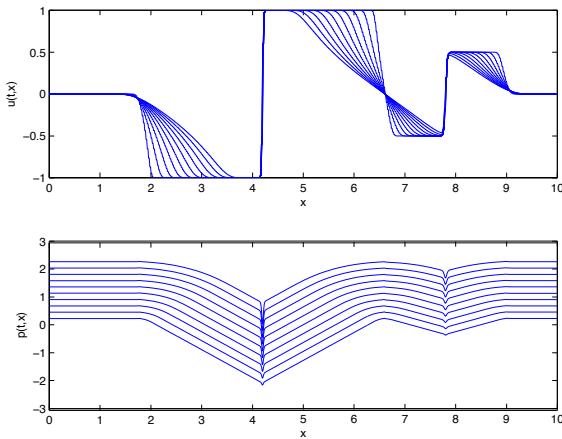
**Figure 3.55.** Evolution of the flame front for the data in Figure 3.52



**Figure 3.56.** Evolution of the flame front for the data in Figure 3.52:  
profiles of  $p$  and  $u$  at times  $t = 0.5632$ ,  $t = 0.6758$ ,  $t = 0.7885$   
 $t = 0.9011$ ,  $t = 1.0138$  and  $t = 1.1264$



**Figure 3.57.** Initial data of the combustion problem



**Figure 3.58.** Evolution of the flame front for the data in Figure 3.57: profiles of  $p$  and  $u$  at times  $t = 0.5632$ ,  $t = 0.6758$ ,  $t = 0.7885$ ,  $t = 0.9011$ ,  $t = 1.0138$  and  $t = 1.1264$

### 3.4.2. Systems of conservation laws

#### 3.4.2.1. Introduction to gas dynamics equations

We saw slightly earlier that the (scalar) wave equation could be interpreted as a linear system of conservation laws. This point of view guided the design of a numerical scheme based on the principles of upwinding; we hence identified a vector version of the Lax–Friedrichs scheme. We will now focus on nonlinear systems of conservation laws, in one spatial dimension:

$$\partial_t u + \partial_x f(u) = 0 \quad [3.64]$$

with  $u : [0, T] \times \mathbb{R} \rightarrow \mathbb{R}^d$  and  $f : \mathbb{R}^d \rightarrow \mathbb{R}^d$ . More specifically, we concentrate on systems motivated by fluid mechanics:

- The isentropic Euler equations. This is a system where  $d = 2$ , the unknowns being the density  $\rho$  and the momentum  $J$ . It is sometimes convenient to set  $J = \rho v$ , which thus defines the material velocity  $v$ . This system involves a *pressure law*  $\rho \mapsto p(\rho)$ . For simplification, we often take  $p(\rho) = \kappa \rho^\gamma$  for  $\kappa > 0$  and  $\gamma \geq 1$ , but the relations used by engineers are of a phenomenological nature and much more complex. The system is written in the form [3.64] with

$$u = (\rho, J), \quad f(u) = (J, J^2/\rho + p(\rho)).$$

– The (complete) Euler equations where this time  $d = 3$ . The unknowns are the density  $\rho$ , the momentum  $J = \rho v$ ,  $v$  being the material velocity, and the total energy  $\mathcal{E}$ . The system is completed by a state law which defines the pressure as a function of the density and internal energy  $e$ ; more specifically, by defining the kinetic energy as  $\rho \frac{v^2}{2}$ , we decompose the total energy into the form

$$\mathcal{E} = \rho \left( \frac{v^2}{2} + e \right),$$

and we have  $p = p(\rho, e)$ . For example, for monatomic gases (in one dimension<sup>6</sup>) we set  $p = \rho\theta$  where  $2e = \theta$ , the temperature.

The phenomena displayed for scalar conservation laws are of course recovered for these systems: it is common for regular data to lead to solutions with discontinuities in finite time. The theory for the systems is much less developed than for scalar equations, notably for multi-dimensional frameworks. Nevertheless, we understand that the challenge of constructing numerical schemes must cover the following points:

- reproducing the discontinuities and propagating them correctly;
- preserving certain *a priori* estimates, which are, in general, critical from a physical point of view (positivity of the density and the internal energy);
- dissipating certain quantities conserved by the regular solutions of the continuous problem.

Although we know that these manipulations are “forbidden” for discontinuous solutions, it is interesting to write the system in the non-conservative form as follows:

$$\partial_t u + f'(u) \partial_x u = 0$$

where  $f'(u)$  represents the Jacobian matrix of the system. From this, we interpret certain structural properties of the system.

- For the isentropic Euler, the conserved quantities are  $u = (\rho, J)$  and the flux is expressed as  $f(u) = (J, J^2/\rho + p(\rho))$ . We thus obtain

$$f'(u) = \begin{pmatrix} 0 & 1 \\ p'(\rho) - \frac{J^2}{\rho^2} & 2\frac{J}{\rho} \end{pmatrix}.$$

---

<sup>6</sup> In fact, the relation between internal energy  $e$  and temperature  $\theta$  and the definition of the pressure law include the spatial dimension  $N$ :  $e = \frac{N}{2}\theta$  and  $p = \rho\theta$ .

The eigenvalues of the system are  $\lambda_{\pm} = \frac{J}{\rho} \pm \sqrt{p'(\rho)}$ . In order that the quantity  $\sqrt{p'(\rho)}$ , which is interpreted as the speed of sound, be real, the pressure  $p$  must be assumed to be a monotonically increasing function of the density.

– For monatomic gases, the conserved quantities are  $u = (\rho, J, \mathcal{E})$  and the flux is expressed as

$$f(u) = \left( J, \frac{J^2}{\rho} + p, (\mathcal{E} + p) \frac{J}{\rho} \right),$$

with

$$\mathcal{E} = \rho \left( \frac{J^2}{2\rho^2} + \frac{\theta}{2} \right), \quad p = \rho\theta,$$

that is to say

$$f(u) = \left( J, 2\mathcal{E}, 3\mathcal{E} \frac{J}{\rho} - \frac{J^3}{\rho^2} \right).$$

We thus arrive at

$$f'(u) = \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 2 \\ \frac{J}{\rho^2} \left( 2 \frac{J^2}{\rho} - 3\mathcal{E} \right) & \frac{3}{\rho} \left( \mathcal{E} - \frac{J^2}{\rho} \right) & 3 \frac{J}{\rho} \end{pmatrix}.$$

The eigenvalues are  $v = \frac{J}{\rho}$  and  $v \pm c$ , with a speed of sound  $c = \sqrt{3\theta}$ .

The eigenvalues of  $f'(u)$  are important: they describe the characteristic speeds of the system. In particular, in these equations, the information propagates at a finite speed and these eigenvalues determine the fastest speeds of propagation, which is useful for setting the numerical parameters.

**NOTE 3.7.–** In these relations, we must be aware that  $f'(u)$  signifies the derivative of the flux function  $f$  with respect to the conservative variables (we could also write this  $\nabla_u f(u)$ ). For the Euler equations, we thus differentiate the components of  $f$  with respect to  $\rho$  and  $J$  (isentropic case) or with respect to  $\rho$ ,  $J$  and  $\mathcal{E}$  and not with respect to the auxiliary variables density, material velocity or density, material velocity and temperature. This distinction is important to avoid confusion when it is sometimes convenient to express the calculations with these auxiliary quantities.

NOTE 3.8.– Of course, a state with a constant density  $\bar{\rho} > 0$  and a null material velocity  $v$  constitutes a particular solution of the isentropic Euler equations. We can focus on states which are a “small” perturbation of this particular solution:  $\rho = \bar{\rho} + \tilde{\rho}$ ,  $v = 0 + \tilde{v}$ . By assuming that the fluctuations  $\tilde{\rho}, \tilde{v}$  and their derivatives are small, we obtain the following linear system:

$$\begin{pmatrix} 1 & 0 \\ \bar{\rho} & 0 \end{pmatrix} \partial_t \begin{pmatrix} \tilde{\rho} \\ \tilde{v} \end{pmatrix} + \begin{pmatrix} 0 & \bar{\rho} \\ p'(\bar{\rho}) & 0 \end{pmatrix} \partial_x \begin{pmatrix} \tilde{\rho} \\ \tilde{v} \end{pmatrix} = 0.$$

We recover the wave equation with speed  $c = \sqrt{p'(\bar{\rho})}$ .

### 3.4.2.2. Lax–Friedrichs and Rusjanov schemes for the Euler equations

The Rusjanov scheme is constructed with the help of the centered flux, which we correct with a numerical diffusion proportional to the greatest characteristic speeds of the system. We thus set

$$u_j^{n+1} = u_j^n - \frac{\Delta t}{\Delta x} (F_{j+1/2}^n - F_{j-1/2}^n)$$

where the numerical fluxes are defined by

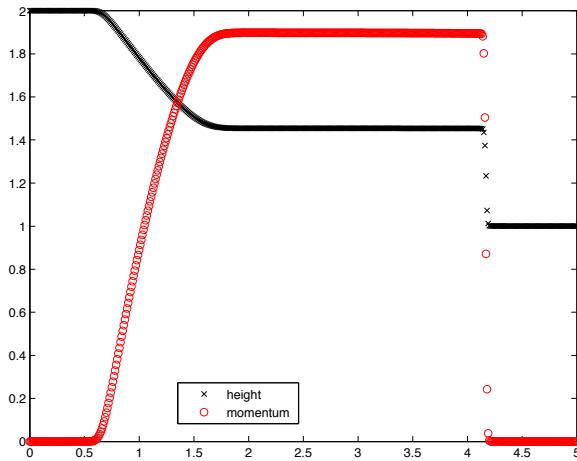
$$F_{j+1/2}^n = \mathbb{F}(u_{j+1}^n, u_j^n) = \frac{f(u_{j+1}^n) + f(u_j^n)}{2} - \frac{1}{2} \Lambda_{j+1/2} (u_{j+1}^n - u_j^n),$$

$$\Lambda_{j+1/2} = \max (\text{Spectral Radius}(f'(u_{j+1}^n)), \text{Spectral Radius}(f'(u_j^n))).$$

We fix the time step for each iteration so as to ensure that

$$\Delta t < \frac{\Delta x}{\max_j(\Lambda_{j+1/2}^n)}.$$

We conduct simulations for the case  $\gamma = 2$ : we often make reference to this system under the name of the *Saint-Venant equations*. It is used to describe the surface run-offs in shallow water (“Shallow–Water Systems”):  $\rho$  is then interpreted as the depth of the water and  $v = J/\rho$  is the velocity field. More specifically, we here set  $p(\rho) = 9.81 \frac{\rho^2}{2}$ . The domain of calculation is the interval  $[0, 5]$ . We set  $\Delta x = 1/100$ . The final time is  $T = 0.4$  (before the run-off interacts with the boundary). The initial data corresponds to a null velocity field  $v(0, x) = 0$  and a discontinuous water depth  $\rho(0, x) = 2 \times \mathbf{1}_{0 \leq x \leq 2.5} + \mathbf{1}_{2.5 \leq x \leq 5}$ . The solution obtained by the Rusjanov scheme is shown in Figure 3.59. We distinguish three states where densities and speeds are constant, connected by a discontinuity in the water depth and speed on the right of the graphic, or by a regular profile on the left of the graphic. These structures develop and propagate, precisely following the characteristic speeds of the model.



**Figure 3.59.** Simulation of the Saint-Venant system by the Rusjanov scheme. For the color version of this figure, see [www.iste.co.uk/goudon/mathmodel.zip](http://www.iste.co.uk/goudon/mathmodel.zip)

We then simulate, still with the Rusjanov scheme, the Euler equations for a monatomic gas over  $]0, 1[$ . We set  $\Delta x = 10^{-3}$  and  $\Delta t = 0.95\Delta x$ . We test different initial data with a discontinuity at  $x = 1/2$ , of the form:

$$\begin{aligned}\rho_{\text{Init}}(x) &= \rho_g \mathbf{1}_{x \leq 0.5} + \rho_d \mathbf{1}_{0.5 < x \leq 1}, \\ u_{\text{Init}}(x) &= u_g \mathbf{1}_{x \leq 0.5} + u_d \mathbf{1}_{0.5 < x \leq 1}, \\ p_{\text{Init}}(x) &= p_g \mathbf{1}_{x \leq 0.5} + p_d \mathbf{1}_{0.5 < x \leq 1}.\end{aligned}$$

We thus have

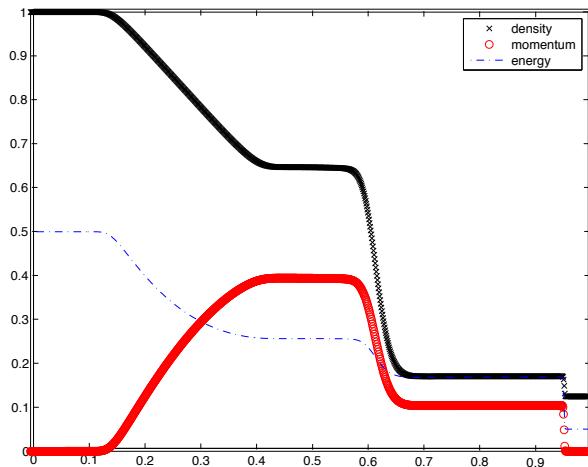
$$U_{\text{Init}}(x) = \left( \rho_{\text{Init}}, \rho_{\text{Init}} u_{\text{Init}}, \frac{1}{2} (\rho_{\text{Init}} u_{\text{Init}}^2 + p_{\text{Init}}) \right).$$

The parameters are fixed in the following manner:

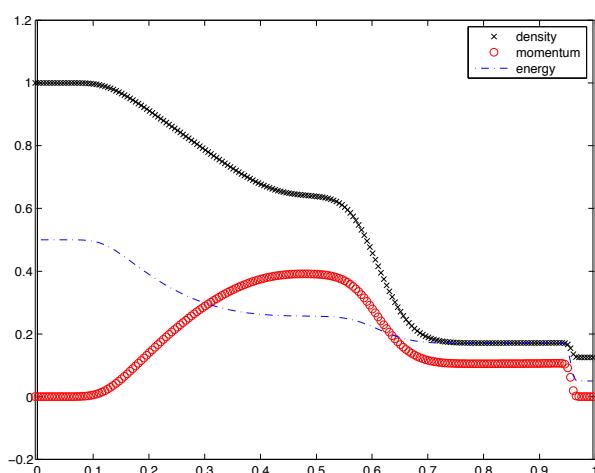
	$\rho_g$	$\rho_d$	$u_g$	$u_d$	$p_g$	$p_d$	$T_{\text{final}}$
Case 1	1	0.125	0	0	1	0.1	0.2
Case 2	0.1709	0.125	0.1044	0	0.1689	0.05	0.2
Case 3	0.125	0.1709	0	-0.1044	0.1689	0.2	0.05
Case 4	1	1	-10	10	0.01	0.01	0.008

For these simulations, we use the so-called “wall” boundary conditions: the unknown values outside of the calculation domain (i.e. for spatial indices  $j = 0$  and  $j = J + 1$ ) are evaluated by keeping the same density and pressure as for the neighboring point of the mesh but by inverting the direction of the velocity.

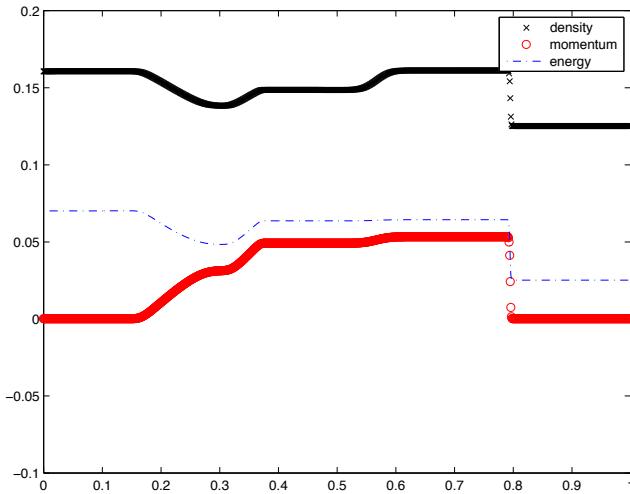
Figure 3.60 gives the result obtained for Case 1. Here, we see the formation of three states where  $\rho$ ,  $\rho u$  and  $\mathcal{E}$  are constant, connected by transition phases which can be discontinuous. Figure 3.61 corresponds to the same case, but with a discretization step size 5 times larger: the discontinuities are hence poorly resolved.



**Figure 3.60.** Simulation of the Euler equations: Case 1 ( $\Delta x = 10^{-3}$ ).  
For the color version of this figure, see  
[www.iste.co.uk/goudon/mathmodel.zip](http://www.iste.co.uk/goudon/mathmodel.zip)



**Figure 3.61.** Simulation of the Euler equations: Case 1 ( $\Delta x = 5 \times 10^{-3}$ ). For the color version of this figure, see  
[www.iste.co.uk/goudon/mathmodel.zip](http://www.iste.co.uk/goudon/mathmodel.zip)



**Figure 3.62.** Simulation of the Euler equations: Case 2 ( $\Delta x = 10^{-3}$ ).

For the color version of this figure, see  
[www.iste.co.uk/goudon/mathmodel.zip](http://www.iste.co.uk/goudon/mathmodel.zip)

Figures 3.62 and 3.63 display the solutions for Cases 2 and 3, where we have simply exchanged the states on the left- and right-hand sides, by changing the sign of the velocity. We observe that the numerical solutions indeed preserve this symmetry, which we can verify over the continuous problem. Figure 3.64 shows the formation of a vacuum in the configuration of Case 4; see in particular the density profile in Figure 3.65. we note the formation of peaks in the transition phases.

### 3.4.3. Kinetic schemes

We shall briefly present a family of schemes, schemes said to be *kinetic*, where the formulation of numerical fluxes is inspired by principles of statistical physics. Consult the seminal articles [COR 91, DES 86a, DES 86b, PER 92, PUL 80] or the review [PER 02] for details regarding the development of these techniques and various applications. The design of these kinetic schemes follows the approach that allows the fluid mechanics equations to be obtained from models of a more microscopic nature, the Boltzmann equation in particular, see [SAI 09]. The idea is to interpret the conservation law [3.43] as originating from the limit  $\varepsilon \rightarrow 0$  of

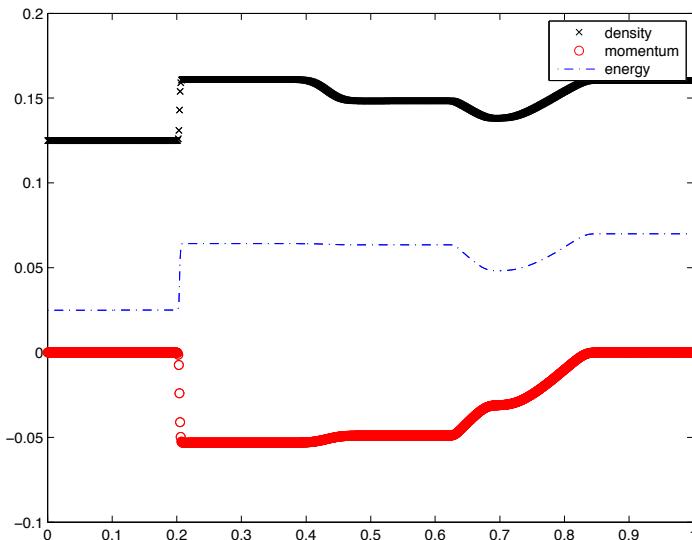
$$\partial_t f_\varepsilon + a(v) \partial_x f_\varepsilon = \frac{1}{\varepsilon} Q(f_\varepsilon) \quad [3.65]$$

In the physical cases, the unknown  $f_\varepsilon(t, x, v)$  in [3.65] is interpreted as a density in phase space: it thus depends not only on the time and space variables but also on an

auxiliary variable  $v \in \mathbb{R}$ . The equation involves a “velocity” function  $a : \mathbb{R} \rightarrow \mathbb{R}$ , we will see how this is defined as a function of flux in [3.43]. The hope is thus to obtain inherently

– a scheme that is simple. In particular, the terms involving the derivatives in [3.65] are linear: the nonlinearities are now contained in the operator  $Q$ , which takes the form of an operator of integral type in the variable  $v$ . The introduction of the additional variable  $v$  can be seen as the price to pay for this simplicity. However, we shall see that the role of this variable is only fictitious and that the scheme obtained can be written without the appearance of this variable.

– a scheme that preserves the crucial properties of model [3.43] (conservation, dissipation, etc.), as a consequence of the fundamental physical properties, contained in the details of the interaction operator  $Q$ .



**Figure 3.63.** Simulation of the Euler equations: Case 3 ( $\Delta x = 10^{-3}$ ).

For the color version of this figure, see  
[www.iste.co.uk/goudon/mathmodel.zip](http://www.iste.co.uk/goudon/mathmodel.zip)

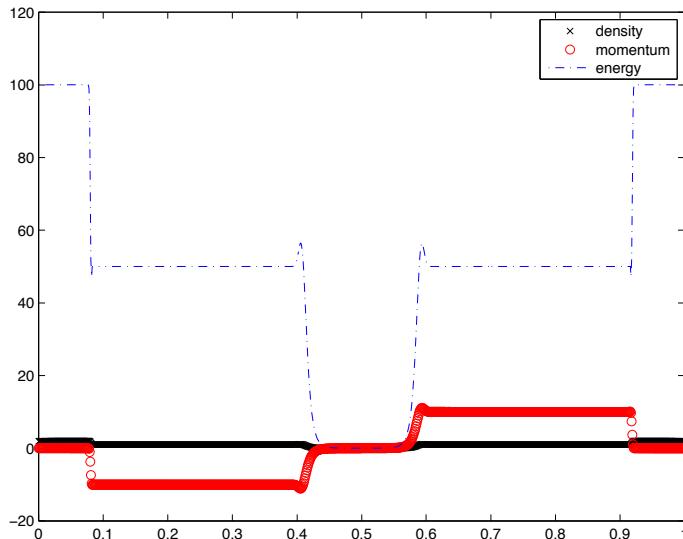
### 3.4.3.1. Scalar conservation laws

We are interested in the following model, introduced in [PER 91],

$$Q(f) = \chi_u - f, \quad u(t, x) = \int f(t, x, v) dv \quad [3.66]$$

where

$$\chi_u(v) = \begin{cases} +1 & \text{if } 0 < v < u, \\ -1 & \text{if } u < v < 0, \\ 0 & \text{otherwise.} \end{cases} \quad [3.67]$$



**Figure 3.64.** Simulation of the Euler equations: Case 4 ( $\Delta x = 10^{-3}$ ).  
For the color version of this figure, see  
[www.iste.co.uk/goudon/mathmodel.zip](http://www.iste.co.uk/goudon/mathmodel.zip)

We note that

$$\int \chi_{u(t,x)} dv = \int \mathbf{1}_{0 < v < u(t,x)} dv - \int \mathbf{1}_{u(t,x) < v < 0} dv = u(t, x) = \int f(t, x, v) dv.$$

We deduce from this the following conservation law, associated with [3.65]–[3.67]

$$\partial_t \int f(t, x, v) dv + \partial_x \int a(v) f(t, x, v) dv = 0.$$

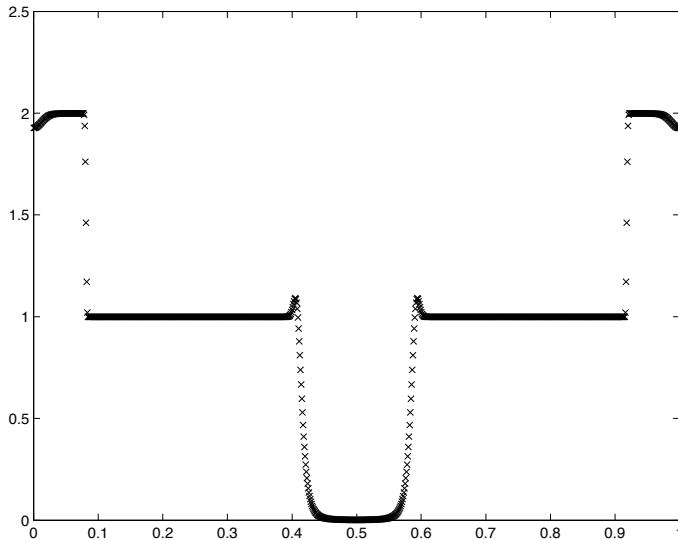
Indeed, this relation is obtained by simply integrating [3.65] over the variable  $v$ . When  $\varepsilon \ll 1$ , we expect that  $Q(f_\varepsilon)$  tends towards 0, i.e. in the present case that

$f_\varepsilon(t, x, v)$  resembles  $\chi_{u_\varepsilon(t, x)}(v)$ , with  $u_\varepsilon(t, x) = \int f_\varepsilon(t, x, v) dv = \int \chi_{u_\varepsilon(t, x)}(v) dv$ . Specifically, we can write

$$\begin{aligned} \partial_t \int f_\varepsilon(t, x, v) dv + \partial_x \int a(v) f_\varepsilon(t, x, v) dv &= 0 \\ &= \partial_t \int \chi_{u_\varepsilon(t, x)}(v) dv + \partial_x \int a(v) \chi_{u_\varepsilon(t, x)}(v) dv \\ &\quad + \partial_x \int a(v) (f_\varepsilon(t, x, v) - \chi_{u_\varepsilon(t, x)}(v)) dv \end{aligned}$$

where the final term is in fact

$$-\varepsilon \partial_x \int a(v) (\partial_t + a(v) \partial_x) f_\varepsilon(t, x, v) dv.$$



**Figure 3.65.** Simulation of the Euler equations: Case 4, graph of the density ( $\Delta x = 10^{-3}$ )

Evidently, it still remains to establish a priori estimates to justify that the term is indeed, in an appropriate sense, of order  $\varepsilon$ . Thus, as  $\varepsilon \rightarrow 0$ , provided that  $f_\varepsilon(t, x, v) \rightarrow \chi_{u(t, x)}(v)$ , we obtain

$$\partial_t \int \chi_{u(t, x)}(v) dv + \partial_x \int a(v) \chi_{u(t, x)}(v) dv = \partial_t u + \partial_x A(u) = 0$$

with

$$A'(u) = a(u).$$

Furthermore, the model induces a dissipation property which arises from the following fact.

LEMMA 3.9.– Let  $v \mapsto f(v)$  be a function, such that  $-1 \leq f(v) \leq 1$  and  $\text{sgn}(v)f(v) \geq 0$ . Let  $H$  be a monotonically increasing function. Then, we have  $\int (M[\rho_f] - f)H \, dv \leq 0$ .

PROOF.– We make use of the conservation property  $\int (\chi_u - f) \, dv$ , where  $u = \int f \, dv$  to write

$$\begin{aligned} I &= \int (\chi_u - f)H \, dv = \int (\chi_u(v) - f(v))(H(v) - H(u)) \, dv \\ &= \int_0^{[u]_+} \underbrace{(1 - f(v))}_{\geq 0} \underbrace{(H(v) - H(u))}_{\leq 0} \, dv + \int_{[u]_+}^{\infty} \underbrace{(-f(v))}_{\leq 0} \underbrace{(H(v) - H(u))}_{\geq 0} \, dv \\ &\quad + \int_{[u]_-}^0 \underbrace{(-1 - f(v))}_{\leq 0} \underbrace{(H(v) - H(u))}_{\geq 0} \, dv + \int_{-\infty}^{[u]_-} \underbrace{(-f(v))}_{\geq 0} \underbrace{(H(v) - H(u))}_{\leq 0} \, dv \\ &\leq 0. \end{aligned}$$

□

We can show that, if the initial data  $f_0$  for [3.66] is contained within  $-1$  and  $+1$  with the same sign as  $\xi$ , then this property propagates with time. As a consequence, lemma 3.9 implies that, for all convex functions  $\eta$ , we have

$$\partial_t \int \eta'(\xi) f_\varepsilon \, d\xi + \partial_x \int a(v) \eta'(\xi) f_\varepsilon \, dv \leq 0.$$

By again allowing the convergence of  $f_\varepsilon$  towards  $\chi_u$ , this leads to the entropy inequalities

$$\partial_t \eta(u) + \partial_x q(u) \leq 0$$

with  $q'(u) = \eta'(u)a(u)$ .

This link between microscopic and macroscopic models, established by B. Perthame and E. Tadmor [PER 91], allows for the designing of numerical schemes for the conservation law  $\partial_t u + \partial_x A(u) = 0$ . The principle rests on an approximation of [3.65] by a *time splitting*: knowing an approximation of  $f$  at time  $n\Delta t$ :

– We first solve the free transport equation

$$\partial_t f + a(v) \partial_x f = 0$$

over the time interval  $[n\Delta t, (n+1)\Delta t]$ . To this end, we can use a simple upwind scheme. We thus obtain, in a version completely discretized (in time and space)

$$f_j^*(v) = f_j^n(v) - \frac{\Delta t}{\Delta x} (a_+(v)(f_j^n - f_{j-1}^n)(v) + a_-(v)(f_{j+1}^n - f_j^n)(v))$$

where we write

$$a_\pm(v) = \frac{a(v) \pm |a(v)|}{2}.$$

– We solve the stiff equation

$$\partial_t f = \frac{1}{\varepsilon} (\chi_u(v) - f)$$

with  $f^*$  as initial data. The key point lies in the fact that the macroscopic quantity  $u(t, x)$  is not modified during this stage since the integration over  $v$  gives

$$\partial_t \int f \, dv = \partial_t u = \frac{1}{\varepsilon} \int (\chi_u(v) - f) \, dv = 0.$$

Thus,  $u$  remains constant, given by  $u^*$ . We thus obtain

$$f_j^{n+1}(v) = f_j^*(v) e^{-\Delta t/\varepsilon} + \chi_{u_j^*}(v) (1 - e^{-\Delta t/\varepsilon}),$$

with  $u_j^{n+1} = \int f_j^{n+1} \, dv = u_j^*$ .

When  $\varepsilon \rightarrow 0$ , the second stage degenerates into a simple projection

$$f_j^{n+1}(v) = \chi_{u_j^*}(v).$$

We thus obtain a numerical scheme for the conservation law by integrating over the variable  $v$  (which is thus just an artifice that does not intervene in the final scheme):

$$\begin{aligned} u_j^{n+1} &= \int f_j^*(v) \, dv \\ &= \int \left( \chi_{u_j^n}(v) - \frac{\Delta t}{\Delta x} (a_+(v)(\chi_{u_j^n} - \chi_{u_{j-1}^n})(v) + a_-(v)(\chi_{u_{j+1}^n} - \chi_{u_j^n})(v)) \right) \, dv \\ &= u_j^{n+1} - \frac{\Delta x}{\Delta t} (F_{j+1/2}^n - F_{j-1/2}^n) \end{aligned}$$

where the numerical fluxes are defined by

$$F_{j+1/2}^n = \mathbb{F}(u_{j+1}^n, u_j^n) = \int (a_-(v)\chi_{u_{j+1}^n}(v) + a_+(v)\chi_{u_j^n}(v)) dv.$$

The scheme thus obtained is indeed flux consistent (in the sense of definition 3.6), since

$$\mathbb{F}(u, u) = \int a(v)\chi_u(v) dv = A(u),$$

by making use of the fact that  $a_+(v) + a_-(v) = a(v)$ . This formula defines a monotone scheme, since

$$\mathbb{F}(u_1, u_2) = A_-(u_1) + A_+(u_2), \quad A'_\pm(z) = [a(z)]_\pm$$

is monotonically decreasing in the first variable, and monotonically increasing in the second variable. Actually, we thus recover the Enquist–Osher scheme.

**PROPOSITION 3.9.–** Over a uniform mesh of step size  $\Delta x$ , we assume that the stability condition

$$\sup |a(v)| \leq \frac{\Delta t}{\Delta x}$$

is satisfied. The kinetic scheme thus exhibits the following properties

- i) Maximum principle:  $\min_j u_j^0 \leq u_j^n \leq \max_j u_j^0$ .
- ii) Conservation of mass:  $\sum_j u_j^n = \sum_j u_j^0$ .
- iii) Stability: for all convex functions  $\eta$ , we have

$$\sum_j \eta(u_j^n) \leq \sum_j \eta(u_j^0),$$

**PROOF.–** By returning to the transport stage of the scheme, we have

$$f_j^*(v) = f_j^n \left( 1 - \frac{\Delta t}{\Delta x} |a(v)| \right) + \frac{\Delta t}{\Delta x} a_+(v) f_{j-1}^n - \frac{\Delta t}{\Delta x} a_-(v) f_{j+1}^n \quad [3.68]$$

where  $f_j^n(v) = \chi_{u_j^n}(v)$ . If  $v \geq 0$  with  $v \geq \sup\{u_j^n, u_{j-1}^n, u_{j+1}^n\}$ , then all the functions  $\chi$  appearing in this formula cancel:  $f_j^n = f_{j-1}^n = f_{j+1}^n = 0$  and thus  $f_j^* = 0$ . The same conclusion applies when  $v \leq 0$  with  $v \leq \inf\{u_j^n, u_{j-1}^n, u_{j+1}^n\}$ . When  $v \geq 0$ ,  $u_j^n, u_{j-1}^n, u_{j+1}^n \geq 0$  and  $v \leq \sup\{u_j^n, u_{j-1}^n, u_{j+1}^n\}$ , then

$f_j^n = f_{j-1}^n = f_{j+1}^n = 1$  and thus  $f_j^\star = 1 - \frac{\Delta t}{\Delta x} |a(v)| + \frac{\Delta t}{\Delta x} (a_+(v) - a_-(v)) = 1$ . The same approach is adapted when  $v \leq 0$ ,  $u_j^n, u_{j-1}^n, u_{j+1}^n \leq 0$  with  $v \geq \inf\{u_j^n, u_{j-1}^n, u_{j+1}^n\}$  and leads to  $f_j^\star = -1$ . In all other cases,  $f_j^n$  appears as a convex combination of functions  $\chi_{u_j^n}$ ,  $\chi_{u_{j-1}^n}$  and  $\chi_{u_{j+1}^n}$ , as a result of the stability condition. The maximum principle follows by integration over the variable  $v$ . The stability stated in iii) is a consequence of lemma 3.9. Indeed, on the one hand, we have

$$\int \eta'(v) f_j^{n+1}(v) dv = \int \eta'(v) \chi_{u_j^{n+1}}(v) dv = \eta(u_j^{n+1})$$

and, on the other hand, following lemma 3.9

$$\eta(u_j^{n+1}) = \eta(u_j^\star) = \int \eta'(v) \chi_{u_j^\star}(v) dv \leq \int \eta'(v) f_j^\star(v) dv,$$

since  $f_j^\star(v) \in [-1, +1]$  has the sign of  $v$  and  $\int f_j^\star(v) dv = u_j^\star$ . We conclude by returning to expression [3.68] and summing over  $j$ .  $\square$

### 3.4.3.2. Kinetic scheme for Euler equations: monatomic case

The Euler equations are written as

$$\partial_t \begin{pmatrix} \rho \\ \rho u \\ \rho E \end{pmatrix} + \partial_x \begin{pmatrix} \rho u \\ \rho u^2 + p \\ (\rho E + p)u \end{pmatrix} = 0$$

where, in the monatomic case, pressure  $p$  and energy  $E$  are defined by the following relations:

$$p = \rho\theta, \quad E = \frac{u^2}{2} + e, \quad e = \frac{\rho\theta}{2} = \frac{p}{2}.$$

This system can be seen, at least formally (see [SAI 09] for the mathematical analysis of these asymptotic regimes), as corresponding to the limit  $\epsilon \rightarrow 0$  of the following BGK equation:

$$\partial_t f_\varepsilon + v \partial_x f_\varepsilon = \frac{1}{\varepsilon} (M[f_\varepsilon] - f_\varepsilon),$$

where  $M[f_\varepsilon]$  is the Maxwellian state

$$M[f_\varepsilon](v) = \frac{\rho_\varepsilon}{\sqrt{2\pi\theta_\varepsilon}} \exp\left(-\frac{|v - u_\varepsilon|^2}{2\theta_\varepsilon}\right)$$

connected to  $f_\varepsilon$  by the relations

$$\int \begin{pmatrix} 1 \\ v \\ v^2 \end{pmatrix} M[f_\varepsilon](v) dv = \begin{pmatrix} \rho_\varepsilon \\ \rho_\varepsilon u_\varepsilon \\ \rho_\varepsilon u_\varepsilon^2 + \rho_\varepsilon \theta_\varepsilon \end{pmatrix} = \int \begin{pmatrix} 1 \\ v \\ v^2 \end{pmatrix} f_\varepsilon(v) dv.$$

Actually, first the model imposes the conservation properties

$$\partial_t \int \begin{pmatrix} 1 \\ v \\ v^2 \end{pmatrix} f_\varepsilon(v) dv + \partial_x \int v \begin{pmatrix} 1 \\ v \\ v^2 \end{pmatrix} f_\varepsilon(v) dv = 0$$

and second, as  $\varepsilon \rightarrow 0$ , we expect that  $f_\varepsilon$  relaxes towards such a Maxwellian state  $M$ . By replacing  $f_\varepsilon$  with  $M(v) = \frac{\rho}{\sqrt{2\pi\theta}} \exp\left(-\frac{|v-u|^2}{2\theta}\right)$  in these relations, we do indeed obtain the Euler system.

We again write a splitting scheme:

– Transport stage where we solve  $(\partial_t + v\partial_x)f = 0$ . By using the upwind fluxes, we obtain

$$f_j^*(v) = f_j^n(v) - \frac{\Delta t}{\Delta x} \left( v_+ (f_j^n - f_{j-1}^n) + v_- (f_{j+1}^n - f_j^n) \right).$$

– Projection stage where we solve the stiff equation  $\partial_t f = \frac{1}{\varepsilon}(M[f] - f)$ , with data  $f^*$ . During this stage, the macroscopic quantities  $\rho, u, \theta$  are not modified, since

$$\partial_t \int \begin{pmatrix} 1 \\ v \\ v^2 \end{pmatrix} f(v) dv = 0.$$

We thus have

$$\int \begin{pmatrix} 1 \\ v \\ v^2 \end{pmatrix} f(v) dv = \int \begin{pmatrix} 1 \\ v \\ v^2 \end{pmatrix} f(v) dv$$

and  $M[f] = M[f^*]$ . We thus obtain

$$f_j^{n+1}(v) = f_j^n(v) e^{-\Delta t/\varepsilon} + M[f_j^*](v) (1 - e^{-\Delta t/\varepsilon}).$$

As  $\varepsilon \rightarrow 0$ , the second stage reduces to a projection over the Maxwellian state

$$f_j^{n+1}(v) = M[f_j^*](v).$$

We thus obtain the following scheme for the Euler equations: by writing

$$M_j^n(v) = \frac{\rho_j^n}{\sqrt{2\pi\theta_j^n}} \exp\left(-\frac{|v - u_j^n|^2}{2\theta_j^n}\right)$$

we arrive at

$$\begin{pmatrix} \rho_j^{n+1} \\ \rho_j^{n+1}u_j^{n+1} \\ \rho_j^{n+1}E_j^{n+1} \end{pmatrix} = \int \begin{pmatrix} 1 \\ v \\ v^2 \end{pmatrix} M_j^{n+1}(v) dv = \begin{pmatrix} \rho_j^n \\ \rho_j^n u_j^n \\ \rho_j^n E_j^n \end{pmatrix} - \frac{\Delta t}{\Delta x} \int \begin{pmatrix} 1 \\ v \\ v^2 \end{pmatrix} (v_+(M_j^n - M_{j-1}^n)(v) + v_-(M_{j+1}^n - M_j^n)(v)) dv.$$

We can rewrite this in a form involving numerical fluxes. We set

$$\mathcal{U}_j^n = \begin{pmatrix} \rho_j^n \\ \rho_j^n u_j^n \\ \rho_j^n E_j^n \end{pmatrix}$$

and we have

$$\mathcal{U}_j^{n+1} = \mathcal{U}_j^n - \frac{\Delta t}{\Delta x} (F_{j+1/2}^n - F_{j-1/2}^n)$$

with

$$F_{j+1/2}^n = \int_{-\infty}^0 v \begin{pmatrix} 1 \\ v \\ v^2 \end{pmatrix} M_{j+1}^n dv + \int_0^{+\infty} v \begin{pmatrix} 1 \\ v \\ v^2 \end{pmatrix} M_j^n dv = \mathbb{F}(\mathcal{U}_{j+1}^n, \mathcal{U}_j^n).$$

The scheme is consistent because

$$\mathbb{F}(\mathcal{U}, \mathcal{U}) = \int v \begin{pmatrix} 1 \\ v \\ v^2 \end{pmatrix} \frac{\rho}{\sqrt{2\pi\theta}} e^{-|v-u|^2/(2\theta)} dv = \begin{pmatrix} \rho u \\ \rho u^2 + p \\ \rho u^2 + 3\rho\theta \end{pmatrix}.$$

The previous construction seems quite natural, the BGK equation and the Maxwellian equilibria having a physical basis, arising from the description of the gases in terms of statistical physics (see [SAI 09]). However, from a numerical point of view, we can find interest in working with different equilibrium states, which give the scheme better properties. In particular, working with states with compact support in the variable  $v$  offers certain advantages. We set

$$\mathcal{M}[\rho, u, \theta](v) = \frac{\rho}{2\sqrt{3\theta}} \mathbf{1}_{|v-u| \leq \sqrt{3\theta}}.$$

We again have

$$\int \begin{pmatrix} 1 \\ v \\ v^2 \end{pmatrix} \mathcal{M}[\rho, u, \theta](v) dv = \begin{pmatrix} \rho \\ \rho u \\ \rho u^2 + 3\rho\theta \end{pmatrix}$$

We thus write the numerical scheme simply by replacing the Maxwellian state by this function  $\mathcal{M}[\rho, u, \theta](v)$ ; in other words, the numerical fluxes are now given by

$$\begin{aligned} F_{j+1/2}^n &= \int_{-\infty}^0 v \begin{pmatrix} 1 \\ v \\ v^2 \end{pmatrix} \mathcal{M}_{j+1}^n dv + \int_0^{+\infty} v \begin{pmatrix} 1 \\ v \\ v^2 \end{pmatrix} \mathcal{M}_j^n dv, \mathcal{M}_j^n(v) \\ &= \mathcal{M}[\rho_j^n u_j^n, \theta_j^n](v). \end{aligned} \quad [3.69]$$

**THEOREM 3.14.**— We assume that

$$\sup_j (|u_j^n| + \sqrt{3\theta_j^n}) \leq \frac{\Delta x}{\Delta t}.$$

Thus, the scheme defined by the fluxes [3.69] satisfies

$$\rho_j^{n+1} \geq 0, \quad \theta_j^{n+1} \geq 0.$$

**PROOF.**— We again have

$$f_j^\star(v) = f_j^n \left( 1 - \frac{\Delta t}{\Delta x} |a(v)| \right) + \frac{\Delta t}{\Delta x} a_+(v) f_{j-1}^n - \frac{\Delta t}{\Delta x} a_-(v) f_{j+1}^n$$

with  $f_j^n(v) = \mathcal{M}_j^n(v)$ . In particular, the term on the right-hand side cancels when  $|v| \geq \sup_j (|u_j^n| + \sqrt{3\theta_j^n})$ . In other cases, the stability condition guarantees that  $f_j^\star(v)$  is a convex combination of the quantities  $f_{j+1}^n, f_{j-1}^n, f_j^n \geq 0$ . Thus,  $f_j^\star(v) \geq 0$  is supported in  $\{|v| \leq \sup_j (|u_j^n| + \sqrt{3\theta_j^n})\}$ . As a consequence, we obtain  $\rho_j^{n+1} = \int f_j^\star(v) dv \geq 0$ . Furthermore, for the total energy, we have

$$\begin{aligned} (\rho E)_j^{n+1} &= \int \frac{v^2}{2} f_j^\star(v) dv \\ &= \frac{1}{2} \int (|v - u_j^{n+1}|^2 + |u_j^{n+1}|^2 + 2(v - u_j^{n+1}) u_j^{n+1}) f_j^\star(v) dv \\ &\geq \frac{1}{2} \int (|u_j^{n+1}|^2 + 2(v - u_j^{n+1}) u_j^{n+1}) f_j^\star(v) dv \quad \text{car } f_j^\star(v) \geq 0 \end{aligned}$$

$$\begin{aligned} &\geq \frac{1}{2} \int (-|u_j^{n+1}|^2 + 2vu_j^{n+1}) f_j^\star(v) dv \\ &\geq \frac{1}{2} \left( -|u_j^{n+1}|^2 \int f_j^\star(v) dv + 2u_j^{n+1} \int vf_j^\star(v) dv \right) = \frac{\rho^{n+1} |u_j^{n+1}|^2}{2} \geq 0. \end{aligned}$$

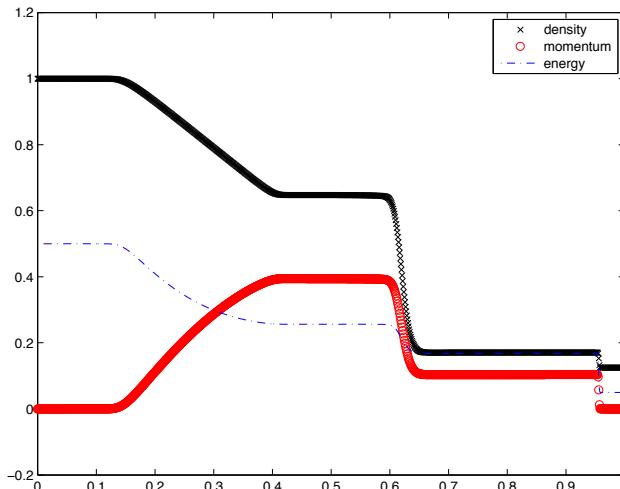
Finally, since  $E = u^2/2 + 3\theta/2$ , we deduce that  $\theta \geq 0$ .  $\square$

We carry out the same simulations as in section 3.4.2.2, with identical numerical parameters. Figure 3.66 reproduces the result for Case 1 (compare with Figure 3.60). Figure 3.67 shows the material speed and pressure; we thus observe that if the conservative quantities display a singularity in the middle of the run-off, these auxiliary quantities are themselves continuous there. (We say that there is a contact discontinuity.) In Figure 3.68, we show the entropy

$$S(t, x) = \frac{1}{\rho^3(t, x)} \left( \mathcal{E}(t, x) - \frac{1}{2} \frac{J^2(t, x)}{\rho(t, x)} \right).$$

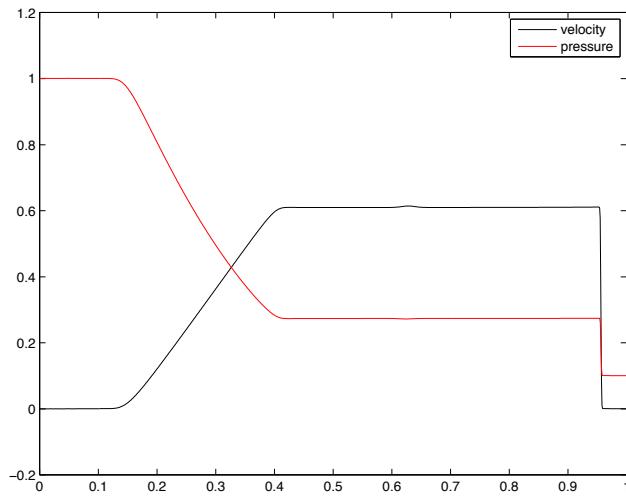
For regular solutions, this quantity is simply transported at speed  $v(t, x) = \frac{J}{\rho}(t, x)$ , but displays discontinuities at the point of shocks.

For Case 4, described by Figure 3.69, the formation of the vacuum zone, in the middle of the calculation domain, is well captured, without excessive oscillations at the points of connection.

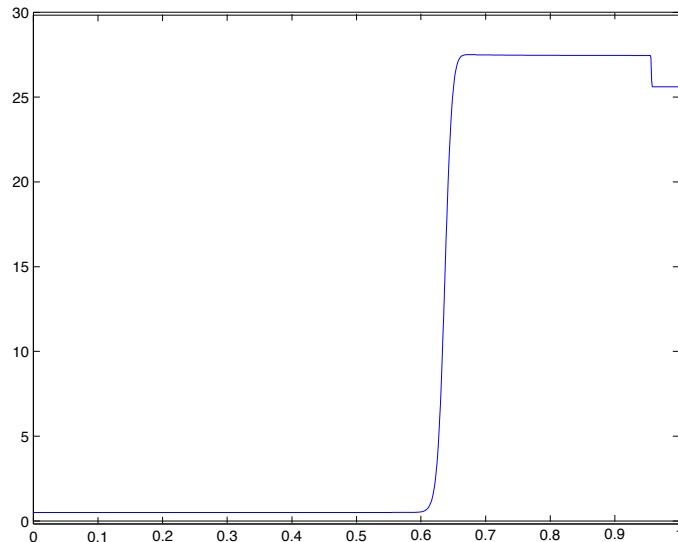


**Figure 3.66.** Simulation of the Euler equations: Case 1 ( $\Delta x = 10^{-3}$ ).

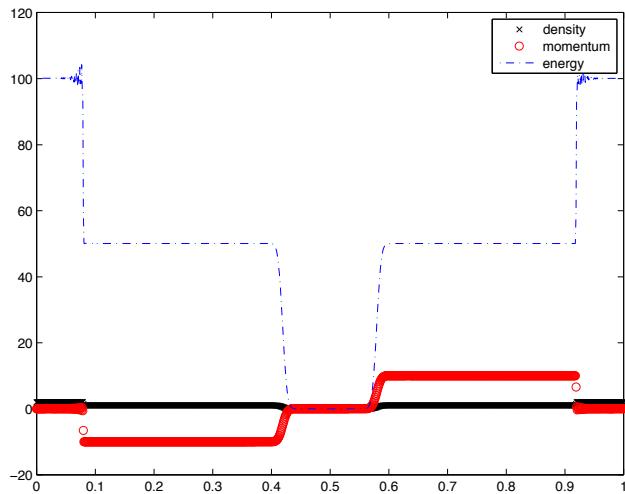
For the color version of this figure, see  
[www.iste.co.uk/goudon/mathmodel.zip](http://www.iste.co.uk/goudon/mathmodel.zip)



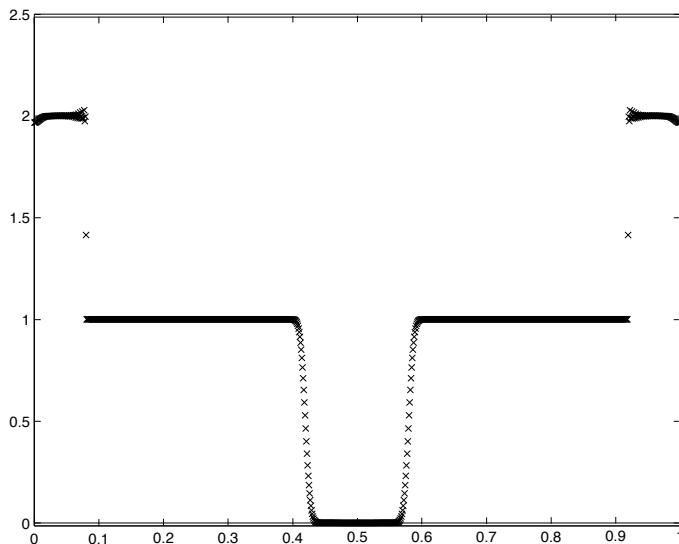
**Figure 3.67.** Simulation of the Euler equations: Case 1, graph of the speed and pressure ( $\Delta x = 10^{-3}$ ). For the color version of this figure, see [www.iste.co.uk/goudon/mathmodel.zip](http://www.iste.co.uk/goudon/mathmodel.zip)



**Figure 3.68.** Simulation of the Euler equations: Case 4, graph of entropy ( $\Delta x = 10^{-3}$ )



**Figure 3.69.** Simulation of the Euler equations: Case 4 ( $\Delta x = 10^{-3}$ ).  
For the color version of this figure, see  
[www.iste.co.uk/goudon/mathmodel.zip](http://www.iste.co.uk/goudon/mathmodel.zip)



**Figure 3.70.** Simulation of the Euler equations: Case 4, graph of the density ( $\Delta x = 10^{-3}$ )

## **APPENDICES**



# Appendix 1

## Solving Linear Systems

### A1.1. Condition number of a matrix

DEFINITION A1.1.– We consider  $\mathbb{R}^N$  equipped with a norm  $|\cdot|$ . We represent the subordinate norm with  $\|\cdot\|$ : for  $A \in \mathbb{M}_N$ , we have

$$\|A\| = \sup \{|Ax|, x \in \mathbb{S}^{N-1}\} = \sup \left\{ \frac{|Ax|}{|x|}, x \neq 0 \right\}.$$

The *condition number* of an invertible matrix  $A \in \mathbb{M}_N$  relative to the norm  $|\cdot|$  is the name we give the quantity

$$\text{cond}(A) = \|A\| \|A^{-1}\|.$$

This value is of interest because it measures how errors in the data affect the solution of a linear system. This is formalized by the following statement.

LEMMA A1.1.– Let  $A \in \mathbb{M}_N$  be an invertible matrix. We have  $\text{cond}(A) \geq 1$ . Let  $x, b \in \mathbb{R}^N$  satisfy  $Ax = b$ . Let  $\delta x, \delta b \in \mathbb{R}^N$  be such that  $A(x + \delta x) = b + \delta b$ . Thus, we have

$$\frac{|\delta x|}{|x|} \leq \text{cond}(A) \frac{|\delta b|}{|b|}$$

PROOF.– We first of all note that  $1 = \|\mathbb{I}\| = \|AA^{-1}\| \leq \|A\|\|A^{-1}\| = \text{cond}(A)$ . Then, by linearity we have  $\delta b = A\delta x$ , i.e.  $\delta x = A^{-1}\delta b$  and the definition of the subordinate norm thus leads to  $|\delta x| = |A^{-1}\delta b| \leq \|A^{-1}\|\|\delta b\|$ . Moreover, we also have  $|b| = |Ax| \leq \|A\||x|$ . We conclude by combining these two relations  $\square$

In this statement,  $\delta b$  is interpreted as the uncertainty, the errors in the data  $b$ . We evaluate the resulting relative error  $\frac{|\delta x|}{|x|}$  in the solution of the linear system defined by the matrix  $A$ . As  $\text{cond}(A) \geq 1$ , these errors can always be amplified. A “well-conditioned” matrix thus has a condition number close to 1 (but unfortunately greater than 1). The following example, proposed in [LAS 04], is especially striking

$$A = \begin{pmatrix} 10 & 7 & 8 & 7 \\ 7 & 5 & 6 & 5 \\ 8 & 6 & 10 & 9 \\ 7 & 5 & 9 & 10 \end{pmatrix}$$

With  $b = (32, 23, 33, 31)$ , we find  $x = (1, 1, 1, 1)$ . With  $\delta b = (0.01, -0.01, 0.01, -0.01)$ , we find  $\delta x = (0.82, -1.36, 0.35, -0.21)$ . In this case,  $|\delta b|/|b| \simeq 3.10^{-3}$ , but  $|\delta x|/|x| \simeq 0.8$ : the relative error in the data is multiplied by 2400. A statement can be made of the same nature with respect to perturbations in the coefficients of the matrix  $A$  itself. Indeed, if we have, on the one hand,  $Ax = b$  and, on the other hand,  $(A + \delta A)(x + \delta x) = b$ , then  $\delta A(x + \delta x) = -A\delta x$ , i.e.  $\delta x = -A^{-1}\delta A(x + \delta x)$ . It immediately follows that

$$\frac{|\delta x|}{|x + \delta x|} \leq \text{cond}(A) \frac{\|\delta A\|}{\|A\|}.$$

Poor evaluation of the data, or sometimes even the structure of the problem, thus has important practical consequences. There exist numerical techniques, called “preconditioning”, for improving the condition numbers of matrices. One way to understand the difficulty is to interpret it in terms of the spectral distribution.

**THEOREM A1.1.–** The condition number of an invertible matrix  $A$  relative to the Euclidean norm is given by the ratio of the extreme singular values of  $A$

$$\text{cond}(A) = \sqrt{\frac{\Sigma}{\sigma}}$$

where  $\sigma$  and  $\Sigma$  are the smallest and greatest eigenvalues of  $A^\top A$ . In the specific case that  $A$  is symmetric  $\text{cond}(A) = \frac{\Lambda}{\lambda}$  with  $\Lambda = \max\{|\mu|, \mu \text{ eigenvalue of } A\}$ ,  $\lambda = \min\{|\mu|, \mu \text{ eigenvalue of } A\}$ .

**PROOF.–** We simply note that, for the Euclidean norm,

$$\|A\|^2 = \sup \left\{ \frac{(Ax|Ax)}{(x|x)}, x \neq 0 \right\} = \sup \left\{ \frac{(A^\top Ax|x)}{(x|x)}, x \neq 0 \right\}.$$

As  $A^\top A$  is symmetric, it is diagonalizable on an orthogonal basis, and its eigenvalues  $\mu_1, \dots, \mu_N$  are positive since  $(A^\top Ax|x) = |Ax|^2 \geq 0$ . We use  $P$  to

represent the transformation matrix, such that  $A^\top A = P^\top \text{diag}(\mu_1, \dots, \mu_N) P$ . Hence,  $(A^\top A x | x) = (P^\top \text{diag}(\mu_1, \dots, \mu_N) P x | x) = \sum_{j=1}^N \mu_j |y_j|^2$ , with  $y = Px$ . As  $|y|^2 = (Px | Px) = (P^\top Px | x) = (x | x) = |x|^2$ , we conclude from this that  $\|A\|^2 = \max_j \{\mu_j\} = \Sigma$ . The same reasoning with  $A^{-1}$  gives  $(A^{-1})^\top A^{-1} = (AA^\top)^{-1}$ . We deduce from this that  $\|A^{-1}\|^2 = \sigma$ , and hence the stated expression for  $\text{cond}(A)$ .  $\square$

## A1.2. Spectral radius

The concept of spectral radius is a critical component of numerous results in numerical analysis. It is thus good to have a clear idea of the subject. We recall that,  $A$  being a matrix of  $\mathcal{M}_N(\mathbb{C})$ , the spectral radius is the real number defined by

$$\rho(A) = \max \{|\lambda|, \lambda \text{ eigenvalues of } A\}.$$

We always have

$$\rho(A) \leq \|A\|.$$

Indeed, if  $\lambda$  is an eigenvalue of  $A$  and  $x \neq 0$  is an associated eigenvector, we have  $|\lambda x| = |\lambda| |x| = |Ax| \leq \|A\| |x|$ , which implies this relation. We note however that  $\rho(A)$  can be null even though  $A \neq 0$ ; for example, this is the case for the matrix

$$A = \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}.$$

Let us now assume that  $S$  is a symmetric real matrix; it is thus diagonalizable on an orthogonal basis and we can write  $S = P \Lambda P^\top$ , with  $\Lambda = \text{diag}(\lambda_1, \dots, \lambda_N)$  where  $\lambda_j$  are the eigenvalues of  $S$ . It follows that

$$|(Sx|x)| = |(\Lambda P^\top x | P^\top x)| \leq \max_{\lambda \in \sigma(S)} |\lambda| |P^\top x| \leq \rho(S) |x|^2.$$

For any matrix  $A \in \mathcal{M}_N(\mathbb{R})$ ,  $S = A^\top A$  is symmetric and positive (in the sense that  $(Sx|x) = |Ax|^2 \geq 0$ ). We thus obtain

$$0 \leq |Ax|^2 = (A^\top A x | x) \leq \rho(A^\top A) |x|^2,$$

which leads to  $\|A\| \leq \sqrt{\rho(A^\top A)}$ . However, if  $\mu \geq 0$  is an eigenvalue of  $A^\top A$ , with an associated eigenvector  $x$ , we arrive at

$$(A^\top A x | x) = \mu |x|^2 = |Ax|^2 \leq \|A\|^2 |x|^2,$$

which proves that  $\rho(A^\top A) \leq \|A\|^2$ . We conclude from this that

$$\|A\|^2 = \rho(A^\top A).$$

In particular, if  $A$  is symmetric then  $A^\top A = A^2$  has  $\{\lambda^2, \lambda \in \sigma(A)\}$  for its spectrum and, from this, we deduce the following statement.

**LEMMA A 1.2.**— Let  $A$  be a symmetric matrix. Thus, we have, for the norm associated with the Euclidean scalar product,  $\|A\| = \rho(A)$ .

Numerous arguments make use of the following characterization of the spectral radius.

**LEMMA A 1.3.**— We have  $\rho(A) = \lim_{k \rightarrow \infty} \|A^k\|^{1/k}$ .

**PROOF.**— We have seen that  $\rho(A) \leq \|A\|$  from which we deduce that

$$\rho(A^k) = \rho(A)^k \leq \|A^k\|.$$

It thus follows that

$$\rho(A) \leq \liminf_{k \rightarrow \infty} \|A^k\|^{1/k}.$$

Next, for  $\epsilon > 0$ , we set

$$A_\epsilon = \frac{A}{\rho(A) + \epsilon}$$

whose spectral radius satisfies

$$\rho(A_\epsilon) = \frac{A}{\rho(A) + \epsilon} < 1.$$

Let us temporarily accept that this property means that  $\lim_{k \rightarrow \infty} A_\epsilon^k = 0$ . It follows that we can find an integer  $k_0$ , such that for all  $k \geq k_0$ , we have  $\|A_\epsilon^k\| \leq 1$  which implies that

$$\|A^k\|^{1/k} \leq \rho(A) + \epsilon.$$

We deduce from this that

$$\limsup_{k \rightarrow \infty} \|A^k\|^{1/k} \leq \rho(A) + \epsilon$$

is satisfied for all  $\epsilon > 0$ , which allows the conclusion by taking  $\epsilon$  tending towards 0.  $\square$

To complete the argument, it remains to justify the following lemma.

LEMMA A1.4.– We have  $\lim_{k \rightarrow \infty} A^k = 0$  if and only if  $\rho(A) < 1$ .

PROOF.– We deduce from  $\rho(A^k) = \rho(A)^k \leq \|A^k\|$  that, if  $\rho(A) > 1$ , then  $(\|A^k\|)_{k \in \mathbb{N}}$  cannot tend towards 0. In order to establish the reciprocal, we use Jordan decomposition, as in the proof of lemma 3.2.  $\square$

We finish this section by establishing a monotony property of the spectral radius.

LEMMA A1.5.– Let  $A$  and  $B$  be two real matrices, such that  $0 \leq A_{ij} \leq B_{ij}$  for all indices  $i, j$ . Thus, we have  $\rho(A) \leq \rho(B)$ .

An immediate recurrence shows that, for all indices  $i, j$  and all integers  $k$ , we again have  $0 \leq [A^k]_{ij} \leq [B^k]_{ij}$  and  $\|A^k\|_\infty \leq \|B^k\|_\infty$ . As the function  $t \mapsto t^{1/k}$  is monotonically increasing, we deduce from this that

$$\rho(A) = \lim_{k \rightarrow \infty} \|A^k\|_\infty^{1/k} \leq \lim_{k \rightarrow \infty} \|B^k\|_\infty = \rho(B). \quad \square$$

### A1.3. Conjugate gradient

One way or another, the solution of linear systems lies at the heart of computation algorithms used to find the approximate solution of complex problems. It is thus important to have a thorough knowledge of the basis of methods which can find the solution of

$$Ax = b \in \mathbb{R}^N.$$

We direct the reader to Chapters 4 and 7 of [LAS 04] for a very detailed presentation of these methods, both direct and iterative. Here, we assume that:

the matrix  $A$  is *symmetric positive definite*

and we present only the conjugate gradient algorithm which works in the following manner.

We arbitrarily choose a vector  $x^{(0)}$  and we set  $e^{(0)} = y^{(0)} = b - Ax^{(0)}$  (which is, in general, non-null!). Then, using  $x^{(k)}, y^{(k)} \neq 0, e^{(k)} \neq 0$ :

– we set:

$$\beta^{(k)} = \frac{e^{(k)} \cdot y^{(k)}}{Ay^{(k)} \cdot y^{(k)}}, \quad x^{(k+1)} = x^{(k)} + \beta^{(k)} y^{(k)}, \quad e^{(k+1)} = b - Ax^{(k+1)}, \quad [\text{A1.1}]$$

– if  $e^{(k+1)} = 0$ , we have found the solution of  $Ax = b$  and the algorithm stops; if not we set

$$\alpha^{(k)} = -\frac{Ay^{(k)} \cdot e^{(k+1)}}{Ay^{(k)} \cdot y^{(k)}}, \quad y^{(k+1)} = e^{(k+1)} + \alpha^{(k)}y^{(k)}. \quad [\text{A1.2}]$$

The vector  $e^{(k)}$  evaluates the error produced at stage  $k$ . In fact, we can show that the gradient algorithm converges in at most  $N$  iterations (see [LAS 04, Corollary 15]). However, in practice, we stop the algorithm when the error  $|e^{(k)}|$  goes beneath a certain threshold  $0 < \epsilon \ll 1$  set beforehand; we hope that this stop condition will be satisfied for a “small” number of iterations  $k$ . The cost of an iteration of the algorithm being of the order of  $N^2$  operations, we will thus have gained in calculation time, the methods based upon the *LU* decomposition having a cost of the order of  $N^3$  operations. We shall see the role that the condition number of the matrix  $A$  plays in the performance analysis of the conjugate gradient algorithm.

A crucial point of the analysis consists of exploiting the fact that the matrices  $A$  and  $A^{-1}$  allow for the definition of scalar products. We have

$$Ay^{(k)} = Ae^{(k)} + \alpha^{(k-1)}Ay^{(k-1)}.$$

First of all, it follows that

$$Ay^{(k)} \cdot y^{(k-1)} = Ae^{(k)} \cdot y^{(k-1)} + \alpha^{(k-1)}Ay^{(k-1)} \cdot y^{(k-1)} = 0$$

by the definition of  $\alpha^{(k-1)}$ : the vectors  $y^{(k)}$  and  $y^{(k-1)}$  are  $A$ –orthogonal. Moreover, we thus also have

$$\begin{aligned} Ay^{(k)} \cdot y^{(k)} &= Ae^{(k)} \cdot y^{(k)} + \alpha^{(k-1)} \times 0 = Ae^{(k)} \cdot e^{(k)} + \alpha^{(k-1)}Ae^{(k)} \cdot y^{(k-1)} \\ &= Ae^{(k)} \cdot e^{(k)} + \alpha^{(k-1)}Ay^{(k)} \cdot y^{(k-1)} - (\alpha^{(k-1)})^2Ay^{(k-1)} \cdot y^{(k-1)} \\ &= Ae^{(k)} \cdot e^{(k)} - (\alpha^{(k-1)})^2Ay^{(k-1)} \cdot y^{(k-1)}. \end{aligned}$$

Similarly, we observe that

$$e^{(k+1)} = e^{(k)} - \beta^{(k)}Ay^{(k)}$$

which implies that

$$e^{(k+1)} \cdot y^{(k)} = e^{(k)} \cdot y^{(k)} - \beta^{(k)}Ay^{(k)} \cdot y^{(k)} = 0$$

by the definition of  $\beta^{(k)}$ : the vectors  $e^{(k+1)}$  and  $y^{(k)}$  are also orthogonal, for the usual scalar product. But, we also have  $A^{-1}e^{(k+1)} = A^{-1}e^{(k)} - \beta^{(k)}y^{(k)}$ , which leads to

$$\begin{aligned} A^{-1}e^{(k+1)} \cdot e^{(k+1)} &= A^{-1}e^{(k)} \cdot e^{(k+1)} - \beta^{(k)} \times 0 \\ &= A^{-1}e^{(k)} \cdot e^{(k)} - \beta^{(k)}A^{-1}e^{(k)} \cdot Ay^{(k)} \\ &= A^{-1}e^{(k)} \cdot e^{(k)} - \beta^{(k)}e^{(k)} \cdot y^{(k)} \\ &= A^{-1}e^{(k)} \cdot e^{(k)} - \beta^{(k)}e^{(k+1)} \cdot y^{(k)} - (\beta^{(k)})^2Ay^{(k)} \cdot y^{(k)} \\ &= A^{-1}e^{(k)} \cdot e^{(k)} - (\beta^{(k)})^2Ay^{(k)} \cdot y^{(k)}. \end{aligned}$$

The idea then consists of evaluating the progression of the error, defined using the norm associated with the scalar product induced by  $A^{-1}$ . We thus introduce

$$\varepsilon^{(k+1)} = \frac{1}{2}A^{-1}e^{(k+1)} \cdot e^{(k+1)}$$

We use the previous relations to calculate

$$\begin{aligned} \varepsilon^{(k+1)} &= \frac{1}{2}A^{-1}e^{(k)} \cdot e^{(k)} - \frac{1}{2}(\beta^{(k)})^2Ay^{(k)} \cdot y^{(k)} \\ &= \varepsilon^{(k)} - \frac{1}{2} \frac{(e^{(k)} \cdot y^{(k)})^2}{Ay^{(k)} \cdot y^{(k)}} \\ &= \varepsilon^{(k)} - \frac{1}{2} \frac{(e^{(k)} \cdot e^{(k)})^2}{Ay^{(k)} \cdot y^{(k)}} \end{aligned}$$

since  $e^{(k)} \cdot y^{(k)} = e^{(k)} \cdot (e^{(k)} + \alpha^{(k-1)}y^{(k-1)}) = e^{(k)} \cdot e^{(k)}$ . We thus obtain

$$\varepsilon^{(k+1)} = \varepsilon^{(k)} - \frac{1}{2} \frac{(e^{(k)} \cdot e^{(k)})^2}{Ae^{(k)} \cdot e^{(k)} - (\alpha^{(k-1)})^2Ay^{(k-1)} \cdot y^{(k-1)}}$$

Now, we have

$$\begin{aligned} Ae^{(k)} \cdot e^{(k)} &\geq Ae^{(k)} \cdot e^{(k)} - (\alpha^{(k-1)})^2Ay^{(k-1)} \cdot y^{(k-1)} \\ &= Ae^{(k)} \cdot e^{(k)} - \frac{(Ay^{(k-1)} \cdot e^{(k)})^2}{Ay^{(k-1)} \cdot y^{(k-1)}} \geq 0, \end{aligned}$$

the positivity being ensured by the Cauchy–Schwarz inequality for the norm  $N_A(x) = Ax \cdot x$ :  $|Ay^{(k-1)} \cdot e^{(k)}| \leq N_A(y^{(k-1)})N_A(e^{(k)})$ . We deduce from this the estimate

$$\begin{aligned}\varepsilon^{(k+1)} &\leq \varepsilon^{(k)} - \frac{1}{2} \frac{(e^{(k)} \cdot e^{(k)})^2}{Ae^{(k)} \cdot e^{(k)}} \\ &\leq \varepsilon^{(k)} - \frac{1}{2} \frac{A^{-1}e^{(k)} \cdot e^{(k)} (e^{(k)} \cdot e^{(k)})^2}{A^{-1}e^{(k)} \cdot e^{(k)} Ae^{(k)} \cdot e^{(k)}} \\ &\leq \varepsilon^{(k)} \left(1 - \frac{(e^{(k)} \cdot e^{(k)})^2}{A^{-1}e^{(k)} \cdot e^{(k)} Ae^{(k)} \cdot e^{(k)}}\right).\end{aligned}$$

We signify with  $\underline{\lambda}, \bar{\lambda} > 0$  the extreme eigenvalues of  $A$ . We recall that

$$\underline{\lambda}|e^{(k)}|^2 \leq Ae^{(k)} \cdot e^{(k)} \leq \bar{\lambda}|e^{(k)}|^2 \quad \text{and} \quad \frac{1}{\bar{\lambda}}|e^{(k)}|^2 \leq A^{-1}e^{(k)} \cdot e^{(k)} \leq \frac{1}{\underline{\lambda}}|e^{(k)}|^2.$$

We conclude from this that

$$\varepsilon^{(k+1)} \leq \varepsilon^{(k)}(1 - \underline{\lambda}/\bar{\lambda}).$$

This relation proves the convergence of the algorithm; this one is faster the closer the ratio  $\underline{\lambda}/\bar{\lambda}$  is to 1, i.e. when the matrix  $A$  is well conditioned.

# Appendix 2

## Numerical Integration

In practice, we rarely know how to explicitly calculate

$$I[f] = \int_a^b f(t) dt$$

given “any” function  $f$  defined over  $[a, b]$ , with two fixed integers  $a, b$ . We would like to define an approximation of  $I[f]$  starting with the evaluation of the function  $f$  over a certain number of points  $a = \xi_1 < \xi_2 < \dots < \xi_n < \xi_{n+1} = b$ . We expect that the approximation will be better the greater the number of points used. As an approximate expression we can use a formula of the form

$$I_n[f] = \sum_{j=1}^{n+1} \omega_j f(\xi_j) \quad [A2.1]$$

and we thus seek the weights  $\omega_1, \dots, \omega_{n+1}$ , so that  $\lim_{n \rightarrow \infty} I_n[f] = I[f]$ .

The simplest example is given by the *rectangle formula*, which in fact forms the basis of the definition of the Riemann integral:

$$I_n^{\text{Rect}}[f] = \sum_{j=1}^n f(\xi_j)(\xi_{j+1} - \xi_j).$$

In other words, here  $\omega_j = \xi_{j+1} - \xi_j$  if  $j \in \{1, \dots, n\}$  and  $\omega_{n+1} = 0$ .

**THEOREM A2.1.**— We set  $h_n = \max \{\xi_{j+1} - \xi_j, j \in \{1, \dots, n\}\}$ . We assume that  $\lim_{n \rightarrow \infty} h_n = 0$ . Thus, for all continuous functions  $f$ , we have  $\lim_{n \rightarrow \infty} I_n^{\text{Rect}}[f] = I[f]$ . If, additionally,  $f$  is a function of class  $C^1$ , we in fact have the inequality  $|I[f] - I_n^{\text{Rect}}[f]| \leq (b - a) \|f'\|_\infty h_n$ .

PROOF.– We must evaluate

$$|I[f] - I_n^{\text{Rect}}[f]| = \left| \sum_{j=1}^n \int_{\xi_j}^{\xi_{j+1}} (f(y) - f(\xi_j)) dy \right|.$$

As  $f$  is continuous, it is uniformly continuous over the compact  $[a, b]$ : for all  $\epsilon > 0$ , there exists  $\eta(\epsilon) > 0$ , such that if  $\xi, \zeta$  are two points of  $[a, b]$  satisfying  $|\xi - \zeta| \leq \eta(\epsilon)$ , then  $|f(\xi) - f(\zeta)| \leq \epsilon$ . Now, presumably, we can find  $N(\epsilon)$ , such that if  $n \geq N(\epsilon)$ , then  $h_n \leq \eta(\epsilon)$ . As  $0 < \xi_{j+1} - \xi_j < h_n$ , we deduce from this that for all  $y \in [\xi_j, \xi_{j+1}]$ , we have  $|f(y) - f(\xi_j)| \leq \epsilon$  when  $n \geq N(\epsilon)$ . Consequentially, we obtain

$$|I[f] - I_n^{\text{Rect}}[f]| \leq \sum_{j=1}^n \epsilon (\xi_{j+1} - \xi_j) = (b - a)\epsilon$$

when  $n \geq N(\epsilon)$ .

When  $f$  is a function of class  $C^1$ , the differentiated function being continuous, it is bounded over the compact  $[a, b]$  and we arrive at

$$|f(y) - f(\xi_j)| = \left| \int_{\xi_j}^y f'(z) dz \right| \leq \|f'\|_\infty |y - \xi_j| \leq \|f'\|_\infty h_n.$$

It follows that

$$|I[f] - I_n^{\text{Rect}}[f]| \leq \|f'\|_\infty h_n \sum_{j=1}^n \int_{\xi_j}^{\xi_{j+1}} dy = (b - a)\|f'\|_\infty h_n. \quad \square$$

NOTE A2.1.– The rectangle formula lies at the basis of the design of the explicit Euler scheme. We would obtain the same result if we use the “right-hand” rectangle formula  $I_n^{\text{Rect,dr}}[f] = \sum_{j=1}^n f(\xi_{j+1})(\xi_{j+1} - \xi_j)$ , which appears in the implicit Euler scheme.

We propose an approximation of the same nature, but slightly different, by first introducing the points  $a = \xi_{1/2} < \xi_{3/2} < \dots < \xi_{n-1/2} < \xi_{n+1/2} = b$ . Then, we set for  $j \in \{1, \dots, n\}$

$$\xi_j = \frac{\xi_{j+1/2} + \xi_{j-1/2}}{2}, \quad \tilde{h}_j = \xi_{j+1/2} - \xi_{j-1/2}, \quad \omega_j = \tilde{h}_j,$$

$$I_n^{\text{PM}}[f] = \sum_{j=1}^n \omega_j f(\xi_j).$$

This is called the *mid-point* formula. These complicated notations simply mean that we approximate the integral of  $f$  over  $[\xi_{j+1/2}, \xi_{j+1/2}]$  by  $\tilde{h}_j f(\xi_j)$ . This formula gives a more precise approximation (assuming a stronger regularity over  $f$ ).

**THEOREM A2.2.** – We set  $h_n = \max \{\tilde{h}_j, j \in \{1, \dots, n\}\}$ . Thus, there exists a constant  $C$ , independent of  $n$ , such that for all functions  $f$  of class  $C^2$ , we have the inequality  $|I[f] - I_n^{\text{PM}}[f]| \leq C \|f''\|_\infty h_n^2$ .

**PROOF.** – We again write the difference to be evaluated as a sum

$$I[f] - I_n^{\text{PM}}[f] = \sum_{j=1}^n \int_{\xi_{j-1/2}}^{\xi_{j+1/2}} (f(y) - f(\xi_j)) dy.$$

We use Taylor's formula

$$f(y) - f(\xi_j) = f'(\xi_j)(y - \xi_j) + \int_0^1 (1-t)f''(\xi_j + t(y - \xi_j)) (y - \xi_j)^2 dt.$$

Now, we have

$$\int_{\xi_{j-1/2}}^{\xi_{j+1/2}} f'(\xi_j)(y - \xi_j) dy = f'(\xi_j) \int_{\xi_j - \tilde{h}_j/2}^{\xi_j + \tilde{h}_j/2} (y - \xi_j) dy = 0.$$

Additionally,  $f''$  is bounded over  $[a, b]$ . We thus obtain

$$|I[f] - I_n^{\text{PM}}[f]| \leq \sum_{j=1}^n \int_{\xi_{j-1/2}}^{\xi_{j+1/2}} \left( h_n^2 \|f''\|_\infty \int_0^1 (1-t) dt \right) dy \leq \frac{b-a}{2} h_n^2 \|f''\|_\infty. \square$$

More generally, we consider a subdivision  $a = \xi_1 < \xi_2 < \dots < \xi_n < \xi_{n+1} = b$ . The step size  $h = \max \{\xi_{j+1} - \xi_j, j \in \{1, \dots, n\}\}$  measures the refinement of this subdivision, to which we associate a certain numerical integration formula  $\mathcal{I}_h[f]$ . The examples discussed above, where  $\mathcal{I}_h$  takes the particular form [A2.1], lead to the introduction of the following concept that quantifies the quality of the approximation.

**DEFINITION A2.1.** – We say that a numerical integration formula  $\mathcal{I}_h$  is of order  $p$  if there exists a constant  $C$ , such that for all functions  $f$  of class  $C^p$ , we have  $|I[f] - \mathcal{I}_h[f]| \leq C \max_{\ell \in \{0, \dots, p\}} \|f^{(\ell)}\|_\infty h^p$ .

Thus, the method of rectangles is of order 1, and the mid-point method is of order 2. To analyze more general methods of numerical integration, we decompose the integral to be evaluated over intervals of the subdivision

$$\int_a^b f(x) dx = \sum_{j=1}^n \int_{\xi_j}^{\xi_{j+1}} f(x) dx = \sum_{j=1}^n (\xi_{j+1} - \xi_j) \int_0^1 f(\xi_j + \tau(\xi_{j+1} - \xi_j)) d\tau. \quad [\text{A2.2}]$$

Thus, the question comes back to approximating the integral over  $[0, 1]$  of  $\tau \mapsto f(\xi_j + \tau(\xi_{j+1} - \xi_j))$ . To this end, assuming that we have a method of type [A2.1] defined for the functions  $g : [0, 1] \rightarrow \mathbb{R}$ :

$$I_M(g) = \sum_{m=1}^M \omega_m g(\tau_m). \quad [\text{A2.3}]$$

Thus, we approximate  $\int_a^b f(x) dx$  by the formula

$$\mathcal{I}_h[f] = \sum_{j=1}^n (\xi_{j+1} - \xi_j) \sum_{m=1}^M \omega_m f(\xi_j + \tau_m(\xi_{j+1} - \xi_j)). \quad [\text{A2.4}]$$

The numerical method of integration thus involves not only the point of the subdivision  $\xi_1, \dots, \xi_{n+1}$ , but also the intermediate points which are of the form  $\xi_j + \tau_m(\xi_{j+1} - \xi_j)$ .

Evidently, we do not need a numerical method of integration to calculate the integral of a polynomial function. Nevertheless, we shall see that we can link the order of approximation of a numerical integration method to its behavior for polynomials.

**DEFINITION A2.2.**— We say that a numerical integration formula is  $\mathbb{P}_r$ —exact if, for all polynomials  $P$  of a degree smaller than or equal to  $r$ , we have  $\mathcal{I}_h(P) = 0$ .

We easily find that the rectangle method is  $\mathbb{P}_0$ —exact (exact for constants) and the mid-point method is  $\mathbb{P}_1$ —exact (exact for linear functions). The following statement gives a practical criterion for determining the order of a numerical integration method.

**THEOREM A2.3.**— If the integration formula [A2.3] is  $\mathbb{P}_r$ —exact, then the integration formula [A2.4] is of order  $r + 1$ .

**PROOF.**— Thanks to relation [A2.2], we start by noting that

$$\begin{aligned} I[f] - \mathcal{I}_h[f] &= \sum_{j=1}^n (\xi_{j+1} - \xi_j) \left( \int_0^1 f(\xi_j + \tau h_{j+1/2}) d\tau \right. \\ &\quad \left. - \sum_{m=1}^M \omega_m f(\xi_j + \tau_m h_{j+1/2}) \right). \end{aligned}$$

where we have written  $h_{j+1/2} = \xi_{j+1} - \xi_j$ . We make use of Taylor's formula to give, up to a remaining term, a polynomial expression in the variable  $\tau$ :

$$\begin{aligned} f(\xi_j + \tau h_{j+1/2}) &= \sum_{\ell=0}^r f^{(\ell)}(\xi_j) \frac{h_{j+1/2}^\ell}{\ell!} \tau^\ell \\ &\quad + \int_0^1 f^{(r+1)}(\xi_j + s\tau h_{j+1/2}) (\tau h_{j+1/2})^{r+1} \frac{(1-s)^r}{r!} ds. \end{aligned}$$

Thus, in  $I[f] - \mathcal{I}_h[f]$ , we identify the difference

$$\sum_{\ell=0}^r f^{(\ell)}(\xi_j) \frac{h_{j+1/2}^\ell}{\ell!} \left( \int_0^1 \tau^\ell d\tau - \sum_{m=1}^M \omega_m \tau_m^\ell \right)$$

where (...) is nothing other than  $I[g] - I_M[g]$ , with [A2.3], applied to the function  $g : \tau \mapsto \tau^\ell$ . Formula [A2.3] being  $\mathbb{P}_r$ -exact, this difference is zero. We deduce from this that

$$\begin{aligned} I[f] - \mathcal{I}_h[f] &= \sum_{j=1}^n (\xi_{j+1} - \xi_j) \left( \int_0^1 \int_0^1 \left( f^{(r+1)}(\xi_j + s\tau h_{j+1/2}) (\tau h_{j+1/2})^{r+1} \frac{(1-s)^r}{r!} ds \right) d\tau \right. \\ &\quad \left. - \sum_{m=1}^M \omega_m \int_0^1 f^{(r+1)}(\xi_j + s\tau_m h_{j+1/2}) (\tau_m h_{j+1/2})^{r+1} \frac{(1-s)^r}{r!} ds \right). \end{aligned}$$

Now, we can rewrite

$$\begin{aligned} &\int_0^1 f^{(r+1)}(\xi_j + s\tau h_{j+1/2}) (\tau h_{j+1/2})^{r+1} (1-s)^r ds \\ &= h_{j+1/2}^{r+1} \int_0^1 f^{(r+1)}(\xi_j + s\tau h_{j+1/2}) (\tau - s\tau)^r \tau ds \\ &= h_{j+1/2}^{r+1} \int_0^\tau f^{(r+1)}(\xi_j + \sigma h_{j+1/2}) (\tau - \sigma)^r d\sigma \\ &= h_{j+1/2}^{r+1} \int_0^1 f^{(r+1)}(\xi_j + \sigma h_{j+1/2}) [\tau - \sigma]_+^r d\sigma. \end{aligned}$$

We thus obtain

$$\begin{aligned} & I[f] - \mathcal{I}_h[f] \\ &= \sum_{j=1}^n (\xi_{j+1} - \xi_j) h_{j+1/2}^{r+1} \\ &\quad \times \int_0^1 \frac{1}{r!} \left( \int_0^1 [\tau - \sigma]_+^r d\tau - \sum_{m=1}^M \omega_m [\tau_m - \sigma]_+^r \right) f^{(r+1)}(\xi_j + \sigma h_{j+1/2}) d\sigma. \end{aligned}$$

The quantity

$$\begin{aligned} \sigma \longmapsto N(\sigma) &= \frac{1}{r!} \left( \int_0^1 [\tau - \sigma]_+^r d\tau - \sum_{m=1}^M \omega_m [\tau_m - \sigma]_+^r \right) \\ &= \frac{(1 - \sigma)^{r+1}}{(r+1)!} - \frac{1}{r!} \sum_{m=1}^M \omega_m [\tau_m - \sigma]_+^r \end{aligned}$$

is called the *Peano kernel* of order  $r$  associated with the method [A2.3]. We thus arrive at the following inequality

$$|I[f] - \mathcal{I}_h[f]| \leq h^{r+1} \times (b-a) \|f^{(r+1)}\|_\infty \int_0^1 |N_r(\sigma)| d\sigma. \quad \square$$

We thus return the analysis of a numerical method of integration to the study of its behavior with polynomials. For example, the trapeze method is defined over  $[0, 1]$  by  $M = 2, \tau_1 = 0, \tau_2 = 1$

$$I^{\text{Trap}}[g] = \frac{g(0) + g(1)}{2}.$$

This formula is  $\mathbb{P}_1$ -exact. With the argument of changing scale, it leads to

$$I_n^{\text{Trap}}[f] = \sum_{j=1}^n (\xi_{j+1} - \xi_j) \frac{f(\xi_j) + f(\xi_{j+1})}{2}.$$

We thus have  $|I[f] - I_n^{\text{Trap}}[f]| \leq C \|f''\|_\infty h^2$ , as for the mid-point method.

**THEOREM A2.4.**— Formula [A2.3] is  $\mathbb{P}_r$ -exact if and only if, for all  $\ell \in \{1, \dots, r\}$ , we have

$$\sum_{m=1}^M \omega_m \tau_m^{\ell-1} = \frac{1}{\ell}.$$

PROOF.– If formula [A2.3] is  $\mathbb{P}_r$ -exact, then we have

$$\int_0^1 \tau^\ell d\tau = \frac{1}{\ell+1} = \sum_{m=1}^M \omega_m \tau_m^\ell$$

for all  $\ell \in \{0, \dots, r\}$ . Conversely, if these relations are satisfied, then, for all polynomials  $P(\tau) = \sum_{\ell=0}^r a_\ell \tau^\ell$ , we have

$$\begin{aligned} \int_0^1 P(\tau) d\tau &= \sum_{\ell=0}^r \frac{a_\ell}{\ell+1} = \sum_{\ell=0}^r a_\ell \left( \sum_{m=1}^M \omega_m \tau_m^\ell \right) \\ &= \sum_{m=1}^M \omega_m \left( \sum_{\ell=0}^r a_\ell \tau_m^\ell \right) = I_M(P). \end{aligned} \quad \square$$

Using these results, we shall seek to improve the order by working with polynomials of a higher degree. A polynomial of degree 2 defined over  $[0, 1]$  can be expressed as

$$P(\tau) = a_0 + a_1 \tau + a_2 \tau^2 = P(0)\psi_0(\tau) + P(1/2)\psi_{1/2}(\tau) + P(1)\psi_1(\tau)$$

where the functions  $\psi_0, \psi_{1/2}, \psi_1$  are defined as in section 2.4.2:

$$\psi_0(\tau) = 2(1-\tau)(1/2-\tau), \quad \psi_{1/2}(\tau) = 4(1-\tau)\tau, \quad \psi_1(\tau) = 2\tau(\tau-1/2).$$

We thus have

$$\int_0^1 P(\tau) d\tau = \frac{1}{6} (P(0) + 4P(1/2) + P(1)).$$

We thus define the formula of numerical integration of type [A2.3], called *Simpson's rule*:

$$I^{\text{Sim}}[g] = \frac{1}{6} (g(0) + 4g(1/2) + g(1)).$$

For this method  $M = 3$ , the nodes  $\tau_m$  are the points  $0, 1/2, 1$  and the weights  $\omega_m$  are, respectively,  $1/6, 2/3, 1/6$ . By definition, this method is  $\mathbb{P}_2$ -exact. However, we note that we also have  $\frac{1}{6}0^3 + \frac{2}{3}\frac{1}{2^3} + \frac{1}{6}1^3 = \frac{1}{4}$ : according to theorem A2.4, Simpson's rule is  $\mathbb{P}_3$ -exact. (But we have  $\frac{1}{6}0^4 + \frac{2}{3}\frac{1}{2^4} + \frac{1}{6}1^4 = \frac{5}{24} \neq \frac{1}{5}$  thus it is not  $\mathbb{P}_4$ -exact.) It thus gives us an approximation method of order 4.

This approach can be generalized. Actually, theorem A2.4 leads to a linear system of  $r$  equations for the  $M$  unknowns  $\omega_1, \dots, \omega_M$ . To have a solution, we thus need  $M \geq r$ . The subsystem of size  $r \times r$  entails the (Vandermonde) matrix:

$$V = \begin{pmatrix} 1 & 1 & 1 & \cdots & 1 \\ \tau_1 & \tau_2 & \tau_3 & \cdots & \tau_r \\ \tau_1^2 & \tau_2^2 & \tau_3^2 & \cdots & \tau_r^2 \\ \vdots & \vdots & \vdots & & \vdots \\ \tau_1^r & \tau_2^r & \tau_3^r & \cdots & \tau_r^r \end{pmatrix}.$$

This system also appears in the framework of *interpolation* problems: given a function  $g : [0, 1] \rightarrow \mathbb{R}$  and the separate points  $\tau_1, \dots, \tau_r$  in  $[0, 1]$ , we seek coefficients  $\omega_1, \dots, \omega_r$ , such that the polynomial  $P_g(\tau) = \sum_{m=1}^r \omega_m \tau^{m-1} \in \mathbb{P}_{r-1}$  takes the values  $g(\tau_j)$  at the points  $\tau_j$ . This problem comes down to solving the linear system  $V^\top \Omega = G$  for  $\Omega = (\omega_1, \dots, \omega_r)$  and  $G = (g(\tau_1), \dots, g(\tau_r))$ . The matrix  $V^\top$  describes the linear mapping  $\Phi : P \in \mathbb{P}_{r-1} \mapsto (P(\tau_1), \dots, P(\tau_r)) \in \mathbb{R}^r$ , which maps a space of dimension  $r$  onto a space which is also of dimension  $r$ . Now,  $\Phi(P) = 0$  signifies that  $P \in \mathbb{P}_{r-1}$  has  $r$  roots  $\tau_1, \dots, \tau_r$ ; such a polynomial is null. The mapping  $\Phi$  is thus injective and in fact bijective. We conclude from this that  $V$  and  $V^\top$  are indeed invertible. For the problem of interpolation, we associate with the points  $\tau_1, \dots, \tau_r$  the basis of  $\mathbb{P}_{r-1}$  formed of the functions

$$\psi_j(\tau) = \prod_{m \neq j} \frac{\tau - \tau_m}{\tau_j - \tau_m},$$

which in particular satisfy  $\psi_j(\tau_m) = \delta_{jm}$ . Hence, the polynomial (*the Lagrange interpolating polynomial*)

$$P_g(\tau) = \sum_{m=1}^r g(\tau_j) \psi_j(\tau)$$

addresses the interpolation problem. Armed with these observations, we return to the definition of numerical integration formulae. We assume now that  $M = r$ . If formula

[A2.3] is  $\mathbb{P}_r$ -exact, then, in particular, it is exact for the basis functions  $\psi_j$  and we deduce the value of the coefficients  $\omega_j$  from this, since

$$\sum_{m=1}^r \omega_m \psi_j(\tau_m) = \omega_j = \int_0^1 \psi_j(\tau) d\tau.$$

We now consider a general function  $g : [0, 1] \rightarrow \mathbb{R}$ . The formula of approximation [A2.3] is given by

$$I_r[g] = \sum_{m=1}^r g(\tau_m) \int_0^1 \psi_m(\tau) d\tau = \int_0^1 \left( \sum_{m=1}^r g(\tau_m) \psi_m(\tau) \right) d\tau = \int_0^1 P_g(\tau) d\tau.$$

In other words, we have replaced the integral of  $g$  by the integral of the associated polynomial of interpolation. According to theorem A2.3, this formula gives an approximation of order  $r + 1$  of  $I[g]$ .

We shall bear in mind that definition A2.1 assumes that the function to be integrated is reasonably regular. The demonstration of theorem A2.3 incidentally makes full use of, via Taylor's formula, the  $C^{r+1}$  character of  $f$ . In practice, the order is degraded when we work with less regular functions.



# Appendix 3

## A Peetre–Tartar Equivalence Theorem

The purpose of this appendix is to study the following lemma and some of its applications.

LEMMA A3.1.– Let  $E$  be a Banach space and let  $F, G$  be normed spaces. Let  $A \in \mathcal{L}(E, F)$  and  $B \in \mathcal{L}(E, G)$ . We assume that:

- i)  $B$  is compact;
- ii) There exist  $\mu, \nu > 0$ , such that for all  $x \in E$ , we have

$$\nu(\|Ax\|_F + \|Bx\|_G) \leq \|x\|_E \leq \mu(\|Ax\|_F + \|Bx\|_G).$$

Thus,  $\text{Ker}(A)$  is of finite dimension and  $\text{Ran}(A)$  is closed.

PROOF.– The inequality on the left-hand side in ii) simply expresses the continuity of the operators  $A$  and  $B$ . The inequality on the right-hand side proves that  $x \mapsto \|Ax\|_F + \|Bx\|_G$  is a norm over  $E$ , equivalent to the norm  $\|\cdot\|_E$ .

We first note that  $\text{Ker}(A)$  is a subspace of  $E$ , closed because  $\text{Ker}(A)$  is the inverse image of the closed set  $\{0\}$  by the continuous mapping  $A$ . Over  $\text{Ker}(A)$ , relation ii) becomes  $\nu\|Bx\|_G \leq \|x\|_E \leq \mu\|Bx\|_G$ . We deduce from this that  $B|_{\text{Ker}(A)}$  is injective and thus 0 is not an eigenvalue of  $B|_{\text{Ker}(A)}$ . Let  $(u_n)_{n \in \mathbb{N}}$  be a normed sequence of elements of  $\text{Ker}(A)$ : for all  $n \in \mathbb{N}$ , we have  $Au_n = 0$  and  $\|u_n\|_E = 1$ . As  $B$  is compact, we can extract from it a subsequence, such that  $Bu_{n_k}$  has a limit in  $G$  as  $k \rightarrow \infty$ . However,  $\|u_{n_k} - u_{n_\ell}\|_E \leq \mu\|Bu_{n_k} - Bu_{n_\ell}\|_G$  proves that  $(u_{n_k})_{k \in \mathbb{N}}$  is a Cauchy sequence in the complete space  $E$ , thus it converges in  $E$ . We have thus shown that the unit sphere of  $\text{Ker}(A)$  is compact, which implies, according to Riesz's lemma (see [GOU 11, Theorem 5.13]), that  $\text{Ker}(A)$  is of finite dimension.

We then set  $T : u \in E \mapsto (Au, Bu) \in F \times G$ . Thus,  $T$  is continuously linear and  $\|Tu\| = \|Au\|_F + \|Bu\|_G$  satisfies  $\nu\|Tu\| \leq \|u\|_E \leq \mu\|Tu\|$ . This proves that:

–  $T$  is injective;

–  $\text{Ran}(T)$  is closed. Indeed, if  $T(u_n)$  converges towards  $t = (t_1, t_2) \in F \times G$ , then  $\|u\|_E \leq \mu\|Tu\|$  implies that  $(u_n)_{n \in \mathbb{N}}$  is Cauchy in  $E$ , and thus has a limit  $u$  there. By continuity of  $A$  and  $B$ , it follows that  $Tu_n$  converges towards  $Tu = (Au, Bu) = (t_1, t_2) \in \text{Ran}(T)$ . As a consequence, we find that  $T$  is invertible from  $E$  over  $\text{Ran}(T)$ , with continuous  $T^{-1}$  of norm  $\leq 1/\nu$ .

Let  $(u_n)_{n \in \mathbb{N}}$  be a sequence in  $E$ , such that  $\lim_{n \rightarrow \infty} Au_n = a \in F$ ; we seek to show that  $a \in \text{Ran}(A)$ , i.e. that there exists  $u \in E$ , such that  $a = Au$ . We set

$$d_n = \text{dist}(u_n, \text{Ker}(A)) = \inf\{\|u_n - w\|_E, w \in \text{Ker}(A)\}.$$

We have seen that  $\text{Ker}(A)$  is of finite dimension, thus there exists  $w_n \in \text{Ker}(A)$ , such that

$$d_n = \|u_n - w_n\|_E$$

(we can return to a continuous function over a compact set). We have  $Au_n = A(u_n - w_n)$ , which tends towards  $a$  in  $E$ . We shall show that  $v_n = (u_n - w_n)$  is bounded. Let us assume that  $\lim_{n \rightarrow \infty} \|v_n\|_E = \infty$ . Thus,  $V_n = \frac{v_n}{\|v_n\|_E}$  is of norm 1 and such that  $\lim_{n \rightarrow \infty} AV_n = 0$ . Furthermore,  $B$  being compact,  $BV_n$  has a subsequence, which we continue to label by the index  $n$ , which converges towards an element  $b$  of  $G$ . We thus have  $\lim_{n \rightarrow \infty} TV_n = (AV_n, BV_n) = (0, b)$ . However, we have seen that  $T$  is of closed image, thus there exists  $V \in E$ , such that  $b = BV$  and  $0 = AV$ . We thus obtain

$$V_n = T^{-1}TV_n \xrightarrow[n \rightarrow \infty]{} T^{-1}(0, b) = T^{-1}(0, BV) = V \in \text{Ker}(A).$$

Now, we remark that

$$\begin{aligned} \text{dist}(V_n, \text{Ker}(A)) &= \inf \left\{ \left\| \frac{u_n - w_n}{\|v_n\|_E} - w \right\|_E, w \in \text{Ker}(A) \right\} \\ &= \frac{1}{\|v_n\|_E} \inf \{ \|u_n - z\|_E, z \in \text{Ker}(A) \} = \frac{d_n}{\|v_n\|_E} \\ &= \frac{1}{\|v_n\|_E} \|u_n - w_n\|_E = 1. \end{aligned}$$

By taking  $n$  tending towards  $+\infty$ , we arrive at  $\text{dist}(V, \text{Ker}(A)) = 1$ , which contradicts the fact that  $V \in \text{Ker}(A)$ . We deduce from this that  $v_n = (u_n - w_n)$  is

bounded. As  $B$  is compact, we can assume (even if it means extracting a subsequence) that  $Bv_n$  converges towards  $b$  in  $G$ . It follows that  $Tv_n = (Av_n, Bv_n) = (Au_n, Bv_n)$  converges towards  $(a, b) \in \overline{\text{Ran}(T)} = \text{Ran}(T)$ , since  $\text{Ran}(T)$  is closed. Thus, there exists  $u \in E$ , such that  $a = Au$  and  $b = Bu$ . We conclude from this that  $\text{Ran}(A)$  is closed.  $\square$

By making use of this statement with  $E = L^2([0, L[)$ ,  $F = G = H^{-1}([0, L[)$ ,  $A = \frac{d}{dx}$  and  $B = \mathbb{I}$ , the canonical embedding, we demonstrate lemma 2.13, useful for the analysis of the Stokes problem. Property ii) results from the demonstration of lemma 2.12. The compactness of the injection operator  $B$  is a consequence of the Rellich–Kondrakov theorem [GOU 11, Corollary 7.58] and the Schauder theorem, which ensures that if  $T \in \mathcal{L}(X, Y)$  is compact, then the adjoint  $T^* \in \mathcal{L}(Y', X')$  is also compact [BRÉ 05, Theorem VI.4].

As is demonstrated in [TAR 87], this statement allows us to establish reasonably directly a number of useful results for the analysis of partial differential equations or for the analysis of numerical methods.

LEMMA A3.2.– In addition to the assumptions of lemma A3.1, we assume that there exists  $L \in \mathcal{L}(E, H)$ , where  $H$  is a normed space, such that

iii)  $\text{Ker}(A) \subset \text{Ker}(L)$ .

Thus, there exists  $C > 0$ , such that

$$\|Lu\|_H \leq C \|L\| \|Au\|_F$$

for all  $u \in E$ .

PROOF.– As  $\text{Ker}(A)$  is of finite dimension, it has a topological supplement<sup>1</sup>, see [BRÉ 05, Section II.4], which we signify with  $Z$ . This results from the preceding analysis that  $A$  is bijective from  $Z$  in  $\text{Ran}(A)$ . We use  $D$  to represent the reciprocal bijection. By returning to the arguments used to prove lemma A3.1, we can show that there exists  $C > 0$ , such that for all  $u \in Z$ , we have<sup>2</sup>  $\|u\|_E \leq C \|Au\|_F$ , i.e.  $\|Dy\|_E \leq C \|y\|_F$  for all  $y \in \text{Ran}(A)$ . Assumption iii) allows us to write  $Lu = LDAu$  for all  $u \in E$ , since  $(\mathbb{I} - DA)u \in \text{Ker}(A)$ . We deduce from this  $\|Lu\|_H \leq C \|L\| \|Au\|_F$ .  $\square$

Here are a few examples of applications of this Lemma:

<sup>1</sup> This signifies that  $Z$  is closed and that all elements  $x \in E$  decompose uniquely into the form  $x = y + z$ , with  $y \in \text{Ker}(A)$  and  $z \in Z$ .

<sup>2</sup> We use *reductio ad absurdum* by taking a sequence of elements of  $Z$ , such that  $\|u_n\|_E = 1$  and  $\lim_{n \rightarrow \infty} Au_n = 0$ . Thus, we can assume that  $Bu_n$  tends towards  $b \in G$ , i.e. that  $Tu_n \rightarrow (0, b)$ . However, according to ii),  $(u_n)_{n \in \mathbb{N}}$  is Cauchy in the complete space  $E$  and thus converges towards  $u \in Z$ , as  $Z$  is closed. It follows that  $Au = 0$ , while  $u \in Z$ , a topological supplement of  $\text{Ker}(A)$ , and  $\|u\|_E = 1$ , a contradiction.

– Numerical integration:

We use an approximate integration formula [A2.3] over  $[0, 1]$ , written as  $I_M$ , which we assume to be  $\mathbb{P}_r$ -exact. We represent with  $L$  the operator

$$L : u \in C^{r+1}([0, 1]) \mapsto \int_0^1 u(y) dy - I_M[u] \in \mathbb{R}.$$

We set  $A : u \mapsto \frac{d^{r+1}u}{dx^{r+1}} \in \mathcal{L}(C^{r+1}([0, 1]), C^0([0, 1]))$ . Assumption iii) is satisfied:  $\text{Ker}(A) = \mathbb{P}_r \subset \text{Ker}(L)$ . Finally, we use  $B$  to represent the canonical embedding of  $C^{r+1}([0, 1])$  in  $C^0([0, 1])$ . The Arzela–Ascoli theorem (see [GOU 11, Theorem 7.49]) ensures that  $B$  is a compact operator: i) is satisfied. Evidently, we have

$$\|u\|_{C^{r+1}} = \max \left\{ \left| \frac{d^k u}{dx^k}(x) \right|, x \in [0, 1], k \in \{0, \dots, r+1\} \right\} \geq \|u\|_{C^0} + \left\| \frac{d^{r+1} u}{dx^{r+1}} \right\|_{C^0}.$$

To obtain ii), the existence of a constant  $C > 0$  remains yet to be demonstrated, such that for all  $u \in C^{r+1}([0, 1])$ , we have

$$\|u\|_{C^{r+1}} \leq C \left( \|u\|_{C^0} + \left\| \frac{d^{r+1} u}{dx^{r+1}} \right\|_{C^0} \right).$$

We use *reductio ad absurdum* by assuming that we can present a sequence of functions of class  $C^{r+1}$ , such that

$$\|u_n\|_{C^{r+1}} = 1 \quad \text{and} \quad \lim_{n \rightarrow \infty} \left( \|u_n\|_{C^0} + \left\| \frac{d^{r+1} u_n}{dx^{r+1}} \right\|_{C^0} \right) = 0. \quad [\text{A3.1}]$$

By again evoking the Arzela–Ascoli theorem, we can assume (potentially at the cost of extracting subsequences) that all the sequences  $(\frac{d^k u_n}{dx^k})_{n \in \mathbb{N}}$ , for  $k \in \{0, \dots, r\}$ , converge uniformly over  $[0, 1]$  and we write their respective limits as  $\ell_k \in C^0([0, 1])$ . However, we have

$$\frac{d^r u_n}{dx^r}(x) = \frac{d^r u_n}{dx^r}(0) + \int_0^x \frac{d^{r+1} u_n}{dx^{r+1}}(y) dy.$$

As  $n \rightarrow \infty$ , we deduce from this that

$$\ell_r(x) = \ell_r(0) + 0$$

is constant. We apply the same reasoning by reducing the order of differentiation to show that the functions  $\ell_k$  are polynomials of degree  $r - k$  and in particular  $\ell_0 \in \mathbb{P}_r$

with  $\ell_k = \frac{d^k \ell_0}{dx^k}$ . Furthermore, assumptions [A3.1] also imply that

$$\max \left\{ \|\ell_k\|_{C^0}, k \in \{0, \dots, r+1\} \right\} = 1 \quad \text{and} \quad \|\ell_0\|_{C^0} = 0,$$

which leads to a contradiction. We conclude from this that ii) is also satisfied.

Lemma A3.2 thus implies that there exists  $C > 0$ , such that for all  $g \in C^{r+1}([0, 1])$ , we have

$$|I[g] - I_M[g]| \leq C \left\| \frac{d^{r+1} g}{dx^{r+1}} \right\|_{C^0}.$$

Returning to the notation of section A2, we apply this inequality to  $g(x) = f(\xi_j + x h_{j+1/2})$  noting that

$$\frac{d^{r+1} g}{dx^{r+1}}(x) = h_{j+1/2}^{r+1} \left( \frac{d^{r+1} f}{dx^{r+1}} \right)(\xi_j + x h_{j+1/2}).$$

We deduce from this the evaluation of the error

$$|I[f] - \mathcal{I}_h[f]| \leq \sum_{j=1}^n h_{j+1/2}^{r+1} (\xi_{j+1} - \xi_j) C \left\| \frac{d^{r+1} f}{dx^{r+1}} \right\|_{C^0} \leq h^{r+1} C(b-a) \left\| \frac{d^{r+1} f}{dx^{r+1}} \right\|_{C^0}.$$

#### *– Poincaré inequalities:*

Over the domain  $[0, L]$ , we consider  $E = H^1(]0, L[)$ ,  $F = G = L^2(]0, L[)$ ,  $A = \frac{d}{dx}$ ,  $B = \mathbb{I}$ , the canonical embedding (which is compact according to the Rellich–Kondrachov theorem) and the operator

$$L : u \in H^1(]0, L[) \longmapsto u - \frac{1}{L} \int_0^L u(y) dy \in L^2(]0, L[).$$

The analysis conducted for the Stokes problem allows us to establish property ii). We have  $\text{Ker}(A) = \mathbb{P}_0 = \text{Ker}(L)$ . It follows that there exists  $C > 0$ , such that for all  $u \in H^1(]0, 1[)$ , we have

$$\left\| u - \frac{1}{L} \int_0^L u(y) dy \right\|_{L^2} \leq C \left\| \frac{du}{dx} \right\|_{L^2}.$$

This inequality can be seen as a version of the Poincaré inequality of lemma 2.2. We also use such a version for the analysis of the diffusion equation in the periodic framework of theorem 2.3 and we could use it to treat the case of the problem with Neumann conditions.



# Appendix 4

## Schauder's Theorem

This appendix is dedicated to the demonstration of theorem 2.4. We first note that in Schauder's fixed-point theorem, the fact that the invariant set  $C$  is bounded is crucial, as the following counter-example shows: the mapping  $f : x \in \mathbb{R} \mapsto x^2 + 1 \in \mathbb{R}$  is continuous and compact; however, it does not have a fixed point. In practice, the compactness is often an "easy" property to demonstrate, as a consequence of *a priori* estimates and functional properties (typically  $f(x)$  is an element of a set that embeds compactly in the space  $E$ , as we saw with the Sobolev space  $H^1$  and its embedding in  $L^2$ ). However, the continuity can cause difficulties; in particular, we cannot be satisfied with using the compactness property, which allows this conclusion only at the convergence of extracted sequences.

We start by recalling the following classic statement which is restricted to finite dimensions.

**THEOREM A4.1** (Brouwer's theorem).— Let  $F$  be a normed vector space of *finite* dimension. Let  $C \subset F$  be a convex, closed, bounded and non-empty set. Let  $f : C \rightarrow C$  be a continuous mapping. Then,  $f$  has a fixed point in  $C$ .

In infinite dimensions, the assumptions of this statement are insufficient to guarantee the existence of a fixed point, as shown in the following counter-example. In  $\ell^2(\mathbb{N})$  equipped with  $\|\cdot\|_2$ , we use  $B(0, 1)$  to signify the unit ball, and we consider the mapping

$$\begin{aligned} f : B(0, 1) \subset \ell^2(\mathbb{N}) &\longrightarrow \ell^2(\mathbb{N}) \\ x = (x_n)_{n \in \mathbb{N}} &\longmapsto f(x) = (\sqrt{1 - \|x\|_2^2}, x_0, x_1, x_2, \dots). \end{aligned}$$

For  $(x, y) \in B(0, 1)^2$ , we have

$$\begin{aligned}\|f(x) - f(y)\|_2^2 &= (\sqrt{1 - \|x\|_2^2} - \sqrt{1 - \|y\|_2^2})^2 + \sum_{k=0}^{+\infty} |x_k - y_k|^2 \\ &= (\sqrt{1 - \|x\|_2^2} - \sqrt{1 - \|y\|_2^2})^2 + \|x - y\|_2^2.\end{aligned}$$

As the mapping that associates  $\sqrt{1 - \|x\|_2^2}$  with  $x \in B(0, 1)$  is continuous,  $f$  is continuous over  $\ell^2$ . Furthermore,  $\|f(x)\|_2^2 = 1 - \|x\|_2^2 + \|x\|_2^2 = 1$ :  $f$  has values in the unit sphere. However,  $f$  does not have a fixed point. Indeed, if  $f$  had a fixed point  $x = f(x)$ , then  $\|x\|_2 = 1$ , and we would have  $x = (x_0, x_1, x_2, \dots) = (0, x_0, x_1, \dots)$ ; in other words,  $x_0 = 0$  and  $\forall n \in \mathbb{N} \setminus \{0\}, x_n = x_{n-1}$ . Thus, this implies  $x_n = 0$  for all  $n$ , in contradiction to  $\|x\|_2 = 1$ .

The supplementary hypothesis that allows treatment of the cases of maps over infinite-dimensional spaces consists of assuming that the mapping considered is *compact*. The link with Brouwer's theorem results from the capability to approximate such a mapping by a sequence of maps of finite rank. It is well known that the limit of continuous, finite rank, *linear* operators is a compact operator [BRÉ 05, Cor. VI.2]. The reciprocal is a delicate question whose answer is generally negative: a compact operator over a Banach space is not necessarily the limit of finite-rank operators (a first counter-example is attributed to [ENF 73] and the case, discovered more recently [SZA 81], of the space  $\mathcal{L}(H)$  of continuous operators over an infinite-dimensional Hilbert space  $H$  is particularly striking). The reciprocal only becomes true with the assumptions of supplementary structure over the Banach space, for example, if this space is a Hilbert space. Here, we shall present a general process of approximating compact maps – linear or not – by maps of finite rank, which are themselves nonlinear.

Let  $E$  be a normed vector space,  $B$  a closed, bounded and non-empty subset of  $E$  and  $f : B \rightarrow E$  a compact (potentially non-linear) mapping. The construction proceeds in two stages:

i) Let  $n \in \mathbb{N} \setminus \{0\}$ . We can thus recover  $\overline{f(B)}$  by a finite number  $N_n \in \mathbb{N} \setminus \{0\}$  of balls of radius  $\frac{1}{n}$ ;  $\overline{f(B)} \subset \bigcup_{i=1}^{N_n} \overset{\circ}{B}(y_i, \frac{1}{n})$  with  $y_i \in \overline{f(B)}$  for all  $i$ . For  $y \in E$ , we set

$$\psi_i(y) = \begin{cases} \frac{1}{n} - \|y - y_i\| & \text{if } y \in B(y_i, \frac{1}{n}), \\ 0 & \text{otherwise.} \end{cases}$$

We show that

$$\Psi : y \in \overline{f(B)} \longmapsto \sum_{i=1}^{N_n} \psi_i(y)$$

is continuous and that there exists  $\delta > 0$ , such that for all  $y \in \overline{f(B)}$ , we have  $\Psi(y) \geq \delta$ .

Indeed, each of these maps  $\psi_i$  is continuous, since  $\psi_i(y) = [1/n - \|y - y_i\|]_+$ , with  $s \in \mathbb{R} \mapsto [s]_+ = \max(s, 0) = \frac{1}{2}(s + |s|)$ , is written as a composition of continuous maps. Thus,  $\Psi$  is continuous. Let  $y \in \overline{f(B)}$ . Then, there exists  $i_0 \in \{1, \dots, N_n\}$ , such that  $y \in \overset{\circ}{B}(y_{i_0}, 1/n)$ , and thus we have  $\psi_{i_0}(y) > 0$  as well as  $\Psi(y) > 0$ . The function  $\Psi$  is continuous over the compact set  $f(B)$ , and thus it reaches its minimum  $\delta$  there, which is strictly positive following what we have just seen. Thus, we have shown that there exists  $\delta > 0$  ( $\delta = \Psi(\bar{y}) > 0$  for a given  $\bar{y} \in \overline{f(B)}$ ), such that for all  $y \in f(B)$ , we have  $\Psi(y) \geq \delta > 0$ .

ii) We introduce the mapping  $f_n : B \rightarrow E$  defined by

$$f_n(x) = \frac{1}{\Psi(f(x))} \sum_{i=1}^{N_n} \psi_i(f(x)) y_i.$$

If  $x \in B$ , then  $f(x) \in f(B)$ , which allows the definition  $f_n(x)$ , since  $\Psi(f(x)) > 0$ . Thus, we have

$$\begin{aligned} \|f(x) - f_n(x)\| &= \left\| \frac{1}{\Psi(f(x))} \sum_{i=1}^{N_n} \psi_i(f(x))(y_i - f(x)) \right\| \\ &\leq \frac{1}{\Psi(f(x))} \sum_{i=1}^{N_n} \psi_i(f(x)) \|y_i - f(x)\| \end{aligned}$$

since  $\psi_i(f(x))$  are  $\geq 0$ . Now, for given  $i \in \{1, \dots, n\}$ , either  $\|y_i - f(x)\| > \frac{1}{n}$  and in this case  $\psi_i(f(x))\|y_i - f(x)\| = 0 \leq \frac{1}{n}\psi_i(f(x))$ , or  $\|y_i - f(x)\| \leq \frac{1}{n}$  and the inequality  $\psi_i(f(x))\|y_i - f(x)\| \leq \frac{1}{n}\psi_i(f(x))$  is clearly satisfied. It follows from this that

$$\|f(x) - f_n(x)\| \leq \frac{1}{\Psi(f(x))} \sum_{i=1}^{N_n} \psi_i(f(x)) \times \frac{1}{n} = \frac{1}{n}.$$

We are now in a position to demonstrate theorem 2.4, by making use of theorem A4.1. We set  $F = \text{Span}(y_1, \dots, y_{N_n})$  and  $C = B \cap F$ . This set  $C$  is

– bounded, because  $B$  is so;

– closed, being the intersection of two closed sets:  $F$ , which is a subspace of finite dimension, and  $B$ ;

– convex, since  $F$  and  $B$  are so;

– non-empty, because  $C$  contains at least  $y_i \in \overline{f(B)}$ , a subset of the closed set  $B$ .

By construction,  $f_n$  maps  $C$  in  $F$  and is a continuous mapping over  $C$  (since  $\Psi$  and  $\psi_i$  are). Furthermore, for all  $x \in C$ ,  $f_n(x)$  arises as a convex combination of  $\{y_1, \dots, y_{N_n}\}$ , which are elements of  $C$ . Thus, by the convexity of  $C$ , we have  $f_n(x) \in C$ . Finally,  $C$  being contained in  $F$ , a finite-dimensional space, we can apply Brouwer's theorem to  $f_n : C \rightarrow C$ : there exists  $x_n \in C \subset B$ , such that  $f_n(x_n) = x_n$ .

Now, the sequence  $(x_n)_{n \in \mathbb{N}}$  is bounded, because its values are in  $B$ ; we can thus extract from it a subsequence, such that  $(f(x_{n_k}))_{k \in \mathbb{N}}$  converges towards  $x \in \overline{f(B)} \subset B$ . As for all  $k \in \mathbb{N}$ , we have  $\|f(x_{n_k}) - x_{n_k}\| = \|f(x_{n_k}) - f_{n_k}(x_{n_k})\| \leq \frac{1}{n_k}$ , we obtain

$$\lim_{k \rightarrow +\infty} x_{n_k} = x.$$

Finally,  $f$  being continuous, we also have

$$\lim_{k \rightarrow +\infty} f(x_{n_k}) = f(x) = x.$$

□

It remains to establish theorem A4.1. This is a classic result for which it is good to know a demonstration. We shall start by demonstrating Brouwer's theorem in the case where the set  $C$  is the closed unit ball  $B(0, 1)$  of  $\mathbb{R}^N$ . The proof is based upon *reductio ad absurdum*. The key point is to first note that

- (\*) if  $f : B(0, 1) \rightarrow B(0, 1)$  is continuous and has no fixed point in  $B(0, 1)$   
then we can construct a continuous mapping  $\varphi$  for which the image  
of  $B(0, 1)$  is the sphere  $\mathbb{S}^{N-1}$  and such that  $\varphi|_{\mathbb{S}^{N-1}}$  is the identity  
transformation.

We shall next show that there are actually no such continuous mappings from the ball over the sphere.

We thus assume that  $f : B(0, 1) \rightarrow B(0, 1)$  is continuous and has no fixed point. Thus,  $x \mapsto |x - f(x)|$  is continuous and strictly positive over the compact set  $B(0, 1)$ . Thus, there exists  $\epsilon > 0$ , such that  $|x - f(x)| \geq \epsilon > 0$  for all  $x \in B(0, 1)$ . We seek  $\varphi$  of the form

$$\varphi(x) = x - \tau(x)(x - f(x)), \quad \tau : B(0, 1) \rightarrow \mathbb{R}.$$

Saying that  $\varphi(x) \in \mathbb{S}^{N-1}$  is the same as requiring

$$|\varphi(x)|^2 = 1 = |x|^2 - 2\tau(x)x \cdot (x - f(x)) + \tau(x)^2|x - f(x)|^2.$$

This relation is simply a polynomial equation of second degree for  $\tau(x)$ . We find

$$\tau(x) = \frac{x \cdot (x - f(x)) - \sqrt{|x \cdot (x - f(x))|^2 + (1 - |x|^2)|x - f(x)|^2}}{|x - f(x)|^2}.$$

In this expression, we note

$$\begin{aligned} I(x) &= |x \cdot (x - f(x))|^2 + (1 - |x|^2)|x - f(x)|^2 \\ &= |x - f(x)|^2 \left( \left| x \cdot \frac{x - f(x)}{|x - f(x)|} \right|^2 + 1 - |x|^2 \right). \end{aligned}$$

As  $x \in B(0, 1)$ , this is the sum of two positive terms, so  $I(x) \geq 0$ . In fact,  $I(x)$  cancels if and only if  $|x| = 1$  and  $x \cdot (x - f(x)) = |x|^2 - x \cdot f(x) = 0$ , i.e.  $x \in \mathbb{S}^{N-1}$  and  $x \cdot f(x) = 1$ . According to the Cauchy–Schwarz inequality, this last relation is possible only for  $f(x)$  collinear with  $x$ , so in fact for  $f(x) = x$ , which is excluded by assumption. We conclude from this that  $I(x) > 0$ . Still, as a result of the Cauchy–Schwarz inequality, we observe that  $x \cdot (x - f(x)) = |x|^2 - x \cdot f(x) \geq 0$ . Thus, the root  $\tau(x)$  has been chosen so that  $\varphi(x) = x$  for all  $x \in \mathbb{S}^{N-1}$ . Finally,  $\varphi$  arises as a composition of continuous functions, so it is also a continuous function. This demonstrates the assertion (\*). We further remark, since  $|x - f(x)| \geq \epsilon$  does not cancel and  $z \mapsto |z|$  is of class  $C^1$  over  $\mathbb{R}^N \setminus \overset{\circ}{B}(0, \epsilon)$ , that  $\varphi$  is a function of class  $C^1$  while  $f$  is a function of  $C^1$ .

We will actually be satisfied to work with  $f$  which is a function of class  $C^1$ . To this end, for continuous  $f$ , we can evoke the Stone–Weierstrass theorem to find a polynomial  $P_\epsilon$ , such that

$$\sup_{x \in B(0, 1)} |f(x) - P_\epsilon(x)| \leq \epsilon/2.$$

We set

$$f_\epsilon(x) = \frac{P_\epsilon(x)}{1 + \epsilon/2},$$

which is indeed a function of class  $C^1$  (and even  $C^\infty$ ). Furthermore, we have  $|f_\epsilon(x)| \leq \frac{1}{1+\epsilon/2} |f(x)| + \frac{|P_\epsilon(x) - f(x)|}{1+\epsilon/2} \leq \frac{1}{1+\epsilon/2} + \frac{\epsilon/2}{1+\epsilon/2} = 1$ :  $f_\epsilon(B(0, 1)) \subset B(0, 1)$ . Finally,  $f_\epsilon$  has no fixed point in  $B(0, 1)$ , since

$$\begin{aligned} |x - f_\epsilon(x)| &= \left| x - f(x) + f_\epsilon(x) - \frac{f(x)}{1 + \epsilon/2} + \frac{\epsilon/2}{1 + \epsilon/2} f(x) \right| \\ &\geq |x - f(x)| - \frac{1}{1 + \epsilon/2} |P_\epsilon(x) - f(x)| - \frac{\epsilon/2}{1 + \epsilon/2} |f(x)| \\ &\geq \epsilon - \frac{\epsilon/2}{1 + \epsilon/2} - \frac{\epsilon/2}{1 + \epsilon/2} = \frac{\epsilon^2}{2(1 + \epsilon/2)} > 0. \end{aligned}$$

Without loss of generality, we can thus assume that the functions involved in  $(*)$  are all of class  $C^1$ . We can thus conclude that  $(*)$  leads to a contradiction thanks to the following lemma, for which we will propose a demonstration inspired by [ROG 80].

**LEMMA A4.1.**— There is no mapping  $\varphi : B(0, 1) \rightarrow \mathbb{S}^{N-1}$  of class  $C^1$  which leaves the points of  $\mathbb{S}^{N-1}$  invariant.

**PROOF.**— We use *reductio ad absurdum* by assuming that such a function  $\varphi$  exists. For  $\theta \in [0, 1]$ , we associate with  $\varphi$  the following function:

$$\varphi_\theta : x \in B(0, 1) \longmapsto \varphi_\theta(x) = (1 - \theta)x + \theta\varphi(x).$$

For all  $\theta \in [0, 1]$ , this function  $\varphi_\theta$  is a mapping of class  $C^1$ ,  $\varphi_\theta(B(0, 1)) \subset B(0, 1)$  is satisfied and, for all  $x \in \mathbb{S}^{N-1}$ , we still have  $\varphi_\theta(x) = x$ . As  $\varphi$  is a function of  $C^1$  over the compact set  $B(0, 1)$ , we can define

$$C = \sup_{z \in B(0, 1)} \|\nabla \varphi(z)\| < \infty.$$

In particular, we have

$$|\varphi(x) - \varphi(y)| \leq C|x - y|.$$

We deduce from this that  $\varphi_\theta$  is injective, provided that  $\theta \in [0, \theta_0[$ , with sufficiently small  $\theta_0$ . Indeed, let us assume that  $\varphi_\theta(x) = \varphi_\theta(y)$ ; we thus have

$$(1 - \theta)(x - y) = \theta(\varphi(y) - \varphi(x)).$$

It follows that

$$|x - y| \leq \frac{C\theta}{1 - \theta}|x - y|.$$

Thus, for  $0 \leq \theta < \theta_0$  with  $\theta_0 = \frac{1}{1+C}$ , this implies that  $x = y$ .

We also observe that

$$\nabla \varphi_\theta(x) = (1 - \theta)\mathbb{I} + \theta\nabla \varphi(x) = (1 - \theta) \left( \mathbb{I} + \frac{\theta}{1 - \theta}\nabla \varphi(x) \right). \quad [\text{A4.1}]$$

Now, the inequality  $\|\frac{\theta}{1-\theta}\nabla \varphi(x)\| \leq \frac{C\theta}{1-\theta} < 1$  for all  $0 \leq \theta < \theta_0$  ensures that the linear mapping  $\nabla \varphi_\theta(x) : h \in \mathbb{R}^N \mapsto \nabla \varphi_\theta(x)h \in \mathbb{R}^N$  is invertible<sup>1</sup>. Theorem 1.8

---

<sup>1</sup> Inverse of  $\frac{1}{1-\theta} \sum_{k=0}^{\infty} (-1)^k \left( \frac{\theta}{1-\theta} \nabla \varphi(x) \right)^k$

allows us to introduce the reciprocal mapping  $\varphi_\theta^{-1}$ , defined and of class  $C^1$  over the image of the open ball  $\mathcal{U}_\theta = \varphi_\theta(\dot{B}(0, 1))$ , which is thus an open set contained in  $B(0, 1)$ .

We shall now establish that  $\varphi_\theta$  is surjective over  $B(0, 1)$ . By definition, we already know that  $\varphi_\theta(\mathbb{S}^{N-1}) = \mathbb{S}^{N-1}$ . It remains to be shown that  $\mathcal{U}_\theta = \dot{B}(0, 1)$ . Let us assume that there exists  $y_0 \in \dot{B}(0, 1) \setminus \mathcal{U}_\theta$ . We set  $y_1 \in \mathcal{U}_\theta \subset B(0, 1)$  and

$$y(t) = ty_1 + (1 - t)y_0.$$

As  $\dot{B}(0, 1)$  is convex, for all  $t \in [0, 1]$ ,  $y(t) \in \dot{B}(0, 1)$ , and we have

$$y(0) = y_0 \in \dot{B}(0, 1) \setminus \mathcal{U}_\theta, \quad y(1) = y_1 \in \mathcal{U}_\theta.$$

We set

$$t_* = \sup \{t \in [0, 1], y(t) \notin \mathcal{U}_\theta\}.$$

As  $\mathcal{U}_\theta$  is an open set, for  $t$  sufficiently close to 1, we know that  $y(t) \in \mathcal{U}_\theta$  and thus  $0 \leq t_* < 1$ . We consider the sequence  $t_n = t_* + 1/n$ . For all  $n \in \mathbb{N} \setminus \{0\}$ ,  $y_n = y(t_n) \in \mathcal{U}_\theta$ ; thus, there exists  $x_n \in \dot{B}(0, 1)$ , such that  $\varphi_\theta(x_n) = y_n$ . The sequence  $(x_n)_{n \in \mathbb{N} \setminus \{0\}}$  being bounded in  $\mathbb{R}^N$ , according to the Bolzano–Weierstrass theorem, we can extract from it a subsequence which converges; we note  $\lim_{k \rightarrow \infty} x_{n_k} = x_* \in B(0, 1)$ . As  $\varphi_\theta$  is continuous, we obtain

$$\lim_{k \rightarrow \infty} \varphi_\theta(x_{n_k}) = \varphi_\theta(x_*) = \lim_{k \rightarrow \infty} y_{n_k}.$$

We thus have

$$y_* = \lim_{k \rightarrow \infty} y_{n_k} = t_* y_1 + (1 - t_*) y_0 = \varphi_\theta(x_*), \quad x_* \in B(0, 1), \quad y_* \in \dot{B}(0, 1).$$

We cannot have  $|x_*| = 1$  otherwise, the sphere being invariant by  $\varphi_\theta$ , we would have  $y_* \in \mathbb{S}^{N-1}$  which contradicts the fact that  $y_* \in \dot{B}(0, 1)$ . Thus,  $|x_*| < 1$ , which implies  $y_* \in \mathcal{U}_\theta$ . Now,  $\mathcal{U}_\theta$  is open, and for all  $t$  close enough to  $t_*$ , the points  $y(t)$  are again in  $\mathcal{U}_\theta$ , which contradicts the definition of  $t_*$ . We conclude from this discussion that  $\varphi_\theta$  is a  $C^1$  diffeomorphism of  $\dot{B}(0, 1)$  over  $\dot{B}(0, 1)$  for all  $0 \leq \theta < \theta_0$ .

We now introduce the quantity

$$\mathcal{V}(\theta) = \int_{B(0,1)} \det(\nabla \varphi_\theta(x)) \, dx.$$

For  $\theta = 0$ ,  $\varphi_0$  is simply the identity transformation. It follows that  $\det(\nabla\varphi_\theta(x)) > 0$  for  $\theta$  close to 0. We carry out the change in variable  $y = \varphi_\theta(x)$ . Provided  $0 \leq \theta < \theta_0$ , the new variable  $y$  still describes the ball  $\mathring{B}(0, 1)$  when  $x$  describes  $\mathring{B}(0, 1)$  and we have  $dy = \det(\nabla\varphi_\theta(x)) dx$ . We thus obtain

$$\mathcal{V}(\theta) = \int_{\mathring{B}(0,1)} dy = |B(0, 1)|, \quad \text{for all } 0 \leq \theta < \theta_0.$$

Returning to [A4.1], we note that, for fixed  $x$ ,  $\theta \mapsto \det(\nabla\varphi_\theta(x))$  is a polynomial of degree of at most  $N$ . Thus,  $\theta \mapsto \mathcal{V}(\theta)$  is also a polynomial function. We have just seen that this function is constant over the interval  $[0, \theta_0]$ ; it is thus constant over the whole of  $[0, 1]$ :

$$\mathcal{V}(\theta) = \int_{\mathring{B}(0,1)} dy = |B(0, 1)|, \quad \text{for all } 0 \leq \theta \leq 1. \quad [\text{A4.2}]$$

The contradiction comes from the fact that  $\varphi_1 = \varphi$  has values in the unit sphere. Indeed, by deriving the relation

$$|\varphi_1(x)|^2 = 1$$

we obtain

$$\nabla\varphi_1(x)\varphi_1(x) = 0,$$

where  $\varphi_1(x) \neq 0$ , since it is an element of the sphere. In other words, 0 is an eigenvalue of the matrix  $\nabla\varphi_1(x)$  with  $\varphi_1(x)$  for the associated eigenvector. In particular, this implies  $\det(\nabla\varphi_1(x)) = 0$  and thus  $V(1) = 0$ , which contradicts [A4.2].  $\square$

We can directly extend theorem A4.1 in the case that  $C$  is a convex, homeomorphic set to the unit ball  $B(0, 1)$  of  $\mathbb{R}^N$ . By representing with  $h$  the homeomorphism  $C \rightarrow B(0, 1)$ , the mapping  $g = h \circ f \circ h^{-1}$  is continuous from  $B(0, 1)$  in  $B(0, 1)$ . It thus has a fixed point  $y \in B(0, 1)$  and  $x = h^{-1}(y) \in C$  is a fixed point of  $f$ . To deal with the general case, we must make use of the following result.

**LEMMA A4.2.–** Let  $C$  be a convex, compact, non-empty set of  $\mathbb{R}^N$ . Then, either  $C$  is reduced to a point or there exists a whole  $M \leq N$ , such that  $C$  is homeomorphic to the unit ball of  $\mathbb{R}^M$ .

**PROOF.–** We remark that in general a convex, closed, compact set of  $\mathbb{R}^N$  is not homeomorphic to the unit ball of  $\mathbb{R}^N$ , as the counter-example of a simple segment in  $\mathbb{R}^2$  shows. Of course, if  $C$  is reduced to a point and  $f$  is a continuous mapping (!) of

$C$  in  $C$ , then  $f$  has a fixed point in  $C$ . This case is of no interest. We next distinguish two cases where either  $\overset{\circ}{C} \neq \emptyset$  or not.

We first consider the situation  $\overset{\circ}{C} \neq \emptyset$ . Without loss of generality, we can assume that  $0 \in C$ . There exists  $r, R > 0$ , such that

$$B(0, r) \subset C \subset B(0, R)$$

since  $C$  is of non-empty interior and  $C$  is compact and thus bounded. We introduce the mapping (called the *gauge* of the convex set  $C$ )

$$j : x \mapsto \inf \left\{ t > 0, \frac{x}{t} \in C \right\}.$$

The limits

$$\frac{|x|}{R} \leq j(x) \leq \frac{|x|}{r}$$

are satisfied for all  $x \in \mathbb{R}^N$ . The set  $C$  is characterized by

$$C = \{x \in \mathbb{R}^N, j(x) \leq 1\}.$$

Actually, if  $x \in C$  and  $t \geq 1$ , then  $x/t + (1 - 1/t) \times 0 \in C$  because  $C$  is convex,  $0, x \in C$  and  $1/t \in ]0, 1]$ . Reciprocally, if  $j(x) \leq 1$ , we can find a sequence of real numbers  $(\alpha_n)_{n \in \mathbb{N}}$  such that  $\lim_{n \rightarrow \infty} \alpha_n = j(x)$  and  $x/\alpha_n \in C$ . We thus write  $x = \alpha_n \times \frac{x}{\alpha_n} + (1 - \alpha_n) \times 0$ . If there exists an index  $n_0$  for which  $\alpha_{n_0} \leq 1$ , then this proves that  $x \in C$ . If  $\alpha_n$  are always  $> 1$ , then  $j(x) = 1$  and  $\lim_{n \rightarrow \infty} \frac{x}{\alpha_n} = x \in C$ , since  $C$  is assumed to be closed. Let us now consider two points  $x, y \in \mathbb{R}^N$ . We assume that  $x/\alpha \in C$  and  $y/\beta \in C$ . Thus, for all  $0 \leq \lambda \leq 1$ , we have  $\frac{\lambda}{\alpha}x + \frac{1-\lambda}{\beta}y \in C$ . In particular, this is satisfied for  $\lambda = \frac{\beta}{\alpha+\beta}$  for which  $\frac{\lambda}{\alpha} = \frac{1-\lambda}{\beta}$  and this relation becomes  $\frac{1}{\alpha+\beta}(x+y) \in C$ . We thus have  $\alpha + \beta \geq j(x+y)$  for all  $\alpha \geq j(x)$  and  $\beta \geq j(y)$ . We deduce from this that  $j$  is sub-additive

$$j(x+y) \leq j(x) + j(y).$$

We use this with  $x+y = X, Y = y$   $j(X) - j(Y) \leq j(X-Y) \leq \frac{|X-Y|}{r}$ . With  $x+y = Y, x = X$  we similarly have  $j(Y) - j(X) \leq j(Y-X) \leq \frac{|X-Y|}{r}$ . By combining these two relations, we arrive at  $|j(X) - j(Y)| \leq \frac{|X-Y|}{r}$ . Thus,  $j$  is continuous over  $\mathbb{R}^N$  (in fact, it is even Lipschitzian and we note that the Lipschitz

constant depends on the fact that  $C$  is of non-empty interior via the radius  $r$ ). Equipped with this characterization of  $C$ , we set

$$\begin{aligned}\Phi : x &\longmapsto \frac{j(x)}{|x|} x \quad \text{if } x \neq 0, \text{ and } 0 \text{ if not,} \\ \Psi : y &\longmapsto \frac{|y|}{j(y)} y \quad \text{if } y \neq 0, \text{ and } 0 \text{ if not,}\end{aligned}$$

(We have seen that the fact that  $C$  is bounded implies that  $j(y) > 0$  for  $y \neq 0$ .) We easily verify that  $\Psi \circ \Phi(x) = x$  and  $\Phi \circ \Psi(y) = y$ . Furthermore,  $|\Phi(x)| = j(x)$  thus  $\Phi$  maps  $C$  in the ball  $B(0, 1)$  of  $\mathbb{R}^N$  while  $j(\Psi(y)) = |y|$  means that  $\Psi$  maps  $B(0, 1)$  in  $C$ . Finally we have, for non-null  $x, x'$ ,

$$\begin{aligned}|\Phi(x) - \Phi(x')| &= \left| \frac{j(x)}{|x|}(x - x') + x' \left( \frac{j(x)}{|x|} - \frac{j(x')}{|x'|} \right) \right| \\ &\leq \frac{j(x)}{|x|} |x - x'| + |j(x) - j(x')| + \frac{|x'| j(x)}{|x| |x'|} ||x'| - |x|| \\ &\leq \frac{1}{r} |x - x'| + \frac{1}{r} |x - x'| + |x - x'|\end{aligned}$$

and for  $x' = 0$ ,  $|\Phi(x) - \Phi(0)| = j(x) \leq \frac{1}{r} |x|$ . This proves that  $\Phi$  is continuous. A similar argument allows us to establish the continuity of  $\Psi$ . We have thus indeed constructed a homeomorphism of  $C$  in the ball  $B(0, 1)$  of  $\mathbb{R}^N$ , in the case where  $\mathring{C} \neq \emptyset$ .

We finally imagine the situation, where  $\mathring{C} = \emptyset$ . In this case, we cannot find  $N$  linearly independent vectors in  $C$ , otherwise  $C$  would contain the simplex produced by the null vector and these  $N$  vectors<sup>2</sup>, a set which is itself of non-empty interior in  $\mathbb{R}^N$ . Furthermore,  $C$  is not reduced to the point  $\{0\}$ , so we can find a set made up of  $1 \leq M < N$  non-null, linearly independent vectors of  $C$ . The convex set  $C$  thus contains the simplex produced by 0 and these  $M$  vectors. This set is a convex set of non-empty interior in  $\mathbb{R}^M$ . We thus return to the previous reasoning in this lower dimension.  $\square$

---

<sup>2</sup> If we write these  $N$  vectors as  $e_1, \dots, e_N$  and the null vector as  $e_0$ , the simplex in question is the set of points  $x = \sum_{j=0}^N \lambda_j e_j$ , with  $0 \leq \lambda_j \leq 1$ ,  $\sum_{j=0}^N \lambda_j = 1$ .

# Appendix 5

## Fundamental Solutions of the Laplacian in Dimension 1 and 2

We saw in section 2.7 that the fundamental solution of the operator  $-\Delta$  in dimension 3 is  $\frac{1}{4\pi} \frac{1}{|x|}$ . More generally, we seek to determine the functions  $x \mapsto E_N(x)$ , such that  $-\Delta E_N = \delta(x = 0)$  over  $\mathbb{R}^N$ . Evidently, since the Dirac measure appears in this definition, we must work with a concept of weak derivatives.

### A5.1. Dimension 1

The function  $x \in \mathbb{R} \mapsto |x|$  is the prototype of a function that is non-differentiable at 0, in the usual sense. Nevertheless, we can define its derivative in the weak sense by transposition: for  $\varphi \in C_c^\infty(\mathbb{R})$ , we calculate

$$\int_{\mathbb{R}} |x| \varphi'(x) dx = \int_0^\infty x \varphi'(x) dx - \int_{-\infty}^0 x \varphi'(x) dx$$

Over each domain  $[0, \infty[$  and  $]-\infty, 0]$ , we can integrate by parts and thus obtain

$$\begin{aligned} \int_{\mathbb{R}} |x| \varphi'(x) dx &= - \int_0^\infty \varphi(x) dx + \int_{-\infty}^0 \varphi(x) dx \\ &= - \int_{\mathbb{R}} \operatorname{sgn}(x) \varphi(x) dx. \end{aligned}$$

We conclude from this that  $\frac{d}{dx}|x| = \operatorname{sgn}(x)$ , the derivative being understood here “in the weak sense”:

$$\left\langle \frac{d}{dx}|x|, \varphi \right\rangle = - \int_{\mathbb{R}} |x| \varphi'(x) dx.$$

(In accordance with the intuition, the derivative of  $x \mapsto |x|$  is positive for positive  $x$  and negative for negative  $x$ , in agreement with the monotony of this function over each of these domains.) We continue this approach next by writing

$$\int_{\mathbb{R}} |x| \psi''(x) dx = - \int_{\mathbb{R}} \operatorname{sgn}(x) \psi'(x) dx$$

which is only a direct application of the previous formula with  $\varphi = \psi'$ , where we have taken  $\psi \in C_c^\infty(\mathbb{R})$ . We thus arrive at

$$\int_{\mathbb{R}} |x| \psi''(x) dx = - \int_0^\infty \psi'(x) dx + \int_{-\infty}^0 \psi'(x) dx = 2\psi(0).$$

We conclude from this that, in dimension 1, the fundamental solution of  $-\frac{d^2}{dx^2}$  is

$$E_1(x) = -\frac{1}{2}|x|.$$

## A5.2. Dimension 2

We remark that  $x \in \mathbb{R}^2 \mapsto \ln(|x|)$  is locally integrable over  $\mathbb{R}^2$ . Indeed, for all  $0 < R < \infty$ , we have, passing into radial coordinates,

$$\int_{|x| \leq R} |\ln(|x|)| dx = 2\pi \int_0^R r |\ln(r)| dr < \infty,$$

since  $\lim_{r \rightarrow 0} r \ln(r) = 0$ . We are thus free to write, for  $\varphi \in C_c^\infty(\mathbb{R}^2)$ ,

$$\int \ln(|x|) \Delta \varphi(x) dx = \lim_{\epsilon \rightarrow 0} \int_{|x| > \epsilon} \ln(|x|) \Delta \varphi(x) dx$$

then integrate by parts to obtain

$$\begin{aligned} \int \ln(|x|) \Delta \varphi(x) dx &= \lim_{\epsilon \rightarrow 0} \left( - \int_{|x| > \epsilon} \frac{1}{|x|} \frac{x}{|x|} \cdot \nabla \varphi(x) dx \right. \\ &\quad \left. + \int_{|x| = \epsilon} \ln(|x|) \nabla \varphi(x) \cdot \left( -\frac{x}{|x|} \right) d\sigma(x) \right) \end{aligned}$$

since the exterior normal to the domain  $\{|y| > \epsilon\}$ , at a point  $x$  is  $-\frac{x}{|x|}$ . Now, the last integral, which we will signify with  $r_\epsilon$ , can be rewritten with a change of variable

$x = \epsilon(\cos(\theta), \sin(\theta))$ ,  $\theta \in [0, 2\pi[$ ,  $dx = \epsilon d\theta$ , see [GOU 11, Section 4.3], in the following form

$$\begin{aligned} r_\epsilon &= \int_{|x|=\epsilon} \ln(|x|) \nabla \varphi(x) \cdot \left( -\frac{x}{|x|} \right) d\sigma(x) \\ &= -\epsilon \ln(\epsilon) \int_0^{2\pi} \nabla \varphi(\epsilon \cos(\theta), \epsilon \sin(\theta)) \cdot \begin{pmatrix} \cos(\theta) \\ \sin(\theta) \end{pmatrix} d\theta. \end{aligned}$$

We deduce from this that

$$|r_\epsilon| \leq \epsilon |\ln(\epsilon)| 2\pi \|\nabla \varphi\|_\infty$$

and thus

$$\lim_{\epsilon \rightarrow 0} r_\epsilon = 0.$$

Furthermore, we note that  $x \mapsto \frac{x}{|x|^2}$  is differentiable over  $\{|x| > \epsilon\}$  and, over this domain, we have

$$\partial_{x_j} \left( \frac{x_i}{|x|^2} \right) = \frac{\delta_{ij}}{|x|^2} - 2 \frac{x_i x_j}{|x|^4}.$$

By integrating by parts, we thus obtain

$$\begin{aligned} \int_{|x|>\epsilon} \frac{x}{|x|^2} \cdot \nabla \varphi(x) dx &= \sum_{i=1}^2 \int_{|x|>\epsilon} \frac{x_i}{|x|^2} \partial_{x_i} \varphi(x) dx \\ &= \sum_{i=1}^2 \left( - \int_{|x|>\epsilon} \partial_{x_i} \left( \frac{x_i}{|x|^2} \right) \varphi(x) dx + \int_{|x|=\epsilon} \frac{x_i}{|x|^2} \left( -\frac{x_i}{|x|} \right) \varphi(x) d\sigma(x) \right) \\ &= - \int_{|x|>\epsilon} \underbrace{\left( \frac{2}{|x|^2} - 2 \frac{|x|^2}{|x|^4} \right)}_{=0} \varphi(x) dx - \int_{|x|=\epsilon} \frac{|x|^2}{|x|^3} \varphi(x) d\sigma(x) \\ &= -\frac{1}{\epsilon} \int_0^{2\pi} \varphi(\epsilon \cos(\theta), \epsilon \sin(\theta)) \epsilon d\theta. \end{aligned}$$

We thus arrive at the relation

$$\int \ln(|x|) \Delta \varphi(x) dx = \lim_{\epsilon \rightarrow 0} \left( \int_0^{2\pi} \varphi(\epsilon \cos(\theta), \epsilon \sin(\theta)) d\theta + r_\epsilon \right) = 2\pi \varphi(0).$$

We conclude from this that

$$E_2(x) = -\frac{1}{2\pi} \ln(|x|).$$

### A5.3. Higher dimensions

The same techniques allow us to obtain the following general formula, in dimension  $N > 2$ :

$$E_N(x) = \frac{\Gamma(N/2)}{2(N-2)\pi^{N/2}} \frac{1}{|x|^{N-2}}.$$

In particular, we note that the fundamental solution of  $-\Delta$  is a positively valued function in dimension 3 or higher, is negative in dimension 1 and has no sign in dimension 2.

---

## Bibliography

---

- [ABG 09] ABGRALL R., SHU C.W., “Development of residual distribution schemes for the discontinuous Galerkin methods: the scalar case”, *Communications in Computational Physics*, vol. 5, pp. 376–390, 2009.
- [ABG 10] ABGRALL R., “A residual distribution method using discontinuous elements for the computation of possibly non smooth flows”, *The Advances in Applied Mathematics and Mechanics*, vol. 2, pp. 32–44, 2010.
- [AIN 04] AINSWORTH M., “Discrete dispersion relation for  $hp$ -version finite element approximation at high wave number”, *SIAM Journal on Numerical Analysis*, vol. 42, pp. 553–575, 2004.
- [ARN 87] ARNAUDIÈS, J.-M., FRAYSSE, H., “Cours de mathématiques. Algèbre”, Dunod, 1987.
- [ARN 88] ARNOLD V., *Équations différentielles ordinaires*, 4th ed., MIR, Moscow, 1988.
- [AW 00] AW A., RASCLE M., “Resurrection of second order models of traffic flow”, *SIAM Journal on Numerical Analysis*, vol. 60, pp. 916–938, 2000.
- [BAT 15] BATEMAN H., “Some recent researches on the motion of fluids”, *Monthly Weather Review*, vol. 43, pp. 163–170, 1915.
- [BEN 07] BENZONI-GAVAGE S., SERRE D., *Multidimensional Hyperbolic Partial Differential Equations*, Oxford University Press, 2007.
- [BEN 10] BENZONI-GAVAGE S., *Calcul différentiel et équations différentielles*, Masson, 2010.
- [BOY 06] BOYER F., FABRIE P., *Eléments d’analyse pour l’étude de quelques modèles d’écoulements de fluides visqueux incompressibles*, Springer-Verlag, Berlin, 2006.
- [BOY 12] BOYER F., “Analysis of the upwind finite volume method for general initial and boundary value transport problems”, *IMA Journal of Numerical Analysis*, vol. 32, pp. 1404–1439, 2012.
- [BRÉ 05] BRÉZIS H., *Analyse fonctionnelle. Théorie et applications*, Masson, 2005.

- [BRU 97] BRUNEAU C.H., FABRIE P., RASETARINERA P., “An accurate finite difference scheme for solving convection-dominated equations”, *International Journal for Numerical Methods in Fluids*, vol. 24, pp. 69–183, 1997.
- [BUR 40] BURGERS J., “Application of a model system to illustrate some points of the statistical theory of free turbulence”, *Proceedings of the Koninklijke Nederlandse Akademie van Wetenschappen*, vol. 43, pp. 2–12, 1940.
- [CAL 09] CALVEZ V., LENUZZA N., OELTZ D. *et al.*, “Size distribution dependence of prion aggregates infectivity”, *Mathematical Biosciences*, vol. 217, pp. 88–99, 2009.
- [CHA 48] CHALLIS J., “On the velocity of sound”, *Philosophical Magazine Series 3*, vol. 32, pp. 494–499, 1848.
- [CHA 43] CHANDRASEKHAR S., “Stochastic problems in physics and astronomy”, *Reviews of Modern Physics*, vol. 15, pp. 1–89, 1943.
- [CHA 12] CHATZIPANTELIDIS P., LAZAROV R., THOMÉE V., “Some error estimates for the lumped mass finite element method for a parabolic problem”, *Mathematics of Computation*, vol. 81, pp. 1–20, 2012.
- [CHE 95] CHEMIN J.Y., *Fluides Parfaits Incompressibles*, SMF, 1995.
- [CHE 00] CHEN G.Q., RASCLE M., “Initial layers and uniqueness of weak entropy solutions to hyperbolic conservation laws”, *Archive for Rational Mechanics and Analysis*, vol. 153, pp. 205–220, 2000.
- [COU 28] COURANT R., FRIEDRICH K., LEWY H., “On the partial difference equations of mathematical physics”, *Mathematische Annalen*, vol. 100, pp. 32–74, 1928.
- [COU 99] COUDIÈRE Y., VILA J.P., VILLEDIEU P., “Convergence rate of a finite volume scheme for a two dimensional diffusion convection problem”, *Mathematical Modelling and Numerical Analysis (M2AN)*, vol. 33, pp. 493–516, 1999.
- [COR 91] CORON F., PERTHAME B., “Numerical passage from kinetic to fluid equations”, *SIAM Journal on Numerical Analysis*, vol. 28, pp. 26–42, 1991.
- [CUS 98] CUSHING J.M., “An introduction to structured population dynamics”, *CBMS-NSFT Regional Conference Series in Applied Mathematics*, SIAM, 1998.
- [DAF 10] DAFERMOS C.M., *Hyperbolic Conservation Laws in Continuum Physics*, 3rd ed., Springer-Verlag, Berlin, 2010.
- [DAU 84] DAUTRAY R., LIONS J.L., *Analyse mathématique et calcul numérique pour les sciences et les techniques*, Masson, 1984.
- [DE 04] DE LELLIS C., OTTO F., WESTDICKENBERG M., “Minimal entropy conditions for, 2004 Burgers equation”, *Quarterly of Applied Mathematics*, vol. 62, pp. 687–700, 2004.
- [DEL 11] DELARUE F., LAGOUTIÈRE F., “Probabilistic analysis of the upwind scheme for transport equations”, *Archive for Rational Mechanics and Analysis*, vol. 199, pp. 229–268, 2011.
- [DES 86a] DESHPANDE S.M., “Kinetic theory based new upwind methods for inviscid compressible flows”, *AIAA 24th Aerospace Science Meeting*, Nevada, USA, AIAA paper 86-0275, 6–9 January 1986.

- [DES 86b] DESHPANDE S.M., On the Maxwellian distribution, symmetric form and entropy conservation for the Euler equations, Technical report, NASA Langley Research Centre, Hampton, VA, 1986.
- [DES 01] DESPRÉS B., LAGOUTIÈRE F., “Contact discontinuity capturing schemes for linear advection and compressible gas dynamics”, *Journal of Scientific Computing*, vol. 16, pp. 479–524, 2001.
- [DES 04a] DESPRÉS B., “An explicit a priori estimate for a finite volume approximation of linear advection on non-Cartesian grids”, *SIAM Journal on Numerical Analysis*, vol. 42, pp. 484–504, 2004.
- [DES 04b] DESPRÉS B., “Lax theorem and finite volume schemes”, *Mathematics of Computation*, vol. 73, pp. 1203–1234, 2004.
- [DOM 05] DOMELEVO K., OMNES P., “A finite volume method for the Laplace equation on almost arbitrary two-dimensional grids”, *Mathematical Modelling and Numerical Analysis (M2AN)*, vol. 39, pp. 1203–1249, 2005.
- [ENF 73] ENFLO P., “A counterexample to the approximation property in Banach spaces”, *Acta Mathematica*, vol. 130, pp. 309–317, 1973.
- [EVA 92] EVANS L.C., GARIEPY R.F., *Measure Theory and Fine Properties of Functions*, CRC Press, 1992.
- [EVA 98] EVANS L.C., “Partial differential equations”, *Graduate Studies in Mathematics*, Volume 19, American Mathematical Society, 1998.
- [EYM 00] EYMARD R., GALLOUËT T., HERBIN R., “Finite volume methods”, in *Handbook of Numerical Analysis*, North-Holland, Amsterdam, 2000.
- [FAI 92] FAILLE I., “A control volume method to solve an elliptic equation on a two-dimensional irregular mesh”, *Computer Methods in Applied Mechanics and Engineering*, vol. 100, pp. 275–290, 1992.
- [FRO 12] FROBENIUS G., *Ueber matrizen aus nicht negativen elementen*, Sitzungsber. Königl. Preuss. Akad. Wiss., 1912.
- [GIG 85] GIGA Y., KOHN R.V., “Asymptotically self-similar blow-up of semilinear heat equations”, *Communications on Pure and Applied Mathematics*, vol. 38, pp. 297–319, 1985.
- [GOD 91] GODLEWSKI E., RAVIART P.A., *Hyperbolic Systems of Conservation Laws*, Ellipses, 1991.
- [GOU 11] GOUDON T., *Intégration. Intégrale de Lebesgue et introduction à l'analyse fonctionnelle*, Ellipses, 2011.
- [GRE 06] GREER M., PUJO-MENJOUET L., WEBB G., “A mathematical analysis of the dynamics of prion proliferation”, *Journal of Theoretical Biology*, vol. 242, pp. 598–606, 2006.
- [HAR 65] HARLOW F., WELCH J.E., “Numerical calculation of time-dependent viscous incompressible flow of fluid with a free surface”, *Physics of Fluids*, vol. 8, pp. 2182–2189, 1965.

- [HAR 84] HARTEN H., “On a class of high resolution total-variation-stable finite difference schemes”, *SIAM Journal on Numerical Analysis*, vol. 21, pp. 1–23, 1984.
- [HES 08] HESTHAVEN J.S., WARBURTON T., *Nodal Discontinuous Galerkin Methods: Algorithms, Analysis, and Applications*, Springer, 2008.
- [HOP 50] HOPF E., “The partial differential equation  $u_t + uu_x = \mu u_{xx}$ ”, *Communications on Pure and Applied Mathematics*, vol. 3, pp. 201–230, 1950.
- [HOR 90] HORN R.A., JOHNSON C.R., *Matrix Analysis*, Cambridge University Press, Cambridge, 1990.
- [HOR 03] HORSTMANN D., “From 1970 until present: the Keller–Segel model in chemotaxis and its consequences. I”, *Jahresbericht der Deutschen Mathematiker-Vereinigung*, vol. 105, 103–165, 2003.
- [HOR 04] HORSTMANN D., “From 1970 until present: the Keller–Segel model in chemotaxis and its consequences. II”, *Jahresbericht der Deutschen Mathematiker-Vereinigung*, vol. 106, pp. 51–69, 2004.
- [HUG 87] HUGONIOT H., “Sur la propagation du mouvement dans les corps et spécialement dans les gaz parfaits, I”, *Journal de l’École Polytechnique*, vol. 57, pp. 3–97, 1887.
- [HUG 89] HUGONIOT H., “Sur la propagation du mouvement dans les corps et spécialement dans les gaz parfaits, II”, *Journal de l’École Polytechnique*, vol. 58, pp. 1–125, 1889.
- [JÄG 92] JÄGER W., LUCKHAUS S., “On explosions of solutions to a system of partial differential equations modelling chemotaxis”, *Transactions of the American Mathematical Society*, vol. 329, pp. 819–824, 1992.
- [JOH 02] JOHNSTON H., LIU J.-G., “Finite difference schemes for incompressible flow based on local pressure boundary conditions”, *Journal of Computational Physics*, vol. 180, pp. 120–154, 2002.
- [KEL 70] KELLER E., SEGEL L., “Initiation of slime mold aggregation viewed as an instability”, *Journal of Theoretical Biology*, vol. 26, pp. 399–415, 1970.
- [KRU 70] KRUZHkov S.N., “First order quasilinear equations in several independent variables”, *Matematicheskii Sbornik*, vol. 123, pp. 228–255, 1970.
- [KUZ 76] KUZNECOV N.N., “The accuracy of certain approximate methods for the computation of weak solutions of a first order quasilinear equation”, *Zhurnal Vychislitel'noi Matematiki i Matematicheskoi Fiziki*, vol. 16, pp. 1489–1502, 1627, 1976.
- [LAD 57] LADYŽENSKAYA O.A., “On the construction of discontinuous solutions of quasi-linear hyperbolic equations as limits of solutions of the corresponding parabolic equations when the ‘coefficient of viscosity’ tends toward zero”, *Proceedings of Moscow Mathematical Society*, vol. 6, pp. 465–480, 1957.
- [LAS 04] LASCAUX P., THÉODOR R., *Analyse numérique matricielle appliquée à l’art de l’ingénieur*, Masson, 2004.
- [LAX 54] LAX P.D., “Weak solutions of nonlinear hyperbolic equations and their numerical computation”, *Communications on Pure and Applied Mathematics*, vol. 7, pp. 159–193, 1954.

- [LAX 57] LAX P.D., “Hyperbolic systems of conservation laws, II”, *Communications on Pure and Applied Mathematics*, vol. 10, pp. 537–566, 1957.
- [LAX 60] LAX P.D., WENDROFF B., “Systems of conservation laws”, *Communications on Pure and Applied Mathematics*, vol. 43, pp. 217–237, 1960.
- [LAX 77] LE ROUX A.-Y., “A numerical conception of entropy for quasi-linear equations”, *Mathematics of Computation*, vol. 31, pp. 848–872, 1977.
- [LE 99] LE ROUX A.-Y., La formule de Hopf-Lax, course notes, Université Bordeaux 1, 1999.
- [LES 45] LESLIE P.H., “On the use of matrices in certain population mathematics”, *Biometrika*, vol. 33, pp. 183–212, 1945.
- [LES 48] LESLIE P.H., “Some further notes on the use of matrices in certain population mathematics”, *Biometrika*, vol. 35, pp. 213–245, 1948.
- [LIG 55] LIGHTHILL M.J., WHITHAM G.B., “On kinematic waves II: a theory of traffic flow on long crowded roads”, *Proceedings of the Royal Society of London. Series A*, vol. 229, pp. 317–345, 1955.
- [MAL 98] MALTHUS T.R., *An Essay on the Principle of Population as it Affects the Future Improvement of Society*, J. Johnson, London, 1798.
- [MER 98] MERLE F., ZAAG H., “Optimal estimates for blow-up rate and behavior for nonlinear heat equations”, *Communications on Pure and Applied Mathematics*, vol. 51, pp. 139–196, 1998.
- [MER 07] MERLET B., VOVELLE J., “Error estimate for finite volume scheme”, *Numerische Mathematik*, vol. 106, pp. 129–155, 2007.
- [MER 07/08] MERLET B., “ $L^\infty$ - and  $L^2$ -error estimates for a finite volume approximation of linear advection”, *SIAM Journal on Numerical Analysis*, vol. 46, pp. 124–150, 2007/08.
- [MIC 05] MICHEL P., Principe d’entropie relative généralisée et dynamique de populations structurées, PhD thesis, University of Paris-Dauphine, 2005.
- [NEC 96] NECAS J., “Sur les normes équivalentes dans  $W_p^k$  et sur la coercivité des formes formellement positives”, *Séminaire de Mathématiques Supérieures*, Université de Montréal, pp. 102–128, 1966.
- [OLE 57] OLEINIK O.A., “Discontinuous solutions of non-linear differential equations”, *Uspekhi Matematicheskikh Nauk*, vol. 12, pp. 3–73, 1957.
- [OSG 98] OSGOOD W.F., “Beweis der Existenz einer Lösung der Differentialgleichung  $dy/dx = f(x, y)$  ohne Hinzunahme der Cauchy-Lipschitzschen Bedingung”, *Monatshefte der Mathematik und Physik*, vol. 9, pp. 331–345, 1898.
- [PAN 94] PANOV E., “Uniqueness of the solution of the Cauchy problem for a first-order quasilinear equation with an admissible strictly convex entropy”, *Matematicheskie Zametki*, vol. 55, pp. 116–129, 159, 1994.
- [PER 02] PERTHAME B., *Kinetic Formulation of Conservation Laws*, Oxford University Press, 2002.
- [PER 07] PERRON O., “Zur Theorie der Matrices”, *Mathematische Annalen*, vol. 64, pp. 248–263, 1907.

- [PER 91] PERTHAME B., TADMOR E., “A kinetic equation with kinetic entropy functions for scalar conservation laws”, *Communications in Mathematical Physics*, vol. 136, pp. 501–517, 1991.
- [PER 92] PERTHAME B., “Second order Boltzmann schemes for compressible Euler equations in one and two space dimension”, *SIAM Journal on Numerical Analysis*, vol. 29, pp. 1–19, 1992.
- [PER 96] PERRIN, D., “Cours d’algèbre”, *Collection CAPES/Agrég Mathématiques*, Ellipses, 1996.
- [PRÜ 06] PRÜSS J., PUJO-MENJOUET L., WEBB G.F. et al., “Analysis of a model for the dynamics of prions”, *Discrete and Continuous Dynamical Systems – Series B*, vol. 6, pp. 215–225, 2006.
- [PUL 80] PULLIN D.I., “Direct simulation methods for compressible gas flow”, *Journal of Computational Physics*, vol. 34, pp. 231–244, 1980.
- [RAN 70] RANKINE W.J.M., “On the thermodynamic theory of waves of finite longitudinal disturbance”, *Philosophical Transactions of the Royal Society of London*, vol. 160, pp. 277–288, 1870.
- [RAP 12] Rapport du jury de l’agrégation externe de mathématiques, Ministère de l’Education Nationale, 2012.
- [RAS 95] RASCLE M., ZITI C., “Finite time blow-up in some models of chemotaxis”, *Journal of Mathematical Biology*, vol. 33, pp. 388–414, 1995.
- [RIC 56] RICHARDS P.I., “Shockwaves on the highway”, *Operations Research*, vol. 4, pp. 42–51, 1956.
- [RIC 63] RICHTMYER R.D., A survey of difference methods for non-steady fluid dynamics, Technical report, National Center for Atmospheric Research, Boulder, Colorado, 1963.
- [RIE 58/59] RIEMANN B., “Über die fortpanzung ebener luftwellen von endlicher schwingungsweite”, *Abhandlungen der Königlichen Gesellschaft der Wissenschaften in Göttingen (Math. Cl.)*, vol. 8, pp. 43–65, 1858/59.
- [ROG 80] ROGERS C.A., “A less strange version of milnor’s proof of Brouwer fixed-point theorem”, *The American Mathematical Monthly*, vol. 87, pp. 525–527, 1980.
- [RUD 87] RUDIN W., *Real and Complex Analysis*, McGraw-Hill, 1987.
- [SAI 02] SAINT-RAYMOND J., “Local inversion for differentiable functions and the Darboux property”, *Mathematika*, vol. 49, pp. 141–158, 2002.
- [SAI 09] SAINT-RAYMOND L., *Hydrodynamic Limits of the Boltzmann Equation*, Springer, 2009.
- [SCH 90] SCHMIDT M., *Hommes de science, 28 portraits*, Herrmann, 1990.
- [SER 96a] SERRE D., *Systèmes de lois de conservation. Hyperbolicité, entropies, ondes de choc*, Diderot, 1996.
- [SER 96b] SERRE D., *Systèmes de lois de conservation. Structures géométriques, oscillations et problèmes mixtes*, Diderot, 1996.
- [SER 01] SERRE D., *Matrices: Théorie et pratique*, Dunod, 2001.

- [STO 48] STOKES G.G., “On a difficulty in the theory of sound”, *Philosophical Magazine Series 3*, vol. 33, pp. 349–356, 1848.
- [SZA 81] SZANKOWSKI A., “ $B(H)$  does not have the approximation property”, *Acta Mathematica*, vol. 147, pp. 89–108, 1981.
- [TAR 82] TARTAR L., *Topics in Nonlinear Analysis*, Publications Mathématiques d’Orsay, vol. 78.13, University of Paris-Sud, 1982.
- [TAR 87] TARTAR L., “Sur un lemme d’équivalence utilisé en analyse numérique”, *Calcolo*, vol. 24, pp. 129–140, 1987.
- [THO 06] THOMÉE V., *Galerkin Finite Element Methods for Parabolic Problems*, 2nd ed., Springer-Verlag, 2006.
- [VAS 01] VASSEUR A., “Strong traces for solutions of multidimensional scalar conservation laws”, *Archive for Rational Mechanics and Analysis*, vol. 160, pp. 181–193, 2001.
- [WEI 85] WEISSLER F.B., “An  $L^\infty$  blow-up estimate for a nonlinear heat equation”, *Communications on Pure and Applied Mathematics*, vol. 38, pp. 291–295, 1985.
- [WIN 46] WINTNER A., “On the convergence of successive approximations”, *American Journal of Mathematics*, vol. 68, pp. 13–19, 1946.
- [ZUI 02] ZUILY C., *Eléments de distributions et d’équations aux dérivées partielles*, Dunod, 2002.



---

# Index

---

## A, C, D

algorithm

Arrow-Hurwicz, 258, 262

descent, 35, 217

gradient, 174, 222, 223, 235–237, 259,  
413, 414

amplification factor, 272, 278, 280, 304

blow up criterion, 14

conditioning, 93, 178, 224, 261, 262,  
266, 410

conditions

Neumann, 142, 152–154, 164, 171,  
172, 224–227

non-homogeneous Dirichlet, 171

convergence of numerical scheme, 273

d'Alembert formula, 345

elementary solution (Laplace equation), 218

## G, H, J

Gibbs phenomenon, 213

Hamilton–Jacobi equation, 381

Jordan–Chevalley decomposition, 16, 17,  
54, 60

## L, M, N

Lax criterion, 375

Lagrange multiplier, 24, 25, 250, 256

lemma

Céa, 201

Grönwall, 15–16, 98, 108, 160, 334

Le Roux–Harten, 371

Necas, 246

Poincaré (general case), 165

Poincaré (1D), 150

light cone, 351, 352, 348, 349

mass condensation, 283, 287

matrix

dominant diagonal, 173

irreducible, 71, 72

Jacobian, 19, 21, 40, 41, 104, 123, 142,  
237, 238, 388

M-, 99, 175, 278, 279, 281

mass, 195, 282, 283, 289, 290

primitive, 57

rigidity, 195

maximal solution, 11–16

maximum principle, 143, 146, 174

method

characteristics, 295–299

Crank–Nicolson, 50, 80, 284

Hölmgren, 299

Newton, 46–49

shooting, 148–149, 160–162

monotone flux, 368–369

## P, R, S

Picard iteration, 8

Poiseuille flow, 239, 240

power method, 51, 73, 77, 78

- Rankine-Hugoniot relation, 359–362  
saddle point, 256, 257  
scheme  
    Crank-Nicolson, 50, 78, 110, 279,  
        287, 288  
    Lax-Friedrichs, 307–316, 354,  
        373–376, 383, 387  
    Lax-Wendroff, 308–316, 375, 376  
    Rusjanov, 366, 390, 391  
    upwind, 299–306, 310–319, 374–376  
Verlet, 124  
semi simple eigenvalue, 17–19, 75, 275  
stability (von Neumann analysis), 269
- Cauchy–Peano, 7, 8  
implicit function, 21, 24 , 41, 42, 124  
inverse function, 21–24  
Osgood, 8  
Perron–Frobenius, 51–77  
Picard-Lindelöf (general case), 1, 5, 21  
Picard-Lindelöf (global Lipschitzian  
case), 1, 5, 21  
Schauder, 158  
Sobolev Embedding (1D), 50  
transformation  
    Hopf–Cole, 364  
    Legendre, 381  
Uzawa iteration algorithm, 259

## T, U

Taylor expansion, 111, 177, 319  
theorem

- Banach, 2
- Brouwer, 23

---

Other titles from



in

Mathematics and Statistics

---

## 2016

CELANT Giorgio, BRONIATOWSKI Michel

*Interpolation and Extrapolation Optimal Designs 1*

CHIASSERINI Carla Fabiana, GRIBAUDO Marco, MANINI Daniele

*Analytical Modeling of Wireless Communication Systems (Stochastic Models in Computer Science and Telecommunication Networks Set – Volume 1)*

KERN Michel

*Numerical Methods for Inverse Problems*

RYKOV Vladimir

*Reliability of Engineering Systems and Technological Risks*

*Stochastic Models in Survival Analysis and Reliability Set – Volume 1*

## 2015

DE SAPORTA Benoîte, DUFOUR François, ZHANG Huilong

*Numerical Methods for Simulation and Optimization of Piecewise Deterministic Markov Processes*

DEVOLDER Pierre, JANSSEN Jacques, MANCA Raimondo

*Basic Stochastic Processes*

LE GAT Yves

*Recurrent Event Modeling Based on the Yule Process*

(*Mathematical Models and Methods in Reliability Set – Volume 2*)

## 2014

COOKE Roger M., NIEBOER Daan, MISIEWICZ Jolanta

*Fat-tailed Distributions: Data, Diagnostics and Dependence*

(*Mathematical Models and Methods in Reliability Set – Volume 1*)

MACKEVICIUS Vigirdas

*Integral and Measure: From Rather Simple to Rather Complex*

PASCHOS Vangelis Th

*Combinatorial Optimization – 3-volume series – 2<sup>nd</sup> edition*

*Concepts of Combinatorial Optimization / Concepts and Fundamentals – volume 1*

*Paradigms of Combinatorial Optimization – volume 2*

*Applications of Combinatorial Optimization – volume 3*

## 2013

COUALLIER Vincent, GERVILLE-RÉACHE Léo, HUBER Catherine, LIMNIOS

Nikolaos, MESBAH Mounir

*Statistical Models and Methods for Reliability and Survival Analysis*

JANSSEN Jacques, MANCA Oronzio, MANCA Raimondo

*Applied Diffusion Processes from Engineering to Finance*

SERICOLA Bruno

*Markov Chains: Theory, Algorithms and Applications*

## 2012

BOSQ Denis

*Mathematical Statistics and Stochastic Processes*

CHRISTENSEN Karl Bang, KREINER Svend, MESBAH Mounir

*Rasch Models in Health*

DEVOLDER Pierre, JANSSEN Jacques, MANCA Raimondo

*Stochastic Methods for Pension Funds*

## **2011**

MACKEVIČIUS Vigirdas

*Introduction to Stochastic Analysis: Integrals and Differential Equations*

MAHJOUB Ridha

*Recent Progress in Combinatorial Optimization – ISCO2010*

RAYNAUD Hervé, ARROW Kenneth

*Managerial Logic*

## **2010**

BAGDONAVIČIUS Vilijandas, KRUOPIS Julius, NIKULIN Mikhail

*Nonparametric Tests for Censored Data*

BAGDONAVIČIUS Vilijandas, KRUOPIS Julius, NIKULIN Mikhail

*Nonparametric Tests for Complete Data*

IOSIFESCU Marius *et al.*

*Introduction to Stochastic Models*

VASSILIOU PCG

*Discrete-time Asset Pricing Models in Applied Stochastic Finance*

## **2008**

ANISIMOV Vladimir

*Switching Processes in Queuing Models*

FICHE Georges, HÉBUTERNE Gérard

*Mathematics for Engineers*

HUBER Catherine, LIMNIOS Nikolaos *et al.*

*Mathematical Methods in Survival Analysis, Reliability and Quality of Life*

JANSSEN Jacques, MANCA Raimondo, VOLPE Ernesto

*Mathematical Finance*

**2007**

HARLAMOV Boris

*Continuous Semi-Markov Processes*

**2006**

CLERC Maurice

*Particle Swarm Optimization*

# **WILEY END USER LICENSE AGREEMENT**

Go to [www.wiley.com/go/eula](http://www.wiley.com/go/eula) to access Wiley's ebook EULA.



This book provides the mathematical basis for investigating numerical equations from physics, life sciences or engineering. Tools for analysis and algorithms are applied to a large set of relevant examples to show the difficulties and the limitations of the most naïve approaches. Not only do these examples give the opportunity to put into practice mathematical statements, but modeling issues are also addressed in detail through the mathematical perspective.

Chapter 1 addresses the solution of ordinary differential equations, and includes an overview of its essential theoretical basis. The analysis of classical schemes is also covered, and various concepts of stability are illustrated primarily by the description of biological systems. Chapter 2 deals with numerical solutions to elliptic boundary value problems, as well as techniques related to optimization. The final chapter concentrates on evolutionary partial differential equations, looking at questions of stability and consistency for the heat equation and for hyperbolic problems.

**Thierry Goudon** is Senior Research Scientist at Université Côte d'Azur (Inria, CNRS, LJAD) in France. His research is mainly motivated by the analysis and numerical simulation of partial differential equations arising in physics.

