SEMA SIMAI Springer Series 13

Mariano Mateos
Pedro Alonso *Editors*

Computational Mathematics, Numerical Analysis and Applications

Lecture Notes of the XVII 'Jacques-Louis Lions' Spanish-French School

SEMA SIMAI SOCIETÀ ITALIANA DI MATEMATICA APPLICATA E INDUSTRIALE
Springer

SEMA SIMAI Springer Series

Volume 13

More information about this series at http://www.springer.com/series/10532

Mariano Mateos • Pedro Alonso
Editors

Computational Mathematics, Numerical Analysis and Applications

Lecture Notes of the XVII 'Jacques-Louis Lions' Spanish-French School

Springer

Editors
Mariano Mateos
Matemáticas
Universidad de Oviedo
Gijón
Asturias, Spain

Pedro Alonso
Matemáticas
Universidad de Oviedo
Gijón
Asturias, Spain

ISSN 2199-3041
SEMA SIMAI Springer Series
ISBN 978-3-319-49630-6
DOI 10.1007/978-3-319-49631-3

ISSN 2199-305X (electronic)
ISBN 978-3-319-49631-3 (eBook)

Library of Congress Control Number: 2017946701

This Springer imprint is published by Springer Nature
The registered company is Springer International Publishing AG
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

Preface

The 17th edition of the Jacques-Louis Lions Spanish-French School, which addressed Numerical Simulation in Physics and Engineering, took place in Gijón, Spain, in June 2016. The School is a biennial event jointly organized by the Spanish Society of Applied Mathematics, SeMA, and the French Society of Applied and Industrial Mathematics, SMAI. This year, we also celebrated the 25th anniversary of SeMA. More than 80 mathematicians of different nationalities came together in Gijón for 5 days in order to attend the courses and participate in the other events organized for the occasion.

Four-hour courses were delivered by experts in the fields of Optimal Control, High Performance Computing, Numerical Linear Algebra, and Computational Physics. During the school, the attendants—graduate students and also some experienced researchers interested in the organized courses—had the opportunity to present their own work with a poster. Almost twenty participated in the poster session.

The lecture notes for the courses are presented in the first part of this book in the form of long review papers. These papers are authored by very experienced researchers and each one is intended to offer a self-contained presentation of the state of the art in the topic under consideration. We hope that they can be used both as a reference for the interested researcher and as a textbook for graduate students.

In the second part of this publication we present a selection of the extended abstracts submitted to the poster session. Together with these works, we have also included an extended abstract of the conference lecture by J. Calvo, winner of the 19th SeMA Antonio Valle Award, presented to the most outstanding young researcher in 2016.

The short papers in this part, all of which relate to different aspects of computational methods and numerical analysis, do not cover only topics concerning Simulation in Physics and Engineering. They also deal with topics ranging from numerical linear algebra or computational methods in group theory to applications of Mathematics to subjects such as biomedical sciences, chemistry, and quantum physics.

We think that both the courses and the short papers evidence that numerical simulation is no longer a field only applicable to physics or engineering and that, as more applications appear, the need for faster and more reliable methods in numerical linear algebra and computational techniques will become more pressing.

The first six papers in the second part correspond to the works presented at the school by J. Calvo, M. Garzon, S. Busto, J.R. Rodríguez-Galván, N. Esteban, and H. Al Rachid. We can say that these works fall into the classical definition of "applied mathematics", where some numerical method is developed and investigated to solve some aspect of a physical model.

The work by J.A. Huidobro et al. investigates different models in Chemistry and compares them with actual experimental data to develop a new simpler model to solve the problem.

The eighth extended abstract, introduced at the school by M.L. Serrano, investigates several aspects of numerical linear algebra, in close connection with the lecture notes of the course delivered by J.M. Peña and also related to the lecture notes of the course delivered by L. Grigori. Solving large scale systems of linear equations has become a necessity for the mathematical community. For instance, in the numerical experiments shown at the end of the course by E. Casas and M. Mateos, the nonlinear system (73)–(76) has more than one million unknowns and to solve it not just one but a sequence of linear systems with a huge number of variables must be solved.

The interesting paper by J. Martínez Carracedo and C. Martínez López shows how computer-based techniques can be applied to prove abstract algebra results.

The last two works, which correspond to posters presented by J.C. Beltrán and M. Loureiro-Ga, deal with applications of Mathematics to medical sciences. Here, we find again the usual language of applied mathematics: least squares, PDEs, discrete approximations. But the focus is on the applications of numerical simulation as another tool to help medical doctors in research and clinical work.

Finally, we want to thank all the contributors (more than forty) who have co-authored the articles contained in this volume, as well as the anonymous referees who have revised the work.

Gijón, Spain
April 2017

Pedro Alonso
Mariano Mateos

Acknowledgements

We would also like to thank the following people and institutions for making possible the edition of this volume:

- The Spanish Society for Applied Mathematics, SeMA, its outgoing president Rafael Bru and its current president, Rosa Donat.
- The French Society for Industrial and Applied Mathematics, SMAI, and its president Fatiha Alabau.
- Our funding sponsors:
  - Gijón Convention Boureau.
  - Accenture Digital.
  - Department of Mathematics of the Universidad de Oviedo.
  - Embassy of France in Madrid.
- The scientific committee, formed by the members of SeMA Inmaculada Higueras (U. de Navarra), Carlos Vázquez (U. de A Coruña), Salim Meddahai (U. de Oviedo) and the members of SMAI Emmanuel Trelat (U. Pierre et Marie Curie P6), Cristophe Prud'homme (U. de Strasbourg) and Bruno Bouchard (U. Paris-Dauphine).
- The other members of the local organizing committee, Rafael Gallego, María Luisa Serrano, Jesús Suárez Pérez-del-Río and Virginia Selgas.
- The staff of the Fundación Universidad de Oviedo.
- The staff of the Hotel Tryp Rey Pelayo.
- The editorial board of the SeMA-SIMAI Springer Series and the staff of Springer.

Contents

Contributors

B.F. Alfonso Departamento de Fisica, Universidad de Oviedo, Gijón, Spain

P. Alonso University of Oviedo, Gijón, Spain

Houssam AlRachid Université Paris-Est Créteil, Créteil, France

Juan Carlos Beltrán Vargas Department of Mathematics, University of Oviedo, Oviedo, Spain

A. Bermúdez Facultade de Matemáticas, Departamento de Matemática Aplicada, Universidade de Santiago de Compostela, Santiago de Compostela, Spain

S. Busto Facultade de Matemáticas, Departamento de Matemática Aplicada, Universidade de Santiago de Compostela, Santiago de Compostela, Spain

Francisco Calvo Instituto de Investigación Sanitaria Galicia Sur, Cardioloxía, Hospital Álvaro Cunqueiro, Vigo, Spain

Juan Calvo Facultad de Ciencias, Universidad de Granada, Campus de Fuentenueva, Granada, Spain

Eric Cancès CERMICS, Ecole des Ponts and Inria Paris, Marne-la-Vallée, France

Jorge Martínez Carracedo Department of Mathematics, University of Oviedo, Oviedo, Spain

E. Carrizosa Departamento de Estadística e Investigación Operativa, Universidad de Sevilla, C/ Tarfia S/N, Sevilla, Spain

Eduardo Casas Departmento de Matemática Aplicada y Ciencias de la Computación, Universidad de Cantabria, Santander, Spain

Ana Cernea Department of Mathematics, University of Oviedo, Oviedo, Spain

Ó. Crego Departamento de Matemática Aplicada, Universidad de Santiago de Compostela, Santiago de Compostela, Spain

Enrique J. deAndrés-Galiana Department of Mathematics, University of Oviedo, Oviedo, Spain

N. Esteban Departamento de Matemática Aplicada, Universidad de Santiago de Compostela, Campus Vida, Santiago de Compostela, Spain

G. Fernández-Manin Departamento de Matemática Aplicada II, Universidade de Vigo, Vigo, Spain

Juan Luis Fernández-Martínez Department of Mathematics, University of Oviedo, Oviedo, Spain

J.L. Ferrín Facultade de Matemáticas, Departamento de Matemática Aplicada, Universidade de Santiago de Compostela, Santiago de Compostela, Spain

J.R. Garcia Departamento de Quimica Organica e Inorganica, Universidade de Oviedo, Oviedo, Spain

Maria F. Garcia Instituto de Investigación Sanitaria Galicia Sur, Cardioloxía, Hospital Álvaro Cunqueiro, Vigo, Spain

Maria Garzon Universidad de Oviedo, Oviedo, Spain

Laura Grigori Inria Paris, Alpines, and UPMC Univ Paris 06, CNRS UMR 7598, Laboratoire Jacques-Louis Lions, Paris, France

F. Guillén-González Departamento EDAN and IMUS, Universidad de Sevilla, Sevilla, Spain

J.A. Huidobro Departamento de Matematicas, Universidad de Oviedo, Gijón, Spain

I. Iglesias Departamento de Fisica, Universidad de Oviedo, Gijón, Spain

Andrés Iñiguez Instituto de Investigación Sanitaria Galicia Sur, Cardioloxía, Hospital Álvaro Cunqueiro, Vigo, Spain

Victor Jimenez Instituto de Investigación Sanitaria Galicia Sur, Cardioloxía, Hospital Álvaro Cunqueiro, Vigo, Spain

August Johansson Center for Biomedical Computing, Simula, Norway

Tony Lelièvre École des Ponts ParisTech, Université Paris Est, Marne-la-Vallée, France

Consuelo Martínez López Department of Mathematics, University of Oviedo, Oviedo, Spain

Marcos Loureiro-Ga Universidade de Santiago de Compostela, Santiago de Compostela, Spain
Instituto de Investigación Sanitaria Galicia Sur, Cardioloxía, Hospital Álvaro Cunqueiro, Vigo, Spain

Mariano Mateos Departmento de Matemáticas, E.P.I. Gijón, Universidad de Oviedo, Campus de Gijón, Gijón, Spain

Emilio Paredes Instituto de Investigación Sanitaria Galicia Sur, Cardioloxía, Hospital Álvaro Cunqueiro, Vigo, Spain

J.M. Peña University of Zaragoza, Zaragoza, Spain

M.V. Redondo-Neble Departamento de Matemáticas, Universidad de Cádiz, Cádiz, Spain

J.F. Rodríguez-Calo Centro de Tecnología, Autovía de Extremadura, Móstoles, Madrid, Spain

J.R. Rodríguez-Galván Departamento de Matemáticas, Universidad de Cádiz, Cádiz, Spain

L. Saavedra Departamento de Matemática Aplicada a la Ingeniería Aeroespacial, Universidad Politécnica de Madrid E.T.S.I. Aeronáuticos, Madrid, Spain

M.L. Serrano University of Oviedo, Oviedo, Spain

James A. Sethian University of Berkeley, Berkeley, CA, USA

E.F. Toro Laboratory of Applied Mathematics, DICAM, University of Trento, Trento, Italy

C. Trobajo Departamento de Quimica Organica e Inorganica, Universidad de Oviedo, Oviedo, Spain

M.E. Vázquez-Cendón Facultade de Matemáticas, Departamento de Matemática Aplicada, Universidade de Santiago de Compostela, Santiago de Compostela, Spain

Cesar Veiga Instituto de Investigación Sanitaria Galicia Sur, Cardioloxía, Hospital Álvaro Cunqueiro, Vigo, Spain

About the Editors

Pedro Alonso is currently a professor at the Department of Mathematics at the University of Oviedo (Spain). He received his PhD in mathematics from the University of Oviedo in 1995. His main interests include numerical linear algebra, error analysis, study of algorithms (complexity, performance, stability, convergence, etc.), high-performance computing, and mathematics education. He has several publications in international journals and communications for international conferences to his credit.

Mariano Mateos graduated from the University of Oviedo with a degree in mathematics in 1995 and in 2000 completed his PhD at the University of Cantabria, where he is currently a member of the research group "Optimal Control of Partial Differential Equations." He is the author of several works on this subject and participates regularly in international conferences on applied mathematics. In 2016, he chaired the Spanish-French School on Numerical Simulation organized by SeMA and SMAI. Currently, he reads numerical methods at the Engineering School of Gijón as an associate professor.

Part I
Theory

Optimal Control of Partial Differential Equations

Eduardo Casas and Mariano Mateos

Abstract In this chapter, we present an introduction to the optimal control of partial differential equations. After explaining what an optimal control problem is and the goals of the analysis of these problems, we focus the study on a model example. We consider an optimal control problem governed by a semilinear elliptic equation, the control being subject to bound constraints. Then we explain the methods to prove the existence of a solution; to derive the first and second order optimality conditions; to approximate the control problem by discrete problems; to prove the convergence of the discretization and to get some error estimates. Finally we present a numerical algorithm to solve the discrete problem and we provide some numerical results. Though the whole analysis is done for an elliptic control problem, with distributed controls, some other control problems are formulated, which show the scope of the field of control theory and the variety of mathematical methods necessary for the analysis. Among these problems, we consider the case of evolution equations, Neumann or Dirichlet boundary controls, and state constraints.

1 Introduction

In an optimal control problem, we find the following basic elements.

1. A *control* u that we can handle according to our interests, that can be chosen among a family of feasible controls $\mathbb{K}$.
2. The *state of the system* y to be controlled, that depends on the control. Some limitations can be imposed on the state, in mathematical terms $y \in \mathbb{C}$, which means that not every possible state of the system is satisfactory.

E. Casas (✉)
Departmento de Matemática Aplicada y Ciencias de la Computación, Universidad de Cantabria, 39005 Santander, Spain
e-mail: eduardo.casas@unican.es

M. Mateos
Departmento de Matemáticas, E.P.I. Gijón, Universidad de Oviedo, Campus de Gijón, 33203 Gijón, Spain
e-mail: mmateos@uniovi.es

M. Mateos, P. Alonso (eds.), *Computational Mathematics, Numerical Analysis and Applications*, SEMA SIMAI Springer Series 13,
DOI 10.1007/978-3-319-49631-3_1

3. A *state equation* that establishes the dependence between the control and the state. In the next sections this state equation will be a partial differential equation, y being the solution of the equation and u a function arising in the equation so that any change in the control u produces a change in the solution y. However the origin of control theory was connected with the control of systems governed by ordinary differential equations and there was a huge activity in this field; see, for instance, the classical books Pontriaguine et al. [41] or Lee and Markus [31].
4. A *function* to be minimized, called the objective function or the cost function, depending on the state and the control (y, u).

The aim is to determine an admissible control that provides a satisfactory state and that minimizes the value of the functional J. It is called the optimal control, and the associated state is the optimal state. The basic questions to study are the existence of a solution and methods for its computation. However to obtain the solution we must use some numerical methods, that leads to some delicate mathematical questions in this numerical analysis. The first step to solve numerically the problem requires the discretization of the control problem that is made usually by finite elements. A natural question is how good the approximation is. Of course we would like to have some error estimates of these approximations. In order to derive the error estimates, some regularity of the optimal control is essential, more precisely, some order of differentiability (at least in a weak sense) is necessary. The regularity of the optimal control can be deduced from the first order optimality conditions. Another key tool in the proof of error estimates is the use of second order sufficient optimality conditions. Therefore, our analysis requires to derive the first and second order conditions for optimality. This will be analyzed in this paper.

Once we have a discrete control problem, we have to use some numerical algorithm of optimization to solve this problem. When the problem is not convex, the optimization algorithms typically provides local minima, the question now is if these local minima are significant for the original control problem.

The following steps must be performed when we study an optimal control problem:

1. Existence of a solution.
2. First and second order optimality conditions.
3. Numerical approximation. Convergence analysis and error estimates.
4. Numerical resolution of the discrete control problem.

We will will consider these issues for a model problem. In this model problem the state equation will be a semilinear elliptic partial differential equation. Though the nonlinearity introduces some complications in the study, we have preferred to consider it to show the role played by the second order optimality conditions. Indeed, if the equation is linear and the cost functional is the typical quadratic functional, then the use of the second order optimality conditions is hidden.

There are not many books devoted to all the questions we are going to study here. Firstly let us mention the book by professor Lions [33], which is an obliged reference in the study of the theory of optimal control problems of partial differential equations. In this text that has left an indelible track, the reader will be able to find some of the methods used in the resolution of the two first questions above indicated. More recent books are those by Li and Yong [32], Fattorini [23], Neittaanmaki et al. [38], Hinze et al. [27] and Tröltzsch [48].

2 Setting of the Model Control Problem

Let Ω be an open and bounded subset of $\mathbb{R}^n$, $n \in \{2, 3\}$, Γ being its boundary that we will assume to be regular; $C^{1,1}$ is enough for us in the whole paper. In Ω we will consider a linear operator A defined by

$$Ay = -\sum_{i,j=1}^{n} \partial_{x_j} \left(a_{ij}(x) \partial_{x_i} y(x) \right) + a_0(x) y(x),$$

where $a_{ij} \in C^{0,1}(\bar{\Omega})$ and $a_0 \in L^\infty(\Omega)$ satisfy

$$\begin{cases} \exists m > 0 \text{ such that } \displaystyle\sum_{i,j=1}^{n} a_{ij}(x)\xi_i\xi_j \geq m|\xi|^2 \ \ \forall \xi \in \mathbb{R}^n \text{ and } \forall x \in \Omega, \\ \\ a_0(x) \geq 0 \ \text{ a.e. } x \in \Omega. \end{cases}$$

Now let $\phi : \mathbb{R} \longrightarrow \mathbb{R}$ be a non decreasing monotone function of class C^2, with $\phi(0) = 0$. For any $u \in L^2(\Omega)$, the Dirichlet problem

$$\begin{cases} Ay + \phi(y) = u \ \text{ in } \ \Omega \\ y = 0 \ \text{ on } \ \Gamma \end{cases} \tag{1}$$

has a unique solution $y_u \in H_0^1(\Omega) \cap L^\infty(\Omega)$.

The control problem associated to this system is formulated as follows

$$\text{(P)} \begin{cases} \text{Minimize } \ J(u) = \displaystyle\int_\Omega L(x, y_u(x), u(x))dx \\ \\ u \in \mathbb{K} = \{u \in L^\infty(\Omega) : \alpha \leq u(x) \leq \beta \ \text{ a.e. } x \in \Omega\}, \end{cases}$$

where $-\infty < \alpha < \beta < +\infty$ and L fulfills the following assumptions:

(H1) $L : \Omega \times \mathbb{R}^2 \longrightarrow \mathbb{R}$ is a Carathéodory function and for all $x \in \Omega$, $L(x, \cdot, \cdot)$ is of class C^2 in $\mathbb{R}^2$. Moreover for every $M > 0$ and all $x, x_1, x_2 \in \Omega$ and $y, y_1, y_2, u, u_1, u_2 \in [-M, +M]$, the following properties hold

$$|L(x, y, u)| \le L_{M,1}(x), \quad |\frac{\partial L}{\partial y}(x, y, u)| \le L_{M,\bar{p}}(x)$$

$$|\frac{\partial L}{\partial u}(x_1, y, u) - \frac{\partial L}{\partial u}(x_2, y, u)| \le C_M |x_1 - x_2|$$

$$|L''_{(y,u)}(x, y, u)|_{\mathbb{R}^{2\times 2}} \le C_M$$

$$|L''_{(y,u)}(x, y_1, u_1) - L''_{(y,u)}(x, y_2, u_2)|_{\mathbb{R}^{2\times 2}} \le C_M(|y_1 - y_2| + |u_1 - u_2|),$$

where $L_{M,1} \in L^1(\Omega)$, $L_{M,\bar{p}} \in L^{\bar{p}}(\Omega)$, $\bar{p} > n$, $C_M > 0$, $L''_{(y,u)}$ is the Hessian matrix of L with respect to (y, u), and $|\cdot|_{\mathbb{R}^{2\times 2}}$ is any matricial norm.
To prove sufficient second order optimality conditions and error estimates, we will need the following additional assumption

(H2) There exists $\Lambda > 0$ such that

$$\frac{\partial^2 L}{\partial u^2}(x, y, u) \ge \Lambda \quad \forall\ (x, y, u) \in \Omega \times \mathbb{R}^2.$$

Remark 1 A typical functional in control theory is the so-called tracking type functional

$$J(u) = \int_\Omega \left\{ |y_u(x) - y_d(x)|^2 + Nu^2(x) \right\} dx, \tag{2}$$

where $y_d \in L^2(\Omega)$ denotes the ideal state of the system and $N \ge 0$. The term $\int_\Omega Nu^2(x)dx$ is called the Tikhonov term. It can be considered as the control cost term, and the control is said expensive if N is big, however the control is cheap if N is small or zero. From a mathematical point of view, the presence of the term $\int_\Omega Nu^2(x)dx$, with $N > 0$, has a regularizing effect on the optimal control. Hypothesis **(H1)** is fulfilled if $y_d \in L^p(\Omega)$. This condition plays an important role in the study of the regularity of the optimal control. Hypothesis **(H2)** holds if $N > 0$.

Remark 2 Other choices for the set of feasible controls are possible, in particular the case $\mathbb{K} = L^2(\Omega)$ is frequent. The important issue is that $\mathbb{K}$ must be closed and convex. Moreover, if $\mathbb{K}$ is not bounded, then some coercivity assumption on the functional J is required to assure the existence of a solution.

Remark 3 In practice, $\phi(0) = 0$ is not a true restriction because it is enough to change ϕ by $\phi - \phi(0)$ and u by $u - \phi(0)$ to get a new problem satisfying the

required assumptions. Nonlinear terms of the form $f(x, y(x))$, with f of class C^2 with respect to the second variable and monotone non decreasing with respect to the same variable, can be considered as an alternative to the term $\phi(y(x))$. We lose some generality in order to avoid technicalities and to get a simplified and more clear presentation of our methods to study the control problem.

Given $\frac{n}{2} < p < +\infty$ and $u \in L^p(\Omega)$, we can prove the existence of a unique solution y_u of (1) in $W^{2,p}(\Omega) \cap H_0^1(\Omega)$ as follows. First, we prove the existence of a solution y_u in $H_0^1(\Omega) \cap L^\infty(\Omega)$: we truncate ϕ to get a bounded function ϕ_k, for instance in the way

$$\phi_k(t) = \begin{cases} \phi(t) & \text{if } |t| \le k, \\ \phi(+k) & \text{if } t > +k, \\ \phi(-k) & \text{if } t < -k. \end{cases}$$

Then, the operator $(A + \phi_k) : H_0^1(\Omega) \longrightarrow H^{-1}(\Omega)$ is monotone, continuous and coercive. Therefore there exists a unique element $y_k \in H_0^1(\Omega)$ satisfying $Ay_k + \phi_k(y_k) = u$ in Ω. By using the usual methods it is easy to prove that $\{y_k\}_{k=1}^\infty$ is uniformly bounded in $L^\infty(\Omega)$ (see, for instance, Stampacchia [46]). Consequently for k large enough $\phi_k(y_k) = \phi(y_k)$ and then $y_k = y_u \in H_0^1(\Omega) \cap L^\infty(\Omega)$ is the solution of problem (1). On the other hand, the $C^{1,1}$ regularity of Γ and the fact that $Ay_u \in L^p(\Omega)$ imply the $W^{2,p}(\Omega)$-regularity of y_u; see Grisvard [24, Chap. 2]. Thus we have the following theorem.

Theorem 4 *For any control $u \in L^p(\Omega)$ with $\frac{n}{2} < p < +\infty$ there exists a unique solution y_u of (1) in $W^{2,p}(\Omega) \cap H_0^1(\Omega)$. Moreover, there exists a constant $C_p > 0$ independent of u such that*

$$\|y_u\|_{W^{2,p}(\Omega)} \le C_p\big(\|u\|_{L^p(\Omega)} + 1\big). \tag{3}$$

Finally, remembering that $\mathbb{K}$ is bounded in $L^\infty(\Omega)$, we deduce the next result.

Corollary 5 *For any control $u \in \mathbb{K}$ there exists a unique solution y_u of (1) in $W^{2,p}(\Omega) \cap H_0^1(\Omega)$, for all $p < \infty$. Moreover, there exists a constant $C_p > 0$ such that*

$$\|y_u\|_{W^{2,p}(\Omega)} \le C_p \quad \forall u \in \mathbb{K}. \tag{4}$$

It is important to remark that the previous corollary implies $C^1(\bar{\Omega})$ regularity of y_u. Indeed, it is enough to remind that $W^{2,p}(\Omega) \subset C^1(\bar{\Omega})$ for any $p > n$.

3 Existence of a Solution

The goal of this section is to study the existence of a solution for problem (P), which is done in the following theorem.

Theorem 6 *Let us assume that L is a Carathéodory function satisfying the following assumptions:*

A1) *For every* $(x, y) \in \Omega \times \mathbb{R}$, $L(x, y, \cdot) : \mathbb{R} \longrightarrow \mathbb{R}$ *is a convex function.*
A2) *For any* $M > 0$, *there exists a function* $\psi_M \in L^1(\Omega)$ *such that*

$$|L(x, y, u)| \leq \psi_M(x) \ \ a.e.\ x \in \Omega, \ \ \forall |y| \leq M, \ \ \forall |u| \leq M.$$

Then problem (P) has at least one solution.

Proof Denote $\inf(\mathrm{P}) = \inf\{J(u) : u \in \mathbb{K}\}$. Let $\{u_k\} \subset \mathbb{K}$ be a minimizing sequence of (P), this means that $J(u_k) \to \inf(\mathrm{P})$. Take a subsequence, again denoted in the same way, converging weakly* in $L^\infty(\Omega)$ to an element $\bar{u} \in \mathbb{K}$. Let us prove that $J(\bar{u}) = \inf(\mathrm{P})$. For this we will use Mazur's Theorem (see, for instance, Ekeland and Temam [22]): given $1 < p < +\infty$ arbitrary, there exists a sequence of convex combinations $\{v_k\}_{k\in\mathbb{N}}$,

$$v_k = \sum_{l=k}^{n_k} \lambda_l u_l, \ \ \text{with} \ \ \sum_{l=k}^{n_k} \lambda_l = 1 \ \ \text{and} \ \ \lambda_l \geq 0,$$

such that $v_k \to \bar{u}$ strongly in $L^p(\Omega)$. Then, using the convexity of L with respect to the third variable, the dominated convergence theorem and the assumption A1), it follows

$$J(\bar{u}) = \lim_{k\to\infty} \int_\Omega L(x, y_{\bar{u}}(x), v_k(x))dx \leq$$

$$\limsup_{k\to\infty} \sum_{l=k}^{n_k} \lambda_l \int_\Omega L(x, y_{\bar{u}}(x), u_l(x))dx \leq \limsup_{k\to\infty} \sum_{l=k}^{n_k} \lambda_l J(u_l) +$$

$$\limsup_{k\to\infty} \int_\Omega \sum_{l=k}^{n_k} \lambda_l \left| L(x, y_{u_l}(x), u_l(x)) - L(x, y_{\bar{u}}(x), u_l(x)) \right| dx =$$

$$\inf(\mathrm{P}) + \limsup_{k\to\infty} \int_\Omega \sum_{l=k}^{n_k} \lambda_l \left| L(x, y_{u_l}(x), u_l(x)) - L(x, y_{\bar{u}}(x), u_l(x)) \right| dx,$$

where we have used the convergence $J(u_k) \to \inf(\mathrm{P})$. To prove that the last term converges to zero, it is enough to remark that, for any given point x, the function $L(x, \cdot, \cdot)$ is uniformly continuous on bounded subsets of $\mathbb{R}^2$, the sequences $\{y_{u_l}(x)\}$

and $\{u_l(x)\}$ are uniformly bounded and $y_{u_l}(x) \to y_{\bar{u}}(x)$ when $l \to \infty$. Therefore

$$\lim_{k\to\infty} \sum_{l=k}^{n_k} \lambda_l \left| L(x, y_{u_l}(x), u_l(x)) - L(x, y_{\bar{u}}(x), u_l(x)) \right| = 0 \ \text{ a.e. } x \in \Omega.$$

Using again the dominated convergence theorem, assumption A2) and the previous convergence, we get

$$\limsup_{k\to\infty} \int_\Omega \sum_{l=k}^{n_k} \lambda_l \left| L(x, y_{u_l}(x), u_l(x)) - L(x, y_{\bar{u}}(x), u_l(x)) \right| dx = 0,$$

which concludes the proof. □

Remark 7 It is possible to formulate other similar problems to (P) by taking $\mathbb{K}$ as a closed and convex subset of $L^p(\Omega)$, with $\frac{n}{2} < p < +\infty$. The existence of a solution of this kind of problems can be proved as above by assuming that $\mathbb{K}$ is bounded in $L^p(\Omega)$ or J is coercive on $\mathbb{K}$. The coercivity holds if the following conditions is fulfilled: $\exists \psi \in L^1(\Omega)$ and $C > 0$ such that

$$L(x, y, u) \geq C|u|^p + \psi(x) \ \ \forall (x, y, u) \in \Omega \times \mathbb{R}^2.$$

This coercivity assumption implies the boundedness in $L^p(\Omega)$ of any minimizing sequence, the rest of the proof being as in Theorem 6.

Remark 8 If there is neither convexity nor compactness, we cannot assure, in general, the existence of a solution. Let us see an example.

$$\text{(P)} \begin{cases} \text{Minimize } \ J(u) = \displaystyle\int_\Omega [y_u(x)^2 + (u^2(x) - 1)^2] dx \\ -1 \leq u(x) \leq +1, \ \ x \in \Omega, \end{cases}$$

where y_u is the solution of the state equation

$$\begin{cases} -\Delta y = u \ \text{ in } \Omega \\ \quad y = 0 \text{ on } \Gamma. \end{cases}$$

Let us take a sequence of controls $\{u_k\}_{k=1}^\infty$ such that $|u_k(x)| = 1$ for every $x \in \Omega$ and satisfying that $u_k \rightharpoonup 0$ weakly* in $L^\infty(\Omega)$. The reader can construct such a sequence (include Ω in a n-cube to simplify the proof). Then, taking into account that $y_{u_k} \to 0$ uniformly in Ω, we have

$$0 \leq \inf_{-1\leq u(x)\leq +1} J(u) \leq \lim_{k\to\infty} J(u_k) = \lim_{k\to\infty} \int_\Omega y_{u_k}(x)^2 dx = 0.$$

But it is obvious that $J(u) > 0$ for any feasible control, which proves the non-existence of an optimal control.

In [4] and [5], some compactness of the control set was used to prove the existence of optimal controls.

To deal with control problems in the absence of convexity and compactness, (P) is sometimes included in a more general problem $(\bar{\mathrm{P}})$, in such a way that $\inf(\mathrm{P}) = \inf(\bar{\mathrm{P}})$, $(\bar{\mathrm{P}})$ having a solution. This leads to the relaxation theory; see Ekeland and Temam [22], Pedregal [40], Roubiček [43], Warga [49], Young [50].

4 Some Other Control Problems

In this section, we are going to present some control problems whose existence of solution can be proved by using the previous methods. First let us start with a very well known problem, which is a particular case of (P).

4.1 The Linear Quadratic Control Problem

Let us assume that ϕ is linear and $L(x, y, u) = (1/2)\{(y - y_d(x))^2 + Nu^2\}$, with $y_d \in L^2(\Omega)$ and $N \geq 0$ fixed, therefore

$$J(u) = \frac{1}{2}\int_\Omega (y_u(x) - y_d(x))^2 dx + \frac{N}{2}\int_\Omega u^2(x)dx.$$

Now (P) is a convex control problem. In fact the objective functional $J : L^2(\Omega) \to \mathbb{R}$ is well defined, continuous and strictly convex. Under these conditions, if $\mathbb{K}$ is a convex and closed subset of $L^2(\Omega)$, we can prove the existence and uniqueness of an optimal control under one of the two following assumptions:

1. $\mathbb{K}$ is a bounded subset of $L^2(\Omega)$.
2. $N > 0$.

For the proof it is enough to take a minimizing sequence as in Theorem 6, and remark that the previous assumptions imply the boundedness of the sequence. Then it is possible to take a subsequence $\{u_k\}_{k=1}^\infty \subset \mathbb{K}$ converging weakly in $L^2(\Omega)$ to $\bar{u} \in \mathbb{K}$. Finally the convexity and continuity of J implies the weak lower semicontinuity of J, then

$$J(\bar{u}) \leq \liminf_{k\to\infty} J(u_k) = \inf(\mathrm{P}).$$

The uniqueness of the solution is an immediate consequence of the strict convexity of J.

If $N > 0$, the term $\frac{N}{2}\int_\Omega u^2(x)dx$ is called Tychonoff regularization term. In this case, it is usually possible to prove that the optimal control is more regular than expected, e.g. it maybe a Lipschitz function, and both the analysis and the numerical approximation of (P) are simpler than in the case $N = 0$.

4.2 A Neumann Boundary Control Problem

Let us consider the Neumann problem

$$\begin{cases} Ay + \phi(y) = f & \text{in } \Omega \\ \partial_{\nu_A} y = u & \text{on } \Gamma, \end{cases}$$

where $f \in L^\rho(\Omega)$, $\rho > n/2$, $u \in L^s(\Gamma)$, $s > n-1$ and

$$\partial_{\nu_A} y = \sum_{i,j=1}^{n} a_{ij}(x)\partial_{x_i} y(x)\nu_j(x),$$

$\nu(x)$ being the unit outward normal vector to Γ at the point x.

The choice $\rho > n/2$ and $s > n-1$ allows us to deduce a theorem of existence and uniqueness analogous to Theorem 4, assuming that $a_0 \not\equiv 0$.

The control problem is defined as follows

$$\text{(P)} \begin{cases} \text{Minimize } J(u) = \displaystyle\int_\Omega L(x, y_u(x))\,dx + \int_\Gamma l(x, y_u(x), u(x))\,d\sigma(x) \\ \\ u \in \mathbb{K} = \{u \in L^\infty(\Gamma) : \alpha \le u(x) \le \beta \text{ a.e. } x \in \Gamma\}. \end{cases}$$

4.3 A Dirichlet Boundary Control Problem

Now we are concerned with the Dirichlet problem

$$\begin{cases} Ay + \phi(y) = f & \text{in } \Omega \\ y = u & \text{on } \Gamma, \end{cases}$$

where $f \in L^\rho(\Omega)$, $\rho > n/2$, $u \in L^\infty(\Gamma)$.

Associated to this boundary value problem we consider the control problem

$$\text{(P)} \begin{cases} \text{Minimize } J(u) = \displaystyle\int_\Omega L(x, y_u(x))\,dx + \frac{N}{2}\int_\Gamma u(x)^2\,d\sigma(x) \\ \\ u \in \mathbb{K} = \{u \in L^\infty(\Gamma) : \alpha \le u(x) \le \beta \text{ a.e. } x \in \Gamma\}. \end{cases}$$

4.4 A Parabolic Control Problem

Let us consider the following parabolic equation:

$$\begin{cases} \dfrac{\partial y}{\partial t}(x,t) + Ay(x,t) + b(x,t,y(x,t)) = u(x,t) \text{ in } \Omega_T = \Omega \times (0,T), \\ y(x,t) = 0 \text{ on } \Sigma_T = \Gamma \times (0,T), \\ y(x,0) = y_0(x) \text{ in } \Omega. \end{cases}$$

Here $y_0 \in L^\infty(\Omega) \cap H_0^1(\Omega)$ and b is a Carathédoroy function, non-decreasing monotone with respect to the third variable and locally bounded. For every $u \in L^\infty(\Omega_T)$, the previous problem has a unique solution $y_u \in L^\infty(\Omega_T) \cap L^2([0,T], H_0^1(\Omega))$.

For $N > 0$ and $y_d \in L^\infty(\Omega_T)$, we can formulate a control problem as follows:

$$\text{(P)} \begin{cases} \text{Minimize } J(u) = \dfrac{1}{2}\displaystyle\int_{\Omega_T} (y_u(x,t) - y_d(x))^2 dxdt + \frac{N}{2}\int_{\Omega_T} u(x)^2 dxdt \\ u \in \mathbb{K} = \{u \in L^\infty(\Omega_T) : \alpha \le u(x,t) \le \beta \text{ a.e. } (x,t) \in \Omega_T\}. \end{cases}$$

4.5 A Problem with State Constraints

Under the same notation and conditions of the previous example, with $y_0 \in C(\bar{\Omega}_T)$, we consider the following state constrained control problem

$$\text{(P)} \begin{cases} \text{Minimize } J(u) \\ u \in \mathbb{K} \text{ and } G(y_u) \in C, \end{cases}$$

where $G : Y \longrightarrow Z$ is a C^1 mapping, $Y = C(\bar{\Omega}_T) \cap L^2([0,T], H^1(\Omega))$, Z being a Banach space, and C is a closed convex subset of Z with nonempty interior. Due to the continuity assumption of y_0, the solution y_u of the above parabolic equation is continuous in $\bar{\Omega}_T$. Let us consider some examples of state constraints $G(y_u) \in C$.

Example 9 Given a continuous function $g : \bar{\Omega}_T \times \mathbb{R} \longrightarrow \mathbb{R}$ of class C^1 respect to the second variable, the constraint $a \le g(x,t,y_u(x,t)) \le b$ for all $(x,t) \in \bar{\Omega}_T$, where $-\infty < a < b < +\infty$ are given real numbers, can be written in the above framework

by setting $Z = C(\bar{\Omega}_T)$, $G : Y \longrightarrow C(\bar{\Omega}_T)$, defined by $G(y) = g(\cdot, y(\cdot))$, and

$$C = \{z \in C(\bar{\Omega}_T) : a \leq z(x,t) \leq b \ \ \forall (x,t) \in \bar{\Omega}_T\}.$$

Example 10 Let $g : \Omega \times [0,T] \times \mathbb{R} \longrightarrow \mathbb{R}$ be a function measurable with respect to the first variable, continuous with respect to the second, of class C^1 with respect to the third and such that $\partial g/\partial y$ is also continuous in the last two variables. Moreover it is assumed that for every $M > 0$ there exists a function $\psi_M \in L^1(\Omega)$ such that

$$|g(x,t,0)| + \left|\frac{\partial g}{\partial y}(x,t,y)\right| \leq \psi_M(x) \ \ a.e.\ x \in \Omega, \ \ \forall t \in [0,T] \text{ and } |y| \leq M.$$

Then the constraint

$$\int_\Omega g(x,t,y_u(x,t))dx \leq \delta \ \ \forall t \in [0,T]$$

is included in the above formulation by taking $Z = C[0,T]$,

$$C = \{z \in C[0,T] : z(t) \leq \delta \ \ \forall t \in [0,T]\},$$

and $G : Y \longrightarrow C[0,T]$ given by

$$G(y) = \int_\Omega g(x,\cdot,y(x,\cdot))dx.$$

Example 11 The constraint

$$\int_{\Omega_T} |y_u(x,t)|dxdt \leq \delta$$

is considered by taking $Z = L^1(\Omega_T)$, $G : Y \longrightarrow L^1(\Omega)$, with $G(y) = y$, and C the closed ball in $L^1(\Omega)$ of center at 0 and radius δ.

Example 12 For every $1 \leq j \leq k$ let $g_j : \Omega_T \times \mathbb{R} \longrightarrow \mathbb{R}$ be a measurable function of class C^1 with respect to the second variable such that for each $M > 0$ there exists a function $\eta_M^j \in L^1(\Omega_T)$ satisfying

$$|g_j(x,t,0)| + \left|\frac{\partial g_j}{\partial y}(x,t,y)\right| \leq \eta_M^j(x,t) \ \ a.e.\ (x,t) \in \Omega_T, \ \ \forall |y| \leq M.$$

Then the constraints

$$\int_\Omega g_j(x,t,y_u(x,t))dxdt \leq \delta_j, \ \ 1 \leq j \leq k,$$

are included in the formulation of (P) by choosing $G = (G_1, \ldots, G_k)^T$, with

$$G_j(y) = \int_\Omega g_j(x, t, y(x, t))dxdt,$$

$Z = \mathbb{R}^k$, and $C = (-\infty, \delta_1] \times \cdots \times (-\infty, \delta_k]$.

Example 13 Integral constraints on the gradient of the state can be considered within our formulation of problem (P):

$$G(y_u) = \int_0^T \int_\Omega |\nabla_x y_u(x, t)|^2 dxdt \leq \delta.$$

In this case we can take $Z = \mathbb{R}$ and $C = (-\infty, \delta]$.

We will not consider state constrained problems in the present work. The interested reader may consult, e.g., the papers [3, 6, 7, 17].

5 First Order Optimality Conditions

The first order optimality conditions are necessary conditions for local optimality. In the case of convex problems, they become also sufficient for global optimality. In absence of convexity, the sufficiency requires the use of second order optimality conditions, which will be the goal of the next section. From the first order necessary conditions we can deduce some properties of the optimal control as we will prove later. Before proving the first order optimality conditions let us recall the meaning of a local minimum.

Definition 14 We will say that $\bar{u}$ is a local minimum of (P) in the $L^p(\Omega)$ sense, $1 \leq p \leq +\infty$, if there exists a ball $B_\varepsilon(\bar{u}) \subset L^p(\Omega)$ such that $J(\bar{u}) \leq J(u)\ \forall u \in \mathbb{K} \cap B_\varepsilon(\bar{u})$. The element $\bar{u}$ will be said a strict local minimum if the inequality $J(\bar{u}) < J(u)$ holds $\forall u \in \mathbb{K} \cap B_\varepsilon(\bar{u})$ with $\bar{u} \neq u$.

Since $\mathbb{K}$ is a bounded subset of $L^\infty(\Omega)$, if $\bar{u}$ is a (strict) local minimum of (P) in the $L^p(\Omega)$ sense, for some $1 \leq p < +\infty$, then $\bar{u}$ is a (strict) local minimum of (P) in the $L^q(Q)$ sense for every $q \in [1, +\infty]$: if $q > p$, this follows directly from the fact that $L^q(\Omega) \hookrightarrow L^p(\Omega)$; if $q < p$ we use the boundedness of $\mathbb{K}$ to get the inequality $|u(x) - \bar{u}(x)|^p \leq |u(x) - \bar{u}(x)|^q |\beta - \alpha|^{p-q}$ for a.e. $x \in \Omega$ to deduce the result. However, if $\bar{u}$ is a local minimum in the $L^\infty(\Omega)$ sense, it is not necessarily a local minimum in the $L^p(\Omega)$ sense for any $p \in [1, +\infty)$. In the sequel, if nothing is precised, when we say that $\bar{u}$ is a local minimum of (P), it should be intended in the $L^p(\Omega)$ sense for some $p \in [1, +\infty]$.

The key tool to get the first order optimality conditions is provided by the next lemma.

Lemma 15 *Let U be a Banach space, $\mathbb{K} \subset U$ a convex subset and $J : U \longrightarrow \mathbb{R}$ a function. Let us assume that $\bar{u}$ is a local solution of the optimization problem*

$$(\mathrm{P}) \begin{cases} \min\ J(u) \\ u \in \mathbb{K} \end{cases}$$

and that J has directional derivatives at $\bar{u}$. Then

$$J'(\bar{u}) \cdot (u - \bar{u}) \geq 0 \quad \forall u \in \mathbb{K}. \tag{5}$$

Conversely, if J is a convex function and $\bar{u}$ is an element of $\mathbb{K}$ satisfying (5), then $\bar{u}$ is a global minimum of (P).

Proof The inequality (5) is easy to get

$$J'(\bar{u}) \cdot (u - \bar{u}) = \lim_{\lambda \searrow 0} \frac{J(\bar{u} + \lambda(u - \bar{u})) - J(\bar{u})}{\lambda} \geq 0.$$

The last inequality follows from the local optimality of $\bar{u}$ and the fact that $\bar{u} + \lambda(u - \bar{u}) \in \mathbb{K}$ for every $u \in \mathbb{K}$ and every $\lambda \in [0, 1]$ due to the convexity of $\mathbb{K}$.

Conversely, if $\bar{u} \in \mathbb{K}$ fulfills (5) and J is convex, then for every $u \in \mathbb{K}$

$$0 \leq J'(\bar{u}) \cdot (u - \bar{u}) = \lim_{\lambda \searrow 0} \frac{J(\bar{u} + \lambda(u - \bar{u})) - J(\bar{u})}{\lambda} \leq J(u) - J(\bar{u}).$$

Therefore $\bar{u}$ is a global solution of (P). □

In order to apply this lemma to the study of problem (P) we need to analyze the differentiability of the functionals involved in the control problem. To this end, taking into account the $C^{1,1}$ regularity of Γ and the regularity result of Theorem 4, we obtain the following result.

Proposition 16 *Let $\frac{n}{2} < p < +\infty$. The mapping $G : L^p(\Omega) \longrightarrow W^{2,p}(\Omega)$ defined by $G(u) = y_u$ is of class C^2. Furthermore if $u, v \in L^p(\Omega)$ and $z = DG(u) \cdot v$, then z is the unique solution in $W^{2,p}(\Omega)$ of the Dirichlet problem*

$$\begin{cases} Az + \phi'(y_u(x))z = v \ \text{ in } \Omega, \\ z = 0 \ \text{ on } \Gamma. \end{cases} \tag{6}$$

Finally, for every $v_1, v_2 \in L^p(\Omega)$, $z_{v_1 v_2} = G''(u)(v_1, v_2)$ is the solution of

$$\begin{cases} Az_{v_1 v_2} + \phi'(y_u(x))z_{v_1 v_2} + \phi''(y_u(x))z_{v_1} z_{v_2} = 0 \ \text{ in } \Omega, \\ z_{v_1 v_2} = 0 \ \text{ on } \Gamma, \end{cases} \tag{7}$$

where $z_{v_i} = G'(u)v_i$, $i = 1, 2$.

Proof To prove the differentiability of G, we apply the implicit function theorem. Let us consider the Banach space $V(\Omega) = W^{2,p}(\Omega) \cap W_0^{1,p}(\Omega)$. Let us mention that $V(\Omega) \subset C(\bar{\Omega})$ with a continuous embedding. Indeed, since $p > \frac{n}{2}$ the continuous embedding $W^{2,p}(\Omega) \subset C(\bar{\Omega})$ follows. Now let us take the function

$$F : V(\Omega) \times L^p(\Omega) \longrightarrow L^p(\Omega)$$

defined by

$$F(y, u) = Ay + \phi(y) - u.$$

It is obvious that F is of class C^2, $y_u \in V(\Omega)$ for every $u \in L^p(\Omega)$, $F(y_u, u) = 0$ and

$$\frac{\partial F}{\partial y}(y, u) \cdot z = Az + \phi'(y)z$$

is an isomorphism from $V(\Omega)$ into $L^p(\Omega)$. By applying the implicit function theorem we deduce that G is of class C^2 and $DG(u) \cdot z$ is given by (6). Finally (7) follows by differentiating twice with respect to u in the equation

$$AG(u) + \phi(G(u)) = u. \qquad \square$$

For every $u \in L^\infty(\Omega)$, we define its related adjoint state $\varphi_u \in W^{2,\bar{p}}(\Omega)$, as the unique solution of the problem

$$\begin{cases} A^*\varphi + \phi'(y_u)\varphi = \dfrac{\partial L}{\partial y}(x, y_u, u) \ \text{ in } \Omega \\ \varphi = 0 \ \text{ on } \Gamma, \end{cases} \tag{8}$$

A^* being the adjoint operator of A and $\bar{p} > n$, the exponent introduced in Assumption **(H1)**. As a consequence of the previous result we get the following proposition.

Proposition 17 *The function $J : L^\infty(\Omega) \to \mathbb{R}$ is of class C^2. Moreover, for every $u, v, v_1, v_2 \in L^\infty(\Omega)$*

$$J'(u)v = \int_\Omega \left(\frac{\partial L}{\partial u}(x, y_u, u) + \varphi_u \right) v \, dx \tag{9}$$

and

$$J''(u)v_1v_2 = \int_\Omega \left[\frac{\partial^2 L}{\partial y^2}(x, y_u, u) z_{v_1} z_{v_2} + \frac{\partial^2 L}{\partial y \partial u}(x, y_u, u)(z_{v_1} v_2 + z_{v_2} v_1) + \right.$$

$$\left. \frac{\partial^2 L}{\partial u^2}(x, y_u, u) v_1 v_2 - \varphi_u \phi''(y_u) z_{v_1} z_{v_2} \right] dx \tag{10}$$

where $z_{v_i} = G'(u)v_i$, $i = 1, 2$.

Proof From hypothesis **(H1)**, Proposition 16 and the chain rule, we deduce that J is of class C^2 and we have

$$J'(u)\cdot v = \int_\Omega \left[\frac{\partial L}{\partial y}(x, y_u(x), u(x))z(x) + \frac{\partial L}{\partial u}(x, y_u(x), u(x))v(x)\right] dx,$$

where $z = G'(u)v$. Using (8) and integrating by parts in this expression, we get

$$\begin{aligned} J'(u)\cdot v &= \int_\Omega \left\{[A^*\varphi_u + \phi'(y_u)\varphi_u]z + \frac{\partial L}{\partial u}(x, y_u(x), u(x))v(x)\right\} dx \\ &= \int_\Omega \left\{[Az + \phi'(y_u)z]\varphi_u + \frac{\partial L}{\partial u}(x, y_u(x), u(x))v(x)\right\} dx \\ &= \int_\Omega \left\{\varphi_u(x) + \frac{\partial L}{\partial u}(x, y_u(x), u(x))\right\} v(x)\, dx, \end{aligned}$$

which proves (9). Finally, (10) follows again by application of the chain rule and Proposition 16. □

Remark 18 Let us note that for any $u \in L^\infty(\Omega)$, the continuous linear and bilinear forms $J'(u) : L^\infty(\Omega) \longrightarrow \mathbb{R}$ and $J''(u) : L^\infty(\Omega) \times L^\infty(\Omega) \longrightarrow \mathbb{R}$ can be readily extended to continuous linear and bilinear forms $J'(u) : L^2(\Omega) \longrightarrow \mathbb{R}$ and $J''(u) : L^2(\Omega) \times L^2(\Omega) \longrightarrow \mathbb{R}$. Indeed, it is enough to use the expressions given by (9) and (10). In addition, we have the following continuity property: if $\{u_k\}_{k=1}^\infty$ is a bounded sequence in $L^\infty(\Omega)$ converging to u in $L^2(\Omega)$, then

$$\lim_{k\to\infty} \|J''(u_k) - J''(u)\| = 0.$$

Here $\|\cdot\|$ denotes the norm in the space of bilinear and continuous forms in $L^2(\Omega)$. To check this identity, we first observe that Proposition 16 implies that $\|G(u_k) - G(u)\|_{L^\infty(\Omega)} \to 0$ and $\|G'(u_k) - G'(u)\| \to 0$, where $\|\cdot\|$ denotes the norm in $\mathscr{L}(L^2(\Omega), H^2(\Omega))$. Moreover, an application of Lebesgue theorem and Assumption **(H1)** leads easily to the above convergence.

Combining Lemma 15 with the previous proposition, we get the first order optimality conditions.

Theorem 19 *Let $\bar u$ be a local minimum of (P). Then there exist $\bar y, \bar\varphi \in H_0^1(\Omega) \cap W^{2,\bar p}(\Omega)$ such that the following relations hold*

$$\begin{cases} A\bar y + \phi(\bar y) = \bar u \ \text{ in } \Omega, \\ \bar y = 0 \ \text{ on } \Gamma, \end{cases} \tag{11}$$

$$\begin{cases} A^*\bar\varphi + \phi'(\bar y)\bar\varphi = \dfrac{\partial L}{\partial y}(x, \bar y, \bar u) \ \text{ in } \Omega, \\ \bar\varphi = 0 \ \text{ on } \Gamma, \end{cases} \tag{12}$$

$$\int_\Omega \left\{ \bar{\varphi}(x) + \frac{\partial L}{\partial u}(x, \bar{y}(x), \bar{u}(x)) \right\} (u(x) - \bar{u}(x))dx \geq 0 \ \forall u \in \mathbb{K}. \tag{13}$$

From this theorem, we can deduce some regularity results of the local minima.

Theorem 20 *Let us assume that $\bar{u}$ is a local minimum of* (P) *and that the hypotheses* ***(H1)*** *and* ***(H2)*** *are fulfilled. Then for any $x \in \bar{\Omega}$, the equation*

$$\bar{\varphi}(x) + \frac{\partial L}{\partial u}(x, \bar{y}(x), t) = 0 \tag{14}$$

has a unique solution $t = \bar{s}(x)$, where $\bar{y}$ is the state associated to $\bar{u}$ and $\bar{\varphi}$ is the adjoint state defined by (12)*. The mapping $\bar{s} : \bar{\Omega} \longrightarrow \mathbb{R}$ is Lipschitz. Moreover $\bar{u}$ and $\bar{s}$ are related by the formula*

$$\bar{u}(x) = \mathrm{Proj}_{[\alpha,\beta]}(\bar{s}(x)) = \max(\alpha, \min(\beta, \bar{s}(x))), \tag{15}$$

and $\bar{u}$ is Lipschitz too.

Proof The existence and uniqueness of solution of Eq. (14) is an immediate consequence of the hypothesis **(H2)**, therefore $\bar{s}$ is well defined. Let us see that $\bar{s}$ is bounded. Indeed, by the mean value theorem and the identity

$$\bar{\varphi}(x) + \frac{\partial L}{\partial u}(x, \bar{y}(x), \bar{s}(x)) = 0,$$

we get that

$$\frac{\partial^2 L}{\partial u^2}(x, \bar{y}(x), \theta(x)\bar{s}(x))\bar{s}(x) = -\bar{\varphi}(x) - \frac{\partial L}{\partial u}(x, \bar{y}(x), 0)$$

for some measurable function $0 \leq \theta(x) \leq 1$. This relationship, **(H2)**, and **(H1)** lead to

$$\Lambda|\bar{s}(x)| \leq |\bar{\varphi}(x)| + \left| \frac{\partial L}{\partial u}(x, \bar{y}(x), 0) \right| \leq C \qquad \forall x \in \Omega.$$

Now let us prove that $\bar{s}$ is Lipschitz. To do this, we use **(H2)**, the properties of L enounced in **(H1)**, the fact that $\bar{y}$ and $\bar{\varphi}$ are Lipschitz functions (due to the inclusion $W^{2,\bar{p}}(\Omega) \subset C^1(\bar{\Omega})$ for $\bar{p} > n$) and the equation above satisfied by $\bar{s}(x)$.

Let $x_1, x_2 \in \bar{\Omega}$

$$\Lambda|\bar{s}(x_2) - \bar{s}(x_1)| \leq |\frac{\partial L}{\partial u}(x_2, \bar{y}(x_2), \bar{s}(x_2)) - \frac{\partial L}{\partial u}(x_2, \bar{y}(x_2), \bar{s}(x_1))| =$$
$$|\bar{\varphi}(x_1) - \bar{\varphi}(x_2) + \frac{\partial L}{\partial u}(x_1, \bar{y}(x_1), \bar{s}(x_1)) - \frac{\partial L}{\partial u}(x_2, \bar{y}(x_2), \bar{s}(x_1))| \leq$$
$$|\bar{\varphi}(x_1) - \bar{\varphi}(x_2)| + C_M\,(|\bar{y}(x_1) - \bar{y}(x_2)| + |x_2 - x_1|) \leq C|x_2 - x_1|.$$

Finally, from (13) and the fact that $(\partial L/\partial u)$ is an increasing function of the third variable we have

$$\alpha < \bar{u}(x) < \beta \Rightarrow \bar{\varphi}(x) + \frac{\partial L}{\partial u}(x, \bar{y}(x), \bar{u}(x)) = 0 \Rightarrow \bar{u}(x) = \bar{s}(x),$$

$$\bar{u}(x) = \beta \quad \Rightarrow \bar{\varphi}(x) + \frac{\partial L}{\partial u}(x, \bar{y}(x), \bar{u}(x)) \leq 0 \Rightarrow \bar{u}(x) \leq \bar{s}(x),$$

$$\bar{u}(x) = \alpha \quad \Rightarrow \bar{\varphi}(x) + \frac{\partial L}{\partial u}(x, \bar{y}(x), \bar{u}(x)) \geq 0 \Rightarrow \bar{u}(x) \geq \bar{s}(x),$$

which implies (15). □

Remark 21 If the assumption **(H2)** does not hold, then the optimal controls can be discontinuous. The most obvious case is the one where L is independent of u. In this case (13) is reduced to

$$\int_\Omega \bar{\varphi}(x)(u(x) - \bar{u}(x))\,dx \geq 0 \quad \forall u \in \mathbb{K},$$

which leads to

$$\bar{u}(x) = \begin{cases} \alpha & \text{if } \bar{\varphi}(x) > 0 \\ \beta & \text{if } \bar{\varphi}(x) < 0 \end{cases} \quad \text{a.e. } x \in \Omega.$$

If $\bar{\varphi}$ vanishes only in a set of points of zero Lebesgue measure, then $\bar{u}$ jumps from α to β. Such a control $\bar{u}$ is called a bang-bang control. The controls of this nature are of great interest in the applications because of the ease to automate the control process. All the results presented previously are valid without the assumption **(H2)**, except Theorem 20.

Remark 22 If we consider the tracking cost functional $L(x, y, u) = [(y - y_d(x))^2 + Nu^2]/2$ with $N > 0$ and $y_d \in L^2(\Omega)$, then (14) leads to $\bar{s} = -\bar{\varphi}/N$, and (15) implies

$$\bar{u}(x) = \mathrm{Proj}_{\mathbb{K}}\left(-\frac{1}{N}\bar{\varphi}\right)(x) = \begin{cases} \alpha & \text{if } -\dfrac{1}{N}\bar{\varphi}(x) < \alpha, \\ \beta & \text{if } -\dfrac{1}{N}\bar{\varphi}(x) > \beta, \\ -\dfrac{1}{N}\bar{\varphi}(x) & \text{if } \alpha \le -\dfrac{1}{N}\bar{\varphi}(x) \le \beta. \end{cases}$$

If, furthermore, we assume that $\mathbb{K} = L^2(\Omega)$, then (13) implies that $\bar{u} = -(1/N)\bar{\varphi}$. Thus $\bar{u}$ has the same regularity than $\bar{\varphi}$. Therefore, $\bar{u}$ will be the more regular as much as greater be the regularity of y_d, Γ, ϕ and the coefficients of operator A. In particular, we can get $C^\infty(\bar{\Omega})$-regularity, if all the data of the problem are of class C^∞.

6 Second Order Optimality Conditions

The material contained in this section is based on the paper by Casas and Tröltzsch [15].

Let $\bar{u} \in \mathbb{K}$ satisfy the first order optimality conditions (11)–(13) along with $\bar{y}$ and $\bar{\varphi}$. In order to simplify the notation we will consider the function

$$\bar{d}(x) = \frac{\partial L}{\partial u}(x, \bar{y}(x), \bar{u}(x)) + \bar{\varphi}(x).$$

From (13) it follows

$$\bar{d}(x) \begin{cases} 0 & \text{a.e. } x \in \Omega \text{ if } \alpha < \bar{u}(x) < \beta, \\ \ge 0 & \text{a.e. } x \in \Omega \text{ if } \bar{u}(x) = \alpha, \\ \le 0 & \text{a.e. } x \in \Omega \text{ if } \bar{u}(x) = \beta. \end{cases} \tag{16}$$

The following cone of critical directions is essential in the formulation of the second order optimality conditions:

$$C_{\bar{u}} = \{v \in L^2(\Omega) \text{ satisfying (17) and } v(x) = 0 \text{ if } \bar{d}(x) \neq 0\},$$

$$v(x) \begin{cases} \ge 0 & \text{a.e. } x \in \Omega \text{ if } \bar{u}(x) = \alpha, \\ \le 0 & \text{a.e. } x \in \Omega \text{ if } \bar{u}(x) = \beta. \end{cases} \tag{17}$$

Now we can formulate the necessary and sufficient conditions for optimality.

Theorem 23 *Under the hypothesis **(H1)**, if $\bar{u}$ is a local minimum of (P), then*

$$J''(\bar{u})v^2 \geq 0 \quad \forall v \in C_{\bar{u}}. \tag{18}$$

*Conversely, if additionally **(H2)** holds and $\bar{u} \in \mathbb{K}$ fulfills the first order optimality conditions (11)–(13) along with the condition*

$$J''(\bar{u})v^2 > 0 \quad \forall v \in C_{\bar{u}} \setminus \{0\}, \tag{19}$$

then there exist $\kappa > 0$ and $\varepsilon > 0$ such that

$$J(u) \geq J(\bar{u}) + \frac{\kappa}{2}\|u - \bar{u}\|^2_{L^2(\Omega)} \quad \forall u \in \mathbb{K} \cap \bar{B}_\varepsilon(\bar{u}), \tag{20}$$

where $\bar{B}_\varepsilon(\bar{u})$ is the closed ball in $L^2(\Omega)$ with center at $\bar{u}$ and radius ε.

Proof

i) Given $v \in C_{\bar{u}}$ we define for every $k \in \mathbb{N}$

$$v_k(x) = \begin{cases} 0 \ \text{ if } \alpha < \bar{u}(x) < \alpha + \dfrac{1}{k} \ \text{ or } \ \beta - \dfrac{1}{k} < \bar{u}(x) < \beta, \\ \text{Proj}_{[-k,+k]}(v(x)) \ \text{ otherwise.} \end{cases}$$

Then we have that $v_k \in C_{\bar{u}} \cap L^\infty(\Omega)$. Moreover $v_k \to v$ in $L^2(\Omega)$ when $k \to \infty$, and $\bar{u} + \rho v_k \in \mathbb{K}$ holds for every $\rho \in (0, \frac{1}{k^2}]$. By using the local optimality of $\bar{u}$ and taking ρ small enough we obtain

$$0 \leq \frac{J(\bar{u} + \rho v_k) - J(\bar{u})}{\rho} = J'(\bar{u})v_k + \frac{\rho}{2}J''(\bar{u} + \theta_k \rho v_k)v_k^2,$$

with $0 < \theta_k < 1$. From this inequality and the identity

$$J'(\bar{u})v_k = \int_\Omega \left(\frac{\partial L}{\partial u}(x, \bar{y}(x), \bar{u}(x)) + \bar{\varphi}(x)\right)v_k(x)\,dx = \int_\Omega \bar{d}(x)v_k(x)\,dx = 0,$$

we deduce by passing to the limit as $\rho \to 0$

$$0 \leq J''(\bar{u} + \theta_k \rho v_k)v_k^2 \to J''(\bar{u})v_k^2.$$

By the expression of the second derivative J'' given by (10), we can pass to the limit in the previous expression when $k \to \infty$ and get that $J''(\bar{u})v^2 \geq 0$.

ii) Now let us assume that (19) holds and prove (20). We argue by contradiction and assume that for any $k \in \mathbb{N}$ we can find an element $u_k \in \mathbb{K}$ such that

$$\|\bar{u} - u_k\|_{L^2(\Omega)} < \frac{1}{k} \ \text{ and } \ J(u_k) < J(\bar{u}) + \frac{1}{2k}\|u_k - \bar{u}\|^2_{L^2(\Omega)}. \tag{21}$$

Let us define

$$\rho_k = \|u_k - \bar{u}\|_{L^2(\Omega)} \text{ and } v_k = \frac{1}{\rho_k}(u_k - \bar{u}).$$

By taking a subsequence if necessary, we can suppose that $v_k \rightharpoonup v$ weakly in $L^2(\Omega)$. The proof is split into three steps.

Step 1 $v \in C_{\bar{u}}$. First we observe that each element v_k obviously satisfies (17). Since the set of elements satisfying (17) is closed and convex in $L^2(\Omega)$, hence weakly closed, we deduce that v satisfies (17) as well. Let us prove that v vanishes at almost all points $x \in \Omega$ where $\bar{d}(x) \neq 0$. From (13) and the fact that $u_k \in \mathbb{K}$, we infer

$$\int_\Omega \bar{d}(x)v(x)\,dx = \lim_{k\to\infty} \frac{1}{\rho_k} \int_\Omega \bar{d}(x)(u_k - \bar{u})\,dx \geq 0. \tag{22}$$

Now, using (21) and the mean value theorem we get

$$J'(\bar{u} + \theta_k(u_k - \bar{u}))(u_k - \bar{u}) < \frac{1}{2k}\|u_k - \bar{u}\|^2_{L^2(\Omega)} = \frac{\rho_k^2}{2k}.$$

Dividing this inequality by ρ_k it follows

$$J'(\bar{u} + \theta_k(u_k - \bar{u}))v_k < \frac{\rho_k}{2k} \to 0 \text{ as } k \to \infty. \tag{23}$$

Let us write $\hat{u}_k = \bar{u} + \theta_k(u_k - \bar{u})$, and let $\hat{y}_k$ and $\hat{\varphi}_k$ be the associated state and adjoint state. Applying Proposition 16 we get that $\hat{y}_k = G(\hat{u}_k) \to G(\bar{u}) = \bar{y}$ in $W^{2,\bar{p}}(\Omega) \subset C(\bar{\Omega})$. From the assumption **(H1)** we infer that

$$\frac{\partial L}{\partial y}(x, \hat{y}_k, \hat{u}_k) \to \frac{\partial L}{\partial y}(x, \bar{y}, \bar{u}) \text{ strongly in } L^{\bar{p}}(\Omega) \text{ as } k \to \infty.$$

Then, from the equation satisfied by $\hat{\varphi}_k$ and the above convergences we deduce that $\hat{\varphi}_k \to \bar{\varphi}$ strongly in $W^{2,\bar{p}}(\Omega) \subset C(\bar{\Omega})$. Hence, we can pass to the limit in (23) and deduce

$$\int_\Omega \bar{d}(x)v(x)\,dx = \lim_{k\to\infty} \int_\Omega \Big(\hat{\varphi}_k + \frac{\partial L}{\partial u}(x, \hat{y}_k, \hat{u}_k)\Big)v_k\,dx = \lim_{k\to\infty} J'(\bar{u} + \theta_k(u_k - \bar{u}))v_k \leq 0.$$

This inequality, along with (22), implies that

$$\int_\Omega \bar{d}(x)v(x)\,dx = 0.$$

Finally, due to the sign conditions (17) satisfied by v, we conclude that

$$\int_\Omega |\bar{d}(x)||v(x)|\,dx = \int_\Omega \bar{d}(x)v(x)\,dx = 0,$$

hence $v \in C_{\bar{u}}$.

Step 2 $v = 0$. Performing a Taylor expansion, we get

$$J(u_k) = J(\bar{u} + \rho_k v_k) = J(\bar{u}) + \rho_k J'(\bar{u})v_k + \frac{\rho_k^2}{2}J''(\hat{u}_k)v_k^2,$$

where $\hat{u}_k = \bar{u} + \vartheta_k(u_k - \bar{u})$ with $0 \le \vartheta_k(x) \le 1$. Since $\bar{u}$ satisfies the first order optimality conditions, we have that $J'(\bar{u})v_k \ge 0$. Using this fact in the above identity and (21), we obtain

$$\frac{\rho_k^2}{2}J''(\hat{u}_k)v_k^2 \le J(u_k) - J(\bar{u}) < \frac{\rho_k^2}{2k},$$

therefore $J''(\hat{u}_k)v_k^2 < 1/k \to 0$ as $k \to \infty$. In the next lines we will prove that $J''(\bar{u})v^2 \le \liminf_{k\to\infty} J''(\hat{u}_k)v_k^2 \le 0$. This inequality, the fact that $v \in C_{\bar{u}}$, and (19) imply that $v = 0$.

With the same notation as above and arguing in the same way, we get that $(\hat{y}_k, \hat{\varphi}_k, \hat{u}_k) \to (\bar{y}, \bar{\varphi}, \bar{u})$ strongly in $L^{\bar{p}}(\Omega)$ and $(\hat{y}_k, \hat{\varphi}_k) \to (\bar{y}, \bar{\varphi})$ strongly in $C(\bar{\Omega})$ as $k \to \infty$. Moreover, from Proposition 16 we have that $\hat{z}_{v_k} = DG(\hat{u}_k)v_k \rightharpoonup DG(\bar{u})v = z_v$ weakly in $W^{2,p}(\Omega)$ for every $p < +\infty$, hence $\hat{z}_{v_k} \to z_v$ strongly in $C(\bar{\Omega})$. Now, from (10) we obtain

$$\begin{aligned} J''(\hat{u}_k)v_k^2 = &\int_\Omega \Big(\Big[\frac{\partial^2 L}{\partial y^2}(x, \hat{y}_k, \hat{u}_k) - \hat{\varphi}_k \phi''(\hat{y}_k) \Big] \hat{z}_{v_k}^2 + 2\frac{\partial^2 L}{\partial y \partial u}(x, \hat{y}_k, \hat{u}_k) v_k \hat{z}_{v_k} \Big)\, dx \\ &+ \int_\Omega \frac{\partial^2 L}{\partial u^2}(x, \hat{y}_k, \hat{u}_k) v_k^2\, dx. \end{aligned} \tag{24}$$

From the convergence properties established for $(\hat{y}_k, \hat{\varphi}_k, \hat{u}_k, \hat{z}_{v_k})$ it is easy to pass to the limit in the first integral towards the corresponding terms of $J''(\bar{u})v^2$; see Remark 18. To deal with the last integral, we use Lemma 24 below to deduce that

$$\int_\Omega \frac{\partial^2 L}{\partial u^2}(x, \bar{y}, \bar{u})v^2\, dx \le \liminf_{k\to\infty} \int_\Omega \frac{\partial^2 L}{\partial u^2}(x, \hat{y}_k, \hat{u}_k)v_k^2\, dx,$$

which concludes the proof of the step 2.

Step 3 *Contradiction.* Since $v = 0$ we have that $\hat{z}_{v_k} \to 0$ in $C(\bar{\Omega})$. Then, the first integral in (24) converges to 0. Now, using the assumption **(H2)** and that

$\|v_k\|_{L^2(\Omega)} = 1$, we infer from (24)

$$\Lambda = \Lambda \int_\Omega v_k^2\,dx \le \liminf_{k\to\infty} \int_\Omega \frac{\partial^2 L}{\partial u^2}(x, \hat{y}_k, \hat{u}_k) v_k^2\,dx = \liminf_{k\to\infty} J''(\hat{u}_k) v_k^2 \le 0,$$

which contradicts the fact that $\Lambda > 0$. □

Lemma 24 ([15]) *Let (X, Σ, μ) be a measure space with $\mu(X) < +\infty$. Suppose that $\{g_k\}_{k=1}^\infty \subset L^\infty(X)$ and $\{v_k\}_{k=1}^\infty \subset L^2(X)$ satisfy the assumptions*

- *$g_k \ge 0$ a.e. in X, $\{g_k\}_{k=1}^\infty$ is bounded in $L^\infty(X)$ and $g_k \to g$ in $L^1(X)$ as $k \to \infty$.*
- *$v_k \rightharpoonup v$ in $L^2(X)$ as $k \to \infty$.*

Then there holds the inequality

$$\int_X g(x) v^2(x)\,d\mu(x) \le \liminf_{k\to\infty} \int_X g_k(x) v_k^2(x)\,d\mu(x). \tag{25}$$

Proof Since $\{g_k\}_{k=1}^\infty$ is bounded in $L^\infty(X)$, it holds $g \in L^\infty(X)$. Denote the lower limit in (25) by λ. Then there exists a subsequence of functions, denoted in the same way, such that the integrals of the right hand side of (25) converge to λ. Again, we can select a new subsequence of this one such that $g_k(x) \to g(x)$ *a.e.* in X. Let $\varepsilon > 0$ be arbitrary. By Egorov's theorem, there exists a measurable set $K_\varepsilon \subset X$ such that $\mu(X \setminus K_\varepsilon) < \varepsilon$ and $\|g - g_k\|_{L^\infty(K_\varepsilon)} \to 0$ as $k \to \infty$. Then we have

$$\begin{aligned}
&\liminf_{k\to\infty} \int_X g_k(x) v_k^2(x)\,d\mu(x) \ge \liminf_{k\to\infty} \int_{K_\varepsilon} g_k(x) v_k^2(x)\,d\mu(x) \\
&\ge \liminf_{k\to\infty} \int_{K_\varepsilon} [g_k(x) - g(x)] v_k^2(x)\,d\mu(x) + \liminf_{k\to\infty} \int_{K_\varepsilon} g(x) v_k^2(x)\,d\mu(x) \\
&= \liminf_{k\to\infty} \int_{K_\varepsilon} g(x) v_k^2(x)\,d\mu(x) \ge \int_{K_\varepsilon} g(x) v^2(x)\,d\mu(x).
\end{aligned}$$

Finally, passing to the limit as $\varepsilon \to 0$ we get (25) □

We will finish this section by proving an interesting result that simplifies the proof of the error estimates of discrete approximations of problem (P).

Theorem 25 *Under the hypotheses **(H1)** and **(H2)**, if $\bar{u} \in \mathbb{K}$ satisfies (11)–(13), the following statements are equivalent:*

$$J''(\bar{u}) v^2 > 0 \quad \forall v \in C_{\bar{u}} \setminus \{0\} \tag{26}$$

and

$$\exists \delta > 0 \text{ and } \exists \tau > 0 : J''(\bar{u}) v^2 \ge \delta \|v\|_{L^2(\Omega)}^2 \ \forall v \in C_{\bar{u}}^\tau, \tag{27}$$

where

$$C_{\bar u}^{\tau} = \{v \in L^2(\Omega) \text{ satisfying (17) and } v(x) = 0 \text{ if } |\bar d(x)| > \tau\}.$$

Proof Since $C_{\bar u} \subset C_{\bar u}^{\tau}$ for all $\tau > 0$, it is obvious that (27) implies (26). Let us prove the converse implication. We proceed by contradiction and assume that for any $\tau > 0$ there exists $v_\tau \in C_{\bar u}^{\tau}$ such that $J''(\bar u)v_\tau^2 < \tau \|v_\tau\|^2_{L^2(\Omega)}$. Dividing v_τ by its norm, and taking a subsequence if necessary, we can assume that

$$\|v_\tau\|_{L^2(\Omega)} = 1, \;\; J''(\bar u)v_\tau^2 < \tau \;\text{ and }\; v_\tau \rightharpoonup v \text{ in } L^2(\Omega). \tag{28}$$

Let us prove that $v \in C_{\bar u}$. Arguing as in the proof of the previous theorem, we get that v satisfies the sign condition (17). On the other hand,

$$\int_\Omega |\bar d(x)v(x)|\,dx = \int_\Omega \bar d(x)v(x)\,dx =$$
$$\lim_{\tau\to 0}\int_\Omega \bar d(x)v_\tau(x)\,dx = \lim_{\tau\to 0}\int_{|\bar d(x)|\le\tau} \bar d(x)v_\tau(x)\,dx \le$$
$$\lim_{\tau\to 0}\tau\int_\Omega |v_\tau(x)|\,dx \le \lim_{\tau\to 0}\tau\sqrt{|\Omega|}\|v_\tau\|_{L^2(\Omega)} = 0,$$

which proves that $v(x) = 0$ if $\bar d(x) \neq 0$. Thus we have that $v \in C_{\bar u}$. Then (26) implies that either $v = 0$ or $J''(\bar u)v^2 > 0$. But (28) leads to

$$J''(\bar u)v^2 \le \liminf_{\tau\to 0} J''(\bar u)v_\tau^2 \le \limsup_{\tau\to 0} J''(\bar u)v_\tau^2 \le 0.$$

Thus we conclude that $v = 0$ and $\lim_{k\to\infty} J''(\bar u)v_\tau^2 = 0$. Moreover, arguing as in the proof of the previous theorem, we deduce that $z_\tau \to 0$ strongly in $C(\bar\Omega)$, therefore

$$0 < \Lambda = \Lambda\int_\Omega v_\tau^2\,dx \le \liminf_{\tau\to 0}\int_\Omega \frac{\partial^2 L}{\partial u^2}(x,\bar y,\bar u)v_\tau^2\,dx = \lim_{\tau\to 0} J''(\bar u)v_\tau^2$$
$$-\lim_{\tau\to 0}\int_\Omega\left[\frac{\partial^2 L}{\partial y^2}(x,\bar y,\bar u)z_\tau^2 + \frac{\partial^2 L}{\partial y\partial u}(x,\bar y,\bar u)v_\tau z_\tau - \bar\varphi\phi''(\bar y)z_\tau^2\right]dx = 0,$$

which leads to the desired contradiction. □

Remark 26 Some comments are necessary to clarify the results stated in the Theorems 23 and 25. If J is a functional in $\mathbb{R}^n$, we know that the first order condition $J'(\bar u) = 0$, together with the second order condition $J''(\bar u)v^2 > 0$ for every $v \in \mathbb{R}^n \setminus \{0\}$, implies that $\bar u$ is a strict local minimum of J. However, this is not true in infinite dimension. Let us confirm this by the following example.

Example 27 Consider the optimization problem

$$\min_{u\in L^\infty(0,1)} J(u) = \int_0^1 [tu^2(t) - u^3(t)]\,dt.$$

The function $\bar{u}(t) \equiv 0$ satisfies the first order necessary condition $J'(\bar{u}) = 0$ and

$$J''(\bar{u})v^2 = \int_0^1 2tv^2(t)\,dt > 0 \ \ \forall v \in L^\infty(0,1) \setminus \{0\}.$$

However, $\bar{u}$ is not a local minimum of J. Indeed, if we define

$$u_k(t) = \begin{cases} 2t \text{ if } t \in (0, \dfrac{1}{k}), \\ 0 \ \text{ otherwise,} \end{cases}$$

then it holds $J(u_k) = -\frac{1}{k^4} < J(\bar{u})$, and $\|u_k - \bar{u}\|_{L^\infty(0,1)} = \frac{2}{k}$, which shows that $\bar{u}$ is not a local minimum of the optimization problem.

The classical theorem for optimization in infinite dimensional spaces state that the second order sufficient condition requires the existence of some $\delta > 0$ such that $J''(\bar{u})v^2 \geq \delta\|v\|^2$. In finite dimension this condition is equivalent to $J''(\bar{u})v^2 > 0$ if $v \neq 0$, but this equivalence is not true, in general, in infinite dimension. However, Theorem 25 proves the equivalence of both conditions for our control problem. The main reasons for this equivalence are the following: on the one hand, the compactness of the relation control-to-state. On the other, the strict convexity of L with respect to u. Indeed, the fact that $\Lambda > 0$ played a crucial role in the proof. The situation is even more complicated for infinite dimensional control problems with constraints. Indeed, the following example by Dunn [21] demonstrates that $J''(\bar{u})v^2 \geq \delta\|v\|^2_{L^2(\Omega)}$ for every $v \in C_{\bar{u}}$ is not in general sufficient for local optimality.

Example 28 We define $J : L^2(0,1) \to \mathbb{R}$ by

$$J(u) = \int_0^1 [2a(x)u(x) - \text{sign}(a(x))u(x)^2]\,dx,$$

where $a(x) = 1 - 2x$. The set of admissible functions u is defined by

$$\mathbb{K} := \{u \in L^\infty(0,1) : \ 0 \leq u(x) \leq 2 \text{ for a.a. } x \in [0,1]\},$$

and the optimization problem is

$$\min_{u\in\mathbb{K}} J(u).$$

Let us set $\bar{u}(x) = \max\{0, -a(x)\}$; then $\bar{u}(x) = 0$ holds on $[0, 1/2]$ and $0 < \bar{u}(x) < 2$ on $(1/2, 2)$. We have

$$J'(\bar{u})v = \int_0^1 2[a(x) - \operatorname{sign}(a(x))\bar{u}(x)]v(x)\,dx = \int_0^1 \bar{d}(x)v(x)dx$$
$$= \int_0^{1/2} 2a(x)v(x)dx \geq 0$$

for all $v \in L^2(0,1)$ with $v(x) \geq 0$ on $[0, 1/2]$. Since $u - \bar{u}$ is nonnegative for all $u \in \mathbb{K}$, $\bar{u}$ satisfies the first order necessary optimality conditions.

In view of the sign conditions (17) and of $\bar{d}(x) > 0$ on $[0, 1/2)$, the critical cone for this example is

$$C_{\bar{u}} = \{v \in L^2(0,1) : v(x) = 0 \text{ on } [0, 1/2)\}.$$

For all $v \in C_{\bar{u}}$, we obtain

$$J''(\bar{u})v^2 = -\int_0^1 2\,\operatorname{sign}(a(x))\,v^2(x)\,dx = 2\int_{1/2}^1 v^2(x)\,dx - 2\int_0^{1/2} v^2(x)\,dx$$
$$= 2\int_{1/2}^1 v^2(x)\,dx = 2\,\|v\|^2_{L^2(0,1)}.$$

Therefore, $J''(\bar{u})v^2 \geq \delta\|v\|^2_{L^2(0,1)}$ $\forall v \in C_{\bar{u}}$ is fulfilled with $\delta = 2$. However, $\bar{u}$ is not a local minimum in $L^2(0,1)$. Indeed, take for $0 < \varepsilon < 1/2$

$$u_\varepsilon(x) = \begin{cases} 3\varepsilon, & \text{if } x \in [\frac{1}{2} - \varepsilon, \frac{1}{2}] \\ \bar{u}(x), & \text{else.} \end{cases}$$

Then we have

$$J(u_\varepsilon) - J(\bar{u}) = \int_{\frac{1}{2}-\varepsilon}^{\frac{1}{2}} [6\varepsilon(1 - 2x) - 9\varepsilon^2]dx = -3\varepsilon^3 < 0.$$

The second order condition $J''(\bar{u})v^2 \geq \delta\|v\|^2_{L^2(0,1)}$ must be assumed on an extended cone. Once again, in our control problem the fact that $\Lambda > 0$ implies that the condition on the extended cone $C^\tau_{\bar{u}}$ is equivalent to the condition (19).

Remark 29 (The Two-Norm Discrepancy)

There is another important issue when we look for second order sufficient conditions in infinite dimensional spaces. Let us consider the following example.

Example 30

$$\min_{u\in L^2(0,1)} J(u) = \int_0^1 \sin(u(t))\,dt.$$

It is obvious that $\bar{u}(t) \equiv -\pi/2$ is a solution. After some fast computations we get

$$J'(\bar{u})v = \int_0^1 \cos(\bar{u}(t))v(t)\,dt = 0 \ \ \forall v \in L^2(0,1)$$

and

$$J''(\bar{u})v^2 = -\int_0^1 \sin(\bar{u}(t))v^2(t)\,dt = \int_0^1 v^2(t)\,dt = \|v\|^2_{L^2(0,1)}.$$

Then we conclude that $\bar{u}$ is a strict local minimum. However, this is not true! Indeed, the functions

$$u_\varepsilon(t) = \begin{cases} -\dfrac{\pi}{2} & \text{if } t \in [0, 1-\varepsilon] \\ \\ +\dfrac{3\pi}{2} & \text{if } t \in (1-\varepsilon, 1] \end{cases}$$

satisfy that $J(\bar{u}) = J(u_\varepsilon)$ and nevertheless we have that $\|\bar{u} - u_\varepsilon\|_{L^2(0,1)} = 2\pi\sqrt{\varepsilon}$, which shows that $\bar{u}$ is not a strict local minimum. Therefore, something is wrong. What is it? The point is that J is not a C^2 function in $L^2(0,1)$. On the other hand, it is immediate to check that $\bar{u}$ is a strict local minimum of J in $L^\infty(0,1)$ and moreover J is of class C^2 in $L^\infty(0,1)$. However, an inequality of type

$$J''(\bar{u})v^2 \geq \delta\|v\|^2_{L^\infty(0,1)}$$

does not hold.

This fact is known as the *two-norm discrepancy*: the function is of class C^2 with respect to one norm, but the second order sufficient condition holds with respect to a different norm. This fact does never occur in finite dimension because all the norms are equivalent. To deal with this situation we can use the following abstract result.

Theorem 31 *Let U be a vector space endowed with two norms, $\|\cdot\|_\infty$ and $\|\cdot\|_2$, such that $J : (U, \|\cdot\|_\infty) \mapsto \mathbb{R}$ is of class C^2 in a $(U, \|\cdot\|_\infty)$-neighborhood $\mathscr{A} \subset U$ of $\bar{u}$ and assume that the following properties hold:*

$$J'(\bar{u}) = 0 \ \ \textit{and} \ \ \exists \delta > 0 \ \textit{such that} \ J''(\bar{u})v^2 \geq \delta\|v\|_2^2 \ \forall v \in U, \tag{29}$$

and there exists some $\varepsilon > 0$ *such that* $\bar{B}_\infty(\bar{u};\varepsilon) \subset \mathscr{A}$ *and*

$$|J''(\bar{u})v^2 - J''(u)v^2| \le \frac{\delta}{2}\|v\|_2^2 \ \forall v \in U \text{ if } \|u-\bar{u}\|_\infty \le \varepsilon. \tag{30}$$

Then there holds

$$\frac{\delta}{4}\|u-\bar{u}\|_2^2 + J(\bar{u}) \le J(u) \ \text{ if } \ \|u-\bar{u}\|_\infty \le \varepsilon \tag{31}$$

so that $\bar{u}$ *is strictly locally optimal with respect to the norm* $\|\cdot\|_\infty$.

In the above theorem $B_\infty(\bar{u};\varepsilon)$ denotes the ball of radius ε and centered at $\bar{u}$ with respect to the norm $\|\cdot\|_\infty$.

The proof of this theorem is quite elementary. To our knowledge, Ioffe [28] was the first who proved a result of this type by using two norms in the context of optimal control for ordinary differential equations.

Theorem 31 can be applied to Example 30 to deduce that $\bar{u}$ is a strict local minimum in the sense of $L^\infty(0,1)$.

Returning to the control problem (P), we observe that despite the two-norm discrepancy occurs, we get strict local optimality of $\bar{u}$ in $L^2(\Omega)$. The situation is completely different if we analyze the case where the assumption **(H2)** does not hold. For instance, if we consider the tracking-type control problem where the Tikhonov term does not appear: $N = 0$. It is not the aim of this paper to study such a case. We only show how the second order sufficient conditions can be formulated.

Theorem 32 *Let us assume that* $\bar{u} \in \mathbb{K}$ *satisfies the first order optimality conditions. We also suppose that there exist* $\delta > 0$ *and* $\tau > 0$ *such that*

$$J''(\bar{u})v^2 \ge \delta\|z_v\|^2_{L^2(\Omega)} \quad \forall v \in C^\tau_{\bar{u}},$$

where $z_v = G'(\bar{u})v$. *Then, there exist* $\varepsilon > 0$ *and* $\kappa > 0$ *such that*

$$J(\bar{u}) + \frac{\kappa}{2}\|y_u - \bar{y}\|^2_{L^2(\Omega)} \le J(u) \quad \forall u \in B_\varepsilon(\bar{u}) \cap \mathbb{K},$$

where $B_\varepsilon(\bar{u})$ *is the* $L^2(\Omega)$ *ball centered at* $\bar{u}$ *and radius* ε.

The proof of this theorem can be found in [9]. It is also proved in [9] that the inequality $J''(\bar{u})v^2 \ge \delta\|v\|^2_{L^2(\Omega)}$ for every $v \in C^\tau_{\bar{u}}$ is never fulfilled.

7 Numerical Approximation

The goal of this section is to prove error estimates for the numerical approximation for the control problem. In the last years, many papers have been devoted to this question; see [1, 2, 7, 8, 10, 12–14, 16, 20, 26, 37]. The reader is also referred to [34, 36, 39] and [35] for the case of control problems of parabolic equations.

In order to simplify the presentation we will assume from now on that Ω is convex.

We consider a finite element based approximation of (P). Associated with a parameter h we consider a family of triangulations $\{\mathscr{T}_h\}_{h>0}$ of $\bar{\Omega}$. To every element $T \in \mathscr{T}_h$ we assign two parameters $\rho(T)$ and $\sigma(T)$, where $\rho(T)$ denotes the diameter of T and $\sigma(T)$ is the diameter of the biggest ball contained in T. The size of the grid is given by $h = \max_{T\in\mathscr{T}_h} \rho(T)$. The following standard regularity assumptions on the triangulation are assumed.

1. There exist two positive constants ρ and σ such that

$$\frac{\rho(T)}{\sigma(T)} \le \sigma, \quad \frac{h}{\rho(T)} \le \rho$$

   for every $T \in \mathscr{T}_h$ and all $h > 0$.
2. Let us set $\bar{\Omega}_h = \cup_{T\in\mathscr{T}_h} T$, where Ω_h and Γ_h are the interior and the boundary of $\bar{\Omega}_h$ respectively. We assume that the vertices of $\mathscr{T}_h$ placed on the boundary Γ_h are points of Γ. We also assume

$$\exists C > 0 \text{ such that } |\Omega \setminus \Omega_h| \le Ch^2, \tag{32}$$

   where $|\cdot|$ denotes the Lebesgue measure. See [42, inequality (5.2.19)] for a proof of this inequality for two dimensional domains with a C^2 boundary.

   Associated to these triangulations we define the spaces

$$U_h = \{u \in L^\infty(\Omega_h) \mid u_{|T} \text{ is constant on each } T \in \mathscr{T}_h\},$$

$$Y_h = \{y_h \in C(\bar{\Omega}) \mid y_{h|T} \in \mathscr{P}_1, \text{ for every } T \in \mathscr{T}_h, \text{ and } y_h = 0 \text{ in } \bar{\Omega} \setminus \Omega_h\},$$

where $\mathscr{P}_1$ is the space formed by the polynomials of degree less than or equal to one. For every $u \in L^2(\Omega_h)$, we denote by $y_h(u)$ the unique element of Y_h satisfying

$$a(y_h(u), w_h) + \int_{\Omega_h} \phi(y_h(u)) w_h \, dx = \int_{\Omega_h} u w_h \, dx \quad \forall w_h \in Y_h, \tag{33}$$

where $a : Y_h \times Y_h \longrightarrow \mathbb{R}$ is the bilinear form defined by

$$a(y_h, w_h) = \int_{\Omega_h} \Big(\sum_{i,j=1}^{n} a_{ij}(x) \partial_{x_i} y_h(x) \partial_{x_j} w_h(x) + a_0(x) y_h(x) w_h(x)\Big) \, dx.$$

To prove the existence of a solution of (33) we truncate ϕ as in the proof of Theorem 4. Then we use Brouwer's fixed point theorem, and we take into account that all the norms are equivalent in a finite dimensional space. The uniqueness is an

immediate consequence of the monotonicity of ϕ and the coercivity of the elliptic operator A.

The set of discrete admissible controls is defined by

$$\mathbb{K}_h = \{u_h \in U_h : \alpha \leq u_{|T} \leq \beta \quad \forall T \in \mathcal{T}_h\}.$$

The finite dimensional approximation of the optimal control problem (P) is defined in the following way

$$(\mathrm{P}_h) \begin{cases} \min J_h(u_h) = \int_{\Omega_h} L(x, y_h(u_h)(x), u_h(x))\, dx, \\ u_h \in \mathbb{K}_h. \end{cases}$$

Let us start the study of problem (P_h) by analyzing the differentiability of the functions involved in the control problem. We just state the differentiability results analogous to the ones of Sect. 5 whose proof is an immediate consequence of the implicit function theorem. The change of the space $L^p(\Omega)$ considered in Proposition 16 by $L^1(\Omega)$ is possible due to the fact that $Y_h \subset C_0(\Omega)$.

Proposition 33 *For every $u \in L^1(\Omega_h)$, problem (33) has a unique solution $y_h(u) \in Y_h$. The mapping $G_h : L^1(\Omega_h) \longrightarrow Y_h$, defined by $G_h(u) = y_h(u)$, is of class C^2 and for all $v, u \in L^1(\Omega_h)$, $z_h(v) = G_h'(u)v$ is the solution of*

$$a(z_h(v), w_h) + \int_{\Omega_h} \phi'(y_h(u)) z_h(v) w_h\, dx = \int_{\Omega_h} v w_h\, dx \quad \forall w_h \in Y_h. \tag{34}$$

Finally, for every $v_1, v_2 \in L^1(\Omega_h)$, $z_h(v_1, v_2) = G_h''(u)(v_1, v_2) \in Y_h$ is the solution of the variational equation:

$$a(z_h, w_h) + \int_{\Omega_h} \phi'(y_h(u)) z_h w_h\, dx + \int_{\Omega_h} \phi''(y_h(u)) z_{h1} z_{h2} w_h dx = 0, \tag{35}$$

for all $w_h \in Y_h$, where $z_{hi} = G_h'(u)v_i$, $i = 1, 2$.

For every $u \in L^\infty(\Omega_h)$. we define its related discrete adjoint state $\varphi_h(u) \in Y_h$, as the unique solution of the problem

$$\begin{aligned} &a(w_h, \varphi_h(u)) + \int_{\Omega_h} \phi'(y_h(u)) \varphi_h(u) w_h\, dx = \\ &\int_{\Omega_h} \frac{\partial L}{\partial y}(x, y_h(u), u) w_h\, dx \quad \forall w_h \in Y_h. \end{aligned} \tag{36}$$

Proposition 34 *The functional $J_h : L^\infty(\Omega_h) \to \mathbb{R}$ is of class C^2. Moreover, for all $u, v, v_1, v_2 \in L^\infty(\Omega_h)$*

$$J_h'(u)v = \int_{\Omega_h} \left(\frac{\partial L}{\partial u}(x, y_h(u), u) + \varphi_h(u) \right) v \, dx \tag{37}$$

and

$$J_h''(u)v_1v_2 = \int_{\Omega_h} \left[\frac{\partial^2 L}{\partial y^2}(x, y_h(u), u) z_{h1} z_{h2} + \right.$$

$$\frac{\partial^2 L}{\partial y \partial u}(x, y_h(u), u)[z_{h1}v_2 + z_{h2}v_1] +$$

$$\left. \frac{\partial^2 L}{\partial u^2}(x, y_h(u), u)v_1v_2 - \varphi_h(u)\phi''(y_h(u)) z_{h1} z_{h2} \right] dx \tag{38}$$

where $y_h(u) = G_h(u)$, $\varphi_h(u) \in Y_h$ is defined in (36) and $z_{hi} = G_h'(u)v_i$, $i = 1, 2$.

We conclude this section by studying the existence of a solution of problem (P_h) and establishing the first order optimality conditions. The second order conditions are analogous to those proved for problem (P) and they can be obtained by the classical methods of finite dimensional optimization.

Theorem 35 *For every $h > 0$, problem (P_h) has at least one solution. If $\bar{u}_h$ is a local minimum of (P_h), then there exist $\bar{y}_h$, $\bar{\varphi}_h \in Y_h$ such that*

$$a(\bar{y}_h, w_h) + \int_{\Omega_h} \phi(\bar{y}_h) w_h(x) \, dx = \int_{\Omega_h} \bar{u}_h(x) w_h(x) \, dx \quad \forall w_h \in Y_h, \tag{39}$$

$$a(w_h, \bar{\varphi}_h) + \int_{\Omega_h} \phi'(\bar{y}_h) \bar{\varphi}_h w_h \, dx = \int_{\Omega_h} \frac{\partial L}{\partial y}(x, \bar{y}_h, \bar{u}_h) w_h \, dx \; \forall w_h \in Y_h, \tag{40}$$

$$\int_{\Omega_h} \left\{ \bar{\varphi}_h + \frac{\partial L}{\partial u}(x, \bar{y}_h, \bar{u}_h) \right\} (u_h - \bar{u}_h) dx \geq 0 \; \forall u_h \in \mathbb{K}_h. \tag{41}$$

Proof The existence of a solution is an immediate consequence of the compactness of $\mathbb{K}_h$ in U_h and the continuity of J_h. The optimality system (39)–(41) follows from Lemma 15 and Proposition 34. □

From this theorem we can deduce a representation formula of local minima of (P_h) analogous to that obtained in Theorem 20.

Theorem 36 *Under the hypotheses **(H1)** and **(H2)**, if $\bar{u}_h$ is a local minimum of* (P_h), *and $\bar{y}_h$ and $\bar{\varphi}_h$ are the state and adjoint state associated to $\bar{u}_h$, then for every $T \in \mathscr{T}_h$ the equation*

$$\int_T [\bar{\varphi}_h(x) + \frac{\partial L}{\partial u}(x, \bar{y}_h(x), t)]\, dx = 0, \tag{42}$$

has a unique solution $\bar{t} = \bar{s}_T$. The mapping $\bar{s}_h \in U_h$, defined by $\bar{s}_{h|T} = \bar{s}_T$, is related with $\bar{u}_h$ by the formula

$$\bar{u}_h(x) = \mathrm{Proj}_{[\alpha,\beta]}(\bar{s}_h(x)) = \max(\alpha, \min(\beta, \bar{s}_h(x))). \tag{43}$$

Proof The existence of a unique solution of (42) is a consequence of hypothesis **(H2)**. Let us denote by $\bar{u}_T$ the restriction of $\bar{u}_h$ to T. From the definition of U_h and (41) we deduce that

$$\int_T \left\{ \bar{\varphi}_h + \frac{\partial L}{\partial u}(x, \bar{y}_h, \bar{u}_T) \right\} dx\, (t - \bar{u}_T) \geq 0 \quad \forall t \in [\alpha, \beta] \text{ and } \forall T \in \mathscr{T}_h.$$

From here we get

$$\alpha < \bar{u}_T < \beta \Rightarrow \int_T \left\{ \bar{\varphi}_h + \frac{\partial L}{\partial u}(x, \bar{y}_h, \bar{u}_T) \right\} dx = 0 \Rightarrow \bar{u}_T = \bar{s}_T,$$

$$\bar{u}_T = \beta \quad \Rightarrow \int_T \left\{ \bar{\varphi}_h + \frac{\partial L}{\partial u}(x, \bar{y}_h, \bar{u}_T) \right\} dx \leq 0 \Rightarrow \bar{u}_T \leq \bar{s}_T,$$

$$\bar{u}_T = \alpha \quad \Rightarrow \int_T \left\{ \bar{\varphi}_h + \frac{\partial L}{\partial u}(x, \bar{y}_h, \bar{u}_T) \right\} dx \geq 0 \Rightarrow \bar{u}_T \geq \bar{s}_T,$$

which implies (43). □

8 Convergence of the Approximations

In this section we will prove that the solutions of the discrete problems (P_h) converge strongly in $L^\infty(\Omega_h)$ to solutions of problem (P). We will also prove that strict local minima of problem (P) can be approximated by local minima of problems (P_h). In order to prove these convergence results we will use two lemmas whose proofs can be found in [2] and [11].

Lemma 37 *Let $(v, v_h) \in L^\infty(\Omega) \times U_h$ satisfy $\|v\|_{L^\infty(\Omega)} \leq M$ and $\|v_h\|_{L^\infty(\Omega_h)} \leq M$. Let us assume that y_v and $y_h(v_h)$ are the solutions of (1) and (33) corresponding to v and v_h respectively. Moreover, let φ_v and $\varphi_h(v_h)$ be the solutions of (8) and (36)*

corresponding to v and v_h respectively. Then the following estimates hold

$$\|y_v - y_h(v_h)\|_{H^1(\Omega_h)} + \|\varphi_v - \varphi_h(v_h)\|_{H^1(\Omega_h)} \le C(h + \|v - v_h\|_{L^2(\Omega_h)}), \tag{44}$$

$$\|y_v - y_h(v_h)\|_{L^2(\Omega_h)} + \|\varphi_v - \varphi_h(v_h)\|_{L^2(\Omega_h)} \le C(h^2 + \|v - v_h\|_{L^2(\Omega_h)}), \tag{45}$$

$$\|y_v - y_h(v_h)\|_{L^\infty(\Omega_h)} + \|\varphi_v - \varphi_h(v_h)\|_{L^\infty(\Omega_h)} \le C(h^2|\log h|^2 + \|v - v_h\|_{L^2(\Omega_h)}), \tag{46}$$

where $C \equiv C(\Omega, n, M)$ is a positive constant independent of h.

Estimate (46) was not proved in [2], but it follows from [2] and the uniform error estimates for the discretization of linear elliptic equations; see for instance [44] and [45].

Lemma 38 *Let $\{u_h\}_{h>0}$ be a sequence, with $u_h \in \mathbb{K}_h$ and $u_h \rightharpoonup u$ weakly in $L^1(\Omega)$. Then $y_h(u_h) \to y_u$ and $\varphi_h(u_h) \to \varphi_u$ in $H_0^1(\Omega) \cap C(\bar\Omega)$ strongly as $h \to 0$. Moreover $J(u) \le \liminf_{h\to 0} J_h(u_h)$.*

Let us remark that u_h is only defined in Ω_h. Therefore, we need to establish what $u_h \rightharpoonup u$ weakly in $L^1(\Omega)$ means. It means that

$$\int_{\Omega_h} \psi u_h \, dx \to \int_\Omega \psi u \, dx \quad \forall \psi \in L^\infty(\Omega).$$

Since the measure of $\Omega \setminus \Omega_h$ tends to zero when $h \to 0$, the above property is equivalent to

$$\int_\Omega \psi \tilde u_h \, dx \to \int_\Omega \psi u \, dx \quad \forall \psi \in L^\infty(\Omega)$$

for any uniformly bounded extension $\tilde u_h$ of u_h to Ω. Analogously we can define the weak* convergence in $L^\infty(\Omega)$.

Theorem 39 *Let us assume that **(H1)** and **(H2)** hold. For every $h > 0$ let $\bar u_h$ be a solution of* (P$_h$). *Then there exist subsequences of $\{\bar u_h\}_{h>0}$ converging in the weak* topology of $L^\infty(\Omega)$ that will be denoted in the same way. If $\bar u_h \overset{*}{\rightharpoonup} \bar u$ in $L^\infty(\Omega)$, then $\bar u$ is a solution of (P) and the following identities hold*

$$\lim_{h\to 0} J_h(\bar u_h) = J(\bar u) = \inf(\mathrm{P}) \quad \text{and} \quad \lim_{h\to 0} \|\bar u - \bar u_h\|_{L^\infty(\Omega_h)} = 0. \tag{47}$$

Proof The existence of subsequences converging in the weak* topology of $L^\infty(\Omega)$ is a consequence of the boundedness of $\{\bar u_h\}_{h>0}$, $\alpha \le \bar u_h(x) \le \beta$, for every $h > 0$. Let $\bar u$ be a limit point of one of these converging subsequences. We are going to prove that $\bar u$ is a solution of (P). Let $\tilde u$ be a solution of (P). From Theorem 20 we deduce that $\tilde u$ is Lipschitz in $\bar\Omega$. Let us consider the operator $\Pi_h : L^1(\Omega) \longrightarrow U_h$

defined by

$$\Pi_h u|_T = \frac{1}{|T|}\int_T u(x)\,dx \quad \forall T \in \mathscr{T}_h.$$

Let $u_h = \Pi_h \tilde{u} \in U_h$. It is easy to prove that $u_h \in \mathbb{K}_h$ and

$$\|\tilde{u} - u_h\|_{L^\infty(\Omega_h)} \leq \Lambda_{\tilde{u}} h,$$

where $\Lambda_{\tilde{u}}$ is the Lipschitz constant of $\tilde{u}$. By applying the Lemmas 37 and 38 we get

$$J(\bar{u}) \leq \liminf_{h\to 0} J_h(\bar{u}_h) \leq \limsup_{h\to 0} J_h(\bar{u}_h) \leq$$
$$\leq \limsup_{h\to 0} J_h(u_h) = J(\tilde{u}) = \inf(\mathrm{P}) \leq J(\bar{u}),$$

which proves that $\bar{u}$ is a solution of (P) and

$$\lim_{h\to 0} J_h(\bar{u}_h) = J(\bar{u}) = \inf(\mathrm{P}).$$

Let us prove now the uniform convergence $\bar{u}_h \to \bar{u}$. From (15) and (43) follows

$$\|\bar{u} - \bar{u}_h\|_{L^\infty(\Omega_h)} \leq \|\bar{s} - \bar{s}_h\|_{L^\infty(\Omega_h)},$$

therefore it is enough to prove the uniform convergence of $\{\bar{s}_h\}_{h>0}$ to $\bar{s}$. On the other hand, from Theorem 36 we have that

$$\int_T [\bar{\varphi}_h(x) + \frac{\partial L}{\partial u}(x, \bar{y}_h(x), \bar{s}_{h|T})]\,dx = 0.$$

From this equality and the continuity of the integrand with respect to x it follows the existence of a point $\xi_T \in T$ such that

$$\bar{\varphi}_h(\xi_T) + \frac{\partial L}{\partial u}(\xi_T, \bar{y}_h(\xi_T), \bar{s}_h(\xi_T)) = 0. \tag{48}$$

Given $x \in \Omega_h$, let $T \in \mathscr{T}_h$ be such that $x \in T$. Since $\bar{s}_h$ is constant in each element T

$$|\bar{s}(x) - \bar{s}_h(x)| \leq |\bar{s}(x) - \bar{s}(\xi_T)| + |\bar{s}(\xi_T) - \bar{s}_h(\xi_T)| \leq$$
$$\Lambda_{\bar{s}}|x - \xi_T| + |\bar{s}(\xi_T) - \bar{s}_h(\xi_T)| \leq \Lambda_{\bar{s}} h + |\bar{s}(\xi_T) - \bar{s}_h(\xi_T)|,$$

where $\Lambda_{\bar{s}}$ is the Lipschitz constant of $\bar{s}$. Thus it remains to prove the convergence $\bar{s}_h(\xi_T) \to \bar{s}(\xi_T)$ for every T. To do this, we will use again the strict positivity of the second derivative of L with respect to u (Hypothesis **(H2)**) along with (48) and the

fact that $\bar{s}(x)$ is the solution of the Eq. (14) to get

$$\Lambda|\bar{s}(\xi_T) - \bar{s}_h(\xi_T)| \le \left|\frac{\partial L}{\partial u}(\xi_T, \bar{y}_h(\xi_T), \bar{s}(\xi_T)) - \frac{\partial L}{\partial u}(\xi_T, \bar{y}_h(\xi_T), \bar{s}_h(\xi_T))\right| \le$$
$$\left|\frac{\partial L}{\partial u}(\xi_T, \bar{y}_h(\xi_T), \bar{s}(\xi_T)) - \frac{\partial L}{\partial u}(\xi_T, \bar{y}(\xi_T), \bar{s}(\xi_T))\right| +$$
$$\left|\frac{\partial L}{\partial u}(\xi_T, \bar{y}(\xi_T), \bar{s}(\xi_T)) - \frac{\partial L}{\partial u}(\xi_T, \bar{y}_h(\xi_T), \bar{s}_h(\xi_T))\right| =$$
$$\left|\frac{\partial L}{\partial u}(\xi_T, \bar{y}_h(\xi_T), \bar{s}(\xi_T)) - \frac{\partial L}{\partial u}(\xi_T, \bar{y}(\xi_T), \bar{s}(\xi_T))\right| + |\bar{\varphi}(\xi_T) - \bar{\varphi}_h(\xi_T)| \to 0$$

thanks to the uniform convergence $\bar{y}_h \to \bar{y}$ and $\bar{\varphi}_h \to \bar{\varphi}$ (Lemma 38). □

In a certain sense, the next result is converse to the previous theorem. The question we formulate now is wether a local minimum u of (P) can be approximated by a local minimum u_h of (P_h). The answer is positive if the local minimum u is strict.

Theorem 40 *Let us assume that **(H1)** and **(H2)** hold. Let $\bar{u}$ be a strict local minimum of (P) in the $L^p(\Omega)$ sense with $1 \le p \le \infty$. Then there exist a ball $B_\varepsilon(\bar{u})$ of $L^p(\Omega)$ and $h_0 > 0$ such that* (P_h) *has a local minimum $\bar{u}_h \in B_\varepsilon(\bar{u})$ for every $h < h_0$. Moreover the convergences* (47) *hold.*

Proof First we analyze the case $p = \infty$. Since $\bar{u}$ is a strict local minimum in the $L^\infty(\Omega)$ sense, there exists $\varepsilon > 0$ such that $\bar{u}$ is the unique solution of problem

$$(P_\varepsilon) \begin{cases} \min J(u) \\ u \in \mathbb{K} \cap \bar{B}_\varepsilon(\bar{u}), \end{cases}$$

where $B_\varepsilon(\bar{u})$ denotes the ball of $L^\infty(\Omega)$. Let us consider the functions

$$\alpha_\varepsilon(x) = \max\{\alpha, \bar{u}(x) - \varepsilon\} \text{ and } \beta_\varepsilon(x) = \min\{\beta, \bar{u}(x) + \varepsilon\} \ \ \forall x \in \Omega.$$

It is easy to check that

$$\begin{aligned} &|\alpha_\varepsilon(x_2) - \alpha_\varepsilon(x_1)| \le |\bar{u}(x_2) - \bar{u}(x_1)| \le \Lambda_u |x_2 - x_1|, \\ &|\beta_\varepsilon(x_2) - \beta_\varepsilon(x_1)| \le |\bar{u}(x_2) - \bar{u}(x_1)| \le \Lambda_u |x_2 - x_1|, \end{aligned} \quad \forall x_1, x_2 \in \Omega. \tag{49}$$

For every $T \in \mathcal{T}_h$ we set

$$\alpha_{\varepsilon T} = \frac{1}{|T|}\int_T \alpha_\varepsilon(x)\,dx, \quad \beta_{\varepsilon T} = \frac{1}{|T|}\int_T \beta_\varepsilon(x)\,dx$$

and

$$\alpha_{\varepsilon h}(x) = \sum_{T\in\mathscr{T}} \alpha_{\varepsilon T}\chi_T \quad \text{and} \quad \beta_{\varepsilon h}(x) = \sum_{T\in\mathscr{T}} \beta_{\varepsilon T}\chi_T,$$

where χ_T denotes the characteristic function of T. Using (49) we infer $\forall x \in T$

$$|\alpha_\varepsilon(x) - \alpha_{\varepsilon h}(x)| = \left|\frac{1}{|T|}\int_T (\alpha_\varepsilon(x) - \alpha_\varepsilon(\xi))\, d\xi\right| \leq \Lambda_u h.$$

A similar inequality is obtained for $\beta_\varepsilon(x) - \beta_{\varepsilon h}(x)$. Hence we get

$$\|\alpha_\varepsilon - \alpha_{\varepsilon h}\|_{L^\infty(\Omega_h)} \leq \Lambda_u h \ \text{ and } \ \|\beta_\varepsilon - \beta_{\varepsilon h}\|_{L^\infty(\Omega_h)} \leq \Lambda_u h. \tag{50}$$

Now, we introduce a discrete set $\mathbb{K}_{\varepsilon h}$ approximating $\mathbb{K}_\varepsilon = \mathbb{K} \cap \bar{B}_\varepsilon(\bar{u})$ in the following way

$$\mathbb{K}_{\varepsilon h} = \{u_h \in U_h : \alpha_{\varepsilon T} \leq u_T \leq \beta_{\varepsilon T} \ \ \forall T \in \mathscr{T}_h\}.$$

Associated to this set we consider the family of discrete control problems

$$(\mathrm{P}_{\varepsilon h}) \begin{cases} \min J_h(u_h) \\ u_h \in \mathbb{K}_{\varepsilon h}. \end{cases}$$

Let $\Pi_h : L^1(\Omega) \longrightarrow U_h$ be the operator introduced in the proof of the previous theorem. Since $\|\Pi_h\bar{u} - \bar{u}\|_{L^\infty(\Omega_h)} \to 0$, it is obvious that $\Pi_h\bar{u} \in \mathbb{K}_{\varepsilon h}$ for every h small enough. Therefore $\mathbb{K}_{\varepsilon h}$ is non empty compact set and consequently $(\mathrm{P}_{\varepsilon h})$ has at least one solution $\bar{u}_h$ for every small h. In the sequel every element $\bar{u}_h$ is extended to Ω by setting $\bar{u}_h(x) = \bar{u}(x)$ if $x \in \Omega \setminus \Omega_h$. Now, let us consider a subsequence, denoted in the same way, such that $\bar{u}_h \overset{*}{\rightharpoonup} \tilde{u}$ in $L^\infty(\Omega)$. Arguing as in the proof of Theorem 39, we have that $\tilde{u}$ is a solution of (P_ε) and $J_h(\bar{u}_h) \to J(\tilde{u})$. But $\bar{u}$ is the unique solution of (P_ε), hence $\tilde{u} = \bar{u}$, and consequently the whole sequence $\{\bar{u}\}_{h>0}$ converges weakly* to $\bar{u}$ in $L^\infty(\Omega)$. Let us prove that this convergence is strong. As in Theorem 36, performing some obvious modifications in the proof, we have that every $\bar{u}_h$ satisfies

$$\bar{u}_h(x) = \mathrm{Proj}_{[\alpha_{\varepsilon h}(x),\beta_{\varepsilon h}(x)]}(\bar{s}_h(x)). \tag{51}$$

Moreover, taking into account (15) and the definition of α_ε and β_ε, it is obvious that

$$\bar{u}(x) = \mathrm{Proj}_{[\alpha,\beta]}(\bar{s}(x)) = \mathrm{Proj}_{[\alpha_\varepsilon(x),\beta_\varepsilon(x)]}(\bar{s}(x)). \tag{52}$$

Finally, combining (50)–(52) and the convergence $\|\bar{s} - \bar{s}_h\|_{L^\infty(\Omega_h)} \to 0$ as $h \to 0$ established in the proof of Theorem 39, we obtain for $x \in \Omega_h$

$$
\begin{aligned}
|\bar{u}(x) - \bar{u}_h(x)| &= \left| \mathrm{Proj}_{[\alpha_\varepsilon(x),\beta_\varepsilon(x)]}(\bar{s}(x)) - \mathrm{Proj}_{[\alpha_{\varepsilon h}(x),\beta_{\varepsilon h}(x)]}(\bar{s}_h(x)) \right| \\
&\le \left| \mathrm{Proj}_{[\alpha_\varepsilon(x),\beta_\varepsilon(x)]}(\bar{s}(x)) - \mathrm{Proj}_{[\alpha_{\varepsilon h}(x),\beta_{\varepsilon h}(x)]}(\bar{s}(x)) \right| \\
&\quad + \left| \mathrm{Proj}_{[\alpha_{\varepsilon h}(x),\beta_{\varepsilon h}(x)]}(\bar{s}(x)) - \mathrm{Proj}_{[\alpha_{\varepsilon h}(x),\beta_{\varepsilon h}(x)]}(\bar{s}_h(x)) \right| \\
&\le \max\{|\alpha_\varepsilon(x) - \alpha_{\varepsilon h}(x)|, |\beta_\varepsilon(x) - \beta_{\varepsilon h}(x)|\} + |\bar{s}(x) - \bar{s}_h(x)| \\
&\le \Lambda_u h + \|\bar{s} - \bar{s}_h\|_{L^\infty(\Omega_h)} \to 0
\end{aligned}
$$

as $h \to 0$, which proves that $\|\bar{u} - \bar{u}_h\|_{L^\infty(\Omega_h)} \to 0$. Therefore there exists $h_0 > 0$ such that $\|\bar{u} - \bar{u}_h\|_{L^\infty(\Omega_h)} < \frac{\varepsilon}{2}$ for every $h \le h_0$. Then, for any $u_h \in \mathbb{K}_h \cap B_{\frac{\varepsilon}{2}}(\bar{u}_h)$ we have that

$$
\|u_h - \bar{u}\|_{L^\infty(\Omega)} \le \|\bar{u}_h - \bar{u}\|_{L^\infty(\Omega)} + \|u_h - \bar{u}_h\|_{L^\infty(\Omega)} < \varepsilon,
$$

hence $u_h \in \mathbb{K}_{\varepsilon h}$ and consequently $J(\bar{u}_h) \le J(u_h)$. This proves that $\bar{u}_h$ is a local minimum of (P_h).

To complete the proof we consider the case $1 \le p < \infty$. We introduce again the problem $(\mathrm{P}_{\varepsilon h})$ with $\mathbb{K}_{\varepsilon h} = \mathbb{K}_h \cap \bar{B}_\varepsilon(\bar{u})$. This time, $\bar{B}_\varepsilon(\bar{u})$ denotes the ball in $L^p(\Omega)$. Once again we have that $\Pi_h \bar{u} \in \mathbb{K}_{\varepsilon h}$ for every h small enough. Now, we consider a sequence $\{\bar{u}_h\}_{h>0}$ of solutions of the problems $(\mathrm{P}_{\varepsilon h})$, where $\bar{u}_h$ is extended to Ω by $\bar{u}$ as before. Then, arguing as above we get that $\bar{u}_h \overset{*}{\rightharpoonup} \bar{u}$ and $J_h(\bar{u}_h) \to J(\bar{u})$ as $h \to 0$. If we prove that $\bar{u}_h \to \bar{u}$ in $L^2(\Omega)$ as $h \to 0$, then the boundedness of $\{\bar{u}_h\}_h$ in $L^\infty(\Omega)$ implies the strong convergence $\bar{u}_h \to \bar{u}$ in $L^p(\Omega)$ as $h \to 0$ for every $p < \infty$. Hence, we deduce as above that $\bar{u}_h$ is a local minimum of (P_h).

Let us prove that $\bar{u}_h \to \bar{u}$ in $L^2(\Omega)$. First we observe that Lemma 38 implies that $\bar{y}_h \to \bar{y}$ in $C(\bar{\Omega})$. Then, using the hypothesis **(H1)**, the convergence $\bar{u}_h \overset{*}{\rightharpoonup} \bar{u}$, (32), and $J_h(\bar{u}_h) \to J(\bar{u})$, we infer

$$
\begin{aligned}
&\int_\Omega [L(x, \bar{y}, \bar{u}_h) - L(x, \bar{y}, \bar{u})]\, dx \\
&= \int_\Omega [L(x, \bar{y}, \bar{u}_h) - L(x, \bar{y}_h, \bar{u}_h)]\, dx + \int_\Omega [L(x, \bar{y}_h, \bar{u}_h) - L(x, \bar{y}, \bar{u})]\, dx \\
&= \int_\Omega \frac{\partial L}{\partial y}(x, \bar{y} + \theta_h(\bar{y}_h - \bar{y}), \bar{u}_h)\, dx + \int_{\Omega \setminus \Omega_h} L(x, 0, \bar{u})\, dx + [J_h(\bar{u}_h) - J(\bar{u})] \to 0.
\end{aligned}
$$

From this convergence and hypothesis **(H2)**, we obtain by a Taylor expansion

$$\begin{aligned}0 &= \lim_{h\to 0}\int_\Omega [L(x,\bar y,\bar u_h) - L(x,\bar y,\bar u)]\,dx\\ &= \lim_{h\to 0}\int_\Omega \frac{\partial L}{\partial u}(x,\bar y,\bar u)(\bar u_h-\bar u)\,dx + \lim_{h\to 0}\frac{1}{2}\int_\Omega \frac{\partial^2 L}{\partial u^2}(x,\bar y,\bar u+\vartheta_h(\bar u_h-\bar u)))(\bar u_h-\bar u)^2\,dx\\ &= \lim_{h\to 0}\frac{1}{2}\int_\Omega \frac{\partial^2 L}{\partial u^2}(x,\bar y,\bar u+\vartheta_h(\bar u_h-\bar u)))(\bar u_h-\bar u)^2\,dx \ge \frac{\Lambda}{2}\limsup_{h\to 0}\|\bar u_h-\bar u\|^2_{L^2(\Omega)},\end{aligned}$$

which concludes the proof. □

9 Error Estimates

In this section we will assume that **(H1)** and **(H2)** hold and that $\bar u$ is a local minimum of (P) satisfying the sufficient second order condition for optimality (19) or equivalently (27). $\{\bar u_h\}_{h>0}$ denotes a sequence of local minima of problems (P_h) such that $\|\bar u-\bar u_h\|_{L^\infty(\Omega_h)}\to 0$; remind Theorems 39 and 40. The goal of this section is to estimate the error $\bar u-\bar u_h$ in the of $L^2(\Omega_h)$ and $L^\infty(\Omega_h)$ norms, respectively. To this aim , we are going to prove three auxiliary lemmas.

For convenience, in this section we will extend $\bar u_h$ to Ω by taking $\bar u_h(x)=\bar u(x)$ for every $x\in\Omega\setminus\Omega_h$.

Lemma 41 *Let $\delta>0$ be as in Theorem 25. Then there exists $h_0>0$ such that*

$$\frac{\delta}{2}\|\bar u-\bar u_h\|^2_{L^2(\Omega_h)} \le (J'(\bar u_h)-J'(\bar u))(\bar u_h-\bar u)\ \ \forall h<h_0. \tag{53}$$

Proof Let us set

$$\bar d_h(x) = \frac{\partial L}{\partial u}(x,\bar y_h(x),\bar u_h(x)) + \bar\varphi_h(x)$$

and take $\delta>0$ and $\tau>0$ as in Theorem 25. We know that $\bar d_h$ converges uniformly to $\bar d$ in Ω, therefore there exists $h_\tau>0$ such that

$$\|\bar d-\bar d_h\|_{L^\infty(\Omega_h)} < \frac{\tau}{4}\ \ \forall h\le h_\tau. \tag{54}$$

For every $T\in\mathscr{T}_h$ we define

$$I_T = \int_T \bar d_h(x)\,dx.$$

From (41), it follows

$$\bar{u}_{h|_T} = \begin{cases} \alpha \text{ if } I_T > 0 \\ \beta \text{ if } I_T < 0. \end{cases}$$

Let us take $0 < h_1 \le h_\tau$ such that

$$|\bar{d}(x_2) - \bar{d}(x_1)| < \frac{\tau}{4} \text{ if } |x_2 - x_1| < h_1.$$

This inequality, along with (54), implies that

$$\xi \in T \text{ and } \bar{d}(\xi) > \tau \Rightarrow \bar{d}_h(x) > \frac{\tau}{2} \ \forall x \in T, \ \forall T \in \hat{\mathscr{T}}_h, \ \forall h < h_1,$$

hence $I_T > 0$, therefore $\bar{u}_{h|_T} = \alpha$, in particular $\bar{u}_h(\xi) = \alpha$. From (16) we also have $\bar{u}(\xi) = \alpha$. Then $(\bar{u}_h - \bar{u})(\xi) = 0$ whenever $\bar{d}(\xi) > \tau$ and $h < h_1$. We can prove the analogous result when $\bar{d}(\xi) < -\tau$. On the other hand, since $\alpha \le \bar{u}_h(x) \le \beta$, it is obvious that $(\bar{u}_h - \bar{u})(x) \ge 0$ if $\bar{u}(x) = \alpha$ and $(\bar{u}_h - \bar{u})(x) \le 0$ if $\bar{u}(x) = \beta$. Thus we have proved that $(\bar{u}_h - \bar{u}) \in C_{\bar{u}}^{\tau}$, remember that $\bar{u} = \bar{u}_h$ in $\Omega \setminus \Omega_h$. Then (27) leads to

$$J''(\bar{u})(\bar{u}_h - \bar{u})^2 \ge \delta \|\bar{u}_h - \bar{u}\|^2_{L^2(\Omega)} = \delta \|\bar{u}_h - \bar{u}\|^2_{L^2(\Omega_h)} \ \forall h < h_1. \tag{55}$$

On the other hand, by applying the mean value theorem, we get for some $0 < \theta_h < 1$ that

$$\begin{gathered} (J'(\bar{u}_h) - J'(\bar{u}))(\bar{u}_h - \bar{u}) = J''(\bar{u} + \theta_h(\bar{u}_h - \bar{u}))(\bar{u}_h - \bar{u})^2 \ge \\ (J''(\bar{u} + \theta_h(\bar{u}_h - \bar{u})) - J''(\bar{u}))(\bar{u}_h - \bar{u})^2 + J''(\bar{u})(\bar{u}_h - \bar{u})^2 \ge \\ \left(\delta - \|J''(\bar{u} + \theta_h(\bar{u}_h - \bar{u})) - J''(\bar{u})\|\right) \|\bar{u}_h - \bar{u}\|^2_{L^2(\Omega)}. \end{gathered}$$

Finally, recalling Remark 18, we can choose $0 < h_0 \le h_1$ such that

$$\|J''(\bar{u} + \theta_h(\bar{u}_h - \bar{u})) - J''(\bar{u})\| \le \frac{\delta}{2} \ \forall h < h_0$$

to deduce (53). □

In the next step the convergence of J'_h to J' is estimated.

Lemma 42 *There exists a constant $C > 0$ independent of h such that for every u_1, $u_2 \in \mathbb{K}$ and every $v \in L^2(\Omega)$, with $v = 0$ on $\Omega \setminus \Omega_h$, the following inequalities are fulfilled*

$$|(J'_h(u_2) - J'(u_1))v| \le C \left\{h^2 + \|u_2 - u_1\|_{L^2(\Omega)}\right\} \|v\|_{L^2(\Omega_h)}. \tag{56}$$

Proof By using the expression of the derivatives given by (9) and (37) along with the inequality (32) we get

$$
\begin{aligned}
&|(J_h'(u_2) - J'(u_1))v| \\
&\leq \int_{\Omega_h} \left| \left(\frac{\partial L}{\partial u}(x, y_h(u_2), u_2) + \varphi_h(u_2) \right) - \left(\frac{\partial L}{\partial u}(x, y_{u_1}, u_1) + \varphi_{u_1} \right) \right| |v| \, dx \\
&\leq C \left\{ \|\varphi_h(u_2) - \varphi_{u_1}\|_{L^2(\Omega_h)} + \|y_h(u_2) - y_{u_1}\|_{L^2(\Omega_h)} \right. \\
&\qquad \left. + \|u_2 - u_1\|_{L^2(\Omega_h)} \right\} \|v\|_{L^2(\Omega_h)}.
\end{aligned}
$$

Now (56) follows from the previous inequality and (45). □

A key point in the derivation of the error estimate is to get a good approximate of $\bar{u}$ by a discrete control $u_h \in \mathbb{K}_h$ satisfying $J'(\bar{u})\bar{u} = J'(\bar{u})u_h$. Let us define this control u_h and prove that it fulfills the required conditions. For every $T \in \mathscr{T}_h$ let us set

$$
I_T = \int_T \bar{d}(x) \, dx.
$$

We define $u_h \in U_h$ with $u_{h|T} = u_{hT}$ for every $T \in \mathscr{T}_h$ by the expression

$$
u_{hT} = \begin{cases} \dfrac{1}{I_T} \displaystyle\int_T \bar{d}(x)\bar{u}(x) \, dx \text{ if } I_T \neq 0 \\ \\ \dfrac{1}{|T|} \displaystyle\int_T \bar{u}(x) \, dx \quad \text{if } I_T = 0. \end{cases} \tag{57}
$$

We extend this function to Ω by taking $u_h(x) = \bar{u}(x)$ for every $x \in \Omega \setminus \Omega_h$. This function u_h satisfies our requirements.

Lemma 43 *There exists $h_0 > 0$ such that for every $0 < h < h_0$ the following properties hold*

1. $u_h \in \mathbb{K}_h$.
2. $J'(\bar{u})\bar{u} = J'(\bar{u})u_h$.
3. *There exists $C > 0$ independent of h such that*

$$
\|\bar{u} - u_h\|_{L^\infty(\Omega_h)} \leq Ch. \tag{58}
$$

Proof Let $\Lambda_{\bar{u}} > 0$ be the Lipschitz constant of $\bar{u}$ and let us take $h_0 = (\beta - \alpha)/(2\Lambda_{\bar{u}})$. Then, for every $T \in \mathscr{T}_h$ and every $h < h_0$, there holds

$$
|\bar{u}(\xi_2) - \bar{u}(\xi_1)| \leq \Lambda_{\bar{u}} |\xi_2 - \xi_1| \leq \Lambda_{\bar{u}} h < \frac{\beta - \alpha}{2} \quad \forall \xi_1, \xi_2 \in T
$$

which implies that $\bar{u}$ cannot take the values α and β in a same element T for any $h < h_0$. Therefore the sign of $\bar{d}$ in T must be constant thanks to (16). Hence $I_T = 0$ if and only if $\bar{d}(x) = 0$ for all $x \in T$. Moreover if $I_T \neq 0$, then $\bar{d}(x)/I_T \geq 0$ for every $x \in T$. As a first consequence of this, we get that $\alpha \leq u_{hT} \leq \beta$, which means that $u_h \in \mathbb{K}_h$. On the other hand

$$J'(\bar{u})u_h = \int_{\Omega\setminus\Omega_h} \bar{d}(x)\bar{u}_h(x)\,dx + \sum_{T\in\mathscr{T}_h} \left(\int_T \bar{d}(x)\,dx\right) u_{hT}$$

$$= \int_{\Omega\setminus\Omega_h} \bar{d}(x)\bar{u}(x)\,dx + \sum_{T\in\mathscr{T}_h} \int_T \bar{d}(x)\bar{u}(x)\,dx = J'(\bar{u})\bar{u}.$$

Finally let us prove (58). Since the sign of $\bar{d}(x)/I_T$ is always nonnegative and $\bar{d}$ is a continuous function, we get for any of the two possible definitions of u_{hT} the existence of a point $\xi_j \in T$ such that $u_{hT} = \bar{u}(\xi_j)$. Hence for all $x \in T$

$$|\bar{u}(x) - u_h(x)| = |\bar{u}(x) - u_{hT}| = |\bar{u}(x) - \bar{u}(\xi_j)| \leq \Lambda_{\bar{u}}|x - \xi_j| \leq \Lambda_{\bar{u}}h,$$

which proves (58). □

Finally we get the desired error estimates.

Theorem 44 *There exists a constant $C > 0$ independent of h such that*

$$\|\bar{u} - \bar{u}_h\|_{L^2(\Omega_h)} \leq Ch. \tag{59}$$

Proof Taking $u = \bar{u}_h$ in (13) we get

$$J'(\bar{u})(\bar{u}_h - \bar{u}) = \int_\Omega \left(\bar{\varphi} + \frac{\partial L}{\partial u}(x, \bar{y}, \bar{u})\right)(\bar{u}_h - \bar{u})\,dx \geq 0. \tag{60}$$

From (41) with u_h defined by (57) it follows

$$J_h'(\bar{u}_h)(u_h - \bar{u}_h) = \int_{\Omega_h} \left(\bar{\varphi}_h + \frac{\partial L}{\partial u}(x, \bar{y}_h, \bar{u}_h)\right)(u_h - \bar{u}_h)\,dx \geq 0,$$

then

$$J_h'(\bar{u}_h)(\bar{u} - \bar{u}_h) + J_h'(\bar{u}_h)(u_h - \bar{u}) \geq 0. \tag{61}$$

Adding (60) and (61) and using Lemma 43-2, we deduce

$$\left(J'(\bar{u}) - J_h'(\bar{u}_h)\right)(\bar{u} - \bar{u}_h) \leq J_h'(\bar{u}_h)(u_h - \bar{u}) = \left(J_h'(\bar{u}_h) - J'(\bar{u})\right)(u_h - \bar{u}).$$

For h small enough, this inequality along with (53) implies

$$\frac{\delta}{2}\|\bar{u}-\bar{u}_h\|^2_{L^2(\Omega_h)} \le \left(J'(\bar{u})-J'(\bar{u}_h)\right)(\bar{u}-\bar{u}_h) \le$$
$$\left(J'_h(\bar{u}_h)-J'(\bar{u}_h)\right)(\bar{u}-\bar{u}_h)+\left(J'(\bar{u}_h)-J'(\bar{u})\right)(u_h-\bar{u}).$$

We estimate the first term of the previous line using (56) with $u_2 = u_1 = \bar{u}_h$ and $v = \bar{u}-\bar{u}_h$. For the second term, we use the expression of J' given by (9) along with (45) for $v = \bar{u}$ and $v_h = \bar{u}_h$. We obtain

$$\frac{\delta}{2}\|\bar{u}-\bar{u}_h\|^2_{L^2(\Omega_h)} \le C_1\left(h^2+\|\bar{u}-u_h\|_{L^2(\Omega_h)}\right)\|\bar{u}-\bar{u}_h\|_{L^2(\Omega_h)}.$$

From (58) we deduce

$$\frac{\delta}{2}\|\bar{u}-\bar{u}_h\|_{L^2(\Omega_h)} \le C_2(h^2+h),$$

which implies (59). □

Finally let us prove the error estimate in $L^\infty(\Omega)$.

Theorem 45 *There exists a constant $C > 0$ independent of h such that*

$$\|\bar{u}-\bar{u}_h\|_{L^\infty(\Omega_h)} \le Ch. \tag{62}$$

Proof Let ξ_T be defined by (48). In the proof of Theorem 35 we obtained

$$\|\bar{u}-\bar{u}_h\|_{L^\infty(\Omega_h)} \le \|\bar{s}-\bar{s}_h\|_{L^\infty(\Omega_h)} \le \Lambda_{\bar{s}}h+$$
$$\max_{T\in\mathscr{T}_h}\left|\frac{\partial L}{\partial u}(\xi_T,\bar{y}_h(\xi_T),\bar{s}(\xi_T))-\frac{\partial L}{\partial u}(\xi_T,\bar{y}(\xi_T),\bar{s}(\xi_T))\right|+|\bar{\varphi}(\xi_T)-\bar{\varphi}_h(\xi_T)|.$$

Using the hypothesis **(H1)**, (46) and (59) we get

$$\|\bar{u}-\bar{u}_h\|_{L^\infty(\Omega_h)} \le \Lambda_{\bar{s}}h+C(\|\bar{y}-\bar{y}_h\|_{L^\infty(\Omega_h)}+\|\bar{\varphi}-\bar{\varphi}_h\|_{L^\infty(\Omega_h)}) \le$$
$$\Lambda_{\bar{s}}h+C(h+\|\bar{u}-\bar{u}_h\|_{L^2(\Omega_h)}) \le Ch.$$ □

10 Piecewise Linear Approximations of the Controls

In this section we are going to use a different approximation of the controls. Instead of using piecewise constant controls, we will consider piecewise linear and continuous functions. More precisely we take

$$U_h = \{u \in C(\bar{\Omega}_h) \mid u_{|T} \in \mathscr{P}_1, \text{ for all } T \in \mathscr{T}_h\},$$

where $\mathscr{P}_1$ is the space of polynomials of degree less or equal than 1. Let us denote by $\{x_j\}_{j=1}^{N(h)}$ the nodes of the triangulation $\mathscr{T}_h$. A basis of U_h is formed by the functions $\{e_j\}_{j=1}^{N(h)} \subset U_h$ defined by their values at the nodes x_j

$$e_j(x_i) = \begin{cases} 1 \text{ if } i = j, \\ 0 \text{ otherwise.} \end{cases}$$

In the sequel we will follow the notation $u_j = u_h(x_j)$ for any function $u_h \in U_h$, so that

$$u_h = \sum_{j=1}^{N(h)} u_j e_j.$$

The finite dimensional approximation of the optimal control problem is defined by

$$(\mathrm{P}_h) \begin{cases} \min J_h(u_h) = \int_{\Omega_h} L(x, y_h(u_h)(x), u_h(x))\, dx, \\ u_h \in \mathbb{K}_h = \{u_h \in U_h : \alpha \le u_j \le \beta \;\; 1 \le j \le N(h)\}. \end{cases}$$

Theorem 35 is still valid, there is no difference in the proof. However the representation of the optimal control given by the formula (43) is not true. The reason is that (43) is a local representation, but we can not change the values of the discrete controls in a triangle without modifying them in the neighbouring triangles. This was possible for piecewise constant controls, but it is not for continuous piecewise linear controls. The representation formula (43) was used in Theorem 39 to prove the uniform convergence of the discretizations; see (47). With new approximations of the controls, Theorem 39 is also valid except for the uniform convergence of the controls. However, we still can prove the strong convergence in $L^2(\Omega)$. Lemma 41 is also valid, but the given proof used the uniform convergence of the discrete controls. In the new framework the proof is completely different. Finally, the function u_h used in Lemma 43 is replaced by $u_h = I_h \bar{u}$, where $I_h : C(\bar{\Omega}) \longrightarrow U_h$ is the interpolation operator:

$$I_h \bar{u} = \sum_{j=1}^{N(h)} \bar{u}(x_j) e_j.$$

The elements u_h are extended to Ω by setting $u_h(x) = \bar{u}(x)$ in $\Omega \setminus \Omega_h$. For the interpolated function u_h it is proved

$$\lim_{h \to 0} \frac{J'(\bar{u})(u_h - \bar{u})}{h^2} = 0.$$

Taking into account these modifications, it is possible to follow the approach used in the proof of Theorem 44 to deduce that

$$\lim_{h\to 0}\frac{1}{h}\|\bar{u}-\bar{u}_h\|_{L^2(\Omega_h)}=0; \tag{63}$$

The reader is referred to Casas [8] for the details. In many practical situations this error estimate can be improved. Indeed, let us set

$$\mathscr{T}_h^+=\{T\in\mathscr{T}_h:|\bar{d}(x)|>0\ \forall x\in T\},$$
$$\mathscr{T}_h^0=\{T\in\mathscr{T}_h:\exists\xi_T\in T \text{ such that } \bar{d}(\xi_T)=0\},$$
$$\mathscr{T}_h^{0,1}=\{T\in\mathscr{T}_h^0:\ \text{such that } \bar{u}\in H^2(T)\},\quad \mathscr{T}_h^{0,2}=\mathscr{T}_h^0\setminus\mathscr{T}_h^{0,1}.$$

Now we assume that

$$\sum_{T\in\mathscr{T}_h^{0,2}}|T|\le Ch, \tag{64}$$

which is a frequent situation. Then we have

$$|J'(\bar{u})(I_h\bar{u}-\bar{u})|=\left|\sum_{T\in\mathscr{T}_h}\int_T\bar{d}(x)(I_h\bar{u}(x)-\bar{u}(x))\,dx\right|$$
$$=\left|\sum_{T\in\mathscr{T}_h^0}\int_T\bar{d}(x)(I_h\bar{u}(x)-\bar{u}(x)\,dx\right|$$
$$\le\sum_{T\in\mathscr{T}_h^0}\int_T|\bar{d}(x)-\bar{d}(\xi_T)||I_h\bar{u}(x)-\bar{u}(x)|\,dx$$
$$\le\Lambda_{\bar{d}}h\sum_{T\in\mathscr{T}_h^0}\int_T|I_h\bar{u}(x)-\bar{u}(x)|\,dx$$
$$\le\Lambda_{\bar{d}}h\left\{\sum_{T\in\mathscr{T}_h^{0,1}}\int_T|I_h\bar{u}(x)-\bar{u}(x)|\,dx+\sum_{T\in\mathscr{T}_h^{0,2}}\int_T|I_h\bar{u}(x)-\bar{u}(x)|\,dx\right\}$$
$$\le C\Lambda_{\bar{d}}h\left\{h^2\left(\sum_{T\in\mathscr{T}_h^{0,1}}\|\bar{u}\|^2_{H^2(T)}\right)^{1/2}+O(h)\|\bar{u}\|_{C^{0,1}(\bar{\Omega})}\sum_{T\in\mathscr{T}_h^{0,2}}|T|\right\}=O(h^3),$$

where $\Lambda_{\bar{d}}$ is the Lipschitz constant of $\bar{d}$. From this inequality we deduce by following the same proof of Theorem 44 that

$$\|\bar{u} - \bar{u}_h\|_{L^2(\Omega_h)} = O(h^{3/2}). \tag{65}$$

11 Semidiscretization of the Problem (P)

Let us consider in this section the very frequent case where

$$L(x, y, u) = L_0(x, y) + \frac{N}{2}u^2.$$

In this situation Hinze [26] suggested the discretization of the state equation by using piecewise linear approximations of the states, but he proposed no discretization for the controls. It means that $U_h = L^\infty(\Omega)$ for any $h > 0$. In this case we deduce from (41)

$$\int_{\Omega_h} (\bar{\varphi}_h(x) + N\bar{u}_h(x))(u_h(x) - \bar{u}_h(x))\, dx \geq 0 \quad \forall u_h \in \mathbb{K}, \tag{66}$$

which leads to

$$\bar{u}_h(x) = \mathrm{Proj}_{[\alpha,\beta]}\big(-\frac{1}{N}\bar{\varphi}_h(x)\big) = \max\{\alpha, \min\{\beta, -\frac{1}{N}\bar{\varphi}_h(x)\} \text{ a.e. in } \Omega_h. \tag{67}$$

Since $\bar{\varphi}_h \in Y_h$, $\bar{u}_h$ is piecewise linear and continuous in Ω_h. The linear structure of $\bar{u}_h$ is not supported on the grid defined by the nodes $\{x_j\}_{j=1}^{N(h)}$, but there is a different grid, which can be computed, where $\bar{u}_h$ is supported. Therefore it is possible to carry out the computations by using the corresponding grid for any iterate u_h^k obtained from the projection formula applied to φ_h^k. In this case we can take $u_h = \bar{u}$ in the proof of Theorem 44. Therefore the inequality used in that proof,

$$\frac{\delta}{2}\|\bar{u} - \bar{u}_h\|^2_{L^2(\Omega_h)} \leq \big(J'(\bar{u}) - J'(\bar{u}_h)\big)(\bar{u} - \bar{u}_h) \leq$$
$$\big(J_h'(\bar{u}_h) - J'(\bar{u}_h)\big)(\bar{u} - \bar{u}_h) + \big(J'(\bar{u}_h) - J'(\bar{u})\big)(u_h - \bar{u}),$$

is reduced to

$$\frac{\delta}{2}\|\bar{u} - \bar{u}_h\|^2_{L^2(\Omega_h)} \leq \big(J_h'(\bar{u}_h) - J'(\bar{u}_h)\big)(\bar{u} - \bar{u}_h).$$

Now using (56) we obtain

$$\|\bar{u} - \bar{u}_h\|_{L^2(\Omega_h)} \leq Ch^2. \tag{68}$$

12 Superconvergence and Postprocessing Step

A different approach to obtain convergence order $O(h^2)$ was shown by Meyer and Rösch in [37]. When the discretization of the state and the adjoint state is done using continuous piecewise linear elements and the discretization of the control is done using piecewise constant elements, a direct application of Theorem 44 and Eq. (45) in Lemma 37 leads to the error estimate

$$\|\bar{y} - \bar{y}_h\|_{L^2(\Omega_h)} + \|\bar{\varphi} - \bar{\varphi}_h\|_{L^2(\Omega_h)} \leq Ch.$$

Nevertheless, a superconvergence phenomenon, which suggests that the order of convergence should be $O(h^2)$, is observed in all the available numerical experiments. Meyer and Rösch proved that the values of the numerical solution $\bar{u}_h$ in the centroids of the elements have a quadratic convergence rate. Using this, they were able to explain the order of convergence that was observed numerically.

Theorem 46 *Suppose that* $L(x, y, u) = \frac{1}{2}(y - y_d(x))^2 + \frac{N}{2}u^2$ *with* $N > 0$ *and* $y_d \in L^p(\Omega)$, $p > 2$. *Suppose also that* (64) *is satisfied. Then*

$$\|\bar{y} - \bar{y}_h\|_{L^2(\Omega_h)} + \|\bar{\varphi} - \bar{\varphi}_h\|_{L^2(\Omega_h)} \leq Ch^2.$$

The proof given in [37] is for problems governed by linear equations. Recently, Krumbiegel and Pfefferer [30] proved that the result was true for Neumann control problems governed by a semilinear elliptic equation. Their technique can also be applied to distributed problems.

Using this, one can construct a new approximation of the optimal control, namely

$$\tilde{u}_h(x) = \mathrm{Proj}_{[\alpha,\beta]}\left(-\frac{1}{N}\bar{\varphi}_h(x)\right) \text{ a.e. in } \Omega_h,$$

that satisfies

$$\|\bar{u} - \tilde{u}_h\|_{L^2(\Omega)} \leq Ch^2.$$

13 Time Dependent Problems

Let us briefly comment on some results about the approximation of the time dependant problem presented in Sect. 4.4.

We will use the discontinuous Galerkin method dG0 to obtain the discretization in time. For this, we consider a quasi-uniform family of partitions of $[0, T]$, $0 = t_0 < t_1 < \cdots < t_{N_\tau} = T$ and denote $I_j = (t_{j-1}, t_j)$, $\tau_j = t_j - t_{j-1}$, $\tau = \max\{\tau_j\}$ and $\sigma = (h, \tau)$. We will also use the space-time cylinder $Q_h = \Omega_h \times (0, T)$.

Now we consider the finite dimensional space

$$\mathscr{Y}_\sigma = \{y_\sigma \in L^2(0,T;Y_h) : y_{\sigma|I_j} \in Y_h \ \forall j = 1,\dots,N_\tau\}.$$

The elements of $\mathscr{Y}_\sigma$ can be written as

$$y_\sigma = \sum_{j=1}^{N_\tau} y_{h,j}\chi_j$$

where $y_{h,j} \in Y_h$ for $j = 1,\dots,N_\tau$ and χ_j denotes the characteristic function of the interval $I_j = (t_{j-1}, t_j)$.

For every $u \in L^\infty(Q_h)$, we define its associated discrete state as the unique element $y_\sigma(u) \in \mathscr{Y}_\sigma$ such that

$$\begin{aligned}
&\int_{\Omega_h} (y_{h,j} - y_{h,j-1}) z_h dx + \tau_j a(y_{h,j}, z_h) + \int_{I_j}\int_{\Omega_h} b(x,t,y_{h,j}) z_h dx dt \\
&= \int_{I_j}\int_{\Omega_h} u z_h dx dt \quad \forall z_h \in Y_h \text{ and all } j = 1,\dots,N_\tau, \\
&\int_{\Omega_h} y_{h,0} z_h dx = \int_{\Omega_h} y_0 z_h dx \ \forall z_h \in Y_h.
\end{aligned} \tag{69}$$

By using the monotonicity of the nonlinear term $b(x,t,y)$, the proof of the existence and uniqueness of a solution for (69) is standard.

To discretize the controls, we will use piecewise constant functions. Consider U_h as in Sect. 7,

$$\mathscr{U}_\sigma = \{u_\sigma \in L^2(0,T;U_h) : \ u_{\sigma|I_j} \in U_h \ \forall j = 1,\dots,N_\tau\}$$

and

$$\mathscr{K}_\sigma = \{u_\sigma \in \mathscr{U}_\sigma : \ \alpha \le u_\sigma(x,t) \le \beta \text{ for a.e. } (x,t) \in Q_h\}.$$

We formulate the discrete problem as

$$(\mathrm{P}_\sigma) \quad \min_{u_\sigma \in \mathscr{K}_\sigma} J_\sigma(u_\sigma) = \frac{1}{2}\int_{Q_h} (y_\sigma(u_\sigma)(x,t) - y_d(x,t))^2 dx\,dt + \frac{N}{2}\int_{Q_h} u_\sigma(x,t)^2 dx\,dt.$$

In [39, Theorem 5.3], it is proved that under adequate second order sufficient conditions, the following error estimate holds:

$$\|\bar{u} - \bar{u}_\sigma\|_{L^2(Q_h)} \le C(\tau + h).$$

In that reference, it is also shown that the spatial order of convergence is improved to h^2 for a variational discretization or a post processing step analogous to those discussed in Sects. 11 and 12.

14 An Optimization Method

One of the most common ways to solve (P) is to use an SQP method; see [48] e.g. In this method, at each step, a control constrained linear-quadratic control problem must be solved; this is accomplished using a primal dual active set strategy, which is equivalent to a semismooth Newton method. Therefore, two nested loops are needed, the outer with quadratic order of convergence and the inner only with superlinear order of convergence. Nevertheless, the semismooth Newton method can be directly applied to solve (P), leading to a superlinear convergent sequence. Let us show how to do this.

To compute a solution of (P) we write the optimality system with the help of a Lagrange multiplier. From Theorems 19 and 20, we have that, if $\bar{u}$ is a local minimum of (P) and we define

$$\bar{\lambda} = -\bar{\varphi} - \frac{\partial L}{\partial u}(x, \bar{y}, \bar{u})$$

then $\bar{\lambda} \in C^{0,1}(\bar{\Omega})$ and for any $c > 0$ the optimality system can be written as

$$\begin{cases} A\bar{y} + \phi(\bar{y}) = \bar{u} \ \text{ in } \Omega, \\ \bar{y} = 0 \ \text{ on } \Gamma, \end{cases} \tag{70}$$

$$\begin{cases} A^*\bar{\varphi} + \phi'(\bar{y})\bar{\varphi} = \dfrac{\partial L}{\partial y}(x, \bar{y}, \bar{u}) \ \text{ in } \Omega, \\ \bar{\varphi} = 0 \ \text{ on } \Gamma, \end{cases} \tag{71}$$

$$\bar{\varphi} + \frac{\partial L}{\partial u}(x, \bar{y}, \bar{u}) + \bar{\lambda} = 0 \text{ in } \ \Omega, \tag{72}$$

$$\bar{\lambda} = \max\{0, \bar{\lambda} + c(\bar{u} - \beta)\} + \min\{0, \bar{\lambda} + c(\bar{u} - \alpha)\}. \tag{73}$$

This is a nonlinear system. We are going to apply Newton's method to solve it. Notice that the nonlinearities appearing in (70) and (71) are smooth, while the $\max(0, z)$ and $\min(0, z)$ functions appearing in (73) are not. To deal with them, we introduce the concept of slantly differentiable function in the sense stated in [18, 25]. For an alternative approach involving semismoothness with respect to Clarke's generalized differential see [27]. In the book [29], the notion of semismooth Newton differentiability is used. In our case, all these approaches are equivalent.

Definition 47 Let X and Y be Banach spaces and consider a function $F : D \subset X \to Y$, where $D \subset X$ is open. We will say that F is slantly differentiable in D if there

exists a family of mappings $M : D \to \mathscr{L}(X, Y)$ such that

$$\lim_{h\to 0} \frac{F(x+h) - F(x) - M(x+h)h}{\|h\|} = 0 \text{ for all } x \in D.$$

A family of mappings M satisfying this condition is called a slant derivative of F. It need not be unique.

Given a slantly differentiable function F with a slant derivative M, to solve the equation $F(x) = 0$ we may apply Newton's method:

1. choose $x_0 \in D$ and set $k = 0$;
2. solve $M(x_k)(x_{k+1} - x_k) = -F(x_k)$;
3. set $k = k + 1$ and go to 2.

Theorem 48 ([18, 25]) *Let $F : D \subset X \to Y$ be a continuous and slantly differentiable function with slant derivative M, and let $\bar{x} \in D$ be a solution of $F(x) = 0$. Suppose further that $M(x)$ is nonsingular for all $x \in D$ and $\|M(x)^{-1}\|$ is uniformly bounded in U. Then there exists $\delta > 0$ such that for $\|x_0 - x\| < \delta$, the sequence $\{x_k\}$ generated by Newton's method converges superlinearly to $\bar{x}$.*

Any function that is differentiable in an open set is slantly differentiable and the family of differentials is a slant derivative, which is unique in this case. We have that $t \in \mathbb{R} \to \max\{0, t\} \in \mathbb{R}$ is slantly differentiable and a slant derivative of it is given by $M(t) = 1$ if $t > 0$, $M(t) = 0$ if $t \le 0$. It is known that $v \in L^2(\Omega) \mapsto \max\{0, v\} \in L^2(\Omega)$ is not slantly differentiable, but for every $q > 2$, $v \in L^q(\Omega) \mapsto \max\{0, v\} \in L^2(\Omega)$ is; see [25, Appendix A]. Since the optimal control is a Lipschitz function, there is no problem in choosing some $q > 2$ and to define

$$F : H_0^1(\Omega) \times H_0^1(\Omega) \times L^q(\Omega) \times L^q(\Omega) \to H^{-1}(\Omega) \times H^{-1}(\Omega) \times L^2(\Omega) \times L^2(\Omega)$$

by

$$F\begin{pmatrix} y \\ \varphi \\ u \\ \lambda \end{pmatrix} = \begin{pmatrix} Ay + \phi(y) - u \\ A^*\varphi + \phi'(y)\varphi - \dfrac{\partial L}{\partial y}(\cdot, y, u) \\ \varphi + \dfrac{\partial L}{\partial u}(\cdot, y, u) + \lambda \\ \lambda - \max\{0, \lambda + c(u - \beta)\} - \min\{0, \lambda + c(u - \alpha)\} \end{pmatrix}.$$

By the use of the chain rule [27], to compute the slant derivative of $\max\{0, \lambda + c(u - \beta)\}$ and of $\min\{0, \lambda + c(u - \alpha)\}$, all we need to know are the so called active and inactive sets. For fixed $c > 0$, given (y, u, φ, λ), let us define the active and inactive sets related to it as

$$A(y, u, \varphi, \lambda) = A_\alpha(y, u, \varphi, \lambda) \cup A^\beta(y, u, \varphi, \lambda),$$
$$I(y, u, \varphi, \lambda) = \Omega \setminus A(y, u, \varphi, \lambda),$$

where

$$A^{\beta}(y,u,\varphi,\lambda) = \{x \in \Omega : \lambda + c(u-\beta) > 0\},$$
$$A_{\alpha}(y,u,\varphi,\lambda) = \{x \in \Omega : \lambda + c(u-\alpha) < 0\}.$$

Sometimes we will abuse notation and just write A^{β}, A_{α}, A and I. So for given $(y,\varphi,u,\lambda) \in H_0^1(\Omega) \times H_0^1(\Omega) \times L^q(\Omega) \times L^q(\Omega)$, a slant derivative of F applied to $(z,\zeta,v,\mu) \in H_0^1(\Omega) \times H_0^1(\Omega) \times L^q(\Omega) \times L^q(\Omega)$ is given by

$$M \begin{pmatrix} y \\ \varphi \\ u \\ \lambda \end{pmatrix} \begin{pmatrix} z \\ \zeta \\ v \\ \mu \end{pmatrix} = \begin{pmatrix} Az + \phi'(y)z - v \\ A^*\zeta + \phi''(y)z\varphi + \phi'(y)\zeta - \dfrac{\partial^2 L}{\partial y^2}(\cdot,y,u)z - \dfrac{\partial^2 L}{\partial y \partial u}(\cdot,y,u)v \\ \zeta + \dfrac{\partial^2 L}{\partial y \partial u}(\cdot,y,u)z + \dfrac{\partial^2 L}{\partial u^2}(\cdot,y,u)v + \mu \\ \mu\chi_I - cv\chi_A \end{pmatrix}$$

where χ_A and χ_I denote the characteristic functions of A and I, respectively. After some simplification, Newton's iteration reads like

$$\begin{cases} Ay_{k+1} + \phi'(y_k)y_{k+1} = u_{k+1} + \phi'(y_k)y_k - \phi(y_k) \ \text{ in } \Omega, \\ y_{k+1} = 0 \ \text{ on } \Gamma, \end{cases} \tag{74}$$

$$\begin{cases} A^*\varphi_{k+1} + \phi'(y_k)\varphi_{k+1} = \left[\dfrac{\partial^2 L}{\partial y^2}(x,y_k,u_k) - \varphi_k\phi''(y_k)\right](y_{k+1}-y_k) \\ \qquad\qquad + \dfrac{\partial L}{\partial y}(x,y_k,u_k) + \dfrac{\partial^2 L}{\partial y \partial u}(x,y_k,u_k)(u_{k+1}-u_k) \ \text{ in } \Omega, \\ \varphi_{k+1} = 0 \ \text{ on } \Gamma, \end{cases} \tag{75}$$

$$\varphi_{k+1} + \frac{\partial^2 L}{\partial y \partial u}(x,y_k,u_k)(y_{k+1}-y_k) + \frac{\partial^2 L}{\partial u^2}(x,y_k,u_k)(u_{k+1}-u_k) + \frac{\partial L}{\partial u}(x,y_k,u_k) + \lambda_{k+1} = 0 \text{ in } \Omega \tag{76}$$

$$\lambda_{k+1} = 0 \text{ in } I_k, \ u_{k+1} = \beta \text{ in } A_k^{\beta}, \ u_{k+1} = \alpha \text{ in } A_{\alpha}^k \tag{77}$$

where $A_k^{\beta} = A^{\beta}(y_k,u_k,\varphi_k,\lambda_k)$, $A_{\alpha}^k = A_{\alpha}(y_k,u_k,\varphi_k,\lambda_k)$, and $I_k = I(y_k,u_k,\varphi_k,\lambda_k)$.

Now, one possibility is to set up a discrete approximation of this system and to solve the resulting linear system. This usually leads to a very large scale problem; see e.g. [19]. Instead, we are going to write it as an unconstrained reduced quadratic program in the inactive part of the control variable.

First, we notice that we can write $u_{k+1} = u_{k+1}\chi_{I_k} + \beta\chi_{A_k^\beta} + \alpha\chi_{A_\alpha^k}$ and define $y_{0k} \in W^{2,p}(\Omega) \cap H_0^1(\Omega)$ the solution of

$$\begin{cases} Ay_{0k} + \phi'(y_k)y_{0k} = \beta\chi_{A_k^\beta} + \alpha\chi_{A_\alpha^k} + \phi'(y_k)y_k - \phi(y_k) \text{ in } \Omega \\ y_{0k} = 0 \text{ on } \Gamma. \end{cases} \tag{78}$$

Next, for $v \in L^2(I_k)$ we introduce the linearized state $z_v^k \in W^{2,p}(\Omega) \cap H_0^1(\Omega)$, the solution of

$$\begin{cases} Az_v^k + \phi'(y_k)z_v^k = v \text{ in } \Omega \\ z_v^k = 0 \text{ on } \Gamma, \end{cases} \tag{79}$$

where, abusing notation, we extend the functions in $L^2(I_k)$ by zero to $A_k = \Omega \setminus I_k$. We have that $y_{k+1} = z^k_{u_{k+1}\chi_{I_k}} + y_{0k}$ and the following relations are satisfied:

$$\begin{cases} Az^k_{u_{k+1}\chi_{I_k}} + \phi'(y_k)z^k_{u_{k+1}\chi_{I_k}} = u_{k+1}\chi_{I_k} \text{ in } \Omega \\ z^k_{u_{k+1}\chi_{I_k}} = 0 \text{ on } \Gamma \end{cases} \tag{80}$$

$$\begin{cases} A^*\varphi_{k+1} + \phi'(y_k)\varphi_{k+1} = \left[\dfrac{\partial^2 L}{\partial y^2}(x, y_k, u_k) - \varphi_k\phi''(y_k)\right](z^k_{u_{k+1}\chi_{I_k}} - (y_k - y_{0k})) \\ \qquad\qquad + \dfrac{\partial L}{\partial y}(x, y_k, u_k) \\ \qquad\qquad + \partial^2_{yu}L(x, y_k, u_k)(u_{k+1}\chi_{I_k} - (u_k - \beta\chi_{A_k^\beta} - \alpha\chi_{A_\alpha^k})) \text{ in } \Omega, \\ \varphi_{k+1} = 0 \text{ on } \Gamma, \end{cases} \tag{81}$$

$$\varphi_{k+1} + \frac{\partial^2 L}{\partial y \partial u}(x, y_k, u_k)(z^k_{u_{k+1}\chi_{I_k}} - (y_k - y_{0k})) + \frac{\partial^2 L}{\partial u^2}(x, y_k, u_k)(u_{k+1}\chi_{I_k} - u_k)$$
$$+ \frac{\partial L}{\partial u}(x, y_k, u_k) = 0 \text{ in } I_k. \tag{82}$$

This is the optimality system of the unconstrained linear-quadratic optimal control problem

$$(\mathrm{P}_k) \quad \text{Min } \{J_k(v) : v \in L^2(I_k)\}, \tag{83}$$

where

$$J_k(v) = \frac{1}{2}\int_\Omega \left[\frac{\partial^2 L}{\partial y^2}(x, y_k, u_k) - \varphi_k \phi''(y_k)\right](z_v^k - (y_k - y_{0k}))^2\, dx$$
$$+ \int_\Omega \frac{\partial L}{\partial y}(x, y_k, u_k) z_v^k\, dx$$
$$+ \int_\Omega \frac{\partial^2 L}{\partial y \partial u}(x, y_k, u_k)(v - (u_k - \beta \chi_{A_k^\beta} - \alpha \chi_{A_\alpha^k})) z_v^k\, dx$$
$$+ \int_{I_k} \left[\frac{\partial^2 L}{\partial y \partial u}(x, y_k, u_k)(z_v^k - (y_k - y_{0k})) + \frac{\partial L}{\partial u}(x, y_k, u_k)\right] v\, dx$$
$$+ \frac{1}{2}\int_{I_k} \frac{\partial^2 L}{\partial u^2}(x, y_k, u_k)(v - u_k)^2\, dx.$$

Remark 49 In the existing literature, see e.g., [47, Corollary 4.3], it is assumed that $J''(\bar{u})v^2 \geq \delta \|v\|^2_{L^2(\Omega)}$ for all $v \in L^2(\Omega)$. Then, for u_k close enough to $\bar{u}$, it is proved J_k is a strictly convex functional and hence (P_k) has a unique solution.

A weaker assumption can be formulated, if we assume that $\bar{u}$ satisfies the sufficient second order condition (19) and the strict complementarity condition $|\bar{d}(x)| > 0$ for a.a. x in the active set $\{\bar{u}(x) = \alpha \text{ or } \bar{u}(x) = \beta\}$. Indeed, taking into account the continuity property of the second derivative stated in Remark 18 and the equivalence result given in Theorem 25, it can be proved that if u_k is close enough to $\bar{u}$ and I_k is close enough to the active set, then J_k is a strictly convex functional and hence (P_k) has a unique solution.

Finally, let us describe with some detail how to solve (P_k). We use the linear solution operator of Eq. (79)

$$S_k : L^2(I_k) \to L^2(\Omega)$$
$$v \mapsto z_v^k$$

and its adjoint operator, given by

$$S_k^* : L^2(\Omega) \to L^2(I_k)$$
$$z \mapsto \zeta_{|I_k},$$

where $\zeta \in H_0^1(\Omega)$ is the unique solution of

$$\begin{cases} A^*\zeta + \phi'(y_k)\zeta = z \text{ in } \Omega, \\ \zeta = 0 \text{ on } \Gamma. \end{cases} \tag{84}$$

Using integration by parts we have that

$$(z, S_k v)_\Omega = (S_k^* z, v)_{I_k} \ \forall\ z \in L^2(\Omega) \text{ and } \forall v \in L^2(I_k),$$

and we can write

$$J_k(v) = \frac{1}{2}(\mathscr{A}_k v, v)_{I_k} - (b_k, v)_{I_k} + C$$

where C is independent of v, $\mathscr{A}_k \in \mathscr{L}(L^2(I_k))$ is given by

$$\begin{aligned}\mathscr{A}_k = S_k^*\big([\frac{\partial^2 L}{\partial y^2}(\cdot, y_k, u_k) - \varphi_k \phi''(y_k)]S_k\big)\\ + 2S_k^* \frac{\partial^2 L}{\partial y \partial u}(\cdot, y_k, u_k)\mathbb{I}_k + \frac{\partial^2 L}{\partial u^2}(\cdot, y_k, u_k)\mathbb{I}_k,\end{aligned}$$

where $\mathbb{I}_k$ is the identity operator in $L^2(I_k)$, and $b_k \in L^2(I_k)$ is

$$\begin{aligned}b_k = -S_k^*\big[\frac{\partial L}{\partial y}(\cdot, y_k, u_k) - \frac{\partial^2 L}{\partial y \partial u}L(\cdot, y_k, u_k)(u_k - \beta\chi_{A_k^\beta} - \alpha\chi_{A_\alpha^k})\\ -\frac{\partial^2 L}{\partial y^2}(\cdot, y_k, u_k)(y_k - y_{0k}) + \varphi_k\phi''(y_k)(y_k - y_{0k})\big]\\ -\frac{\partial L}{\partial u}(\cdot, y_k, u_k) + \frac{\partial^2 L}{\partial y \partial u}(\cdot, y_k, u_k)(y_k - y_{0k}) + \frac{\partial^2 L}{\partial u^2}(\cdot, y_k, u_k)u_k.\end{aligned}$$

At each step of Newton's method we have to solve the unconstrained quadratic program

$$\min_{v \in L^2(I_k)} \frac{1}{2}(\mathscr{A}_k v, v)_{I_k} - (b_k, v)_{I_k}.$$

Under the assumptions of Remark 49, $\mathscr{A}_k$ is a self-adjoint, positive definite operator, so the quadratic problem can be solved using the conjugate gradient method. We cannot compute $\mathscr{A}_k$ explicitly, but at each step, for a given descent direction d, we can compute $\mathscr{A}_k d$ just solving two partial differential equations, namely (79) and (84). In practice, of course, we compute appropriate FEM discretizations $S_{k,h}$ and $S_{k,h}^*$ of the operators.

An alternative way of addressing this problem consists in replacing the discretized version of $\mathscr{A}_k$ by a BGFS quasi-Newton approximation. If the size of the problem is very big, a limited-memory BGFS quasi-Newton method may be used. This may be necessary if J_k is not convex or u_k is not close enough to $\bar{u}$.

15 An Example

We will consider the tracking functional with Tikhonov regularization

$$L(x, y, u) = \frac{1}{2}(y - y_d(x))^2 + \frac{N}{2}u^2,$$

We can simplify the above expression for $\mathscr{A}_k$ and b_k and obtain

$$\mathscr{A}_k = S_k^*\big([1 - \varphi_k\phi''(y_k)]S_k\big) + N\mathbb{I}_k \text{ and } b_k = -S_k^*\big(y_{0k} - y_d + \varphi_k\phi''(y_k)(y_k - y_{0k})\big).$$

To build an example with explicit known optimal control, we simply define $\bar{\varphi} \in W^{2,p}(\Omega) \cap H_0^1(\Omega)$ and compute $\bar{u}$ using the projection formula of Remark 22. If we define $y_d = \Delta\bar{\varphi} + \bar{y} - \bar{\varphi}\phi'(\bar{y})$, we have that $(\bar{y}, \bar{u}, \bar{\varphi})$ satisfies the first order optimality conditions for (P). Since we cannot compute $\bar{y}$ explicitly, the FEM approximation $y_h(\bar{u})$ will be used in the computations when necessary.

For our example we have set $\Omega = B(0, 1) \subset \mathbb{R}^n$, $n = 2$ or $n = 3$, $\alpha = -0.5$, $\beta = 0.5$, $N = 1$, $\phi(y) = |y|^3 y$ and $\bar{\varphi} = -1 + |x|^2$. Notice that $\Delta\bar{\varphi} = 2n$.

We approximate both the control and the state by continuous piecewise linear elements and solve the finite element approximations of the problem using the semismooth Newton method. We select the parameter $c = 1$ in all cases. The method stops when the difference between two iterations, measured in $L^2(\Omega)$ is smaller than 5×10^{-15} for the control, state and the adjoint state variables. In all the experiments, the method terminated after about 10 iterations, independently of the mesh size.

We report on the error on the $L^2(\Omega)$ norm of the control. At each level j, the mesh is obtained by regular dyadic refinement from the previous mesh (this must be done very carefully in the case of dimension $n = 3$ in order to obtain a quasi-uniform family of triangulations of the unit ball). In this way, we have that $h \sim 2^{-j}$.

In our problem (64) holds, and using (65) we expect an order of convergence of 1.5. We have computed the experimental order of convergence (EOC) using a linear regression of the loglog graph of the mesh size vs. the error. In dimension 2, we obtain the order 1.56 and in dimension 3, we obtain 1.62. These values are quite in agreement with the theoretical value 1.5. In Tables 1 and 2 we show the mesh data and the errors for the 2D and the 3D problems respectively. The optimal value of the discrete problem is also included for reference and possible double-check. We show loglog graphs of the results as well as the linear regression lines together with lines with slope 1.5. Results for dimension 2 are shown in Fig. 1 and for dimension 3 in Fig. 2.

Table 1 Results for dimension 2

j	♯cells	♯nodes	$\|\bar{u} - \bar{u}_h\|_{L^2(\Omega_h)}$	$J_h(\bar{u}_h)$
1	32	25	2.15E−01	24.6762
2	128	81	3.09E−02	25.2333
3	512	289	1.46E−02	25.3645
4	2048	1089	7.06E−03	25.3988
5	8192	4225	1.36E−03	25.4080
6	32,768	16,641	6.59E−04	25.4101
7	131,072	66,049	2.00E−04	25.4107
8	524,288	263,169	8.38E−05	25.4108
EOC			1.56	

Table 2 Results for dimension 3

j	♯cells	♯nodes	$\|\bar{u} - \bar{u}_h\|_{L^2(\Omega_h)}$	$J_h(\bar{u}_h)$
1	160	55	3.36 E−01	65.9926
2	1280	309	5.91E−02	73.1047
3	10,240	2057	2.71E−02	75.0302
4	81,920	14,993	9.21E−03	75.5253
5	655,360	114,465	3.11E−03	75.6498
EOC			1.62	

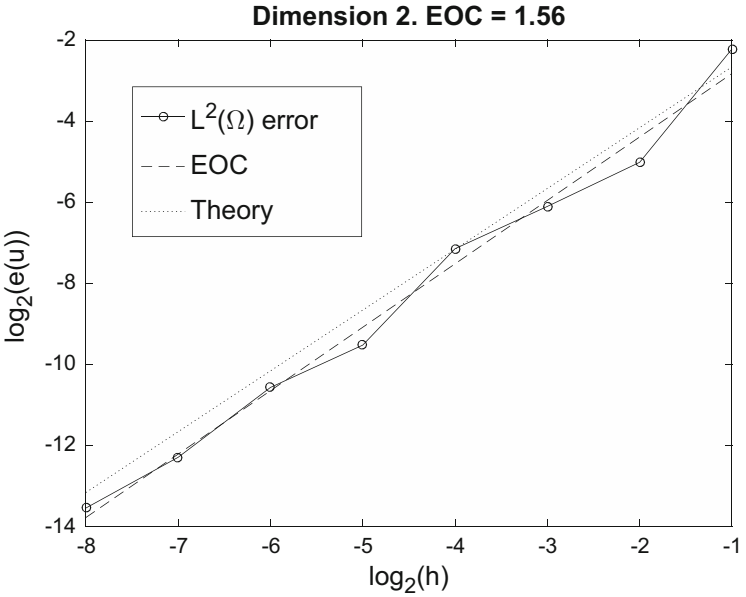


Fig. 1 Results for dimension 2

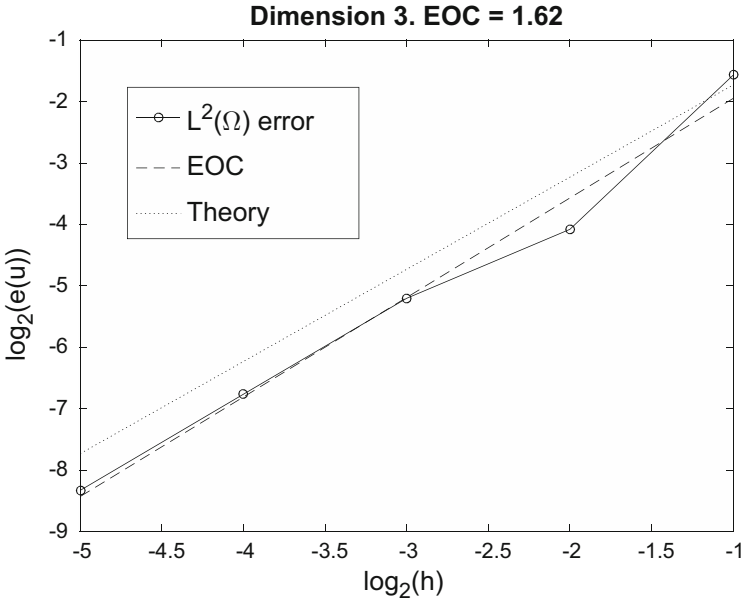


Fig. 2 Results for dimension 3

Acknowledgements The authors were partially supported by Ministerio Español de Economía y Competitividad under research project MTM2014-57531-P

References

1. Apel, T., Rösch, A., Winkler, G.: Optimal control in non-convex domains: a priori discretization error estimates. Calcolo **44**(3), 137–158 (2007)
2. Arada, N., Casas, E., Tröltzsch, F.: Error estimates for the numerical approximation of a semilinear elliptic control problem. Comput. Optim. Appl. **23**(2), 201–229 (2002)
3. Casas, E.: Control of an elliptic problem with pointwise state constraints. SIAM J. Control. Optim. **24**(6), 1309–1318 (1986)
4. Casas, E.: Optimality conditions and numerical approximations for some optimal design problems. Control. Cybern. **19**(3–4), 73–91 (1990)
5. Casas, E.: Optimal control in coefficients with state constraints. Appl. Math. Optim. **26**, 21–37 (1992)
6. Casas, E.: Pontryagin's principle for state-constrained boundary control problems of semilinear parabolic equations. SIAM J. Control. Optim. **35**(4), 1297–1327 (1997)
7. Casas, E.: Error estimates for the numerical approximation of semilinear elliptic control problems with finitely many state constraints. ESAIM: Control Optim. Calc. Var. **8**, 345–374 (2002)
8. Casas, E.: Using piecewise linear functions in the numerical approximation of semilinear elliptic control problems. Adv. Comput. Math. **26**, 137–153 (2007)
9. Casas, E.: Second order analysis for bang-bang control problems of PDEs. SIAM J. Control. Optim. **50**(4), 2355–2372 (2012)

10. Casas, E., Dhamo, V.: Error estimates for the numerical approximation of Neumann control problems governed by a class of quasilinear elliptic equations. Comput. Optim. Appl. **52**, 719–756 (2012)
11. Casas, E., Mateos, M.: Uniform convergence of the FEM. Applications to state constrained control problems. Comput. Appl. Math. **21**(1), 67–100 (2002)
12. Casas, E., Mateos, M.: Error estimates for the numerical approximation of Neumann control problems. Comput. Optim. Appl. **39**, 265–295 (2008)
13. Casas, E., Raymond, J.-P.: Error estimates for the numerical approximation of Dirichlet boundary control for semilinear elliptic equations. SIAM J. Control. Optim. **45**(5), 1586–1611 (2006)
14. Casas, E., Tröltzsch, F.: Numerical analysis of some optimal control problems governed by a class of quasilinear elliptic equations. ESAIM: Control Optim. Calc. Var. **17**, 771–800 (2010)
15. Casas, E., Tröltzsch, F.: Second order analysis for optimal control problems: improving results expected from abstract theory. SIAM J. Optim. **22**(1), 261–279 (2012)
16. Casas, E., Mateos, M., Tröltzsch, F.: Error estimates for the numerical approximation of boundary semilinear elliptic control problems. Comput. Optim. Appl. **31**, 193–219 (2005)
17. Casas, E., Mateos, M., Vexler, B.: New regularity results and improved error estimates for optimal control problems with state constraints. ESAIM: Control Optim. Calc. Var. **20**(3), 803–822 (2014)
18. Chen, X., Nashed, Z., Qi, L.: Smoothing methods and semismooth methods for nondifferentiable operator equations. SIAM J. Numer. Anal. **38**(4), 1200–1216 (2000). (Electronic), MR 1786137 (2001h:65064)
19. De los Reyes, J.C.: Numerical PDE-Constrained Optimization. Springer Briefs in Optimization. Springer, Cham (2015). MR 3308473
20. Deckelnick, K., Hinze, M.: Convergence of a finite element approximation to a state constraint elliptic control problem. SIAM J. Numer. Anal. **45**(5), 1937–1953 (2007)
21. Dunn, J.C.: On second order sufficient optimality conditions for structured nonlinear programs in infinite-dimensional function spaces. In: Fiacco, A. (ed.) Mathematical Programming with Data Perturbations, pp. 83–107. Marcel Dekker, New York (1998)
22. Ekeland, I., Temam, R.: Analyse convexe et problèmes variationnels. Dunod-Gauthier Villars, Paris (1974)
23. Fattorini, H.O.: Infinite Dimensional Optimization and Control Theory. Cambridge University Press, New York (1998)
24. Grisvard, P.: Elliptic Problems in Nonsmooth Domains. Pitman, Boston, London, Melbourne (1985)
25. Hintermüller, M., Ito, K., Kunisch, K.: The primal-dual active set strategy as a semismooth Newton method. SIAM J. Optim. **13**(3), 865–888 (2003). MR 1972219 (2004b:90123)
26. Hinze, M.: A variational discretization concept in control constrained optimization: the linear-quadratic case. Comput. Optim. Appl. **30**, 45–61 (2005)
27. Hinze, M., Pinnau, R., Ulbrich, M., Ulbrich, S.: Optimization with PDE Constraints. Mathematical Modelling: Theory and Applications, vol. 23. Springer, New York (2009)
28. Ioffe, A.D.: Necessary and sufficient conditions for a local minimum. III. Second order conditions and augmented duality. SIAM J. Control. Optim. **17**(2), 266–288 (1979). MR 525027 (82j:49005c)
29. Ito, K., Kunisch, K.: Lagrange Multiplier Approach to Variational Problems and Applications. Advances in Design and Control, vol. 15. Society for Industrial and Applied Mathematics (SIAM), Philadelphia (2008)
30. Krumbiegel, K., Pfefferer, J.: Superconvergence for Neumann boundary control problems governed by semilinear elliptic equations. Comput. Optim. Appl. **61**, 373–408 (2015)
31. Lee, E.B., Marcus, L.: Foundations of Optimal Control Theory. Wiley, New York (1967)
32. Li, X., Yong, J.: Optimal Control Theory for Infinite Dimensional Systems. Birkhäuser, Boston (1995)
33. Lions, J.L.: Contrôle optimal de systèmes gouvernés par des equations aux dérivées partielles. Dunod, Paris (1968)

34. Meidner, D., Vexler, B.: A priori error estimates for the space-time finite element discretization of parabolic optimal control problems. Part I: problems without control constraints. SIAM J. Control. Optim. **47**(3), 1150–1177 (2008)
35. Meidner, D., Vexler, B.: A priori error estimates for the space-time finite element discretization of parabolic optimal control problems. Part II: problems with control constraints. SIAM J. Control. Optim. **47**(3), 1301–1329 (2008)
36. Meidner, D., Rannacher, R., Vexler, B.: A priori error estimates for finite element discretizations of parabolic optimization problems with pointwise state constraints in time. SIAM J. Control. Optim. **49**(5), 1961–1997 (2011)
37. Meyer, C., Rösch, A.: Superconvergence properties of optimal control problems. SIAM J. Control. Optim. **43**(3), 970–985 (2004)
38. Neittaanmaki, P., Sprekels, J., Tiba, D.: Optimization of Elliptic Systems. Springer Monographs in Mathematics. Springer, New York (2006)
39. Neitzel, I., Vexler, B.: A priori error estimates for space-time finite element discretization of semilinear parabolic optimal control problems. Numer. Math. **120**, 345–386 (2012)
40. Pedregal, P.: Parametrized Measures and Variational Principles. Progress in Nonlinear Differential Equations and Their Applications. Birkhäuser Verlag, Basel (1997)
41. Pontriaguine, L., Boltianski, V., Gamkrélidzé, R., Michtchenko, E.: Théorie mathématique des processus optimaux. Editions MIR, Moscou (1974)
42. Raviart, P.A., Thomas, J.M.: Introduction à l'analyse numérique des equations aux dérivées partielles. Masson, Paris (1983)
43. Roubíček, T.: Relaxation in Optimization Theory and Variational Calculus. Walter de Gruyter, Berlin (1997)
44. Schatz, A.H.: Pointwise error estimates and asymptotic error expansion inequalities for the finite element method on irregular grids: part I. Global estimates. Math. Comput. **67**(223), 877–899 (1998)
45. Schatz, A.H., Wahlbin, L.B.: On the quasi-optimality in L_∞ of the $\overset{o}{H}{}^1$ -projection into finite element spaces. Math. Comput. **38**(157), 1–22 (1982)
46. Stampacchia, G.: Le problème de Dirichlet pour les équations elliptiques du second ordre à coefficients discontinus. Ann. Inst. Fourier (Grenoble) **15**, 189–258 (1965)
47. Tröltzsch, F.: An SQP method for the optimal control of a nonlinear heat equation. Control. Cybern. **15**(1/2), 267–288 (1994)
48. Tröltzsch, F.: Optimal control of partial differential equations: theory, methods and applications. Graduate Studies in Mathematics, vol. 112. American Mathematical Society, Philadelphia (2010)
49. Warga, J.: Optimal Control of Differential and Functional Equations. Academic, New York-London (1972)
50. Young, L.C.: Lectures on the Calculus of Variations and Optimal Control Theory. W.B. Saunders, Philadelphia (1969)

Introduction to First-Principle Simulation of Molecular Systems

Eric Cancès

Abstract First-principle molecular simulation aims at computing the physical and chemical properties of a molecule, or more generally of a material system, from the fundamental laws of quantum mechanics. It is widely used in various application fields ranging from quantum chemistry to materials science and molecular biology, and is the source of many very interesting and challenging mathematical and numerical problems. This chapter is an elementary introduction to this field, covering some modeling, mathematical, and numerical aspects.

1 Introduction

This chapter contains lecture notes of a 4 h introductory course to first-principle molecular simulation, delivered in June 2016 in Gijón, on the occasion of the XVII Jacques-Louis Lions Spanish-French School on Numerical Simulation in Physics and Engineering. First-principle molecular simulation aims at computing the physical and chemical properties of a molecule, or more generally of a material system, from the fundamental laws of quantum mechanics. Its power is that it can be used in principle to compute *any* property of *any* molecule or materials from its chemical formula. Its limitations are on the one hand that approximations are required to deal with the curse of dimensionality (see Sect. 5), and on the other hand that the computational costs of the approximate models increase fast with the size and complexity of the simulated system.

First-principle molecular simulation is used by thousands of physicists, chemists, biologists, materials scientists, and nanoscientists on a daily basis. Such simulations are reported in over 20,000 scientific articles published in 2015, and are the matter of about 15% of the high-performance computing (HPC) projects funded by PRACE (Partnership for Advanced Computing in Europe) in 2016. The importance of

E. Cancès (✉)
CERMICS, Ecole des Ponts and Inria Paris, 6 & 8 avenue Blaise Pascal, 77455 Marne-la-Vallée, France
e-mail: cances@cermics.enpc.fr

M. Mateos, P. Alonso (eds.), *Computational Mathematics, Numerical Analysis and Applications*, SEMA SIMAI Springer Series 13,
DOI 10.1007/978-3-319-49631-3_2

molecular simulation for the applications was acknowledged by the 1998 and 2013 Nobel prizes in Chemistry [43, 45, 50, 64, 77].

From a mathematical point of view, first-principle molecular simulation is an extremely rich field, which gives rise to a variety of interesting modeling, mathematical analysis, and numerical problems of different natures, ranging from easy to extremely difficult. The many mathematical models encountered in this field involve linear and nonlinear partial differential equations (PDEs), optimization problems, spectral theory, stochastic processes, high-performance computing, machine learning, as well as some tools of differential geometry (Berry curvature), non-commutative geometry (C*-algebras), or algebraic topology (Chern classes). This is therefore a fantastic playground for mathematicians.

This chapter is organized as follows. In Sects. 2 and 3, we briefly present two fundamental mathematical tools, namely optimization in Hilbert spaces, and the spectral theory of self-adjoint operators, which are useful in many fields of mathematics, and are heavily relied upon in Sects. 4–7. The reader familiar with these tools can directly proceed to Sect. 4. In the latter, we introduce the (non-relativistic) quantum many-problem and the N-body Schrödinger equation, and we then apply this formalism to the special case of a molecular system in Sect. 5. In Sect. 6, we present the Hartree-Fock model, which is the simplest variational approximation of the central equation in first-principle molecular simulation, that is the N-electron Schrödinger equation. As will be seen throughout these notes, (linear and nonlinear) elliptic eigenvalue problems play a key role in this field. Section 7 is devoted to the numerical approximation of the eigenvalues of (linear) elliptic eigenvalue problems.

2 Optimization in Hilbert Spaces

It is well-known that if $J : \mathbb{R} \to \mathbb{R}$ is differentiable, the set of the local minimizers of J is included in the set $\mathscr{C} = \{x \in \mathbb{R} \mid J'(x) = 0\}$ of the critical points of J. The latter set contains all the local minimizers and maximizers of J, as well as points which are neither minimizers nor maximizers (see Fig. 1).

The purpose of this section is to extend this elementary result to unconstrained and constrained optimization problems in finite or infinite dimensional Hilbert spaces. Let us first recall some basic definitions.

In this section, V and W are two real Hilbert spaces. We denote by $(\cdot,\cdot)_V$ and $(\cdot,\cdot)_W$ the scalar products on V and W respectively, by $\|\cdot\|_V$ and $\|\cdot\|_W$ the associated norms, and by $\mathscr{B}(V,W)$ the vector space of the continuous (also called bounded) linear maps from V to W. Recall that $\mathscr{B}(V,W)$, endowed with the norm defined by

$$\|A\|_{\mathscr{B}(V,W)} := \sup_{v \in V \setminus \{0\}} \frac{\|Av\|_W}{\|v\|_v},$$

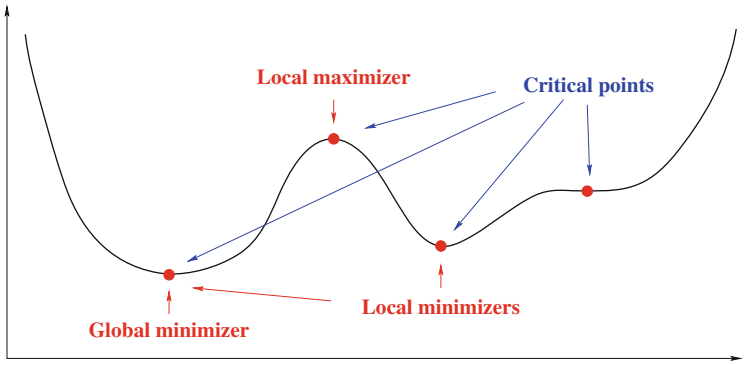


Fig. 1 Critical points of a simple differentiable function $J : \mathbb{R} \to \mathbb{R}$

is a Banach space. The adjoint of a continuous linear map $A \in \mathscr{B}(V, W)$ is the continuous linear map $A^* \in \mathscr{B}(W, V)$ characterized by

$$\forall (v, w) \in V \times W, \quad (A^* w, v)_V = (w, Av)_W.$$

The above definition makes sense by virtue of Riesz representation theorem [69, Theorem II.4].

Definition 1 Let U be an open subset of V, $F : U \to W$, and $v \in U$. The function F is called differentiable at v, if there exists $d_v F \in \mathscr{B}(V, W)$ such that in the vicinity of v,

$$F(v + h) = F(v) + d_v F(h) + o(h),$$

which means

$$\forall \varepsilon > 0, \ \exists \eta > 0 \text{ s.t. } \forall h \in V \text{ s.t. } \|h\|_V \leq \eta, \text{ we have } v + h \in U \text{ and}$$
$$\|F(v + h) - F(v) - d_v F(h)\|_W \leq \varepsilon \|h\|_V.$$

If such a linear map $d_v F$ exists, it is unique. It is called the derivative of F at v.

Definition 2 The function F is called differentiable on U if F is differentiable at each point of U. In this case, the mapping

$$dF : U \to \mathscr{B}(V, W)$$
$$v \mapsto d_v F$$

is called the derivative of F. The function F is called of class C^1 on U if dF is continuous.

Definition 3 Let U be an open subset of V and $J : U \to \mathbb{R}$ a function differentiable at $v \in U$. The unique vector of V denoted by $\nabla J(v)$ and uniquely defined[1] by

$$\forall h \in V, \qquad d_v J(h) = (\nabla J(v), h)_V,$$

is called the gradient of J at v.

Note that the above abstract definition of the gradient of agrees with the usual one when V is the space $\mathbb{R}^d$ endowed with the Euclidean scalar product:

$$\forall h \in \mathbb{R}^d, \quad J(x+h) = J(x) + \sum_{i=1}^{d} \frac{\partial J}{\partial x_i}(x)\, h_i + o(h) = J(x) + \nabla J(x) \cdot h + o(h),$$

where

$$\nabla J(x) = \begin{pmatrix} \dfrac{\partial J}{\partial x_1}(x) \\ \cdot \\ \cdot \\ \cdot \\ \dfrac{\partial J}{\partial x_d}(x) \end{pmatrix}.$$

It is important to keep in mind the geometric interpretation of the gradient. Let $J : V \to \mathbb{R}$ be a function of class C^1, $v \in V$ and $\alpha = J(v)$. If $\nabla J(v) \neq 0$, then

- in the vicinity of v, the level set

$$\mathscr{C}_\alpha := \{w \in V \,|J(w) = \alpha\}$$

  is a C^1 hypersurface (a codimension one C^1 manifold);
- the vector $\nabla J(v)$ is orthogonal to the affine hyperplane tangent to $\mathscr{C}_\alpha$ at v and points toward the steepest ascent direction.

The first-order optimality condition for smooth unconstrained optimization problems in Hilbert spaces, that is for problems consisting in minimizing some differentiable real-valued function on an open subset of a Hilbert space, is a direct extension of the basic result for the one-dimensional case recalled at the beginning of the present section.

[1] Again by Riesz representation theorem.

Theorem 4 (Optimality Condition for Unconstrained Optimization Problems) *Let $J : V \to \mathbb{R}$ be a differentiable function. The set of the local minima of J is included in the set*

$$\mathscr{C} = \{v \in V \mid d_v J = 0\} = \{v \in V \mid \nabla J(v) = 0\}$$

of the critical points of J.

The proof of this result is elementary and is left to the reader.

As a first example, consider $V = \mathbb{R}^2$, endowed with the Euclidean scalar product, and $J : \mathbb{R}^2 \to \mathbb{R}$ defined by

$$\forall \mathbf{x} = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \in \mathbb{R}^2, \quad J(\mathbf{x}) = (x_1^3 + x_2^2)\, e^{-(x_1^2+x_2^2)}. \tag{1}$$

We have

$$\nabla J(\mathbf{x}) = \begin{pmatrix} x_1 \left(3x_1 - 2x_1^3 - 2x_2^2\right) e^{-(x_1^2+x_2^2)} \\ 2x_2 \left(1 - x_1^3 - x_2^2\right) e^{-(x_1^2+x_2^2)} \end{pmatrix} = 0 \quad \Leftrightarrow \quad (x_1, x_2) = \begin{vmatrix} (0,0), \\ (\pm\sqrt{3/2}, 0), \\ (0, \pm 1), \\ (2/3, \pm\sqrt{19/27}). \end{vmatrix}$$

We can clearly see the positions of these seven critical points on the level set representation of the function J plotted on Fig. 2.

The second example is concerned with an infinite dimensional optimization problem in the Sobolev space

$$V = H^1(\mathbb{R}^d) = \left\{v \in L^2(\mathbb{R}^d) \mid \nabla v \in (L^2(\mathbb{R}^d))^d\right\},$$

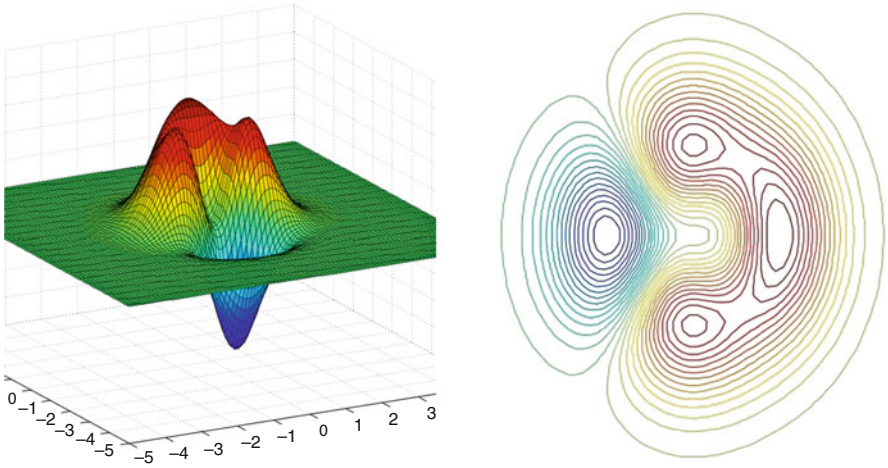


Fig. 2 Graphical representations of the function J defined by (1): 3D plot (*left*) and level sets (*right*)

endowed with its usual scalar product

$$(u,v)_{H^1} = \int_{\mathbb{R}^d} uv + \int_{\mathbb{R}^d} \nabla u \cdot \nabla v,$$

and the quadratic functional $J : H^1(\mathbb{R}^d) \to \mathbb{R}$ defined by

$$\forall v \in H^1(\mathbb{R}^d), \quad J(v) = \frac{1}{2}\int_{\mathbb{R}^d} |\nabla v|^2 + \frac{1}{2}\int_{\mathbb{R}^d} v^2 - \int_{\mathbb{R}^d} fv,$$

where f is a given function of $L^2(\mathbb{R}^d)$. To compute the derivative of J, we proceed as follows. For $v \in V$ and $h \in V$, we have

$$\begin{aligned} J(v+h) &= \frac{1}{2}\int_{\mathbb{R}^d} |\nabla(v+h)|^2 + \frac{1}{2}\int_{\mathbb{R}^d} (v+h)^2 - \int_{\mathbb{R}^d} f(v+h) \\ &= \frac{1}{2}\int_{\mathbb{R}^d} |\nabla v|^2 + \int_{\mathbb{R}^d} \nabla v \cdot \nabla h + \frac{1}{2}\int_{\mathbb{R}^d} |\nabla h|^2 + \frac{1}{2}\int_{\mathbb{R}^d} v^2 + \int_{\mathbb{R}^d} vh + \frac{1}{2}\int_{\mathbb{R}^d} h^2 \\ &\qquad - \int_{\mathbb{R}^d} fv - \int_{\mathbb{R}^d} fh \\ &= J(v) + \int_{\mathbb{R}^d} \nabla v \cdot \nabla h + \int_{\mathbb{R}^d} vh - \int_{\mathbb{R}^d} fh + \frac{1}{2}\int_{\mathbb{R}^d} |\nabla h|^2 + \frac{1}{2}\int_{\mathbb{R}^d} h^2, \end{aligned}$$

with

$$\left| \int_{\mathbb{R}^d} \nabla v \cdot \nabla h + \int_{\mathbb{R}^d} vh - \int_{\mathbb{R}^d} fh \right| \leq C_{v,f} \|h\|_{H^1},$$

and

$$\left| \frac{1}{2}\int_{\mathbb{R}^d} |\nabla h|^2 + \frac{1}{2}\int_{\mathbb{R}^d} h^2 \right| = \frac{1}{2}\|h\|_{H^1}^2 = o(h).$$

This allows one to conclude that J is differentiable at v and that

$$\forall h \in V, \quad d_v J(h) = \int_{\mathbb{R}^d} \nabla v \cdot \nabla h + \int_{\mathbb{R}^d} vh - \int_{\mathbb{R}^d} fh.$$

By definition, the gradient of J at v is the function $w \in H^1(\mathbb{R}^d)$ characterized by

$$\forall h \in V = H^1(\mathbb{R}^3), \quad (w,h)_{H^1} = d_v J(h) = \int_{\mathbb{R}^3} \nabla v \cdot \nabla h + \int_{\mathbb{R}^3} vh - \int_{\mathbb{R}^d} fh.$$

To compute $w = \nabla J(v)$, we therefore have to solve the linear elliptic problem

$$\begin{cases} \text{seek } w \in V \text{ such that} \\ \forall h \in V, \quad a(w,h) = L(h), \end{cases}$$

where

$$a(w,h) = \int_{\mathbb{R}^3} \nabla w \cdot \nabla h + \int_{\mathbb{R}^3} wh \qquad \text{and} \qquad L(h) = \int_{\mathbb{R}^3} \nabla v \cdot \nabla h + \int_{\mathbb{R}^3} vh - \int_{\mathbb{R}^3} fh,$$

or equivalently the PDE

$$\text{seek } w \in H^1(\mathbb{R}^3) \text{ such that } -\Delta w + w = -\Delta v + v - f \text{ in } \mathscr{D}'(\mathbb{R}^3),$$

where $\mathscr{D}'(\mathbb{R}^3)$ is the space of distributions in $\mathbb{R}^3$. The integral kernel of the operator $(-\Delta + 1)^{-1}$ being the Green function $G(x,y) = \frac{e^{-|x-y|}}{4\pi|x-y|}$, we therefore have

$$\nabla_v J(x) = v(x) - \int_{\mathbb{R}^3} \frac{e^{-|x-y|}}{4\pi|x-y|} f(y)\, dy.$$

Let us now turn to the more interesting case of equality constrained optimization problems. Let V and W be real Hilbert spaces such that $\dim(W) < \infty$, $J : V \to \mathbb{R}$, and $F : V \to W$. We consider the optimization problem

$$\inf_{v \in K} J(v) \quad \text{where} \quad K = \{v \in V \mid F(v) = 0\}.$$

The first-order optimality conditions for the above problem are easy to state when the constraints $F = 0$ are qualified in the following sense.

Definition 5 (Qualification of the Constraints) The equality constraints $F = 0$ are called qualified at $u \in K$ if $d_u F : V \to W$ is surjective.

We are now in position to write down the central result of this section.

Theorem 6 *Let V and W be real Hilbert spaces such that $\dim(W) < \infty$, $J : V \to \mathbb{R}$, and $F : V \to W$. Let $u \in K$ be a local minimum of J on*

$$K = \{v \in V \mid F(v) = 0\}.$$

Assume that

1. *J is differentiable at u and F is C^1 in the vicinity of u;*
2. *the equality constraint $F = 0$ is qualified at u.*

Then, there exists a unique $\lambda \in W$ such that

$$\forall h \in V,\ d_u J(h) + (\lambda, d_u F(h))_W = 0 \quad \textit{or equivalently} \quad \nabla J(u) + d_u F^*(\lambda) = 0,$$

where $d_u F^*$ is the adjoint of $d_u F$. The vector $\lambda \in W$ is called the Lagrange multiplier of the constraint $F = 0$.

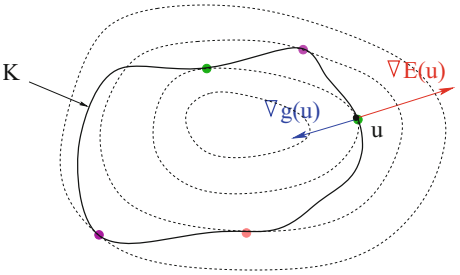


Fig. 3 Graphical illustration of Theorem 6 for $V = \mathbb{R}^2$ and $W = \mathbb{R}$. Some level sets of J are represented in *dashed closed curves*, while $K = F^{-1}(0)$ is represented by a *solid closed curve*. The five critical points of J on K are represented by *bullets*

Assume that the constraints are qualified at any point of K. The solutions of the Euler-Lagrange equations

$$\begin{cases} \text{seek } (u, \lambda) \in V \times W \text{ such that} \\ \nabla J(u) + d_u F^*(\lambda) = 0, \\ F(u) = 0, \end{cases} \tag{2}$$

are called the critical points of J on K. The set of critical points contains in particular the local minimizers and the local maximizers of J on K.

Remark 7 If $\dim(V) = d < \infty$ and $\dim(W) = m < \infty$, then the above problem consists of $(d + m)$ scalar equations with $(d + m)$ scalar unknowns.

A simple case when $V = \mathbb{R}^2$ and $W = \mathbb{R}$ is depicted on Fig. 3. On $K = F^{-1}(0) = \{v \in V \,|\, F(v) = 0\}$, the function J possesses

- two local minimizers, both global
- two local maximizers, among which the global maximizer
- one critical point which is neither a local minimizer not a local maximizer.

Sketch of the Proof of Theorem 6 Let u be a local minimizer of J on $K = F^{-1}(0) = \{v \in V \,|\, F(v) = 0\}$ and $\alpha = J(u)$. If the constraint $F = 0$ is qualified at u (i.e. if $d_u F : \mathscr{H} \to \mathscr{K}$ is surjective), then, in the vicinity of u, K is a C^1 manifold and its affine tangent subspace at u is

$$u + T_u K = u + \{h \in V \mid d_u F(h) = 0\} = u + \mathrm{Ker}(d_u F).$$

Since u is a minimizer of J on K, the vector $\nabla J(u)$ must be orthogonal to $T_u K$. Indeed, for any $h \in T_u K$, there exists a C^1 curve $\phi : [-1, 1] \to V$ drawn on K such that $\phi(0) = u$ and $\phi'(0) = h$, and we have

$$0 \leq J(\phi(t)) - J(u) = J(u + th + o(t)) - J(u) = t \nabla J(u) \cdot h + o(t).$$

Therefore, $\nabla J(u) \cdot h = 0$. In addition, it holds

$$\nabla J(u) \in (T_u K)^{\perp} = (\mathrm{Ker}(d_u F))^{\perp} = \overline{\mathrm{Ran}(d_u F^*)} = \mathrm{Ran}(d_u F^*) \text{ since } \dim(W) < \infty.$$

Therefore, there exists $\lambda \in W$ such that $\nabla J(u) + d_u F^*(\lambda) = 0$. □

Most often, Lagrange multipliers have a "physical" interpretation:

- in statistical mechanics [9], the equilibrium state of a chemical system interacting with its environment is obtained by maximizing the entropy (which is equivalent to minimizing minus the entropy) under the constraints that the energy, the volume and the concentration of chemical species are given on average: the corresponding Lagrange multipliers are respectively $1/T$, P/T and μ_i/T, where T is the temperature, P the pressure, and μ_i the chemical potential of species i;
- in fluid mechanics [25], the admissible dynamics of an incompressible fluid are the critical points of some action under the constraint that the density of the fluid remains constant ($\mathrm{div}\,(u) = 0$). The Lagrange multiplier of the incompressibility constraint is the pressure field;
- in microeconomics [66], prices are Lagrange multipliers arising in the optimization of utility functions under the constraints that some goods have limited availability.

Let us conclude this section with a result on the differentiability of functions defined by equality constrained optimization problems. Such a situation is encountered in many fields of science and engineering, and is very useful in first-principle molecular simulation to compute atomic forces (see Sects. 5 and 6) or molecular properties such as polarizabilities or hyperpolarizabilities [40]. Consider the function $W : \mathbb{R}^d \to \mathbb{R}$ defined as

$$\forall \mathbf{x} \in \mathbb{R}^d, \quad W(\mathbf{x}) = \inf \{E(\mathbf{x}, v),\ v \in V,\ F(\mathbf{x}, v) = 0\}, \tag{3}$$

where $E : \mathbb{R}^d \times V \to \mathbb{R}$, $F : \mathbb{R}^d \times V \to W$, V and W being real Hilbert spaces such that $\dim(W) < \infty$.

Assume that for each $x \in \mathbb{R}^d$, problem (3) has a unique minimizer $v(\mathbf{x})$, and that the function $\mathbf{x} \mapsto v(\mathbf{x})$ is regular. Then,

$$W(\mathbf{x}) = E(\mathbf{x}, v(\mathbf{x})) \quad \Rightarrow \quad \frac{\partial W}{\partial x_i}(\mathbf{x}) = \frac{\partial E}{\partial x_i}(\mathbf{x}, v(\mathbf{x})) + \frac{\partial E}{\partial v}(\mathbf{x}, v(\mathbf{x})) \left(\frac{\partial v}{\partial x_i}(\mathbf{x}) \right),$$

$$F(\mathbf{x}, v(\mathbf{x})) = 0 \quad \Rightarrow \quad \frac{\partial F}{\partial x_i}(\mathbf{x}, v(\mathbf{x})) + \frac{\partial F}{\partial v}(\mathbf{x}, v(\mathbf{x})) \left(\frac{\partial v}{\partial x_i}(\mathbf{x}) \right) = 0.$$

On the other hand, the Euler-Lagrange equations associated with the constrained optimization problem (3) give

$$\forall h \in V, \quad \frac{\partial E}{\partial v}(\mathbf{x}, v(\mathbf{x}))\,(h) + \left(\frac{\partial F}{\partial v}(\mathbf{x}, v(\mathbf{x}))(h), \lambda(\mathbf{x}) \right)_W = 0.$$

Therefore

$$\frac{\partial W}{\partial x_i}(\mathbf{x}) = \frac{\partial E}{\partial x_i}(\mathbf{x}, v(\mathbf{x})) + \left(\frac{\partial F}{\partial x_i}(\mathbf{x}, v(\mathbf{x})), \lambda(\mathbf{x})\right). \tag{4}$$

This formula is very important for practical purposes: it implies that it is possible to compute the derivatives of W at $\mathbf{x}$ without computing the derivatives of the minimizer $v(\mathbf{x})$. Only the state itself $v(\mathbf{x})$ and the Lagrange multiplier $\lambda(\mathbf{x})$ are necessary, and those quantities can be obtained by solving the Euler-Lagrange equations.

3 Introduction to the Spectral Theory of Self-adjoint Operators

The purpose of this section is to transpose to the case of self-adjoint operators on infinite-dimensional separable complex Hilbert spaces, the following well-known results on Hermitian[2] matrices:

1. the spectrum $\sigma(A) = \left\{z \in \mathbb{C} \,|\, (z-A) \in \mathbb{C}^{d\times d} \text{ non-invertible}\right\}$ of a Hermitian matrix $A \in \mathbb{C}^{d\times d}$ consists of the set

$$\begin{aligned}\sigma_{\mathrm{p}}(A) &= \left\{z \in \mathbb{C} \,|\, (z-A) \in \mathbb{C}^{d\times d} \text{ non-injective}\right\} \\ &= \left\{z \in \mathbb{C} \,|\, \exists \mathbf{x} \in \mathbb{C}^d \setminus \{0\} \text{ s.t. } A\mathbf{x} = z\mathbf{x}\right\}\end{aligned}$$

   of the eigenvalues of A, and $\sigma(A) \subset \mathbb{R}$;
2. any Hermitian matrix $A \in \mathbb{C}^{d\times d}$ can be diagonalized in an orthonormal basis:

$$A = \sum_{i=1}^{d} \lambda_i \mathbf{x}_i \mathbf{x}_i^*, \quad \lambda_i \in \sigma(A) \subset \mathbb{R}, \quad \mathbf{x}_i \in \mathbb{C}^d, \quad \mathbf{x}_i^* \mathbf{x}_j = \delta_{ij}, \quad A\mathbf{x}_i = \lambda_i \mathbf{x}_i. \tag{5}$$

   Here $\lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_d$ denote the d eigenvalues of A (counting multiplicities), and $(\mathbf{x}_1, \cdots, \mathbf{x}_d)$ an orthonormal basis of associated eigenvectors;
3. there exists a functional calculus for Hermitian matrices: for any Hermitian matrix A, and any $f : \mathbb{R} \to \mathbb{C}$, the matrix

$$f(A) := \sum_{i=1}^{d} f(\lambda_i) \mathbf{x}_i \mathbf{x}_i^* \tag{6}$$

[2] Recall that a matrix $A \in \mathbb{C}^{d\times d}$ is called Hermitian if $A^* = A$ (i.e. $\overline{A_{ij}} = A_{ji}$, $\forall 1 \leq i,j \leq d$). If $z \in \mathbb{C}$ and $A \in \mathbb{C}^{d\times d}$, we use the shorthand notation $z - A$ to denote the matrix $zI_d - A$, where I_d is the rank-d identity matrix. We proceed similarly with linear operators on complex Hilbert spaces.

is independent of the choice of the spectral decomposition of A, that is on the choice of the basis $(\mathbf{x}_1, \cdots, \mathbf{x}_d)$ of eigenvectors. This definition agrees with the usual definition of $f(A)$ for polynomial functions f. Indeed, if $f(\lambda) = \sum_{k=0}^{n} \alpha_k \lambda^k$, then

$$f(A) = \sum_{i=1}^{d} f(\lambda_i) \mathbf{x}_i \mathbf{x}_i^* = \sum_{i=1}^{d} \left(\sum_{k=0}^{n} \alpha_k \lambda_i^k \right) x_i x_i^* = \sum_{k=0}^{n} \alpha_k \left(\sum_{i=1}^{d} \lambda_i^k x_i x_i^* \right) = \sum_{k=0}^{n} \alpha_k A^k.$$

The strength of formula (6) is that it makes sense for any function $f : \mathbb{R} \to \mathbb{C}$, while the definition based on a polynomial expansion of f only works for polynomial functions, and in the limit, for continuous functions by virtue of Weierstrass approximation theorem.

In this section, $\mathscr{H}$ denotes a separable complex Hilbert space, $\langle \cdot | \cdot \rangle$ its scalar product, and $\| \cdot \|$ the associated norm.

3.1 Linear Operators on Hilbert Spaces

Let us first review some basic properties of *bounded* linear operators on Hilbert spaces.

Definition 8 (Bounded Linear Operator) A bounded operator on $\mathscr{H}$ is a linear map $A : \mathscr{H} \to \mathscr{H}$ such that

$$\|A\| := \sup_{u \in \mathscr{H} \setminus \{0\}} \frac{\|Au\|}{\|u\|} < \infty.$$

In other words, a bounded operator on $\mathscr{H}$ is an element of $\mathscr{B}(\mathscr{H}) := \mathscr{B}(\mathscr{H}, \mathscr{H})$.

Theorem 9 *The set $\mathscr{B}(\mathscr{H})$ of bounded operators on $\mathscr{H}$ is a non-commutative algebra and $\| \cdot \|$ is a norm on the algebra $\mathscr{B}(\mathscr{H})$:*

$$\forall (A, B) \in \mathscr{B}(\mathscr{H}) \times \mathscr{B}(\mathscr{H}), \quad \|AB\| \leq \|A\| \, \|B\|. \tag{7}$$

Endowed with the norm $\| \cdot \|$, $\mathscr{B}(\mathscr{H})$ is a Banach algebra.

The proof that $\| \cdot \|$ is a norm on $\mathscr{B}(\mathscr{H})$ is elementary, as well as the one of (7). Regarding the completeness of $\mathscr{B}(\mathscr{H})$ for the resulting topology, we refer e.g. to [69, Theorem III.2].

Note that, in view of Riesz representation theorem, a bounded linear operator B is uniquely defined by the values of the sesquilinear form $\mathscr{H} \times \mathscr{H} \ni (u, v) \mapsto \langle u | Bv \rangle \in \mathbb{C}$. This is the reason why the following definition makes sense.

Definition 10 (Adjoint of a Bounded Linear Operator) Let $A \in \mathscr{B}(\mathscr{H})$. The operator $A^* \in \mathscr{B}(\mathscr{H})$ defined by

$$\forall (u, v) \in \mathscr{H} \times \mathscr{H}, \quad \langle u | A^* v \rangle = \langle Au | v \rangle, \tag{8}$$

is called the adjoint of A. The operator A is called self-adjoint if $A^* = A$.

Endowed with its norm $\| \cdot \|$ and the $*$ operation, $\mathscr{B}(\mathscr{H})$ is in fact a C*-algebra [5]:

$$(A^*)^* = A, \quad \|A^*\| = \|A\|, \quad \text{and} \quad \|A^* A\| = \|A\|^2.$$

Many linear operators arising in quantum mechanics are not bounded operators on some Hilbert space. This is the case for instance of the kinetic energy operator, formally defined as $T = -\frac{\hbar^2}{2m}\Delta$. We therefore have to introduce the concept of (non-necessarily bounded) linear operators on Hilbert spaces.

Definition 11 (Linear Operator) A linear operator on $\mathscr{H}$ is a linear map $A : D(A) \to \mathscr{H}$, where $D(A)$ is a subspace of $\mathscr{H}$ called the domain of A.

Note that bounded linear operators are special linear operators, for which $D(A) = \mathscr{H}$ and $A : \mathscr{H} \to \mathscr{H}$ is continuous.

Definition 12 (Extensions of Operators) Let A_1 and A_2 be operators on $\mathscr{H}$. A_2 is called an extension of A_1 if $D(A_1) \subset D(A_2)$ and if $\forall u \in D(A_1), A_2 u = A_1 u$.

Definition 13 (Unbounded Linear Operator) An operator A on $\mathscr{H}$ which does not possess a bounded extension is called an unbounded operator on $\mathscr{H}$.

A possible way to extend the notion of bounded self-adjoint operator to the case of unbounded operators is the following.

Definition 14 (Symmetric Operator) A linear operator A on $\mathscr{H}$ with dense domain $D(A)$ is called symmetric if

$$\forall (u, v) \in D(A) \times D(A), \quad \langle Au | v \rangle = \langle u | Av \rangle. \tag{9}$$

Criterion (9) is simple and usually quite easy to check, but, unfortunately, symmetric operators are not very interesting. Only self-adjoint operators—which we are going to introduce—represent physical observables and have nice mathematical properties reminiscent of those of Hermitian matrices (real spectrum, spectral decomposition, functional calculus).

Definition 15 (Adjoint of a Linear Operator with Dense Domain) Let A be a linear operator on $\mathscr{H}$ with dense domain $D(A)$, and $D(A^*)$ the vector space defined as

$$D(A^*) = \{ v \in \mathscr{H} \mid \exists w_v \in \mathscr{H} \text{ s.t. } \forall u \in D(A),\ \langle Au | v \rangle = \langle u | w_v \rangle \}.$$

The linear operator A^* on $\mathscr{H}$, with domain $D(A^*)$, defined by

$$\forall v \in D(A^*), \quad A^* v = w_v,$$

(if w_v exists, it is unique since $D(A)$ is dense) is called the adjoint of A.

Note that this definition agrees with definition (8) for bounded operators.

Definition 16 (Self-adjoint Operator) A linear operator A with dense domain is called self-adjoint if $A^* = A$ (that is if A symmetric *and* $D(A^*) = D(A)$).

Any self-adjoint operator is symmetric, but the converse is not true. As mentioned previously, only self-adjoint operators have interesting mathematical properties. While it is usually easy to check that a given operator is symmetric, proving self-adjointness is not trivial and often relies on deep theorems of linear operator theory. We will not elaborate on these technicalities in these lectures notes and refer the reader to the literature [67]. We will only provide a short list of self-adjoint operators commonly encountered in first-principle molecular simulation:

- free-particle Hamiltonian (or kinetic energy operator)

$$\mathscr{H} = L^2(\mathbb{R}^d), \quad D(T) = H^2(\mathbb{R}^d), \quad \forall u \in D(T),\ Tu = -\frac{\hbar^2}{2m}\Delta u,$$

  where $m > 0$ is the mass of the particle, and $\hbar$ the reduced Planck constant;
- Schrödinger operators with confining potential $V \in C^0(\mathbb{R}^d)$ s.t. $V(x) \underset{|x|\to+\infty}{\longrightarrow} +\infty$

$$\mathscr{H} = L^2(\mathbb{R}^d), \quad D(H) = \left\{ u \in L^2(\mathbb{R}^d) \mid -\frac{\hbar^2}{2m}\Delta u + Vu \in L^2(\mathbb{R}^d) \right\}$$

$$\forall u \in D(H),\ Hu = -\frac{\hbar^2}{2m}\Delta u + Vu\,;$$

- Schödinger operators with uniformly locally L^2 potentials in dimension 3, i.e.

$$V \in L^2_{\text{unif}}(\mathbb{R}^3) := \left\{ u \in L^2_{\text{loc}}(\mathbb{R}^3) \mid \sup_{x\in\mathbb{R}^3} \int_{x+[0,1]^3} |u|^2 < \infty \right\},$$

$$\mathscr{H} = L^2(\mathbb{R}^d), \quad D(H) = H^2(\mathbb{R}^3), \quad \forall u \in D(H),\ Hu = -\frac{\hbar^2}{2m}\Delta u + Vu.$$

3.2 Spectrum

The following definition is a natural extension of the definition of the spectrum of a square matrix $A \in \mathbb{C}^{d\times d}$.

Definition 17 (Spectrum of a Linear Operator) Let A be a closed[3] linear operator on $\mathscr{H}$. Then

- the set $\rho(A) = \{z \in \mathbb{C} \mid (z - A) : D(A) \to \mathscr{H} \text{ invertible}\}$ is called the resolvent set of A;
- the set $\sigma(A) = \mathbb{C} \setminus \rho(A)$ is called the spectrum of A.

As for Hermitian matrices, the spectrum of a self-adjoint operator A is always a subset of $\mathbb{R}$. On the other hand, it does not only contains the set of the eigenvalues of A, that is the set of the complex numbers z such that $(z - A) : D(A) \to \mathscr{H}$ is *injective*. Indeed, even in the case when $D(A) = \mathscr{H}$, the linear map $(z - A)$ can be injective and not surjective since $\mathscr{H}$ is infinite dimensional.

Theorem 18 (Spectrum and Resolvent) *Let A be a closed linear operator on $\mathscr{H}$. Then*

- *the resolvent set $\rho(A)$ is an open subset of $\mathbb{C}$ and the function*

$$\rho(A) \ni z \mapsto R_z(A) := (z - A)^{-1} \in \mathscr{B}(\mathscr{H})$$

is analytic. It is called the resolvent of A. It holds

$$\forall (z, z') \in \rho(A) \times \rho(A), \quad R_z(A) - R_{z'}(A) = (z' - z) R_z(A) R_{z'}(A).$$

The above equality is called the resolvent identity;
- *the spectrum $\sigma(A)$ of A is a closed subset of $\mathbb{C}$.*

Theorem 19 (Spectrum of a Self-adjoint Operator) *Let A be a self-adjoint operator on $\mathscr{H}$. Then A is closed, $\sigma(A) \subset \mathbb{R}$, and it holds*

$$\sigma(A) = \sigma_{\mathrm{p}}(A) \cup \sigma_{\mathrm{c}}(A),$$

where $\sigma_{\mathrm{p}}(A)$ and $\sigma_{\mathrm{c}}(A)$ are respectively

- *the point spectrum of A*

$$\sigma_{\mathrm{p}}(A) = \{z \in \mathbb{C} \mid (z - A) : D(A) \to \mathscr{H} \textit{ non-injective}\} = \{\textit{eigenvalues of } A\};$$

- *the continuous spectrum of A*

$$\sigma_{\mathrm{c}}(A) = \overline{\{z \in \mathbb{C} \mid (z - A) : D(A) \to \mathscr{H} \textit{ injective but non surjective}\}}.$$

The mathematical decomposition of the spectrum of a self-adjoint operator into point and continuous spectra has an interesting physical counterpart, which will

[3] The operator A is called closed if its graph $\Gamma(A) := \{(u, Au),\ u \in D(A)\}$ is a closed subspace of $\mathscr{H} \times \mathscr{H}$.

be presented in the next section. The following alternative decomposition of the spectrum is fundamental both for theoretical and numerical purposes, as will be seen in Sect. 7.

Definition 20 Let A be a closed linear operator on $\mathscr{H}$. Then $\sigma(A) = \sigma_{\mathrm{d}}(A) \cup \sigma_{\mathrm{ess}}(A)$ where

$$\sigma_{\mathrm{d}}(A) = \{\text{isolated eigenvalues of } A \text{ with finite multiplicities}\} \quad \text{(discrete spectrum)};$$

$$\sigma_{\mathrm{ess}}(A) = \sigma(A) \setminus \sigma_{\mathrm{d}}(A) \quad \text{(essential spectrum)}.$$

The essential spectrum therefore consists of

- the continuous spectrum;
- the eigenvalues of infinite multiplicities;
- the eigenvalues embedded in the continuous spectrum.

Theorem 21 (Weyl) *Let A be a self-adjoint operator on $\mathscr{H}$ and B a symmetric operator on $\mathscr{H}$ with domain $D(A)$ such that $B(A+i)^{-1} \in \mathscr{L}(\mathscr{H})$ is compact. Then $A + B$, with $D(A+B) = D(A)$ is self-adjoint and $\sigma_{\mathrm{ess}}(A+B) = \sigma_{\mathrm{ess}}(A)$.*

Weyl theorem allows in particular to prove the following result, which covers many interesting cases arising in first-principle molecular simulation.

Corollary 22 *Let $\alpha > 0$ and $V \in L^2(\mathbb{R}^3) + L^\infty_\varepsilon(\mathbb{R}^3)$, where*

$$L^2(\mathbb{R}^3) + L^\infty_\varepsilon(\mathbb{R}^3) := \Big\{ V \in L^2_{\mathrm{loc}}(\mathbb{R}^3) \mid \forall \varepsilon > 0,\ \exists (V_2, V_\infty) \in L^2(\mathbb{R}^3) \times L^\infty(\mathbb{R}^3) \text{ such that } V = V_2 + V_\infty,\ \|V_\infty\|_{L^\infty} \leq \varepsilon \Big\}.$$

Then the operator $H = -\alpha\Delta + V$ is self-adjoint on $L^2(\mathbb{R}^3)$ with domain $H^2(\mathbb{R}^3)$ and $\sigma_{\mathrm{ess}}(H) = [0, +\infty)$.

We conclude this brief introduction to spectral theory, with the famous min-max principle, which gives a variational characterization of the discrete eigenvalues (with their multiplicities) located below the bottom of the essential spectrum of a bounded below self-adjoint operator.

Theorem 23 (Min-Max Principle, Courant-Fisher Formula) *Let A be a bounded below self-adjoint operator on $\mathscr{H}$, $Q(A)$ its form domain,[4] and a its associated quadratic form. For each $j \in \mathbb{N}^*$, we define*

$$\lambda_j(A) = \inf_{W_j \in \mathscr{E}_j} \sup_{w \in W_j \setminus \{0\}} \frac{a(w,w)}{\|w\|^2},$$

[4] Since A is bounded below, there exists $C \in \mathbb{R}$ s.t. $(u,v)_{Q(A)} := \langle u|Av\rangle + C\langle u|v\rangle$ is a scalar product on $D(A)$. The Cauchy closure of $D(A)$ for the associated norm is a Hilbert space, independent of C, called the form domain of A. The quadratic form associated with A is the unique continuous extension of $(u,v) \mapsto \langle u|Av\rangle$ to $Q(A)$.

where $\mathscr{E}_j$ is the set of the j-dimensional subspaces of $Q(A)$. Then,

- *if A has at least j eigenvalues lower than $\min\sigma_{\rm ess}(A)$ (taking multiplicities into account), then $\lambda_j(A)$ is the smallest jth eigenvalue of A;*
- *otherwise, $\lambda_j(A) = \min\sigma_{\rm ess}(A)$.*

4 The Quantum Many-Body Problem

According to the first principles of quantum mechanics, an isolated quantum system is described by

- a state space $\mathscr{H}$ (a complex Hilbert space);
- a Hamiltonian H (a self-adjoint operator on $\mathscr{H}$);
- other observables (i.e. self-adjoint operators on $\mathscr{H}$) allowing one to connect theory and experiments.

The state[5] of the system at time t is completely characterized by a wavefunction $\Psi(t) \in \mathscr{H}$ such that $\|\Psi(t)\|_{\mathscr{H}} = 1$. Its dynamics is governed by the time-dependent Schrödinger equation

$$i\hbar\frac{d\Psi}{dt}(t) = H\Psi(t), \tag{10}$$

where we recall that $\hbar$ is the reduced Planck constant. The steady states are by definition states of the form $\Psi(t) = f(t)\psi$, where $f(t) \in \mathbb{C}$ and $\psi \in \mathscr{H}$. Inserting the Ansatz $\Psi(t) = f(t)\psi$ in (10) and separating the variables, we obtain that the function f is just a physically irrelevant phase factor[6]: $f(t) = e^{-iEt/\hbar}$, with $E \in \mathbb{R}$ is homogenous to an energy. The function ψ satisfies the time-independent Schrödinger equation

$$H\psi = E\psi, \quad \|\psi\|_{\mathscr{H}} = 1.$$

The energy E is therefore an eigenvalue of the Hamiltonian H and ψ an associated normalized eigenvector.

[5] We limit ourselves to pure states in these lectures notes.

[6] It may seem weird that steady states explicitly depend on time. This apparent paradox is due to the fact that a state is in fact an element of the projective space $(\mathscr{H} \setminus \{0\})/\mathbb{C}^*$, so that $f(t)\psi$ and ψ actually represent the exact same state.

4.1 One-Particle Systems

The above formalism is completely general, and valid for any isolated quantum system. Let us now deal with specific systems of physical interest, starting with a very simple one: a spinless particle of mass m subjected to an external potential V_{ext}. In this case, the state space is $\mathscr{H} = L^2(\mathbb{R}^3, \mathbb{C})$ and the Hamiltonian

$$H = -\frac{\hbar^2}{2m}\Delta + V_{\text{ext}},$$

which, under assumptions on V_{ext} (see some examples in Sect. 3.1), is a self-adjoint operator on $\mathscr{H}$. In the so-called position representation, the wavefunction has a clear physical meaning: $|\Psi(t, \mathbf{r})|^2$ is the probability density of observing the particle at point $\mathbf{r}$ at time t. Note that it follows from the normalization condition that

$$\int_{\mathbb{R}^3} |\Psi(t, \mathbf{r})|^2 \, d\mathbf{r} = \|\Psi(t)\|^2_{\mathscr{H}} = 1.$$

The time-dependent Schrödinger equation then takes the form of a partial differential equation (PDE):

$$i\hbar \frac{\partial \Psi}{\partial t}(t, \mathbf{r}) = -\frac{\hbar^2}{2m}\Delta\Psi(t, \mathbf{r}) + V_{\text{ext}}(\mathbf{r})\Psi(t, \mathbf{r}).$$

Likewise, the time-independent Schrödinger equation reads in this case as an elliptic linear eigenvalue problem:

$$-\frac{\hbar^2}{2m}\Delta\psi(\mathbf{r}) + V_{\text{ext}}(\mathbf{r})\psi(\mathbf{r}) = E\psi(\mathbf{r}).$$

The spectrum of H strongly depends on V_{ext} (see Sect. 3.2). The spectrum of the free Hamiltonian ($V_{\text{ext}} = 0$) is purely continuous and equal to $\mathbb{R}_+$. For confining potentials, the spectrum of H is purely discrete and consists of an increasing sequence of real eigenvalues of finite multiplicities going to $+\infty$. For one-particle or mean-field Hamiltonians usually encountered in first-principle models of finite molecular systems, the potential V_{ext} vanishes at infinity, $\sigma_{\text{ess}}(H) = \mathbb{R}_+$, and $\sigma_{\text{d}}(H)$ can be either empty (no bound states), or consist of a finite or infinite increasing sequence of negative eigenvalues of finite multiplicities. If H has negative eigenvalues, the lowest one is called the ground state energy. If V_{ext} is not too singular (see [68, Theorem XIII.46] for details), it is non-degenerate. The higher eigenvalues are called excited state energies. If H has infinitely many negative eigenvalues, then they necessarily accumulate at 0, the bottom of the essential spectrum. This is the case for instance for the Hamiltonian of the hydrogen atom: the discrete spectrum of the Hamiltonian

$$H = -\frac{\hbar^2}{2m_e} - \frac{e^2}{4\pi\varepsilon_0|\mathbf{r}|}$$

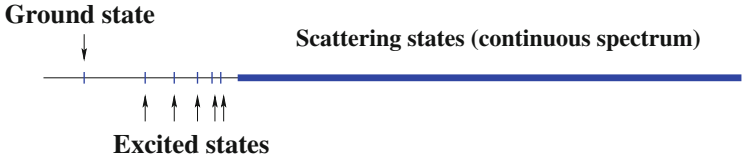


Fig. 4 Typical spectrum of one-particle Hamiltonians encountered in first-principle simulation of finite molecular systems

Fig. 5 Emission spectrum of atomic hydrogen in the visible range

on $L^2(\mathbb{R}^3)$ is the Rydberg series $(E_n)_{n\in\mathbb{N}^*}$, where $E_n = -\frac{E_{\rm Ryd}}{n^2}$, and where

$$E_{\rm Ryd} := \frac{m_e}{2}\left(\frac{e^2}{4\pi\varepsilon_0\hbar}\right)^2$$

is the Rydberg energy. Here m_e is the electron mass, e the elementary charge, and ε_0 the dielectric permittivity of the vacuum (Fig. 4).

When this model is coupled to a quantized electromagnetic field, transitions between electronic energy levels may occur. The electron of the hydrogen atom may jump from a higher energy level E_m to a lower one E_n $(m > n)$ by emitting a photon of energy $h\nu_{m\to n} = E_m - E_n$ ($h = 2\pi\hbar$ is the Planck constant and $\nu_{m\to n}$ the frequency of the photon), or, conversely, absorb a photon of energy $h\nu_{m\to n}$ and jump from the energy level E_n to the energy level E_m. As a consequence, the transitions between electronic levels are quantized. This is the reason why the emission and absorption spectra of molecular gases consist of rays (see Fig. 5). In the case of the hydrogen atom, four rays lay in the visible spectrum (wavelengths between 400 and 700 nm). They are part of the Balmer series (transitions between E_m and E_2) and can be easily measured experimentally:

$$\lambda^{\rm exp}_{6\to2} = 410.17\ \text{nm},\ \lambda^{\rm exp}_{5\to2} = 434.05\ \text{nm},\ \lambda^{\rm exp}_{4\to2} = 486.13\ \text{nm},\ \lambda^{\rm exp}_{3\to2} = 656.28\ \text{nm}.$$

Using the relation $\lambda_{m\to n} = c/\nu_{m\to n}$, where c is the speed of light, the wavelengths of the electronic transitions are given by

$$\lambda_{m\to n} = \frac{8\pi\hbar c}{E_{\rm Ryd}}\left(\frac{1}{n^2} - \frac{1}{m^2}\right)^{-1},$$

which leads to the following numerical results

$$\lambda_{6\to 2} = 410.07 \text{ nm},\ \lambda_{5\to 2} = 433.94 \text{ nm},\ \lambda_{4\to 2} = 486.01 \text{ nm},\ \lambda_{3\to 2} = 656.11 \text{ nm}.$$

The slight discrepancies between these results and the experimental ones are due to the fact that the motion of the nucleus and the relativistic effects have not been taken into account. Replacing the electron mass m_e with the reduced mass $m_e m_p/(m_e + m_p)$, where m_p is the proton mass, and adding to the non-relativistic Hamiltonian the so-called Breit terms [30], experimental values can be recovered with a very high relative accuracy of the order of 10^{-8}.

The above discussion provides a physical interpretation of the discrete spectrum of the Hamiltonian of the hydrogen atom. Let us now turn to the continuous spectrum.

Theorem 24 (RAGE Theorem, Ruelle [70], Amrein and Georgescu [3], Enss [33]) *Let H be a locally compact[7] self-adjoint operator on $L^2(\mathbb{R}^d)$. Let*

$$\mathscr{H}_p = \overline{Span\{eigenvectors\ of\ H\}} \quad and \quad \mathscr{H}_c = \mathscr{H}_p^{\perp}.$$

Let χ_{B_R} be the characteristic function of the ball $B_R = \{\mathbf{r} \in \mathbb{R}^d \mid |\mathbf{r}| < R\}$. Then,

$$\left(\phi_0 \in \mathscr{H}_p\right) \Leftrightarrow \forall \varepsilon > 0,\ \exists R > 0,\ \forall t \geq 0,\ \left\| (1 - \chi_{B_R}) e^{-itH/\hbar} \phi_0 \right\|_{L^2}^2 \leq \varepsilon;$$

$$(\phi_0 \in \mathscr{H}_c) \Leftrightarrow \forall R > 0,\ \lim_{T\to+\infty} \frac{1}{T} \int_0^T \left\| \chi_{B_R} e^{-itH/\hbar} \phi_0 \right\|_{L^2}^2 dt = 0.$$

The physical meaning of this result is the following: if the particle is in the quantum state ϕ_0 at $t = 0$, then its state at time t is the solution at time t of the time-dependent Schrödinger equation (10) with initial datum ϕ_0, that is $\psi(t) = e^{-itH/\hbar}\phi_0$. In view of the physical interpretation of the wavefunction in the position representation,

$$\left\| \chi_{B_R} e^{-itH/\hbar} \phi_0 \right\|_{L^2}^2 = \int_{B_R} |\psi(t, \mathbf{r})|^2 \, d\mathbf{r}$$

[7] An operator A on $L^2(\mathbb{R}^d)$ such that $\rho(A) \neq \emptyset$ is called locally compact if for any bounded set B, the operator $\chi_B(z - A)^{-1}$ is a compact operator on $L^2(\mathbb{R}^d)$ for some (and then all by virtue of the resolvent formula) $z \in \rho(A)$. Here, χ_B is the characteristic function of B; in the expression $\chi_B(z - A)^{-1}$, χ_B should be understood as the multiplication operator by the bounded function χ_B, which is a bounded self-adjoint operator on $\mathscr{H}$. The Hamiltonian of the hydrogen atom is a locally compact self-adjoint operator on $L^2(\mathbb{R}^3)$, and for this operator, $\dim(\mathscr{H}_p) = \dim(\mathscr{H}_c) = \infty$.

is the probability that the particle lays inside the ball B_R at time t, while

$$\left\|(1-\chi_{B_R})e^{-itH/\hbar}\phi_0\right\|_{L^2}^2 = \int_{\mathbb{R}^3\setminus B_R} |\psi(t,\mathbf{r})|^2\, d\mathbf{r}$$

is the probability that the particle lays outside the ball B_R at time t.

The subspace $\mathscr{H}_{\rm p}$ can therefore be seen as a set of bound states, and the subspace $\mathscr{H}_{\rm c}$ as a set of scattering states:

- if $\phi_0 \in \mathscr{H}_p$, then the particle essentially remains in the vicinity of the nucleus at all times;
- if $\phi_0 \in \mathscr{H}_c$, then the particle scatters away from the nucleus. Note that in the case of the hydrogen atom, which has no singular continuous spectrum [68, Section XIII.10], the convergence is stronger: $\left\|\chi_{B_R}e^{-itH/\hbar}\phi_0\right\|_{L^2}^2$ goes to zero when t goes to infinity.

4.2 Many-Particle Systems

The state space $\mathscr{H}$ of a quantum system consisting of two spinless particles is always a closed subspace of $L^2(\mathbb{R}^3,\mathbb{C}) \otimes L^2(\mathbb{R}^3,\mathbb{C}) \equiv L^2(\mathbb{R}^6,\mathbb{C})$, and, in the position representation, if the system is in the pure state $\Psi(t)$ at time t, then $|\Psi(t,\mathbf{r}_1,\mathbf{r}_2)|^2$ is the probability density of observing at time t particle 1 at $\mathbf{r}_1$ and particle 2 at $\mathbf{r}_2$. The precise structure of $\mathscr{H}$ depends of the natures of the two particles[8]:

- for two different particles: $\mathscr{H} = L^2(\mathbb{R}^3,\mathbb{C}) \otimes L^2(\mathbb{R}^3,\mathbb{C})$;
- for two identical bosons (e.g. two carbon 12 nuclei), $\mathscr{H} = L^2(\mathbb{R}^3,\mathbb{C}) \vee L^2(\mathbb{R}^3,\mathbb{C})$, where $\vee$ denotes the symmetrized tensor product. Otherwise stated, the wavefunction Ψ must satisfy the symmetry condition

$$\Psi(t,\mathbf{r}_2,\mathbf{r}_1) = \Psi(t,\mathbf{r}_1,\mathbf{r}_2);$$

- for two identical fermions (e.g. two electrons), $\mathscr{H} = L^2(\mathbb{R}^3,\mathbb{C}) \wedge L^2(\mathbb{R}^3,\mathbb{C})$, where $\wedge$ denotes the antisymmetrized tensor product. In other words, the wavefunction Ψ must satisfy the antisymmetry condition, also called Pauli principle,

$$\Psi(t,\mathbf{r}_2,\mathbf{r}_1) = -\Psi(t,\mathbf{r}_1,\mathbf{r}_2).$$

[8] For simplicity, we omit the spin variables. See Remark 25 below for more details.

Note that for two identical particles, whatever they are bosons or fermions, the particle density is given by

$$\rho(t,\mathbf{r}) = \int_{\mathbb{R}^3} |\Psi(t,\mathbf{r},\mathbf{r}_2)|^2\, d\mathbf{r}_2 + \int_{\mathbb{R}^3} |\Psi(t,\mathbf{r}_1,\mathbf{r})|^2\, d\mathbf{r}_1 = 2\int_{\mathbb{R}^3} |\Psi(t,\mathbf{r},\mathbf{r}_2)|^2\, d\mathbf{r}_2.$$

Consider now N quantum particles of masses $m_1,\cdots,m_N$ subjected to an external potential $V_{\text{ext}}(\mathbf{r})$ and pair-interaction potentials $W_{ij}(\mathbf{r}_i,\mathbf{r}_j)$. The state space $\mathscr{H}$ then is a closed subspace of $L^2(\mathbb{R}^3,\mathbb{C})\otimes\cdots\otimes L^2(\mathbb{R}^3,\mathbb{C}) \equiv L^2(\mathbb{R}^{3N},\mathbb{C})$, whose precise structure depends on the natures of the N particles. In the case of N identical bosons, $\mathscr{H} = \vee^N L^2(\mathbb{R}^3,\mathbb{C})$, while in the case of N identical fermions, $\mathscr{H} = \wedge^N L^2(\mathbb{R}^3,\mathbb{C})$. Likewise, if the state of the system at time t is characterized by the wavefunction $\Psi(t)\in\mathscr{H}$ in the position representation, then $|\Psi(t,\mathbf{r}_1,\cdots,\mathbf{r}_N)|^2$ is the probability density of observing at time t particle 1 at $\mathbf{r}_1$, particle 2 at $\mathbf{r}_2$, etc. The time-independent Schrödinger equation of such a system reads

$$\left(-\sum_{i=1}^N \frac{\hbar^2}{2m_i}\Delta_{\mathbf{r}_i} + \sum_{i=1}^N V_{\text{ext}}(\mathbf{r}_i) + \sum_{1\le i<j\le N} W_{ij}(\mathbf{r}_i,\mathbf{r}_j)\right)\Psi(\mathbf{r}_1,\cdots,\mathbf{r}_N) = E\,\Psi(\mathbf{r}_1,\cdots,\mathbf{r}_N)$$

and therefore has the structure of a $3N$-dimensional linear elliptic eigenvalue problem.

In general, such an equation is extremely difficult to solve. However, in the special case of N *non-interacting* identical particles of mass m subjected to an external potential $V_{\text{ext}}(\mathbf{r})$, the Hamiltonian becomes separable

$$H = -\sum_{i=1}^N \frac{\hbar^2}{2m}\Delta_{\mathbf{r}_i} + \sum_{i=1}^N V_{\text{ext}}(\mathbf{r}_i) = \sum_{i=1}^N \mathfrak{h}_{\mathbf{r}_i}$$

and all the bound states of H can be easily computed from the bound states of the three-dimensional Schrödinger operator $\mathfrak{h}$:

$$\begin{cases} \mathfrak{h}\,\phi_i = \varepsilon_i\phi_i, \qquad \varepsilon_1\le\varepsilon_2\le\cdots, \\ \displaystyle\int_{\mathbb{R}^3}\phi_i\phi_j = \delta_{ij}, \\ \mathfrak{h} = -\dfrac{\hbar^2}{2m}\Delta + V_{\text{ext}}. \end{cases}$$

In particular, if $\mathfrak{h}$ is bounded below and has at least one (for bosons) or N (for fermions) eigenvalues below the bottom of the essential spectrum, then H has a ground state:

- the bosonic ground state energy is $E_0 = N\varepsilon_1$ and the ground state wavefunction and density are given by

$$\psi(\mathbf{r}_1,\cdots,\mathbf{r}_N) = \prod_{i=1}^N \phi_1(\mathbf{r}_i) \quad\text{and}\quad \rho(\mathbf{r}) = N|\phi_1(\mathbf{r})|^2;$$

- the fermionic ground state energy is $E_0 = \sum_{i=1}^N \varepsilon_i$, a ground state wavefunction is the Slater determinant

$$\psi(\mathbf{r}_1, \cdots, \mathbf{r}_N) = \frac{1}{\sqrt{N!}} \det(\phi_i(\mathbf{r}_j)) = \frac{1}{\sqrt{N!}} \begin{vmatrix} \phi_1(\mathbf{r}_1) & \phi_1(\mathbf{r}_2) & \cdots & \phi_1(\mathbf{r}_N) \\ \phi_2(\mathbf{r}_1) & \phi_2(\mathbf{r}_2) & \cdots & \phi_2(\mathbf{r}_N) \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & \cdots & \cdot \\ \cdot & \cdot & & \cdot \\ \phi_N(\mathbf{r}_1) & \phi_N(\mathbf{r}_2) & \cdots & \phi_N(\mathbf{r}_N) \end{vmatrix},$$

  and the corresponding density is $\rho(\mathbf{r}) = \sum_{i=1}^N |\phi_i(\mathbf{r})|^2$.

5 First-Principle Molecular Simulation

First-principle molecular simulation is based on a simple observation:

- any molecule is a set of M nuclei and N electrons in Coulomb interaction;
- the state space $\mathscr{H}$ is the subset of $L^2(\mathbb{R}^{3(M+N)}, \mathbb{C})$ defined by the suitable symmetry and antisymmetry constraints for identical bosons and fermions;
- the Hamiltonian of the molecule is

$$\begin{aligned} H = &- \sum_{k=1}^M \frac{1}{2m_k} \Delta_{\mathbf{R}_k} - \sum_{i=1}^N \frac{1}{2} \Delta_{\mathbf{r}_i} \\ &- \sum_{i=1}^N \sum_{k=1}^M \frac{z_k}{|\mathbf{r}_i - \mathbf{R}_k|} + \sum_{1 \le i < j \le N} \frac{1}{|\mathbf{r}_i - \mathbf{r}_j|} + \sum_{1 \le k < l \le M} \frac{z_k z_l}{|\mathbf{R}_k - \mathbf{R}_l|}. \end{aligned} \tag{11}$$

Here, we have used atomic units, that is the set of units such that

$$\hbar = 1, \quad m_e = 1, \quad e = 1, \quad 4\pi\varepsilon_0 = 1.$$

Remarkably, the Hamiltonian (11) is free of empirical parameters specific to the molecular system, and it can be deduced from the mere chemical formula of the latter. Likewise, any physical observable associated with the system and can be written down from the first-principles of quantum mechanics. Quoting Dirac [28],

> *The underlying physical laws necessary for the mathematical theory of a large part of physics and the whole of chemistry are thus completely known, and the difficulty is only that the exact application of these laws leads to equations much too complicated to be solved.*

The model described above is extremely accurate, at least for light atoms, for which relativistic effects can be neglected. As a matter of example, let us consider the computation of the ionization energy of the helium atom. The ionization process is the reaction

$$\mathrm{He} + \mathrm{h}\nu \;\rightarrow\; \mathrm{He}^{+} + \mathrm{e}^{-},$$

in which a helium atom absorbs a photon; if the energy of the photon is larger than a threshold value $\Delta E = h\Delta\nu$, one of the two electrons of the atom is kicked out of its bound state and escapes to infinity. The threshold frequency $\Delta\nu$ can be measured experimentally with high accuracy. Two different experiments on helium 4 (the most common isotope of helium, whose nucleus contains four nucleons: two protons and two neutrons) performed in 1997 and 1998 respectively lead to the following results:

$$\Delta\nu^{\mathrm{exp.1}} \simeq 5{,}945{,}204{,}238\,\mathrm{MHz}\ [32] \quad \text{and} \quad \Delta\nu^{\mathrm{exp.2}} \simeq 5{,}945{,}204{,}356\,\mathrm{MHz}\ [11].$$

From a theoretical point of view, $\Delta\nu$ can be computed as $\Delta\nu = \Delta E/h$, where $\Delta E = \min(\sigma(H_{\mathrm{He}^+})) - \min(\sigma(H_{\mathrm{He}}))$, where $\sigma(H_{\mathrm{He}^+})$ and $\sigma(H_{\mathrm{He}})$ are the spectra of the operators

$$H_{\mathrm{He}} = -\frac{1}{2m}\Delta_{\mathbf{R}} - \frac{1}{2}\Delta_{\mathbf{r}_1} - \frac{1}{2}\Delta_{\mathbf{r}_2} - \frac{2}{|\mathbf{r}_1 - \mathbf{R}|} - \frac{2}{|\mathbf{r}_2 - \mathbf{R}|} + \frac{1}{|\mathbf{r}_1 - \mathbf{r}_2|},$$

and

$$H_{\mathrm{He}^+} = -\frac{1}{2m}\Delta_{\mathbf{R}} - \frac{1}{2}\Delta_{\mathbf{r}_1} - \frac{2}{|\mathbf{r}_1 - \mathbf{R}|},$$

respectively (see Fig. 6), where m denotes the mass of the Helium 4 nucleus. It can be shown that $\min(\sigma(H_{\mathrm{He}^+})) = -2$. Using translational and rotational invariance, the quantity $\min(\sigma(H_{\mathrm{He}}))$ can be obtained by solving a three-dimensional linear elliptic eigenvalue problem. A careful calculation reported in [47] gives: $\Delta E^{\mathrm{calc.1}} = 5{,}945{,}262{,}288$ MHz. Taking relativistic corrections (Breit terms) into account gives $\Delta E^{\mathrm{calc.2}} = 5{,}945{,}204{,}223\,\mathrm{MHz}$, to be compared with the experimental

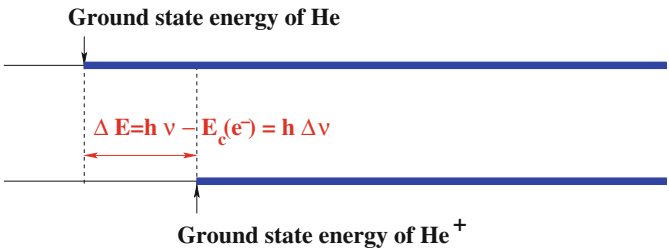


Fig. 6 Spectra of the Hamiltonians H_{He} and H_{He^+}

values reported above. The agreement between theory and experiment is therefore exceptionally good.

Let us now turn to the more complicated case of a polyatomic system. As a matter of example, we will deal with a water molecule H_2O, which consists of $M = 3$ atomic nuclei (1 oxygen 16 nucleus, and 2 hydrogen 1 nuclei[9]) and $N = 10$ electrons in Coulomb interaction. Such a system can be fully described by the laws of quantum mechanics (many-body Schrödinger equation) and statistical physics. The only parameters of these models are

- a few fundamental constants of physics

$$\hbar = 1, \quad m_e = 1, \quad e = 1, \quad \varepsilon_0 = (4\pi)^{-1},$$

$$c \simeq 137.0359996287515\ldots, \quad k_{\mathrm{B}} = 3.16681537\ldots \times 10^{-6},$$

  where c is the speed of light and k_{B} the Boltzmann constant (all the values are in atomic units);
- the charges and masses of the hydrogen 1 and oxygen 16 nuclei

$$z_{\mathrm{H}} = 1, \quad z_{\mathrm{O}} = 8, \quad m_{\mathrm{H}} = 1836.152701\ldots, \quad m_{\mathrm{O}} = 29156.944123\ldots$$

We then observe that the ratio m_e/m_n (electron mass/nucleus mass) is very small, even for the lightest nucleus (hydrogen 1). Following Born and Oppenheimer, this suggests to use this ratio as a small parameter to approximate the many-body Schrödinger equation. The procedure described in the sequel can be justified to some point with mathematically rigorous arguments; we refer the interested reader to the literature cited below. The so-called Born-Oppenheimer method can be decomposed in two steps:

- step 1: definition of the potential energy surfaces;
- step 2: analysis of the potential energy surfaces.

Let us first detail the first step. Assuming that the M nuclei are clamped point-like particles located at positions $\mathbf{R}_1, \cdots, \mathbf{R}_M$, $\mathbf{R}_k \in \mathbb{R}^3$, the electronic problem for the nuclear configuration $\{\mathbf{R}_k\}_{1 \le k \le M}$ consists in computing the bound states of the N electrons in the electrostatic potential

$$V^{\mathrm{ne}}_{\{\mathbf{R}_k\}}(\mathbf{r}) = -\sum_{k=1}^{M} \frac{z_k}{|\mathbf{r} - \mathbf{R}_k|}$$

[9] These are the most common isotopes of oxygen and hydrogen.

generated by the nuclei. For the water molecule, we have: $M = 3$, $N = 10$, $z_1 = 8$, $z_2 = 1$, $z_3 = 1$. The electronic bound states are obtained by solving the time-independent Schrödinger equation

$$\left(-\frac{1}{2}\sum_{i=1}^{N}\Delta_{\mathbf{r}_i} + \sum_{i=1}^{N} V^{\mathrm{ne}}_{\{\mathbf{R}_k\}}(\mathbf{r}_i) + \sum_{1\le i<j\le N}\frac{1}{|\mathbf{r}_i-\mathbf{r}_j|}\right)\psi(\mathbf{r}_1,\cdots,\mathbf{r}_N) = E\,\psi(\mathbf{r}_1,\cdots,\mathbf{r}_N).$$

Since electrons are fermions, the wavefunction ψ must satisfy the antisymmetry condition

$$\forall p \in \mathfrak{S}_N, \quad \psi(\mathbf{r}_{p(1)},\cdots,\mathbf{r}_{p(N)}) = \varepsilon(p)\psi(\mathbf{r}_1,\cdots,\mathbf{r}_N), \qquad \text{(Pauli principle)}.$$

The electronic density associated with ψ is

$$\rho_\psi(\mathbf{r}) = N\int_{\mathbb{R}^{3(N-1)}} |\psi(\mathbf{r},\mathbf{r}_2,\cdots,\mathbf{r}_N)|^2\, d\mathbf{r}_2\cdots d\mathbf{r}_N, \tag{12}$$

and the normalization condition $\|\psi\|_{L^2} = 1$ ensures that

$$\int_{\mathbb{R}^3}\rho_\psi(\mathbf{r})\,d\mathbf{r} = N.$$

Remark 25 For simplicity, we omit here the spin variables. In fact, electrons are particles of spin $s = 1/2$, so that the one-electron state space is not $L^2(\mathbb{R}^3,\mathbb{C})$ but

$$L^2(\mathbb{R}^3,\mathbb{C}^{2s+1}) = L^2(\mathbb{R}^3,\mathbb{C}^2) \equiv L^2(\mathbb{R}^3\times\{|\uparrow\rangle,|\downarrow\rangle\},\mathbb{C}),$$

where $|\uparrow\rangle$ and $|\downarrow\rangle$ respectively denote the spin-up and spin-down states. An N-electron wavefunction therefore is a vector of $\mathscr{H}_N = \bigwedge^N L^2(\mathbb{R}^3\times\{|\uparrow\rangle,|\downarrow\rangle\},\mathbb{C})$, that is a complex-valued function of the variables $(\mathbf{r}_1,\sigma_1;\cdots,\mathbf{r}_N,\sigma_N) \in \left(\mathbb{R}^3\times\{|\uparrow\rangle,|\downarrow\rangle\}\right)^N$ satisfying the antisymmetry condition

$$\forall p \in \mathfrak{S}_N, \quad \psi(\mathbf{r}_{p(1)},\sigma_{p(1)};\cdots;\mathbf{r}_{p(N)},\sigma_{p(N)}) = \varepsilon(p)\psi(\mathbf{r}_1,\sigma_1;\cdots;\mathbf{r}_N,\sigma_N).$$

In this framework, $|\psi(\mathbf{r}_1,\sigma_1;\cdots;\mathbf{r}_N,\sigma_N)|^2$ represents the probability density of observing electron 1 at r_1 in the spin state σ_1, electron 2 at r_2 in the spin state σ_2, etc.

The structure of the spectrum of the electronic Hamiltonian

$$H_N^{\{\mathbf{R}_k\}} = -\sum_{i=1}^{N}\frac{1}{2}\Delta_{\mathbf{r}_i} - \sum_{i=1}^{N} V^{\mathrm{ne}}_{\{\mathbf{R}_k\}}(\mathbf{r}_i) + \sum_{1\le i<j\le N}\frac{1}{|\mathbf{r}_i-\mathbf{r}_j|}$$

Fig. 7 Graphical illustration of Zhislin's theorem

on $\mathscr{H}_N = \bigwedge^N L^2(\mathbb{R}^3, \mathbb{C})$ with domain $\mathscr{H}_N \cap H^2(\mathbb{R}^{3N})$, which can be proved to be self-adjoint, is given by Zhislin's theorem (illustrated by Fig. 7).

Theorem 26 (Zhislin [79]) *If* $N \leq \sum_{k=1}^{M} z_k$ *(neutral or positively charged system), then*

$$\sigma_{\mathrm{d}}(H_N^{\{\mathbf{R}_k\}}) = \left\{E_n^{\{\mathbf{R}_k\}}\right\}_{n\in\mathbb{N}} \quad \text{and} \quad \sigma_{\mathrm{ess}}(H_N^{\{\mathbf{R}_k\}}) = [\Sigma_N^{\{\mathbf{R}_k\}}, +\infty),$$

where $(E_n^{\{\mathbf{R}_k\}})_{n\in\mathbb{N}}$ *is a nondecreasing sequence of negative eigenvalues*[10] *converging to* $\Sigma^{\{\mathbf{R}_k\}}$*, the bottom of the essential spectrum of* $H_N^{\{\mathbf{R}_k\}}$*. Besides* $\Sigma_N^{\{\mathbf{R}_k\}} = 0$ *if* $N = 1$ *and* $\Sigma^{\{\mathbf{R}_k\}} < 0$ *if* $N \geq 1$.

It can also be shown (HVZ theorem [42, 76, 78]) that $\Sigma_N^{\{\mathbf{R}_k\}} = \min \sigma(H_{N-1}^{\{\mathbf{R}_k\}})$.

The lowest eigenvalue $E_0^{\{\mathbf{R}_k\}}$ is called the ground state energy of $H_N^{\{\mathbf{R}_k\}}$, while the eigenvalues $E_n^{\{\mathbf{R}_k\}} > E_0^{\{\mathbf{R}_k\}}$ are called the excited state energies of $H_N^{\{\mathbf{R}_k\}}$. For each $n \in \mathbb{N}$, the function $\mathbb{R}^{3M} \ni (\mathbf{R}_1, \cdots, \mathbf{R}_M) \mapsto E_n^{\{\mathbf{R}_k\}} \in \mathbb{R}$ is continuous. This can be proved using e.g. the minmax principle (Theorem 23), or Kato's perturbation theory of self-adjoint operators [44]. Using the latter approach, it can be shown in addition that this function is C^1 at $(\mathbf{R}_1, \cdots, \mathbf{R}_M)$ whenever $E_n^{\{\mathbf{R}_k\}}$ is a nondegenerate eigenvalue of $H_N^{\{\mathbf{R}_k\}}$.

The potential energy surfaces are then defined as the real-valued functions W_n on $\mathbb{R}^{3M}$, $n \in \mathbb{N}$, defined by

$$W_n(\mathbf{R}_1, \cdots, \mathbf{R}_M) = E_n^{\{\mathbf{R}_k\}} + \sum_{1\leq k<l\leq M} \frac{z_k z_l}{|\mathbf{R}_k - \mathbf{R}_l|}. \tag{13}$$

The function W_0 is called the ground state potential energy surface (PES), the function W_1 the first excited state PES, etc. (Fig. 8).

[10] Eigenvalues are counted with their multiplicities, so that $E_0^{\{\mathbf{R}_k\}} \leq E_1^{\{\mathbf{R}_k\}} \leq E_2^{\{\mathbf{R}_k\}} \cdots$, with a priori large inequalities.

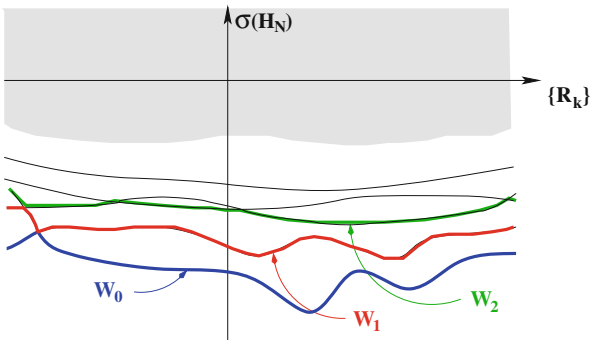


Fig. 8 Sketch of the potential energy surfaces W_n

Let us now turn to the second step, that is the analysis of the potential energy surfaces. Usually,[11] the Born-Oppenheimer approximation is invoked at this point. This approximation is based on the fact that

1. the ratio m_e/m_n (electron mass/nucleus mass) is small, which allows one to somehow decouple the electronic and nuclear dynamics by means of an adiabatic limit [61]. At low enough temperature (usually from 0 K to room temperature or more), it can be considered for most systems that the wave function of the molecular system at time t can be approximated by a wave function of the form

$$\Psi^{\mathrm{BO}}(t;\mathbf{R}_1,\cdots,\mathbf{R}_M;\mathbf{r}_1,\cdots,\mathbf{r}_N) = \Phi(t;\mathbf{R}_1,\cdots,\mathbf{R}_M)\psi^{(\mathbf{R}_1,\cdots,\mathbf{R}_M)}(\mathbf{r}_1,\cdots,\mathbf{r}_N),$$

   where $\psi^{(\mathbf{R}_1,\cdots,\mathbf{R}_M)}(\mathbf{r}_1,\cdots,\mathbf{r}_N)$ is a normalized ground state of $H_N^{\{\mathbf{R}_k\}}$, that is a L^2-normalized eigenfunction of the electronic Hamiltonian $H_N^{\{\mathbf{R}_k\}}$ associated with the ground state eigenvalue $E_0^{\{\mathbf{R}_k\}}$, assumed here to be non-degenerate;
2. nuclei are heavy particles, so that their dynamics can be well-approximated by the classical Newton equation

$$m_k \frac{d^2\mathbf{R}_k}{dt^2}(t) = -\nabla_{\mathbf{R}_k} W_0(\mathbf{R}_1(t),\cdots,\mathbf{R}_M(t)), \qquad 1 \le k \le M. \tag{14}$$

   This equation is obtained from the Schrödinger equation on $\Phi(t;\mathbf{R}_1,\cdots,\mathbf{R}_M)$ resulting from the adiabatic approximation, by letting the reduced Planck constant $\hbar$ go to zero (semiclassical limit, see [1, 2] and references therein).

Equation (14), together with the definition (13) for $n = 0$ of the ground state PES, are the fundamental equations of first-principle molecular dynamics. According to this model, the nuclei behave as point-like classical particles interacting via the effective M-body potential W_0.

[11] Breakdowns of the adiabatic approximation are studied in [14, 22, 35].

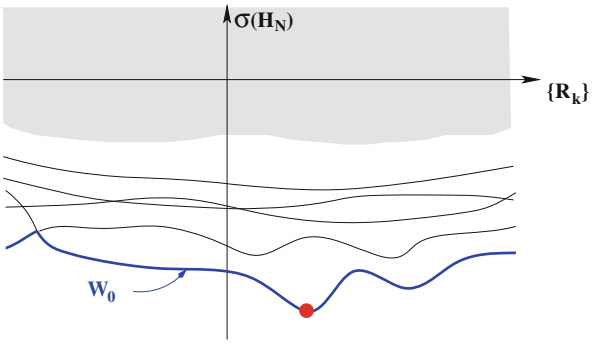


Fig. 9 Within the Born-Oppenheimer approximation, the global minimizers of W_0 correspond to the most stable configurations of the system

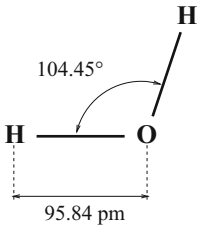


Fig. 10 Equilibrium configuration of the water molecule (experimental values)

It follows from (14) that the local minima of W_0 correspond to equilibrium configurations of the system. In particular, global minima of W_0 correspond to the most stable configurations of the molecular system under consideration (Fig. 9). The water molecule has a single global minimum (up to translations and rotations), corresponding to the configuration depicted on Fig. 10.

The limiting step for integrating numerically the first-principle molecular dynamics Eq. (14) is the computation of the effective forces $-\nabla_{\mathbf{R}_k} W_0(\mathbf{R}_1, \cdots, \mathbf{R}_M)$ experienced by the nuclei. The nucleus-nucleus interaction is explicit and easy to deal with. The main issue is the computation of $-\nabla_{\mathbf{R}_k} E_0^{\{\mathbf{R}_1,\cdots,\mathbf{R}_M\}}$. Since $E_0^{\{\mathbf{R}_1,\cdots,\mathbf{R}_M\}}$ is the ground state eigenvalue of $H_N^{\{\mathbf{R}_k\}}$, il can be obtained by solving the constrained optimization problem

$$E_0^{\{\mathbf{R}_1,\cdots,\mathbf{R}_M\}} = \inf \left\{ \langle \psi | H_N^{\{\mathbf{R}_k\}} | \psi \rangle,\ \psi \in \bigwedge^N L^2(\mathbb{R}^3) \cap H^1(\mathbb{R}^{3N}),\ \|\psi\|_{L^2} = 1 \right\}. \tag{15}$$

This problem has the same structure as problem (3), which implies that it is not necessary to compute the first derivatives of the minimizers with respect to the $\mathbf{R}_k$'s to compute $-\nabla_{\mathbf{R}_k} E_0^{\{\mathbf{R}_1,\cdots,\mathbf{R}_M\}}$. In addition, since the constraint $\|\psi\|_{L^2} = 1$ does

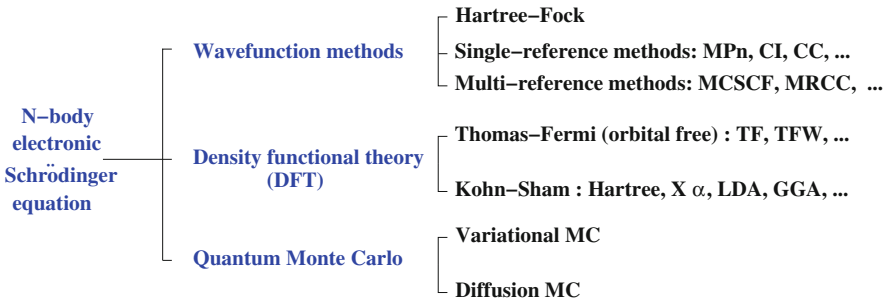


Fig. 11 Classification of the main electronic structure methods

not depend explicitly on the $\mathbf{R}_k$'s, the gradients $-\nabla_{\mathbf{R}_k} E_0^{\{\mathbf{R}_1,\cdots,\mathbf{R}_M\}}$ can be computed explicitly from the minimizer $\psi_0^{\{\mathbf{R}_k\}}$. A simple calculation shows that

$$-\nabla_{\mathbf{R}_k} W_0(\mathbf{R}_1,\cdots,\mathbf{R}_M) = z_k \int_{\mathbb{R}^3} \rho_0^{\{\mathbf{R}_k\}}(\mathbf{r}) \frac{\mathbf{r}-\mathbf{R}_k}{|\mathbf{r}-\mathbf{R}_k|^3}\, d\mathbf{r} + \sum_{l\neq k} z_k z_l \frac{\mathbf{R}_k-\mathbf{R}_l}{|\mathbf{R}_k-\mathbf{R}_l|^3},$$

where the ground state density

$$\rho_0^{\{\mathbf{R}_k\}}(\mathbf{r}) = N \int_{\mathbb{R}^{3(N-1)}} |\psi_0^{\{\mathbf{R}_k\}}(\mathbf{r},\mathbf{r}_2,\cdots,\mathbf{r}_N)|^2\, d\mathbf{r}_2\cdots\mathbf{r}_N,$$

is the electronic density associated with the ground state wavefunction $\psi_0^{\{\mathbf{R}_k\}}$. Since the electronic Schrödinger equation is a 3N-dimensional PDE, it is not possible to solve it accurately for systems containing more than a couple of electrons. Several approximation have been proposed along the past 80 years, which can be classified in three main groups (see Fig. 11). Describing all these methods is out of the scope of this introductory lecture notes. We will only focus on the simplest of them, namely the Hartree-Fock method, which will be presented in the next section. We refer the reader to [40] for a comprehensive monograph on wavefunction methods, to [29, 34] for reference textbooks on DFT, to [7] for a several relevant contributions, including a mathematical introduction to quantum Monte Carlo methods, and to [4, 17–19, 24, 26, 37–39, 48, 51, 53–60, 71, 74] and reference therein for various mathematical and numerical works on these models.

Let us mention that the various avatars of the Kohn-Sham model [10, 46, 62, 63, 73, 75] are the most widely used models in the present time, since it is generally considered as the best compromise between computational efficiency and accuracy. The mathematical structure of the Kohn-Sham LDA model is quite similar to the one of the Hartree-Fock model we are now going to discuss.

6 Hartree-Fock Approximation

In this section, we assume that the nuclear configuration $\{\mathbf{R}_k\}$ is given, and we focus on the calculation of the electronic ground state energy $E_0^{\{\mathbf{R}_k\}}$ and of the electronic components $-\nabla_{\mathbf{R}_k} E_0^{\{\mathbf{R}_k\}}$ of the atomic forces. In order to simplify the notation, we set $E_0 := E_0^{\{\mathbf{R}_k\}}$, $\rho_0 := \rho_0^{\{\mathbf{R}_k\}}$,

$$H_N := -\frac{1}{2}\sum_{i=1}^{N} \Delta_{\mathbf{r}_i} + \sum_{i=1}^{N} V^{\mathrm{ne}}(\mathbf{r}_i) + \sum_{1\le i<j\le N} \frac{1}{|\mathbf{r}_i - \mathbf{r}_j|} \quad \text{and} \quad V^{\mathrm{ne}}(\mathbf{r}) := -\sum_{k=1}^{M} \frac{z_k}{|\mathbf{r} - \mathbf{R}_k|}.$$

Recall that the operator H_N is self-adjoint on $\mathscr{H}_N = \bigwedge^N L^2(\mathbb{R}^3)$ with domain $D(H_N) = \mathscr{H}_N \cap H^2(\mathbb{R}^{3N})$ and form domain $Q(H_N) = \mathscr{H}_N \cap H^1(\mathbb{R}^{3N})$, and that the ground state energy can be obtained as

$$E_0 = \inf\{\langle\psi|H_N|\psi\rangle,\ \psi \in \mathscr{W}_N\},$$

where

$$\mathscr{W}_N = \left\{\psi \in \bigwedge^N L^2(\mathbb{R}^3) \cap H^1(\mathbb{R}^{3N}),\ \|\psi\|_{L^2} = 1\right\}.$$

The Hartree-Fock approximation is a variational approximation consisting in minimizing the exact energy functional $\langle\psi|H_N|\psi\rangle$ on the subset of $\mathscr{W}_N$ defined as

$$\left\{\psi_\Phi,\ \Phi = (\phi_1, \cdots, \phi_N) \in (H^1(\mathbb{R}^3))^N,\ \int_{\mathbb{R}^3} \phi_i\phi_j = \delta_{ij}\right\}$$

where

$$\psi_\Phi(\mathbf{r}_1, \cdots, \mathbf{r}_N) := \frac{1}{\sqrt{N!}} \begin{vmatrix} \phi_1(\mathbf{r}_1) & \phi_1(\mathbf{r}_2) & \cdots & \phi_1(\mathbf{r}_N) \\ \phi_2(\mathbf{r}_1) & \phi_2(\mathbf{r}_2) & \cdots & \phi_2(\mathbf{r}_N) \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \\ \phi_N(\mathbf{r}_1) & \phi_N(\mathbf{r}_2) & \cdots & \phi_N(\mathbf{r}_N) \end{vmatrix}$$

is the Slater determinant constructed with the functions $\phi_1, \cdots, \phi_N$.

Rewriting $\langle\psi_\Phi|H_N|\psi_\Phi\rangle$ as a function of $\Phi = (\phi_1, \cdots, \phi_N)$, we obtain after some technical manipulations that the Hartree-Fock ground state energy is

$$E_0^{\mathrm{HF}} = \inf\left\{E^{\mathrm{HF}}(\Phi),\ \Phi = (\phi_1, \cdots, \phi_N) \in (H^1(\mathbb{R}^3))^N,\ \int_{\mathbb{R}^3} \phi_i\phi_j = \delta_{ij}\right\},$$

where the Hartree-Fock energy functional is defined by

$$E^{\mathrm{HF}}(\Phi) = \frac{1}{2}\sum_{i=1}^{N}\int_{\mathbb{R}^3}|\nabla\phi_i|^2 + \int_{\mathbb{R}^3}\rho_\Phi V^{\mathrm{ne}} + \underbrace{\frac{1}{2}\int_{\mathbb{R}^3}\int_{\mathbb{R}^3}\frac{\rho_\Phi(\mathbf{r})\,\rho_\Phi(\mathbf{r}')}{|\mathbf{r}-\mathbf{r}'|}\,d\mathbf{r}\,d\mathbf{r}'}_{\text{Coulomb term}} \underbrace{-\frac{1}{2}\int_{\mathbb{R}^3}\int_{\mathbb{R}^3}\frac{|\gamma_\Phi(\mathbf{r},\mathbf{r}')|^2}{|\mathbf{r}-\mathbf{r}'|}\,d\mathbf{r}\,d\mathbf{r}'}_{\text{exchange term}},$$

with

$$V^{\mathrm{ne}}(\mathbf{r}) = -\sum_{k=1}^{M}\frac{z_k}{|\mathbf{r}-\mathbf{R}_k|}, \qquad \gamma_\Phi(\mathbf{r},\mathbf{r}') = \sum_{i=1}^{N}\phi_i(\mathbf{r})\,\phi_i(\mathbf{r}'), \qquad \rho_\Phi(\mathbf{r}) = \sum_{i=1}^{N}|\phi_i(\mathbf{r})|^2.$$

Since the Hartree-Fock approximation is variational, we have $E_0 \le E_0^{\mathrm{HF}}$. The function ρ_Φ is the electronic density associated with Φ. It is easy to check that $\rho_\Phi = \rho_{\psi_\Phi}$, where ρ_{ψ_Φ} is the density associated with the N-body wavefunction ψ_Φ by (12). The function γ_Φ is called the (one-electron) density matrix associated with Φ. It holds

$$\gamma_\Phi(\mathbf{r},\mathbf{r}') = N\int_{\mathbb{R}^{3(N-1)}}\psi_\Phi(\mathbf{r},\mathbf{r}_2,\cdots,\mathbf{r}_N)\psi_\Phi(\mathbf{r}',\mathbf{r}_2,\cdots,\mathbf{r}_N)\,d\mathbf{r}_2\cdots d\mathbf{r}_N.$$

The Hartree-Fock model enjoys a gauge invariance property: if $\Phi \in (H^1(\mathbb{R}^3))^N$ satisfies the L^2-orthonormality constraints, then so does ΦU for all $U \in O(N)$ and $E(\Phi U) = E(\Phi)$. This is due to the fact that $\psi_{\Phi U} = \det(U)\psi_\Phi$. This property is used in the proof of the fifth statement of the following theorem.

Theorem 27 *Assume that $N \le Z := \sum_{k=1}^{M} z_k$ (neutral or positively charged molecular system). Then*

1. *the Hartree-Fock model has a ground state $\Phi^0 = (\phi_1^0,\cdots,\phi_N^0)$ [56];*
2. *Euler-Lagrange equations: there exists $\lambda \in \mathbb{R}^{N\times N}$ symmetric such that*

$$\begin{cases} \Phi^0 = (\phi_1^0,\cdots,\phi_N^0) \in (H^1(\mathbb{R}^3))^N \\ -\frac{1}{2}\Delta\phi_i^0 + V^{\mathrm{ne}}\phi_i^0 + \left(\rho_{\Phi^0}\star|\cdot|^{-1}\right)\phi_i^0 - \int_{\mathbb{R}^3}\frac{\gamma_{\Phi^0}(\cdot,\mathbf{r}')}{|\cdot-\mathbf{r}'|}\,\phi_i^0(\mathbf{r}')\,d\mathbf{r}' = \sum_{j=1}^{N}\lambda_{ij}\phi_j^0 \\ \int_{\mathbb{R}^3}\phi_i^0\phi_j^0 = \delta_{ij}; \end{cases}$$

3. *elliptic regularity: $\phi_i^0 \in H^2(\mathbb{R}^3)\cap C^{0,1}(\mathbb{R}^3)\cap C^\infty(\mathbb{R}^3\setminus\{\mathbf{R}_k\})$;*

4. *Fock operator:*

$$\mathscr{F}_{\Phi^0} := -\frac{1}{2}\Delta + V^{\mathrm{ne}} + \rho_{\Phi^0} \star |\cdot|^{-1} + \mathscr{K}_{\Phi^0},$$

where

$$(\mathscr{K}_{\Phi^0}\phi)(\mathbf{r}) = -\int_{\mathbb{R}^3} \frac{\gamma_{\Phi^0}(\mathbf{r},\mathbf{r}')}{|\mathbf{r}-\mathbf{r}'|}\,\phi(\mathbf{r}')\,d\mathbf{r}',$$

defines a self-adjoint operator on $L^2(\mathbb{R}^3)$ with domain $H^2(\mathbb{R}^3)$ and form domain $H^1(\mathbb{R}^3)$. It is bounded below and $\sigma_{\mathrm{ess}}(H_0) = [0, +\infty)$;

5. *Hartree-Fock equations: up to replacing Φ^0 by $\Phi^0 U$ for some $U \in O(N)$, it holds*

$$\mathscr{F}_{\Phi^0}\phi_i^0 = \varepsilon_i\phi_i^0, \qquad \int_{\mathbb{R}^3} \phi_i^0\phi_j^0 = \delta_{ij}, \qquad \varepsilon_1 \leq \cdots \leq \varepsilon_N < 0;$$

6. *Aufbau principle: $\varepsilon_1 \leq \varepsilon_2 \leq \cdots \leq \varepsilon_N$ are the lowest N eigenvalues of $\mathscr{F}_{\Phi^0}$, counting multiplicities;*
7. *no unfilled-shell property [8]: $\varepsilon_N < \varepsilon_{N+1}$ where $\varepsilon_{N+1} = \lambda_{N+1}(\mathscr{F}_{\Phi^0})$ is the $(N+1)^{\mathrm{st}}$ eigenvalue of $\mathscr{F}_{\Phi^0}$ (counting multiplicities) if $\mathscr{F}_{\Phi^0}$ has at least $(N+1)$ negative eigenvalues and 0 otherwise.*

The Hartree-Fock model can be solved numerically by means of a Galerkin approximation. Let $\mathscr{X} = \mathrm{Span}(\chi_1, \cdots, \chi_{N_b})$ be a subspace of $H^1(\mathbb{R}^3)$ of finite dimension N_b. An upper bound $E^{\mathrm{HF}}_{0,\mathscr{X}}$ of the exact Hartree-Fock ground state energy E^{HF}_0, which is itself an upper bound of the exact ground state energy E_0 of the electronic Hamiltonian, is obtained by minimizing the Hartree-Fock energy functional on the sets of orbitals in $\mathscr{X} = \mathrm{Span}(\chi_1, \cdots, \chi_{N_b})$ satisfying the L^2 orthonormality conditions:

$$E_0 \leq E^{\mathrm{HF}}_0 \leq E^{\mathrm{HF}}_{0,\mathscr{X}} = \inf\left\{ E^{\mathrm{HF}}(\Phi),\, \Phi = (\phi_1, \cdots, \phi_N) \in \mathscr{X}^N,\ \int_{\mathbb{R}^3} \phi_i\phi_j = \delta_{ij} \right\}.$$

Denoting by $C = [C_{\mu i}]_{1\leq\mu\leq N_n,\, 1\leq i\leq N}$ the matrix collecting the coefficients of the orbitals $\phi_1, \cdots, \phi_N$ in the basis $(\chi_1, \cdots, \chi_{N-b})$,

$$\phi_i(\mathbf{r}) = \sum_{\mu=1}^{N_b} C_{\mu i}\chi_\mu(\mathbf{r}),$$

the discretized Hartree-Fock model can be written as

$$E^{\mathrm{HF}}_{0,\mathscr{X}} = \inf\left\{ E^{\mathrm{HF}}(CC^T),\ C \in \mathbb{R}^{N_b\times N},\ C^T S C = I_N \right\},$$

where

$$E^{\mathrm{HF}}(D) = \mathrm{Tr}(hD) + \frac{1}{2}\mathrm{Tr}(G(D)D), \qquad [G(D)]_{\mu\nu} = \sum_{\kappa\lambda} \left[(\mu\nu|\kappa\lambda) - (\mu\lambda|\kappa\nu)\right] D_{\kappa\lambda},$$

and where the entries of the overlap matrix S, the core Hamiltonian matrix h, and the two-electron integrals $(\mu\lambda|\kappa\nu)$ are defined as

$$S_{\mu\nu} = \int_{\mathbb{R}^3} \chi_\mu \chi_\nu, \qquad h_{\mu\nu} = \frac{1}{2}\int_{\mathbb{R}^3} \nabla\chi_\mu \cdot \nabla\chi_\nu - \sum_{k=1}^{M} z_k \int_{\mathbb{R}^3} \frac{\chi_\mu(\mathbf{r})\chi_\nu(\mathbf{r})}{|\mathbf{r}-\mathbf{R}_k|}\, d\mathbf{r}, \tag{16}$$

and

$$(\mu\nu|\kappa\lambda) = \int_{\mathbb{R}^3}\int_{\mathbb{R}^3} \frac{\chi_\mu(\mathbf{r})\chi_\nu(\mathbf{r})\chi_\kappa(\mathbf{r}')\chi_\lambda(\mathbf{r}')}{|\mathbf{r}-\mathbf{r}'|}\, d\mathbf{r}\, d\mathbf{r}'. \tag{17}$$

A fundamental observation made by Boys in the 1950s [15] is that if the χ_μ's are gaussian-polynomial functions, i.e. functions of the form

$$\chi_\mu(\mathbf{r}) = p(\mathbf{r})\ \exp(-\alpha|\mathbf{r}|^2),$$

where p is a polynomial function and α a positive real number, then all the integrals in (16) and (17) can be computed analytically.

In practice, most calculations in quantum chemistry are performed using gaussian atomic orbital basis sets, which are built as follows:

1. a collection $\left\{\xi_\mu^A\right\}_{1\le\mu\le n_A}$ of n_A linearly independent linear combinations of gaussian polynomials are associated with each chemical element A of the periodic table: these are the atomic orbitals of A;
2. to perform a calculation on a given chemical system, one builds a basis $\left\{\chi_\mu\right\}$ by putting together all the atomic orbitals related to all the atoms of the system.

This approach is reminiscent of the reduced basis method used in other fields of science and engineering (see e.g. [41, 65] and references therein). For instance, still in the case of a water molecule H_2O, we have

$$\begin{aligned}\left\{\chi_\mu\right\} = \big\{&\xi_1^H(\mathbf{r}-\mathbf{R}_{H_1}), \cdots, \xi_{n_H}^H(\mathbf{r}-\mathbf{R}_{H_1}); \xi_1^H(\mathbf{r}-\mathbf{R}_{H_2}), \cdots, \xi_{n_H}^H(\mathbf{r}-\mathbf{R}_{H_2});\\ &\xi_1^O(\mathbf{r}-\mathbf{R}_O), \cdots, \xi_{n_O}^O(\mathbf{r}-\mathbf{R}_O)\big\},\end{aligned}$$

where $\mathbf{R}_{H_1}$, $\mathbf{R}_{H_2}$ and $\mathbf{R}_O$ denote the positions in $\mathbb{R}^3$ of the hydrogen nuclei and of the oxygen nucleus respectively.

To better understand the geometric nature of the discretized Hartree-Fock model, let us assume that the family $(\chi_\mu)_{1\le\mu\le N_b}$ is orthonormal. The discretized Hartree-Fock model can be written in two different ways:

- molecular orbital formulation

$$E^{\mathrm{HF}}_{0,\mathscr{X}} = \inf\left\{E^{\mathrm{HF}}(CC^T),\ C \in \mathscr{C}\right\}, \tag{18}$$

where

$$E^{\mathrm{HF}}(D) = \mathrm{Tr}(hD) + \frac{1}{2}\mathrm{Tr}(G(D)D),$$

and where

$$\mathscr{C} = \left\{C \in \mathbb{R}^{N_b \times N},\ C^T C = I_N\right\}$$

is a so-called Stiefel manifold;

- density matrix formulation

$$E^{\mathrm{HF}}_{0,\mathscr{X}} = \inf\left\{E^{\mathrm{HF}}(D),\ D \in P\right\}, \tag{19}$$

where

$$\mathscr{P} = \left\{D \in \mathbb{R}^{N_b \times N_b},\ D = D^T,\ \mathrm{Tr}(D) = N,\ D^2 = D\right\}$$

is the set of rank-N orthogonal projectors of $\mathbb{R}^{N_b \times N_b}$ and is called a Grassmann manifold.

The equivalence between (18) and (19) comes from the fact that when C varies in the set $\mathscr{C}$, $D = CC^T$ spans $\mathscr{P}$.

The Euler-Lagrange equations associated with (18) can be transformed as in the fifth statement of Theorem 27 by a unitary transform to diagonalize the Lagrange multiplier λ of the orthonormality constraints (λ is an $N \times N$ real symmetric matrix). We thus obtain the discretized Hartree-Fock equations (for the general case of a non-orthogonal basis)

$$\begin{cases} D = \displaystyle\sum_{i=1}^{N} \Phi_i \Phi_i^T, \\ \\ F = h + G(D), \\ \\ F\Phi_i = \varepsilon_i S\Phi_i, \quad \varepsilon_1 \leq \cdots \leq \varepsilon_N, \quad \Phi_i^T \Phi_j = \delta_{ij}, \end{cases} \tag{20}$$

where $\varepsilon_1 \leq \cdots \leq \varepsilon_N$ are the lowest N generalized eigenvalues (counting multiplicities) of the generalized eigenvalue problem

$$F\Phi = \varepsilon S\Phi,$$

and where

$$D \in \mathbb{R}^{N_b \times N_b}_{\text{sym}}, \quad F \in \mathbb{R}^{N_b \times N_b}_{\text{sym}}, \quad \Phi_i \in \mathbb{R}^{N_b},$$

respectively denote the discretizations of the density matrix, of the Fock operator, of the Hartree-Fock orbitals in the discretization basis $(\chi_1, \cdots, \chi_{N_b})$.

Solutions to the discretized Hartree-Fock problem can be obtained

- either by solving a constrained optimization problem (on a Stiefel or a Grassmann manifold [31, 49]);
- or by solving the above equations by means of a self-consistent field (SCF) algorithm (see [16] and references therein).

The design of more efficient methods, in particular for very large molecular systems, is still an active field of research.

Since the Hartree-Fock ground state energy for the nuclear configuration $\{\mathbf{R}_k\}$ is obtained by solving a constrained optimization problem depending parametrically on the $\{\mathbf{R}_k\}$, it also falls into the scope of formulas (3) and (4). A simple calculation shows that the effective forces in the discretized Hartree-Fock model are given by

$$-\nabla_{\mathbf{R}_k} W_0^{\text{HF}}(\mathbf{R}_1, \cdots, \mathbf{R}_M) = -\text{Tr}\left(\nabla_{\mathbf{R}_k} h D\right) - \text{Tr}\left(\nabla_{\mathbf{R}_k} S D_E\right) + \sum_{l \neq k} z_k z_l \frac{\mathbf{R}_k - \mathbf{R}_l}{|\mathbf{R}_k - \mathbf{R}_l|^3},$$

where D is the ground state density of the discretized Hartree-Fock model for the nuclear configuration $\{\mathbf{R}_k\}$ obtained by solving (20) and D_E is the energy weighted ground state density matrix defined by

$$D_E = \sum_{i=1}^{N} \varepsilon_i \Phi_i \Phi_i^T,$$

where the ε_i's and the Φ_i's are solutions to (20).

7 Numerical Approximation of Eigenvalues of Self-adjoint Operators

Let A be a self-adjoint operator on a Hilbert space $\mathscr{H}$ with domain $D(A)$ and form domain $Q(A)$, and a the associated quadratic form. The typical example we have in mind is the three-dimensional Schrödinger operator

$$\mathscr{H} = L^2(\mathbb{R}^3), \quad D(A) = H^2(\mathbb{R}^3), \quad A = -\frac{1}{2}\Delta + V, \quad V \in L^2_{\text{unif}}(\mathbb{R}^3).$$

The quadratic form associated with A is defined on the form domain $Q(A) = H^1(\mathbb{R}^3)$ by

$$\forall (u,v) \in Q(A) \times Q(A), \quad a(u,v) = \frac{1}{2}\int_{\mathbb{R}^3} \nabla u \cdot \nabla v + \int_{\mathbb{R}^3} Vuv.$$

Let $(V_n)_{n\in\mathbb{N}}$ be a sequence of finite-dimensional subspaces of $Q(A)$ such that

$$\forall v \in Q(A), \quad \inf_{v_n \in V_n} \|v - v_n\|_{Q(A)} \underset{n\to\infty}{\longrightarrow} 0.$$

For each n, we denote by $A|_{V_n}$ the self-adjoint operator on V_n defined by

$$\forall (u_n, v_n) \in V_n \times V_n, \quad (A|_{V_n} u_n, v_n)_{\mathscr{H}} = a(u_n, v_n).$$

The spectrum of $A|_{V_n}$ is obtained by solving the variational problem

$$\begin{cases} \text{search } (u_n, \lambda_n) \in V_n \times \mathbb{R} \text{ such that} \\ \forall v_n \in V_n, \quad a(u_n, v_n) = \lambda_n (u_n, v_n)_{\mathscr{H}} \\ \|u_n\|_{\mathscr{H}} = 1 \end{cases}$$

The question we would like to investigate in this section is the following: does $\sigma(A|_{V_n})$, the spectrum of $A|_{V_n}$, converge to $\sigma(A)$, the spectrum of A? Quite surprisingly, the answer to this question is no, in general.

Recall that, according to Theorem 23, if A is bounded below, then the real number

$$\lambda_j(A) = \inf_{W_j \in \mathscr{E}_j} \sup_{w \in W_j \setminus \{0\}} \frac{a(w,w)}{\|w\|^2},$$

where $\mathscr{E}_j$ is the set of the d-dimensional subspaces of $Q(A)$, is equal to

- the smallest jth eigenvalue of A if A has at least j eigenvalues lower than $\min \sigma_{\text{ess}}(A)$ (taking multiplicities into account);
- $\min \sigma_{\text{ess}}(A)$ otherwise.

From this result, we can infer the following classical results (see e.g. [6, 23]).

Theorem 28 *Let A be a bounded below self-adjoint operator on $\mathscr{H}$. Then*

$$\forall j \in \mathbb{N}^*, \qquad \lambda_j(A|_{V_n}) \underset{n\to\infty}{\downarrow} \lambda_j(A).$$

Theorem 29 *Let A be a bounded below self-adjoint operator on $\mathscr{H}$, $\lambda < \min \sigma_{\text{ess}}(A)$ a discrete eigenvalue of A of multiplicity m, and $\varepsilon > 0$ such that*

$$[\lambda - \varepsilon, \lambda + \varepsilon] \cap \sigma(A) = \{\lambda\}.$$

Let $P := \mathbb{1}_{\{\lambda\}}(A)$ *and* $P_n := \mathbb{1}_{[\lambda-\varepsilon/2,\lambda+\varepsilon/2]}(A|_{V_n})$. *Then, for* n *large enough,* $Rank(P_n) = m$ *and there exists* $C \in \mathbb{R}_+$ *such that*

$$\|(P - P_n)P\|_{\mathscr{B}(\mathscr{H},Q(A))} \leq C\|(1 - \Pi_{V_n}^{Q(A)})P\|_{\mathscr{B}(\mathscr{H},Q(A))},$$

$$\|(P - P_n)P_n\|_{\mathscr{B}(\mathscr{H},Q(A))} \leq C\|(1 - \Pi_{V_n}^{Q(A)})P\|_{\mathscr{B}(\mathscr{H},Q(A))},$$

$$\max_{\lambda_n \in \sigma(A|_{V_n}) \cap [\lambda-\varepsilon,\lambda+\varepsilon]} |\lambda_n - \lambda| \leq C\|(1 - \Pi_{V_n}^{Q(A)})P\|^2_{\mathscr{B}(\mathscr{H},Q(A))},$$

where $\Pi_{V_n}^{Q(A)}$ *is the orthogonal projection of* $Q(A)$ *on* V_n *for the* $Q(A)$*-scalar product.*

As previously mentioned, the spectrum of the discretized operator $A|_{V_n}$ does not, in general, converge to the spectrum of the original operator A. However, everything goes well if A is a bounded operator with compact resolvent.[12]

Theorem 30 *Assume that* A *is bounded below with compact resolvent. Then*

$$\lim_{n\to\infty} \sigma(A|_{V_n}) = \sigma(A).$$

More precisely,

- *the spectrum of* A *is purely discrete and the sequence* $(\lambda_j)_{j\in\mathbb{N}^*}$ *of the eigenvalues of* A *(counted with their multiplicities) forms a non-decreasing sequence going to* $+\infty$*;*
- *let* $\lambda_1^n \leq \lambda_2^n \leq \cdots \leq \lambda_{N_n}^n$ *denote the eigenvalues of* $A|_{V_n}$ *(counted with their multiplicities). Then*

$$\forall j \in \mathbb{N}^*, \quad \lambda_j^n \geq \lambda_j \text{ for all } n \in \mathbb{N} \text{ such that } N_n \geq j, \quad \text{and} \quad \lim_{n\to\infty} \lambda_j^n = \lambda_j.$$

Example 31 Let $\mathscr{H} = L^2(\mathbb{R}^d)$ and $V \in C^0(\mathbb{R}^d)$ such that $\lim_{|x|\to+\infty} V(x) = +\infty$ (confining potential). Consider the operator A defined as

$$D(A) = \left\{ u \in L^2(\mathbb{R}^d) \mid -\frac{1}{2}\Delta u + Vu \in L^2(\mathbb{R}^d) \right\},$$

and

$$\forall u \in D(A),\ Au = -\frac{1}{2}\Delta u + Vu.$$

[12]The operator A has a compact resolvent if, for some $z \in \rho(A)$ (and therefore for all $z \in \rho(A)$ by virtue of the resolvent formula), $z - A$, considered as a bounded operator on $\mathscr{H}$, is compact.

Then A is bounded below and has a compact resolvent. The spectrum of A therefore is an increasing sequence of eigenvalues of finite multiplicities going to $+\infty$, and Theorem 23 can be applied.

If more general situations, two different problems may occur. First, it may happen that

$$\sigma(A) \not\subseteq \liminf_{n\to\infty} \sigma(A|_{V_n}).$$

This is referred to as the lack of approximation problem. It may also happen that

$$\limsup_{n\to\infty} \sigma(A|_{V_n}) \not\subseteq \sigma(A).$$

This is called the spectral pollution problem.

Example 32 (Lack of Approximation Problem) Let $\mathscr{H} = L^2_{\rm per}((0,2\pi),\mathbb{C})$, $D(A) = H^1_{\rm per}((0,2\pi),\mathbb{C})$ and $A = -i\dfrac{d}{dx}$. The operator A is the momentum operator in one-dimensional quantum mechanics. Let $(e_k)_{k\in\mathbb{Z}}$ be the basis of the Fourier modes $(e_k(x) = (2\pi)^{-1/2}e^{ikx})$, and

$$V_n = \mathbb{C}e_{0,n} \oplus \mathbb{C}\tilde{e}_{0,n} \oplus \mathrm{Span}\,\{e_k,\ 1 \le |k| \le n-1\},$$

where

$$e_{0,n} := \cos(1/n)e_0 + \frac{\sin(1/n)}{\sqrt{2}}e_n + \frac{\sin(1/n)}{\sqrt{2}}e_{-n}, \qquad \tilde{e}_{0,n} = \frac{1}{\sqrt{2}}e_n - \frac{1}{\sqrt{2}}e_{-n}.$$

Then

$$\sigma(A) = \mathbb{Z} \qquad \text{and} \qquad \lim_{n\to\infty} \sigma(A|_{V_n}) = \mathbb{Z}^*,$$

which reveals a lack of approximation problem: the eigenvalue 0 of the operator A is missed by the numerical approximation.

The lack of approximation and spectral solutions problems are investigated from a mathematical point of view in the references [13, 27, 52, 72], from which we have extracted some important general results.

First, there is no risk of lack of approximation whenever the operator is semibounded, that is bounded from above, or bounded from below.

Theorem 33 *If A is semibounded, then $\sigma(A) \subset \liminf_{n\to\infty} \sigma(A|_{V_n})$.*

The following nice example of spectral pollution is due to Szegö.

Example 34 Let $V_{\rm per} \in L^\infty_{\rm per}((0,2\pi),\mathbb{R})$, $\mathscr{H} = L^2_{\rm per}((0,2\pi),\mathbb{C})$,

$$(Au)(x) = V_{\rm per}(x)u(x).$$

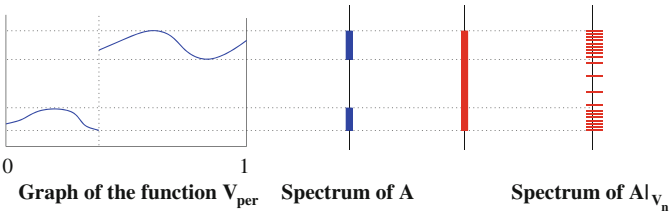


Fig. 12 A case of spectral pollution (Example 34)

Let $V_n = \text{Span}\{e_k,\ |k| \leq n\}$, where $(e_k)_{k\in\mathbb{Z}}$ is the Fourier basis. Then (see Fig. 12),

$$\sigma(A) = \text{ess-range}(V_{\text{per}}) \qquad \text{and} \qquad \lim_{n_0\to\infty} \overline{\bigcup_{n\geq n_0} \sigma(A|_{V_n})} = \text{CH}(\sigma(A)),$$

where $\text{CH}(B)$ denotes the convex hull of the set B.

Definition 35 A real number $\lambda \notin \sigma(A)$ such that there exists a sequence $(V_n)_{n\in\mathbb{N}}$ of finite-dimensional subspaces of $Q(A)$ such that

- $\forall v \in Q(A),\ \inf_{v_n\in V_n} \|v - v_n\|_{Q(A)} \underset{n\to\infty}{\longrightarrow} 0$
- $\lambda \in \lim_{n\to\infty} \sigma(A|_{V_n})$

is called a spurious eigenvalue of A. The set of the spurious eigenvalues of A is denoted by $\text{Spu}(A)$.

Theorem 36 *It holds*

$$\text{Spu}(A) = \text{CH}\left(\overline{\sigma(A)}^{\mathbb{R}} \setminus \sigma_{\text{d}}(A)\right) \setminus \sigma(A).$$

Let us illustrate the spectral pollution problem and the above theorem on the more physical case of a perturbed periodic Schrödinger operators on $L^2(\mathbb{R}^d)$. Such a situation notably arises in the modeling of crystals with point defects within density functional theory (DFT). Consider a periodic lattice $\mathscr{R}$ of $\mathbb{R}^d$, and the operator

$$H = -\frac{1}{2}\Delta + V_{\text{per}} + W, \tag{21}$$

with

$$V_{\text{per}} \in L^\infty(\mathbb{R}^d)\mathscr{R}\text{-periodic} \quad \text{and} \quad W \in L^\infty(\mathbb{R}^d), \quad \lim_{|x|\to\infty} W(x) = 0.$$

The operator H can be seen as a perturbation of the periodic Schrödinger operator

$$H_0 = -\frac{1}{2}\Delta + V_{\text{per}}.$$

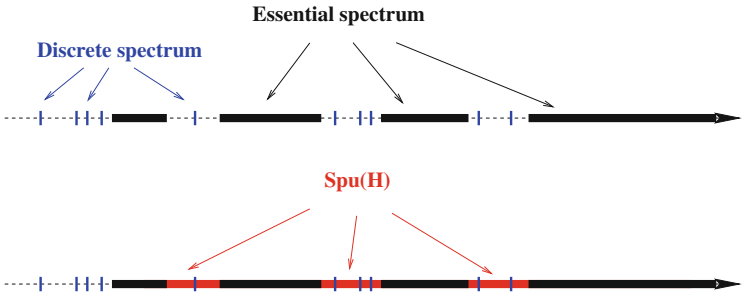


Fig. 13 Sketch of the spectrum of the perturbed periodic Schrödinger operator H defined by (21) (*top*) and of the set of the spurious eigenvalues of H (*bottom*) given by Theorem 36

It can be shown that the multiplication operator by the function W is H_0-compact. It therefore follows from Weyl's theorem (Theorem 21) that $\sigma_{\text{ess}}(H) = \sigma_{\text{ess}}(H_0)$. Besides, the spectrum of H_0 can be studied using Bloch's theory (see e.g. [68, Section XIII.16]). It turns out that H_0 is bounded below and that its spectrum is purely continuous (i.e. H_0 has no eigenvalues), and consists of bands: it is a countable union of possibly overlapping closed bounded intervals of $\mathbb{R}$.

In view of the previous results, there is no risk of lack of approximation since H is bounded below (cf. Theorem 33), but spectral pollution may be a problem (Fig. 13).

Quoting Boulton and Levitin [12], *the natural approach of truncating $\mathbb{R}^d$ to a large compact domain and applying the projection method to the corresponding Dirichlet problem is prone to spectral pollution*. Consider for instance the case when $d = 2$, $\mathscr{R} = 2\pi\mathbb{Z}^2$ (so that a unit cell is $[-\pi, \pi)^2$),

$$V_{\text{per}}(x, y) = \cos(x) + 3\sin(2(x + y) + 1),$$
$$W(x, y) = -(x + 2)^2(2y - 1)^2 \exp(-(x^2 + y^2)),$$

and the approximation spaces

$$V_n = \left\{ v_n \in C^0(\mathbb{R}^2) \mid \text{Supp}(v_n) \subset \Omega_n,\ \forall K_n \in \mathscr{T}_n^\infty,\ v_n|_{K_n} \in \mathbb{P}_1 \right\},$$

where the computational domain is defined as $\Omega_n = [-L_n/2, L_n/2]$ with $L_n \to \infty$, and the mesh $\mathscr{T}_n^\infty$ is a uniform $\mathscr{R}$-periodic mesh of $\mathbb{R}^2$ with $2n^2$ triangles per unit cell (see Fig. 14).

Numerical simulations using Bloch theory show that there is a gap $(-0.341, 0.016)$ between the first and second bands of the unperturbed operator $H^0_{\text{per}} = -\Delta + V_{\text{per}}$, and that $H = H^0_{\text{per}} + W$ has exactly one eigenvalue $\lambda \simeq -0.105$ in this gap. The spectral pollution problem can be clearly observed on Fig. 15. The eigenfunction associated with the approximation of the eigenvalue λ in the circle on Fig. 15 is plotted on Fig. 16. The one associated with the spurious approximation in the square on Fig. 15 is plotted on Fig. 17. We can see that the spurious eigenfunction

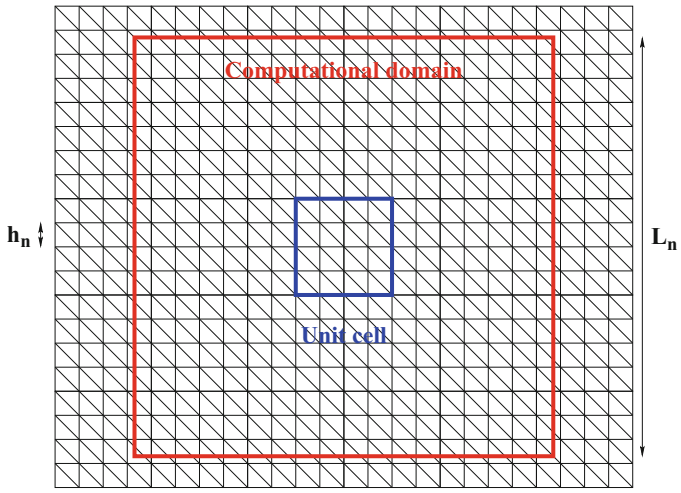


Fig. 14 Computation domain Ω_n and mesh $\mathcal{T}_n^\infty$

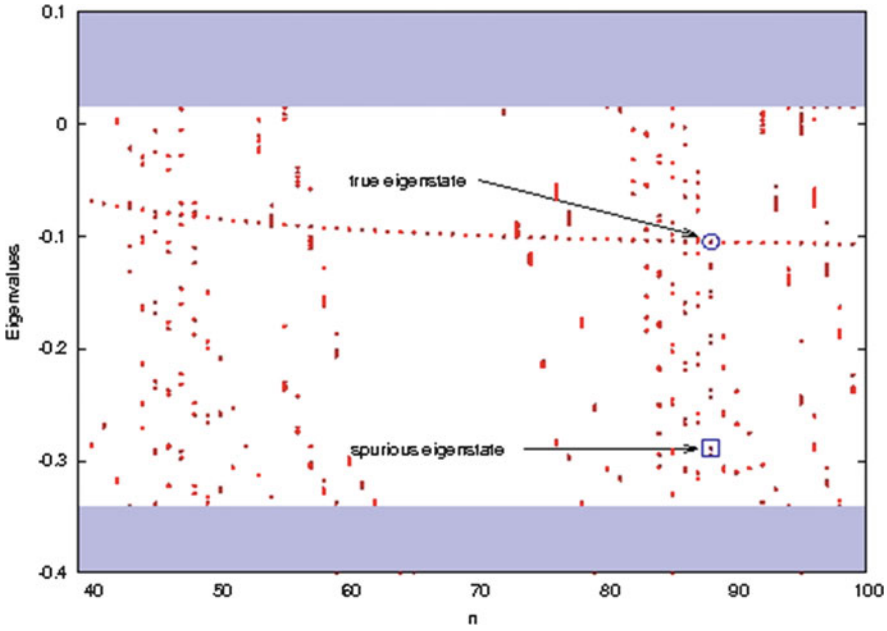


Fig. 15 Spectrum of $H|_{V_n}$ in the gap for $40 \leq n \leq 100$

seems to concentrate on the boundary of the simulation domain. A mathematical explanation of this phenomenon is given in [21] (see also [20]). These calculations were performed with FreeFEM++ [36].

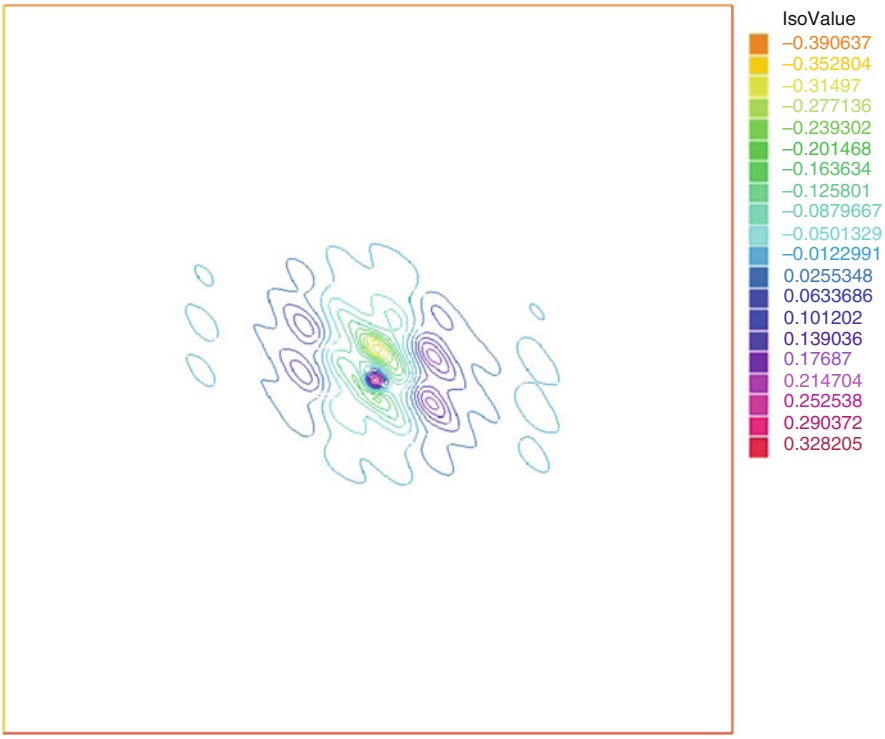


Fig. 16 Profile of a "true" eigenvector

Let us summarize the main messages of this section:

- variational approximations work well if the operator A is bounded below and has a purely discrete spectrum;
- if the operator is bounded below (resp. bounded above), variational approximations allow one to approximate the eigenvalues which are below the bottom (resp. above the top) of the essential spectrum;
- if the operator is bounded neither from below nor from above, variational approximations can lead to lack of approximation (some eigenvalues can be missed);
- variational approximation can give rise to spectral pollution in the "gaps" of the essential spectrum;
- in the latter two cases, the approximation spaces must be chosen very carefully.

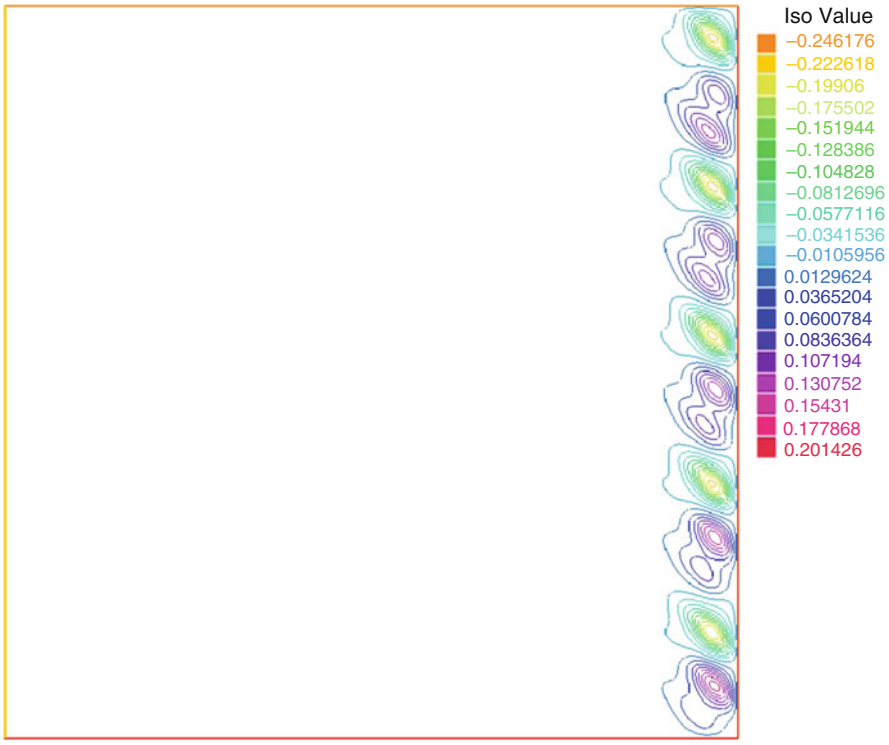


Fig. 17 Profile of a "spurious" eigenvector

Acknowledgements I am grateful to the organizers of the XVII Jacques-Louis Lions Spanish-French School on Numerical Simulation in Physics and Engineering for inviting me to deliver a course. Special thanks to Mariano Mateos Alberdi for the great local organization.

References

1. Ambrosio, L., Friesecke, G., Giannoulis, J.: Passage from quantum to classical molecular dynamics in the presence of Coulomb interactions. Commun. Partial Diff. Eqs. **35**, 1490–1515 (2010)
2. Ambrosio, L., Figalli, A., Friesecke, G., Giannoulis, J., Paul, T.: Semiclassical limit of quantum dynamics with rough potentials and well posedness of transport equations with measure initial data. Commun. Pure Appl. Math. **64**, 1199–1242 (2011)
3. Amrein, W.O., Georgescu, V.: On the characterization of bound states and scattering states in quantum mechanics. Helv. Phys. Acta **46**, 635–658 (1973/1974)
4. Anantharaman, A., Cancès, E.: Existence of minimizers for Kohn-Sham models in quantum chemistry. Ann. Inst. Henri. Poincaré, **26**, 2425–2455 (2009)
5. Arveson, W.: An Invitation to C*-Algebra. Springer, Berlin (1976)
6. Babuška, I., Osborn, J.: Eigenvalue problems. In Ciarlet, P.G., Lions, J.-L. (eds.) Handbook of Numerical Analysis, vol. II, pp. 641–787. North-Holland, Amsterdam (1991)

7. Bach, V., Delle Site, L. (eds.): Many-Electron Approaches in Physics, Chemistry and Mathematics. A Multidisciplinary View. Springer, Heidelberg/NewYork (2014)
8. Bach, V., Lieb, E.H., Loss, M., Solovej, J.P.: There are no unfilled shells in unrestricted Hartree-Fock theory. Phys. Rev. Lett. **72**, 2981–2983 (1994)
9. Balian, R.: From Microphysics to Macrophysics, Methods and Applications of Statistical Physics. Springer, Berlin (2007)
10. Becke, A.D.: Density-functional thermochemistry. III. The role of exact exchange. J. Chem. Phys. **98**, 5648–5652 (1993)
11. Bergeson, S.D. et al.: Measurement of the He Ground State Lamb Shift via the Two-Photon 1S1-2S1 Transition. Phys. Rev. Lett. **80**, 3475–3478 (1998)
12. Boulton, L., Levitin, M.: On the approximation of the eigenvalues of perturbed periodic Schrödinger operators. J. Phys. A **40**, 9319–9329 (2007)
13. Boulton, L., Boussaïd, N., Lewin, M.: Generalised Weyl theorems and spectral pollution in the Galerkin method. J. Spectral Theory **2**, 329–354 (2012)
14. Bourquin, R., Gradinaru, V., Hagedorn, G.A.: Non-adiabatic transitions near avoided crossings: theory and numerics. J. Math. Chem. **50**, 602–619 (2012)
15. Boys, S.F: Electronic wave functions: I. A general method of calculation for the stationary states of any molecular system. Proc. R. Soc. Lond. **A 200**, 542–554 (1950)
16. Cancès, E.: Self-consistent field (SCF) algorithms. In Engquist, B. (ed.) Encyclopedia of Applied and Computational Mathematics. Springer, Berlin/Heidelberg (2015)
17. Cancès, E., Le Bris, C.: On the convergence of SCF algorithms for the Hartree-Fock equations. ESAIM: M2AN **34**, 749–774 (2000)
18. Cancès, E., Defranceschi, M., Kutzelnigg, W., Le Bris, C., Maday, Y.: Computational quantum chemistry: A primer. In Ciarlet, P., Le Bris, C. (eds.) Handbook of Numerical Analysis. Computational chemistry, vol. X, pp. 3–270. North-Holland, Amsterdam (2003)
19. Cancès, E., Chakir, R., Maday, Y.: Numerical analysis of the planewave discretization of orbital-free and Kohn-Sham models. ESAIM: M2AN **46**, 341–388 (2012)
20. Cancès, E., Ehrlacher, V., Maday, Y.: Periodic Schrödinger operators with local defects and spectral pollution. SIAM J. Numer. Anal. **46**, 3016–3035 (2012)
21. Cancès, E., Ehrlacher, V., Maday, Y.: Non-consistent approximations of self-adjoint eigenproblems: application to the supercell method. Numer. Math. **28**, 663–706 (2014)
22. Ceperley, D.M.: Path integrals in the theory of condensed helium. Rev. Mod. Phys. **67**, 279–355 (1995)
23. Chatelin, F.: Spectral Approximation of Linear Operators. Academic Press, New York (1983)
24. Chen, H., Gong, X., He, L, Yang, Z., Zhou, A.: Numerical analysis of finite dimensional approximations of Kohn-Sham models. Adv. Comput. Math. **38**, 225–256 (2013)
25. Chorin, A.J., Marsden, J.E.: A Mathematical Introduction to Fluid Mechanics. Springer, New York (1979)
26. Cotar, C., Friesecke, G., Klüppelberg, C.: Density functional theory and optimal transportation with Coulomb cost. Commun. Pure Appl. Math. **66**, 548–599 (2013)
27. Davis, E.B., Plum, M.: Spectral pollution. IMA J. Numer. Anal. **24**, 417–438 (2004)
28. Dirac, P.A.M.: Quantum mechanics of many-electron systems. Proc. Royal Soc. Lond. Ser. A **123**, 714–733 (1929)
29. Dreizler, R., Gross, E.K.U: Density Functional Theory. Springer, Berlin/Heidelberg (1990)
30. Dyall, K.G., Faegri, K.: Introduction to Relativistic Quantum Chemistry. Oxford University Press, New York (2007)
31. Edelman, A., Arias, T.A., Smith, S.T.: The geometry of algorithms with orthogonality constraints. SIAM J. Matrix Anal. Appl. **20**, 303–353 (1998)
32. Eikema, K.S.E., Ubachs, W., Vassen, W., Hogervorst, W.: Lamb shift measurement in the $1t^1S$ ground state of helium. Phys. Rev. **A 55**, 1866–1884 (1997)
33. Enss V.: Asymptotic completeness for quantum mechanical potential scattering. I. Short range potentials. Commun. Math. Phys. **61**, 285–291 (1978)
34. Fiolhais, C. Nogueira, F., Marques, M.A.L. (eds.): A Primer in Density Functional Theory. Lecture Notes in Physics, vol. 620. Springer, Berlin/New York (2003)

35. Fermanian Kammerer, C., Gérard, P., Lasser, C.: Wigner measure propagation and Lipschitz conical singularity for general initial data. Arch. Ration. Mech. Anal. **209**, 209–236 (2013)
36. FreeFEM++ Finite Element Software, Version v.53-1. http://www.freefem.org/. Released on 10 May 2017
37. Friesecke, G.: The multiconfiguration equations for atoms and molecules: charge quantization and existence of solutions. Arch. Rat. Mech. Anal. **169**, 35–71 (2003)
38. García-Cervera, C.J., Lu, J., Xuan, Y., E, W.: A linear scaling subspace iteration algorithm with optimally localized non-orthogonal wave functions for Kohn-Sham density functional theory. Phys. Rev. B **79**, 115110 (2009)
39. Griesemer, M., Hantsch, F.: Unique solutions to Hartree-Fock equations for closed shell atoms. Arch. Ration. Mech. Anal. **203**, 883–900 (2012)
40. Helgaker, T., Jorgensen, P., Olsen, J.: Molecular Electronic-Structure Theory. Wiley, New York (2000)
41. Hesthaven, J.S., Rozza, G., Stamm, B.: Certified Reduced Basis Methods for Parametrized Partial Differential Equations. Springer, New York (2016)
42. Hunziker, M.: On the spectra of Schrodinger multiparticle Hamiltonians. Helv. Phys. Acta **39**, 451–462 (1966)
43. Karplus, M.: Development of multiscale models for complex chemical systems from H+H2 to biomolecules. Nobel Lecture delivered on December 8, 2013
44. Kato, T.: Perturbation Theory for Linear Operators. Springer, Berlin (1995)
45. Kohn, W.: Electronic structure of matter – Wave functions and density functionals. In Nobel Lectures, Chemistry 1996–2000. World Scientific Publishing, Singapore (2003)
46. Kohn, W., Sham, L.J.: Self-consistent equations including exchange and correlation effects. Phys. Rev. **140**, A1133 (1965)
47. Korobov, V., Yelkhovsky, A.: Ionization potential of the helium atom. Phys. Rev. Lett. **87**, 193003 (2001)
48. Le Bris, C.: A general approach for multiconfiguration methods in quantum molecular chemistry. Ann. Inst. Henri. Poincaré **11**, 441–484 (1994)
49. Levitt, A.: Convergence of gradient-based algorithms for the Hartree-Fock equations. ESAIM: M2AN **46**, 1321–1336 (2012)
50. Levitt M.: Birth and future of multiscale modeling for macromolecular systems. Nobel Lecture delivered on December 8, 2013
51. Lewin, M.: Solutions of the multiconfiguration equations in quantum chemistry. Arch. Ration. Mech. Anal. **171**, 83–114 (2004)
52. Lewin, M., Séré, E.: Spectral pollution and how to avoid it (with applications to Dirac and periodic Schrödinger operators). Proc. Lond. Math. Soc. **100**, 864–900 (2010)
53. Lieb, E.H., Thomas-Fermi and related theories of atoms and molecules. Rev. Mod. Phys. **53**, 603–641 (1981)
54. Lieb, E.H.: Density functional for coulomb systems. Int. J. Quantum Chem. **24**, 243–277 (1983)
55. Lieb, E.H., Simon, B.: The Thomas-Fermi theory of atoms, molecules and solids. Adv. Math. **23**, 22–116 (1977)
56. Lieb, E.H., Simon, B.: The Hartree-Fock theory for Coulomb systems. Commun. Math. Phys. **53**, 185–194 (1977)
57. Lin, L., Lu, J., Ying, L., Car, R., E, W.: Fast algorithm for extracting the diagonal of the inverse matrix with application to the electronic structure analysis of metallic systems. Commun. Math. Sci. **7**, 755–777 (2009)
58. Lin, L., Chen, M., Yang, C., He, L.: Accelerating atomic orbital-based electronic structure calculation via pole expansion and selected inversion. J. Phys. Condens. Matter **25**, 295501 (2013)
59. Lions, P.-L.: Solutions of Hartree-Fock equations for Coulomb systems. Commun. Math. Phys. **109**, 33–97 (1987)
60. Lu, J., Otto, F.: Nonexistence of minimizer for Thomas-Fermi-Dirac-von Weizsacker model. Commun. Pure Appl. Math. **67**, 1605–1617 (2014)

61. Panati, G., Spohn, H., Teufel, S.: The time-dependent Born-Oppenheimer approximation. ESAIM: M2AN **41**, 297–314 (2007)
62. Perdew, J.P., Zunger, A.: Self-interaction correction to density-functional approximations for many-electron systems. Phys. Rev. B **23**, 5048–5079 (1981)
63. Perdew, J.P., Burke, K., Ernzerhof, M.: Generalized gradient approximation made simple. Phys. Rev. Lett. **77**, 3865–3868 (1996)
64. Pople, J.: Quantum Chemical Models. In Nobel Lectures, Chemistry 1996–2000. World Scientific Publishing, Singapore (2003)
65. Quarteroni, A., Manzoni, A., Negri, F.: Reduced Basis Methods for Partial Differential Equations: An introduction. Springer, Berlin (2016)
66. Rader, T.: Theory of Microeconomics. Academic Press, New York (1972)
67. Reed, M., Simon, B.: Methods of Modern Mathematical Physics II: Fourier Analysis, Self-adjointness. Academic Press, New York/London (1975)
68. Reed, M., Simon, B.: Methods of Modern Mathematical Physics IV: Analysis of Operators. Academic Press, New York/London (1978)
69. Reed, M., Simon, B.: Methods of Modern Mathematical Physics I: Functional Analysis. Academic Press, New York/London (1980)
70. Ruelle D.: A remark on bound states in potential-scattering theory. Nuovo Cimento **A 61**, 655–662 (1969)
71. Schneider, R.: Analysis of the projected coupled cluster method in electronic structure calculation. Numer. Math. **113**, 433–471 (2009)
72. Shargorodsky, E.: Geometry of higher order relative spectra and projection methods. J. Operator Theory **44**, 43–62 (2000)
73. Tao, J.M., Perdew, J.P., Staroverov, V.N., Scuseria, G.E.: Climbing the density functional ladder: nonempirical meta-generalized gradient approximation designed for molecules and solids. Phys. Rev. Lett. **91**, 146401 (2003)
74. Teller, E.: On the stability of molecules in the Thomas-Fermi theory. Rev. Mod. Phys. **34**, 627–631 (1962)
75. Toulouse, J., Colonna, F., Savin, A.: Long-range/short-range separation of the electron-electron interaction in density-functional theory. Phys. Rev. A **70**, 062505 (2005)
76. Van Winter, C.: Theory of finite systems of particles. I. Mat.-Fys. Skr. Danske Vid. Selsk **1**, 1–60 (1960)
77. Warshel, A.: Multiscale modeling of biological functions: from enzymes to molecular machines. Nobel Lecture delivered on December 8, 2013
78. Zhislin, G.M.: Investigation of the spectrum of the Schrodinger operator for a many particle system. Trudy Moskov. Mat. Ob-va **9**, 81–120 (1960)
79. Zhislin, G.M., Sigalov, A.G.: The spectrum of the energy operator for atoms with fixed nuclei on subspaces corresponding to irreducible representations of the group of permutations. Izv. Akad. Nauk SSSR Ser. Mat. **29**, 835–860 (1965)

Accurate Computations and Applications of Some Classes of Matrices

J.M. Peña

Abstract Performing an algorithm with high relative accuracy is a very desirable goal. High relative accuracy means that the relative errors of the computations are of the order of machine precision, independently of the size of the condition number. This goal is difficult to assure although in recent years there have been some advances, in particular in the field of Numerical Linear Algebra. Up to now, computations with high relative accuracy are guaranteed only for a few classes of matrices, mainly for some subclasses of M-matrices and for some subclasses of totally positive matrices. Previously, a reparametrization of the matrices is needed. We review this procedure related with the high relative accuracy computations of these matrices. We also present some recent applications of the two classes of matrices mentioned previously. On the one hand, applications of M-matrices to the linear complementarity problem. On the other hand, applications of totally positive matrices to Computer Aided Geometric Design.

1 Introduction

This paper surveys some recent advances on high relative accuracy when working with some classes of matrices. Performing an algorithm with high relative accuracy (HRA) is a very desirable goal. Recent research in Numerical Linear Algebra has shown that certain classes of matrices allow us to perform many computations to HRA, independently of the size of the condition number. For instance, the computation of their singular values, eigenvalues or inverses. These classes of matrices are defined by special sign or other structure and require to know some natural parameters to HRA, and they are related to some subclasses of P-matrices. Let us recall that a square matrix is called a P-matrix if all its principal minors are positive (the principal minors use the same rows and columns). Subclasses of P-matrices with many applications are the nonsingular totally positive matrices and

J.M. Peña (✉)
Department of Applied Mathematics/IUMA, Universidad de Zaragoza, Zaragoza, Spain
e-mail: jmpena@unizar.es

M. Mateos, P. Alonso (eds.), *Computational Mathematics, Numerical Analysis and Applications*, SEMA SIMAI Springer Series 13,
DOI 10.1007/978-3-319-49631-3_3

the nonsingular M-matrices. Some recent applications of these matrices are also recalled in this paper.

Let us now present the layout of the paper. Section 2 recalls some basic concepts related with the errors obtained when computing with floating point arithmetic. In general, the forward error bound (difference between the computed and the exact solution) is obtained through the backward error and the condition number of our problem. However, in some problems it is possible to find a parametrization of the data and an algorithm leading to small forward error bounds in spite of a bad conditioning with its initial parametrization. In these cases, we can assure HRA. We mention simple algorithms that cannot be computed with HRA and simple structured classes for which we cannot guarantee computations with HRA. We also present a simple sufficient condition (NIC: no inaccurate cancellation) to assure that an algorithm can be performed with HRA. The remaining sections deal with classes of matrices for which an adequate parametrization permits NIC algorithms.

Section 3 presents the class of P-matrices and some subclasses of P-matrices (see [86]). In a Linear Complementarity (LC) problem, there exists a unique solution if and only if the associated matrix is a P-matrix. We shall also present some subclass of P-matrices, including the class of nonsingular M-matrices (matrices with nonpositive off-diagonal entries and nonnegative inverse). M-matrices present important applications in Numerical Analysis, dynamic systems, Economy or Optimization (including the LC problem mentioned above). Several characterizations of nonsingular M-matrices are recalled. We also present other related classes of matrices such as diagonally dominant matrices and H-matrices (which are generalized diagonally dominant matrices). Finally, we recall the LC problem and we recall some recent error bounds for this problem obtained when the associated matrix A is an H-matrix with positive diagonal entries, which are valid in particular when A is a nonsingular M-matrix. Let us also mention that error bounds for the LC problem have been improved for some subclasses of nonsingular M-matrices (see [50, 51]) and have also obtained for other subclasses of M-matrices different of nonsingular M-matrices (see [47, 49, 52]).

Section 4 considers one of the classes of matrices for which algorithms with HRA have been obtained: the class of diagonally dominant M-matrices. First we recall the important concept of rank revealing decomposition. Recall that if we have a rank revealing decomposition of a matrix A with HRA, then we can apply the results of [39] to derive algorithms for finding the singular values of A with HRA. In the case of a diagonally dominant M-matrix, an LDU-decomposition obtained with an adequate pivoting strategy will provide the rank revealing decomposition. We recall several ways of obtaining such LDU-factorization with HRA and how the pivoting strategy can be implemented in a very economic way.

A matrix is said to be totally positive (TP) if all its minors are nonnegative. These matrices present important applications (see [8, 43, 46, 53, 64, 91]) in many fields such as Approximation Theory, Economics, Combinatorics, Mechanics, Statistics, Differential Equations or Computer Aided Geometric Design (CAGD). Section 5 presents TP matrices and considers some problems related with their parametrization to obtain algorithms with HRA. There are many remarkable

properties of TP matrices, such as spectral properties or variation diminishing properties. However, the important property in the context of accurate computations is their bidiagonal factorization. We recall the history of this factorization and its role to parametrize nonsingular TP matrices. Given the bidiagonal factorization of a nonsingular TP matrix with HRA, then one can derive algorithms with HRA to obtain all eigenvalues, all singular values and even the inverse of the matrix. In order to construct the bidiagonal factorization of a TP matrix, Neville elimination can be used. Neville elimination is a procedure to create zeroes in a matrix alternative to Gaussian elimination. The elementary operations of Neville elimination always add to a row a multiple of the previous one, instead of a multiple of the pivot row as in Gaussian elimination. Neville elimination allows us to check if a given matrix is TP with a computational cost similar to that Gaussian elimination (see [55] and [87]). Section 5.1 presents Neville elimination and shows how it can be used to obtain the bidiagonal factorization of a nonsingular TP matrix. In fact, we provide an explicit bidiagonal factorization of a nonsingular TP matrix in terms of the multipliers and diagonal pivots of the Neville elimination. Error analysis of Neville elimination has been considered in [4, 5] and [6]. In Sect. 5.2 we show how to extend the accurate computation for nonsingular TP matrices to the larger class of signed bidiagonal decomposition (SBD) matrices, which contains nonsingular TP matrices as well as their inverses.

Section 6 shows some applications of TP matrices to the field of CAGD. We recall that shape preserving representations of curves are associated with (normalized) totally positive bases, which are bases such that their collocation matrices are (stochastic) TP (see [81]). We also recall corner cutting algorithms, which form the main source of algorithms in CAGD and that present a matrix form as a bidiagonal factorization. Other important applications of tTP matrices to CAGD is related with the problem of finding bases with optimal shape preserving properties (which correspond to the concept of normalized B-bases) and with the problem of recognizing normalized totally positive bases (and so, shape preserving representations). As for this last problem, we have a shape preserving representation (which would be associated to a normalized totally positive basis) if the matrix of change of basis with respect to the normalized B-basis of the space is totally positive. So, checking the total positivity of a unique matrix implies the total positivity of the basis and so of its infinite collocation matrices.

Examples of normalized B-bases are the Bernstein basis and the B-spline basis. Normalized B-bases satisfy more optimal properties (see [18, 24]). In the particular case of the Bernstein basis, we also have optimal conditioning of its collocation matrices with respect to the collocation matrices of another normalized totally positive basis (see [27]). As for the important problem of the numerical evaluation of curves and surfaces in CAGD, B-bases also present important advantages. In fact, B-bases and beyond Total Positivity, more general bases present optimal stability properties for the evaluation among all bases of nonnegative functions (see [72, 82–84]). The stability of the corresponding evaluation algorithms has been also deeply studied (see [13, 14, 23, 25, 26, 28–30, 33, 35, 73, 88, 89]).

It is well known that, if we have the bidiagonal decomposition $\mathscr{BD}(A)$ of a nonsingular TP matrix with HRA, then we can perform many computations of A with HRA, such as computing its inverse or computing its eigenvalues or its singular values (cf. [68]). Therefore, the entries of the bidiagonal factorization (17) are the adequate parameters for nonsingular TP matrices. There are several subclasses of nonsingular TP matrices for which this factorization can be obtained to HRA (and so, the computations mentioned previously, too). For instance, the mentioned algebraic computations can be performed with HRA for the following subclasses of TP matrices: Vandermonde positive matrices [38], Bernstein-Vandermonde matrices [74], Said-Ball-Vandermonde matrices [75], Pascal matrices [7], Jacobi-Stirling matrices [32], some rational collocation matrices [31], q-Bernstein-Vandermonde matrices [34] (these last three cases are considered in Sect. 7 and Schoenmakers-Coffey matrices [36].

Rational Bernstein bases play a very important role in CAGD and their collocation matrices are called rational Bernstein-Vandermonde matrices. In Sect. 7.1 we present the construction of the bidiagonal factorization of rational Bernstein-Vandermonde matrices with HRA.

The basis of q-Bernstein polynomials has been introduced recently with some advantages for the curve design. Its collocation matrices are called q-Bernstein-Vandermonde matrices. In Sect. 7.2 we present the construction of the bidiagonal factorization of conversion of a q-Bernstein-Vandermonde matrix with HRA, and we show how it can be used to compute its inverse or its eigenvalues and singular values with HRA.

Finally, in Sect. 7.3 we consider Jacobi-Stirling matrices, which play an important role in Combinatorics. We present the construction of the bidiagonal factorization of with HRA.

2 Errors and High Relative Accuracy

If x is a real number that can be calculated through an algorithm, let us denote by $\hat{x}$ the corresponding computed number with floating point arithmetic. Then the absolute error performed to compute $\hat{x}$ is given by $E_{abs}(\hat{x}) = |x - \hat{x}|$ and the corresponding relative error when $x \neq 0$ is

$$E_{rel}(\hat{x}) = \frac{|x - \hat{x}|}{|x|}.$$

Analogously, if x is a real vector that can be calculated through an algorithm, let us denote by $\hat{x}$ the corresponding computed vector with floating point arithmetic. Then the relative error performed to compute $\hat{x}$ (with x a nonzero vector) is given by

$$E_{rel}(\hat{x}) = \frac{\|x - \hat{x}\|}{\|x\|}.$$

However, if x has nonzero components and we do not want to miss the computed error corresponding to the components of x with least absolute value, then we can consider a componentwise relative error given by $\max_i \frac{|x_i - \hat{x}_i|}{|x_i|}$.

Since we do not know the exact error performed with our computations, it is convenient to try to derive upper bounds of this error, usually known as *forward* error bounds. However, it is usually difficult to obtain directly such bounds. An alternative approach that has been very successful in the field of Numerical Linear Algebra and other fields tries to obtain the forward error bounds through the backward errors. Let us introduce this last concept. This alternative approach considers that our computed solution is the exact solution of a perturbed problem, and the backward error measures the distance between the perturbed problem and the initial problem. For instance, consider $y = f(x)$, a continuous real function, and $\hat{y}$ a numerical approximation to f at a point x. Then let us consider the set of values $x + \triangle x$ for which $\hat{y}$ is the exact value:

$$\hat{y} = f(x + \triangle x),$$

and we consider the least $|\triangle x|$, which is called *backward* error. If for all x, the value $|\triangle x|$ is small (in the context of our problem), then we say that our method is *backward* stable. Backward stability plays an important role for designing a good algorithm.

It is well known that the growth factor of an algorithm is an indicator of its stability (cf. [58]). The *growth factor* of a numerical algorithm is usually defined as the quotient between the maximal absolute value of all the elements that occur during the performance of the algorithm and the maximal absolute value of all the initial data.

Backward and forward errors are related by the conditioning of the problem, which measures the effect of data perturbations on the solution of the problem.

In general, when for a given problem we have defined the corresponding forward error, backward error and the condition number, one tries to prove the relation:

$$\text{forward error} \ \leq \ \text{condition number} \times \text{backward error},$$

which allows us to derive a forward error bound through the backward error. Although the computed solution has a small backward error, it can be amplified by the condition number leading to a large forward error. So, in contrast to the backward error, which depends of the used method, the conditioning can become an intrinsic cause to obtain a nice forward error bound. However, in some problems it is possible to find a parametrization of the data and an algorithm leading to small forward error bounds in spite of a bad conditioning with its initial parametrization. The desired goal is to guarantee *high relative accuracy* (HRA). We say that we have performed an algorithm with HRA if the following formula holds:

$$\text{relative forward error} \leq Ku, \ \text{for some constant } K,$$

where u is the unit roundoff.

Is it always possible to guarantee HRA for a given problem? The answer is NO. An example of a simple problem for which an HRA algorithm cannot be found is provided by the evaluation of three real numbers $x + y + z$ (see [40]). We have announced that for some structured classes of matrices, HRA algorithms can be found. However, this is not always possible. For instance, accurate linear algebra for the problem of calculating determinants or minors is impossible on the class of Toeplitz matrices (see Corollaries 3.43 and 3.45 of [40]). Let us recall that a Toeplitz matrix B has the following simple structure:

$$B = \begin{pmatrix} a_0 & a_1 & \cdots & a_{n-2} & a_{n-1} \\ a_{-1} & a_0 & \ddots & & a_{n-2} \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ a_{-n+2} & & \ddots & \ddots & a_1 \\ a_{-n+1} & a_{-n+2} & \cdots & a_{-1} & a_0 \end{pmatrix}.$$

There exists a sufficient condition to assure the HRA of an algorithm that we now recall. Given an algorithm using only additions of numbers of the same sign, multiplications and divisions, and assuming that each initial real datum is known to HRA, then it is well-known that the output of that algorithm can be computed to HRA (cf. [39, p. 52]). Moreover, in (well-implemented) floating point arithmetic HRA is also preserved even when we perform true subtractions when the operands are original (and so, exact) data (cf. p. 53 of [39]). So, the sufficient condition to assure the HRA of an algorithm is satisfied if it only uses additions of numbers of the same sign, multiplications, divisions and subtractions (additions of numbers of different sign) of the initial data. This condition is called "no inaccurate cancellation" (NIC).

In order to find algorithms satisfying the NIC condition for some classes of matrices, an idea that has played a crucial role in some recent works has been the need to reparametrize matrices belonging to these special classes. This topic will be considered in the following sections.

3 *P*-Matrices, *M*-Matrices, Diagonal Dominance and Applications to LC Problems

Recent research in Numerical Linear Algebra has shown that certain classes of matrices allow us to perform many computations to HRA, independently of the size of the condition number. For instance, the computation of their singular values, eigenvalues or inverses. These classes of matrices are defined by special sign or other structure and require to know some natural parameters to HRA, and they are related to some subclasses of P-matrices. Let us recall that a square matrix is

called a P-matrix if all its principal minors are positive (the principal minors use the same rows and columns). Subclasses of P-matrices with many applications are the nonsingular TP matrices and the nonsingular M-matrices. Usually, accurate spectral computation (eigenvalues, singular values) or accurate inversion is assured when an accurate matrix factorization with a suitable pivoting is provided. For instance, the bidiagonal decomposition in the case of TP matrices (see [68]) or an LDU factorization after a symmetric pivoting in the case of diagonally dominant matrices (cf. [37, 85]).

Let us now introduce some other classes of matrices used in this section. A real matrix with nonpositive off-diagonal entries is called a *Z-matrix*. We say that a matrix $A = (a_{ij})_{1\le i,j\le n}$ is *(row) diagonally dominant* (resp., *strictly (row) diagonally dominant*) if, for each $i = 1, \ldots, n$, $|a_{ii}| \ge \sum_{j\ne i} |a_{ij}|$ (reps., $|a_{ii}| > \sum_{j\ne i} |a_{ij}|$). If A^T is row diagonally dominant, then we say that A is column diagonally dominant. Given a matrix $M = (m_{ij})_{1\le i,j\le n}$, its *comparison matrix* $\tilde{M} = (\tilde{M}_{ij})_{1\le i,j\le n}$ is the Z-matrix defined by $m_{ii} := |m_{ii}|$ and $m_{ij} := -|m_{ij}|$ if $i \ne j$, $1 \le i,j \le n$. Let us recall that if a Z-matrix A can be expressed as $A = sI - B$, with $B \ge 0$ and $s \ge \rho(B)$ (where $\rho(B)$ is the spectral radius of B), then it is called an *M-matrix*. Let us also recall that a Z-matrix A is a nonsingular M-matrix if and only if A^{-1} is nonnegative. Nonsingular M-matrices have important applications, for instance, in iterative methods in numerical analysis, in the analysis of dynamical systems, in economics and in mathematical programming. Finally, we say that a matrix is an *H-matrix* if its comparison matrix is a nonsingular M-matrix.

Nonsingular M-matrices have many equivalent definitions. In fact, Berman and Plemmons (see Theorem 2.3 in Chap. 6 of [15]) list 50 equivalent definitions. We shall use the following equivalent definitions:

Definition 1 Let A be a real $n \times n$ matrix with nonpositive off-diagonal elements. Then the following concepts are equivalent:

(i) A is an *M-matrix*.
(ii) A^{-1} is nonnegative.
(iii) The principal minors of A are strictly positive.
(iv) $Ax \ge 0$ implies $x \ge 0$ for all $x \in \mathbf{R}^n$.

A remarkable property of P-matrices is that the linear complementarity problem has always a unique solution if and only if the associate matrix M is a P-matrix. Let us now present the linear complementarity (LC) problem. The LC problem consists of finding vectors $x \in \mathbf{R}^n$ satisfying

$$Mx + q \ge 0, \quad x \ge 0, \quad x^T(Mx + q) = 0, \tag{1}$$

where M is an $n \times n$ real matrix and $q \in \mathbf{R}^n$. We denote this problem by LCP(M, q) and its solutions by x^*. Many problems can be posed in the form (1). For instance, problems in linear and quadratic programming, the problem of finding a Nash equilibrium point of a bimatrix game or some free boundary problems of fluid mechanics (see Chap. 10 of [15, 77] and [21], and references therein).

It is well-known that an H-matrix with positive diagonals is a P-matrix (see, for instance, Theorem 2.3 of Chap. 6 of [15]) and that a strictly diagonally dominant matrix is an H-matrix. In [77], error bounds for $\|x - x^*\|$ were derived when M in (1) is a P-matrix. When M in (1) is an H-matrix with positive diagonals, sharper error bounds were obtained in [20], as we now recall.

Let M be an H-matrix with positive diagonal entries. Since M is a P-matrix, we can apply the third inequality of Theorem 2.3 of [20] and obtain for any $x \in \mathbf{R}^n$ the inequality:

$$\|x - x^*\|_\infty \leq \max_{d\in[0,1]^n} \|(I - D + DM)^{-1}\|_\infty \|r(x)\|_\infty,$$

where I is the $n \times n$ identity matrix, D the diagonal matrix $D = \text{diag}(d_i)$ with $0 \leq d_i \leq 1$ for all $i = 1, \ldots, n$, x^* is the solution of the LCP(M, q) and $r(x) := \min(x, Mx + q)$, where the min operator denotes the componentwise minimum of two vectors.

By (2.4) of [20], given in Theorem 2.1 of [20], when $M = (m_{ij})_{1\leq i,j\leq n}$ is an H-matrix with positive diagonals, then

$$\max_{d\in[0,1]^n} \|(I - D + DM)^{-1}\|_\infty \leq \|\tilde{M}^{-1}\max(\Lambda, I)\|_\infty, \tag{2}$$

where $\tilde{M}$ is the comparison matrix of M, Λ is the diagonal part of M ($\Lambda := \text{diag}(m_{ii})$) and $\max(\Lambda, I) := \text{diag}(\max\{m_{11}, 1\}, \ldots, \max\{m_{nn}, 1\})$.

Computing the bound (2) requires $O(n^3)$ elementary operations because it involvers the inverse of an $n \times n$ matrix. As we shall see, the bound of the following theorem (which corresponds to Theorem 2.1 of [48]) can be obtained with lower computational cost than the bound (2). Moreover, it can be much smaller than (2) as we show later.

Theorem 2 *Let us assume that $M = (m_{ij})_{1\leq i,j\leq n}$ is an H-matrix with positive diagonal entries. Let $\bar{D} = \text{diag}(\bar{d}_1, \ldots, \bar{d}_n)$, $\bar{d}_i > 0$, for all $i = 1, \ldots, n$, be a diagonal matrix such that $M\bar{D}$ is strictly diagonally dominant by rows. For any $i = 1, \ldots, n$, let $\bar{\beta}_i := m_{ii}\bar{d}_i - \sum_{j\neq i} |m_{ij}|\bar{d}_j$. Then*

$$\max_{d\in[0,1]^n} \|(I - D + DM)^{-1}\|_\infty \leq \max\{\frac{\max_i\{\bar{d}_i\}}{\min_i\{\bar{\beta}_i\}}, \frac{\max_i\{\bar{d}_i\}}{\min_i\{\bar{d}_i\}}\}. \tag{3}$$

A first way to obtain the matrix $\bar{D}$ of Theorem 2 can be described as follows. We form $\tilde{M}$, the comparison matrix of M, and consider any positive vector $p > 0$ (for instance, $p = e := (1, 1, \ldots, 1)^T$). Since $\tilde{M}$ is an M-matrix, $\tilde{M}^{-1} \geq 0$ and then system $\tilde{M}\bar{d} = p$ has the nonnegative solution $\bar{d} = \tilde{M}^{-1}p$ and then we take $\bar{D} = \text{diag}(\bar{d}_1, \bar{d}_2, \ldots, \bar{d}_n)$. Observe that $\bar{\beta}_i$ of Theorem 2 coincides with the i-th component of p. Since this procedure involves the solution of a linear system associated to the $n \times n$ matrix $\tilde{M}$, it requires $O(n^3)$ elementary operations and so, the complexity is similar to that of the bound (2). However there is a second alternative to obtain our bound (3) of Theorem 2 with a complexity of lower order.

There are several recent iterative methods to compute the matrix $\bar{D}$ with at most $O(n^2)$ elementary operations per iteration (see [1, 70] and [80]), which lead to a computational cost much lower than computing $\tilde{M}^{-1}$ in (2), in particular in the case of sparse matrices. Then, we can observe that the vector $\bar{\beta} := (\bar{\beta}_1, \ldots, \bar{\beta}_n)^T$ satisfies $\bar{\beta} = \tilde{M}\bar{D}e$ and so its calculation requires $O(n^2)$ additional elementary operations. In conclusion, this alternative procedure has less computational cost than that of (2.4) of [20].

If the matrix M of Theorem 2 is strictly diagonally dominant by rows, then we can take $\bar{D} = I$ and so formula (3) becomes

$$\max_{d\in[0,1]^n} \|(I - D + DM)^{-1}\|_\infty \leq \max\{\frac{1}{\min_i\{\bar{\beta}_i\}}, 1\}.$$

4 HRA for Diagonally Dominant M-Matrices

A crucial tool to derive accurate algorithms for the computation of the singular values of a matrix is provided by the concept of rank revealing decomposition. Let us recall that a *rank revealing decomposition* of a matrix A is defined in [39] as a decomposition $A = XDY^T$, where X, Y are well conditioned and D is a diagonal matrix. In [39] Demmel et al. showed that the singular value decomposition can be computed accurately and efficiently for matrices possessing accurate rank revealing decompositions.

Let us also recall that an idea that has played a crucial role in some recent works on accurate computations has been the need to reparametrize matrices belonging to some special classes. In the class of M-matrices, the natural parameters that permit obtaining accurate and efficient algorithms are the off-diagonal entries and the row sums (or the column sums): see [2, 3] and [37], where the class of M-matrices row diagonally dominant was considered. Furthermore, the parameters can have a meaningful interpretation when the matrix arises in a "real" problem. In the field of digital electrical circuits, the column sums are given by the quotient between the conductance and capacitance of each node (see [2]).

An algorithm of [3] computed to HRA the LDU factorization of an $n \times n$ row diagonally dominant M-matrix A when the off-diagonal entries and the row sums are given. The trick was to modify Gaussian elimination to compute the off-diagonal entries and the row sums of each Schur complement without performing subtractions. On the other hand, let us recall that a symmetric pivoting leading to an LDU-decomposition of A is equivalent to the following factorization of A: $PAP^T = LDU$, where P is the permutation matrix associated to the pivoting strategy. Symmetric complete pivoting was used in [37] in order to obtain well conditioned L and U factors because U is row diagonally dominant and the off-diagonal entries of L have absolute value less than 1. This factorization is a special case of a rank revealing decomposition. To implement symmetric complete pivoting, the algorithm in [37] computes all the diagonal entries and all Schur complements

and this increases the cost in $\mathcal{O}(n^3)$ flops with respect to standard Gaussian elimination. In [85] another symmetric pivoting strategy (called diagonally dominant pivoting) was used, also with a subtraction-free implementation and a similar computational cost, but leading to both triangular matrices L and U column and row diagonally dominant, respectively. In [10], an accurate algorithm for the same LDU-decomposition of [85], but requiring $\mathcal{O}(n^2)$ elementary operations beyond the cost of Gaussian elimination, is presented. This method is also valid for diagonally dominant matrices satisfying certain sign patterns: with off-diagonal entries of the same sign or satisfying a chessboard pattern. The problem of computing an accurate LDU decomposition of diagonally dominant matrices has been solved by Ye in [95]. Finally, for a class of $n \times n$ nonsingular almost row diagonally dominant Z-matrices and given adequate parameters, an efficient method to compute its LDU decomposition with HRA is provided in [12]. It adds an additional cost of $\mathcal{O}(n^2)$ elementary operations over the computational cost of Gaussian elimination.

Now we recall the accurate algorithm of [10] mentioned above. We start with some notations and definitions. As usual, an LDU factorization of a square matrix $A = LDU$ means that L is a lower triangular matrix with unit diagonal (unit lower triangular), D is a diagonal matrix and U is an upper triangular matrix with unit diagonal (unit upper triangular). Given $k \in \{1, 2, \ldots, n\}$, let α, β be two increasing sequences of k positive integers less than or equal to n. Then we denote by $A[\alpha|\beta]$ the $k \times k$ submatrix of A containing rows numbered by α and columns numbered by β. For principal submatrices, we use the notation $A[\alpha] := A[\alpha|\alpha]$. Gaussian elimination with a given pivoting strategy, for nonsingular matrices $A = (a_{ij})_{1\leq i,j\leq n}$, consists of a succession of at most $n-1$ major steps resulting in a sequence of matrices as follows:

$$A = A^{(1)} \longrightarrow \tilde{A}^{(1)} \longrightarrow A^{(2)} \longrightarrow \tilde{A}^{(2)} \longrightarrow \cdots \longrightarrow A^{(n)} = \tilde{A}^{(n)} = DU, \tag{4}$$

where $A^{(t)} = (a_{ij}^{(t)})_{1\leq i,j\leq n}$ has zeros below its main diagonal in the first $t-1$ columns and DU is upper triangular with the pivots on its main diagonal. The matrix $\tilde{A}^{(t)} = (\tilde{a}_{ij}^{(t)})_{1\leq i,j\leq n}$ is obtained from the matrix $A^{(t)}$ by reordering the rows and/or columns $t, t+1, \ldots, n$ of $A^{(t)}$ according to the given pivoting strategy and satisfying $\tilde{a}_{tt}^{(t)} \neq 0$. To obtain $A^{(t+1)}$ from $\tilde{A}^{(t)}$ we produce zeros in column t below the *pivot element* $\tilde{a}_{tt}^{(t)}$ by subtracting multiples of row t from the rows beneath it. If the matrix A is singular, in this paper we allow the resulting matrices in (4) to have $\tilde{a}_{tt}^{(t)} = 0$, but (as we shall see later) in this case its corresponding column and row are null,

$$A^{(t)}[t, \ldots, n|t] = 0, \quad A^{(t)}[t|t, \ldots, n] = 0, \tag{5}$$

and we continue the elimination process with $A^{(t+1)}[t+1, \ldots n] = A^{(t)}[t+1, \ldots n]$.

We say that we carry out a *symmetric pivoting strategy* when we perform the same row and column exchanges, that is, $PAP^T = LDU$, where P is the associated permutation matrix. Let us present several symmetric pivoting strategies for row diagonally dominant matrices that either have been used in other papers or

will be used in this paper. Since row diagonal dominance is inherited by Schur complements in the Gaussian elimination, Gaussian elimination with symmetric pivoting preserves it, that is, all matrices $A^{(t)}$ of (4) are row diagonally dominant (and, in particular, DU and so U). Therefore, it is sufficient to describe the choice of the first pivot $\tilde{a}_{11} = a_{kk}$. On the one hand, the symmetric pivoting that selects the maximum entry on the diagonal for the pivot will be equivalent to complete pivoting and was used in [37]. It leads to U row diagonally dominant, and so well conditioned, and to L, which is usually well conditioned as well. On the other hand, since A is row diagonally dominant, we have

$$\sum_{i=1}^{n} |a_{ii}| \geq \sum_{i=1}^{n} \sum_{j=1, j\neq i}^{n} |a_{ij}|,$$

and there exists k such that column k is diagonally dominant, that is,

$$|a_{kk}| \geq \sum_{i=1, i\neq k}^{n} |a_{ik}|. \tag{6}$$

The symmetric pivoting strategy that chooses the first pivot $\tilde{a}_{11} = a_{kk}$ was called in [95] *column diagonal dominance pivoting*. In [85] the first pivot $\tilde{a}_{11} = a_{kk}$ was chosen so that it gives the most diagonal dominance in (6) (i.e., the largest difference between the absolute value of a diagonal entry and the sum of the absolute values of the off-diagonal entries of the corresponding row), and this strategy is a particular case of column diagonal dominance pivoting. In this paper we shall use a strategy that we call *weak column diagonal dominance pivoting*: it is a symmetric pivoting strategy that chooses the first pivot $\tilde{a}_{11} = a_{kk}$ satisfying (6), and without the necessity of being nonzero. If $\tilde{a}_{11} = 0$, then its row and column diagonal dominance implies that its row and column are null, and we continue the elimination process with $A^{(2)}[2, \ldots, n] = A[2, \ldots, n]$ (as we had announced for the t-th pivot in (5)). In order to uniquely determine this strategy, we can choose the first index k satisfying (6).

Column diagonal dominance pivoting and weak column diagonal dominance pivoting lead to U row diagonally dominant and to L column diagonally dominant. Then both triangular matrices are always well conditioned. In fact, since L is unit lower triangular column diagonally dominant, we know by Peña [85, Proposition 2.1, Remark 2.2] that

$$\kappa_\infty(L) = \|L\|_\infty \, \|L^{-1}\|_\infty \leq n^2 \quad \text{and} \quad \kappa_1(L) = \|L\|_1 \, \|L^{-1}\|_1 \leq 2n. \tag{7}$$

Analogously, with U unit upper triangular and row diagonally dominant, we have

$$\kappa_\infty(U) \leq 2n \quad \text{and} \quad \kappa_1(U) \leq n^2.$$

In contrast, symmetric complete pivoting leads to L that is usually well conditioned, but it is not necessarily column diagonally dominant. Finally, let $e := (1, \ldots, 1)^T$ and let

$$r := Ae \tag{8}$$

be the vector of row sums.

It is well known (cf. [3]) that we can carry out the Gaussian elimination of a diagonally dominant M-matrix with HRA because there is no subtraction involved throughout the process. Summarizing the process of [3, Algorithm 1], it starts with (8) and at each step of the Gaussian elimination it is only necessary to update the vector r. Diagonal entries of the matrix are not computed at each step (except the pivot) and so, the computational cost is of order $\mathcal{O}(n^2)$ beyond the cost of Gaussian elimination. We can also conclude that it is possible to compute the inverse of a nonsingular diagonally dominant M-matrix, A, with HRA, by the following procedure. We obtain the LDU factorization of A accurately. Then, it is well known (cf. [63, Sect. 13.2]) that we can compute the inverse of L and U without subtraction in the process. Thus, we can compute $A^{-1} = U^{-1}D^{-1}L^{-1}$ with HRA.

If A is a row and column diagonally dominant M-matrix, then no pivoting strategy is necessary to compute an accurate LDU factorization with L and U column and row diagonally dominant, respectively, because L also inherits through Gaussian elimination the column diagonal dominance from A. In fact, Gaussian elimination can be applied without row or column exchanges and so, for each $t = 1, \ldots, n-1$, $A^{(t)} = \tilde{A}^{(t)}$ (see (4)) and all matrices $A^{(t)}[t, \ldots, n]$ are row and column diagonally dominant. In conclusion, given the off-diagonal elements of a row and column diagonally dominant M-matrix $A = (a_{ij})_{1 \leq i,j \leq n}$ and the vector r of row sums (see (8)), we can calculate with HRA the LDU decomposition of A, where L is column diagonally dominant and U is row diagonally dominant. Moreover, this computation is subtraction-free and so can be performed with HRA.

Now we provide an accurate and efficient method for obtaining the LDU factorization (with L column diagonally dominant and U row diagonally dominant) of a row diagonally dominant M-matrix provided its off-diagonal entries and its row sums. Using A^T instead of A, we have also an accurate method for obtaining the LDU factorization of a column diagonally dominant M-matrix provided its off-diagonal entries and its column sums. The comparison with the computational cost of the methods presented in [37] and in [85, Sect. 4] can be seen in Remark 3.1 of [10]. This method produces a matrix U with a similar conditioning as in those papers because it is also row diagonally dominant and a matrix L that can be better conditioned than that of [37] (as the matrices of [10] show) and satisfies bounds (7) because it is column diagonally dominant, as commented previously. We start by presenting our algorithm (which corresponds to Algorithm 1 of [10]) to compute the LDU decomposition of a row diagonally dominant M-matrix.

In **output**, the algorithm produces the factorization $PAP^T = LDU$ (nontrivial entries of U are stored in $N = (n_{ij})_{1 \leq i < j \leq n}$).

The following result, which corresponds to Theorem 3.1 of [10] proves the interesting properties of the previous algorithm.

Algorithm 1

Input: $A = [a_{ij}]$ $(i \neq j)$ and $r = [r_i] \geq 0$
For $i = 1 : n$
  $p_i = \sum_{j=1, j\neq i}^{n} a_{ij}$
  $a_{ii} = r_i - p_i$
  $s_i = \sum_{j=1, j\neq i}^{n} a_{ji}$
  $h_i = a_{ii}$
End For
Choose an interchange permutation P_1 such that $A = P_1 A P_1^T$ satisfies $h_1 \geq -s_1$, where $h = P_1 h$, $s = P_1 s$
Initialize: $P = P_1$; $L = I$; $D = \text{diag}(d_i)_{i=1}^{n} = \text{diag}(h_1, 0 \ldots, 0)$; $r = P_1 r$
For $k = 1 : (n-1)$
  If $d_k = 0$
    For $i = (k+1) : n$
      $l_{ik} = 0$
      $n_{ki} = 0$
    End For
  Else
    For $i = (k+1) : n$
      $l_{ik} = a_{ik}/a_{kk}$
      $n_{ki} = a_{ki}/a_{kk}$
      $r_i = r_i - l_{ik} r_k$
      $h_i = h_i - n_{ki} h_k$
      $s_i = s_i - n_{ki} s_k$
      For $j = (k+1) : n$
        If $i \neq j$
          $a_{ij} = a_{ij} - l_{ik} a_{kj}$
        End If
      End For
    End For
  End If
  Choose interchange permutation P_2 such that $A = P_2 A P_2^T$ satisfies $h_{k+1} \geq -s_{k+1}$, where $h = P_2 h$, $s = P_2 s$
  $P = P_2 P$; $L = P_2 L P_2$; $r = P_2 r$
  $p_{k+1} = \sum_{j=k+2}^{n} a_{k+1,j}$
  $a_{k+1,k+1} = r_{k+1} - p_{k+1}$
  $d_{k+1} = a_{k+1,k+1}$
End For

Theorem 3 *Given the off-diagonal elements of a row diagonally dominant M-matrix $A = (a_{ij})_{1\leq i,j\leq n}$ and the vector r of row sums (see (8)), we can compute, by Algorithm 1, with HRA the LDU decomposition of PAP^T, where P the permutation matrix associated to a weak column diagonal dominance pivoting strategy applied when performing Gaussian elimination of A and such that L is column diagonally dominant and U is row diagonally dominant. Moreover, this computation is subtraction-free and can be performed with a computational cost that exceeds that of the Gaussian elimination by at most $(7n^2 - 11n + 6)/2$ additions, $n(n-1)$ multiplications, $n(n-1)/2$ quotients and $n(n-1)/2$ comparisons.*

The method of the previous theorem has less computational cost than those of [37] (symmetric complete pivoting) and [85, Sect. 4] because it requires $\mathcal{O}(n^2)$ (instead of $\mathcal{O}(n^3)$) elementary operations beyond the cost of Gaussian elimination. The reason for the lower computational cost comes from the fact that the method of Theorem 3 does not require, for each $t > 1$, the calculation of all diagonal elements $a_{jj}^{(t)}$ $(j \geq t)$ of the matrices $A^{(t)}[t, \ldots, n]$ in order to choose the pivot $\tilde{a}_{tt}^{(t)}$. However, in the case of symmetric complete pivoting, Ye suggested in [95, p. 2202], that we can use the diagonal entries as computed by standard Gaussian elimination to determine the pivot and permutation and then compute the pivot $a_{tt}^{(t)}$. With this procedure, symmetric complete pivoting also requires $\mathcal{O}(n^2)$ elementary operations beyond the cost of Gaussian elimination, although the possible pivots are not then computed accurately for the choice.

Theorem 3.1 can be applied to any row diagonally dominant matrix $A = (a_{ij})_{1 \leq i,j \leq n}$ satisfying

$$\operatorname{sign}(a_{ij}) \leq 0, \quad j \neq i, \qquad \operatorname{sign}(a_{ii}) \geq 0, \qquad i = 1, \ldots, n, \tag{9}$$

given its off-diagonal entries and its vector r of row sums (see (8)) and so the method of [39] allows us to calculate accurately all its singular values. Let us observe that we can also apply the method of Theorem 3.1 (and so the method of [39] allows us to calculate accurately all its singular values) to any row diagonally dominant matrix A satisfying any of the following sign patterns:

$$\operatorname{sign}(a_{ij}) = (-1)^{i+j+1}, \quad j \neq i, \qquad \operatorname{sign}(a_{ii}) \geq 0, \qquad i = 1, \ldots, n, \tag{10}$$

$$\operatorname{sign}(a_{ij}) \geq 0, \quad j \neq i, \qquad \operatorname{sign}(a_{ii}) \leq 0, \qquad i = 1, \ldots, n. \tag{11}$$

$$\operatorname{sign}(a_{ij}) = (-1)^{i+j}, \quad j \neq i, \qquad \operatorname{sign}(a_{ij}) \leq 0, \qquad i = 1, \ldots, n, \tag{12}$$

assuming that we know its off-diagonal entries and the vector of row sums of its comparison matrix $\mathcal{M}(A)$. In fact, let us define the diagonal $n \times n$ matrix $J = \operatorname{diag}(1, -1, \ldots, (-1)^{n-1})$ and observe that $J^{-1} = J$ and that, if A satisfies (10), then the matrix $J^{-1}AJ = JAJ = \mathcal{M}(A)$ satisfies (9), has the same singular values as A and we can calculate them with the method of [39] after obtaining the accurate LDU factorization of $\mathcal{M}(A)$ by the method of Theorem 3. Analogously, if A satisfies either (11) or (12), then we apply the procedure of Theorem 3 to $-A$ or to $J(-A)J$, respectively. Diagonally dominant matrices with arbitrary sign patterns were considered in [41] and [95], as commented above.

5 Totally Positive Matrices and Bidiagonal Factorizations

Let us recall that TP matrices are real, nonnegative matrices whose minors are all nonnegative. They are also called totally nonnegative matrices and they present many applications to several fields, including CAGD. In the next section we present

some applications to this last field. If all minors of a matrix are positive, then the matrix is called *strictly totally positive matrix* (STP matrix). TP and STP matrices have a long history and many applications (see [8, 43, 46, 53, 64, 91]) and have been studied mainly by researchers of those applications. In spite of their interesting algebraic properties, they have not yet received much attention from linear algebrists, including those working specifically on nonnegative matrices. One of the aims of the masterful survey [8] by T. Ando, which presents a very complete list of results on TP matrices until 1986, was to attract this attention, which lead to a very active research on the properties of these matrices.

The parametrization of TP matrices leading to HRA algorithms is provided by their bidiagonal factorizations, which are in turn closely related to an elimination procedure known as Neville elimination. In some papers by M. Gasca and G. Mühlbach ([54], for example) on the connection between interpolation formulas and elimination techniques it became clear that what they called *Neville elimination* had special interest for TP matrices. It is a procedure to make zeros in a column of a matrix by adding to each row an appropriate multiple of the precedent one and had been already used in some of the first papers on TP matrices (see [55]). However, in [55, 56] and [57] a better knowledge of the properties of Neville elimination was developed, which permitted to improve many previous results on those matrices. In this paper we shall use this elimination technique to get the factorization of a nonsingular TP matrix as a product of bidiagonal matrices. This provides a useful representation of such matrices which allows us to identify some important subclasses, as for example that of STP matrices (that is, TP matrices whose minors are all positive). Under some conditions on the zero pattern of the bidiagonal matrices that representation is unique.

A direct consequence of the well-known Cauchy-Binet identity for determinants is that the product of TP matrices is again a TP matrix. Consequently, one of the topics in the literature of TP matrices has been their decomposition as products of simpler TP matrices. In particular, in view of applications, the most interesting factorization seems to be in terms of bidiagonal nonnegative matrices which, obviously, are always TP matrices. Let us give a brief overview of some of the different approximations to this question.

Square TP matrices of order n form a multiplicative semigroup S_n, and the nonsingular matrices of S_n form a semigroup s_n of the group of all real nonsingular square matrices of order n (see [76]). In [71], Loewner used some notions from the theory of Lie groups which we briefly recall for the study of S_n and s_n. If $U(t)$ is a differentiable matrix function of the real parameter t in an interval $[0, t_0]$, representing for each t an element of s_n, and $U(0)$ is the identity matrix I (which belongs to s_n), then the matrix $(\frac{dU(t)}{dt})_{t=0}$ is called an *infinitesimal element* of s_n. The first task in [71] was to prove that the set σ_n of all infinitesimal elements of s_n consists of the Jacobi (i.e., tridiagonal) matrices with nonnegative off-diagonal elements.

As in Lie-group theory, it can be shown that if $\Omega(t)$ $(0 \leq t \leq t_0)$ is any one-parameter family of elements of σ_n which is piecewise continuous in t, the differential equation

$$\frac{dU(t)}{dt} = \Omega(t)U(t)$$

has a unique continuous solution $U(t)$ in s_n satisfying $U(0) = I$. In this case we say that $U(t_0)$ is generated by the infinitesimal elements $\Omega(t)$ $(0 \leq t \leq t_0)$. In general, a semigroup cannot be completely generated by its infinitesimal elements. However, Loewner proved in [71] that this is not the case for s_n. He used the following reformulation of a result due to Whitney [94].

Let E_{ij} $(1 \leq i, j \leq \ldots, n\})$ be the $n \times n$ matrix with all elements zero with the exception of a one at the place (i, j) and denote $F_{ij}(\omega) = I + \omega E_{ij}$. Then every nonsingular TP matrix U can be written as a product

$$U = U_1 U_2 \cdots U_{n-1} D V_1 V_2 \cdots V_{n-1}, \tag{13}$$

where, for $i = 1, 2, \ldots, n-1$,

$$U_i = F_{n,n-1}(\omega^i_{n,n-1}) F_{n-1,n-2}(\omega^i_{n-1,n-2}) \cdots F_{i+1,i}(\omega^i_{i+1,i}), \tag{14}$$

$$V_i = F_{n-i,n-i+1}(\omega^i_{n-i,n-i+1}) F_{n-i+1,n-i+2}(\omega^i_{n-i+1,n-i+2}) \cdots F_{n-1,n}(\omega^i_{n-1,n}), \tag{15}$$

with all the ω-s nonnegative, and D represents a diagonal matrix with positive diagonal elements.

Observe that the matrices U_i and V_i are products of bidiagonal elementary TP matrices but neither U_i nor V_i are bidiagonal and so the above factorization (13) of U uses $n(n-1)$ bidiagonal factors.

The conclusion of [71] is that, by using infinitesimal generators, any nonsingular TP matrix of order n can be generated from the identity by the solutions of the above differential equation.

In 1979 Frydman and Singer ([45], Theorem 1) showed that the class of transition matrices for the finite state time-inhomogeneous birth and death processes coincides with the class of nonsingular TP stochastic matrices. This result was based upon a factorization of nonsingular TP stochastic matrices S in terms of bidiagonal matrices ([45], Theorem 1') similar to (13) without the diagonal matrix D:

$$S = U_1 U_2 \cdots U_{n-1} V_1 V_2 \cdots V_{n-1}, \tag{16}$$

and with the elementary matrices F scaled to be stochastic. As in (13) the matrices U_i, V_i are not bidiagonal and (16) contains $n(n-1)$ bidiagonal factors. The fact that those transition matrices for birth and death processes are all had been pointed out in 1959 by Karlin and Mc Gregor (see [65] and [66]) with probabilistic arguments. All these results have been surveyed in 1986 by G. Goodman [59], who extended

them to compound matrices, that is matrices whose elements are the values of the minors of a certain order m of a given matrix A.

In [22], Remark 4.1, Cryer pointed out that a matrix A is TP if and only if it can be written in the form

$$A = \prod_{r=1}^{N} L_r \prod_{s=1}^{M} U_s$$

where each L_r (resp. U_s) is a TP bidiagonal lower (upper) triangular matrix. In that remark the author did not give any relation between N, M, and the order n of A.

On the other hand, factorizations of TP matrices as product of bidiagonal matrices are important in CAGD and, in particular, in corner cutting algorithms. In [60], Goodman and Micchelli showed that the existence of a corner cutting algorithm transforming a control polygon of a curve into another one with the same number of vertices was equivalent to the fact that both polygons were related by a nonsingular stochastic matrix. The key tool to obtain this result was again (see [60], Theorem 1) the characterization of a nonsingular TP stochastic matrix of order n as the product of $n-1$ bidiagonal lower triangular stochastic matrices by other $n-1$ matrices which are bidiagonal upper triangular, with an adequate zero pattern. Observe that in this case the factorization is formed by $2n-2$ bidiagonal matrices and compare with (16). What has happened is that the set of TP elementary matrices which appeared in (16), by replacing the factors U_i, V_i by their corresponding decompositions (14),(15), has been reordered to give rise to a short number of bidiagonal (in general nonelementary) matrices. In [78], Theorem 3.1, Micchelli and Pinkus obtained a factorization theorem for rectangular TP matrices as a product of bidiagonal matrices in order to extend the previous interpretation to general corner cutting algorithms. For more details related with this matter and for concrete factorizations associated with corner cutting algorithms to the Bezier polygon, see [16, 17, 61]. The use of Neville elimination (which will be recalled later) was crucial to prove the results on optimality obtained in these last two papers.

In conclusion, bidiagonal factorizations of TP matrices have played a crucial role in their study and applications since several decades ago. More recently, the bidiagonal factorization has been used to perform accurately many computations with these matrices (see [68]).

Let us define bidiagonal matrices $L^{(k)}$, $U^{(k)}$ by

$$L^{(k)} = \begin{pmatrix} 1 & & & & & & \\ 0 & 1 & & & & & \\ & \ddots & \ddots & & & & \\ & & 0 & 1 & & & \\ & & & l^{(k)}_{n-k} & 1 & & \\ & & & & \ddots & \ddots & \\ & & & & & l^{(k)}_{n-1} & 1 \end{pmatrix}, U^{(k)} = \begin{pmatrix} 1 & 0 & & & & & \\ & \ddots & \ddots & & & & \\ & & 1 & 0 & & & \\ & & & 1 & u^{(k)}_{n-k} & & \\ & & & & \ddots & \ddots & \\ & & & & & 1 & u^{(k)}_{n-1} \\ & & & & & & 1 \end{pmatrix},$$

where $k = 1, \ldots, n-1$.

In this section we shall consider matrices with bidiagonal decompositions of the form presented in the following definition.

Definition 4 Let A be a nonsingular $n \times n$ matrix. Suppose that we can write A as a product of bidiagonal matrices

$$A = L^{(1)} \cdots L^{(n-1)} D U^{(n-1)} \cdots U^{(1)}, \tag{17}$$

where $D = \operatorname{diag}(d_1, \ldots, d_n)$, and, for $k = 1, \ldots, n-1$, $L^{(k)}$ and $U^{(k)}$ are lower and upper bidiagonal matrices with unit diagonal respectively, with off-diagonal entries $l_i^{(k)} := (L^{(k)})_{i+1,i}$ and $u_i^{(k)} := (U^{(k)})_{i,i+1}$, $(i = 1, \ldots, n-1)$ satisfying

1. $d_i \neq 0$ for all i,
2. $l_i^{(k)} = u_i^{(k)} = 0$ for $i < n-k$,
3. $l_i^{(k)} = 0 \Rightarrow l_{i+s}^{(k-s)} = 0$ for $s = 1, \ldots, k-1$ and
   $u_i^{(k)} = 0 \Rightarrow u_{i+s}^{(k-s)} = 0$ for $s = 1, \ldots, k-1$.

Then we denote (17) by $\mathscr{BD}(A)$, a bidiagonal decomposition of A satisfying the conditions of this definition.

A matrix that can be decomposed in terms of bidiagonal matrices can also admit many other bidiagonal factorizations (cf. Chap. 6 of [91]). But the next result of [11] shows that a bidiagonal factorization as in Definition 2.1 is unique.

Theorem 5 *If a $\mathscr{BD}(A)$ exists for some matrix A, then it is unique.*

The following result provides the unique bidiagonal decomposition of a nonsingular TP matrix and it is a consequence of Theorem 4.2 of [57].

Theorem 6 *A nonsingular $n \times n$ matrix A is TP if and only if there exists a (unique) $\mathscr{BD}(A)$ such that*

1. $d_i > 0$ *for all i,*
2. $l_i^{(k)} \geq 0$, $u_i^{(k)} \geq 0$ *for* $1 \leq k \leq n-1$ *and* $n-k \leq i \leq n-1$.

It is well known that, if we have the $\mathscr{BD}(A)$ of a nonsingular tall matrix with HRA, then we can perform many computations of A with HRA, such as computing its inverse or computing its eigenvalues or its singular values (cf. [68]). Therefore, the entries of the bidiagonal factorization (17) are the adequate parameters for nonsingular TP matrices. There are several subclasses of nonsingular TP matrices for which this factorization can be obtained to HRA (and so, the computations mentioned previously, too). For instance, the mentioned algebraic computations can be performed with HRA for the following subclasses of TP matrices: Vandermonde positive matrices [38], Bernstein-Vandermonde matrices [74], Said-Ball-Vandermonde matrices [75], Pascal matrices [7], Jacobi-Stirling matrices [32], some rational collocation matrices [31], q-Bernstein-Vandermonde matrices [34] (these last three cases will be considered in Sect. 7 and Schoenmakers-Coffey matrices [36]. The bidiagonal factorization is obtained through an elimination procedure called Neville elimination and described below.

Now let us denote by ε the vector $\varepsilon = (\varepsilon_1, \ldots, \varepsilon_m)$ with $\varepsilon_j \in \{\pm 1\}$ for $j = 1, \ldots, m$, which will be called a *signature*.

Definition 7 Given a signature $\varepsilon = (\varepsilon_1, \ldots, \varepsilon_{n-1})$ and a nonsingular $n \times n$ matrix A, we say that A has a signed bidiagonal decomposition with signature ε if there exists a $\mathscr{BD}(A)$ (unique by Theorem 5) such that

1. $d_i > 0$ for all i,
2. $l_i^{(k)} \varepsilon_i \geq 0$, $u_i^{(k)} \varepsilon_i \geq 0$ for $1 \leq k \leq n-1$ and $n-k \leq i \leq n-1$.

Bidiagonal decompositions satisfying the properties of Definition 2.1 have been considered in [11] and [10] and it was proved that the class of matrices satisfying this definition contains nonsingular TP matrices and their inverses. Moreover, in [11] it has been shown that if we have the $\mathscr{BD}(A)$ of a matrix with HRA, then we can perform many computations of A with HRA, assuming that A belongs to the class of matrices satisfying the previous definition.

5.1 Neville Elimination and Bidiagonal Factorizations

We now present Neville elimination, which provides a constructive way of obtaining bidiagonal factorizations. Neville elimination is an alternative procedure to Gaussian elimination to eliminate nonzeros in a column of a matrix by adding to each row a multiple of the previous one (see [55]). If A is a square matrix of order n, $A = (a_{ij})_{1 \leq i,j \leq n}$ this elimination procedure consists of at most $n-1$ successive major steps, resulting in a sequence of matrices as follows:

$$A = A^{(1)} \to \widetilde{A}^{(1)} \to A^{(2)} \to \widetilde{A}^{(2)} \to \cdots \to A^{(n)} = \widetilde{A}^{(n)} = U, \tag{18}$$

where U is an upper triangular matrix.

The matrix $\widetilde{A}^{(t)}$ can be obtained by a reordering of the rows of the matrix $A^{(t)}$, moving the rows with a zero entry in column t to the bottom such that $\widetilde{a}_{it}^{(t)} = 0$ for $i \geq t$ implies that $\widetilde{a}_{ht}^{(t)} = 0$ for $\forall\, h \geq i$. Besides, $A^{(t+1)}$ is obtained from $\widetilde{A}^{(t)}$ eliminating nonzeros in the column t below the main diagonal by adding an adequate multiple of the ith row to the $(i+1)$th for $i = n-1, n-2, \ldots, t$ according to the following formula

$$a_{ij}^{(t+1)} = \begin{cases} \widetilde{a}_{ij}^{(t)}, & \text{if } 1 \leq i \leq j \leq t, \\ \widetilde{a}_{ij}^{(t)} - \dfrac{\widetilde{a}_{it}^{(t)}}{\widetilde{a}_{i-1,t}^{(t)}} \widetilde{a}_{i-1,j}^{(t)}, & \text{if } t+1 \leq i,j \leq n \text{ and } \widetilde{a}_{i-1,t}^{(t)} \neq 0, \\ \widetilde{a}_{ij}^{(t)}, & \text{if } t+1 \leq i \leq n \text{ and } \widetilde{a}_{i-1,t}^{(t)} = 0, \end{cases} \tag{19}$$

for all $t \in \{1, \ldots, n-1\}$.

The element

$$p_{ij} = \widetilde{a}_{ij}^{(j)}, \quad 1 \leq j \leq i \leq n, \tag{20}$$

is called the (i,j) *pivot* of Neville elimination of A. The Neville elimination can be performed without row exchanges if all the pivots are nonzero. The pivots p_{ii} are called *diagonal pivots*. Let us notice that when no rows exchanges are needed, then $A^{(t)} = \widetilde{A}^{(t)}$ for all t. If all the pivots p_{ij} are nonzero then, by Lemma 2.6 of [55], $p_{i1} = a_{i1}$ for $1 \leq i \leq n$ and

$$p_{ij} = \frac{\det A[i-j+1,\ldots,i|1,\ldots,j]}{\det A[i-j+1,\ldots,i-1|1,\ldots,j-1]} \tag{21}$$

for $1 \leq j \leq i \leq n$. The element

$$m_{ij} = \begin{cases} \dfrac{\widetilde{a}_{ij}^{(j)}}{\widetilde{a}_{i-1,j}^{(j)}} = \dfrac{p_{ij}}{p_{i-1,j}}, & \text{if } \widetilde{a}_{i-1,j}^{(j)} \neq 0, \\ 0, & \text{if } \widetilde{a}_{i-1,j}^{(j)} = 0, \end{cases} \tag{22}$$

is called the (i,j) *multiplier* of Neville elimination of A, where $1 \leq j < i \leq n$.

Neville elimination characterizes nonsingular TP matrices, as the following result shows. It follows from Theorem 4.2 and p. 116 of [57].

Theorem 8 *A matrix A is nonsingular TP if and only if the Neville elimination of A and A^T can be performed without row exchanges, all the multipliers of the Neville elimination of A and A^T are nonnegative and all the diagonal pivots of the Neville elimination of A are positive.*

Using the previous result as well as results of results of [56] and [57], we can describe bidiagonal decompositions of nonsingular TP matrices and their inverses in terms of the diagonal pivots and multipliers of their Neville elimination and the multipliers of the Neville elimination of their transposes.

Theorem 9 *Let A be a nonsingular TP matrix. Then A and A^{-1} admit factorizations in the form*

$$A^{-1} = G_1 G_2 \cdots G_{n-1} D^{-1} F_{n-1} \cdots F_1 \quad \text{and} \quad A = \overline{F}_{n-1} \cdots \overline{F}_1 D \overline{G}_1 \cdots \overline{G}_{n-1}, \tag{23}$$

respectively, where F_i and $\overline{F}_i$, $i \in \{1, \ldots, n-1\}$, are the lower triangular bidiagonal matrices given by

$$F_i = \begin{pmatrix} 1 & & & & & & & & \\ 0 & 1 & & & & & & & \\ & \ddots & \ddots & & & & & & \\ & & 0 & 1 & & & & & \\ & & & -m_{i+1,i} & 1 & & & & \\ & & & & -m_{i+2,i} & 1 & & & \\ & & & & & \ddots & \ddots & & \\ & & & & & & -m_{n,i} & 1 \end{pmatrix}$$

and

$$\overline{F}_i = \begin{pmatrix} 1 & & & & & & & & \\ 0 & 1 & & & & & & & \\ & \ddots & \ddots & & & & & & \\ & & 0 & 1 & & & & & \\ & & & m_{i+1,1} & 1 & & & & \\ & & & & m_{i+2,2} & 1 & & & \\ & & & & & \ddots & \ddots & & \\ & & & & & & m_{n,n-i} & 1 \end{pmatrix},$$

G_i and $\overline{G}_i$, $i \in \{1, \ldots, n-1\}$, are the upper triangular bidiagonal matrices whose transposes are given by

$$G_i^T = \begin{pmatrix} 1 & & & & & & & & \\ 0 & 1 & & & & & & & \\ & \ddots & \ddots & & & & & & \\ & & 0 & 1 & & & & & \\ & & & -\widetilde{m}_{i+1,i} & 1 & & & & \\ & & & & -\widetilde{m}_{i+2,i} & 1 & & & \\ & & & & & \ddots & \ddots & & \\ & & & & & & -\widetilde{m}_{n,i} & 1 \end{pmatrix}$$

and

$$\overline{G}_i^T = \begin{pmatrix} 1 & & & & & & & & \\ 0 & 1 & & & & & & & \\ & \ddots & \ddots & & & & & & \\ & & 0 & 1 & & & & & \\ & & & \widetilde{m}_{i+1,1} & 1 & & & & \\ & & & & \widetilde{m}_{i+2,2} & 1 & & & \\ & & & & & \ddots & \ddots & & \\ & & & & & & \widetilde{m}_{n,n-i} & 1 \end{pmatrix},$$

and D the diagonal matrix diag$(p_{11} \ldots, p_{nn})$*. The entries* m_{ij}*,* $\widetilde{m}_{ij}$ *are the multipliers of the Neville elimination of A and* A^T*, respectively, and the entries* p_{ii} *are the diagonal pivots of A.*

The results obtained until now assuring accurate computations with some subclasses of TP matrices have used the multipliers of Neville elimination as a natural parametrization of the matrices (cf. [7, 31, 32, 34, 36, 38, 67, 68, 74, 75]).

5.2 HRA for SBD Matrices

We finish this section by introducing the class of matrices analyzed in [11] and that will generalize the class of nonsingular TP matrices. For this class of matrices, HRA can be constructed.

Definition 10 Given a signature $\varepsilon = (\varepsilon_1, \ldots, \varepsilon_{n-1})$ and a nonsingular $n \times n$ matrix A, we say that A has a signed bidiagonal decomposition with signature ε if there exists a $\mathscr{BD}(A)$ (unique by Theorem 5) such that

1. $d_i > 0$ for all i,
2. $l_i^{(k)}\varepsilon_i \geq 0$, $u_i^{(k)}\varepsilon_i \geq 0$ for $1 \leq k \leq n-1$ and $n-k \leq i \leq n-1$.

We say that a matrix A has a *signed bidiagonal decomposition* (or, equivalently, that it is SBD) if it has a signed bidiagonal decomposition with some signature ε. Observe that, by Theorem 6, a nonsingular TP matrix has a signed bidiagonal decomposition with signature $\varepsilon = (1, \ldots, 1)$.

Let us now give the characterization of the matrices introduced in the last definition and the characterization of their inverses. First, we need some notations. Given a signature sequence $\varepsilon = (\varepsilon_1, \ldots, \varepsilon_{n-1})$, let us define a diagonal matrix $K = \text{diag}(k_1, \ldots, k_n)$ with k_i satisfying

$$k_i \in \{\pm 1\}\ \forall\, i = 1, \ldots, n, \qquad k_i k_{i+1} = \varepsilon_i\ \forall\, i = 1, \ldots, n-1. \tag{24}$$

Observe that, given ε, there are only two possible diagonal matrices K satisfying (24), depending on the two possibilities for $k_1 = \varepsilon_1$ or $k_1 = -\varepsilon_1$. Finally, given a matrix $A = (a_{ij})_{1\le i,j\le n}$, we define $|A| := (|a_{ij}|)_{1\le i,j\le n}$.

The following theorem, which correspond to Theorem 3.1 of [11], provides several characterizations of matrices with signed bidiagonal decomposition with a given signature.

Theorem 11 *Let $A = (a_{ij})_{1\le i,j\le n}$ be a nonsingular matrix and let $\varepsilon = (\varepsilon_1, \ldots, \varepsilon_{n-1})$ be a signature sequence. Then the following properties are equivalent:*

1. *A has a signed bidiagonal decomposition with signature ε.*
2. *$KAK = |A|$ is TP, where K is any diagonal matrix satisfying (24).*
3. *A^{-1} has a signed bidiagonal decomposition with signature $-\varepsilon = (-\varepsilon_1, \ldots, -\varepsilon_{n-1})$.*
4. *$|A|$ is TP and, for all $1 \le i, j \le n$, $\operatorname{sign}(a_{ij})$ is given by $\varepsilon_j \cdots \varepsilon_{i-1}$ if $i > j$, by 1 if $i = j$ and by $\varepsilon_i \cdots \varepsilon_{j-1}$ if $i < j$, respectively.*

Theorem 11 allows us to characterize the matrices with a signed bidiagonal decomposition in terms of TP matrices, as the following result shows. This characterization will allow us to use the accurate methods for TP matrices of [68] in order to assure accurate computations for matrices with signed bidiagonal decomposition.

Corollary 12 *Let A be a nonsingular $n \times n$ matrix. Then A has a signed bidiagonal decomposition if and only if there exists a diagonal matrix $K = \operatorname{diag}(k_1, \ldots, k_n)$ with $k_i \in \{\pm 1\}$ for all $i = 1, \ldots, n$ such that $KAK = |A|$ is a TP matrix.*

A particular case of the Theorem 11 corresponds to the case of nonsingular TP matrices, as shown in the following corollary, which corresponds to Corollary 3.3 of [11].

Corollary 13 *Let A be a nonsingular matrix. Then the following properties are equivalent:*

1. *A has a signed bidiagonal decomposition with signature $(1, \ldots, 1)$.*
2. *A is TP.*
3. *A^{-1} has a signed bidiagonal decomposition with signature $(-1, \ldots, -1)$.*

As recalled above, inverses of TP matrices are very important in applications. Corollary 13 has proved that they are matrices with signed bidiagonal decomposition. Observe that Theorem 11 also proves that the class of matrices with a signed bidiagonal decomposition is closed for the inversion of matrices.

If a matrix is opposite in sign to a matrix satisfying the properties of Definition 10, then it also satisfies all properties of Definition 10 except property 1, which is replaced by

$1'$. $d_i < 0$ for all i.

So, the bidiagonal decomposition of these matrices has the diagonal matrix with negative diagonal entries. Then, by Corollary 12, a nonsingular matrix A is opposite in sign to a matrix with a signed bidiagonal decomposition if and only if there exists a diagonal matrix $K = \text{diag}(k_1, \ldots, k_n)$ with $k_i \in \{\pm 1\}$ for all $i = 1, \ldots, n$ such that $-KAK = |A|$ is TP.

Let us see that if we assume it for the matrices of this subsection, then we can find algorithms with HRA to perform some computations with these matrices, such as the computation of their singular values, the computation of their eigenvalues, the computation of their inverses or solving certain linear systems $Ax = b$ (those with Kb with a chessboard pattern).

For all the mentioned computations, we can apply a similar procedure, which can be summarized as follows:

Step 1. From $\mathscr{BD}(A)$, we obtain $\mathscr{BD}(|A|)$, given by (6) of [11].
Step 2. We can apply known algorithms with HRA for TP matrices to $\mathscr{BD}(|A|)$. Recall that, by Corollary 12 and Remark 3.4 of [11], $|A|$ is TP if A belongs to the class of matrices characterized by these results.
Step 3. From the information obtained for $|A|$, we can get the corresponding result for A.

Let us now explain how to perform each of the previous steps.

As for Step 1, let us assume that we know the $\mathscr{BD}(A)$ (see (17)) with HRA for a given matrix A either satisfying the properties of Corollary 12 or satisfying them by $-A$. Then either $|A| = KAK$ or $|A| = -KAK$ for a diagonal matrix K satisfying (24) and so we can deduce from (6) of [11] that

$$|A| = |L^{(1)}| \cdots |L^{(n-1)}||D||U^{(n-1)}| \cdots |U^{(1)}| \tag{25}$$

is the $\mathscr{BD}(|A|)$. In fact, $\mathscr{BD}(A)$ and $\mathscr{BD}(KAK)$ are given by (17) and (5) of [11] respectively. The proof of Theorem 11 shows that, if $KAK = |A|$, then all factors of $\mathscr{BD}(KAK)$ (in (5) of [11]) are nonnegative and so $|L^{(j)}| = KL^{(j)}K$, $|U^{(j)}| = KU^{(j)}K$ for all $j = 1, \ldots, n-1$. Thus, (25) follows from (5) of [11]. If $|A| = -KAK$, then by the same argument all factors of $\mathscr{BD}(-KAK)$ are nonnegative and, taking into account that $\mathscr{BD}(KAK)$ and $\mathscr{BD}(-KAK)$ only differ in the fact that the diagonal factor KDK is changed by $-KDK$, we deduce that $-KDK = -D = |D|$ and we can also derive (25) from (5) of [11].

As for Step 2, we apply to (25) the corresponding algorithm for TP matrices with HRA. In particular, we consider the following accurate computations with TP matrices:

A. The eigenvalues of $|A|$ can be obtained by the method of Sect. 5 of [67] (`TNEigenvalues` in [69] is an implementation in MATLAB of this method).
B. The singular values of $|A|$ can be obtained by the method of Sect. 6 of [67] (`TNSingularValues` in [69] is an implementation in MATLAB of this method).
C. The inverse of $|A|$ can be obtained by the method of p.736 of [68].

D. Observe that $Ax = b$ is equivalent to solving $(KAK)(Kx) = Kb$, that is, $|A|(Kx) = Kb$. Then, $|A|^{-1}$ can be calculated accurately by the procedure of the previous case. By Theorem 3.3 of [8], $|A|^{-1}$ has a chessboard pattern of signs and so, since Kb has also a chessboard pattern of signs, $Kx = |A|^{-1}(Kb)$ can be calculated without subtractions and therefore with HRA as recalled in the introduction.

As for Step 3, we have the following cases corresponding to each of the cases of Step 2:

A. If A has a signed bidiagonal decomposition, then $|A| = KAK = K^{-1}AK$ and so they are similar matrices and have the same eigenvalues. If $-A$ has a signed bidiagonal decomposition, then the eigenvalues of A are opposite in sign to the eigenvalues of $|A|$.
B. The singular values of A and $|A|$ coincide because $|A| = \pm KAK$, that is, $|A|$ and A coincide up to unitary matrices.
C. If A has a signed bidiagonal decomposition, then $|A|^{-1} = (KAK)^{-1} = KA^{-1}K$ and so $A^{-1} = K|A|^{-1}K$. Analogously if $-A$ has a signed bidiagonal decomposition, then $|A|^{-1} = -KA^{-1}K$ and $A^{-1} = -K|A|^{-1}K$.
D. If we know Kx, then $x = K(Kx)$.

In addition, let us show that if we have the $\mathscr{BD}(A)$ (see (17)) with HRA, then we can also calculate the LDU decomposition of A with HRA, and even obtain the matrix A with HRA. In fact, by the uniqueness of the LDU decomposition of a matrix, it can be checked that the matrices

$$L = L^{(1)} \cdots L^{(n-1)}, \qquad U = U^{(n-1)} \cdots U^{(1)} \tag{26}$$

can be calculated without subtractions and so with HRA. Then we can also compute $A = LDU$ with HRA.

6 Applications of Totally Positive Matrices to CAGD

Let us start by recalling some basic facts on methods of CAGD using control polygons. We shall focus on the problem of finding bases with (optimal) shape preserving properties.

Given a sequence $U_0, \ldots, U_n$ of points in $\mathbf{R}^k$, we define a curve $\gamma(t) = \sum_{i=0}^{n} U_i u_i(t)$, $t \in I$. We shall denote by $U_0 \cdots U_n$ the polygonal arc with vertices $U_0, \ldots, U_n$. This is usually called the *control polygon* of γ and the points U_i, $i = 0, \ldots, n$, are called control points.

In CAGD it is usually required that the functions u_i, $i = 0, \ldots, n$, are nonnegative and $\sum_{i=0}^{n} u_i(t) = 1$ for all $t \in I$ (that is, the system $U = (u_0, \ldots, u_n)$ is *normalized*, or equivalently, the functions form a partition of unity). A normalized system of nonnegative functions is usually called a *blending* system. Now we shall recall some

properties which are convenient for design purposes, following mainly the notation of [44].

An important property for curve design is the *convex hull property*: for any control polygon, the curve lies always in the convex hull of the control polygon. Let us remark that the convex hull property holds if and only if U is a blending system. For blending systems, *affine invariance* also holds: computing a point of the curve $\gamma(t)$ and then applying an affine map to it gives the same result as applying first an affine map to the control polygon and then evaluating the mapped polygon at t.

These geometric properties correspond to some properties of the collocation matrices of the system of functions U. Given a system of functions $U = (u_0, \ldots, u_n)$ defined on $I \subseteq \mathbf{R}$, the *collocation matrix* of U at $t_0 < \cdots < t_m$ in I is given by

$$M\begin{pmatrix} u_0, \ldots, u_n \\ t_0, \ldots, t_m \end{pmatrix} := (u_j(t_i))_{i=0,\ldots,m;j=0,\ldots,n}.$$

Clearly, U is blending if and only if all its collocation matrices are stochastic (that is, nonnegative and such that the sum of each row is one).

In order to have a more precise guide of the curve, and to put together several pieces of curves, it is desirable for the designer to have a very precise control of what happens at the ends of the curve. This leads to the *endpoint interpolation property*: the first control point always coincides with the start point of the curve and the last control point always coincides with the final point of the curve.

In interactive design we also want that the shape of a parametrically defined polynomial curve mimics the shape of its control polygon; thus we can predict or manipulate the shape of the curve by suitably choosing or changing the control polygon. It is well-known (cf. [16]) that when the basis is normalized totally positive the curve imitates the shape of its control polygon, due to the variation diminishing properties of the TP matrices. A system of functions is *totally positive* if all its collocation matrices are TP. If U is normalized totally positive then the curve γ inherits many shape properties of the control polygon. For instance, any line intersects the curve no more often than it intersects the control polygon. In particular, a planar curve or polygon is convex if and only if it crosses any line of the plane no more than two times. So, if the control polygon is planar and convex then the curve generated is also planar and convex.

For normalized totally positive bases it also holds that the length, number of inflections and angular variation of the curve are bounded above by those of the control polygon (see [16]) for a more precise statement of these properties. Furthermore, totally positive bases also satisfy generalized convexity preserving properties [19] and can be characterized in terms of these properties. From Bemerkung II.4 of [92], we obtain that a blending system of functions satisfying the variation diminishing and the endpoint interpolation properties is necessarily totally positive. Therefore, if a system of functions is such that the curves generated by any control polygon satisfy the convex hull, variation diminishing and endpoint interpolation properties simultaneously, then it is a normalized totally positive system.

Now we shall deal with the problem of comparing the shape preserving properties of different blending systems. Our goal is to find a basis with optimal shape preserving properties: we want a normalized totally positive basis such that the control polygon of a curve with respect to it is closer in shape to the curve than the control polygon with respect to any other normalized totally positive basis of the space. Due again to the variation diminishing property of the TP matrices (see Sect. 2 of [16] and Sect. 2 of [18]), the precise formulation of this problem is finding a normalized totally positive basis $B = (b_0, \ldots, b_n)$ such that any other (normalized) totally positive basis U is of the form $U = (b_0, \ldots, b_n)K$, where K is a (stochastic) TP matrix. The concept of (normalized) *B-basis* (introduced in [17]) gives always an affirmative answer to the previous problem by Theorem 4.2 of [17]: a space with a normalized totally positive basis has always a unique normalized B-basis B, which is the basis with optimal shape preserving properties, as we shall explain later. For the space of polynomials of degree less than or equal to n on $[a, b]$, the Bernstein basis is the optimal (this was proved in [16]), and, for the corresponding space of polynomial splines, the B-spline basis is the optimal (see Theorem 4.6 of [17]).

In general, a space with a totally positive basis has always B-bases (see Remark 3.8 of [17]). By Proposition 3.11 of [17], a B-basis can be characterized as a basis $B = (b_0, \ldots, b_n)$ such that

$$\{\text{totally positive bases}\} = \{BK | K \text{ is a nonsingular TP matrix}\}. \tag{27}$$

Since nonsingular TP matrices can be characterized in terms of products of bidiagonal nonnegative matrices (see [56], where even the uniqueness of such factorizations is studied), let us observe that (27) allows us to construct all the totally positive bases of the space if we know a B-basis.

As for the problem of uniqueness of B-bases, we have the following result, which corresponds to Corollary 3.9 (iii) of [17]:

Proposition 14 *Let $(c_0, \ldots, c_n)$ be a B-basis of a space of functions $\mathscr{U}$. A basis of $\mathscr{U}$ is a B-basis if and only if it is of the form $(d_0c_0, \ldots, d_nc_n)$ with $d_i > 0$ for all $i = 0, \ldots, n$.*

On the other hand, by Theorem 4.2 (ii) of [17], we have that, if $U = (u_0, \ldots, u_n)$ is a normalized B-basis, then the set

$$\{\text{normalized totally positive bases}\}$$

coincides with the set

$$\{UH | H \text{ is a nonsingular stochastic TP matrix}\}.$$

Thus, since a nonsingular stochastic TP matrix can be characterized in terms of products of bidiagonal stochastic nonnegative matrices (cf. Theorem 1 of [60]), we can construct all the normalized totally positive bases of the space if we know the normalized B-basis.

As for the problem of uniqueness of normalized B-bases, we have an affirmative answer, which corresponds to Theorem 4.2 (i) of [17]:

Proposition 15 *If a space of functions has a normalized totally positive basis, then it has a (unique) normalized B-basis.*

Now let us justify the optimal shape preserving properties of the normalized B-basis. As shown in [18], a basis is a normalized B-basis if and only if it satisfies the least variation diminishing, the endpoint interpolation and the convex hull properties simultaneously.

Let us show now that the unique normalized B-basis of a given space has optimal properties in this geometric framework. Since the normalized B-basis is least variation diminishing, we may hope that the control polygon of a curve with respect to the normalized B-basis is the closest in shape to the curve than the control polygons with respect to any other reasonable basis for curve design. Let us collect now some shape properties established in [61] and [16], which applied to the normalized B-basis, confirm that it has optimal shape preserving properties.

Let γ be a curve generated by the control polygon $P_0 \cdots P_n$ with respect to a normalized totally positive basis. Let $B_0 \cdots B_n$ be the control polygon with respect to the normalized B-basis. Then the following properties hold:

(i) If $P_0 \cdots P_n$ is convex, then so are $B_0 \ldots B_n$ and the curve γ, and $B_0 \cdots B_n$ lies between $P_0 \cdots P_n$ and γ.
(ii) Length $\gamma \leq$ length $B_0 \cdots B_n \leq$ length $P_0 \cdots P_n$.
(iii) If $P_0 \cdots P_n$ turns through an angle $< \pi$, then $I(\gamma) \leq I(B_0 \cdots B_n) \leq I(p_0 \cdots P_n)$, where $I(\beta)$ denotes the number of inflexions of a curve β.
(iv) $\theta(\gamma) \leq \theta(B_0 \cdots B_n) \leq \theta(p_0 \cdots P_n)$, where $\theta(\beta)$ denotes the angular variation of a curve β.

In [18] there is a survey of other optimal properties which are satisfied by B-bases.

7 HRA for Some Subclasses of TP Matrices

We have recalled in Sect. 5 that if we have the $\mathscr{BD}(A)$ of a nonsingular TP matrix with HRA, then we can perform many computations of A with HRA, such as computing its inverse or computing its eigenvalues or its singular values.

In this section we shall illustrate some subclasses of TP matrices for which the $\mathscr{BD}(A)$ can be computed with HRA or for which we can perform computations with HRA. The first two subsections considers subclasses of matrices with applications to CAGD and the third subsection a subclass of matrices with applications to Combinatorics.

7.1 HRA with Rational Bernstein-Vandermonde Matrices

Let us start with the example of rational Bernstein-Vandermonde matrices considered in [31].

Given a basis $u = (u_0^n, \ldots, u_n^n)$ of nonnegative functions on $[a, b]$ and a sequence of strictly positive weights $(w_i)_{i=0}^n$, we can construct a rational basis $r = (r_0^n, \ldots, r_n^n)$ defined by

$$r_i^n(t) = \frac{w_i\, u_i^n(t)}{W(t)}, \quad t \in [a, b], \quad i \in \{0, 1, \ldots, n\}, \tag{28}$$

where $W(t) = \sum_{j=0}^n w_j\, u_j^n(t)$.

In CAGD the usual representation of a polynomial curve is the so called Bernstein-Bézier form, that is, these curves are expressed in terms of the Bernstein bases $(b_0^n, b_1^n, \ldots, b_n^n)$ defined by

$$b_i^n(t) = \binom{n}{i} t^i (1-t)^{n-i}, \quad i \in \{0, 1, \ldots, n\}, \quad t \in [0, 1].$$

The square collocation matrices $B := (b_j^n(t_i))_{0 \le i,j \le n}$ of the Bernstein basis of polynomials $(b_0^n, b_1^n, \ldots, b_n^n)$ at a sequence of parameters $0 < t_0 < t_1 < \ldots < t_n < 1$, are STP. From now on we will refer to these matrices as Bernstein-Vandermonde (BV) matrices. The corresponding square collocation matrices of the rational Bernstein basis at a sequence of parameters $0 < t_0 < t_1 < \ldots < t_n < 1$, given by $(r_j^n(t_i))_{0 \le i,j \le n}$, where functions r_i^n are given by (28) with $u_i^n = b_i^n$ for $i = 0, 1, \ldots, n$, will be called rational Bernstein-Vandermonde (RBV) matrices.

In Theorem 3.3 and Sect. 4 of [74] an algorithm for computing with HRA the bidiagonal decompositions of BV matrices and their inverses was presented. Here, we present the bidiagonal decompositions of RBV matrices and its inverses and an algorithm to compute them with HRA.

In Proposition 3.1 of [74], taking into account the relation between BV and Vandermonde matrices, the determinant of BV matrices was computed. Taking into account the result previously mentioned and the relation between RBV and BV matrices, the following result (which corresponds to Theorem 3.1 of [31]) computes the determinant of RBV matrices showing that they are STP.

Theorem 16 *Let $A = (w_j b_j^n(t_i)/W(t_i))_{0 \le i,j \le n}$ be a RBV matrix whose nodes satisfy $0 < t_0 < t_1 < \cdots < t_n < 1$. Then:*

1. $\det A = \binom{n}{0}\binom{n}{1} \cdots \binom{n}{n} \frac{w_0 w_1 \cdots w_n}{W(t_0) W(t_1) \cdots W(t_n)} \prod_{0 \le i < j \le n} (t_j - t_i) (> 0)$.
2. $$\det \begin{pmatrix} (1-t_0)^n & t_0(1-t_0)^{n-1} & \cdots & t_0^n \\ (1-t_1)^n & t_1(1-t_1)^{n-1} & \cdots & t_1^n \\ \vdots & \vdots & \ddots & \vdots \\ (1-t_n)^n & t_n(1-t_n)^{n-1} & \cdots & t_n^n \end{pmatrix} = \prod_{0 \le i < j \le n} (t_j - t_i).$$

3. A is STP.

We now describe bidiagonal decompositions of RBV matrices and their inverses in terms of the diagonal pivots and multipliers of their Neville elimination and the multipliers of the Neville elimination of their transposes (which corresponds to Theorem 3.2 of [31]).

Theorem 17 *Let $A = (w_j b_j^n(t_i)/W(t_i))_{0\leq i,j\leq n}$ be a RBV matrix whose nodes satisfy $0 < t_0 < t_1 < \cdots < t_n < 1$. Then A admits a factorization of the form*

$$A = \overline{F}_n \overline{F}_{n-1} \cdots \overline{F}_1 D \overline{G}_1 \cdots \overline{G}_n, \tag{29}$$

where $\overline{F}_i$ and $\overline{G}_i$, $i \in \{1, \ldots, n\}$, are the lower and upper triangular bidiagonal matrices given by

$$\overline{F}_i = \begin{pmatrix} 1 & & & & & & & \\ 0 & 1 & & & & & & \\ & \ddots & \ddots & & & & & \\ & & 0 & 1 & & & & \\ & & & m_{i0} & 1 & & & \\ & & & & m_{i+1,1} & 1 & & \\ & & & & & \ddots & \ddots & \\ & & & & & & m_{n,n-i} & 1 \end{pmatrix},$$

$$\overline{G}_i^T = \begin{pmatrix} 1 & & & & & & & \\ 0 & 1 & & & & & & \\ & \ddots & \ddots & & & & & \\ & & 0 & 1 & & & & \\ & & & \widetilde{m}_{i0} & 1 & & & \\ & & & & \widetilde{m}_{i+1,1} & 1 & & \\ & & & & & \ddots & \ddots & \\ & & & & & & \widetilde{m}_{n,n-i} & 1 \end{pmatrix},$$

and D the diagonal matrix diag$(p_{00}, p_{11} \ldots, p_{nn})$. The entries m_{ij}, $\widetilde{m}_{ij}$ and p_{ii} are given by

$$m_{ij} = \frac{W(t_{i-1})}{W(t_i)} \frac{(1-t_i)^{n-j}(1-t_{i-j-1})}{(1-t_{i-1})^{n-j+1}} \frac{\prod_{k=i-j}^{i-1}(t_i - t_k)}{\prod_{k=i-j-1}^{i-2}(t_{i-1} - t_k)}, \quad \text{for } 0 \leq j < i \leq n, \tag{30}$$

$$\widetilde{m}_{ij} = \frac{w_i}{w_{i-1}} \frac{n-i+1}{i} \frac{t_j}{1-t_j}, \quad \text{for } 0 \leq j < i \leq n, \tag{31}$$

$$p_{ii} = \frac{w_i}{W(t_i)} \binom{n}{i} \frac{(1-t_i)^{n-i}}{\prod_{k=0}^{i-1}(1-t_k)} \prod_{k=0}^{i-1}(t_i - t_k), \quad \text{for } 0 \leq i \leq n. \tag{32}$$

Given the bidiagonal decomposition $\mathscr{BD}(A)$ of a RBV matrix A presented in Theorem 17, we can obtain, from Theorem 9 and the results in [55, 56] and [57], the bidiagonal decomposition of A^{-1}, which will be denoted by $\mathscr{BD}(A^{-1})$, given by

$$A^{-1} = G_1 G_2 \cdots G_n D^{-1} F_n F_{n-1} \cdots F_1,$$

where F_i and G_i, $i \in \{1, \ldots, n\}$, are the lower and upper triangular bidiagonal matrices of the form of $\overline{F}_i$ and $\overline{G}_i$ (respectively), but replacing the off-diagonal entries

$$\{m_{i0}, m_{i+1,1}, \ldots, m_{n,n-i}\}$$

and

$$\{m_{i0}, m_{i+1,1}, \ldots, m_{n,n-i}\}$$

by the entries

$$\{-m_{i,i-1}, -m_{i+1,i-1}, \ldots, -m_{n,i-1}\}$$

and

$$\{-m_{i,i-1}, -m_{i+1,i-1}, \ldots, -m_{n,i-1}\}$$

(respectively).

Now we present a sequence of algorithms to compute the previous bidiagonal factorizations $\mathscr{BD}(A)$ and $\mathscr{BD}(A^{-1})$ with HRA. These factorizations will be used in the following section in order to solve certain linear systems $Ax = b$, and to compute A^{-1}, and the eigenvalues and singular values of A accurately.

In order to compute those bidiagonal decompositions we need to calculate accurately the multipliers m_{ij} and $\widetilde{m}_{ij}$, and the diagonal pivots p_{ii} in Theorem 17. Let us start with the computation of the multipliers m_{ij}. In the case of the computation of the multipliers m_{ij} it is necessary to evaluate the polynomial $W(t) = \sum_{i=0}^{n} w_i b_i^n(t)$ at the points $(t_i)_{i=0}^n$ accurately. It is well known that Horner algorithm consists of $\mathscr{O}(n)$ elementary operations to evaluate a polynomial of n degree (see [26]). But Horner algorithm uses the monomial representation instead of the Bernstein representation, so we rule it out. The usual form of evaluating a polynomial represented with the Bernstein basis is the de Casteljau algorithm (see [26]). But this algorithm evaluates a polynomial of degree n with $\mathscr{O}(n^2)$ elementary operations. An alternative algorithm for the evaluation of polynomials represented in the Bernstein basis is the VS algorithm, presented in [93] for the evaluation of multivariate polynomials and also adapted for the evaluation of univariate polynomials (see [26] for example). It

evaluates a polynomial of degree n

$$p(t) = \sum_{i=0}^{n} c_i v_i^n(t),$$

with $\mathscr{O}(n)$ elementary operations (see [26]), where the basis $(v_0^n, \ldots, v_n^n)$, given by $v_i^n(t) = t^i(1-t)^{n-i}$ $i = 0, 1, \ldots, n$, coincides with the Bernstein basis up to scaling. Algorithm 1 of [31] reminds the VS algorithm. This algorithm has a nested nature like Horner algorithm. As we shall justify later, for the particular case where the weights are positive, VS algorithm evaluates accurately polynomial $W(t)$ at points $t \in [0, 1]$. By the same reason, we can also consider to evaluate accurately the polynomial $W(t)$ in a straightforward way, that is, to evaluate the basis functions $b_i^n(t)$ and then compute the linear combination of the obtained values with the corresponding weights w_i, $i = 0, 1, \ldots, n$. Since $b_{i+1}^n(t) = \frac{t}{1-t}\frac{n-i}{i+1} b_i^n(t)$, a carefully programming of this approach uses $\mathscr{O}(n)$ elementary operations to compute $W(t_i)$ accurately. Taking into account that we need to evaluate polynomial $W(t) = \sum_{i=0}^{n} w_i b_i^n(t)$, $t \in [0, 1]$, with $w_i > 0$ for all $i \in \{0, 1, \ldots, n\}$, we have applied the VS algorithm above with $c_i = \binom{n}{i} w_i > 0$ for $i = 0, 1, \ldots, n$. In this case the algorithm evaluates $W(t)$ for all $t \in (0, 1)$ accurately because it does not use subtractions.

Now we state in Algorithm 2 of [31] the procedure for the computation of the multipliers m_{ij} accurately. In this algorithm, $W(t_i)$ in this algorithm will be computed by the VS algorithm, that is, by Algorithm 1 of [31]. The algorithm can be computed accurately because we only perform subtractions with the initial data.

Algorithm 3 of [31] provides the multipliers $\widetilde{m}_{ij}$ with HRA. Finally, the diagonal elements p_{ii} of D are computed by Algorithm 4 of [31] accurately.

Observe that the computation of Algorithms 2–4 of [31] is clearly of $\mathscr{O}(n^2)$ elementary operations.

7.2 HRA with q-Bernstein-Vandermonde Matrices

Let us recall that a basis of univariate functions $(u_0, \ldots, u_n)$ on an interval I is called STP (respectively, TP) if all its collocation matrices $(u_j(t_i))_{0 \le i,j \le n}$ at points $t_0 < t_1 < \ldots < t_n$ in I are STP (respectively, TP). A very important example of STP basis is the basis formed by the Bernstein polynomials on $(0, 1)$ (see [16]). In fact, these polynomials provide one of the most important bases in CAGD. Let us recall that the Bernstein polynomials of degree n are defined as

$$b_i^n(x) = \binom{n}{i} x^i (1-x)^{n-i}, \quad x \in [0, 1] \quad i = 0, \ldots, n.$$

The polynomials $\mathscr{B} = (b_0^n(x), \ldots, b_n^n)$ form the Bernstein basis of the space of polynomials of degree less than or equal to n, Π_n.

Now let us introduce the bases considered in this paper. Given $q > 0$ and any nonnegative integer r, we define a *q-integer* $[r]$ as

$$[r] = \begin{cases} (1-q^r)/(1-q), & q \neq 1, \\ r, & q = 1, \end{cases}$$

A *q-factorial* $[r]!$, where r is a nonnegative integer, is defined as

$$[r]! = \begin{cases} [r][r-1]\cdots[1], & r \geq 1, \\ 1, & r = 0, \end{cases}$$

We define the *q-binomial* coefficient as

$$\begin{bmatrix} n \\ r \end{bmatrix} = \frac{[n][n-1]\cdots[n-r+1]}{[r]!} = \frac{[n]!}{[r]![n-r]!}$$

for integers $n \geq r \geq 0$ and as zero otherwise. The q-Bernstein polynomials of degree n for $0 < q \leq 1$ were introduced in the mathematical literature by Phillips in [90]. They are defined as

$$b_{i,q}^n(x) = \begin{bmatrix} n \\ i \end{bmatrix} x^i \prod_{s=0}^{n-i-1} (1-q^s x), \quad x \in [0,1], \quad i = 0, 1, \ldots, n.$$

These polynomials $\mathscr{B}^q = (b_{0,q}^n, b_{1,q}^n, \ldots, b_{n,q}^n)$ also form a basis of Π_n. For $q = 1$ the q-Bernstein basis coincides with the Bernstein basis.

The following equivalence between TP and STP bases of the space of polynomials of degree not greater than n follows from Proposition 3.4 of [16] and will be applied in Sect. 3.1 to the q-Bernstein basis.

Proposition 18 *A basis of the space of polynomials of degree not greater than n is TP on an interval if and only if it is STP on its interior.*

The collocation matrices of the q-Bernstein basis $\mathscr{B}^q$ $(0 < q \leq 1)$ at a sequence of points $x_0 < x_1 < \cdots < x_n$, given by

$$B_q = \begin{pmatrix} b_{0,q}^n(x_0) & b_{1,q}^n(x_0) & \cdots & b_{n,q}^n(x_0) \\ b_{0,q}^n(x_1) & b_{1,q}^n(x_1) & \cdots & b_{n,q}^n(x_1) \\ \vdots & \vdots & \ddots & \vdots \\ b_{0,q}^n(x_n) & b_{1,q}^n(x_n) & \cdots & b_{n,q}^n(x_n) \end{pmatrix},$$

will be called *q-Bernstein-Vandermonde matrices*, qBV matrices from now on. In [62] the total positivity of this matrix for $q \in (0,1)$ was proved (for more details see Sect. 3.1).

In [68], assuming that the multipliers and diagonal pivots of a nonsingular TP matrix A and its transpose are known with HRA (that is, $\mathscr{BD}(A)$), Koev presented algorithms for computing:

- the eigenvalues of the matrix A,
- the singular values of the matrix A,
- the solution of linear systems of equations $Ax = b$ where b has a chessboard pattern of alternating signs,
- the inverse of the matrix A, i.e., A^{-1},

with HRA. In [69] we can get a software library called *TNTool*, which contains an implementation of the mentioned algorithms for Matlab and Octave. The name of the corresponding functions are `TNEigenvalues`, `TNSingularValues`, `TNSolve` and `TNJInverse`, respectively. These functions require as input argument the bidiagonal decomposition (23) of the matrix A, $\mathscr{BD}(A)$. `TNSolve` also requires a second argument, the vector of independent coefficients b of the linear system $Ax = b$ to be solved. Let us observe that `TNJInverse` provides the bidiagonal decomposition of $C = JA^{-1}J$, where $J = \operatorname{diag}(((-1)^i)_{i=0}^n)$, and so we compute $A^{-1} = JCJ$ to HRA. So, if we were able to obtain a bidiagonal decomposition of a qBV matrix to HRA, then we would solve the algebraic problems mentioned above to HRA with the help of the software library in [69].

Observe that, if one has the unique bidiagonal factorization $\mathscr{BD}(A)$ of a matrix A, then $\mathscr{BD}(A^T)$ is given by transposing the factorization (23), as remarked in [68]. In conclusion, if we were able to obtain a bidiagonal decomposition of a qBV matrix A to HRA, we would solve the algebraic problems mentioned above to HRA also for A^T. For instance, an application of solving linear systems with A^T comes from obtaining numerical differentiation and integration formulae with the q-Bernstein basis.

We shall now determine the bidiagonal decomposition (23) of a qBV matrix B_q in terms of the diagonal pivots and multipliers of its Neville elimination and those of their transposes. Previously, let us justify the strict total positivity of these matrices and let us show some relations involving these matrices, which will be very useful later.

Since the q-Bernstein polynomials of degree n, $\mathscr{B}^q$, form a basis of Π_n for any $q \in (0, 1]$, for any q, r with $0 < q, r \leq 1$ there exists a nonsingular matrix $M^{n,q,r}$ such that

$$(b_{0,q}^n(x), \ldots, b_{n,q}^n(x))^T = M^{n,q,r}(b_{0,r}^n(x), \ldots, b_{n,r}^n(x))^T. \tag{33}$$

From Theorems 4.1 and 4.3 of [62] the following result is deduced.

Theorem 19 *For $0 < q \leq r$ the matrix $M^{n,q,r}$ has all its entries positive. In addition, for $0 < q \leq r^{n-1}$ the matrix $M^{n,q,r}$ is TP.*

From the previous theorem, taking $r = 1$, the following results follow (which correspond to Theorem 3.1 and Corollary 3.2 of [34]).

Corollary 20 *For $q \in (0, 1)$ the matrix $M^{n,q,1}$ is TP.*

Since the Bernstein basis is TP we have the following result, which corresponds to Corollary 3.3 of [34]. Observe that the Bernstein basis corresponds to the case $q = 1$.

Corollary 21 *For any $q \in (0, 1]$ the basis formed by the q-Bernstein polynomials is TP on $[0, 1]$ and STP on $(0, 1)$.*

We now present the bidiagonal factorization of qBV matrices, which corresponds to Theorem 3.4 of [34].

Theorem 22 *Let $B_q = (b^n_{j,q}(x_i))_{0\le i,j\le n}$ be a qBV matrix whose nodes satisfy $0 < x_0 < x_1 < \cdots < x_n < 1$. Then B_q admits a factorization of the form (29), where the entries m_{ij}, $\widetilde{m}_{ij}$ and p_{ii} are given by*

$$m_{ij} = \frac{\prod_{s=0}^{n-(j+1)}(1-q^s x_i)\cdot(1-q^{n-j}x_{i-1-j})}{\prod_{s=0}^{n-j}(1-q^s x_{i-1})} \cdot \frac{\prod_{k=i-j}^{i-1}(x_i - x_k)}{\prod_{k=i-1-j}^{i-2}(x_{i-1} - x_k)}, \quad \text{for } 0 \le j < i \le n, \tag{34}$$

$$\widetilde{m}_{ij} = \frac{[n-i+1]}{[i]} \cdot \frac{x_j}{1-q^{n-i}x_j} \cdot \frac{\prod_{k=0}^{j-1}(1-q^{n-i+1}x_k)}{\prod_{k=0}^{j-1}(1-q^{n-i}x_k)}, \quad \text{for } 0 \le j < i \le n, \tag{35}$$

$$p_{ii} = \begin{bmatrix} n \\ i \end{bmatrix} \frac{\prod_{s=0}^{n-(i+1)}(1-q^s x_i)}{\prod_{k=0}^{i-1}(1-q^{n-i}x_k)} \cdot \prod_{k=0}^{i-1}(x_i - x_k), \quad \text{for } 0 \le i \le n. \tag{36}$$

As we can observe in the previous theorem, in order to compute the multipliers and the pivots, expressions of the form $1 - q^s x_k$ must be calculated. These expressions involve products before a true subtraction. Hence, a subtraction of inexact data must be calculated up to the case $q = 1$. Then we conclude that the multipliers and the pivots cannot be computed directly to HRA and $\mathscr{BD}(B_q)$, $q \in (0, 1)$, cannot be computed to HRA. So the algebraic problems detailed above cannot be solved to HRA using this bidiagonal factorization. Now we provide an alternative approach that allows us to solve algebraic problems with q-Bernstein-Vandermonde matrices to HRA.

Let us consider the collocation matrices of the Bernstein basis $\mathscr{B}$, that is, the q-Bernstein basis for $q = 1$ at a sequence of points $x_0 < x_1 < \cdots < x_n$, $B := B_1$. Since these matrices generalize Vandermonde matrices they are usually called *Bernstein-Vandermonde matrices* (see [74]).

In Theorem 4.2 of [62] a bidiagonal factorization of the matrix $T^{n,q,r}$ such that $M^{n,q,r} = D_q T^{n,q,r} D$ (see (33)) was given, where

$$D_q = diag\left(\begin{bmatrix} n \\ 0 \end{bmatrix}, \ldots, \begin{bmatrix} n \\ n \end{bmatrix}\right) \quad \text{and} \quad D^{-1} = diag\left(\binom{n}{0}, \ldots, \binom{n}{n}\right).$$

For our purposes we are interested in the case that $r = 1$.

Theorem 23 *For $n \ge 2$ and for any $q \in (0, 1)$ $T^{n,q,1} = G_1 G_2 \cdots G_{n-1}$, where G_i is the upper triangular bidiagonal matrix coinciding with the identity matrix up to the entries $(1, 2), (2, 3), \ldots, (i, i + 1)$ which are all $1 - q^{n-i}$.*

Observe that the zero pattern of the previous bidiagonal decomposition is different from that of $\mathscr{BD}(T^{n,q,1})$. If we want to use Koev's results and his software library *TNTool* in [69], then we need to overcome this difficulty, and we will use conversion matrices. Given a matrix $A = (a_{ij})_{1\leq i,j\leq n}$ we define the *conversion matrix* of A as $A^{\#} = (a^{\#}_{ij})_{1\leq i,j\leq n} := (a_{n+1-i,n+1-j})_{1\leq i,j\leq n}$, which can be written as $A^{\#} = PAP$ where P is obtained from the identity matrix by reversing the order of its rows. Transposing (33) and taking collocation matrices at a given sequence of points, we have that $B_q = B\,D\,(T^{n,q,1})^T D_q$. Then, using the conversion of this matrix, we obtain

$$(B_q)^{\#} = B^{\#} D^{\#} ((T^{n,q,1})^T)^{\#} (D_q)^{\#}. \tag{37}$$

The bidiagonal decomposition $\mathscr{BD}((T^{n,q,1})^{\#})$ of $(T^{n,q,1})^{\#}$ is given by the product of the conversion of the bidiagonal factors in the decomposition in Theorem 23 in the same order. Since $1 - q^k = (1-q)(1+q+q^2+\cdots+q^{k-1})$ the decomposition $\mathscr{BD}((T^{n,q,1})^{\#})$ can be computed to HRA. Then, the bidiagonal decomposition $\mathscr{BD}(((T^{n,q,1})^T)^{\#})$ is given by the transpose of the previous bidiagonal decomposition, and so it can be computed to HRA. We have implemented the bidiagonal decomposition of $(T^{n,q,1})^{\#}$ to HRA in Matlab function `TNBDCCM` (see Algorithm 1 of [34]).

Using the results and algorithm in Sect. 5.2 of [68], implemented in function `TNProduct` of the library *TNTool*, $\mathscr{BD}((B_q)^{\#})$ to HRA can be obtained if we are able to compute $\mathscr{BD}(B^{\#})$ to HRA. Before studying the possibility of computing the bidiagonal decomposition of $B^{\#}$, let us show that using *TNTool* with $\mathscr{BD}((B_q)^{\#})$ to HRA the algebraic problems with B_q can be solved to HRA.

Since $B_q = P(B_q)^{\#}P$, let us analyze the corresponding four algebraic problems for B_q in terms of $(B_q)^{\#}$:

- **Eigenvalues** and **singular values** of B_q. Taking into account that P is a unitary matrix and that $P^{-1} = P$, the eigenvalues and singular values of B_q coincide with those of $(B_q)^{\#}$. Then, using the functions `TNEigenvalues` and `TNSingularValues` of *TNTool* with the bidiagonal decomposition of $(B_q)^{\#}$ to HRA, we obtain the eigenvalues and the singular values of B_q with HRA.
- **Solution of a system of linear equations** $B_q x = b$ such that b has a chessboard sign pattern. Since $b = B_q x = P(B_q)^{\#}Px$ and $P^{-1} = P$, we deduce that the system is equivalent to $(B_q)^{\#}y = Pb$ where $y = Px$. So, taking into account that, if b has a chessboard sign pattern, then Pb also has a chessboard sign pattern, using `TNSolve` with the bidiagonal decomposition $\mathscr{BD}((B_q)^{\#})$ to HRA, we obtain $y = Px$ to HRA. So, performing Py, which is just reversing the order of the entries of y, we obtain x to HRA.
- **Inverse** of the matrix B_q, $(B_q)^{-1}$. Since $P^{-1} = P$ and $B_q = P(B_q)^{\#}P$, we have $(B_q)^{-1} = P((B_q)^{\#})^{-1}P$. So, using `TNJInverse` with $\mathscr{BD}((B_q)^{\#})$ to HRA we obtain $\mathscr{BD}(j((B_q)^{\#})^{-1}J)$ to HRA. Then, using `TNExpand` of [69] with this last bidiagonal decomposition and the usual matrix product we obtain $(B_q)^{-1}$ to HRA.

In conclusion, we need to obtain the bidiagonal decomposition of $B^{\#}$ to HRA. First we need the following auxiliary result, which corresponds to Proposition 3.6 of [34].

Proposition 24 *Let B be a BV matrix whose nodes satisfy $0 < x_0 < x_1 < \cdots < x_n < 1$. Then:*

1. $\det B^{\#} = \binom{n}{0}\binom{n}{1}\cdots\binom{n}{n}\prod_{0\le i<j\le n}(x_j - x_i)$.
2. $$\det \begin{pmatrix} (1-x_0)^n & x_0(1-x_0)^{n-1} & \cdots & x_0^n \\ (1-x_1)^n & x_1(1-x_1)^{n-1} & \cdots & x_1^n \\ \vdots & \vdots & \ddots & \vdots \\ (1-x_n)^n & x_n(1-x_n)^{n-1} & \cdots & x_n^n \end{pmatrix}^{\#} = \prod_{0\le i<j\le n}(x_j - x_i).$$
3. *$B^{\#}$ is STP.*

We now describe the bidiagonal decompositions of the conversion of BV matrices and their inverses in terms of the diagonal pivots and multipliers of their Neville elimination and the multipliers of the Neville elimination of their transposes, which corresponds to Theorem 3.7 of [34].

Theorem 25 *Let $B^{\#}$ denote the conversion of the BV matrix $B = (b_j^n(x_i))_{0\le i,j\le n}$ whose nodes satisfy $0 < x_0 < x_1 < \cdots < x_n < 1$. Then $B^{\#}$ admits a factorization of the form (29), where the entries m_{ij}, $\widetilde{m}_{ij}$ and p_{ii} are given by*

$$m_{ij} = \frac{x_{n-i}^{n-j}x_{n-(i-j)+1}}{x_{n-(i-1)}^{n-(j-1)}}\frac{\prod_{l=n-(i-1)}^{n-(i-j)}(x_l - x_{n-i})}{\prod_{l=n-(i-2)}^{n-(i-j)+1}(x_l - x_{n-(i-1)})}, \quad \text{for } 0 \le j < i \le n, \tag{38}$$

$$\widetilde{m}_{ij} = \frac{n-i+1}{i}\cdot\frac{1-x_{n-j}}{x_{n-j}}, \quad \text{for } 0 \le j < i \le n, \tag{39}$$

$$p_{ii} = \binom{n}{n-i}\frac{x_{n-i}^{n-i}}{\prod_{k=n-(i-1)}^{n} x_k}\prod_{l=n-(i-1)}^{n}(x_l - x_{n-i}), \quad \text{for } 0 \le i \le n. \tag{40}$$

In [74] Marco and Martínez provide an algorithm for computing the bidiagonal decomposition of a BV matrix with HRA. Analogously, taking into account the previous theorem we can design an algorithm to compute $\mathscr{B}\mathscr{D}(B^{\#})$ to HRA, which corresponds to Algorithm 2 of [34] with the accurate computation of the multipliers m_{ij}. The algorithm is accurate because we only perform subtractions with the initial data. Algorithm 3 of [34] provides the multipliers $\widetilde{m}_{ij}$ with HRA. Finally, the diagonal elements p_{ii} of D are computed accurately by Algorithm 4 of [34]. Observe that the computation of these algorithms is clearly of $\mathscr{O}(n^2)$ elementary operations. We have implemented these algorithms in the function `TNBDConvBV`.

As we have mentioned above, another interesting function of the library for our purpose is `TNProduct`. Given the bidiagonal decompositions to HRA, A and B, of two $m \times m$ TP matrices, F and G, respectively, `TNProduct(A,B)` computes with

$\mathcal{O}(m^3)$ elementary operations the bidiagonal decomposition of the TP matrix FG to HRA. So, from (37), using `TNProduct`, `TNBDConvBV` and `TNBDCCM`, we have implemented the computation of the bidiagonal decomposition of the conversion of the qBV matrix $B_q^{\#}$ to HRA in function `TNBDConvqBV`.

7.3 HRA with Jacobi-Stirling Matrices

In [42] (see also [79] and [9]) the Jacobi-Stirling numbers were introduced as the coefficients of the integral composite powers of the Jacobi differential operator

$$I_{\alpha,\beta}[y](t) = \frac{1}{(1-t)^{\alpha}(1+t)^{\beta}}\left(-(1-t)^{\alpha+1}(1+t)^{\beta+1}y'(t)\right)', \tag{41}$$

with α, β real numbers greater than -1. The Jacobi-Stirling numbers $JS_n^{(j)}(z)$ of the second kind depend only on the parameter $z = \alpha + \beta + 1 (> -1)$ and satisfy the following recurrence relation

$$JS_n^{(j)}(z) = JS_{n-1}^{(j-1)}(z) + j(j+z)\,JS_{n-1}^{(j)}(z) \quad (n, j \geq 1), \tag{42}$$

$$JS_n^{(0)}(z) = JS_0^{(j)}(z) = 0, \quad JS_0^{(0)}(z) = 1. \tag{43}$$

The Jacobi-Stirling numbers $Jc_n^{(j)}(z)$ of the first kind also depend only on the parameter $z = \alpha + \beta + 1$ and satisfy the following recurrence relation

$$Jc_n^{(j)}(z) = Jc_{n-1}^{(j-1)}(z) + (n-1)(n-1+z)\,Jc_{n-1}^{(j)}(z) \quad (n, j \geq 1), \tag{44}$$

$$Jc_n^{(0)}(z) = Jc_0^{(j)}(z) = 0, \quad Jc_0^{(0)}(z) = 1. \tag{45}$$

The Jacobi-Stirling numbers $Jc_n^{(j)}(z)$ of the first kind are a generalization of the Legendre-Stirling numbers: for $z = 1$ we obtain the Legendre-Stirling numbers.

In Theorem 4.2 of [79] the Jacobi-Stirling numbers of the second kind $JS_n^{(j)}$ were defined via the following expansion of the n-th composite power of $I_{\alpha,\beta}[y](t)$:

$$(1-t)^{\alpha}(1+t)^{\beta}I_{\alpha,\beta}[y](t) = \sum_{j=0}^{n}(-1)^j(jS_n^{(j)}(\alpha+\beta+1)(1-t)^{\alpha+j}(1+t)^{\beta+j}y^{(j)}(t))^{(k)},$$

where $I_{\alpha,\beta}[y](t)$ is the Jacobi differential operator (41).

The Jacobi-Stirling numbers $JS_n^{(j)}(z)$ of the second kind satisfy

$$x^n = \sum_{j=0}^{n} JS_n^{(j)}(z)\, \langle x \rangle_j (z) \quad (n \in \mathbf{N}),$$

where

$$\langle x \rangle_j (z) := \prod_{i=0}^{j-1} (x - i(i+z))$$

for all $j \geq 1$ and $\langle x \rangle_0 (z) := 1$. The (unsigned) Jacobi-Stirling numbers of the first kind $Jc_n^{(j)}(z)$ are defined via

$$\langle x \rangle_n (z) = \sum_{j=0}^{n} (-1)^{n+j} Jc_n^{(j)}(z) x^j \quad (n \in \mathbf{N}).$$

In this subsection we consider the infinite matrices $JS(z) = (jS_i^{(j)}(z))_{i,j\geq 0}$ and $Jc(z) = (jc_i^{(j)}(z))_{i,j\geq 0}$ and their corresponding truncated matrices given by $JS_n(z) = (jS_i^{(j)}(z))_{0\leq i,j\leq n-1}$ and $Jc_n(z) = (jc_i^{(j)}(z))_{0\leq i,j\leq n-1}$ formed by the Jacobi-Stirling numbers of the first and second kind, respectively. We provide a technique that guarantees that the computation of the singular values and the inverses of the matrices $JS_n(z)$ and $Jc_n(z)$ can be performed with HRA, recalling the results presented in [32].

In Theorem 5 of [79] the following result was proved.

Theorem 26 *The matrices JS and Jc are TP.*
Taking into account the definition of TP matrix, from the previous theorem the following result follows (see last line of Sect. 4 in [79]).

Corollary 27 *The matrices JS_n and Jc_n are TP.*

The following result is a consequence of Proposition 4 of [79] and states the bidiagonal decomposition of the matrices JS_n.

Theorem 28 *The Jacobi-Stirling matrix JS_n, $n \in \mathbf{N}$, admits a factorization of the form*

$$JS_n = \overline{G}_1^2 \cdots \overline{G}_{n-1}^2, \tag{46}$$

where $\overline{G}_i^2$, $i \in \{1, \ldots, n-1\}$, *are the* $n \times n$ *upper bidiagonal triangular matrices given by*

$$\overline{G}_i^2 = \begin{pmatrix} 1 & 0 & \cdots & \cdots & \cdots & \cdots & \cdots & 0 \\ & \ddots & \ddots & & & & & \vdots \\ & & 1 & 0 & & & & \vdots \\ & & & 1 & m_{i+1,1} & & & \vdots \\ & & & & 1 & m_{i+2,2} & & \vdots \\ & & & & & \ddots & \ddots & 0 \\ & & & & & & 1 & m_{n,n-i} \\ & & & & & & & 1 \end{pmatrix} \tag{47}$$

where $m_{ij} = j(z+j)$ *for* $1 \leq j < i \leq n$.

The following result is also a consequence of Proposition 4 of [79] and provides the bidiagonal decomposition of the matrices Jc_n.

Theorem 29 *The Jacobi-Stirling matrix* Jc_n, $n \in \mathbf{N}$, *admits a factorization of the form*

$$Jc_n = \overline{G}_1^1 \cdots \overline{G}_{n-1}^1, \tag{48}$$

where $\overline{G}_i^1$, $i \in \{1, \ldots, n-1\}$, *are the* $n \times n$ *upper bidiagonal triangular matrices given by*

$$\overline{G}_i^1 = \begin{pmatrix} 1 & 0 & \cdots & \cdots & \cdots & \cdots & \cdots & 0 \\ & \ddots & \ddots & & & & & \vdots \\ & & 1 & 0 & & & & \vdots \\ & & & 1 & \overline{m}_{i+1,1} & & & \vdots \\ & & & & 1 & \overline{m}_{i+2,2} & & \vdots \\ & & & & & \ddots & \ddots & 0 \\ & & & & & & 1 & \overline{m}_{n,n-i} \\ & & & & & & & 1 \end{pmatrix}, \tag{49}$$

where $\overline{m}_{ij} = (i-j)(z+i-j)$ *for all* $1 \leq j < i \leq n$.

Given a nonsingular TP matrix A, the $\mathscr{BD}(A)$ (see [68]) is a compact form for representing the bidiagonal decomposition of A (for instance, for matrices JS_n and Jc_n see (48) and (46)). Moreover, $\mathscr{BD}(A)$ is the starting point for the algorithms to HRA recalled in the next section. From now on, we will denote the compact form of the bidiagonal factorization (46) of JS_n by $\mathscr{BD}(jS_n)$ and that of (48) of Jc_n by $\mathscr{BD}(jc_n)$.

Finally, let us mention that the entries $m_{j,j+1}$'s and $\overline{m}_{j,j+1}$'s of (47) and (49) are the multipliers of the Neville elimination (see Sect. 5).

Given the bidiagonal decomposition of a Jacobi-Stirling matrix $\mathscr{BD}(jS_n)$ or $\mathscr{BD}(jc_n)$ presented in Theorems 28 and 29, respectively, we can obtain from applying Theorem 2.6 of [56] to the transposes of the Jacobi-Stirling matrices (see also the results of [57]) a bidiagonal decomposition of the inverses of the Jacobi-Stirling matrices given by

$$(jS_n)^{-1} = G_1^2 \cdots G_n^2 \quad \text{and} \quad (jc_n)^{-1} = G_1^1 \cdots G_n^1 \tag{50}$$

where G_i^2 and G_i^1, $i \in \{1, \ldots, n\}$, are the upper triangular bidiagonal matrices of the form of $\overline{G}_i^2$ and $\overline{G}_i^1$, respectively, but replacing the off-diagonal entries $\{m_{i+1,1}, \ldots, m_{n,n-i}\}$ and $\{\overline{m}_{i+1,1}, \ldots, \overline{m}_{n,n-i}\}$ by $\{-m_{i+1,i}, -m_{i+2,i}, \ldots, -m_{ni}\}$ and $\{-\overline{m}_{i+1,i}, -\overline{m}_{i+2,i}, \ldots, -\overline{m}_{ni}\}$, respectively, where the m_{ij}'s and $\overline{m}_{ij}$'s are given in Theorems 28 and 29, respectively.

Let us recall (see Sect. 2) that an algorithm can be performed with HRA if it does not include subtractions (except of the initial data), that is, if it only includes products, divisions, sums of numbers of the same sign and subtractions of the initial data (cf. [38]). Observe that Theorems 28 and 29 guarantee that we know $\mathscr{BD}(jS_n)$ and $\mathscr{BD}(jc_n)$ with HRA and that we know bidiagonal factorizations of $(jS_n)^{-1}$ and $(jc_n)^{-1}$ with HRA. In fact, these last bidiagonal factorizations can be used directly to obtain $(jS_n)^{-1}$ and $(jc_n)^{-1}$ because the matrix products can be performed without subtractions.

In [67] and [68] Koev presented algorithms for solving with HRA some usual linear algebra problems for an $n \times n$ TP matrix A, assuming that the entries of the bidiagonal decomposition $\mathscr{BD}(A)$ are known with HRA. In particular, we have the computation of the singular values and the eigenvalues of the matrix A. Koev has also developed a library available in [69], which contains an implementation of the algorithms devised in [67] and [68] for using them with Matlab and Octave. The name of the functions we are interested in are `TNSingularValues` and `TNSolve`, as we will comment in the next section. The computational cost of the corresponding algorithms is of $\mathscr{O}(n^3)$ elementary operations for an $n \times n$ matrix. The functions require as input argument the bidiagonal decompositions $\mathscr{BD}(jc_n)$ (given by (48) and (49)) or $\mathscr{BD}(jS_n)$ (given by (46) and (47)) of the Jacobi-Stirling matrix of first or second kind, respectively. In addition, `TNSolve` need a second argument, the vector of independent terms of the system to be solved.

Previously we have deduced how to compute the bidiagonal decompositions of the $n \times n$ Jacobi-Stirling matrices of both kinds with high accuracy and a total cost of $\mathscr{O}(n^2)$ elementary operations. We have implemented them in the functions named `TNBDJS1` and `TNBDJS2`, respectively, for Matlab and Octave, which take as input argument $z = \alpha + \beta + 1$ and the order n of the matrix. So, the accurate bidiagonal decompositions of Jacobi-Stirling matrices obtained by `TNBDJS1` and `TNBDJS2` can be used with the functions `TNSingularValues` and `TNSolve` of the TNTool library in [69] in order to obtain very accurate approximations of all the singular values of those matrices and of the solution of the corresponding linear systems.

Acknowledgements This work has been partially supported by the Spanish Research Grant MTM2015-65433-P (MINECO/FEDER) and by Gobierno de Aragón and Fondo Social Europeo.

References

1. Alanelli, M., Hadjidimos, A.: A new iterative criterion for H-matrices. SIAM J. Matrix Anal. Appl. **29**, 160–176 (2006/2007)
2. Alfa, A.S., Xue, J., Ye, Q.: Entrywise perturbation theory for diagonally dominant M-matrices with applications. Numer. Math. **90**, 401–414 (1999)
3. Alfa, A.S., Xue, J., Ye, Q.: Accurate computation of the smallest eigenvalue of a diagonally dominant M-matrix. Math. Comp. **71**, 217–236 (2001)
4. Alonso, P., Gasca, M., Peña, J.M.: Backward error analysis of Neville elimination. Appl. Numer. Math. **23**, 193–204 (1997)
5. Alonso, P., Delgado, J., Gallego, R., Peña, J.M.: Iterative refinement for Neville elimination. Int. J. Comput. Math. **86**, 341–353 (2009)
6. Alonso, P., Delgado, J., Gallego, R., Peña, J.M.: Growth factors of pivoting strategies associated to Neville elimination. J. Comput. Appl. Math. **235**, 1755–1762 (2011)
7. Alonso, P., Delgado, J., Gallego, R., Peña, J.M.: Conditioning and accurate computations with Pascal matrices. J. Comput. Appl. Math. **252**, 21–26 (2013)
8. Ando, T.: Totally positive matrices. Linear Algebra Appl. **90**, 165–219 (1987)
9. Andrews, G.G., Egge, E.S., Gawronnski, W., Littlejohn, L.L.: The Jacobi-Stirling numbers. J. Combin. Theory Ser. A **120** 288–303 (2013)
10. Barreras, A., Peña, J.M.: Accurate and efficient LDU decompositions of diagonally dominant M-matrices. Electron. J. Linear Algebra **24**, 153–167 (2012)
11. Barreras, A., Peña, J.M.: Accurate computations of matrices with bidiagonal decomposition using methods for totally positive matrices. Numer. Linear Algebra Appl. **20**, 413–424 (2013)
12. Barreras, A., Peña, J.M.: Accurate and efficient LDU decomposition of almost diagonally dominant Z–matrices. BIT Numer. Math. **54**, 343–356 (2014)
13. Barrio, R. Peña, J.M.: Numerical evaluation of the p-th derivative of Jacobi series. Appl. Numer. Math. **43**, 335–357 (2002)
14. Barrio, R., Peña, J.M.: Evaluation of the derivative of a polynomial in Bernstein form. Appl. Math. Comput. **167**, 125–142 (2005)
15. Berman, A., Plemmons, R.J.: Nonnegative Matrices in the Mathematical Sciences. Classics in Applied Mathematics, vol. 9, SIAM, Philadelphia (1994)
16. Carnicer, J.M., Peña, J.M.: Shape preserving representations and optimality of the Bernstein basis. Adv. Comput. Math. **1**, 173–196 (1993)
17. Carnicer, J.M., Peña, J.M.: Totally positive bases for shape preserving curve design and optimality of B-splines. Comput. Aided Geom. Des. **11**, 633–654 (1994)
18. Carnicer, J.M., Peña, J.M.: Total positivity and optimal bases. In: Gasca, M., Micchelli, C.A. (eds.) Total Positivity and Its Applications. Mathematics and Its Applications, vol. 359, pp. 133–155. Kluwer Academic Publishers, Dordrecht (1996)
19. Carnicer, J.M., García, M., Peña, J.M.: Generalized convexity preserving transformations. Comput. Aided Geom. Des. **13**, 179–197 (1995)
20. Chen, X., Xiang, S.: Computation of error bounds for P-matrix linear complementarity problems. Math. Program. Ser. A **106**, 513–525 (2006)
21. Cottle, R.W., Pang, J.S., Stone, R.E.: The Linear Complementarity Problems. Academic Press, Boston (1992)
22. Cryer, C.W.: Some properties of totally positive matrices. Linear Algebra Appl. **15**, 1–25 (1976)
23. Delgado, J., Peña, J.M.: A corner cutting algorithm for evaluating rational Bézier surfaces and the optimal stability of the basis. SIAM J. Sci. Comput. **29**, 1668–1682 (2007)

24. Delgado, J., Peña, J.M.: Progressive iterative approximation and bases with the fastest convergence rates. Comput. Aided Geom. Des. **24**, 10–18 (2007)
25. Delgado, J., Peña, J.M.: Error analysis of efficient evaluation algorithms for tensor product surfaces. J. Comput. Appl. Math. **219**, 156–169 (2008)
26. Delgado, J., Peña, J.M.: Running relative error for the evaluation of polynomials. SIAM J. Sci. Comput. **31** 3905–3921 (2009)
27. Delgado, J., Peña, J.M.: Optimal conditioning of Bernstein collocation matrices. SIAM J. Matrix Anal. Appl. **31**, 990–996 (2009)
28. Delgado, J., Peña, J.M.: Running error for the evaluation of rational Bézier surfaces. J. Comput. Appl. Math. **233**, 1685–1696 (2010)
29. Delgado, J., Peña, J.M.: Running error for the evaluation of rational Bézier surfaces through a robust algorithm. J. Comput. Appl. Math. **235**, 1781–1789 (2011)
30. Delgado, J., Peña, J.M.: On the evaluation of rational triangular Bézier surfaces and the optimal stability of the basis. Adv. Comput. Math. **38**, 701–721 (2013)
31. Delgado, J., Peña, J.M.: Accurate computations with collocation matrices of rational bases. Appl. Math. Comput. **219**, 4354–4364 (2013)
32. Delgado, J., Peña, J.M.: Fast and accurate algorithms for Jacobi-Stirling matrices. Appl. Math. Comput. **236**, 253–259 (2014)
33. Delgado, J., Peña, J.M.: Accurate evaluation of Bézier curves and surfaces and the Bernstein-Fourier algorithm. Appl. Math. Comput. **271**, 113–122 (2015)
34. Delgado, J., Peña, J.M.: Accurate computations with collocation matrices of q-Bernstein polynomials. SIAM J. Matrix Anal. Appl. **36**, 880–893 (2015)
35. Delgado, J., Peña, J.M.: Algorithm 960: POLYNOMIAL: an object-oriented Matlab library of fast and efficient algorithms for polynomials. Trans. Math. Softw. **42**, 19 (2016)
36. Delgado, J., Peña, G., Peña, J.M.: Accurate and fast computations with positive extended Schoenmakers-Coffey matrices. Numer. Linear Algebra Appl. **23**, 1023–1031 (2016)
37. Demmel, J., Koev, P.: Accurate SVDs of weakly diagonally dominant M-matrices. Numer. Math. **98**, 99–104 (2004)
38. Demmel, J., Koev, P.: The accurate and efficient solution of a totally positive generalized Vandermonde linear system. SIAM J. Matrix Anal. Appl. **27**, 142–152 (2005)
39. Demmel, J., Gu, M., Eisenstat, S., Slapnicar, I., Veselic, K., Drmac, Z.: Computing the singular value decomposition with high relative accuracy. Linear Algebra Appl. **299**, 21–80 (1999)
40. Demmel, J., Dumitriu, I., Holtz, O., Koev, P.: Accurate and efficient expression evaluation and linear algebra. Acta Numer. **17**, 87–145 (2008)
41. Dopico, F.M., Koev, P.: Perturbation theory for the LDU factorization and accurate computations for diagonally dominant matrices. Numer. Math. **119**, 337–371 (2001)
42. Everitt, W.N., Kwon, K.H., Littlejohn, L.L., Wellman R., Yoon, G.J.: Jacobi-Stirling numbers, Jacobi polynomials, and the left-definite analysis of the classical Jacobi differential expression. J. Comput. Appl. Math. **208**, 29–56 (2007)
43. Fallat, S.M., Johnson, C.R.: Totally Nonnegative Matrices. Princeton University Press, Princeton/Oxford (2011)
44. Farin, G.: Curves and Surfaces for Computer Aided Geometric Design. Academic Press, Boston (1988)
45. Frydman, H., Singer, B.: Total positivity and the embedding problem for Markov chains. Math. Proc. Camb. Philos. Soc. **85**, 339–344 (1979)
46. Gantmacher, F.P., Krein, M.G.: Oscillation Matrices and Kernels and Small Vibrations of Mechanical Systems (revised ed.). AMS Chelsea, Providence (2002)
47. García-Esnaola, M., Peña, J.M.: Error bounds for linear complementarity problems of B-matrices. Appl. Math. Lett. **22**, 1071–1075 (2009)
48. García-Esnaola, M., Peña, J.M.: A comparison of error bounds for linear complementarity problems of H-matrices. Linear Algebra Appl. **433**, 956–964 (2010)
49. García-Esnaola, M., Peña, J.M.: Error bounds for linear complementarity problems of B^S-matrices. Appl. Math. Lett. **25**, 1379–1383 (2012)

50. García-Esnaola, M., Peña, J.M.: Error bounds for the linear complementarity problem with a Σ-SDD matrix. Linear Algebra Appl. **438**, 1339–1346 (2013)
51. García-Esnaola, M., Peña, J.M.: Error bounds for linear complementarity problems of Nekrasov matrices. Numer. Algorithms **67**, 655–667 (2014)
52. García-Esnaola, M., Peña, J.M.: B-Nekrasov matrices and error bounds for linear complementarity problems. Numer. Algorithms **72**, 435–445 (2016)
53. Gasca, M., Micchelli, C.A. (eds.): Total Positivity and Its Applications. Mathematics and Its Applications, vol. 359. Kluwer Academic Publishers, Dordrecht (1996)
54. Gasca, M., Mühlbach, G.: Generalized Schur complements and a test for total positivity. Applied Numer. Math. **3**, 215–232 (1987)
55. Gasca, M., Peña, J.M.: Total positivity and Neville elimination. Linear Algebra Appl. **165**, 25–44 (1992)
56. Gasca, M., Peña, J.M.: A matricial description of Neville elimination with applications to total positivity. Linear Algebra Appl. **202**, 33–45 (1994)
57. Gasca, M., Peña, J.M.: On factorizations of totally positive matrices. In: Gasca, M., Micchelli, C.A. (eds). Total Positivity and Its Applications. Mathematics and Its Applications, vol. 359, pp. 109–130. Kluwer Academic Publishers, Dordrecht (1996)
58. Golub, G.H., Van Loan, C.F.: Matrix Computations, 3rd edn. The John Hopkins University Press, Baltimore (1996)
59. Goodman, G.: A probabilistic representation of totally positive matrices. Adv. Appl. Math. **7**, 236–252 (1986)
60. Goodman, T.N.T., Micchelli, C.A.: Corner cutting algorithms for the Bézier representation of free form curves. Linear Alg. Appl. **99**, 225–252 (1988)
61. Goodman, T.N.T., Said, H.B.: Shape preserving properties of the generalized Ball basis. Comput. Aided Geom. Des. **8**, 115–121 (1991)
62. Goodman, T.N.T., Oruç, H., Phillips, G.M.: Convexity and generalized Bernstein polynomials. Proc. Edinb. Math. Soc. **42**, 179–190 (1999)
63. Higham, N.J.: Accuracy and Stability of Numerical Algorithms, 2nd edn. SIAM, Philadelphia (2002)
64. Karlin, S.: Total Positivity, vol. 1. Stanford University Press, Stanford (1968)
65. Karlin, S., McGregor, J.L.: Coincidence probabilities of birth and death processes. Pac. J. Math. **9**, 1109–1140 (1959)
66. Karlin, S., McGregor, J.L.: Coincidence probabilities. Pac J. Math. **9**, 1141–1164 (1959)
67. Koev, P.: Accurate eigenvalues and SVDs of totally nonnegative matrices. SIAM J. Matrix Anal. Appl. **27**, 1–23 (2005)
68. Koev, P.: Accurate computations with totally nonnegative matrices. SIAM J. Matrix Anal. Appl. **29**, 731–751 (2007)
69. Koev, P.: http://math.mit.edu/~plamen/software/TNTool.html
70. Li, L.: On the iterative criterion for generalized diagonally dominant matrices. SIAM J. Matrix Anal. Appl. **24**, 17–24 (2002)
71. Loewner, C.: On totally positive matrices. Math. Z. **63**, 338–340 (1955)
72. Lyche, T., Peña, J.M.: Optimally stable multivariate bases. Adv. Comput. Math. **20**, 149–159 (2004)
73. Mainar, E., Peña, J.M.: Error analysis of corner cutting algorithms. Numer. Algorithms **22**, 41–52 (1999)
74. Marco, A., Martínez, J.J.: A fast and accurate algorithm for solving Bernstein-Vandermonde linear systems. Linear Algebra Appl. **422**, 616–628 (2007)
75. Marco, A., Martínez, J.J.: Accurate computations with Said-Ball-Vandermonde matrices. Linear Algebra Appl. **432**, 2894–2908 (2010)
76. Markham, T.L.: A semigroup of totally nonnegative matrices. Linear Algebra Appl. **3**, 157–164 (1970)
77. Mathias, R., Pang, J.S.: Error bounds for the linear complementarity problem with a *P*-matrix. Linear Algebra Appl. **132**, 123–136 (1990)

78. Micchelli, C.A., Pinkus A.: Descartes systems from corner cutting. Constr. Approx. **7**, 195–208 (1991)
79. Mongelli, P.: Total positivity properties of Jacobi-Stirling numbers. Adv. Appl. Math. **48**, 354–364 (2012)
80. Ojiro, K., Niki, H., Usui, M.: A new criterion for the H-matrix property. J. Comput. Appl. Math. **150**, 293–302 (2003)
81. Peña, J.M. (ed.): Shape Preserving Representations in Computer Aided Geometric Design. Nova Science Publishers, Commack (1999)
82. Peña, J.M.: B-splines and optimal stability. Math. Comput. **66**, 1555–1560 (1997)
83. Peña, J.M.: On the optimal stability of bases of univariate functions. Numer. Math. **91**, 305–318 (2002)
84. Peña, J.M.: A note on the optimal stability of bases of univariate functions. Numer. Math. **103**, 151–154 (2006)
85. Peña, J.M.: LDU decompositions with L and U well conditioned. Electron. Trans. Numer. Anal. **18**, 198–208 (2004)
86. Peña, J.M.: Eigenvalue bounds for some classes of P–matrices. Numer. Linear Algebra Appl. **16**, 871–882 (2009)
87. Peña, J.M.: Tests for the recognition of total positivity. SeMA J. **62**, 61–73 (2013)
88. Peña, J.M., Sauer, T.: On the multivariate Horner scheme. SIAM J. Numer. Anal. **37**, 1186–1197 (2000)
89. Peña, J.M., Sauer, T.: On the multivariate Horner scheme II: running error analysis. Computing **65**, 313–322 (2000)
90. Phillips, G.M.: Bernstein polynomials based on the q-integers. The heritage of P. L. Chebyshev: a Festschrift in honor of the 70th birthday of T. J. Rivlin. Ann. Numer. Math. **4**, 511–518 (1997)
91. Pinkus, A.: Totally Positive Matrices. Cambridge Tracts in Mathematics, vol. 181. Cambridge University Press, Cambridge (2010)
92. Schmeltz, G.: Variationsreduzierende Kurvendarstellungen und Krümmungskriterien für Bézierflächen, Thesis, Fachbereich Mathematik, Technische Hochschule Darmstadt (1992)
93. Schumaker, L.L., Volk, W.: Efficient evaluation of multivariate polynomials. Comput. Aided Geom. Des. **3**, 149–154 (1986)
94. Whitney, A.M.: A reduction theorem for totally positive matrices. J. d'Analyse Math. **2**, 88–92 (1952)
95. Ye, Q.: Computing singular values of diagonally dominant matrices to high relative accuracy. Math. Comp. **77**, 2195–2230 (2008).

Introduction to Communication Avoiding Algorithms for Direct Methods of Factorization in Linear Algebra

Laura Grigori

Abstract Modern, massively parallel computers play a fundamental role in a large and rapidly growing number of academic and industrial applications. However, extremely complex hardware architectures, which these computers feature, effectively prevent most of the existing algorithms to scale up to a large number of processors. Part of the reason behind this is the exponentially increasing divide between the time required to communicate a floating-point number between two processors and the time needed to perform a single floating point operation by one of the processors. Previous investigations have typically aimed at overlapping as much as possible communication with computation. While this is important, the improvement achieved by such an approach is not sufficient. The communication problem needs to be addressed also directly at the mathematical formulation and the algorithmic design level. This requires a shift in the way the numerical algorithms are devised, which now need to reduce, or even minimize when possible, the number of communication instances. Communication avoiding algorithms provide such a perspective on designing algorithms that minimize communication in numerical linear algebra. In this document we describe some of the novel numerical schemes employed by those communication avoiding algorithms, with a particular focus on direct methods of factorization.

1 Introduction

This document discusses one of the main challenges in high performance computing which is the increased communication cost, the fact that the time needed to communicate a floating-point number between two processors exceeds by huge factors the time required to perform a single floating point operation by one of the processors. Several works have shown that this gap has been increasing exponentially (see e.g. [36]) and it is predicted that it will continue to do so in the foreseeable future!

L. Grigori (✉)
Inria Paris, Alpines, and UPMC Univ Paris 06, CNRS UMR 7598, Laboratoire Jacques-Louis Lions, Paris, France
e-mail: Laura.Grigori@inria.fr

M. Mateos, P. Alonso (eds.), *Computational Mathematics, Numerical Analysis and Applications*, SEMA SIMAI Springer Series 13,
DOI 10.1007/978-3-319-49631-3_4

The memory wall problem, the disparity between the time required to transfer data between different levels of the memory hierarchy and the time required to perform floating point operations, was predicted already in 1995 by Wulf and McKee [71]. However, we are also facing now the inter-processor communication wall. Because of this, most of the algorithms are not able to scale to a large number of processors of these massively parallel machines. The slow rate of improvement of latency is mainly due to physical limitations, and it is not expected that the hardware research will find a solution to this problem soon. Hence the communication problem needs to be addressed also at the algorithmic and software level.

The communication gap is already seen and felt in the current, highly optimised applications, as illustrated by the top panel of Fig. 1, which displays the performance of a linear solver based on iterative methods used in the cosmic microwave background (CMB) data analysis application from astrophysics. This performance result is extracted from [40][1] where a more detailed description of the algorithms can be found. It shows the cost of a single iteration of conjugate gradient iterative solver preconditioned by a block diagonal preconditioner, together with the time spent on computation and communication. These runs were performed on a Cray XE6 system, each node of the system is composed of two twelve-cores AMD MagnyCours. It can be seen that the communication becomes quickly very costly, potentially dominating the runtime of the solver when more than 6000 cores are used (each MPI process uses 6 cores). The bottom part of Fig. 1 displays the performance estimated on a model of an exascale machine of a dense solver based on Gaussian elimination with partial pivoting (GEPP) factorization[2] (see also [41]). The plot displays the computation to communication ratio as a function of the problem size, vertical axis, and the number of used nodes, horizontal axis. The plot shows two regimes, at the top left corner this is the computation which dominates the run time, while at the bottom right this is the communication. The white region marks the regime where the problem is too large to fit in memory. We note that the communication-dominated regime is reached very fast, even for such a computationally intensive operation requiring $O(n^3)$ floating point operations (flops) as shown here (where the matrix to be factored is of size $n \times n$).

1.1 Communication Avoiding Algorithms

New communication avoiding algorithms have been introduced in the recent years that minimize communication and are as stable as classic algorithms. We describe in more details the communication complexity of direct methods of factorization in Sect. 2. Then in the following sections we describe communication avoiding algorithms for direct methods of factorization that attain the lower bounds on

[1] Courtesy of M. Szydlarski.

[2] Courtesy of M. Jacquelin.

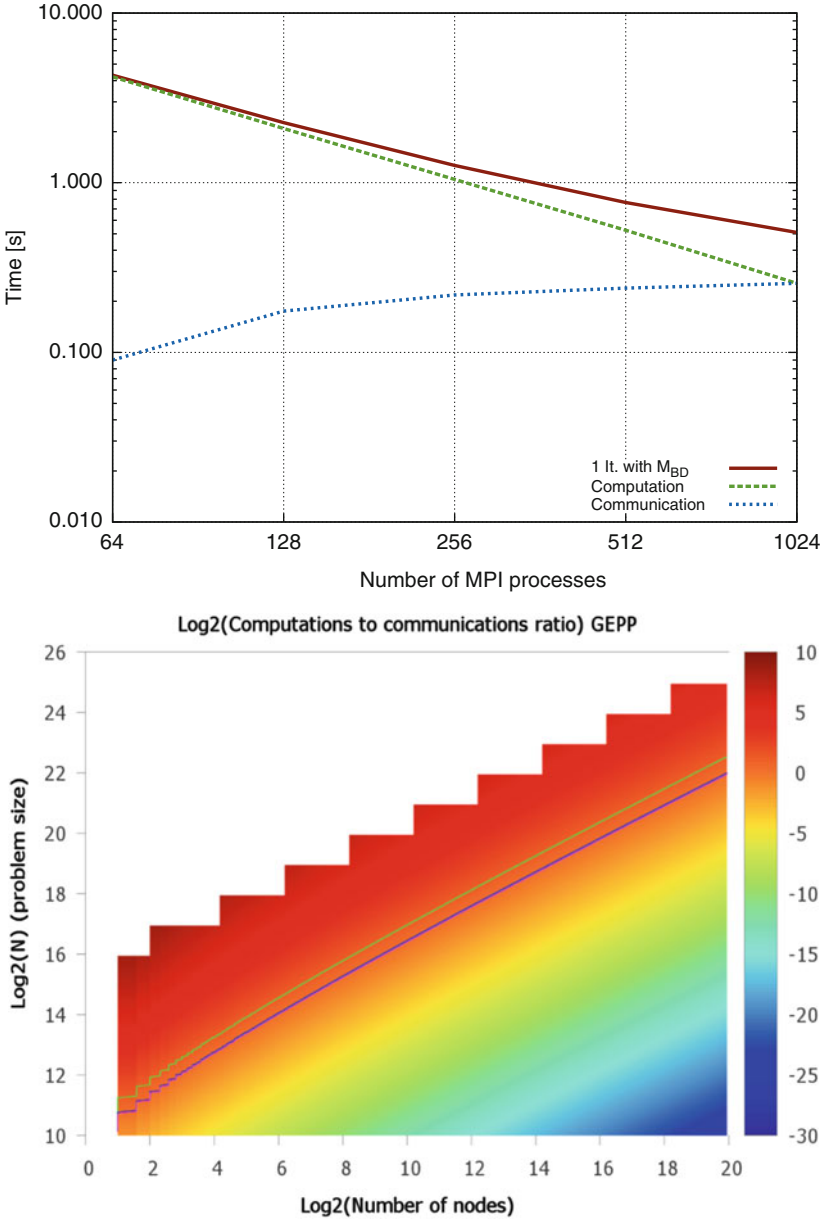


Fig. 1 Communication bottleneck of two algorithms, a dense linear solver based on LU factorization with partial pivoting (*bottom figure*) and a sparse iterative solver applied to the map-making problem in astrophysics (*top figure*, using data extracted from [40])

communication discussed in Sect. 2 (up to polylogarithmic factors). Section 3 describes CALU, a communication avoiding LU factorization. Section 4 presents CAQR, a communication avoiding QR factorization, while Sect. 5 discusses a communication avoiding rank revealing factorization. Section 5.3 focuses on computing a low rank matrix approximation. We assume for simplicity real matrices, but the algorithms can be generalized to complex matrices.

This document follows the presentation of the communication avoiding algorithms from the original papers that introduced them. The communication avoiding LU factorization is introduced in [37, 39], while the communication avoiding QR factorization is introduced in [21], and with many more details in the technical report [19]. A follow-up paper [9] allows to reconstruct Householder vectors such that it is sufficient to replace the panel factorization in a QR factorization to obtain a communication avoiding algorithm. A communication avoiding rank revealing QR factorization is presented in [23], while an LU factorization more stable than Gaussian elimination with partial pivoting is presented in [54]. When executed in parallel, these algorithms reduce significantly the number of messages exchanged with respect to classic algorithms as for example implemented in LAPACK [1] and ScaLAPACK [12]. They sometimes perform redundant computations, however these computations represent lower order terms with respect to the computational complexity of classic algorithms. In practice, when used with advanced scheduling techniques, the new algorithms lead to important speedups over existing algorithms [25, 26].

We cite here several other communication avoiding algorithms that were introduced in the recent years, but they are not described in this document. Communication avoiding algorithms for singular value decomposition (SVD) and eigenvalue problems are described in [2]. Bounds on communication for fast matrix multiplication are introduced in [3] and communication optimal algorithms for Strassen matrix multiplication are discussed in [6]. For sparse matrices, the communication complexity of the Cholesky factorization is studied in [38], while a communication optimal sparse matrix matrix multiplication algorithm is presented in [7].

Let's now give an example of classic algorithms that do not attain the lower bounds on communication. Several direct methods of factorization require some form of pivoting to preserve numerical stability, or reveal the rank of a matrix. The classic pivoting schemes, as partial pivoting in LU factorization or column pivoting in rank revealing QR, imply that the subsequent parallel algorithm cannot attain the lower bounds on communication. For a machine with one level of parallelism, the number of messages exchanged is on the order of $n \log P$, where n is the number of columns of the matrix and P is the number of processors used in the algorithm. For square matrices and when the memory per processor is of size $O(n^2/P)$, the lower bound on number of messages is $\Omega(\sqrt{P})$ (see Eq. (4) in Sect. 2). Hence in this case minimizing communication requires to invent novel pivoting schemes. There are examples in the literature of pivoting schemes, as for example proposed by Barron and Swinnerton-Dyer in their notable work [10], that minimize communication on sequential machines. At that time the matrices were of dimension 100×100 and the pivoting scheme was stable. But as shown in [39], this method can become

unstable for sizes of the matrices we encounter nowadays. The solution that we have developed for LU factorization is described in Sect. 3.

For iterative methods of factorization, most of the research around communication avoiding algorithms focuses on Krylov subspace methods. Those methods, as Conjugate Gradient (CG) [47], Generalized Minimal RESidual (GMRES) [62], Bi-Conjugate Gradient STABilized (Bi-CGSTAB) [69] are the most used iterative methods for solving linear systems of the form $Ax = b$, where A is very large and sparse. Starting from an initial solution x_0 and an initial residual r_0, a new approximate solution x_k is computed at iteration k by minimizing a measure of the error over $x_0 + K_k(A, r_0)$, where $K_k(A, r_0) = \text{span}[r_0, Ar_0, \ldots, A^{k-1}r_0,]$ is the Krylov subspace of dimension k. Every iteration requires computing the product of A (and in some cases of A^T) with a vector and several other operations as dot products related to the orthogonalization of the vectors of the basis. In the parallel case, the input matrix and the vectors are distributed over processors. Hence every iteration requires point to point communications for multiplying A with a sparse vector and collective communications for the dot products. On a large number of processors, the collective communications start dominating the overall cost of the iterative process. There are two main approaches used to reduce communication. The first approach relies on so called s-step methods [15, 17, 31, 50] that compute s vectors of the Krylov basis with no communication and then orthogonalize them against the previous vectors of the basis and against themselves. With this approach, the communication is performed every s iterations and this results in an overall reduction of the communication cost of the iterative method [15, 20, 50]. A second approach, described in [42], relies on enriching the subspace used in these methods that allows, at the cost of some extra computation, to reduce communication, while ensuring theoretically that the convergence is at least as fast as the convergence of the corresponding existing Krylov method. For this, first the problem is partitioned into P domains, and at each iteration of the iterative method, P dimensions are added to the search space instead of one dimension as in classic methods. Experimental results presented in [42] show that enlarged CG converges faster than CG on matrices arising from several different applications. This method is related to block Krylov subspace methods [58]. There are few preconditioners developed in this context, one of them is the communication avoiding incomplete LU preconditioner described in [42].

1.2 Different Previous Approaches for Reducing Communication

Most of the approaches investigated in the past to address this problem rely on changing the schedule of the computation such that the communication is overlapped as much as possible with the computation. However such an approach can lead to limited improvements. Ghosting is a different technique for reducing communication, in which a processor ghosts some data and performs redundantly

some computation, thus avoiding waiting to receive the results of this computation from other processors. But the dependency between computations in linear algebra operations prevents a straightforward application of ghosting. There are operations for which ghosting would require storing and performing on one processor an important fraction of the entire computation. Cache-oblivious algorithms represent a different approach introduced in 1999 for Fast Fourier Transforms [33], and then extended to graph algorithms, dynamic programming, etc. They were also applied to several operations in linear algebra (see e.g. [30, 45, 67]) as dense LU and QR factorizations. These cache-oblivious factorizations are computed through recursive calls of linear algebra operations on sub-blocks of the matrix to be factored. Since the sub-blocks become smaller and smaller, at some level of the recursion they fit in memory, and overall the amount of data transferred between different levels of the memory hierarchy is reduced. However there are cases in which the number of messages is not reduced and they perform asymptotically more floating-point operations.

1.3 Notations

We use Matlab like notation. We refer to the element of A at row i and column j as $A(i, j)$. The submatrix of A formed by rows from i to j and columns from k to s is referred to as $A(i : j, k : s)$. The matrix formed by concatenating two matrices A_1, A_2 stacked atop one another is referred to as $[A_1; A_2]$. The matrix formed by concatenating two matrices one next to another is referred to as $[A_1, A_2]$. The matrix formed by the absolute value of the elements of A is referred to as $|A|$. The identity matrix of size $n \times n$ is referred to as I_n.

To estimate the performance of an algorithm, we use the $\alpha - \beta - \gamma$ model. With this model, the time required for transferring one message of n words between two processors is estimated as $\beta \cdot n + \alpha$, where β is the interprocessor bandwidth cost per word and α is the interprocessor latency. Given the time required to compute one floating point operation (flop) γ, the time of a parallel algorithm is estimated as,

$$T = \gamma \cdot \#\,\text{flops} + \beta \cdot \#\,\text{words} + \alpha \cdot \#\,\text{messages}, \tag{1}$$

where #*flops* represents the computation, #*words* the volume of communication, and #*messages* the number of messages exchanged on the critical path of the parallel algorithm.

2 Lower Bounds on Communication for Dense Linear Algebra

In this section we review recent results obtained on the communication complexity of dense linear algebra operations. In the sequential case, these results consider a machine with two levels of memory, at the first level the memory has size M words, at the second level, the memory has infinite size but the access to the data is much slower. In the parallel case, they assume one level of parallelism, that is a parallel machine with P processing units connected through a fast network. One notable previous theoretical result on communication complexity is a result derived by Hong and Kung [51] providing lower bounds on the volume of communication of dense matrix multiplication for sequential machines. These bounds are extended to dense parallel matrix multiplication in [52] (with a different approach used for the proofs). It was shown in [19] that these bounds hold for LU and QR factorizations (under certain assumptions) and that they can be used to also identify lower bounds on the number of messages. General proofs that hold for almost all direct dense linear algebra operations are given in [4]. Consider a matrix of size $m \times n$ and a direct dense linear algebra algorithm as matrix multiplication, LU, QR, or rank revealing QR factorization, executed on a sequential machine with fast memory of size M words and slow memory of infinite size. The number of words and the number of messages transferred between slow and fast memory is bounded as,

$$\#\text{ words} \geq \Omega\left(\frac{mn^2}{M^{1/2}}\right), \quad \#\text{ messages} \geq \Omega\left(\frac{mn^2}{M^{3/2}}\right). \tag{2}$$

The bounds can be obtained by using the Loomis-Whitney inequality, as proven in [4, 52], which allows to bound the number of flops performed given an amount of data available in the memory of size M. Equation (2) can be used to derive bounds for a parallel program executed on P processors. For simplicity we consider in the following square dense matrices of size $n \times n$. Assuming that at least one processor does n^3/P floating point operations, and that the size of the memory of each processor M has a value between n^2/P and $n^2/P^{2/3}$, the lower bounds become

$$\#\text{ words} \geq \Omega\left(\frac{n^3}{P \cdot M^{1/2}}\right), \quad \#\text{ messages} \geq \Omega\left(\frac{n^3}{P \cdot M^{3/2}}\right). \tag{3}$$

When the memory of each processor is on the order of n^2/P, that is each processor has enough memory to store $/1/P$-th of the matrices involved in the linear algebra operation and there is no replication of the data, the lower bounds become

$$\#\text{ words} \geq \Omega\left(\frac{n^2}{\sqrt{P}}\right), \quad \#\text{ messages} \geq \Omega\left(\sqrt{P}\right). \tag{4}$$

Algorithms that attain these bounds are referred to as $2D$ algorithms. Cannon's matrix multiplication [14] is such an algorithm that attains the lower bounds on communication from (4). The lower bounds from (3) become smaller when the memory size is increased, and this until M is on the order of $n^2/P^{2/3}$. Indeed, even in the case of infinite memory M, it is shown in e.g. [5] that at least one processor must communicate $\Omega(n^2/P^{2/3})$ words of data. This leads to the following lower bounds,

$$\#\text{ words} \geq \Omega\left(\frac{n^2}{P^{2/3}}\right), \quad \#\text{ messages} \geq \Omega\left(1\right). \tag{5}$$

Algorithms that attain the lower bounds on communication in the case when M is larger than n^2/P are referred to as $3D$ algorithms. In these algorithms, the matrices are replicated over a 3D grid of processors.

These lower bounds on communication allow to identify that most of the existing algorithms as implemented in well-known numerical libraries as ScaLAPACK and LAPACK do not minimize communication. In the rest of this document we will discuss 2D algorithms that store only one copy of the matrices involved in the computation and use a memory on the order of n^2/P per processor (for square matrices). We discuss only the parallel case, however the algorithms can be adapted to minimize communication between two levels of memory in the sequential case.

3 Communication Avoiding LU Factorization

Given a matrix $A \in \mathbb{R}^{m\times n}$, $m \geq n$, the LU factorization decomposes the matrix A into the product $L \cdot U$, where L is a lower triangular matrix of size $m \times n$ with unit diagonal and U is an upper triangular matrix of size $n \times n$. This algorithm can be written as three nested loops, whose order can be interchanged. A so-called right-looking version of the algorithm is presented in Algorithm 1. To avoid division by small elements and preserve numerical stability, this algorithm uses partial pivoting. During the factorization, for each column k, the element of maximum magnitude in $A(k:n,k)$ is permuted to the diagonal position before the column is factored. Then, multiples of row k are added to all subsequent rows $k+1$ to m to annihilate all the nonzero elements below the diagonal. This algorithm requires $mn^2 - n^3/3$ flops. An in-place version can be easily obtained by overwriting the matrix A with the matrices L and U.

Typically, this factorization is implemented by using a block algorithm, in which the matrix is partitioned into blocks of columns of size b. In the remaining of this document, without loss of generality, we consider that n and m are multiples of b. At the first iteration, the matrix A is partitioned as

$$A = \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix}, \tag{6}$$

Algorithm 1 LU factorization with partial pivoting (GEPP)

Require: $A \in \mathbb{R}^{m \times n}$
Let $L \in \mathbb{R}^{m \times n}$ be initialized with identity matrix and $U \in \mathbb{R}^{n \times n}$ with zero matrix.
for $k = 1$ to $n - 1$ **do**
  Let $A(i, k)$ be the element of maximum magnitude in $A(k : m, k)$
  Permute row i and row k
  $U(k, k : n) = A(k, k : n)$
  $L(k + 1 : m, k) = A(k + 1 : m, k)/A(k, k)$
  for $i = k + 1 : m$ **do**
    for $j = k + 1 : n$ **do**
      $A(i, j) = A(i, j) - A(i, k)A(k, j)$
    end for
  end for
end for
$U(n, n) = A(n, n)$

where A_{11} is of size $b \times b$, A_{21} is of size $(m - b) \times b$, A_{12} is of size $b \times (n - b)$, and A_{22} is of size $(m - b) \times (n - b)$. With a right looking approach, the block algorithm computes the LU factorization with partial pivoting of the first block-column (panel), it determines the block U_{12}, and then it updates the trailing matrix A_{22}. The factorization obtained after the first iteration is

$$\Pi_1 A = \begin{bmatrix} L_{11} & \\ L_{21} & I_{m-b} \end{bmatrix} \cdot \begin{bmatrix} U_{11} & U_{12} \\ & A_{22}^1 \end{bmatrix}, \tag{7}$$

where $A_{22}^1 = A_{22} - L_{21}U_{12}$. The algorithm continues recursively on the trailing matrix A_{22}^1.

3.1 Parallel Block LU Factorization

We describe now briefly a parallel block LU algorithm by following its implementation in ScaLAPACK (`PDGETRF` routine). The input matrix is distributed over a $P_r \times P_c$ grid of processors using a bidimensional (2D) block cyclic layout with blocks of size $b \times b$. As an example, with a 2×2 grid of processors, the blocks of the matrix are distributed over processors as

$$\begin{bmatrix} P_0 & P_1 & P_0 & P_1 & \dots \\ P_2 & P_3 & P_2 & P_3 & \dots \\ \vdots & \vdots & \vdots & \vdots & \ddots \end{bmatrix}.$$

Algorithm 2 presents the main operations executed at each iteration of the block LU factorization. In terms of number of messages, it can be seen that, except for the panel factorization, all the other operations rely on collective communications

Algorithm 2 LU factorization with partial pivoting using a block algorithm

Require: $A \in \mathbb{R}^{m \times n}$ distributed over a 2D grid of processors $P = P_r \times P_c$

1: Let $L \in \mathbb{R}^{m \times n}$ be initialized with identity matrix and $U \in \mathbb{R}^{n \times n}$ with zero matrix

2: **for** k = 1 to n/b **do**

3: $k_b = (k-1) \cdot b + 1$, $k_e = k_b + b - 1$

4: Compute panel factorization using partial pivoting (processors in the same column of the process grid)

$$\Pi_k A(k_b : m, k_b : k_e) = L(k_b : m, k_b : k_e) \cdot U(k_b : k_e, k_b : k_e)$$

5: Broadcast pivot information along the rows of the process grid, pivot by applying the permutation matrix Π_k on the entire matrix (all processors)

$$A = \Pi_k A$$

6: Broadcast right $L(k_b : k_e, k_b : k_e)$, compute block row of U (processors in the same row of the process grid)

$$U(k_b : k_e, k_e + 1 : n) = L(k_b : k_e, k_b : k_e)^{-1} A(k_b : k_e, k_e + 1 : n)$$

7: Broadcast along rows of the process grid $L(k_e + 1 : m, k_b : k_e)$, broadcast along columns of the process grid $U(k_b : k_e, k_e + 1 : n)$, update trailing matrix (all processors)

$$A(k_e + 1 : m, k_e + 1 : n) = A(k_e + 1 : m, k_e + 1 : n) - L(k_e + 1 : m, k_b : k_e) \cdot U(k_b : k_e, k_e + 1 : n)$$

8: **end for**

which require exchanging $O(\log P_r)$ or $O(\log P_c)$ messages. Hence, the latency bottleneck lies in the panel factorization, where the LU factorization is performed column by column as in Algorithm 1. For each column, finding the element of maximum magnitude requires a reduce-type communication based on exchanging $\log P_r$ messages. In other words, partial pivoting requires performing a number of $O(n \log P_r)$ collective communications, which depends on n, the number of columns of the matrix. Since the lower bound on number of messages in Eq. (4) is $\Omega(\sqrt{P})$ for square matrices, LU factorization with partial pivoting as implemented in ScaLAPACK does not allow to minimize communication on a parallel machine. However we note that recently it has been shown that with a sophisticated data layout, it is possible to minimize data movement on a sequential machine for LU with partial pivoting [8].

3.2 Tournament Pivoting

Communication avoiding LU based on tournament pivoting was introduced in [37, 39] where a more detailed description can be found. We refer to this factorization as CALU. As in a classic LU factorization, the matrix is partitioned in blocks of b

columns. At the first iteration, consider the matrix A partitioned as

$$A = \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix},$$

where A_{11} is of size $b \times b$, A_{21} is of size $(m-b) \times b$, A_{12} is of size $b \times (n-b)$, and A_{22} is of size $(m-b) \times (n-b)$. With tournament pivoting, the panel factorization is performed as following. A preprocessing step plays a tournament to find at low communication cost b pivots that can be used to factor the entire panel. The selected b rows are permuted into the leading positions and they are used as pivots for the LU factorization of the entire panel (which is performed hence with no permutation). The preprocessing step is performed as a reduction operation where at each node of the reduction tree Gaussian elimination with partial pivoting (GEPP) is used to select b pivot rows. This strategy has the property that the communication for computing the panel factorization does not depend on the number of columns, but depends only on the number of processors. We refer to this procedure for computing the LU factorization of the panel as TSLU. The communication avoiding LU algorithm computes then the block U_{12}, updates the trailing matrix A_{22}, and a factorization as in Eq. (7) is obtained. It then continues recursively on the updated block A_{22}^1.

We explain now in more details tournament pivoting. Given P processors, the panel is partitioned into P block rows. We consider here the simple case $P = 4$, a binary reduction tree, and we suppose that m is a multiple of 4. The first panel is partitioned as $A(:, 1:b) = [A_{00}\ ;\ A_{10}\ ;\ A_{20}\ ;\ A_{30}]$. Each processor p has associated a block row A_{p0}. At the first step of the reduction, b rows are selected from each block A_{p0} by using GEPP. The selected rows correspond to the pivot rows used during the LU factorization. After this step we obtain 4 sets of b candidate rows. In the second step, the sets are combined two by two, we obtain two matrices of size $2b \times b$ each. From each matrix we select b rows by using again GEPP. In the last step of tournament pivoting, the two sets of candidate rows form a new matrix of size $2b \times b$ from which the final b rows are selected. This algorithm is illustrated in Fig. 2 from [39], where the function $f(A_{ij})$ computes the GEPP factorization of A_{ij} and returns the b pivot rows used by partial pivoting. The input matrix A_{ij} of dimension $2b \times b$ is formed by the two sets of candidate rows selected by the previous steps of tournament pivoting.

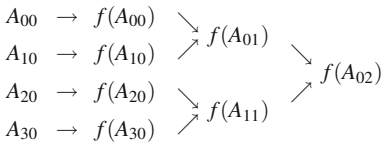


Fig. 2 TSLU with binary tree based tournament pivoting. This figure is from [39]. Copyright ©[2011] Society for Industrial and Applied Mathematics. Reprinted with permission. All rights reserved

Algorithm 3 Parallel TSLU factorization

Require: P processors, i is my processor's index, all reduction tree with height $L = \log P$

Require: $A \in \mathbb{R}^{m \times n}$, $m \gg n$, distributed in block row layout; $A_{i,0}$ is the block of rows belonging to my processor i

1: Compute $\Pi_{i,0} A_{i,0} = L_{i,0} U_{i,0}$ using GEPP
2: **for** each level k in the reduction tree from 1 to L **do**
3: $s = \lfloor i/2^k \rfloor, f = 2^k \lfloor i/2^k \rfloor, j = f + (i + 2^{k-1}) \mod 2^k$
4: $s_i = \lfloor i/2^{k-1} \rfloor, s_j = \lfloor j/2^{k-1} \rfloor$
5: Non-blocking send $(\Pi_{s_i,k-1} A_{s_i,k-1})(1:n, 1:n)$ to processor j
6: Non-blocking receive $(\Pi_{s_j,k-1} A_{s_j,k-1})(1:n, 1:n)$ from processor j
7: Wait until the previous send and receive have completed
8: Form the matrix $A_{s,k}$ of size $2n \times n$ as $A_{s,k} = \begin{bmatrix} (\Pi_{s_i,k-1} A_{s_i,k-1})(1:n, 1:n) \\ (\Pi_{s_j,k-1} A_{s_j,k-1})(1:n, 1:n) \end{bmatrix}$
9: Compute $\Pi_{s,k} A_{s,k} = L_{s,k} U_{s,k}$ using GEPP
10: **end for**
11: Determine the final permutation Π, such that $(\Pi A)(1:n, :)$ are the k selected rows at the end of tournament
12: All P processors compute the Gaussian elimination with no pivoting of their blocks, $\Pi A = LU$

Ensure: $U_{0,L}$ is the U factor obtained at step 12 for all processors i.

Algorithm 3 presents a pseudo-code for the parallel implementation of TSLU on P processors. It follows the presentation of TSLU in [39], where a more detailed description can be found. For simplicity, we consider that P is a power of 2. We consider here an all reduction tree based on a butterfly scheme, whose height is $L = \log P$. The matrix A is distributed block row-wise over processors. The levels of the tree are numbered from 0 to L, where the first level 0 corresponds to the phase with no communication and each leaf node represents a processor. At the first level $k = 1$, each node s has associated two processors i and $i - 1$, where i is an odd number. The two processors exchange their set of candidate rows. Then each processor forms a matrix with the two sets of candidate rows and selects a new set of candidate rows using GEPP. In general, at a given level k, processor i participates to the computation associated with node numbered $s = \lfloor i/2^k \rfloor$. The first processor associated with this node is $f = 2^k \lfloor i/2^k \rfloor$ and the processor exchanging information with this processor is numbered $f + 2^{k-1}$. Processor i exchanges information with processor $j = f + (i + 2^{k-1}) \mod 2^k$. They exchange the candidate rows that were selected at the previous level $k-1$ in the reduction tree at the children nodes s_i and s_j.

TSLU requires exchanging $\log P$ messages among processors. This allows the overall CALU algorithm to attain the lower bounds on communication in terms of both number of messages and volume of communication. When the LU factorization of a matrix of size $n \times n$ is computed by using CALU on a grid of $P = P_r \times P_c$ processors, as shown in [39] where a more detailed description can be found,

the parallel performance of CALU in terms of number of messages, volume of communication, and flops, is

$$\begin{aligned} & T_{CALU}(m,n,P) \\ & \approx \gamma \cdot \left(\frac{1}{P}\left(mn^2 - \frac{n^3}{3}\right) + \frac{1}{P_r}\left(2mn - n^2\right)b + \frac{n^2 b}{2P_c} + \frac{nb^2}{3}(5\log_2 P_r - 1)\right) \\ & \quad + \beta \cdot \left(\left(nb + \frac{3n^2}{2P_c}\right)\log_2 P_r + \frac{1}{P_r}\left(mn - \frac{n^2}{2}\right)\log_2 P_c \right) \\ & \quad + \alpha \cdot \left(\frac{3n}{b}\log_2 P_r + \frac{3n}{b}\log_2 P_c \right). \end{aligned} \tag{12}$$

To attain the lower bounds on communication, an optimal layout can be chosen with $P_r = P_c = \sqrt{P}$ and $b = \log^{-2}\left(\sqrt{P}\right) \cdot \frac{n}{\sqrt{P}}$. The blocking parameter b is chosen such that the number of messages attains the lower bound on communication from Eq. (4), while the number of flops increases only by a lower order term. With this layout, the performance of CALU becomes,

$$\begin{aligned} T_{CALU}(m,n,P = \sqrt{P} \times \sqrt{P}) \approx \gamma \cdot & \left(\frac{1}{P}\frac{2n^3}{3} + \frac{5n^3}{2P\log^2 P} + \frac{5n^3}{3P\log^3 P}\right) \\ & + \beta \cdot \frac{n^2}{\sqrt{P}}\left(2\log^{-1} P + 1.25\log P\right) \\ & + \alpha \cdot 3\sqrt{P}\log^3 P. \end{aligned} \tag{13}$$

We note that GEPP as implemented for example in ScaLAPACK (`PDGETRF` routine) has the same volume of communication as CALU, but requires exchanging a factor on the order of b more messages than CALU.

3.3 Pivoting Strategies and Numerical Stability

The backward stability of the LU factorization depends on the growth factor g_W, defined as,

$$g_W = \frac{\max_{i,j,k} |A^{(k)}(i,j)|}{\max_{ij} |A(i,j)|}, \tag{14}$$

where $A^{(k)}(i,j)$ denotes the entry in position (i,j) obtained after k steps of elimination. This is illustrated by the following Lemma 1.

Table 1 Bounds for the growth factor g_W obtained from different pivoting strategies for a matrix of size $m \times n$

	CALU	GEPP	CALU_PRRP	LU_PRRP
Upper bound	$2^{n(\log P+1)-1}$	2^{n-1}	$(1+\tau b)^{(n/b-1)\log P} \cdot 2^{b-1}$	$(1+\tau b)^{n/b-1} \cdot 2^{b-1}$

CALU_PRRP and LU_PRRP select pivots using strong rank revealing QR (that uses a parameter τ typically equal to 2). The reduction tree used during tournament pivoting is of height $\log P$

Lemma 1 (Lemma 9.6, Sect. 9.3 of [48]) *Let $A = LU$ be the Gaussian elimination without pivoting of A. Then $\||L||U|\|_\infty$ is bounded using the growth factor g_W by the relation $\||L||U|\|_\infty \leq (1 + 2(n^2 - n)g_W)\|A\|_\infty$.*

A comparison of the upper bound of the growth factors obtained by different pivoting strategies is given in Table 1. All the results discussed in this section hold in exact arithmetic. The growth factor of CALU is obtained by using the fact that performing CALU on a matrix A is equivalent with performing GEPP on a larger matrix formed by blocks from the original matrix A and blocks of zeros. In addition to partial pivoting (GEPP) and CALU, we also include in this table the growth factor of the LU factorization with panel rank revealing pivoting (LU_PRRP) and its communication avoiding version (CALU_PRRP), presented in [54]. We observe that the upper bound of the growth factor is larger for CALU than for GEPP. However many experiments presented in [39] show that in practice CALU is as stable as GEPP. There is one particular case of nearly singular matrices in which CALU can lead to a large growth factor, and a solution to this case is presented in a paper in preparation [24].

3.4 Selection of References for LU Factorization

The LU factorization has been largely studied in the literature, and we give here only several references. One of the first references (if not the first) to a block algorithm is [10], a paper by Barron and Swinnerton-Dyer. The authors were interested in solving a linear system of equations on EDSAC 2 computer, by using a magnetic-tape store. Hence they were interested in using as much as possible the data in main store, and reduce the number of transfers between magnetic tape and main store. They introduce two algorithms, the first one uses a pivoting strategy referred to nowadays as pairwise pivoting, the second one is the block LU factorization presented at the beginning of this section. The numerical stability of the LU factorization is studied for example in [48, 49, 68, 70]. Techniques as pairwise pivoting and block pivoting are studied in [64, 68]. In [68] it is shown experimentally that two factors are important for the numerical stability of the LU factorization, the elements of L are bounded in absolute value by a small number and the correction introduced at each step of the factorization is of rank 1. The latter property is satisfied by GEPP, CALU, LU_PRRP, and CALU_PRRP. Pairwise pivoting, parallel pivoting and their block versions do not satisfy this property, and block parallel pivoting can lead to

an exponential growth factor [68]. As shown in [39], for matrices with more than 2^{12} rows and columns, block pairwise pivoting leads to a growth of g_W which is faster than linear. Potentially this pivoting strategy can become unstable for very large matrices.

4 Communication Avoiding QR Factorization

The QR factorization decomposes a matrix $A \in \mathbb{R}^{m\times n}$ as $A = QR$, where $Q \in \mathbb{R}^{m\times m}$ is orthogonal and $R \in \mathbb{R}^{m\times n}$. We can further decompose the factors into $Q_1 \in \mathbb{R}^{m\times n}$, $Q_2 \in \mathbb{R}^{m\times(m-n)}$, and the upper triangular matrix $R_1 \in \mathbb{R}^{n\times n}$ to obtain the factorization

$$A = QR = \begin{bmatrix} Q_1 & Q_2 \end{bmatrix} \begin{bmatrix} R_1 \\ 0 \end{bmatrix} = Q_1 R_1. \tag{15}$$

If A is full rank, the thin factorization Q_1R_1 is unique (modulo signs of diagonal elements of R). We consider in this document the QR factorization based on Householder transformations. Algorithm 4 presents such a factorization. A Householder transformation is a symmetric and orthogonal matrix of the form $H = I - \frac{2}{y^T y} yy^T$, which is independent of the scaling of the vector y. When applied to a vector x, it reflects x through the hyperplane $\text{span}(y)^{\perp}$.

At each iteration k of the QR factorization from Algorithm 4, the Householder matrix $H_k = I - \tau_k y_k y_k^T$ is chosen such that all the elements of $A(k : m, k)$ are annihilated, except the first one, $H_k A(k : m, k) = \pm\|A(k : m, k)\|_2 e_1$. For more

Algorithm 4 QR factorization based on Householder transformations

Require: $A \in \mathbb{R}^{m\times n}$

1: Let $R \in \mathbb{R}^{n\times n}$ be initialized with zero matrix and $Y \in \mathbb{R}^{m\times n}$ with identity matrix
2: **for** k = 1 to n **do**
▷ Compute Householder matrix $H_k = I - \tau_k y_k y_k^T$ s.t. $H_k A(k : m, k) = \pm\|A(k : m, k)\|_2 e_1$. Store y_k in $Y()$ and τ_k in $\mathscr{T}(k)$
3: $R(k, k) = -sgn(A(k, k)) \cdot \|A(k : m, k)\|_2$
4: $Y(k + 1 : m, k) = \frac{1}{R(k,k)-A(k,k)} \cdot A(k + 1 : m, k)$ ▷ vector y_k
5: $\mathscr{T}(k) = \frac{R(k,k)-A(k,k)}{R(k,k)}$ ▷ scalar τ_k
▷ Update trailing matrix $A(k : m, k + 1 : n)$
6: $A(k : m, k + 1 : n) = (I - Y(k + 1 : m, k)\mathscr{T}(k)Y(k + 1 : m, k)^T) \cdot A(k : m, k + 1 : n)$
7: $R(k, k + 1 : n) = A(k, k + 1 : n)$
8: **end for**

Ensure: $A = QR$, where $Q = H_1 \ldots H_n = (I - \tau_1 y_1 y_1^T) \ldots (I - \tau_n y_n y_n^T)$, the Householder vectors y_k are stored in Y, and $\mathscr{T}$ is an array of size n.

details on how to compute the Householder matrix, the reader can refer to [35, 48] or to the LAPACK implementation [1]. We obtain

$$Q^T A = H_n H_{n-1} \dots H_1 A = R,$$
$$Q = (I - \tau_1 y_1 y_1^T) \dots (I - \tau_n y_n y_n^T).$$

A block version of this algorithm can be obtained by using a storage efficient representation of Q [63],

$$Q = (I - \tau_1 y_1 y_1^T) \dots (I - \tau_n y_n y_n^T) = I - YTY^T, \tag{16}$$

where Y is the matrix containing the Householder vectors as obtained in Algorithm 4 and T is computed from Y and the scalars τ_k stored in $\mathscr{T}$. As example, for $n = 2$, the compact representation is obtained as follows,

$$Y = [y_1, y_2], \quad T = \begin{bmatrix} \tau_1 & -\tau_1 y_1^T y_2 \tau_2 \\ 0 & \tau_2 \end{bmatrix}.$$

The product of two compact representations can be represented by one compact representation as follows [30],

$$Q = (I - Y_1 T_1 Y_1^T)(I - Y_2 T_2 Y_2^T) = (I - YTY^T),$$
$$Y = [Y_1, Y_2],$$
$$T = \begin{bmatrix} T_1 & -T_1 Y_1^T Y_2 T_2 \\ 0 & T_2 \end{bmatrix}.$$

A block algorithm for computing the QR factorization is obtained by partitioning the matrix A of size $m \times n$ as

$$A = \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix}, \tag{17}$$

where A_{11} is of size $b \times b$, A_{21} is of size $(m - b) \times b$, A_{12} is of size $b \times (n - b)$, and A_{22} is of size $(m - b) \times (n - b)$. The first step of the block QR factorization algorithm computes the QR factorization of the first b columns $[A_{11}; A_{21}]$ to obtain the following factorization,

$$A = \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix} = Q_1 \begin{bmatrix} R_{11} & R_{12} \\ & A_{22}^1 \end{bmatrix}.$$

The algorithm continues recursively on the trailing matrix A_{22}^1. The algebra of block QR factorization is presented in Algorithm 5.

Algorithm 5 QR factorization based on Householder transformations using a block algorithm

Require: $A \in \mathbb{R}^{m \times n}$

1: Let $R \in \mathbb{R}^{m \times n}$ be initialized with zero matrix
2: **for** k = 1 to n/b **do**
3: $k_b = (k-1) \cdot b + 1$, $k_e = k_b + b - 1$
4: Compute by using Algorithm 4 the factorization

$$A(k_b : m, k_b : k_e) = Q_k R(k_b : k_e, k_b : k_e)$$

5: Compute the compact representation $Q_k = I - Y_k T_k Y_k^T$
6: Apply Q_k^T on the trailing matrix

$$\begin{aligned} A(k_b : m, k_e + 1 : n) &= (I - Y_k T_k^T Y_k^T) A(k_b : m, k_e + 1 : n) \\ &= A(k_b : m, k_e + 1 : n) - Y_k \left(T_k^T \left(Y_k^T \left(A(k_b : m, k_e + 1 : n) \right) \right) \right) \end{aligned}$$

7: $R(k_b : k_e, k_e + 1 : n) = A(k_b : k_e, k_e + 1 : n)$
8: **end for**

Ensure: $A = QR$, where $Q = (I - Y_1 T_1 Y_1^T) \ldots (I - Y_{n/b} T_{n/b} Y_{n/b}^T)$

A parallel implementation of the QR factorization as implemented in ScaLAPACK, `PDGEQRF` routine, considers that the matrix A is distributed over a grid of processors $P = P_r \times P_c$. We do not describe here in detail the parallel algorithm. We note that similarly to the LU factorization, the latency bottleneck lies in the QR factorization of each panel, that is based on Algorithm 4. The computation of a Householder vector at each iteration k of Algorithm 4 requires computing the norm of column k. Given that the columns are distributed over P_r processors, computing the norm of each column requires a reduction among P_r processors. Hence overall a number of messages proportional to the number of columns of A needs to be exchanged during `PDGEQRF`. Such an algorithm cannot attain the lower bounds on the number of messages. We note however that `PDGEQRF` attains the lower bound on the volume of communication.

4.1 Communication Avoiding QR Factorization for a Tall and Skinny Matrix: TSQR

Consider a matrix $A \in \mathbb{R}^{m \times n}$ for which $m \gg n$. TSQR is a QR factorization algorithm that allows to minimize communication between different processors or between different levels of the memory hierarchy. It is performed as a reduction operation, in which the operator used at each step of the reduction is a QR factorization. We describe here the parallel case, for more details the reader is referred to [19]. We assume that the matrix A is distributed over P processors by

using a block row distribution. We consider in the following that $P = 4$, m is a multiple of 4, and the matrix A is partitioned among processors as,

$$A = \begin{bmatrix} A_{00} \\ A_{10} \\ A_{20} \\ A_{30} \end{bmatrix}, \tag{18}$$

where $A_{i0}, i = 0, \ldots, 3$ is of dimension $m/4 \times n$. At the first step of TSQR, each processor computes locally a QR factorization,

$$A = \begin{bmatrix} A_{00} \\ A_{10} \\ A_{20} \\ A_{30} \end{bmatrix} = \begin{bmatrix} Q_{00}R_{00} \\ Q_{10}R_{10} \\ Q_{20}R_{20} \\ Q_{30}R_{30} \end{bmatrix} = \begin{bmatrix} Q_{00} & & & \\ & Q_{10} & & \\ & & Q_{20} & \\ & & & Q_{30} \end{bmatrix} \begin{bmatrix} R_{00} \\ R_{10} \\ R_{20} \\ R_{30} \end{bmatrix}. \tag{19}$$

At the second step, the upper triangular factors $R_{i0}, i = 1 : 4$ are grouped into pairs, and each pair is factored in parallel as,

$$\begin{bmatrix} R_{00} \\ \hline R_{10} \\ R_{20} \\ R_{30} \end{bmatrix} = \begin{bmatrix} Q_{01}R_{01} \\ Q_{11}R_{11} \end{bmatrix} = \begin{bmatrix} Q_{01} & \\ & Q_{11} \end{bmatrix} \begin{bmatrix} R_{01} \\ R_{11} \end{bmatrix}. \tag{20}$$

At the last step the resulting upper triangular factors are factored as,

$$\begin{bmatrix} R_{01} \\ R_{11} \end{bmatrix} = Q_{02}R_{02}. \tag{21}$$

The QR factorization obtained by TSQR based on a binary tree is,

$$A = QR_{02}, \tag{22}$$

where

$$Q = \begin{bmatrix} Q_{00} & & & \\ & Q_{10} & & \\ & & Q_{20} & \\ & & & Q_{30} \end{bmatrix} \cdot \begin{bmatrix} Q_{01} & \\ & Q_{11} \end{bmatrix} \cdot Q_{02}. \tag{23}$$

The matrix Q is an orthogonal matrix formed by the product of three orthogonal matrices (the dimensions of the intermediate factors are chosen such that their product can be written as above). Unless it is required, the matrix Q is not formed explicitly, but it is stored implicitly. The QR factorization used at each step of

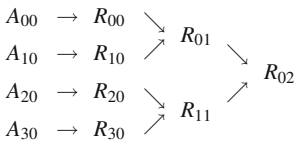


Fig. 3 Binary tree based TSQR. This figure is from [21]. Copyright ©[2012] Society for Industrial and Applied Mathematics. Reprinted with permission. All rights reserved

Algorithm 6 Parallel TSQR factorization

Require: P processors, i is my processor's index, all reduction tree with height $L = \log P$
Require: $A \in \mathbb{R}^{m \times n}$, $m \gg n$, distributed in a block row layout; $A_{i,0}$ is the block of rows belonging to my processor i
1: Compute QR factorization $A_{i,0} = Q_{i,0}R_{i,0}$
2: **for** each level k in the reduction tree from 1 to L **do**
3: $s = \lfloor i/2^k \rfloor, f = 2^k \lfloor i/2^k \rfloor, j = f + (i + 2^{k-1}) \mod 2^k$
4: $s_i = \lfloor i/2^{k-1} \rfloor, s_j = \lfloor j/2^{k-1} \rfloor$
5: Non-blocking send $R_{s_i,k-1}$ to processor j
6: Non-blocking receive $R_{s_j,k-1}$ from processor j
7: Wait until the previous send and receive have completed
8: Compute $\begin{bmatrix} R_{s_i,k-1} \\ R_{s_j,k-1} \end{bmatrix} = Q_{s,k}R_{s,k}$
9: **end for**
Ensure: $A = QR_{0,L}$, $R_{0,L}$ is available on all processors i
Ensure: Q is implicitly represented by the intermediate Q factors $\{Q_{s,k}\}$, for each node s and each level k in the all reduction tree

TSQR can be performed by using Algorithm 4 or any other efficient sequential QR factorization (as recursive QR [30]).

By using an arrow notation similar to CALU, a binary tree based parallel TSQR factorization is represented in Fig. 3. Algorithm 6 presents parallel TSQR by following its presentation from [19]. The notation used for the nodes and the levels of the all reduction tree is the same as in Algorithm 3. It can be easily seen that parallel TSQR requires exchanging only $\log P$ messages, and thus it minimizes communication. It exchanges the same volume of communication as the ScaLAPACK implementation of Householder QR (`PDGEQR2` routine), $(n^2/2) \cdot \log P$ words. In terms of floating point operations, TSQR performs $2mn^2/P + (2n^3/3) \cdot \log P$ flops, while `PDGEQR2` performs $2mn^2/P - (2n^3)/(3P)$ flops.

We note also that it is possible to reconstruct the Householder vectors of the classic Householder QR factorization (Algorithm 4) from TSQR. Let $A = QR$ be the factorization obtained from Householder QR, where A is of size $m \times n$, and let

$$Q = I - YTY^T = I - \begin{bmatrix} Y_1 \\ Y_2 \end{bmatrix} T \begin{bmatrix} Y_1^T Y_2^T \end{bmatrix} \tag{24}$$

be the compact representation of Q, where Q is of size $m \times m$. Let $Q = [Q_1, Q_2]$, where Q_1 is formed by the first n columns of Q. This is also called a *basis-kernel* representation of an orthogonal matrix, and as described in [66], there are several different basis-kernel representations. The reconstruction of Householder vectors introduced in [9] relies on the observation that

$$Q_1 - S = Y(-TY_1^T), \tag{25}$$

where S is a sign matrix which reflects the sign choice of the diagonal of R made in line 3 of Algorithm 4. Since Y is unit lower triangular and $(-TY_1^T)$ is upper triangular, this represents the unique LU decomposition of $Q_1 - S$. In other words, Y and T can be reconstructed by computing the LU decomposition of $Q_1 - S$. With this approach, denoted as TSQR-HR in [9], the performance of the algorithm becomes:

$$T_{TSQR-HR}(m, n, P) = \gamma \cdot \left(\frac{4mn^2}{P} + \frac{4n^3}{3} \log P \right) + \beta \cdot n^2 \log P + \alpha \cdot 2 \log P. \tag{26}$$

We note that this algorithm performs 2.5 times more floating point operations than TSQR. However, in practice it leads to a faster algorithm than `PDGEQR2`, as shown in [9]. It can also be used to obtain a communication avoiding QR factorization by only replacing the panel factorization in `PDGEQRF`. Faster approaches are possible, but they could be less stable. For example the Householder vectors can be reconstructed from the LU factorization of $A - R$, and this approach is stable when A is well conditioned.

4.2 Communication Avoiding QR Factorization

We consider now the case of general matrices. CAQR was introduced in [19] and it relies on using TSQR for its panel factorization. Each QR factorization performed during TSQR induces an update of the trailing matrix. Hence the update of the trailing matrix is driven by the reduction tree used during TSQR. CAQR exchanges the same volume of communication as `PDGEQRF`. But the number of messages with an optimal layout is $(3/8)\sqrt{P}\log^3 P$ for CAQR, while for `PDGEQRF` is $(5n/4)\log^2 P$. The number of floating point operations remains the same (only lower order terms change).

Another approach [9] consists in reconstructing the Householder vectors from TSQR. A communication avoiding version can be obtained by replacing the panel factorization in a classic algorithm such that the update of the trailing matrix does not change. This leads to a simpler algorithm to implement, and better performance on parallel machines, as described in [9].

5 Communication Avoiding Rank Revealing Factorization and Low Rank Matrix Approximation

In this section we consider the problem of estimating the singular values of a matrix or computing its numerical rank, a problem with many diverse applications in both scientific computing and data analytics, a detailed description can be found in [16]. One such application is computing the rank-k approximation $\tilde{A}_k$ of a matrix $A \in \mathbb{R}^{m\times n}$, $\tilde{A}_k = ZW^T$, where $Z \in \mathbb{R}^{m\times k}$, $W^T \in \mathbb{R}^{k\times n}$, and k is much smaller than m and n. Very often, this low rank approximation is used in the context of an iterative process which involves multiplying a matrix with a vector. Hence instead of computing the product Ax, which requires computing $2mn$ flops when A is dense, one could compute the product ZW^Tx with $2(m+n)k$ flops.

The best rank-k approximation of A is the rank-k truncated singular value decomposition (SVD) of A. The singular value decomposition of A is

$$A = U\Sigma V^T = \begin{bmatrix} U_1 & U_2 \end{bmatrix} \cdot \begin{bmatrix} \Sigma_1 & 0 \\ 0 & \Sigma_2 \end{bmatrix} \cdot \begin{bmatrix} V_1 & V_2 \end{bmatrix}^T,$$

where U is $m \times m$ orthogonal matrix, the left singular vectors of A, U_1 is formed by the first k vectors, U_2 is formed by the last $m-k$ vectors. Σ is of dimension $m \times n$, its diagonal is formed by $\sigma_1(A) \geq \ldots \geq \sigma_n(A)$, Σ_1 is of dimension $k \times k$ and contains the first k singular values, Σ_2 is of dimension $(m-k) \times (n-k)$ and contains the last $n-k$ singular values. V is $n \times n$ orthogonal matrix, the right singular vectors of A, V_1 is formed by the first k vectors, V_2 is formed by the last $n-k$ vectors. The rank-k truncated singular value decomposition of A is $A_k = U_1\Sigma_1 V_1^T$. Eckart and Young [29] have shown that

$$\min_{\text{rank}(\tilde{A}_k)\leq k} ||A - \tilde{A}_k||_2 = ||A - A_k||_2 = \sigma_{k+1}(A), \tag{27}$$

$$\min_{\text{rank}(\tilde{A}_k)\leq k} ||A - \tilde{A}_k||_F = ||A - A_k||_F = \sqrt{\sum_{j=k+1}^{n} \sigma_j^2(A)}. \tag{28}$$

Since computing the SVD of a matrix is very expensive, several different approaches exist in the literature to approximate the singular value decomposition which trade-off accuracy for speed. Those include the Lanczos algorithm [18, 61], rank revealing factorizations as the rank revealing QR or LU factorizations, and more recently randomized algorithms. For an overview of randomized algorithms the reader can refer to [56].

5.1 Rank Revealing QR Factorization

In this section we consider the rank revealing QR factorization based on QR factorization with column pivoting. Given a matrix $A \in \mathbb{R}^{m\times n}$, its QR factorization with column pivoting is

$$A\Pi_c = QR = Q\begin{bmatrix} R_{11} & R_{12} \\ & R_{22} \end{bmatrix}, \tag{29}$$

where Π_c is a column permutation matrix, $Q \in \mathbb{R}^{m\times m}$ is orthogonal, $R_{11} \in \mathbb{R}^{k\times k}$ is upper triangular, $R_{12} \in \mathbb{R}^{k\times(n-k)}$, $R_{22} \in \mathbb{R}^{(m-k)\times(n-k)}$. We say that this is a rank revealing factorization (RRQR) if the column permutation matrix Π_c is chosen such that

$$1 \leq \frac{\sigma_i(A)}{\sigma_i(R_{11})}, \frac{\sigma_j(R_{22})}{\sigma_{k+j}(A)} \leq q(k,n), \tag{30}$$

for any $1 \leq i \leq k$ and $1 \leq j \leq \min(m,n) - k$, where $q(k,n)$ is a low degree polynomial in n and k, and $\sigma_1(A) \geq \ldots \geq \sigma_n(A)$ are the singular values of A (we assume in this document that the singular values of A and R are all nonzero). In other words, the column permutation allows to identify a submatrix of k columns whose singular values provide a good approximation of the largest k singular values of A, while the singular values of R_{22} provide a good approximation of the $\min(m,n) - k$ smallest singular values of A. If $||R_{22}||_2$ is small and since $\sigma_{k+1}(A) \leq \sigma_{\max}(R_{22}) = ||R_{22}||_2$, then the numerical rank of A is k. Then $Q(:, 1:k)$ forms an approximate orthogonal basis for the range of A. Since $A\Pi_c \begin{bmatrix} R_{11}^{-1}R_{12} \\ -I \end{bmatrix} = Q\begin{bmatrix} 0 \\ -R_{22} \end{bmatrix}$ then $\Pi_c \begin{bmatrix} R_{11}^{-1}R_{12} \\ -I \end{bmatrix}$ are approximate null vectors.

The usage of a QR factorization to reveal the rank of a matrix was introduced in [34] and the first algorithm to compute it was introduced in [13]. With this algorithm, the absolute value of the entries of $R_{11}^{-1}R_{12}$ is bounded by $O(2^k)$ and it might fail sometimes to satisfy (30), for example on the so-called Kahan matrix [53]. However, in most cases it provides a good approximation to the SVD and it is the method of choice for estimating the singular values of a matrix through a pivoted QR factorization. We refer to this algorithm as QRCP, which stands for QR with Column Pivoting. It chooses at each step of the QR factorization the column of maximum norm and permutes it to the leading position before proceeding with the factorization.

The *strong RRQR factorization* was introduced in [44]. For a given k and a parameter $f > 1$, the results in [44] show that there exists a permutation Π_c such that

$$\left(R_{11}^{-1}R_{12}\right)_{i,j}^2 + \omega_i^2\left(R_{11}\right)\chi_j^2\left(R_{22}\right) \leq f^2, \tag{31}$$

for any $1 \leq i \leq k$ and $1 \leq j \leq n-k$, where $\omega_i(R_{11})$ denotes the 2-norm of the i-th row of R_{11}^{-1} and $\chi_j(R_{22})$ denotes the 2-norm of the j-th column of R_{22}. This inequality bounds the absolute values of the elements of $R_{11}^{-1}R_{12}$ and leads to the following bounds on singular values.

Theorem 2 (Gu and Eisenstat [44]) *Let the factorization in Eq. (29) satisfy inequality (31). Then*

$$1 \leq \frac{\sigma_i(A)}{\sigma_i(R_{11})}, \frac{\sigma_j(R_{22})}{\sigma_{k+j}(A)} \leq \sqrt{1+f^2k(n-k)}, \tag{32}$$

for any $1 \leq i \leq k$ *and* $1 \leq j \leq \min(m,n)-k$.

A strong RRQR factorization can be obtained by computing first a QR factorization with column pivoting to choose a rank k. For this rank k and a given f, additional permutations are performed until the inequality in (31) is satisfied, for a cost of $O(mnk)$ floating point operations [44].

When executed on a distributed memory computer, the matrix A is distributed over a 2D grid of processors $P = P_r \times P_c$. Finding the column of maximum norm at each step of QRCP requires a reduction operation among P_r processors, which costs $O(\log P_r)$ messages. After k steps of factorization, this requires exchanging $O(k \cdot \log P_r)$ messages. Hence, when run to completion, QRCP and its strong variant cannot attain the lower bound on communication $\Omega(\sqrt{P})$.

5.2 Tournament Pivoting for Selecting a Set of k Columns

A communication avoiding rank revealing QR factorization, referred to as CARRQR, was introduced in [23]. This factorization is based on tournament pivoting, and performs a block algorithm which computes the factorization by traversing blocks of k columns (where k is small). At each iteration, it selects k columns that are as well conditioned as possible by using a tournament which requires only $O(\log P_r)$ messages. The selected columns are permuted to the leading positions before the algorithm computes k steps of a QR factorization with no more pivoting.

Algorithm 7 describes the selection of k columns from a matrix A by using binary tree based tournament pivoting. This selection is displayed in Fig. 4, in which the matrix A is partitioned into 4 subsets of columns, $A = [A_{00}, A_{10}, A_{20}, A_{30}]$. At the leaves of the reduction tree, for each subset of columns A_{0j}, $f(A_{0j})$ selects k columns by using strong rank revealing QR factorization of A_{0j}. Then at each node of the reduction tree, a new matrix A_{ij} is obtained by adjoining the columns selected by the children of the node, and $f(A_{ij})$ selects k columns by using strong rank revealing QR factorization of A_{ij}.

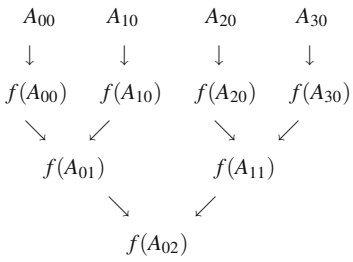


Fig. 4 Binary tree based QR factorization with tournament pivoting. This figure is from [23]. Copyright ©[2015] Society for Industrial and Applied Mathematics. Reprinted with permission. All rights reserved

Algorithm 7 QR_TP (A,k): Select k linearly independent columns from a matrix A by using QR factorization with binary tree based tournament pivoting

Require: $A \in \mathbb{R}^{m\times n}$, number of columns to select k

1: Partition the matrix $A = [A_{00}, \ldots, A_{n/k,0}]$, where $A_{i0} \in \mathbb{R}^{m\times 2k}$, $i = 1, \ldots n/(2k)$ // Assume n is a multiple of $2k$
2: **for** each level in the reduction tree $j = 0$ to $\log_2 n/(2k) - 1$ **do**
3: **for** each node i in the current level j **do**
4: **if** $j = 0$ (at the leaves of the reduction tree) **then**
5: A_{i0} is the i-th block of $2k$ columns of A
6: **else** Form A_{ij} by putting next to each other the two sets of k column candidates selected by the children of node j
7: **end if**
8: Select k column candidates by computing $A_{ij} = Q_1 R_1$ and then computing a RRQR factorization of R_1, $R_1 P_{c_2} = Q_2 \begin{bmatrix} R_2 & * \\ & * \end{bmatrix}$
9: **if** j is the root of the reduction tree **then**
10: Return Π_c such that $(A\Pi_c)(:, 1:k) = (A_{ij}\Pi_{c_2})(:, 1:k)$
11: **else** Pass the k selected columns, $A\Pi_{c_2}(:, 1:k)$ to the parent of i
12: **end if**
13: **end for**
14: **end for**

Ensure: Π_c such that $(A\Pi_c)(:, 1:k)$ are the k selected columns

It is shown in [23] that the factorization as in Eq. (29) computed by CARRQR satisfies the inequality

$$\chi_j^2\left(R_{11}^{-1}R_{12}\right) + \left(\chi_j\left(R_{22}\right)/\sigma_{\min}\left(R_{11}\right)\right)^2 \leq F_{TP}^2, \text{ for } j = 1, \ldots, n-k, \tag{33}$$

where $\chi_j(B)$ denotes the 2-norm of the j-th column of B. This inequality is very similar to the one characterizing a strong RRQR factorization. The following Theorem 3 shows that CARRQR reveals the rank by satisfying an inequality similar to (31), where the constant f is replaced by F_{TP}, a quantity which depends on the number of columns n, the rank k, and the depth of the tree used during tournament pivoting. More details can be found in [23].

Theorem 3 *Assume that there exists a permutation Π_c for which the QR factorization*

$$A\Pi_c = Q \begin{bmatrix} R_{11} & R_{12} \\ & R_{22} \end{bmatrix}, \tag{34}$$

where R_{11} is $k \times k$ and satisfies (33). Then

$$1 \leq \frac{\sigma_i(A)}{\sigma_i(R_{11})}, \frac{\sigma_j(R_{22})}{\sigma_{k+j}(A)} \leq \sqrt{1 + F_{TP}^2(n-k)}, \tag{35}$$

for any $1 \leq i \leq k$ and $1 \leq j \leq \min(m,n) - k$.

If only one step of QR with binary tree based tournament pivoting is used to select k columns of the $m \times n$ matrix A, Corollaries 2.6 and 2.7 from [23] show that the rank of A is revealed by satisfying inequality (35), with bound

$$F_{TP-BT} \leq \frac{1}{\sqrt{2k}} \left(\sqrt{2}fk\right)^{\log_2(n/k)} = \frac{1}{\sqrt{2k}} (n/k)^{\log_2(\sqrt{2}fk)}. \tag{36}$$

Given that f is a small constant and k in general is small compared to n, this bound can be seen as a polynomial in n. If tournament pivoting uses a flat tree, then the bound becomes

$$F_{TP-FT} \leq \frac{1}{\sqrt{2k}} \left(\sqrt{2}fk\right)^{n/k}, \tag{37}$$

exponential in n/k. The exponent of both bounds has an additional factor on the order of n/k if multiple steps of QR with tournament pivoting are required to reveal the rank (which is hence larger than k). However, the extensive numerical experiments performed in [23] show that both binary tree and flat tree are effective in approximating the singular values of A. For a large set of matrices, the singular values approximated with CARRQR are within a factor of 10 of the singular values computed with the highly accurate routine `dgesvj` [27, 28]. Figure 5 shows the results obtained for two matrices (from [22]), EXPONENT, a matrix whose singular values follow an exponential distribution $\sigma_1 = 1$, $\sigma_i = \alpha^{i-1}$ $(i = 2, \ldots, n)$, $\alpha = 10^{-1/11}$ [11], and SHAW, a matrix from an 1D image restoration model [46]. The plots display the singular values computed by SVD and their approximations computed by QR factorizations with column permutations (given by the diagonal values of the R factor): QR with column pivoting (QRCP), CARRQR based on binary tree tournament pivoting (CARRQR-B), and flat tree tournament pivoting (CARRQR-F). The plots also display bounds for trustworthiness,

$$\varepsilon \min\{||(A\Pi_0)(:,i)||_2, ||(A\Pi_1)(:,i)||_2, ||(A\Pi_2)(:,i)||_2\} \tag{38}$$

$$\varepsilon \max\{||(A\Pi_0)(:,i)||_2, ||(A\Pi_1)(:,i)||_2, ||(A\Pi_2)(:,i)||_2\} \tag{39}$$

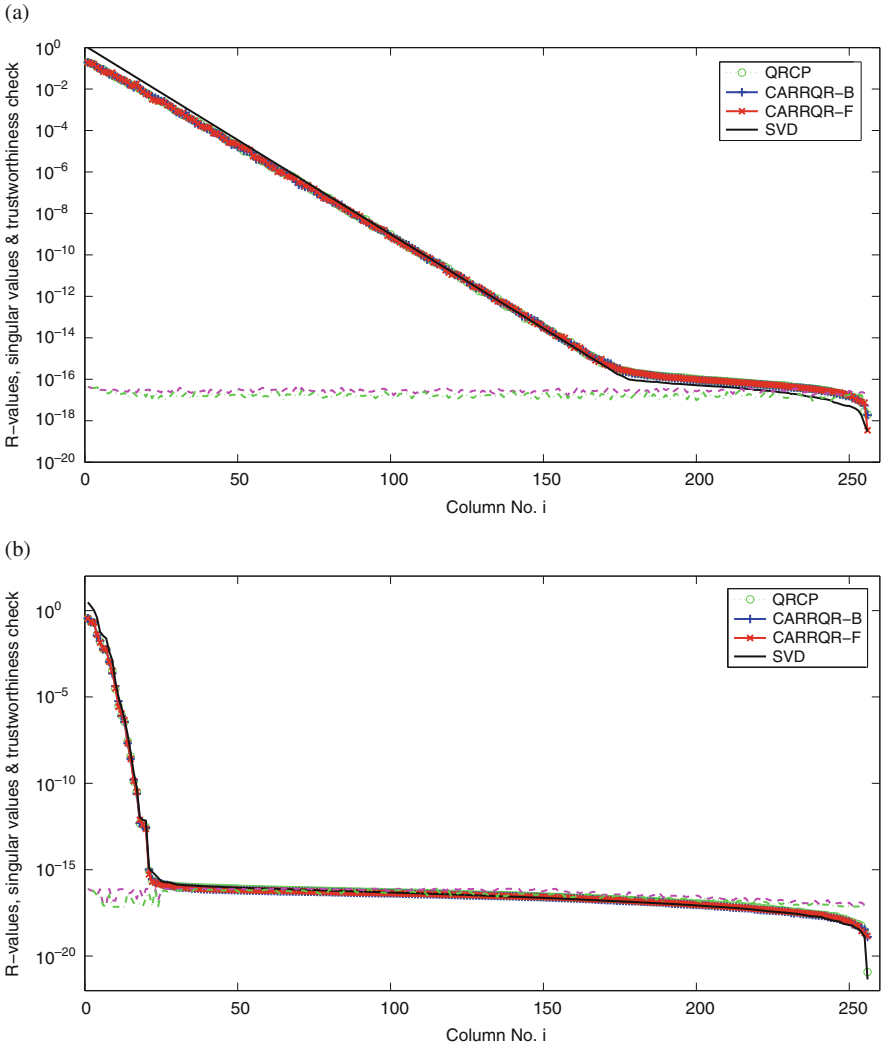


Fig. 5 Singular values as computed by SVD and approximations obtained from QRCP, CARRQR-B, and CARRQR-F. (**a**) EXPONENT. (**b**) SHAW

where $\Pi_j (j = 0, 1, 2)$ are the permutation matrices obtained by QRCP, CARRQR-B, and CARRQR-F respectively, and ε is the machine precision. Those bounds display for each column an estimate of uncertainty in any entry of that column of R as computed by the three pivoting strategies.

On a distributed memory computer, CARRQR is implemented by distributing the matrix over a 2D grid of processors $P = P_r \times P_c$. By using an optimal layout,

$P_r = \sqrt{mP/n}$, $P_c = \sqrt{nP/m}$, and $b = B \cdot \sqrt{mn/P}$, $B = 8^{-1} \log_2^{-1}(P_r) \log_2^{-1}(P_c)$, the overall performance of CARRQR (some lower order terms are ignored) is:

$$\begin{aligned} T_{CARRQR}(m,n,P) \approx \gamma \cdot \left(\frac{6mn^2 - 6n^3/3}{P} + cmn^2 \right) \\ + \beta \cdot 2 \frac{\sqrt{mn^3}}{\sqrt{P}} \left(\log_2 \sqrt{\frac{mP}{n}} + \log_2 \sqrt{\frac{nP}{m}} \right) \\ + \alpha \cdot 2^7 \sqrt{\frac{nP}{m}} \log_2^2 \sqrt{\frac{mP}{n}} \log_2^2 \sqrt{\frac{nP}{m}}, \end{aligned}$$

where $c < 1$. This shows that parallel CARRQR performs three times more floating point operations than QRCP as implemented in ScaLAPACK (routine `pdgeqpf`), and it is communication optimal, modulo polylogarithmic factors.

5.3 Low Rank Matrix Approximation for Sparse Matrices

In this section we focus on computing the low rank approximation of a sparse matrix by using rank revealing factorizations. In this case, the factors obtained by using Cholesky, LU, or QR factorizations have more nonzeros than the matrix A. The R factor obtained from the QR factorization is the Cholesky factor of A^TA, and since A^TA can be much denser than A, it is expected that its Cholesky factor has more nonzeros than the Cholesky factor of A. Hence, the QR factorization can lead to denser factors than the LU factorization. Similarly, a rank revealing QR factorization can be more expensive in terms of both memory usage and floating point operations than a rank revealing LU factorization. We present in the following LU_CRTP, a rank revealing LU factorization that also minimizes communication cost. A detailed presentation can be found in [43]. Given a desired rank k, the factorization is written as

$$\Pi_r A \Pi_c = \begin{bmatrix} \bar{A}_{11} & \bar{A}_{12} \\ \bar{A}_{21} & \bar{A}_{22} \end{bmatrix} = \begin{bmatrix} I \\ \bar{A}_{21}\bar{A}_{11}^{-1} & I \end{bmatrix} \begin{bmatrix} \bar{A}_{11} & \bar{A}_{12} \\ & S(\bar{A}_{11}) \end{bmatrix}, \tag{40}$$

where $A \in \mathbb{R}^{m \times n}$, $\bar{A}_{11} \in \mathbb{R}^{k,k}$, $S(\bar{A}_{11}) = \bar{A}_{22} - \bar{A}_{21}\bar{A}_{11}^{-1}\bar{A}_{12}$. The rank-k approximation matrix $\tilde{A}_k$ is

$$\tilde{A}_k = \begin{bmatrix} I \\ \bar{A}_{21}\bar{A}_{11}^{-1} \end{bmatrix} \begin{bmatrix} \bar{A}_{11} & \bar{A}_{12} \end{bmatrix} = \begin{bmatrix} \bar{A}_{11} \\ \bar{A}_{21} \end{bmatrix} \bar{A}_{11}^{-1} \begin{bmatrix} \bar{A}_{11} & \bar{A}_{12} \end{bmatrix}. \tag{41}$$

The second formulation of $\tilde{A}_k$ from (41) is referred to as CUR decomposition (see [56, 65, 72] and references therein), since the first factor is formed by columns of

A and the third factor is formed by rows of A. This decomposition is of particular interest for sparse matrices because its factors C and R remain sparse as the matrix A.

In LU_CRTP, the first k columns are selected by using QR with tournament pivoting of the matrix A. This leads to the factorization

$$A\Pi_c = Q \begin{bmatrix} R_{11} & R_{12} \\ & R_{22} \end{bmatrix} = \begin{bmatrix} Q_{11} & Q_{12} \\ Q_{21} & Q_{22} \end{bmatrix} \begin{bmatrix} R_{11} & R_{12} \\ & R_{22} \end{bmatrix}. \tag{42}$$

After tournament pivoting we have the QR factorization of the first k columns, $A(:, 1:k) = Q(:, 1:k)R_{11}$. The first k rows are then obtained by using QR factorization with tournament pivoting of the rows of the thin Q factor, $Q(:, 1:k)^T$,

$$\Pi_r Q = \begin{bmatrix} \bar{Q}_{11} & \bar{Q}_{12} \\ \bar{Q}_{21} & \bar{Q}_{22} \end{bmatrix},$$

such that $||\bar{Q}_{21}\bar{Q}_{11}^{-1}||_{\max} \leq F_{TP}$ and bounds for the singular values of $\bar{Q}_{11}$ with respect to the singular values of Q are governed by a low degree polynomial. This leads to the factorization,

$$\begin{aligned} \Pi_r A \Pi_c &= \begin{bmatrix} \bar{A}_{11} & \bar{A}_{12} \\ \bar{A}_{21} & \bar{A}_{22} \end{bmatrix} = \begin{bmatrix} I \\ \bar{A}_{21}\bar{A}_{11}^{-1} & I \end{bmatrix} \begin{bmatrix} \bar{A}_{11} & \bar{A}_{12} \\ & S(\bar{A}_{11}) \end{bmatrix} \\ &= \begin{bmatrix} I \\ \bar{Q}_{21}\bar{Q}_{11}^{-1} & I \end{bmatrix} \begin{bmatrix} \bar{Q}_{11} & \bar{Q}_{12} \\ & S(\bar{Q}_{11}) \end{bmatrix} \begin{bmatrix} R_{11} & R_{12} \\ & R_{22} \end{bmatrix} \end{aligned} \tag{43}$$

where

$$\begin{aligned} \bar{Q}_{21}\bar{Q}_{11}^{-1} &= \bar{A}_{21}\bar{A}_{11}^{-1}, \\ S(\bar{A}_{11}) &= S(\bar{Q}_{11})R_{22} = \bar{Q}_{22}^{-T}R_{22}. \end{aligned}$$

The following theorem from [43] shows that LU_CRTP (A, k) factorization reveals the singular values of A, and in addition also bounds the absolute value of the largest element of $S(\bar{A}_{11})$. This is important for the backward stability of the LU factorization.

Theorem 4 ([43]) *Let A be an $m \times n$ matrix. The LU_CRTP(A, k) factorization,*

$$\bar{A} = \Pi_r A \Pi_c = \begin{bmatrix} \bar{A}_{11} & \bar{A}_{12} \\ \bar{A}_{21} & \bar{A}_{22} \end{bmatrix} = \begin{bmatrix} I \\ \bar{Q}_{21}\bar{Q}_{11}^{-1} & I \end{bmatrix} \begin{bmatrix} \bar{A}_{11} & \bar{A}_{12} \\ & S(\bar{A}_{11}) \end{bmatrix} \tag{44}$$

where

$$S(\bar{A}_{11}) = \bar{A}_{22} - \bar{A}_{21}\bar{A}_{11}^{-1}\bar{A}_{12} = \bar{A}_{22} - \bar{Q}_{21}\bar{Q}_{11}^{-1}\bar{A}_{12}, \tag{45}$$

satisfies the following properties

$$\rho_l(\bar{A}_{21}\bar{A}_{11}^{-1}) = \rho_l(\bar{Q}_{21}\bar{Q}_{11}^{-1}) \leq F_{TP}, \tag{46}$$

$$||S(\bar{A}_{11})||_{\max} \leq \min\left((1 + F_{TP}\sqrt{k})||A||_{\max}, F_{TP}\sqrt{1 + F_{TP}^2(m-k)}\sigma_k(A)\right) \tag{47}$$

$$1 \leq \frac{\sigma_i(A)}{\sigma_i(\bar{A}_{11})}, \frac{\sigma_j(S(\bar{A}_{11}))}{\sigma_{k+j}(A)} \leq q(m,n,k), \tag{48}$$

for any $1 \leq l \leq m-k$, $1 \leq i \leq k$, *and* $1 \leq j \leq \min(m,n)-k$. *Here* $\rho_l(B)$ *denotes the 2-norm of the l-th row of B,* F_{TP} *is the bound obtained from QR with tournament pivoting, as in Eq.* (36), *and* $q(m,n,k) = \sqrt{\left(1 + F_{TP}^2(n-k)\right)\left(1 + F_{TP}^2(m-k)\right)}$.

The existence of a rank revealing LU factorization has been proven by Pan in [59], who shows that there are permutation matrices Π_r, Π_c such that the factorization from (40) satisfies

$$1 \leq \frac{\sigma_k(A)}{\sigma_{\min}(\bar{A}_{11})}, \frac{\sigma_{\max}(S(\bar{A}_{11}))}{\sigma_{k+1}(A)} \leq k(n-k) + 1. \tag{49}$$

The existence of a stronger LU factorization has been proven by Miranian and Gu in [57], which in addition to (49) also upper bounds $||\bar{A}_{11}^{-1}\bar{A}_{12}||_{\max}$ by a low degree polynomial in k, n, and m. Pan also introduces two algorithms for computing such a factorization which are based on the notion of local maximum volume, where the volume of a square matrix refers to the absolute value of its determinant. The first algorithm starts by performing LU factorization with conventional column pivoting (chooses as pivot the element of largest magnitude in the current row) followed by a block pivoting phase. The second algorithm relies on using the LU factorization of A^TA to perform symmetric pivoting. Experiments presented in [32] show that when there is a sufficiently large gap in the singular values of the matrix A, pivoting strategies as rook pivoting or complete pivoting produce good low rank approximations. However, they can fail for nearly singular matrices, as shown by examples given in [60].

The bounds on the approximation of singular values from (48) are worse than those from (49) showing the existence of a rank revealing LU factorization. However LU_CRTP is a practical algorithm that also minimizes communication. The bounds from (48) are also slightly worse than those obtained by CARRQR for which $q(m,n,k) = \sqrt{1 + F_{TP}^2(n-k)}$ (see Theorem 3 for more details). But for sparse matrices, CARRQR requires significantly more computations and memory, as the experimental results in [43] show. A better bound than (48) can be obtained by using strong rank revealing QR for selecting the rows from the thin Q factor in Eq. (43), in which case $||\bar{A}_{21}\bar{A}_{11}^{-1}||_{\max} \leq f$, similar to the LU factorization with panel rank revealing pivoting from [55].

The bound on the growth factor from (47) is the minimum of two quantities. The first quantity has similarities with the bound on the growth factor obtained by the

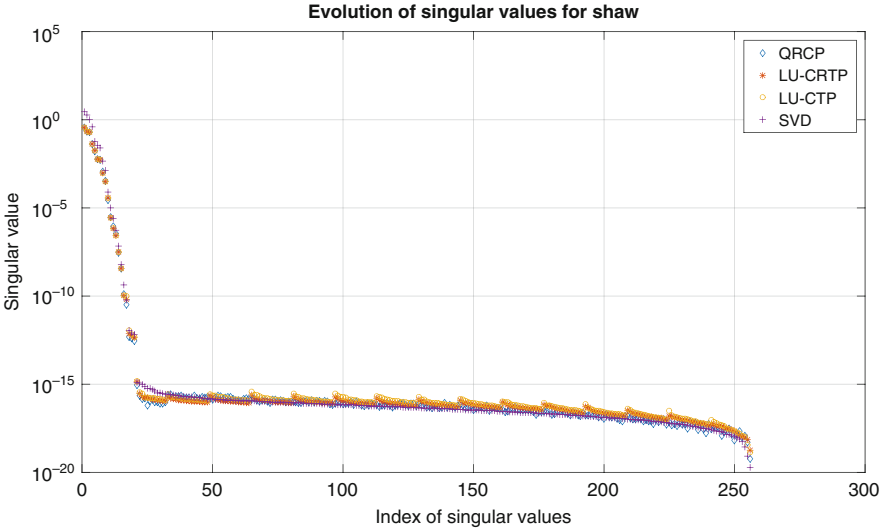


Fig. 6 Singular values as computed by SVD and as approximated by LU_CRTP (LU with column and row tournament pivoting) and LU_CTP (LU with column tournament pivoting and row partial pivoting) for SHAW matrix

LU factorization with panel rank revealing pivoting from [55]. The second quantity is new and it relates the growth factor obtained after k steps of factorization to $\sigma_k(A)$.

Experiments reported in [43] show that LU_CRTP approximates well the singular values. For the matrices considered in that paper, the ratio of the singular values approximated by LU_CRTP to the singular values computed by SVD is at most 13 (and 27 for the devil's stairs, a more difficult matrix). Figure 6 [43] shows the results obtained for SHAW matrix, the 1D image restoration matrix also used in Sect. 5.2.

References

1. Anderson, E., Bai, Z., Bischof, C., Blackford, S., Demmel, J., Dongarra, J., Du, J. Croz, Greenbaum, A., Hammarling, S., McKenney, A., Sorensen, D.: LAPACK Users' Guide. SIAM, Philadelphia, PA (1999)
2. Ballard, G., Demmel, J., Dumitriu, I.: Communication-optimal parallel and sequential eigenvalue and singular value algorithms. Tech. Report EECS-2011-14, UC Berkeley, Feb 2011
3. Ballard, G., Demmel, J., Holtz, O., Schwartz, O.: Graph expansion and communication costs of fast matrix multiplication. In: Proceedings of the 23rd ACM Symposium on Parallelism in Algorithms and Architectures, SPAA'11, pp. 1–12. ACM, New York (2011)
4. Ballard, G., Demmel, J., Holtz, O., Schwartz, O.: Minimizing communication in numerical linear algebra. SIAM J. Matrix Anal. Appl. **32**(3), 866–901 (2011). http://epubs.siam.org/doi/10.1137/090769156
5. Ballard, G., Demmel, J., Holtz, O., Lipshitz, B., Schwartz, O.: Brief announcement: strong scaling of matrix multiplication algorithms and memory-independent communication lower

bounds. In: Proceedings of the 24th ACM Symposium on Parallelism in Algorithms and Architectures, SPAA '12, pp. 77–79. ACM, New York (2012)
6. Ballard, G., Demmel, J., Holtz, O., Lipshitz, B., Schwartz, O.: Communication-optimal parallel algorithm for Strassen's matrix multiplication. In: Proceedings of the 24th ACM Symposium on Parallelism in Algorithms and Architectures, SPAA '12, pp. 193–204. ACM, New York (2012)
7. Ballard, G., Buluc, A., Demmel, J., Grigori, L., Schwartz, O., Toledo, S.: Communication optimal parallel multiplication of sparse random matrices. In: Proceedings of ACM SPAA, Symposium on Parallelism in Algorithms and Architectures (2013)
8. Ballard, G., Demmel, J., Lipshitz, B., Schwartz, O., Toledo, S.: Communication efficient gaussian elimination with partial pivoting using a shape morphing data layout. In: Proceedings of 25th Annual ACM Symposium on Parallelism in Algorithms and Architectures (SPAA) (2013)
9. Ballard, G., Demmel, J., Grigori, L., Jacquelin, M., Nguyen, H.D., Solomonik, E.: Reconstructing Householder Vectors from Tall-Skinny QR. In: Proceedings of IEEE International Parallel and Distributed Processing Symposium IPDPS (2014)
10. Barron, D.W., Swinnerton-Dyer, H.P.F.: Solution of simultaneous linear equations using a magnetic-tape store. Comput. J. **3**, 28–33 (1960)
11. Bischof, C.H.: A parallel QR factorization algorithm with controlled local pivoting. SIAM J. Sci. Stat. Comput. **12**, 36–57 (1991)
12. Blackford, L.S., Choi, J., Cleary, A., D'Azevedo, E., Demmel, J.W., Dhillon, I., Dongarra, J.J., Hammarling, S., Henry, G., Petitet, A., Stanley, K., Walker, D., Wha ley, R.C.: ScaLAPACK Users' Guide. SIAM, Philadelphia, PA (1997)
13. Businger, P.A., Golub, G.H.: Linear least squares solutions by Householder transformations. Numer. Math. **7**, 269–276 (1965)
14. Cannon, L.E.: A cellular computer to implement the Kalman filter algorithm. Ph.D. thesis, Montana State University (1969)
15. Carson, E.: Communication-avoiding Krylov subspace methods in theory and practice. Ph.D. thesis, EECS Department, University of California, Berkeley (2015)
16. Chan, T.F., Hansen, P.C.: Some applications of the rank revealing QR factorization. SIAM J. Sci. Stat. Comput. **13**, 727–741 (1992)
17. Chronopoulos, A.T., Gear, W.: S-step iterative methods for symmetric linear systems. J. Comput. Appl. Math. **25**, 153–168 (1989)
18. Cullum, J.K., Willoughby, R.A.: Lanczos Algorithms for Large Symmetric Eigenvalue Computations, vol. I: Theory. SIAM, Philadelphia (2002)
19. Demmel, J.W., Grigori, L., Hoemmen, M., Langou, J.: Communication-optimal parallel and sequential QR and LU factorizations. Tech. Report UCB/EECS-2008-89, UC Berkeley. LAPACK Working Note 204 (2008)
20. Demmel, J.W., Hoemmen, M., Mohiyuddin, M., Yelick, K.: Avoiding communication in sparse matrix computations. In: IEEE International Symposium on Parallel and Distributed Processing, pp. 1–12 (2008)
21. Demmel, J.W., Grigori, L., Hoemmen, M., Langou, J.: Communication-optimal parallel and sequential QR and LU factorizations. SIAM J. Sci. Comput. 206–239 (2012). Short version of technical report UCB/EECS-2008-89 from 2008
22. Demmel, J.W., Grigori, L., Gu, M., Xiang, H.: Communication avoiding rank revealing QR factorization with column pivoting. Tech. Report UCB/EECS-2013-46, EECS Department, University of California, Berkeley, May 2013
23. Demmel, J.W., Grigori, L., Gu, M., Xiang, H.: Communication-avoiding rank-revealing QR decomposition. SIAM J. Matrix Anal. Appl. **36**, 55–89 (2015)
24. Demmel, J.W., Grigori, L., Gu, M., Xiang, H.: TSLU for nearly singular matrices (in preparation, 2017)
25. Donfack, S., Grigori, L., Kumar Gupta, A.: Adapting communication-avoiding LU and QR factorizations to multicore architectures. In: Proceedings of IPDPS (2010)

26. Donfack, S., Grigori, L., Gropp, W.D., Kale, V.: Hybrid static/dynamic scheduling for already optimized dense matrix factorization. In: IEEE International Parallel and Distributed Processing Symposium IPDPS (2012)
27. Drmač, Z., Veselic, K.: New fast and accurate Jacobi SVD algorithm I. SIAM J. Matrix Anal. Appl. **29**, 1322–1342 (2008)
28. Drmač, Z., Veselic, K.: New fast and accurate Jacobi SVD algorithm II. SIAM J. Matrix Anal. Appl. **29**, 1343–1362 (2008)
29. Eckart, C., Young, G.: The approximation of one matrix by another of lower rank. Psychometrika **1**, 211–218 (1936)
30. Elmroth, E., Gustavson, F., Jonsson, I., Kagstrom, B.: Recursive blocked algorithms and hybrid data structures for dense matrix library software. SIAM Rev. **46**, 3–45 (2004)
31. Erhel, J.: A parallel GMRES version for general sparse matrices. Electron. Trans. Numer. Anal. **3**, 160–176 (1995)
32. Foster, L.V., Liu, X.: Comparison of rank revealing algorithms applied to matrices with well defined numerical ranks. www.math.sjsu.edu/~foster/rank/rank_revealing_s.pdf (2006)
33. Frigo, M., Leiserson, C.E., Prokop, H., Ramachandran, S.: Cache-oblivious algorithms. In: FOCS '99: Proceedings of the 40th Annual Symposium on Foundations of Computer Science. IEEE Computer Society, New York (1999)
34. Golub, G.H.: Numerical methods for solving linear least squares problems. Numer. Math. **7**, 206–216 (1965)
35. Golub, G.H., Van Loan, C.F.: Matrix Computations, 3rd edn. Johns Hopkins University Press, Baltimore, MD (1996)
36. Graham, S.L., Snir, M., Patterson, C.A. (eds.): Getting Up to Speed: The Future of Supercomputing. National Academies Press, Washington, DC (2005)
37. Grigori, L., Demmel, J.W., Xiang, H.: Communication avoiding Gaussian elimination. In: Proceedings of the ACM/IEEE SC08 Conference (2008)
38. Grigori, L., David, P.-Y., Demmel, J., Peyronnet, S.: Brief announcement: lower bounds on communication for direct methods in sparse linear algebra. In: Proceedings of ACM SPAA (2010)
39. Grigori, L., Demmel, J., Xiang, H.: CALU: a communication optimal LU factorization algorithm. SIAM J. Matrix Anal. Appl. **32**, 1317–1350 (2011)
40. Grigori, L., Stompor, R., Szydlarski, M.: A parallel two-level preconditioner for cosmic microwave background map-making. In: Proceedings of the ACM/IEEE Supercomputing SC12 Conference (2012)
41. Grigori, L., Jacquelin, M., Khabou, A.: Performance predictions of multilevel communication optimal LU and QR factorizations on hierarchical platforms. In: Proceedings of International Supercomputing Conference. LNCS (2014)
42. Grigori, L., Moufawad, S., Nataf, F.: Enlarged Krylov subspace conjugate gradient methods for reducing communication. SIAM J. Matrix Anal. Appl. **37**(2), 744–773 (2016). http://epubs.siam.org/doi/10.1137/140989492. Preliminary version published as Inria TR 8597
43. Grigori, L., Cayrols, S., Demmel, J.W.: Low rank approximation of a sparse matrix based on LU factorization with column and row tournament pivoting. Research Report RR-8910, INRIA. Submitted to SIAM Journal on Scientific Computing (2016, in revision)
44. Gu, M., Eisenstat, S.C.: Efficient algorithms for computing a strong rank-revealing QR factorization. SIAM J. Sci. Comput. **17**, 848–869 (1996)
45. Gustavson, F.: Recursion leads to automatic variable blocking for dense linear-algebra algorithms. IBM J. Res. Dev. **41**, 737–755 (1997)
46. Hansen, P.C.: Regularization tools version 4.0 for Matlab 7.3. Numer. Algorithms **46**(2), 189–194 (2007)
47. Hestenes, M.R., Stiefel, E.: Methods of conjugate gradients for solving linear systems. J. Res. Nat. Bur. Stand. **49**, 409–436 (1952)
48. Higham, N.: Accuracy and Stability of Numerical Algorithms, 2nd edn. SIAM, Philadelphia (2002)

49. Higham, N., Higham, D.J.: Large growth factors in Gaussian elimination with pivoting. SIAM J. Matrix Anal. Appl. **10**, 155–164 (1989)
50. Hoemmen, M.F.: Communication-avoiding Krylov subspace methods. Ph.D. thesis, EECS Department, University of California, Berkeley (2010)
51. Hong, J.-W., Kung, H.T.: I/O complexity: the Red-Blue Pebble Game. In: STOC '81: Proceedings of the 13th Annual ACM Symposium on Theory of Computing, pp. 326–333. ACM, New York (1981)
52. Irony, D., Toledo, S., and Tiskin, A.: Communication lower bounds for distributed-memory matrix multiplication. J. Parallel Distrib. Comput. **64**, 1017–1026 (2004)
53. Kahan, W.M.: Numerical linear algebra. Can. Math. Bull. **9**, 757–801 (1966)
54. Khabou, A., Demmel, J., Grigori, L., Gu, M.: Communication avoiding LU factorization with panel rank revealing pivoting. SIAM J. Matrix Anal. Appl. **34**, 1401–1429 (2013). Preliminary version published as INRIA TR 7867
55. Khabou, A., Demmel, J.W., Grigori, L., Gu, M.: Communication avoiding LU factorization with panel rank revealing pivoting. SIAM J. Matrix Anal. Appl. **34**, 1401–1429 (2013)
56. Mahoney, M.W.: Randomized algorithms for matrices and data. Found. Trends Mach. Learn. **3**, 123–224 (2011)
57. Miranian, L., Gu, M.: Strong rank revealing LU factorizations. Linear Algebra Appl. **367**, 1–16 (2003)
58. O'Leary, D.P.: The block conjugate gradient algorithm and related methods. Linear Algebra Appl. **29**, 293–322 (1980)
59. Pan, C.-T.: On the existence and computation of rank-revealing LU factorizations. Linear Algebra Appl. **316**, 199–222 (2000)
60. Peters, G., Wilkinson, J.H.: The least squares problem and pseudo-inverses. Comput. J. **13**, 309–316 (1970)
61. Saad, Y.: Numerical Methods for Large Eigenvalue Problems, 2nd edn. SIAM, Philadelphia (2011)
62. Saad, Y., Schultz, M.H.: GMRES: a generalized minimal residual algorithm for solving non-symmetric linear systems. SIAM J. Sci. Stat. Comput. **7**, 856–869 (1986)
63. Schreiber, R., Van Loan, C.: A storage-efficient WY representation for products of Householder transformations. SIAM J. Sci. Stat. Comput. **10**, 53–57 (1989)
64. Sorensen, D.C.: Analysis of pairwise pivoting in Gaussian elimination. IEEE Trans. Comput. **3**, 274–278 (1985)
65. Stewart, G.W.: Four algorithms for the efficient computation of truncated QR approximations to a sparse matrix. Numer. Math. **83**, 313–323 (1999)
66. Sun, X., Bischof, C.: A basis-kernel representation of orthogonal matrices. SIAM J. Matrix Anal. Appl. **16**, 1184–1196 (1995)
67. Toledo, S.: Locality of reference in LU decomposition with partial pivoting. SIAM J. Matrix Anal. Appl. **18**(4), 1065–1081 (1997)
68. Trefethen, L.N., Schreiber, R.S.: Average-case stability of Gaussian elimination. SIAM J. Matrix Anal. Appl. **11**, 335–360 (1990)
69. van der Vorst, H.A.: Bi-CGSTAB: a fast and smoothly converging variant of Bi-CG for the solution of nonsymmetric linear systems. SIAM J. Sci. Stat. Comput. **13**, 631–644 (1992)
70. Wilkinson, J.H.: Error analysis of direct methods of matrix inversion. J. Assoc. Comput. Mach. **8**, 281–330 (1961)
71. Wulf, W., McKee, S.: Hitting the wall: implications of the obvious. ACM SIGArch Comput. Archit. News **23**, 20–24 (1995)
72. Zamarashkin, N.L., Goreinov, S.A., Tyrtyshnikov, E.E.: A theory of pseudoskeleton approximations. Linear Algebra Appl. **261**, 1–21 (1997)

Part II
Applications

Singular Traveling Waves and Non-linear Reaction-Diffusion Equations

Juan Calvo

Abstract We review some recent results on singular traveling waves arising as solutions to reaction-diffusion equations combining flux saturation mechanisms and porous media type terms. These can be regarded as toy models in connection with some difficulties arising on the mathematical modelization of several scenarios in Developmental Biology, exemplified by pattern formation in the neural tube of chick's embryo.

1 Pattern Formation in Morphogenesis

Morphogenic proteins play a key role in Developmental Biology, acting as signaling molecules mediating intracellular communication. In particular they mediate cellular differentiation processes like those taking place during embryonic development. Understanding how morphogens induce distinct cell fates becomes then a paramount issue.

Morphogenic proteins are usually issuing from localized sources in the extracellular medium, originating a concentration gradient. Several mathematical models have been proposed to explain how morphogens are transported through the extracellular matrix; these have been usually based on reaction-diffusion equations after the pioneering works of Turing, Crick and Meinhardt [13, 19, 24]. Reaction terms account for the set of chemical reactions (known as the signaling pathway) taking place inside each cell after morphogens attach to their membrane receptors; the final result of these intracellular processes is a specific change in gene transcription.

An important scenario which has been the subject of intensive research is that of the neural tube (particularly in chick embryos), which is the precursor of the spinal cord in the adult individual. Owing to the natural propagation direction in this structure, one-dimensional reaction-diffusion models have been widely used to describe how gradients of morphogen concentration are dynamically created in the neural tube, see e.g. [22] and references therein. Such mathematical models assume

J. Calvo (✉)
Facultad de Ciencias, Universidad de Granada, Campus de Fuentenueva, 18071 Granada, Spain
e-mail: juancalvo@ugr.es

M. Mateos, P. Alonso (eds.), *Computational Mathematics, Numerical Analysis and Applications*, SEMA SIMAI Springer Series 13,
DOI 10.1007/978-3-319-49631-3_5

that morphogens are transported through the medium by means of linear diffusion, whereas chemical reactions taking place within individual cells can be described in terms of a system of ordinary differential equations.

Several recent experimental findings have come to question the validity of the linear diffusion assumption in this context (see [23, 26] for an account of this). Here we focus on the results by Dessaud et al. [14], stating that the concentration of morphogen that cells receive and the exposure time have the same importance (e.g. very small morphogen concentration can exert noticeable effects if the exposure time is long enough). To see why this property cannot be replicated by a linear diffusion model, let us consider the one-dimensional FKPP equation [15, 17],

$$u_t = \nu u_{xx} + k\,u(1-u)$$

which displays classical, C^∞-smooth traveling waves $u(t,x) = u(x-\sigma t)$ for wavespeeds $\sigma \geq 2\sqrt{k\nu}$. These traveling profiles are supported in the whole real line, matching with zero by means of an exponentially decaying profile. This is rooted in the fact that the linear diffusion equation has the property of infinite speed of propagation. It entails the fact that traveling waves as such propagate some (chemical) information instantaneously, which spoils any attempt to track exposure times on the sole basis of this model. In fact, no a posteriori engineering procedures seem to quantify in a reasonable way experimental observations [23, 26].

It is mandatory to have mathematical descriptions allowing to track in a very precise way exposure times for the sake of having accurate models for morphogenesis (and specifically for the case of the neural tube). It is a natural idea to test if nonlinear diffusion can perform better in this setting, particularly when models having finite propagation speed are used. We try to get some clues dealing with simplified settings in the following section.

2 Nonlinear Reaction-Diffusion Models

Describing traveling wave solutions in nonlinear reaction-diffusion equations constitutes a full research area in itself. It is tempting to think that traveling waves for reaction diffusion equations having finite speed of propagation will be supported on half lines. The actual scenario is a bit more complicated. As a prototypical example, we may consider the porous medium equation (see e.g. [25]) coupled with a logistic reaction term,

$$u_t = \nu(u^{m-1}u_x)_x + k\,u(1-u), \quad m > 1. \tag{1}$$

For each value of $m > 1$ there is a one-parametric family of traveling wave solutions. All the members of this family are supported on the whole real line, except for the slowest wave of each family, which is a continuous profile which is supported on a half line. See [20] for details.

This subject has been treated in great generality to find that this behavior is not specific of porous media equations but rather of parabolic equations with finite propagation speed, see the book [16]. Here we want to draw attention on a family of degenerate parabolic equations having the property of finite speed of propagation which does not fall in the scope of [16]. These are know as flux saturated or flux limited diffusion equations, arguably introduced in the works by Rosenau [21] and Levermore and Pomraning [18]. A prototypical example is

$$\frac{\partial u}{\partial t} = \nu \operatorname{div}\left(\frac{|u|\nabla_x u}{\sqrt{u^2 + \frac{\nu^2}{c^2}|\nabla_x u|^2}}\right). \tag{2}$$

Note that when $c \to \infty$ we get the heat equation. This model is known in the mathematical literature as the "relativistic heat equation" after [5], in which a connection with optimal transport theory was found. In fact, the form of the cost function hints that (2) should have a finite speed of propagation given by c above, a fact that was also pointed out in [21]. This is proved in [3]. In fact, this model is also able to propagate discontinuous interfaces, which is a desirable feature for the morphogen transport problem, as we explain below. These properties are somewhat natural in the light of the degenerate functional framework which is needed to tackle such models, see e.g. [1, 2] (where an entropy solution framework is introduced). For a recent account on the research done on flux-saturated equations we refer to [8].

In the light of the previous considerations, we want to probe what sort of traveling waves arise in connection with these degenerate diffusion mechanisms. As a test case, we can mix the mechanisms of flux saturation and porous-media-type diffusion and consider the following equation for $m \geq 1$:

$$u_t = \nu \left(\frac{u^m u_x}{\sqrt{|u|^2 + \frac{\nu^2}{c^2}|u_x|^2}}\right)_x + k\,u(1-u). \tag{3}$$

It was shown in [9, 11] that this family of equations admits traveling wave solutions which are supported on a half line and whose interfaces are discontinuous. When $m = 1$ such singular waves exist only for wavespeeds equal to c, any other (necessarily faster) traveling waves are classical. If $m > 1$ then there exist two bifurcation values $\sigma_{smooth} > \sigma_{ent}$. If the wavespeed coincides with σ_{ent} then the corresponding traveling waves are again discontinuous and supported on a half line. If the wavespeed exceeds σ_{smooth} then the associated waves are classical, while those with speeds between the two bifurcation values consist on two smooth branches joined by a jump discontinuity and their support is the whole line. These results are proved by reducing the problem of constructing wave solutions to describing the orbits of a planar dynamical system. Singular traveling waves arise as admissible concatenation of several orbits of the planar diagram, being the compatibility

conditions at the matchin points given by Rankine–Hugoniot's jump conditions plus some geometric information coming from the entropy solution framework, see [12].

These results in [9, 11] are no isolated phenomena but part of a robust framework. It has been tested that similar families of traveling profiles are obtained under a number of generalizations of (3), see [6, 8–10].

3 Nonlinear Models for Morphogen Propagation

The analytical findings in the previous section suggest that replacing the linear diffusion mechanism on mophogen propagation models with a nonlinear mechanism having finite speed of propagation may allow to get a better description of the overall dynamics. One of the concerns we need to address is to track exposure times carefully, a task for which saturation mechanisms like that in (2) seem quite suited -note that propagation speed is universal, in contrast with the porous medium case, for which it depends on the initial datum [25]. This specific feature was tested on a simplified model for morphogen propagation in the neural tube introduced in [7],

$$\begin{cases} \dfrac{\partial u}{\partial t} = \left(\dfrac{|u|u_x}{\sqrt{u^2 + \frac{v^2}{c^2}|u_x|^2}} \right)_x & \text{in} \quad [0,T] \times [0,L]\,, \\ \\ -\mathbf{a}(u(t,0), u_x(t,0)) = \beta > 0 \text{ and } u(t,L) = 0 \quad \text{on } t \in [0,T]\,, \end{cases} \tag{4}$$

The non-homogeneous Neumann boundary condition encodes the fact that there is an incoming morphogenic signal. It was shown in [4] that the incoming flux of morphogens propagates with speed c in the form of a sharp traveling front, quite related with the results mentioned in the previous section. This supports the proposal of a complete model in [26]. Morphogen propagation along the neural tube would be described by an equation like (4), with the addition of suitable reaction terms accounting for attachment and detachment effects linked with the availability of membrane receptors at each individual cell. This partial differential equation is coupled with a system of ordinary differential equations representing the signaling pathway at each cell according to the amount of attached morphogens and the time of exposure to their action. In such a way, the chemical signal is propagated as a traveling front (as shown by numerical simulations), thus allowing different biological responses at different times. We refer to [26] for a detailed exposition. We think that the model proposed in [26] opens a new perspective on the subject, since its qualitative behavior is in close correspondence with biological observations, opposed to what is predicted by linear diffusion models. Therefore, morphogen propagation seems to be an inherently nonlinear process, of which some features could be well approximated by some nonlinear diffusion mechanisms. The proposal

in [26] can be an interesting departing point from which we may develop more accurate theories and models.

References

1. Andreu, F., Caselles, V., Mazón, J.M.: A strongly degenerate quasilinear elliptic equation. Nonlinear Anal. **61**, 637–669 (2005)
2. Andreu, F., Caselles, V., Mazón, J.M.: The cauchy problem for a strongly degenerate quasilinear equation. J. Eur. Math. Soc. (JEMS) **7**, 361–393 (2005)
3. Andreu, F., Caselles, V., Mazón, J.M., Moll, S.: Finite propagation speed for limited flux diffusion equations. Arch. Ration. Mech. Anal. **182**, 269–297 (2006)
4. Andreu, F., Calvo, J., Mazón, J.M., Soler, J.: On a nonlinear flux–limited equation arising in the transport of morphogens. J. Differ. Equ. **252**, 5763–5813 (2012)
5. Brenier, Y.: Extended Monge-Kantorovich theory. In: Caffarelli, L.A., Salsa, S. (eds.) Lecture Notes in Mathematics, vol. 1813, pp. 91–122. Springer, New York (2003)
6. Calvo, J.: Analysis of a class of diffusion equations with a saturation mechanism. SIAM J. Math. Anal. **47**, 2917–2951 (2015)
7. Calvo, J., Mazón, J.M., Soler, J., Verbeni, M.: Qualitative properties of the solutions of a nonlinear flux-limited equation arising in the transport of morphogens. Math. Models Methods Appl. Sci. **21**, 893–937 (2011)
8. Calvo, J., Campos, J., Caselles, V., Sánchez, O., Soler, J.: Qualitative behavior for fluz-saturated mechanisms: traveling waves, waiting times and smoothing effects. EMS Surv. Math. Sci. **2**, 2917–2951 (2015)
9. Calvo, J., Campos, J., Caselles, V., Sánchez, O., Soler, J.: Pattern formation in a flux limited reaction-diffusion equation of porous media type. Invent. Math. **206**, 57–108 (2016)
10. Campos, J., Soler, J.: Qualitative behavior and traveling waves for flux-saturated porous media equations arising in optimal mass transportation. Nonlinear Anal. **137**, 266–290 (2016)
11. Campos, J., Guerrero, P., Sánchez, O., Soler, J.: On the analysis of traveling waves to a nonlinear flux limited reaction-diffusion equation. Ann. Inst. H. Poincaré Anal. Non Linéaire **30**, 141–155 (2013)
12. Caselles, V.: On the entropy conditions for some flux limited diffusion equations. J. Differ. Equ. **250**, 3311–3348 (2011)
13. Crick, F.: Diffusion in embryogenesis. Nature **40**, 561–563 (1970)
14. Dessaud, E., Yang, L.L., Hill, K., Cox., B., Ulloa, F., Ribeiro, A., Mynett, A., Novitch, B.G., Briscoe, J.: Interpretation of the sonic hedgehog morphogen gradient by a temporal adaptation mechanism. Nature **450**, 717–720 (2007)
15. Fisher, R.A.: The wave of advance of advantageous genes. Ann. Eugen. **7**, 335–369 (1037)
16. Gilding, B.H., Kersner, R.: Traveling Waves in Nonlinear Diffusion-Convection-Reaction. Birkhäuser Verlag, Basel (2004)
17. Kolmogoroff, A.N., Petrovsky, I.G., Piscounoff, N.S., Étude de l'equation de la diffusion avec croissance de la quantité de matiére et son application á un probléme biologique. Bull. Univ. de Etatá Moscou Ser. Int. A **1**, 1–26 (1937)
18. Levermore, C.D., Pomraning, G.C.: A flux-limited diffusion theory. Astrophys. J. **248**, 321–334 (1981)
19. Meinhardt, H.: Space-dependent cell determination under the control of a morphogen gradient. J. Theor. Biol. **74**, 307–321 (1978)
20. Newman, W.I.: Some exact solutions to a non-linear diffusion problem in population genetics and combustion. J. Theor. Biol. **85**, 325–334 (1980)
21. Rosenau, P.: Tempered diffusion: a transport process with propagating front and inertial delay. Phys. Rev. A **46**, 7371–7374 (1992)

22. Saha, K., Schaffer, D.V.: Signal dynamics in sonic hedgehog tissue patterning. Development **133**, 889–900 (2006)
23. Sánchez, O., Calvo, J., Ibáñez, C., Guerrero, I., Soler, J.: Modeling hedgehog signaling through flux-saturated mechanisms. In: Riobo, N.A. (ed.) Methods in Molecular Biology, vol. 1322, pp. 19–33. Springer, New York (2015)
24. Turing, A.M.: The chemical basis of Morphogenesis. Philos. Trans. R. Soc. Lond. Ser. B Biol. Sci. **237**, 37–72 (1952)
25. Vazquez, J.L.: The Porous Medium Equation: Mathematical Theory. Oxford University Press, Oxford (2007)
26. Verbeni, M., Sánchez, O., Mollica, E., Siegl–Cachedenier, I., Carleton, A., Guerrero, I., Ruiz i Altaba, A., Soler, J.: Morphogenetic action through flux-limited spreading. Phys. Life Rev. **10**, 457–475 (2013)

Numerical Simulation of Flows Involving Singularities

Maria Garzon, James A. Sethian, and August Johansson

Abstract Many interesting fluid interface problems involve singular events, as breaking-up or merging of the physical domain. In particular, wave propagation and breaking, droplet and bubble break-up, electro-jetting, rain drops, etc. are good examples of such processes. All these mentioned problems can be modeled using the potential flow assumptions, in which an interface needs to be advanced by a velocity determined by the solution of a surface partial differential equation posed on this moving boundary. The standard approach, the Lagrangian-Eulerian formulation together with some sort of front tracking method, is prone to fail when break-up or merging processes appear. The embedded formulation using level sets seamlessly allows topological breakup or merging of the fluid domain. In this work we present the numerical approximation of the embedded model and some computational results regarding electrohydrodynamic applications.

1 The Embedded Model Equations

Let $\Omega_1(t)$ be a fluid domain immersed in an infinite exterior fluid $\Omega_2(t)$, Γ_t be the free boundary separating both domains, and Ω_D be a fixed domain that should contain the free boundary for all $t \in [0, T]$. The level set/extended potential flow model, [3, 4], may be then written as:

$$\mathbf{u} = \nabla\phi \ \text{ in } \ \Omega_1(t) \tag{1}$$

$$\Delta\phi = 0 \ \text{ in } \ \Omega_1(t) \tag{2}$$

M. Garzon (✉)
Universidad de Oviedo, Oviedo, Spain
e-mail: maria.garzon.martin@gmail.com

J.A. Sethian
University of Berkeley, Berkeley, CA, USA
e-mail: sethian@math.berkeley.edu

A. Johansson
Center for Biomedical Computing, Simula, Norway
e-mail: august@simula.no

M. Mateos, P. Alonso (eds.), *Computational Mathematics, Numerical Analysis and Applications*, SEMA SIMAI Springer Series 13,
DOI 10.1007/978-3-319-49631-3_6

$$\Psi_t + \mathbf{u}_{\text{ext}} \cdot \nabla \Psi = 0 \text{ in } \Omega_D \tag{3}$$

$$G_t + \mathbf{u}_{\text{ext}} \cdot \nabla G = f_{\text{ext}} \text{ in } \Omega_D. \tag{4}$$

Here, ϕ is the velocity potential, $\mathbf{u}$ the velocity field, Ψ the level set function, G the extended potential function, f accounts for the surface forces, and the subscript "ext" refers to the extended quantities off the front into Ω_D. This hydrodynamic problem can be coupled with any other exterior problem on $\Omega_2(t)$. In particular, assuming a uniform electric field $\mathbf{E}$ in $\Omega_2(t)$, acting in the direction of the z axis and $\mathbf{E} = 0$ in $\Omega_1(t)$ (perfect conductor fluid) then:

$$\mathbf{E} = -\nabla U \text{ in } \Omega_2(t) \tag{5}$$

$$\Delta U = 0 \text{ in } \Omega_2(t) \tag{6}$$

$$U = U_0 \text{ on } \Gamma_t \tag{7}$$

$$U = -E_\infty z \text{ at the far field,} \tag{8}$$

where U is the electric potential and E_∞ is the electric field intensity.

2 Numerical Approximation

The semidiscretization in time of the model equations is:

$$\mathbf{u}^n = \nabla \phi^n \text{ in } \Omega_1(t_n) \tag{9}$$

$$\Delta \phi^n(r,z) = 0 \text{ in } \Omega_1(t_n) \tag{10}$$

$$\frac{\Psi^{n+1} - \Psi^n}{\Delta t} = -\mathbf{u}^n_{\text{ext}} \cdot \nabla \Psi^n \text{ in } \Omega_D \tag{11}$$

$$\frac{G^{n+1} - G^n}{\Delta t} = -\mathbf{u}^n_{\text{ext}} \cdot \nabla G^n + f^n_{\text{ext}} \text{ in } \Omega_D, \tag{12}$$

$$\Delta U^n(r,z) = 0 \text{ in } \Omega_2(t_n) \tag{13}$$

where a first order explicit scheme has been applied. For the space discretization of Eqs. (11) and (12) a first order or second order upwind scheme can be used. The approximation of (10) and (13) is crucial in this numerical method, as it provides the velocity to advance the free boundary and also the velocity potential evolution within this front. We have coupled the following solvers for the interior and exterior Laplace equations:

- For 2D and 3D axisymmetric geometries a Galerkin boundary integral solution is established, where the boundary element method with linear elements have been used to approximate the integral equations, see [6, 8].

- For the fully 3D approximation a non conforming Nitsche finite element method has been used together with stabilization techniques of the bilinear forms, as the jump stabilization or the ghost penalty stabilization, see [1, 2, 9].

3 Numerical Results

Several physical scenarios can be simulated using the assumptions and the numerical method presented here. In the case of pure hydrodynamic problems, Eqs. (1)–(4), results for the wave breaking phenomena in a 2D geometry have been presented in [3], where splitting of the fluid domain was not considered. The first simulation involving computations through singular events was presented in [4], where the pinch-off of an infinite fluid jet and subsequent cascade of drop formation was reproduced in a seamless 3D axi-symmetric computation. In Fig. 1 we present the comparison of the satellite break up simulation with laboratory photographs. The interaction of two inviscid fluids of different densities was studied in [5]. The only parameter in the non-dimensional model is the fluid density ratio and simulations of the breaking up transition patterns from air bubbles to water droplets have been

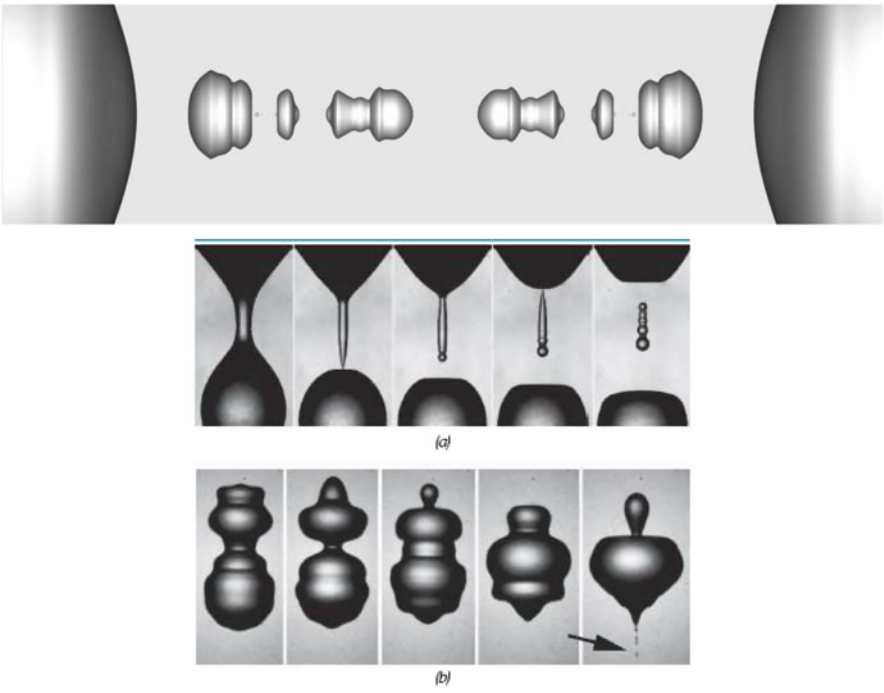


Fig. 1 Satellite drop breaking up, computed profiles (**a**) and Laboratory photographs (**b**), see [10]. Reproduced from [4] with permission from Elsevier

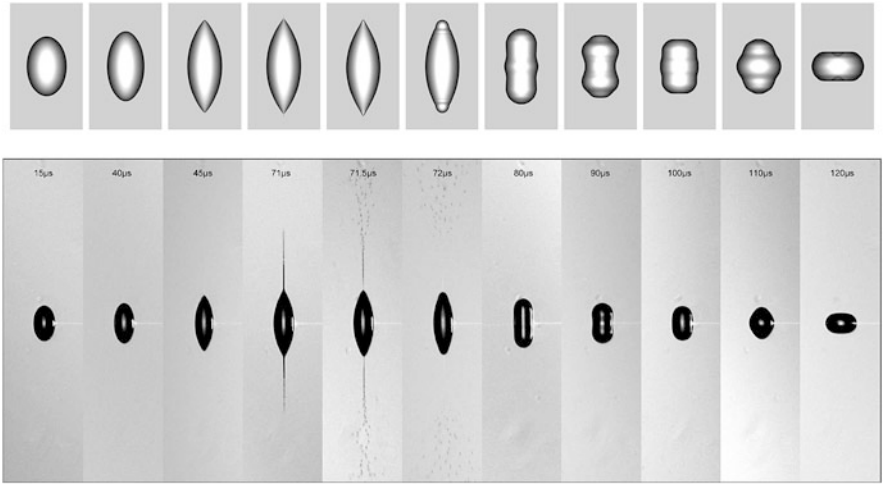


Fig. 2 Laboratory snapshots at indicated times of the evolution of a surface charged super-cooled water droplet, reprinted figure with permission from E. Giglio, D. Duft and T. Leisner, Phys. Rev. E, 77, 036319 (2008). Copyright (2008) by the American Physical Society (*bottom*); and computed profiles at times 80, 101.2, 108.1, 108.5, 109.8, 112.1, 124.2, 133.4, 138, 142, 154.1 μs (*top*)

computed. When electrical forces acting on the free surface are also considered, Eqs. (1)–(8), the flow gets even more interesting: a charged water droplet will elongate until Taylor cones are formed, from which fine filaments will be ejected from both drop tips. As soon as the drop losses enough charge, it will recoil and oscillate back to equilibrium. In Fig. 2 we show also a comparison between computed profiles on top and Laboratory experiments on bottom at corresponding times. See [7].

Acknowledgements This work was supported by the U.S. Department of Energy, under contract Number DE-AC02-05CH11231, the Spanish Ministry of Science and Innovation, Project Number MTM2013-43671-P. The third author was also supported by the Research Council of Norway through a Centers of Excellence grant to the Center for Biomedical Computing at Simula Research Laboratory, Project Number 179578.

References

1. Burman, E., Hansbo, P.: Fictitious domain finite element methods using cut elements: I. A stabilized Lagrange multiplier method. Comput. Methods Appl. Mech. Eng. **199**, 2680–2686 (2010)
2. Burman, E., Hansbo, P.: Fictitious domain finite element methods using cut elements: II. A stabilized Nitsche method. Appl. Numer. Math. **62**(4), 328–341 (2012)
3. Garzon, M., Adalsteinsson, D., Gray, L.J., Sethian, J.A.: A coupled level set-boundary integral method for moving boundary simulations. Interfaces Free Bound. **7**, 277–302 (2005)

4. Garzon, M., Gray, L.J., Sethian, J.A.: Numerical simulation of non-viscous liquid pinch-off using a coupled level set-boundary integral method. J. Comput. Phys. **228**, 6079–6106 (2009)
5. Garzon, M., Gray, L.J., Sethian, J.A.: Simulation of the droplet-to-bubble transition in a two-fluid system. Phys. Rev. E **83**, 046318 (2011)
6. Garzon, M., Gray, L.J., Sethian, J.A.: Axisymmetric boundary integral formulation for a two-fluid system. Int. J. Numer. Methods Fluids **69**, 1124–1134 (2012)
7. Garzon, M., Gray, L.J., Sethian, J.A.: Numerical simulations of electrostatically driven jets from nonviscous droplets. Phys. Rev. E **89**, 033011 (2014)
8. Gray, L.J., Garzon, M., Mantic, V., Graciani, E.: Int. J. Numer. Methods Eng. **66**, 2014–2034 (2006)
9. Johansson, A., Garzon, M., Sethian, J.A.: A three-dimensional coupled Nitsche and level set method for electrohydrodynamic potential flows in moving domains. J. Comput. Phys. **309**, 88–111 (2016)
10. Thoroddsen, S.T., Etoh, T.G., Takehara, K.: Micro-jetting from wave focusing on oscillating drops. Phys. Fluids **19**, 052101–052116 (2007)

A Projection Hybrid Finite Volume-ADER/Finite Element Method for Turbulent Navier-Stokes

A. Bermúdez, S. Busto, J.L. Ferrín, L. Saavedra, E.F. Toro, and M.E. Vázquez-Cendón

Abstract We present a second order finite volume/finite element projection method for low-Mach number flows. Moreover, transport of species law is also considered and turbulent regime is solved using a $k - \varepsilon$ standard model. Starting with a 3D tetrahedral finite element mesh of the computational domain, the momentum equation is discretized by a finite volume method associated with a dual finite volume mesh where the nodes of the volumes are the barycenter of the faces of the initial tetrahedra. The resolution of Navier-Stokes equations coupled with a $k - \varepsilon$ turbulence model requires the use of a high order scheme. The ADER methodology is extended to compute the flux terms with second order accuracy in time and space. Finally, the order of convergence is analysed by means of academic problems and some numerical results are presented.

1 Mathematical Model

The system of equations described in this section corresponds to a model for low-Mach number flows. The underlying assumption is that the Mach number M is sufficiently small so that pressure p can be written as the sum of a spatially constant known function $\overline{\pi}$ and a small perturbation π. The perturbation will be neglected

A. Bermúdez • S. Busto (✉) • J.L. Ferrín • M.E. Vázquez-Cendón
Departamento de Matemática Aplicada, Universidade de Santiago de Compostela, Facultade de Matemáticas, ES-15782 Santiago de Compostela, Spain
e-mail: alfredo.bermudez@usc.es; saray.busto@usc.es; joseluis.ferrin@usc.es; elena.vazquez.cendon@usc.es

L. Saavedra
Departamento de Matemática Aplicada a la Ingeniería Aeroespacial, Universidad Politécnica de Madrid E.T.S.I. Aeronáuticos, ES-28040 Madrid, Spain
e-mail: laura.saavedra@upm.es

E.F. Toro
Laboratory of Applied Mathematics, DICAM, University of Trento, IT-38100 Trento, Italy
e-mail: eleuterio.toro@unitn.it

M. Mateos, P. Alonso (eds.), *Computational Mathematics, Numerical Analysis and Applications*, SEMA SIMAI Springer Series 13,
DOI 10.1007/978-3-319-49631-3_7

in the state equation but it has to be retained in the momentum equation (see [2] for further details). We also consider the conservative law of transport of species and the $k-\varepsilon$ standard model (see [3]). Then, the system of equations to be solved reads

$$\frac{\partial \mathbf{w_u}}{\partial t} + \mathrm{div}(\mathscr{F}^{\mathbf{w_u}}(\mathbf{w_u}, \rho)) + \nabla \pi - \mathrm{div}(\tau) = \mathbf{f_u}, \quad (1)$$

$$\mathrm{div}\mathbf{w_u} = q, \quad q := -\frac{\partial}{\partial t}\left(\frac{\overline{\pi}}{R\theta}\right), \quad (2)$$

$$\tau = (\mu + \mu_t)\nabla \mathbf{u} - \frac{2}{3} w_k \mathbf{I}, \quad \mu_t = \rho C_\mu \frac{w_k^2}{w_\varepsilon}, \quad (3)$$

$$\frac{\partial \mathbf{w_Y}}{\partial t} + \mathrm{div}\mathscr{F}^{\mathbf{w_Y}}(\mathbf{w}, \rho) - \mathrm{div}\left[\left(\rho \mathscr{D} + \frac{\mu_t}{Sc_t}\right)\nabla\left(\frac{1}{\rho}\mathbf{w_Y}\right)\right] = \mathbf{f_Y}, \quad (4)$$

$$\frac{\partial w_k}{\partial t} + \mathrm{div}\mathscr{F}^{w_k}(\mathbf{w}, \rho) - \mathrm{div}\left[\left(\mu + \frac{\mu_t}{\sigma_k}\right)\nabla\left(\frac{w_k}{\rho}\right)\right] + w_\varepsilon = G_k + f_k, \quad (5)$$

$$\frac{\partial w_\varepsilon}{\partial t} + \mathrm{div}\mathscr{F}^{w_\varepsilon}(\mathbf{w}, \rho) - \mathrm{div}\left[\left(\mu + \frac{\mu_t}{\sigma_\varepsilon}\right)\nabla\left(\frac{w_\varepsilon}{\rho}\right)\right] + C_{2\varepsilon}\frac{w_\varepsilon^2}{w_k} = C_{1\varepsilon}\frac{w_\varepsilon}{w_k}G_k + f_\varepsilon, \quad (6)$$

where ρ denotes the density and $\mathbf{w}$ is the conservative variables vector with $\mathbf{w_u} = \rho u$ and u being the velocity. The flux tensor is given by $\mathscr{F}(\mathbf{w}, \rho) = u_i \mathbf{w}$, $\mathscr{F} = (\mathscr{F}^{\mathbf{w_u}}, \mathscr{F}^{\mathbf{w_Y}}, \mathscr{F}^{w_k}, \mathscr{F}^{w_\varepsilon})^T$. The Cauchy stress tensor is denoted by τ and $\mathbf{f} = (\mathbf{f}^{\mathbf{w_u}}, \mathbf{f}^{\mathbf{w_Y}}, f^{w_k}, f^{w_\varepsilon})^T$ is a generic source term. In the equation of state, $R = \mathscr{R}/\mathscr{M}$ denotes the gas constant, where $\mathscr{R}$ is the universal constant ($\mathscr{R} = 8314\,\mathrm{J/(kmol\ K)}$), $\mathscr{M}$ is the molecular mass and θ is the absolute temperature which is supposed to be given. The remaining conservative variables are: $\mathbf{w_Y} = \rho \mathbf{Y}$ ($\mathbf{Y}$ mass fraction vector), $w_k = \rho k$ (k viscosity dissipation rate), $w_\varepsilon = \rho\varepsilon$ (ε turbulent kinetic energy). Finally, G_k represents the production of turbulent kinetic energy and $C_{1\varepsilon}$, $C_{1\varepsilon}$ and C_μ are the closure coefficients of the turbulence model.

2 Numerical Discretization

The developed numerical method solves, at each time step, Eqs. (1) and (4)–(6) with a finite volume method and, so, an approximation of $\mathbf{w}$ is obtained. The next item is the projection step applied to system (2) that provides the pressure correction by a piecewise linear finite element method. Finally, an approximation of $\mathbf{w_u}$ verifying the divergence condition (2) is computed (see [2]).

Focusing on the finite volume method, the Local ADER scheme is developed to approximate the advection terms with second order accuracy in time and space (see [4]). Four main steps are considered:

Step 1. Data reconstruction. First-degree polynomial of each conservative variable, w, for a cell i are used, $p_i(N) = w_i + (N - N_i)\,(\nabla w_i)_N$.

Step 2. Computation of boundary extrapolated values at the barycenter of the faces Γ_{ij}, $w_{iN_{ij}} = p_i(N_{ij}) = w_i + (N_{ij} - N_i)\,(\nabla w_i)_{N_{ij}} = w_i + (N_{ij} - N_i)\nabla\left(\mathbf{W}^n_{|T_{ijL}}\right)$.

Step 3. Computation of the flux terms with second order of accuracy using the mid-point rule. Taylor series expansion in time and Cauchy - Kovalevskaya procedure are applied to locally approximate the conservative variables at time $\tau = \frac{\Delta t}{2}$, (see [5]). Two different options are considered:

OP1 Contribution of the advection term to the time evolution of the normal flux term, $\mathscr{Z}(\mathbf{W}, \eta) = \mathscr{F}(\mathbf{W}, rho)\,\eta$:

$$\overline{\mathbf{W}_{iN_{ij}}} = \mathbf{W}_{iN_{ij}} - \frac{\Delta t}{2\mathscr{L}_{ij}}\left(\mathscr{Z}(\mathbf{W}_{iN_{ij}}, \boldsymbol{\eta}_{ij}) + \mathscr{Z}(\mathbf{W}_{jN_{ij}}, \boldsymbol{\eta}_{ij})\right).$$

OP2 Contribution of the advection and diffusion terms to the time evolution of the flux term:

$$\overline{\mathbf{W}_{iN_{ij}}} = \mathbf{W}_{iN_{ij}} - \frac{\Delta t}{2\mathscr{L}_{ij}}\left(\mathscr{Z}(\mathbf{W}_{iN_{ij}}, \boldsymbol{\eta}_{ij}) + \mathscr{Z}(\mathbf{W}_{jN_{ij}}, \boldsymbol{\eta}_{ij})\right) + \frac{\Delta t}{2\mathscr{L}^2_{ij}}\left(\alpha_{iN_{ij}}\nabla\mathbf{W}^n_{|T_{ijL}}\boldsymbol{\eta}_{ij} + \alpha_{jN_{ij}}\nabla\mathbf{W}^n_{|T_{ijR}}\boldsymbol{\eta}_{ij}\right).$$

We have denoted $\mathscr{L}_{ij} = \min\left\{\frac{\mathrm{vol}(C_i)}{\mathrm{S}(C_i)}, \frac{\mathrm{vol}(C_j)}{\mathrm{S}(C_j)}\right\}$ and α_i the diffusion coefficient of each conservative variable W.

3 Numerical Results

In this section, we present the results obtained for two test problems. To check the order of the error, we compute the norms l^2 in time and L^2 in space. Δt is computed at each time step from a fixed CFL number.

Table 1 Gaussian bells test

Mesh	Elements	Vertex	Nodes	h
M_1	11, 664	2527	24, 408	0.1
M_2	18, 522	3872	38, 514	0.0857
M_3	54, 000	10, 571	111, 000	0.06
M_4	93, 312	17, 797	190, 944	0.05
M_5	182, 250	33, 856	256, 711	0.04

Mesh features

Table 2 Gaussian bells test

Variable	E_{M_1}	E_{M_2}	E_{M_3}	E_{M_4}	E_{M_5}
π	2.59E−04	1.67E−04	5.89E−05	3.47E−05	1.83E−05
$\mathbf{w_u}$	6.75E−04	4.61E−04	1.87E−04	1.18E−04	6.69E−05
w_y	1.61E−03	1.16E−03	5.56E−04	3.85E−04	2.48E−04

Variable	o_{M_1/M_2}	o_{M_2/M_3}	o_{v_3/MC_4}	o_{M_4/M_5}
π	2.87	2.92	2.90	2.87
$\mathbf{w_u}$	2.47	2.54	2.53	2.53
w_y	2.12	2.07	2.01	1.97

Observed errors and convergence rates. Flux terms were computed neglecting the diffusion terms contribution (OP1). $\mathrm{CFL}_y = 0.1$

3.1 Gaussian Bells

The first test problem studied is the Gaussian Bell Problem discussed, for instance, in [1]. The analytical expression of its solution reads

$$\pi(x,y,z,t) = 1, \quad \mathbf{u}(x,y,z,t) = (-y,\ x,\ 0)^T, \quad y(x,y,z,t) = \left(\frac{\sigma_0}{\sigma}\right)^3 \exp\left(\frac{-r}{2\sigma^2}\right),$$

$$r(x,y,z,t) = ((x\cos(t) + y\sin(t)) + 0.25)^2 + (-x\sin(t) + y\cos(t))^2 + z^2,$$

$$\rho = 1, \quad \sigma(t) = \sqrt{\sigma_0^2 + 2t\mathscr{D}}, \quad \sigma_0 = 0.08, \quad \mu = 0.01, \quad \mathscr{D} = 0.01.$$

We define the computational domain $\Omega = [-0.9, 0.9] \times [-0.9, 0.9] \times [-0.3, 0.3]$ and consider the mesh depicted in Table 1 where h denotes the size of the cubes used to generate the tetrahedra of the finite element mesh. After a complete rotation of the bell, second order of accuracy is attained (see Table 2).

3.2 MMS Test

The second test was defined using the method of manufactured solutions (MMS). We consider the computational domain $\Omega = [0, 1]^3$ and we assume the flow being

Table 3 MMS test

Mesh	N	Elements	Vertices	Nodes	v_h^m (m^3)	v_h^M (m^3)
M_1	4	384	125	864	6.51E−04	1.30E−03
M_2	8	3072	729	6528	8.14E−05	1.63E−04
M_3	16	24,576	4913	50,688	1.02E−05	2.03E−05

Mesh features

Table 4 MMS test

Variable	E_{M_1}	E_{M_2}	E_{M_3}	o_{M_1/M_2}	o_{M_2/M_3}
π	8.83E−02	1.94E−02	5.05E−03	2.18	1.94
$\mathbf{w_u}$	8.37E−03	2.30E−03	6.17E−04	1.86	1.90
w_y	6.19E−03	1.44E−03	3.52E−04	2.10	2.03
w_k	7.91E−03	1.79E−03	4.32E−04	2.14	2.05
w_ε	5.27E−03	1.13E−03	2.55E−04	2.22	2.15

Observed errors and convergence rates for the meshes introduced in Table 3. Flux terms were computed accounting for the diffusion terms contribution (OP2). $\mathrm{CFL}_y = 10$

defined by

$$\mu = 1.e-2, \quad \mathscr{D} = 1.e-3, \quad \rho(x,y,z,t) = 1,$$

$$\pi(x,y,z,t) = \cos(\pi t(x+y+z)),\ \mathbf{u}(x,y,z,t) = (\sin(\pi yt),\ -\cos(\pi zt),\ \exp(-\pi xt))^T,$$

$$y(x,y,z,t) = \sin(\pi xt) + 2,\ k(x,y,z,t) = \sin(\pi xt) + 2,\ \varepsilon(x,y,z,t) = \exp(-\pi zt) + 1.$$

The needed source terms corresponding to the former manufactured solution are obtained using symbolic calculus. In order to analyse the accuracy in time and space, three uniform meshes with different cell sizes are used. We consider the computational domain $\Omega = [0,1]^3$.The properties of these meshes can be seen in Table 3 where $N+1$ is the number of points along the edges of the domain and $h = 1/N$. v_h^m and v_h^M denote the minimum and maximum volume of the finite volumes, respectively. The numerical results presented in Table 4 confirm second order of accuracy.

Acknowledgements The authors are indebted to M. Dumbser, from the Laboratory of Applied Mathematics, University of Trento, for the useful discussions on the subject.

This project was partially supported by Spanish MECD under grant FPU13/00279; by Consellería de Cultura Educación e Ordenación Universitaria, Xunta de Galicia, under grant PRE/2013/031; by Spanish MICINN project MTM2013-43745-R; by Xunta de Galicia and FEDER under project GRC2013-014 and by Fundación Barrié under grant *Becas de posgrado en el extranjero*.

References

1. Bermejo, R., Saavedra, L.: Modified lagrange-galerkin methods of first and second order in time for convection-diffusion problems. Numer. Math. **120**, 601–638 (2012)
2. Bermúdez, A., Ferrín, J.L., Saavedra, L., Vázquez-Cendón, M.E.: A projection hybrid finite volume/element method for low-Mach number flows. J. Comput. Phys. **271**, 360–378 (2014)
3. Bermúdez, A., Busto, S., Cobas, M., Ferrín, J., Saavedra, L., Vázquez-Cendón, M.E.: Paths from mathematical problem to technology transfer related with finite volume methods. In: Díaz Moreno, J.M., Medina Moreno, J., Ortegón, F., Pérez Martínez, C., Redondo, M.V., Díaz Moreno, J.C., García Vázquez, C., Rodríguez Galván, J.R. (eds.) Proceedings of the XXIV Congress on Differential Equations and Aplications/XIV Congress on Applied Mathematics (2015), pp. 43–54
4. Busto, S., Toro, E.F., Vázquez-Cendón, M.E.: Design and analysis of ADER-type schemes for model advection-diffusion-reaction equations. J. Comput. Phys. **327**, 553–575 (2016)
5. Toro, E.F.: Riemann Solvers and Numerical Methods for Fluid Dynamics: A Practical Introduction. Springer, Berlin (2009)

Stable Discontinuous Galerkin Approximations for the Hydrostatic Stokes Equations

F. Guillén-González, M.V. Redondo-Neble, and J.R. Rodríguez-Galván

Abstract We propose a Discontinuous Galerkin scheme for the numerical solution of the Anisotropic (in particular, Hydrostatic) Stokes equations in Oceanography. The key is the introduction of interior penalties into the usual Stokes bilinear forms and, moreover, in the anisotropy (with respect to the horizontal and vertical directions) of these forms. Using $\mathscr{P}_k$ discontinuous finite elements for velocity and pressure, we obtain discrete inf-sup stability independently on the ratio ε between the horizontal and vertical domain scales. Numerical tests are provided.

1 Anisotropic (Hydrostatic) Stokes Equations in Oceanography

Anisotropic Stokes (and Hydrostatic) equations are the centerpiece for more complex models in Oceanography in large scale domains, where $\varepsilon =$ (vertical scale) / (horizontal scale) is very small:

$$\begin{cases} -\nu\Delta\mathbf{u} + \nabla_{\mathbf{x}} p = \mathbf{f}, & \text{in } \Omega, \\ -\varepsilon^2\Delta v + \partial_z p = g, & \text{in } \Omega, \\ \nabla_{\mathbf{x}} \cdot \mathbf{u} + \partial_z v = 0, & \text{in } \Omega. \end{cases} \tag{AnisStokes}$$

In particular, the limit case, $\varepsilon = 0$ (the hydrostatic Stokes problem) gives rise to the well known *Primitive Equations of the ocean*. The approximation of (AnisStokes) has been studied by the authors in [3–6]. In the three later ones, two underlying inf-sup constraints for (AnisStokes) were shown:

F. Guillén-González
Departamento EDAN and IMUS, Universidad de Sevilla, Sevilla, Spain
e-mail: guillen@us.es

M.V. Redondo-Neble • J.R. Rodríguez-Galván (✉)
Departamento de Matemáticas, Universidad de Cádiz, Cádiz, Spain
e-mail: victoria.redondo@uca.es; rafael.rodriguez@uca.es

M. Mateos, P. Alonso (eds.), *Computational Mathematics, Numerical Analysis and Applications*, SEMA SIMAI Springer Series 13,
DOI 10.1007/978-3-319-49631-3_8

1. The well-know LBB-like inf-sup condition for Stokes
2. A new Hydrostatic inf-sup constraint, related to the vertical velocity of (Anis-Stokes):

$$\sup_{0\neq p\in P}\frac{\int_\Omega p\partial_z v}{\|p\|_{L^2(\Omega)}}\geq \beta_v\|\partial_z v\|_{L^2(\Omega)},\ \forall\, v\in H^1_z(\Omega). \tag{IS}^V$$

In order to avoid the last constraint, making possible the use of standard finite elements for (AnisStokes), different techniques are proposed in [4–6]. Here we show a different technique, based in DG methods.

2 Discontinuous Galerkin Approximation

We consider Interior Penalty (IP) methods that, for second order elliptic equations, were introduced in [1] and for the isotropic Stokes equations (i.e. for $\varepsilon = 1$) have been recently studied by different authors (see e.g. [2] and references therein).

Specifically, we consider $P_k(\mathscr{T}_h)$ discontinuous spaces $\mathbf{U}_h$, V_h and P_h, for approximation of the velocity field $\mathbf{w}_h = (\mathbf{u}_h, v_h) \in \mathbf{U}_h \times V_h$ and the pressure $p_h \in P_h$. The key for the well posedness of the discrete problem is in the introduction of the anisotropic bilinear form (depending on a penalty parameter $\mu > 0$):

$$a_h(\mathbf{w}_h,\overline{\mathbf{w}}_h) = \nu\sum_{i=1}^{d-1} a^{\mathrm{sip},\mu}(u_{h,i},\overline{u}_{h,i}) + \varepsilon^2\, a^{\mathrm{sip},\mu/\varepsilon^2}(v,\overline{v}_h),$$

where $a^{\mathrm{sip},\mu}$ is the well-known symmetric IP bilinear form,

$$a^{\mathrm{sip},\mu}(v,\overline{v}) = \int_\Omega \nabla_h v\cdot\nabla_h\overline{v} - \sum_{e\in\mathscr{E}_h}\int_e\Big(\{\!\{\nabla_h v\}\!\}\cdot n_e[\![\overline{v}]\!] + [\![v]\!]\{\!\{\nabla_h\overline{v}\}\!\}\cdot n_e\Big) + \mu\sum_{e\in\mathscr{E}_h}\frac{1}{h_e}\int_e [\![v]\!][\![\overline{v}]\!].$$

Here, ∇_h denotes the broken (defined by elements) divergence operator, $[\![\cdot]\!]$ and $\{\!\{\cdot\}\!\}$ are the jump and average operators on the edges (or faces), denoted $e \in \mathscr{E}_h$. We must also consider the anisotropic norm for velocity

$$\|\mathbf{w}_h\|_{\mathrm{vel}} = \Big(\|\mathbf{u}_h\|^2_{\mathrm{sip}} + \|\partial_{z,h}v_h\|^2_{L^2(\Omega)} + |v_h|^2_J\Big)^{1/2},$$

where $\|\cdot\|_{\text{sip}}$ is the norm associated to $a^{\text{sip},\mu}(\cdot,\cdot)$, $\partial_{z,h}$ is the broken partial derivative and $|\cdot|_J$ is the jump seminorm:

$$|v|_J = \Big(\sum_{e\in\mathscr{E}_h} h_e^{-1} \|[\![v]\!]\|_{L^2(e)}{}^2 \Big)^{1/2}.$$

Lemma 1 (Partial Coercivity for $a_h(\cdot,\cdot)$) *There exist $\overline{\mu}$ and $\alpha > 0$ (independent of h and ε), such that $\forall \mu > \overline{\mu}$,*

$$a_h(\mathbf{w}_h, \mathbf{w}_h) \geq \alpha \Big(\|\mathbf{u}_h\|_{\text{sip}}{}^2 + |v_h|_J{}^2 \Big).$$

Note that Lemma 1 does not provide control for $\|\partial_{z,h} v_h\|_{L^2(\Omega)}$. It shall be recovered in Theorem 3. Let us consider the standard IP velocity-pressure coupling

$$b_h(\mathbf{w}_h, p_h) = -\int_\Omega p_h \, \nabla_h \cdot \mathbf{w}_h + \sum_{e\in\mathscr{E}_h} \int_e [\![\mathbf{w}_h]\!] \cdot n_e \, \{\!\{ p_h \}\!\}$$

and pressure seminorm, defined in $H^1(\mathscr{T}_h) \supset P_k(\mathscr{T}_h)$:

$$|p|_P = \Big(\sum_{e\in\mathscr{E}_h^0} h_e \, \|[\![p]\!]\|^2_{L^2(e)} \Big)^{1/2}.$$

Lemma 2 (Stability for b_h) *There Exists $\beta > 0$ independent of h, such that*

$$\beta \, \|p_h\|_{L^2(\Omega)} \leq \sup_{\mathbf{w}_h\in\mathbf{W}_h\setminus\{0\}} \frac{b_h(\mathbf{w}_h, p_h)}{\|\mathbf{w}_h\|_{\text{vel}}} + |p_h|_P, \quad \forall p_h \in P_h.$$

We consider the discrete formulation for (AnisStokes): find $(\mathbf{w}_h, p_h) \in \mathbf{W}_h \times P_h$ such that, $\forall \overline{\mathbf{w}}_h \in \mathbf{W}_h$, $\overline{p}_h \in P_h$,

$$\begin{cases} a_h(\mathbf{w}_h, \overline{\mathbf{w}}_h) + b_h(\overline{\mathbf{w}}_h, p_h) = \int_\Omega f \, \overline{\mathbf{w}}_h + \int_{\Gamma_s} \mathbf{g}_s \, \overline{\mathbf{w}}_h, \\ -b_h(\mathbf{w}_h, \overline{p}_h) + s_h(p_h, \overline{p}_h) = 0, \end{cases} \tag{P}$$

where $s_h(q_h, r_h) = \sum_{e\in\mathscr{E}_h^0} h_e \int_e [\![q_h]\!] \, [\![r_h]\!]$. Let $c((\mathbf{w}_h, p_h), (\overline{\mathbf{w}}_h, \overline{p}_h))$ be the corresponding mixed bilinear form and let

$$\|(\mathbf{w}_h, p_h)\|_{\mathbf{X}_h} = \Big(\|\mathbf{w}_h\|_{\text{vel}}{}^2 + \|p_h\|^2_{L^2} + |p_h|^2_p \Big)^{1/2}.$$

The following result implies well posedness of former discrete problem:

Theorem 3 (Discrete Inf-Sup Stability) *There exists* $\overline{\mu} > 0$ *and* $\gamma > 0$ *(independent of* h *and* ε*) such that, for every* $\mu > \overline{\mu}$*, one has for all* $(\mathbf{w}_h, p_h) \in \mathbf{X}_h = \mathbf{U}_h \times V_h \times P_h$*:*

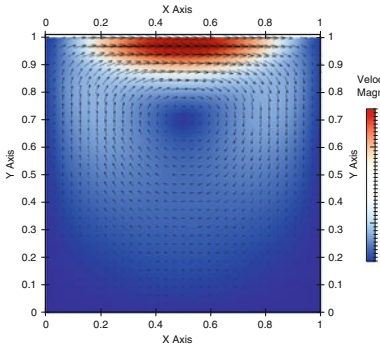
$$\gamma \, \|(\mathbf{w}_h, p_h)\|_{\mathbf{X}_h} \leq \sup_{(\overline{\mathbf{w}}_h, \overline{p}_h) \in \mathbf{X}_h \setminus \{0\}} \frac{c_h((\mathbf{w}_h, p_h), (\overline{\mathbf{w}}_h, \overline{p}_h))}{\|(\overline{\mathbf{w}}_h, \overline{p}_h)\|_{\mathbf{X}_h}}.$$

Proof (Idea) Lemma 1 provides control of $\|\mathbf{u}_h\|_{\text{sip}}{}^2$ and $|v_h|_{\text{J}}{}^2$. Lemma 2 provides control of $\|p_h\|_{L^2(\Omega)}$. Discrete inf-sup condition $(IS)^V$ is satisfied for P_k elements, hence one has control of $\|\partial_{z,h} v_h\|_{L^2(\Omega)}$.

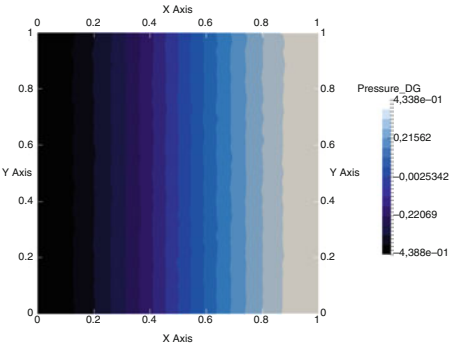
3 Numerical Tests

A cavity test for the discrete problem (P) was programmed in FreeFem++ [7], using the following data: Physical domain V/H ratio: $\varepsilon = 10^{-7}$. Viscosity: $\nu = 1$. RHS functions: $\mathbf{f} = 0$, $g = 0$. Adimensional domain $\Omega = [0,1]^2$, structured 32×32 mesh ($h \sim 10^{-2}$). P_1 elements for velocity & pressure.

Dirichlet boundary conditions (B.C.): Let Γ_S = surface boundary, we take: $u = x(x-1)$ on Γ_S, $\quad u = 0$ on $\partial\Omega \setminus \Gamma_S$, $\quad v = 0$ on $\partial\Omega$. B.C. are imposed weakly, using the Nitsche method. To determinate adequate IP and B.C. (Nitsche) penalty parameters was not easy. We set: IP Penalty parameter $\mu = 10^2$, B.C. Penalty $\eta = 10^2$.



Velocity field



Pressure iso-values

Future Research Convergence shall be studied in a future work. In the Stokes case ($\varepsilon = 1$), optimal estimates in energy norm can be obtained for smooth exact solutions (see e.g. [2]). We conjecture that they can be extended to the Hydrostatic case. Results for non-smooth solutions are more difficult.

Acknowledgements First author was partially financed by MINECO grants MTM2015-69875-P (Spain) with the participation of FEDER. Second and third ones are partially supported by the research group FQM-315 of Junta de Andalucía (Spain).

References

1. Arnold, D.N.: An interior penalty finite element method with discontinuous elements. SIAM J. Numer. Anal. **19**(4), 742–760 (1982)
2. Di Pietro, D.A., Ern, A.: Mathematical aspects of discontinuous Galerkin methods. Springer, Berlin/New York (2012)
3. Guillén-González, F., Redondo-Neble, M.V.: Convergence and error estimates of a viscosity-splitting finite-element schemes for the primitive equations. Appl. Numer. Math. **111**, 219–245 (2017). http://dx.doi.org/10.1016/j.apnum.2016.09.011
4. Guillén-González, F., Rodríguez-Galván, J.R.: Analysis of the hydrostatic Stokes problem and finite-element approximation in unstructured meshes. Numer. Math. **130**(2), 225–256 (2015)
5. Guillén-González, F., Rodríguez Galván, J.R.: Stabilized schemes for the Hydrostatic stokes equations. SIAM J. Numer. Anal. **53**(4), 1876-1896 (2015)
6. Guillén-González, F., Rodríguez Galván, J.R.: On the stability of approximations for the Stokes problem using different finite element spaces for each component of the velocity. Appl. Numer. Math. **99**, 51–76 (2016)
7. Hecht, F.: New development in FreeFem++. J. Numer. Math. **20**(3–4), 251–265 (2012)

A Two-Step Model Identification for Stirred Tank Reactors: Incremental and Integral Methods

A. Bermúdez, E. Carrizosa, Ó. Crego, N. Esteban, and J.F. Rodríguez-Calo

Abstract In this work we present a new methodology for solving an inverse identification problem with application in chemistry, using two approaches in cascade. More precisely, we are interested in the identification of kinetic models and their corresponding parameters in stirred tank reactors, using a set of experimental data and the reactions taking place. A catalogue of kinetic models containing the parameters to be identified will be provided too. In order to solve it, we use a combination of an incremental and an integral method.

1 Introduction

Nowadays the study of chemical process in industry makes extensive use of mathematical modelling. Building models needs the identification of the reactions taking place and their kinetics. The latter represents a challenging task in the reactor description.

A. Bermúdez • Ó. Crego • N. Esteban (✉)
Departamento de Matemática Aplicada, Universidad de Santiago de Compostela, Campus Vida, C/ Lope Gómez de Marzoa S/N, 15786 Santiago de Compostela, Spain
e-mail: alfredo.bermudez@usc.es; oscar.crego@usc.es; noemi.esteban@usc.es

E. Carrizosa
Departamento de Estadística e Investigación Operativa, Universidad de Sevilla, C/ Tarfia S/N, 41012 Sevilla, Spain
e-mail: ecarrizosa@us.es

J.F. Rodríguez-Calo
Centro de Tecnología, Autovía de Extremadura, S/N 28935 Móstoles, Madrid, Spain
e-mail: jfrodriguezc@repsol.com

M. Mateos, P. Alonso (eds.), *Computational Mathematics, Numerical Analysis and Applications*, SEMA SIMAI Springer Series 13,
DOI 10.1007/978-3-319-49631-3_9

2 The General Model

An important family of chemical reactors is the so-called stirred tank reactors (STR). We assume that the mixture inside these reactors is homogeneous because of stirring so the physico-chemical magnitudes do not depend on position. Then they are modelled as (usually stiff) ordinary differential equations which are non-linear and coupled.

We consider a model with mole balance and heat balance equations. In addition, we have an equation for volume variation. The entire model can be written as

$$\begin{cases} \dfrac{d\mathbf{y}}{dt} = \mathbf{f}(\theta, \mathbf{y}, \mathbf{z}) \text{ in } [0,T], & \text{mole balance system} \\ \dfrac{d\theta}{dt} = h(\theta, \mathbf{y}, \mathbf{z}), & \text{heat balance equation} \\ \dfrac{dV}{dt} = \mathrm{f}^2 - \mathrm{f}^3, & \text{volume equation} \\ \mathbf{y}(0) = \mathbf{y}_0, \theta(0) = \theta_0 \text{ and } V(0) = V_0, \end{cases} \tag{1}$$

with source terms

$$\mathbf{f} = A\boldsymbol{\delta}(\theta, \mathbf{y}, \mathbf{z}) + \frac{1}{V}(\mathrm{F}\mathbf{f}^1 - \mathrm{f}^2\mathbf{y}),$$

$$h = \frac{\boldsymbol{\Delta}\mathbf{H}(\theta)\cdot\boldsymbol{\delta}(\theta, \mathbf{y}, \mathbf{z}) - \frac{g}{V}(\theta_{out} - \theta) - \mathbf{w}'(\theta)\cdot\left(\mathrm{F}\sum_{p=1}^{P}\mathrm{f}_p^1(\theta_p^s - \theta)\mathbf{e}_p\right)}{\mathbf{w}'(\theta)\cdot\mathbf{y}}.$$

$$\boldsymbol{\Delta}\mathbf{H}(\theta) = A^t\mathbf{w}(\theta),\ w_i(\theta) = \mathscr{M}_i e_i \text{ and } e_i(\theta) = e_i^* + \int_{\theta^*}^{\theta} c_i(s)ds \text{ fot the } i\text{th species.}$$

$\mathbf{y}$ represents the vector of species concentrations.
θ represents the vector of catalysts.
z represents the temperature of mixture.
V represents the volume of mixture.
A is the stoichiometric matrix.
$\boldsymbol{\delta}$ represents the vector of reaction velocities.
F represents the inlet composition.
$\mathbf{f}_1$ is the vector of inlet flow rates.
f_2 is the sum of components in $\mathbf{f}_1$.
f_3 is the outlet flow rate.

$\boldsymbol{\Delta H}$ is the vector with heat of the reactions.
g is the heat transfer coefficient.
θ_{out} is the outside temperature.
θ_p^s is the temperature of the pth stream where P is the number of streams.
$\mathscr{M}_i$ is the molecular mass of the ith species.
c_i is the specific heat of the ith species.
e_i is the internal energy of the ith species.
e_i^* is the internal energy of formation of the ith species at temperature θ^*.

In the most general case, we have inlet and outlet streams and inlet composition matrix of the mixture and the reactor is called continuous STR, when we only have inlet streams and inlet composition the reactor is called semi-batch STR, and when we have no any of these the reactor is called batch STR.

3 Methodology

For solving this problem, several techniques can be considered, such as differential, integral and incremental methods [2]. All these methods are based on the optimization of their corresponding functional costs which provide kinetics and their parameters. They can be used independently.

We have some controlled experiments and measurements of species at some time instants.

The differential method uses cubic spline functions interpolating the data and taking their derivatives at time measurements. The error in these derivatives may affect to the accuracy of the solution.

In some cases the solution given by the incremental method is good enough and we may conclude the identification process, but this is not always the case. The integral method can improve the initial solution and this is important when the experiments are affected by noise.

- *The incremental method. Initial approach.*

  The incremental method for STR is described in [4]. We introduce an alternative method where the heat balance equation is treated independently.

  The main features of this method are the decoupling of the reaction equations using algebraic procedures and obtaining the direct solution of the transformed equations. Thus, the kinetic models and their parameters can be identified in parallel for all reactions.

  Volume equation can be solved firstly and independently, but the ODEs system (1) is coupled. That is why we work in two stages: the concentrations system is rewritten as a decoupled extents system and the temperature equation is treated apart.

Finally, we minimize the following functional cost in terms of extents

$$J_{m,l}(\boldsymbol{\Theta}_l^m) = \sum_{e\in\mathscr{E}} \sum_{s\in S^e} |\hat{e}_{sl}^e - e_l^{(m)}(t_s^e, \boldsymbol{\Theta}_l^m)|^2, \forall m = 1,\ldots,M_l \text{ and } l = 1,\ldots,L.$$

where $\boldsymbol{\Theta}_l^m$ is the parameter vector, $e_l^{(m)}(t_s^e, \boldsymbol{\Theta}_l^m)$ and $\hat{e}_{sl}^e$ are the lth component of the extents model and of measurements respectively at time $t_s^e \in S^e$ and experiment $e \in \mathscr{E}$. M_l is the set of kinetics for the lth reaction.

- *The integral method. Improvements in the solution.*

  The integral method is based on a direct comparison of measurements and computed concentrations.

  The main difficulties lie in the huge number of parameters, solving numerically the model, and computing the derivatives with respect to these parameters. We propose an heuristic based on the variable neighbourhood search (VNS) [3]. This method uses as initial values of the parameters those previously computed by the incremental method. A new solution is generated by doing successive perturbations both in kinetics and in parameters.

  We use a finite differences scheme (BDF2 initializated with a BDF1) to solve the problem derived from the reactor model and the derivatives of functional cost are computed using the adjoint-state method [1]. The functional cost is the following:

$$J(\boldsymbol{\Theta}) := \sum_{e\in\mathscr{E}} \sum_{i\in\mathscr{S}} \sum_{s\in S^e} \omega_{ien}(y_i^e(t_s^e, \boldsymbol{\Theta}) - \hat{y}_{si}^e)^2,$$

  where $\boldsymbol{\Theta}$ is the parameter vector, $y_{si}^e(t_s^e, \boldsymbol{\Theta})$ and $\hat{y}_{si}^e$ are the ith component of the solution of the model and of measurements respectively at time $t_s^e \in S^e$ and experiment $e \in \mathscr{E}$.

4 Example

We consider an academic example that represents a batch type reactor with known temperature. The reaction system is described by 12 species, involved in 6 reactions and 1 catalysts with constant concentration.

$$A+B \to C+D,\ B \to C+0.5E,\ B \to F+G,\ B \to H+I,\ B \to J+K,\ D+G \to L$$

We have 10 experiments with different initial conditions and time measurements from 0 to 100 s at each 10 s. The list of kinetics for each reaction is:

Reaction 1

$$\delta_1^{(1)}(\theta,\mathbf{y},\mathbf{z}) = B_1 e^{\left(-\frac{Ea_1}{R\theta}\right)} y_1^{\alpha_1} y_2^{\alpha_2} z_1^{\alpha_3},$$
$$\delta_1^{(2)}(\theta,\mathbf{y},\mathbf{z}) = B_1 e^{\left(-\frac{Ea_1}{R\theta}\right)} y_1^{\alpha_1} z_1^{\alpha_2},$$
$$\delta_1^{(3)}(\theta,\mathbf{y},\mathbf{z}) = B_1 e^{\left(-\frac{Ea_1}{R\theta}\right)} y_2^{\alpha_1} z_1^{\alpha_2},$$
$$\delta_1^{(4)}(\theta,\mathbf{y},\mathbf{z}) = B_1 e^{\left(-\frac{Ea_1}{R\theta}\right)} y_1^{\alpha_1} y_2^{\alpha_2},$$
$$\delta_1^{(5)}(\theta,\mathbf{y},\mathbf{z}) = B_1 e^{\left(-\frac{Ea_1}{R\theta}\right)} y_1^{\alpha_1},$$
$$\delta_1^{(6)}(\theta,\mathbf{y},\mathbf{z}) = B_1 e^{\left(-\frac{Ea_1}{R\theta}\right)} y_2^{\alpha_1},$$
$$\delta_1^{(7)}(\theta,\mathbf{y},\mathbf{z}) = B_1 e^{\left(-\frac{Ea_1}{R\theta}\right)} z_1^{\alpha_1},$$
$$\delta_1^{(8)}(\theta,\mathbf{y},\mathbf{z}) = B_1 e^{\left(-\frac{Ea_1}{R\theta}\right)} y_1^{\alpha_1^{int}} y_2^{\alpha_2^{int}} z_1^{\alpha_3^{int}}.$$

Reaction 2

$$\delta_2^{(1)}(\theta,\mathbf{y},\mathbf{z}) = B_2 e^{\left(-\frac{Ea_2}{R\theta}\right)} y_2^{\alpha_1} z_1^{\alpha_2},$$
$$\delta_2^{(2)}(\theta,\mathbf{y},\mathbf{z}) = B_2 e^{\left(-\frac{Ea_2}{R\theta}\right)} y_2^{\alpha_1},$$
$$\delta_2^{(3)}(\theta,\mathbf{y},\mathbf{z}) = B_2 e^{\left(-\frac{Ea_2}{R\theta}\right)} z_1^{\alpha_1},$$
$$\delta_2^{(4)}(\theta,\mathbf{y},\mathbf{z}) = B_2 e^{\left(-\frac{Ea_2}{R\theta}\right)} y_2^{\alpha_1^{int}} z_1^{\alpha_2}.$$

Reaction 3

$$\delta_3^{(1)}(\theta,\mathbf{y},\mathbf{z}) = B_3 e^{\left(-\frac{Ea_3}{R\theta}\right)} y_2^{\alpha_1} z_1^{\alpha_2},$$
$$\delta_3^{(2)}(\theta,\mathbf{y},\mathbf{z}) = B_3 e^{\left(-\frac{Ea_3}{R\theta}\right)} y_2^{\alpha_1},$$
$$\delta_3^{(3)}(\theta,\mathbf{y},\mathbf{z}) = B_3 e^{\left(-\frac{Ea_3}{R\theta}\right)} z_1^{\alpha_1},$$
$$\delta_3^{(4)}(\theta,\mathbf{y},\mathbf{z}) = B_3 e^{\left(-\frac{Ea_3}{R\theta}\right)} y_2^{\alpha_1^{int}} z_1^{\alpha_2},$$
$$\delta_3^{(5)}(\theta,\mathbf{y},\mathbf{z}) = B_3 e^{\left(-\frac{Ea_3}{R\theta}\right)} y_2^{\alpha_1} z_1^{\alpha_2^{int}},$$
$$\delta_3^{(6)}(\theta,\mathbf{y},\mathbf{z}) = B_3 e^{\left(-\frac{Ea_3}{R\theta}\right)} y_2^{\alpha_1^{int}} z_1^{\alpha_2^{int}}.$$

Reaction 4

$$\delta_4^{(1)}(\theta,\mathbf{y},\mathbf{z}) = B_4 e^{\left(-\frac{Ea_4}{R\theta}\right)} y_2^{\alpha_1} z_1^{\alpha_2},$$
$$\delta_4^{(2)}(\theta,\mathbf{y},\mathbf{z}) = B_4 e^{\left(-\frac{Ea_4}{R\theta}\right)} y_2^{\alpha_1},$$
$$\delta_4^{(3)}(\theta,\mathbf{y},\mathbf{z}) = B_4 e^{\left(-\frac{Ea_4}{R\theta}\right)} z_1^{\alpha_1},$$
$$\delta_4^{(4)}(\theta,\mathbf{y},\mathbf{z}) = B_4 e^{\left(-\frac{Ea_4}{R\theta}\right)} y_2^{\alpha_1^{int}} z_1^{\alpha_2}.$$

Reaction 5

$$\delta_5^{(1)}(\theta,\mathbf{y},\mathbf{z}) = B_5 e^{\left(-\frac{Ea_5}{R\theta}\right)} y_2^{\alpha_1} z_1^{\alpha_2},$$
$$\delta_5^{(2)}(\theta,\mathbf{y},\mathbf{z}) = B_5 e^{\left(-\frac{Ea_5}{R\theta}\right)} y_2^{\alpha_1},$$
$$\delta_5^{(3)}(\theta,\mathbf{y},\mathbf{z}) = B_5 e^{\left(-\frac{Ea_5}{R\theta}\right)} z_1^{\alpha_1},$$
$$\delta_5^{(4)}(\theta,\mathbf{y},\mathbf{z}) = B_5 e^{\left(-\frac{Ea_5}{R\theta}\right)} y_2^{\alpha_1^{int}} z_1^{\alpha_2}.$$

Reaction 6

$$\delta_6^{(1)}(\theta,\mathbf{y},\mathbf{z}) = B_6 e^{\left(-\frac{Ea_6}{R\theta}\right)} y_4^{\alpha_1} y_7^{\alpha_2} z_1^{\alpha_3},$$
$$\delta_6^{(2)}(\theta,\mathbf{y},\mathbf{z}) = B_6 e^{\left(-\frac{Ea_6}{R\theta}\right)} y_4^{\alpha_1} y_7^{\alpha_2},$$
$$\delta_6^{(3)}(\theta,\mathbf{y},\mathbf{z}) = B_6 e^{\left(-\frac{Ea_6}{R\theta}\right)} y_7^{\alpha_1} z_1^{\alpha_2},$$
$$\delta_6^{(4)}(\theta,\mathbf{y},\mathbf{z}) = B_6 e^{\left(-\frac{Ea_6}{R\theta}\right)} y_4^{\alpha_1} z_1^{\alpha_2},$$
$$\delta_6^{(5)}(\theta,\mathbf{y},\mathbf{z}) = B_6 e^{\left(-\frac{Ea_6}{R\theta}\right)} y_4^{\alpha_1},$$
$$\delta_6^{(6)}(\theta,\mathbf{y},\mathbf{z}) = B_6 e^{\left(-\frac{Ea_6}{R\theta}\right)} y_7^{\alpha_1},$$
$$\delta_6^{(7)}(\theta,\mathbf{y},\mathbf{z}) = B_6 e^{\left(-\frac{Ea_6}{R\theta}\right)} z_1^{\alpha_1},$$
$$\delta_6^{(8)}(\theta,\mathbf{y},\mathbf{z}) = B_6 e^{\left(-\frac{Ea_6}{R\theta}\right)} y_4^{\alpha_1^{int}} y_7^{\alpha_2^{int}} z_1^{\alpha_3}.$$

R is the universal gas constant. $B \in [0, 10^{14}]$ and $Ea \in [0, 200000]$ represent the pre-exponential factor and the activation energy respectively in the Arrhenius law, and $\alpha_i \in [0, 2]$ $\forall i = 1, 2, 3$. The super index *int* in the exponents means that we do integer optimization on these parameters.

The incremental method select the following kinetics after computing the parameters of all the kinetics in the list in about 4732 s

$$\delta_1^{(4)}(\theta, \mathbf{y}, \mathbf{z}) = 1.76\ 10^5 e^{\left(-\frac{4.60\ 10^5}{R\theta}\right)} y_1^{0.92} y_2^{0.90},\ \text{with}\ J_{4,1}(\boldsymbol{\Theta}_1^4) = 5.48\ 10^{-2}$$

$$\delta_2^{(2)}(\theta, \mathbf{y}, \mathbf{z}) = 1.06\ 10^4 e^{\left(-\frac{3.80\ 10^4}{R\theta}\right)} y_2^{1.87},\ \text{with}\ J_{2,2}(\boldsymbol{\Theta}_2^2) = 5.48\ 10^{-2}$$

$$\delta_3^{(2)}(\theta, \mathbf{y}, \mathbf{z}) = 201.29 e^{\left(-\frac{3.80\ 10^4}{R\theta}\right)} y_2^{1.88},\ \text{with}\ J_{2,3}(\boldsymbol{\Theta}_3^2) = 5.48\ 10^{-2}$$

$$\delta_4^{(2)}(\theta, \mathbf{y}, \mathbf{z}) = 1.16\ 10^4 e^{\left(-\frac{1.14\ 10^5}{R\theta}\right)} y_2^{0.93},\ \text{with}\ J_{2,4}(\boldsymbol{\Theta}_4^2) = 5.48\ 10^{-2}$$

$$\delta_5^{(1)}(\theta, \mathbf{y}, \mathbf{z}) = 1.62\ 10^8 e^{\left(-\frac{1.14\ 10^5}{R\theta}\right)} y_2^{0.95} z_1^{0.96},\ \text{with}\ J_{1,5}(\boldsymbol{\Theta}_5^1) = 5.48\ 10^{-2}$$

$$\delta_6^{(2)}(\theta, \mathbf{y}, \mathbf{z}) = 2.27\ 10^6 e^{\left(-\frac{6.86\ 10^4}{R\theta}\right)} y_4^{0.98} y_7^{0.97},\ \text{with}\ J_{2,6}(\boldsymbol{\Theta}_6^2) = 5.48\ 10^{-2}.$$

The objective value of the integral method for these kinetics is 0.2184.

The integral method provides a value of the functional cost of 0.2071 after 11,755 s. The kinetics selected are the following:

$$\delta_1^{(8)} = 1.97\ 10^8 e^{\left(-\frac{4.60\ 10^4}{R\theta}\right)} y_1 y_2 z_1,$$

$$\delta_2^{(1)} = 1.05\ 10^7 e^{\left(-\frac{3.80\ 10^4}{R\theta}\right)} y_2^{1.87} z_1^{0.99},$$

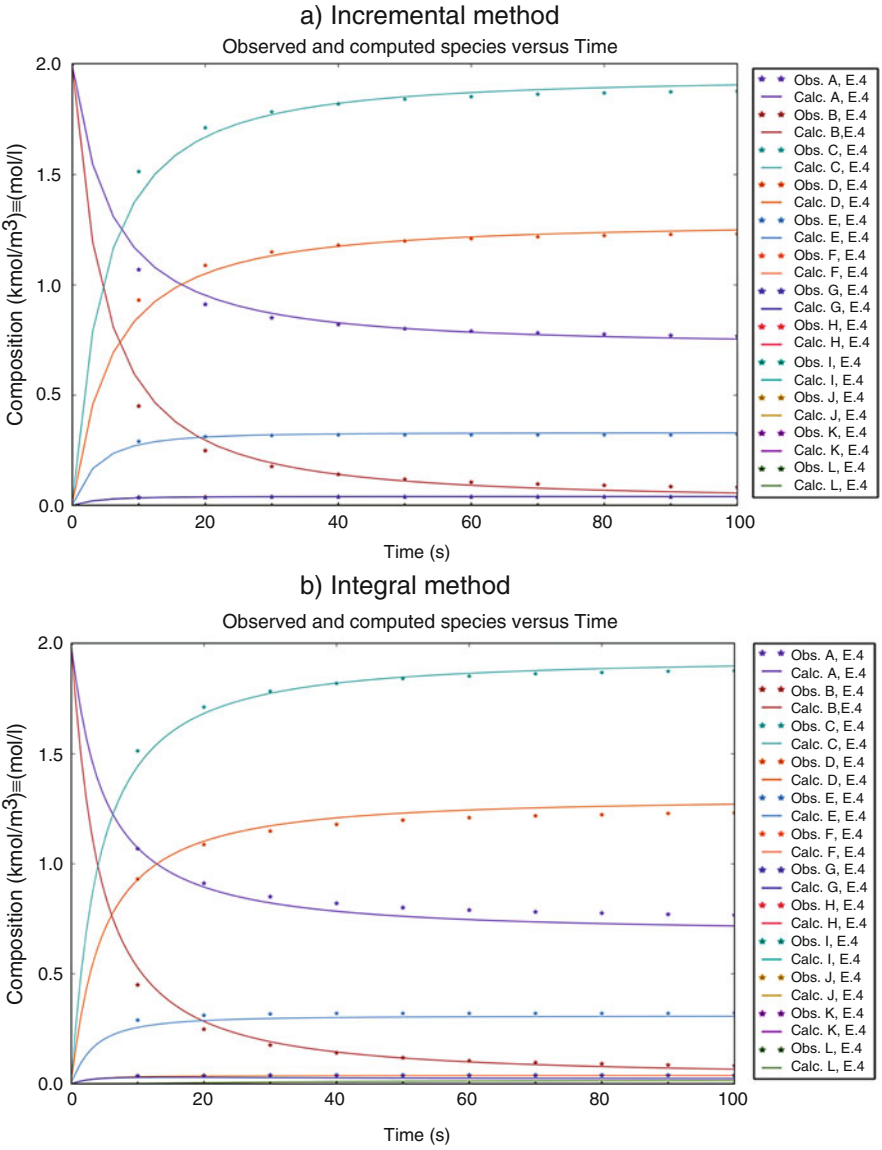
$$\delta_3^{(4)} = 6.46\ 10^4 e^{\left(-\frac{3.44\ 10^4}{R\theta}\right)} y_2^2 z_1^{0.84},$$

$$\delta_4^{(2)} = 1.16\ 10^4 e^{\left(-\frac{6.9710^4}{R\theta}\right)} y_2^{0.93},$$

$$\delta_5^{(4)} = 7.99\ 10^8 e^{\left(-\frac{1.18\ 10^5}{R\theta}\right)} y_2 z_1^{1.02},$$

$$\delta_6^{(8)} = 4.96\ 10^9 e^{\left(-\frac{6.91\ 10^4}{R\theta}\right)} y_4 z_1^{1.09}.$$

Now we can see the comparison between the data and the numerical solution of the model with the selected kinetics and their parameters in both incremental and integral methods in one of the experiments.



The incremental method provides us a good solution as it can be seen in figure a). However in this method experimental data are not compared directly because an algebraic transformation of the data is previously done.

Then, the integral method is used to correct these possible differences between data and numerical solution. The largest differences appear in species A, E and F. It is not recommendable to use only the integral method by itself because it is computationally expensive.

5 Conclusions

In conclusion, the described methods generate better results when used together. The incremental method gives good results when enough measurements and/or the experiments are not affected by noise are given. In other cases, incremental method generates an initial solution for the integral method which is essential in order to obtain a better adjustment. Moreover, this last method is computationally expensive and so, to improve this, an adjoint method is considered for computing functional cost derivatives and a VNS heuristic is too considered in the optimization process.

Acknowledgements Part of this research was developed as an activity in the Joint Research Unit Repsol-ITMATI (code file: IN853A 2014/03) which is funded by FEDER, the Galician Agency for Innovation (GAIN) and the Ministry of Economy and Competitiveness in the framework of the Spanish Strategy for Innovation in Galicia.

References

1. Benítez, M., Bermúdez, A., Rodrǵuez-Calo, J.F.: Adjoint method for inverse problems of chemical reaction systems. Chem. Eng. Res. Des. (2017). In press
2. Bhatt, N., et al.: Incremental identification of reaction systems-a comparison between rate-based and extent-based approaches. Chem. Eng. Sci. **84**, 24–38 (2012)
3. Mladenović, N., Hansen, P.: Variable neighborhood search. Comput. Oper. Res. **24**(11), 1097–1100 (1997)
4. Rodrigues, D., Srinivasan, S., Billeter, J., Bonvin, D.: Variant and invariant states for chemical reaction systems. Comput. Chem. Eng. **73**, 23–33 (2015)

Variance Reduction Result for a Projected Adaptive Biasing Force Method

Houssam AlRachid and Tony Lelièvre

Abstract This paper is committed to investigate an extension of the classical adaptive biasing force method, which is used to compute the free energy related to the Boltzmann-Gibbs measure and a reaction coordinate function. The issue of this technique is that the approximated gradient of the free energy, called biasing force, is not a gradient. The commitment to this field is to project the estimated biasing force on a gradient using the Helmholtz decomposition. The variance of the biasing force is reduced using this technique, which makes the algorithm more efficient than the standard ABF method. We prove exponential convergence to equilibrium of the estimated free energy, with a precise rate of convergence in function of Logarithmic Sobolev inequality constants.

1 Introduction

Let us consider the Boltzmann-Gibbs measure:

$$\mu(dx) = Z_\mu^{-1} \mathrm{e}^{-\beta V(x)} dx, \tag{1}$$

where $x \in \mathscr{D}^N$ denotes the position of N particles in $\mathscr{D} \subset \mathbb{R}^n$ (or the n-dimensional torus $\mathbb{T}^n$). The potential energy function $V : \mathscr{D} \longrightarrow \mathbb{R}$ associates with the positions of the particles $x \in \mathscr{D}$, Z_μ is the normalization constant and β is a constant proportional to the inverse of the temperature.

H. AlRachid (✉)
Université Paris-Est Créteil, 61 Avenue du Général de Gaulle, 94000 Créteil, France
e-mail: houssam.alrachid@u-pec.fr; alrachid.houssam@gmail.com

T. Lelièvre
École des Ponts ParisTech, Université Paris Est, 6-8 Avenue Blaise Pascal Cité Descartes, F-77455 Marne-la-Vallée, France
e-mail: tony.lelievre@enpc.fr

M. Mateos, P. Alonso (eds.), *Computational Mathematics, Numerical Analysis and Applications*, SEMA SIMAI Springer Series 13,
DOI 10.1007/978-3-319-49631-3_10

The equilibrium probability measure μ can be sampled through the Overdamped Langevin Dynamics:

$$dX_t = -\nabla V(X_t)dt + \sqrt{\frac{2}{\beta}}dW_t, \tag{2}$$

where $X_t \in \mathscr{D}^N$ and W_t is a Nn-dimensional standard Brownian motion. Under loose assumptions on V, the dynamics $(X_t)_{t\geq 0}$ is ergodic with respect to the equilibrium measure μ.

Because of the metastability, trajectorial averages converge very slowly to their ergodic limit. To overcome this difficulty, we focus in this paper on the Adaptive Biasing Force (denoted ABF) method (see [2, 3]). In order to introduce the ABF method, we require another ingredient: a reaction coordinate ξ describing the metastable zones of the dynamics associated with the potential energy V. For sake of simplicity, take $\xi : (x_1, \ldots, x_n) \in \mathbb{T}^n \mapsto (x_1, x_2) \in \mathbb{T}^2$. The associated free energy:

$$A(x_1, x_2) = -\beta^{-1} \ln(Z_{\Sigma_{(x_1,x_2)}}) = -\beta^{-1} \ln \int_{\mathbb{T}^{n-2}} e^{-\beta V(x)} dx_3 \ldots dx_n.$$

The idea of the ABF method is that, for a well chosen ξ, the dynamics associated with the potential $V - A \circ \xi$ is less metastable than the dynamics associated with V. The so called mean force $\nabla A(z)$, can be obtained as:

$$\nabla A(x_1, x_2) = Z^{-1}_{\Sigma_{(x_1,x_2)}} \int_{\mathbb{T}^{n-2}} f(x) \mathrm{e}^{-\beta V} dx_3 \ldots dx_n = \mathbb{E}_\mu(f(X)|\xi(X) = (x_1, x_2)),$$

where $f = (f_1, f_2) = (\partial_1 V, \partial_2 V)$. At time t, the mean force is approximated by $F^i_t(z) = \mathbb{E}_\mu[f_i(X_t)|\xi(X_t) = (x_1, x_2)]$, which also, under appropriate assumptions, converges exponentially fast to ∇A (see [4–6]). Despite the fact that F_t converges to a gradient, there is no reason why F_t would be a gradient at time t. In this paper, we propose an alternative method, where we approximate ∇A, at any time t, by a gradient denoted ∇A_t.

2 Projected Adaptive Biasing Force Method (PABF)

In this section, we present the PABF method, by reconstructing the mean force from the estimated one used in the ABF method.

In practice, A_t is obtained from F_t by solving the Poisson problem:

$$\operatorname{div}(\nabla A_t \psi^\xi(t,.)) = \operatorname{div}(F_t \psi^\xi(t,.)) \text{ on } \mathbb{T}^2, \tag{3}$$

where $\psi^\xi(t,\cdot)$ denotes the density of the random variables $\xi(X_t)$ which is the Euler equation associated to the minimization problem:

$$A_t = \underset{g \in H^1(\mathbb{T}^2)/\mathbb{R}}{\operatorname{argmin}} \int_{\mathbb{T}^2} |\nabla g - F_t|^2.$$

Solving (3) amounts to computing the so-called Helmholtz-Hodge decomposition of the vector field F_t as:

$$F_t \psi^\xi = \nabla A_t \psi^\xi + R_t \quad \text{on } \mathbb{T}^2,$$

with $\operatorname{div}(R_t) = 0$. In the following we denote by

$$\nabla A_t = \mathscr{P}_{\psi^\xi}(F_t), \text{ on } \mathbb{T}^2$$

the projection of F_t onto a gradient. We will study the longtime convergence of the following Projected adaptive biasing force (PABF) dynamics:

$$\begin{cases} dX_t = -\nabla\big(V - A_t \circ \xi\big)(X_t)dt + \sqrt{2\beta^{-1}}dW_t, \\ \nabla A_t = \mathscr{P}_{\psi^\xi}(F_t), \\ F_t^i(x_1,x_2) = \mathbb{E}[\partial_i V(X_t) | \xi(X_t) = (x_1,x_2)], \ i = 1,2, \end{cases} \tag{4}$$

Using entropy techniques, we study the longtime behavior of the nonlinear Fokker-Planck equation which rules the evolution of the density of X_t solution to (4). The following theorem shows exponential convergence to equilibrium of A_t to A, with a precise rate of convergence in terms of the Logarithmic Sobolev inequality constants of the conditional measures $d\mu_{\Sigma(x_1,x_2)} = Z^{-1}_{\Sigma_{(x_1,x_2)}} \mathrm{e}^{-\beta V} dx_3 \ldots dx_n$.

The assumptions we need to prove the longtime convergence of the biasing force ∇A_t to the mean force ∇A are the following:

H1 $V \in C^2(\mathbb{T}^n)$, $\exists \gamma > 0$, $\forall\, 3 \le j \le n, \forall x \in \mathbb{T}^n, \max(|\partial_1\partial_j V(x)|, |\partial_2\partial_j V(x)|) \le \gamma$.

H2 V is such that $\exists \rho > 0$, the conditional probability measures $\mu_{\Sigma(x_1,x_2)}$ satisfy a Logarithmic Sobolev inequality with constant ρ.

The proof of the following main theorem is provided in [1].

Theorem 1 *Let us assume **H1** and **H2**. The biasing force ∇A_t converges to the mean force ∇A in the following sense:*

$$\exists C > 0, \exists \lambda > 0, \forall\, t \geq 0, \int_{\mathbb{T}^2} |\nabla A_t - \nabla A|^2 \psi^{\xi}(t, x_1, x_2) dx_1 dx_2 \leq \frac{8C\gamma^2}{\rho} \mathrm{e}^{-\lambda t}.$$

Since, numerically, we use Monte-Carlo methods to approximate F_t and ∇A_t, the variance is an important quantity to assess the quality of the result. The following second main result is a variance reduction result and proved in [1].

Proposition 2 *For any time $t > 0$, the variance of $\mathscr{P}_{\psi^{\xi}}(F_t)$ is smaller than the variance of F_t in the sense:*

$$\forall t > 0, \int_{\mathbb{T}^2} \mathrm{Var}(\mathscr{P}_{\psi^{\xi}}(F_t)) \leq \int_{\mathbb{T}^2} \mathrm{Var}(F_t),$$

where $\mathrm{Var}(F_t) = \mathbb{E}(|F_t|^2) - \mathbb{E}(|F_t|)^2$ *and* $|F_t|$ *being the Euclidian norm.*

3 Numerical Experiments

This section is devoted to a numerical illustration of the practical value of the projected ABF compared to the standard ABF approach.

We consider a system composed of 100 particles in a two-dimensional periodic box. Among these particles, three particles are designated to form a trimer, while the others are solvent particles. All particles interact through several potential functions such as the Lennard-Jones potential, the double-well potential and a potential on the angle formed by the trimer. We choose the reaction coordinate to be the transition from compact to stretched state in each bond of the trimer. We apply now ABF and PABF dynamics to the trimer problem described above. One can refer to [1] for more detailed descriptions of the model and the used ABF and PABF algorithms.

First, we illustrate the improvement of the projected ABF method in terms of the variances of the biasing forces by comparing $\int \mathrm{Var}(\nabla A_t) = \int \mathrm{Var}(\partial_1 A_t) + \int \mathrm{Var}(\partial_2 A_t)$ (for the PABF method) with $\int \mathrm{Var}(F_t) = \int \mathrm{Var}(F_t^1) + \int \mathrm{Var}(F_t^2)$ (for the ABF method). Figure 1 shows that the variance for the projected ABF method is smaller than for the standard ABF method, where $\int \mathrm{Var}(\nabla A_t)$ (respectively $\int \mathrm{Var}(F_t)$) is represented by $Var(F1) + Var(F1)$ (respectively $Var(A1) + Var(A1)$).

We now present, the variation, as a function of time, of the normalized averages L^2-distance between the real free energy and the estimated one. As can be seen in Fig. 2, in both methods, the error decreases as time increases. Moreover, this error is always smaller for the projected ABF method than for the ABF method.

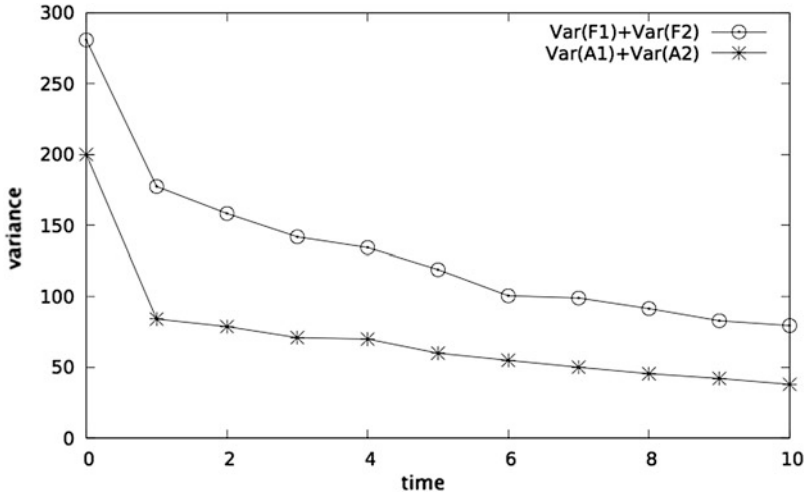


Fig. 1 Variances as a function of time. Reproduced courtesy SMAI-JCM [1]

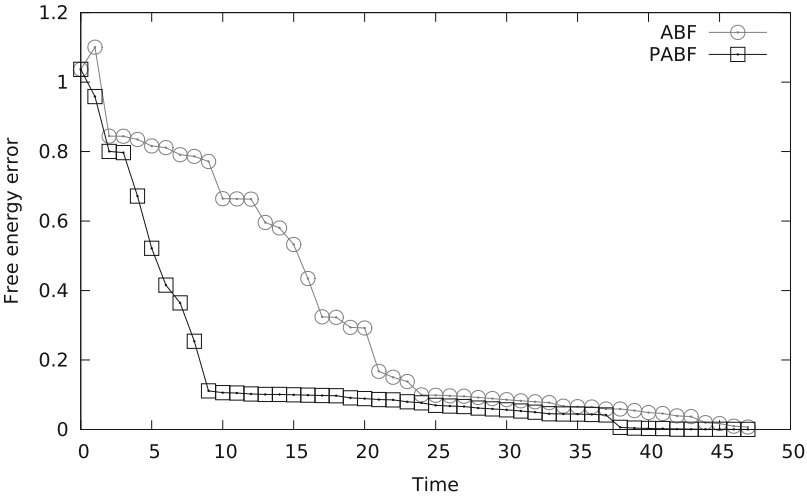


Fig. 2 Free energy error as a function of time. Reproduced courtesy SMAI-JCM [1]

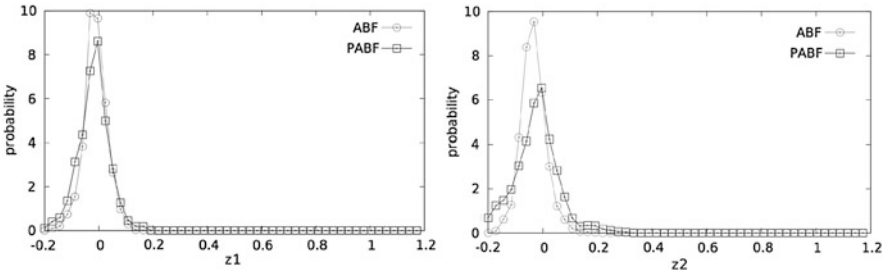


Fig. 3 At time 0.025. *Left*: $\int \psi^{\xi}(z_1, z_2)dz_2$; *Right*: $\int \psi^{\xi}(z_1, z_2)dz_1$. Reproduced courtesy SMAI-JCM [1]

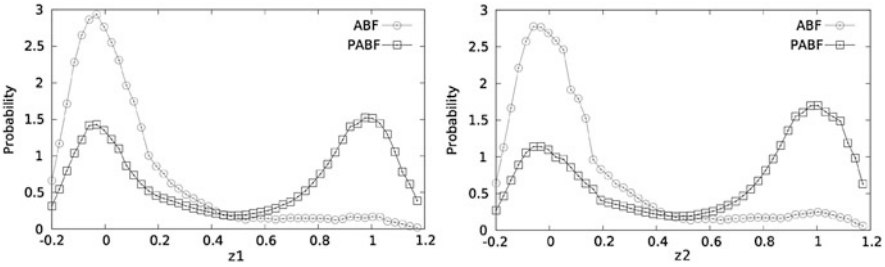


Fig. 4 At time 5. *Left*: $\int \psi^{\xi}(z_1, z_2)dz_2$; *Right*: $\int \psi^{\xi}(z_1, z_2)dz_1$. Reproduced courtesy SMAI-JCM [1]

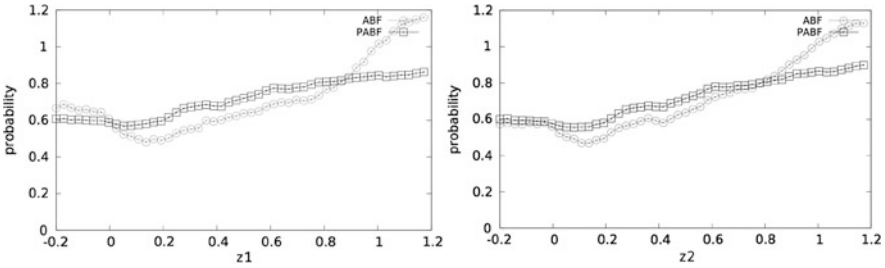


Fig. 5 At time 25. *Left*: $\int \psi^{\xi}(z_1, z_2)dz_2$; *Right*: $\int \psi^{\xi}(z_1, z_2)dz_1$. Reproduced courtesy SMAI-JCM [1]

Another way to illustrate that the projected ABF method converges faster than the standard ABF method is to plot the density function ψ^{ξ} as a function of time. It is clearly observed (see Figs. 3, 4, and 5) that, for the PABF method, the convergence of ψ^{ξ} to uniform law along (ξ_1, ξ_2) is faster with the projected ABF method.

References

1. Alrachid, H., Lelièvre, T.: Long-time convergence of an adaptive biasing force method: variance reduction by Helmholtz projection. SMAI J. Comput. Math. **1**, 55–82 (2015)
2. Darve, E., Pohorille, A.: Calculating free energy using average forces. J. Chem. Phys. **115**, 9169–9183 (2001)
3. Hénin, J., Chipot, C.: Overcoming free energy barriers using unconstrained molecular dynamics simulations. J. Chem. Phys. **121**, 2904 (2004)
4. Lelièvre, T., Rousset M., Stoltz, G.: Computation of free energy profiles with adaptive parallel dynamics. J. Chem. Phys. **126**, 134111 (2007)
5. Lelièvre, T., Rousset M., Stoltz, G.: Long-time convergence of an adaptive biasing force method. Nonlinearity **21**, 1155–1181 (2008)
6. Lelièvre, T., Rousset, M., Stoltz, G.: Free Energy Computations: A Mathematical Perspective. Imperial College Press, London (2010)

Modeling Chemical Kinetics in Solid State Reactions

J.A. Huidobro, I. Iglesias, B.F. Alfonso, C. Trobajo, and J.R. Garcia

Abstract This work deals with the kinetics of thermally stimulated processes which take place in the solid state phases. The activation energy of the solid is calculated using several methods of different families of isoconversional methods (differential, integral and incremental). A model of the kinetics is obtained by a method independent from the procedure used to compute the activation energy and it is analysed in three theoretical simulations as well as the thermal degradation of $FeNH_4(HPO_4)_2$. The reconstructed $\alpha - T$ curves of the simulations and the experimental case indicates that the model works properly.

1 Modeling Kinetics

The study of kinetics in solid state reactions explains the mechanism of the chemical processes as well as the meaning of the related parameters. It provides qualitative and quantitative information on phase transformations, crystallization, thermal decomposition, etc. Several analysis techniques such as Thermogravimetric Analysis (TGA), have been developed to measure physical and chemical changes.

A simple stimulated thermal reaction follows a scheme in the form $Re \rightarrow P + S$ where Re is the reactant, P the solid product and S is the solvent or water vapour. The reaction progress is given by the extent of conversion $\alpha(t) = (m_0 - m(t))/(m_0 - m_f)$

J.A. Huidobro (✉)
Departamento de Matematicas, Universidad de Oviedo, Gijón, Spain
e-mail: jahuidobro@uniovi.es

I. Iglesias • B.F. Alfonso
Departamento de Fisica, Universidad de Oviedo, Gijón, Spain
e-mail: iis@uniovi.es; mbafernandez@uniovi.es

C. Trobajo • J.R. Garcia
Departamento de Quimica Organica e Inorganica, Universidad de Oviedo, Oviedo, Spain
e-mail: ctf@uniovi.es; jrgm@uniovi.es

M. Mateos, P. Alonso (eds.), *Computational Mathematics, Numerical Analysis and Applications*, SEMA SIMAI Springer Series 13,
DOI 10.1007/978-3-319-49631-3_11

where $m(t)$ is the mass of Re at time t and m_0 and m_f are the initial and final masses, respectively. These reactions are commonly described by the equation

$$\frac{d\alpha}{dt} = A\exp(-\frac{E}{RT})f(\alpha) \tag{1}$$

where T is the temperature, R the universal gas constant, A the pre-exponential factor, E the activation energy and $f(\alpha)$ the model function [5].

The knowledge of A, E and $f(\alpha)$, the called kinetic triplet, allows solving Eq. (1) and so a description of the process can be obtained. In the model-fitting methods, the obtention of the kinetic triplet is based on the determination of the model function by fitting several reaction models to the experimental data and then the coefficients A and E are computed. But different forms of $f(\alpha)$ with disparate values of A and E can be fitted to the data and then these methods are not recommended [6].

In isoconversional methods the activation energy is computed without knowing the model function or the pre-exponential factor [7]. Consequently, a model of the process based on Eq. (1) cannot be obtained. Some authors [4, 8] have proposed different methods in order to calculate the product $Af(\alpha)$, considered as a sole factor, depending on how the activations energy has been computed. The main purpose of this study is to analyse the behaviour of a method to compute $Af(\alpha)$, independent of the procedure used to compute the activation energy, when it is applied to three theoretical simulations as well as the thermal degradation of $FeNH_4(HPO_4)_2$.

Mechanisms of chemical transformations are indeed complicated, they usually involve more than a single reaction. Then, Eq. (1) must be understood as an approximation to describe the process, the kinetic parameters are considered as apparent parameters and their physical meaning should be carefully analysed.

One of the simplest isoconversional methods is that proposed by Friedman (FR) [2], which is a differential isoconversional method. For a constant heating rate program of temperature $T = T_0 + \beta t$ and taking logarithms, Eq. (1) turns into

$$\ln\left(\frac{d\alpha(T)}{dT}\beta\right) = \ln(Af(\alpha(T))) - \frac{E}{RT} \tag{2}$$

where now $\alpha(T)$ represents the dependence of the extent of conversion respect to the temperature. Several runs with different heating rates β_i, $i = 1, \ldots, n$ with $n \geq 3$ are carried out and n experimental $\alpha - T$ curves are obtained. Thus, for a fixed value of α and from each experimental curve, values for T_i and $d\alpha(T_i)/dT$ are obtained. Then, from Eq. (2), the points $(1/T_i, \ln(d\alpha(T_i)/dT)\beta_i))$ belong to a straight line whose slope is $-E_\alpha/R$. The activation energy E_α can be obtained by fitting to the experimental data.

Wu et al. [8] extended this method by computing not only E but also $\ln(Af(\alpha(T)))$ and then, the product $Af(\alpha)$ is known and the differential Eq. (1) can be solved.

Generally, one drawback of this method is its sensitivity to noise that can come from numerical differentiation or experimental measures. A method (MFR) to

diminish this effect was proposed in [3] where Eq. (2) is considered for more values of α. Then, by fitting to the experimental data, the activation energy is computed.

Less sensitive to noise are integral isoconversional methods that consider an integral form of Eq. (1) but $\exp(-E/(RT))$ does not have a suitable antiderivative and some approximations have been proposed. They are based on assuming E is constant over the whole process and this is not very common. One of these is the generalized Kissinger method [1] (KAS), widely used. Vyazovkin [5] introduced an non-linear method (Vyaz) by integration of Eq. (1) over $[\alpha^* - \Delta\alpha^*, \alpha^*]$ and Samuelsson [4] computed $Af(\alpha)$, assuming it is constant over the interval.

A different idea, where this assumption is not necessary, is to compute the factor $Af(\alpha)$ directly from Eq. (1). Assuming the activation energy is known, for a fixed value of α, the product $Af(\alpha)$ can be obtained by fitting to the experimental data. Then, the differential Eq. (1) can be solved and a model is obtained. In this work, the four aforementioned methods were used to determine the activation energy and four set of values $E_\alpha - \alpha$ were obtained. For each method, the corresponding values of $Af(\alpha)$ were computed and model of the kinetics was achieved.

2 Results and Conclusions

This procedure to obtain a kinetic model has been implemented in Matlab and it was applied to three theoretical simulations. In all cases four constant heating rates were applied to generate the simulated data. In the first case, a one-step model with a first-order model function $f(\alpha) = 1 - \alpha$ and Arrhenius parameters $A = 10^9\ \text{min}^{-1}$ and $E = 10^2\ \text{kJ mol}^{-1}$ were considered. The four methods used to compute the activation energy provide similar values. Using them, the product $Af(\alpha)$ was computed and then $\alpha - T$ curves were plotted by solving the general kinetic differential equation.

In the second simulation, a parallel two-step, equally weighted, case was analysed being $f_1(\alpha) = 1 - \alpha$, $A_1 = 10^{12}\ \text{min}^{-1}$ and $E_1 = 167\ \text{kJ mol}^{-1}$ and $f_2(\alpha) = 1 - \alpha$, $A_2 = 10^{26}\ \text{min}^{-1}$ and $E_2 = 352\ \text{kJ mol}^{-1}$. The third simulation analysed an Avrami-Erofeev model function $f(\alpha) = 4(1-\alpha)[-\ln(1-\alpha)^{3/4}]$ with $A = 10^2\ \text{min}^{-1}$ and $E = 20.9\ \text{kJ mol}^{-1}$. Figure 1 shows the simulated and the reconstructed $\alpha - T$ curves for the three simulations. In all cases a good agreement was achieved.

Finally, thermogravimetric analysis of the thermal degradation of $FeNH_4(HPO_4)_2$ was conducted in a N_2 dynamic atmosphere, using a Mettler-Toledo TGA/SDTA851^e, at four different heating rates. As can be seen in Fig. 2a, the process occurs in two consecutive stages, the on-line mass spectrometric analysis indicates that the material firstly loses water about 600 K and secondly, at about 900 K, water and ammonia. The $E - \alpha$ plot, displayed in Fig. 2b, shows this behaviour. The experimental data are satisfactorily reconstructed (Fig. 2c)

In conclusion, the product $Af(\alpha)$ has been computed by a method independent of how the activation energy has been obtained. Then, the general kinetic differential

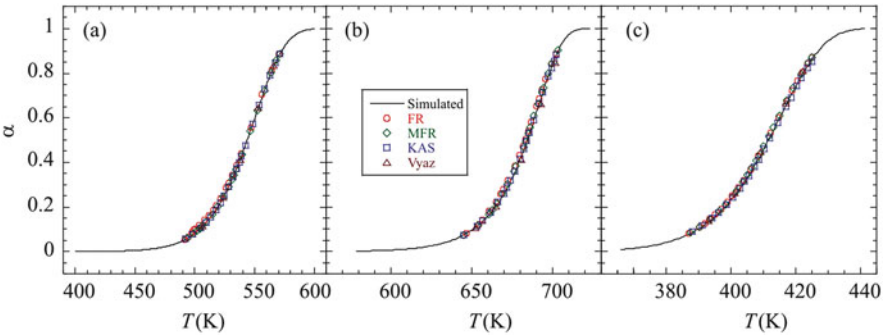


Fig. 1 Comparison of the simulated and reconstructed $\alpha - T$ curves for $\beta = 8\,\mathrm{K\,min^{-1}}$. Simulations 1, 2 and 3 in (**a**), (**b**) and (**c**), respectively

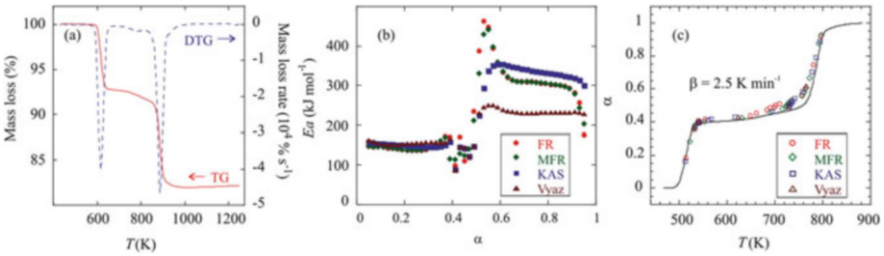


Fig. 2 TG and DTG curves of $FeNH_4(HPO_4)_2$ obtained at $10\,\mathrm{K\,min^{-1}}$ heating rate (**a**); activation energy versus extent of conversion (**b**); experimental (*line*) and reconstructed (*points*) $\alpha - T$ curves for $\beta = 2.5\,\mathrm{K\,min^{-1}}$ (**c**)

equation can be solved overcoming the ambiguity of the model-fitting methods. In this way, a discrete model that can be used to describe kinetics in solid state processes. This model has worked efficiently in the description of the theoretical simulations studied and in the thermal decomposition of $FeNH_4(HPO_4)_2$.

Acknowledgements This work was supported by Ministerio de Economia y Competitividad (MAT2013-40950-R, MAT2011-27573-C04-02), Gobierno del Principado de Asturias (GRUPIN14-060 and GRUPIN14-037), and FEDER.

References

1. Akahira, T., Sunose, T.: Joint Convention of four electrical institutes. Research report (Chiba Institute of Technology) Sci. Technol. **16**, 22–31 (1971)
2. Friedman, H.: Kinetics of thermal degradation of charge-forming plastics from thermogravimetry. Application to a phenolic plastic. J. Polym. Sci. Part C **6**, 183–195 (1964)
3. Huidobro, J.A., Iglesias, I., Alfonso, B.F., Trobajo, C., Garcia, J.R.: Reducing the effects of noise in the calculation of activation energy by the Friedman method. Chemom. Intell. Lab. Syst. (Chiba Institute of Technology) Sci. Technol. **16**, 22–31 (1971)

4. Samuelsson, L.N., Moriana, R., Babler, M.U., Ek, M., Engvall, K.: Model-free rate expression for thermal decomposition processes: The case of microcrystalline cellulose pyrolysis. Fuel **143**, 438–447 (2015)
5. Vyazovkin, S., Dollimore, D.: Linear and nonlinear procedures in isoconversional computations of the activation energy of nonisothermal reactions in solids. J. Chem. Inf. Comput. Sci. **151**, 146–152 (1996)
6. Vyazovkin, S., Wight, C.A.: Kinetics in solids. Annu. Rev. Phys. Chem. **48**, 125–149 (1997)
7. Vyazovkin, S., Burnham, A.K., Criado, J.M., Pérez-Maqueda, L.A., Popescu, C., Sbirrazzuoli, N.: ICTAC Kinetic Committee recommendations for performing kinetic computations on thermal analysis data. Thermochim. Acta **520**, 1–19 (2011)
8. Wu, W., Cai, J., Liu, R.: Isoconversional kinetic analysis of distributed activation energy. Model processes for pyrolysis of solid fuels. Ind. Eng. Chem. Res. **52**, 14376–14383 (2013)

ASSR Matrices and Some Particular Cases

P. Alonso, J.M. Peña, and M.L. Serrano

Abstract A real matrix is said Almost Strictly Sign Regular (ASSR) if all its nontrivial minors of the same order have the same strict sign. In this research, nonsingular ASSR matrices are characterized through the Neville elimination (NE). In addition, the algorithm is simplified for two important subclases: almost strictly totally negative (ASTN) matrices and Jacobi (tridiagonals) ASSR matrices.

1 Introduction

For $k, n \in \mathbb{N}$, with $1 \leq k \leq n$, $Q_{k,n}$ denotes the set of all increasing sequences of k natural numbers not greater than n. For $\alpha = (\alpha_1, \ldots, \alpha_k)$, $\beta = (\beta_1, \ldots, \beta_k) \in Q_{k,n}$ and A an $n \times n$ real matrix, we denote by $A[\alpha|\beta]$ the $k \times k$ submatrix of A containing rows $\alpha_1, \ldots, \alpha_k$ and columns $\beta_1, \ldots, \beta_k$ of A. If $\alpha = \beta$, we denote by $A[\alpha] := A[\alpha|\alpha]$ the corresponding principal submatrix. In addition, $Q^0_{k,n}$ denotes the set of increasing sequences of k consecutive natural numbers not greater than n.

The ASSR matrices present grouped null elements in certain positions, and can be classified in two classes which are defined below, type-I and type-II staircase.

A matrix $A = \left(a_{ij}\right)_{1 \leq i,j \leq n}$ is called type-I staircase if it satisfies simultaneously the following conditions

- $a_{11} \neq 0,\ a_{22} \neq 0,\ \ldots, a_{nn} \neq 0$;
- $a_{ij} = 0, i > j \Rightarrow a_{kl} = 0, \forall l \leq j,\ i \leq k$;
- $a_{ij} = 0, i < j \Rightarrow a_{kl} = 0, \forall k \leq i,\ j \leq l$.

P. Alonso • M.L. Serrano (✉)
University of Oviedo, Oviedo, Spain
e-mail: palonso@uniovi.es; mlserrano@uniovi.es

J.M. Peña
University of Zaragoza, Zaragoza, Spain
e-mail: jmpena@unizar.es

M. Mateos, P. Alonso (eds.), *Computational Mathematics, Numerical Analysis and Applications*, SEMA SIMAI Springer Series 13,
DOI 10.1007/978-3-319-49631-3_12

So, A is a type-II staircase matrix if it satisfies that P_nA is a type-I staircase matrix, where P_n is the backward identity matrix $n \times n$, whose element (i, j) is defined as

$$p_{ij} = \begin{cases} 1 \text{ if } i + j = n + 1, \\ 0 \qquad \text{otherwise.} \end{cases}$$

To describe the zero-pattern of this kind of matrices it is necessary to introduce several sets of indices that we denote as I, J, $\widehat{I}$ and $\widehat{J}$ (see, for instance, p. 482 of [1]). Besides, it is necessary to introduce the concepts of nontrivial matrices and signature sequence.

Given a matrix A of type-I (type-II) staircase, we say that a submatrix $A[\alpha|\beta]$, with $\alpha, \beta \in Q_{m,n}$ is nontrivial if all its main (secondary) diagonal elements are nonzero.

A vector $\varepsilon = (\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n) \in \mathbb{R}^n$ is a signature sequence, or simply, a signature, if $\varepsilon_i = \pm 1$, $\forall i \in \mathbb{N}$, $1 \leq i \leq n$.

Taking into account the previous results, we define the ASSR matrices:

Definition 1 A real matrix $A = (a_{ij})_{1 \leq i,j \leq n}$, is said to be ASSR, with signature $\varepsilon = (\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n)$, if it is either type-I or type-II staircase and all its nontrivial minors $\det A[\alpha|\beta]$ satisfy that

$$\varepsilon_m \det A[\alpha|\beta] > 0, \quad \alpha, \beta \in Q_{m,n}, \qquad m \leq n. \tag{1}$$

Notice that an ASSR matrix is nonsingular.

The characterizations presented here are based on the signs of the pivots of the NE, so we will introduce briefly this procedure (see [4]). If A is a nonsingular $n \times n$ matrix, NE consists of at most $n - 1$ successive major steps, resulting in a sequence of matrices as follows:

$$A = \widetilde{A}^{(1)} \to A^{(1)} \to \cdots \to \widetilde{A}^{(n)} = A^{(n)} = U \tag{2}$$

where U is an upper triangular matrix.

For each t, $1 \leq t \leq n$, $A^{(t)} = \left(a_{ij}^{(t)}\right)_{1 \leq i,j \leq n}$ has zeros in the positions $a_{ij}^{(t)}$, for $1 \leq j < t, j < i \leq n$. Besides, it holds that

$$a_{it}^{(t)} = 0, \ i \geq t \Rightarrow a_{ht}^{(t)} = 0, \ \forall h \geq i. \tag{3}$$

Matrix $A^{(t)}$ is obtained from $\widetilde{A}^{(t)}$ reordering rows $t, t + 1, \ldots, n$ according to a row pivoting strategy which satisfies (3).

To obtain $\widetilde{A}^{(t+1)}$ from $A^{(t)}$, zeros are introduced below the main diagonal of the tth column by subtracting a multiple of the ith row from the $(i+1)$th, for $i = n-1, \ldots, t$. The elements $\widetilde{a}_{ij}^{(t+1)}$ are obtained according to the following formula

$$\begin{cases} a_{ij}^{(t)}, & 1 \le i \le t, \\ a_{ij}^{(t)} - \frac{a_{it}^{(t)}}{a_{i-1,t}^{(t)}} a_{i-1,j}^{(t)}, & \text{if } a_{i-1,t}^{(t)} \neq 0,\ t+1 \le i \le n, \\ a_{ij}^{(t)}, & \text{if } a_{i-1,t}^{(t)} = 0,\ t+1 \le i \le n. \end{cases} \tag{4}$$

The element $p_{ij} = a_{ij}^{(j)}$, $1 \le i,j \le n$, is called the (i,j) pivot of NE of A.

2 Characterization of ASSR Matrices Through NE

In this section we present a characterization of ASSR matrices through the NE. First, we present necessary conditions for nonsingular type-I and type-II staircase matrices.

Theorem 2 *Let $B = (b_{ij})_{1 \le i,j \le n}$ be a nonsingular type-I staircase matrix, with zero pattern defined by I, J, $\widehat{I}$ and $\widehat{J}$. If B is ASSR with signature $\varepsilon = (\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n)$, then the NE of B and B^T can be performed without row exchanges and:*

- *the pivots p_{ij} of NE of B satisfy, for any $1 \le j \le i \le n$,*

$$p_{ij} = 0 \Leftrightarrow b_{ij} = 0 \tag{5}$$

$$\varepsilon_{j-j_t} \varepsilon_{j-j_t+1} p_{ij} > 0 \Leftrightarrow b_{ij} \neq 0 \tag{6}$$

where $\varepsilon_0 := 1$,

$$j_t := \max \{ j_s \,/\, 0 \le s \le k-1,\ j - j_s \le i - i_s \} \tag{7}$$

and k is the only index satisfying that $j_{k-1} \le j < j_k$,
- *and the pivots q_{ij} of NE of B^T satisfy, for any $1 \le i < j \le n$,*

$$q_{ij} = 0 \Leftrightarrow b_{ij} = 0 \tag{8}$$

$$\varepsilon_{i-\widehat{i}_{t'}} \varepsilon_{i-\widehat{i}_{t'}+1} q_{ij} > 0 \Leftrightarrow b_{ij} \neq 0 \tag{9}$$

where

$$\widehat{i}_{t'} := \max \left\{ \widehat{i}_s \,/\, 0 \le s \le k'-1,\ i - \widehat{i}_s \le j - \widehat{j}_s \right\} \tag{10}$$

and k' is the only index satisfying that $\widehat{i}_{k'-1} \le i < \widehat{i}_{k'}$.

Following, we characterize the type-I staircase matrices:

Theorem 3 *A nonsingular matrix* $A = \left(a_{ij}\right)_{1\leq i,j\leq n}$ *is ASSR with signature* $\varepsilon = (\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n)$*, with* $\varepsilon_2 = 1$ *if and only if for every* $h = 1, \ldots, n-1$ *the following properties hold simultaneously:*

(i) *A is type-I staircase;*
(ii) *the NE of the matrices* $A_h := A[h, \ldots, n]$ *and* $A_h^T := A^T[h, \ldots, n]$ *can be performed without row exchanges;*
(iii) *the pivots* p_{ij}^h *of the NE of* A_h *satisfy conditions corresponding to (5), (6), and the pivots* q_{ij}^h *of the NE of* A_h^T *satisfy (8) and (9);*
(iv) *for the positions* (i^h, j^h) *of matrix* A_h*:*

- *if* $i^h \geq j^h$ *and* $i^h - j^h = i_t^h - j_t^h$ *then* $\varepsilon_{j^h - j_t^h}\varepsilon_{j^h - j_t^h + 1} = \varepsilon_{j^h - 1}\varepsilon_{j^h}$,
- *if* $i^h < j^h$ *and* $i^h - j^h = \widehat{i}_{t'}^h - \widehat{j}_{t'}^h$ *then* $\varepsilon_{i^h - \widehat{i}_{t'}^h}\varepsilon_{i^h - \widehat{i}_{t'}^h + 1} = \varepsilon_{i^h - 1}\varepsilon_{i^h}$,

where indices i_t, j_t, $\widehat{i}_{t'}$ *and* $\widehat{j}_{t'}$ *are given by (7) and (10).*

When an ASSR matrix is multiplied by the backward identity matrix, the product is also an ASSR matrix. In the following result the relationship between the signature of A and P_nA is given.

Corollary 4 *A matrix* $A = (a_{ij})_{1\leq i,j\leq n}$ *is ASSR if and only if* P_nA *is also ASSR. Furthermore, if the signature of* A *is* $\varepsilon = (\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n)$*, then the signature of* P_nA *is* $\varepsilon' = (\varepsilon'_1, \varepsilon'_2, \ldots, \varepsilon'_n)$*, with* $\varepsilon'_m = (-1)^{\frac{m(m-1)}{2}}\varepsilon_m$*, for all* $m = 1, \ldots, n$.

Observe that, if the second signature of A is $\varepsilon_2 = -1$, then, the second signature of P_nA is given by $\varepsilon'_2 = (-1)^{\frac{2(2-1)}{2}}(-1) = 1$. This allow us to apply the Theorem 3 to the matrix P_nA and all the ASSR matrices are characterized.

3 Characterization of ASTN Matrices

A real matrix $A = (a_{ij})_{1\leq i,j\leq n}$ is said almost strictly totally negative (ASTN) if it is ASSR with signature $(-1, -1, \ldots, -1)$. In this section we present the obtained characterization for this kind of matrices, which allows us to reduce the computational cost of testing the ASTN characteristic.

Theorem 5 *Given a nonsingular matrix* $A = (a_{ij})_{1\leq i,j\leq n}$*, with* $n \geq 2$*, A is ASTN if and only if the following properties hold simultaneously:*

(a) $a_{ij} \neq 0$ *if* $(i,j) \notin \{(1,1), (n,n)\}$.
(b) *The NE of* $B = P_nA$ *and* B^T *can be performed without row exchanges.*
(c) *The pivots* p_{ij} *of the NE of B, with* $i \geq j$ *verify:*

$$p_{n1} = 0 \Leftrightarrow b_{n1} = 0, \tag{11}$$

$$\text{if } j = j_t, \text{ then } p_{ij} < 0 \Leftrightarrow b_{ij} \neq 0, \tag{12}$$

$$\text{if } j > j_t, \text{ then } (-1)^{j-j_t}p_{ij} > 0 \Leftrightarrow b_{ij} \neq 0, \tag{13}$$

and the pivots q_{ij} of B^T with $i < j$ verify

$$q_{1n} = 0 \Leftrightarrow b_{1n} = 0, \tag{14}$$

$$\text{if } i = \widehat{i}_t, \text{ then } q_{ij} < 0 \Leftrightarrow b_{ij} \neq 0, \tag{15}$$

$$\text{if } i > \widehat{i}_t, \text{ then } (-1)^{i-\widehat{i}_t} q_{ij} > 0 \Leftrightarrow b_{ij} \neq 0, \tag{16}$$

where indices i_t, j_t, $\widehat{i}_t$ and $\widehat{j}_t$ are given by conditions corresponding to (7) and (10).

(d) The matrix $M = A[1, \ldots, n-1|2, \ldots n]$ is strictly totally negative.

Notice that a strictly totally negative matrix is a matrix whose minors are all strictly negative. In [3], the authors present a result to test whether a matrix is STN, checking the signs of the pivots elements.

4 Tridiagonal Matrices

The tridiagonal matrices or Jacobi matrices, often appear to solve problems by numerical methods, see for example the finite element method in one dimension. The Jacobi ASSR matrices are characterized in this section.

Definition 6 Given a matrix $A = \left(a_{ij}\right)_{1 \leq i,j \leq n}$, we say that it is a tridiagonal matrix if $a_{ij} = 0$ when $|i-j| > 1$. In addition, if $a_{ij} \neq 0$ when $|i-j| \leq 1$, we say that A is strictly tridiagonal matrix.

In [2], the authors shown that for a tridiagonal ASSR matrix, with $A \geq 0$, the only feasible signatures are $(1, \ldots, 1, 1)$ or $(1, \ldots, 1, -1)$. In the next Theorem a characterization is given.

Theorem 7 *Let $A = (a_{ij})_{1 \leq i,j \leq n}$ be a real nonnegative tridiagonal matrix and nonsingular. Then A is ASSR with $\varepsilon = (1, 1, \ldots, 1, \varepsilon_n)$ if and only if it holds that*

(a) A is type-I staircase,

(b) the NE of the matrices A and A^T can be performed without row changes,

(c) the pivots p_{ij} of the NE of A and the pivots q_{ij} of the NE of A^T satisfy:

- *If $i \geq j$,*

$$p_{ij} = 0 \Leftrightarrow a_{ij} = 0, \tag{17}$$

$$\left.\begin{array}{l} j < n, \quad p_{ij} > 0 \\ j = n, \, \varepsilon_n p_{in} > 0 \end{array}\right\} \Leftrightarrow a_{ij} \neq 0, \tag{18}$$

- *If $i < j$,*

$$q_{ij} = 0 \Leftrightarrow a_{ij} = 0, \tag{19}$$

$$q_{ij} > 0 \Leftrightarrow a_{ij} \neq 0, \tag{20}$$

(d) $A_2 = A[2, \ldots, n]$ is ASTP.

A definition of Almost Strictly Totally Positive (ASTP) matrices can be found in [2].

If an ASSR matrix $A = (a_{ij})_{1 \leq i,j \leq n}$ verifies that $a_{ij} \leq 0$ for all $1 \leq i, j \leq n$, then $-A$ is an ASSR matrix with $\varepsilon'_1 = 1$. Then we can apply the Theorem 7 to $-A$ and all the strictly tridiagonal matrices are characterized.

Acknowledgements This work has been partially supported by the Spanish Research Grant MTM2015-65433-P (MINECO/FEDER) and MTM2015-68805-REDT.

References

1. Alonso, P., Peña, J.M., Serrano, M.L.: On the characterization of almost strictly sign regular matrices. J. Comput. Appl. Math. **275**, 480–488 (2015)
2. Alonso, P., Peña, J.M., Serrano, M.L.: Characterizations of M-banded ASSR matrices. In: Trends in Differential Equations and Applications, pp. 33–49. Springer International Publishing, Cham (2016)
3. Gasca, M., Peña, J.M.: A test for strictly sign-regularity. Linear Algebra Appl. **198**, 133–142 (1994)
4. Gasca, M., Peña, J.M.: A matricial description of Neville elimination with applications to total positivity. Linear Algebra Appl. **202**, 33–54 (1994)

A Computational Approach to Verbal Width in Alternating Groups

Jorge Martínez Carracedo and Consuelo Martínez López

Abstract We know that every element in an Alternating group A_n, $n \geq 5$, can be written as a Engel word of length two (Carracedo, Extracta Math. **30**(2), 251–262, 2015 and J. Algebra Appl. **16**(2), 1750021, 10 p., 2017). There is a conjecture that every element in an Alternating group A_n, $n \geq 5$, can be written as an Engel word of arbitrary length. We give here a computational approach to this problem, what allows to prove the conjecture for $5 \leq n \leq 14$.

1 Engel Graphs

Given an arbitrary group G and a word in the free group of rank r, $\omega \in F_r$, with r a natural number, we can consider the word map

$$\omega_G : \overbrace{G \times \cdots \times G}^{r} \longrightarrow G$$

that maps each r-tuple $(g_1, g_2, \ldots, g_r)$ to $\omega_G(g_1, g_2, \ldots, g_r)$.

Several questions can be formulated: What is the size of the set $\omega_G(G)$? Is the map ω_G surjective? Is $\langle \omega_G(G) \rangle = G$? Can we find a constant k such that $\omega_G(G)^k = \langle \omega(G) \rangle$?. See [1].

An Engel word of length m is the element of the free group of rank 2 given by:

$$E_m(x, y) := [\ldots [x, \overbrace{y], y], \ldots, y}^{m}].$$

Our aim is to study the verbal width of an Engel word of arbitrary length in an Alternating group A_n, that is, for each $m \geq 1$ we want to find a constant $k \geq 1$ such that every element σ in A_n can be written as a product of at most k Engel words of length m.

J.M. Carracedo (✉) • C.M. López
Department of Mathematics, University of Oviedo, Oviedo, Spain
e-mail: beleragor@gmail.com; chelo@orion.ciencias.uniovi.es

M. Mateos, P. Alonso (eds.), *Computational Mathematics, Numerical Analysis and Applications*, SEMA SIMAI Springer Series 13,
DOI 10.1007/978-3-319-49631-3_13

In [2] it is proved that this constant is at most 2 for every $m \geq 1$ and every $n \geq 5$, while for Engel words of length 2 it is 1, ie, every element in an Alternating group A_n, $n \geq 5$, can be written as an Engel word of length 2.

Technics used to prove these results fail when we consider Engel words of higher length. So we have intended a computational approach. We define the Engel Graph (V_n^y, E) depending on an element y in A_n and we use GAP to study this graph.

If we denote the set of Engel words of length m by $E_m(y) := \{E_m(x, y) \mid x \in A_n\}$ and by Ω_m^y the set $\{C_{A_n}(y)x \mid x \in E_{m-1}(y)\}$, we can construct a map

$$\begin{array}{rccc} \varphi_m : & \Omega_m^y & \longrightarrow & E_m(y) \\ & C_{A_n}(y)x & \mapsto & [x, y] \end{array} \tag{1}$$

Theorem 1 *For every $m \geq 1$ and every element $y \in A_n$, $n \geq 5$, the map φ_m is well defined and bijective.*

Now, let's construct a directed graph taking as set of nodes the set $V_n^y := \Omega_1^y = \{C_{A_n}(y)x \mid x \in A_n\}$ and whose arrows are defined by:

- Given $C_{A_n}(y)z_1, C_{A_n}(y)z_2 \in V_n^y$, there exists an arrow from $C_{A_n}(y)z_1$ to $C_{A_n}(y)z_2$ if and only if $C_{A_n}(y)[z_1, y] = C_{A_n}(y)z_2$.

Definition 2 Let y be an element in an Alternating group A_n, the graph (V_n^y, E) is called Engel graph associated to the element y and the group A_n.

Some of the reasons why we have defined this graph are:

- If we want to compute $E_k(x, y)$, it is enough to start with the node $C_{A_n}(y)x$ and, in each step k_i, to compute the commutator with y of an arbitrary element of the coset $C_{A_n}(y)z_{k_i}$.
- We can study the dynamic of the set $\{E_m(\cdot, y)\}_{m\geq 0}$ by studying the dynamic of the graph (V_n^y, E)

Once we have build the graph, we want to know whether or not an element in the Alternating group A_n, $n \geq 5$, can be written as an Engel word of type $E_m(\cdot, y)$ for $m \geq 1$. We will study the directed cycles of the Engel graph (V_n^y, A).

Theorem 3 *Let φ_1 be the map given in (1) with $m = 1$. If (W, β) is a directed cycle of (V_n^y, E), then every element in the set $\varphi_1(W)$ can be written as an Engel word of arbitrary length.*

Proof Consider (W, β) a directed cycle in the Engel graph (V_n^y, E).

Given an arbitrary element $C_{A_n}(y)x$ in W, we have that

$$\varphi_1(W) := \{E_l(x, y) \mid l \in N\}.$$

Since W is a directed cycle, there exists an integer $k \geq 1$ such that

$$C_{A_n}(y)E_{k-1}(x, y) = C_{A_n}(y)x.$$

and so $E_k(x, y) = [x, y]$, where $k - 1$ is the length of the cycle (W, β).

Let m be an arbitrary integer, $m \geq 1$. For any permutation σ in $\varphi_1(W)$, we have that $\sigma = [x, y]$ for $C_{A_n}(y)x \in W$ and then

$$\sigma = [x, y] = E_k(x, y) = E_{2k}(x, y) = \ldots E_{rk}(x, y),$$

for every $r \geq 1$.

It is enough to take r such that $rk > m$ to get that $\sigma = E_m(\tau, y)$ for some $\tau \in A_n$. □

Example Consider the element $y := (1, 2, 3, 4, 5)$ in A_5. We have that $C_{A_5}(y) = \langle y \rangle$ is a cyclic group of order 5, so $V_5^y = \{\langle y \rangle x \mid x \in A_5\}$ is a set whose cardinal is $| A_5/\langle y \rangle | = 12$.

Let's construct the Engel graph (V_5^y, E). As we know, each node is associated to a coset module $C_{A_5}(y)$. We will denote each node $C_{A_5}(y)\sigma$ by a permutation of the set $\{y^j\sigma \mid 1 \leq j \leq 4\}$ (Fig. 1).

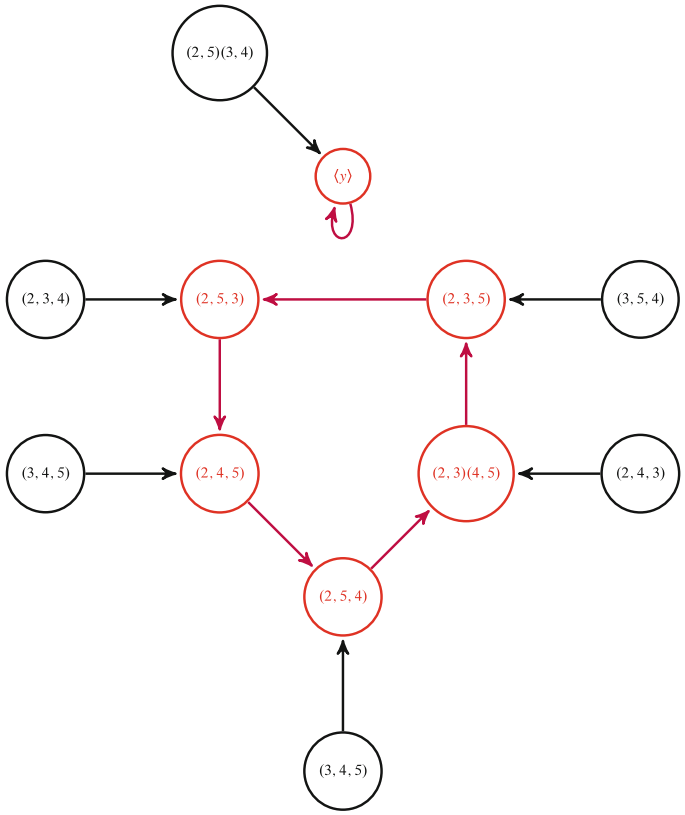


Fig. 1 We can see here the Engel graph (V_5^y, A). There are two directed cycles (*drawn in red*) in the graph. The first one, W_1, is a cycle with five elements, and the other one, W_2, has only one node, $C_{A_5}(y)$

Table 1 Computational results using GAP

Group	Conjugacy classes not represented	Runtime
A_5	$\{(1,2)(3,4)^{S_5}, (1,2,3)^{S_5}\}$	7 mm s
A_6	$\{(1,2)(3,4)^{S_6}, (1,2,3)^{S_6}, (1,2,3)(4,5,6)^{S_6}\}$	18 mm s
A_7	$\{(1,2)(3,4)^{S_7}\}$	40 mm s
A_8	$\{(1,2)(3,4)^{S_8}\}$	201 mm s
A_9	$\{(1,2)(3,4)^{S_9}\}$	4 s 12 mm s
A_{10}	$\{(1,2)(3,4)^{S_{10}}\}$	40 s 809 mm s
A_{11}	$\{(1,2)(3,4)^{S_{11}}\}$	5 min 37 s 139 mm s
A_{12}	$\{(1,2)(3,4)^{S_{12}}\}$	63 min 38 s 210 mm s
A_{13}	$\{(1,2)(3,4)^{S_{13}}\}$	21 h 6 min 54 s
A_{14}	$\{(1,2)(3,4)^{S_{14}}\}$	>12 days

We have used GAP [3] to compute the directed cycles $\{W_k\}_{1\leq k\leq r}$ of the Engel graph (V_5^y, E) for $5 \leq n \leq 14$, where $y = (1,2,\ldots,n)$ if n is odd and $y = (1,2,\ldots,n-1)$ if n is even. Later we have computed

$$\bigcup_{k=1}^{r} \varphi_1(W_k),$$

and we see which conjugacy classes of S_n are not represented in this set. The obtained results are given in Table 1.

Using these computational results and Theorem 3 the following theorem can be proved.

Theorem 4 *Every element in an Alternating group A_n, $5 \leq n \leq 14$, can be written as an Engel word of arbitrary length. That is*

$$A_n = E_m(A_n), \quad 5 \leq n \leq 14, \quad \forall m \geq 2$$

Let us highlight that we could not have got this result computationally through a Brute-force attack, not only because the huge order of A_n when n is big $(n!/2)$, but also because the length of the Engel word is not bounded.

References

1. Carracedo, J.M.: Powers in alternating simple groups. Extracta Math. **30**(2), 251–262 (2015)
2. Carracedo, J.M.: Engel words in alternating groups. J. Algebra Appl. **16**(2), 1750021, 10 p. (2017)
3. GAP Bibliography. www.gap-system.org/

Improvements in Resampling Techniques for Phenotype Prediction: Applications to Neurodegenerative Diseases

Juan Carlos Beltrán Vargas, Enrique J. deAndrés-Galiana, Ana Cernea, and Juan Luis Fernández-Martínez

Abstract Searching for new biomarkers, biological networks and pathways is crucial in the solution of neurodegenerative diseases. In this research we have compared three different algorithms and resampling techniques to find possible genetic causes in patients with Alzheimer's and Parkinson's diseases, providing some interesting insights about the main causes involved in these diseases.

1 The Phenotype Prediction Problem

The study of major neurodegenerative diseases is a clear priority since 16% of the population in Europe is over 65 years old, they affect more than seven million Europeans, and it is expected that this figure will double in the next 20 years. Despite all the research done, the number of existing treatments is very limited and only address symptoms, instead of finding causes. Gene expression analysis studies can provide a snapshot of actively expressed genes and transcripts under various conditions. The solution requires the use of advanced bioinformatics algorithms able to guide the analysis of genetic data relative to this kind of diseases, allowing the discovery of new biomarkers, biological networks and pathways,orphan drugs, and new therapeutic targets. For that purpose, the corresponding phenotype prediction is formulated as a binary supervised classification problem. The ingredients are:

1. A matrix $E \in M_{m\times n}(\mathbb{R})$ of n genes for a set of m samples with $m << n$, where E_{ij} is the expression of gene j in sample i; and the vector of observed phenotype classes $\mathbf{c}^{obs} \in \mathbb{R}^m$.

J.C. Beltrán Vargas (✉) • E.J. deAndrés-Galiana • A. Cernea • J.L. Fernández-Martínez
Departament of Mathematics, University of Oviedo, Calle Calvo Sotelo s/n, 33007 Oviedo, Spain
e-mail: uo226711@uniovi.es; beltrancito@gmail.com; eag@aic.uniovi.es; ana.cernea@uniovi.es; jlfm@uniovi.es

M. Mateos, P. Alonso (eds.), *Computational Mathematics, Numerical Analysis and Applications*, SEMA SIMAI Springer Series 13,
DOI 10.1007/978-3-319-49631-3_14

2. A classifier $L^*(\mathbf{g}) : \mathbf{g} \in \mathbb{R}^s \rightarrow C$ where $\mathbf{g}$ is the set of genetic signatures of size $s << n$, and C is the set of binary classes. The classifier L^* is built ad-hoc and it is just a mathematical abstraction used to discover the genes/pathways that are involved in the phenotype discrimination.

The relevant features would be the ones that minimize the cost function $O(\mathbf{g})$ related to the class prediction vector: $O(\mathbf{g}) = ||\mathbf{L}^*(\mathbf{g}) - \mathbf{c}^{obs}||_p$, where $\mathbf{L}^*(\mathbf{g}) = (L^*(\mathbf{g}_1), \ldots, L^*(\mathbf{g}_i), \ldots, L^*(\mathbf{g}_m))$, p is the norm applied in the distance criterion, $\mathbf{L}^*(\mathbf{g})$ is the set of predicted classes by the classifier $L^*(\mathbf{g})$, $\mathbf{g}_i \in \mathbb{R}^s$ is the set of genes of size s corresponding to sample i, and $L^*(\mathbf{g}_k)$ is the predicted class for sample k. These genetics signatures will be used to predict the class of new incoming samples. Due to noise in data and modeling errors, the phenotype prediction problem is ill-posed [2]. In presence of noise the set of genes with the highest predictive accuracy will never perfectly coincide with the set(s) of genes that explains the disease. For that reason it is desirable to also look for other sets of features with lower predictive accuracy than the optimum. A robust solution consists in finding the genetic networks and pathways that are involved, and performing ensemble-based predictions [5]. For that purpose we have used in all the cases the leave-one-out-crossvalidation (LOOCV) accuracy.

2 Methodology

In this contribution we have compared different methodologies:

1. Adaboost [3] with Adaptive Resampling. The adaboost version that it is proposed consists in building a *strong classifier*, $\mathbf{L}_e^*(\mathbf{g}) = \sum_{k=1}^{N_w} w_k l_k^*(\mathbf{g}_k)$, where l_k^* are the *weak classifiers* and w_k the corresponding weights, such as it is possible to build a new $\mathbf{L}_e^*(\mathbf{g})$ exploring the classification cost function landscape [2] to choose an improved combination of $l_k^*(\mathbf{g}_k)$.
2. K-NN [1] with Network Resampling, following a prior probability distribution that is based in the Fisher's Ratio of the most discriminatory genes. This algorithm works as follows: (1) First of all, the smallest-scale gene signature is found via Backwards Feature Elimination of the ranked list of genes with Fisher's ratio greater than a given cut-off. (2) Secondly, a random sampling is performed using as prior probability distribution induced by the Fisher's ratio. For that purpose the genes are divided into two different categories: headers and helpers. Headers expand the low frequency details in the classification, while helpers provide high frequency details in phenotype discrimination. (3) The final prediction is given by majority voting using the high predictive signatures sampled in the previous step.
3. SVM with LASSO [7], that is, as support vector machines algorithm with a L_1 regularization in the genetic signature. This algorithm tries to find the maximum-margin hyperplane for a binary classification. This problem can be formulated as a regularized estimation problem, corresponding to a prediction error $O(\mathbf{g})$ plus a

regulation term in $||\mathbf{g}||_1$. The L_1 regularization provides the additional property to look for sparse genetic signatures (small-scale genetic signatures). Similarly to k-NN with Network Resampling, we first obtain the sparse coefficients ω (weights) for each gene through a cross validation experiment with twofolds and 5 repetitions. In each of the steps of the cross validation we determine the best complexity value c with a grid search algorithm. The genes are ranked according to ω, and finally, a Recursive Feature Elimination algorithm with linear SVM was used to find the smallest-scale gene signature having the maximum predictive accuracy.

3 Results and Discussion

These algorithms were applied to the analysis of two genetic datasets concerning Parkinson's disease **(PD)** [6] and Alzheimer's disease **(AD)** [4] and their respective control patients. The Parkinson's dataset contained 22,164 genetic probes and 114 samples (59 with PD), while the Alzheimer's dataset had 38,323 genetic probes and 329 samples (225 with Alzheimer and Mild Cognitive Impairment-MCI). Table 1 show the most predictive genes for AD and PD according to their respective Fisher's Ratio.

The best results obtained by these 3 methodologies are shown in Table 2. We also provide the LOOCV accuracy and the number of genes of the predictive signatures used to attain these results. The main results were:

1. SVM with LASSO and KNN-NR have obtained better results than adaboost. SVM-LASSO also provided the shortest high-predictive genetic signatures, but it was highly computationally intensive. KNN-NR improves in both cases the accuracy provided by KNN without resampling. In the case of Alhzaimer a signature of 72 genes with accuracy LOOCV = 77.2% was found. The resampling found a genetic signature with 12 genes and LOOCV = 81.5%. Finally, the majority voting improves accuracy up to 84.5%. In the case of Parkinson, these figures

Table 1 Best discriminatory genes in Alzheimer and Parkinson according to the Fisher's Ratio

Alzheimer		Parkinson	
Gene name	FR	Gene name	FR
LOC401206	1.29	GRHL1	1.35
MRPL51	1.19	SBDS	1.32
THX1BP1	1.17	RPS4Y1	1.28
RPS25	1.16	JARID1D	1.10
LOC650276	1.09	FAM29A	1.09
RPL36AL	1.04	UNQ1940	1.03
RPA3	0.99	CD27	1.09
LOC6462001	0.96	GPR142	1.01
LOC648000	0.93	LELP1	1.00
RPL17	0.92	FAM83C	1.00

Table 2 Accuracy (Acc %) and number of genes (#) of the small-scale gene signature obtained for the different methodologies: Adaboost with adaptive resampling (Adaboost-AR), K-NN with Network Resampling (KNN-NR), SVN with LASSO

	Parkinson		Alzheimer	
Methodology	Acc (%)	# of genes	Acc (%)	# of genes
Adaboost-AR	82.23	158	93.82	328
KNN-NR	84.50	12	97.40	46
SVM-LASSO	90.00	2	99.00	2

were 90.35% with 46 genes, 92.1% with 36 genes and a final majority voting of 97.40%. This algorithm is very fast and accurate. Adaboost with Adaptive Resampling provided good results at the expenses of increasing the length of the genetic signatures used. It is also highly computationally intensive.

2. The pathway analysis has shown the importance of several mechanisms concerning oxidative stress and transcriptions factor concerning hypoxia in the case of Parkinson Disease, and the role of Ribosomal and Mitochondrial Ribosomal proteins, involved in Influenza Viral RNA Transcription and Replication and Viral mRNA Translation in Alzheimer. The Parkinson's disease (99%) was better predicted than Alzheimer's disease (90%). This result suggests that some important genetic mechanisms in Alzheimer have not been sampled and/or the presence of behavioral outliers.

References

1. De Andrés, G.E., Fernández-Martínez, J.L., Sonis, S.T.: Design of biomedical robots for phenotype prediction problems. J. Comput. Biol. **23**, 678–692 (2016, to appear)
2. Fernández-Martínez, J.L., Fernández-Muñiz, M., Tompkins, M.: On the topography of the cost functional in linear and nonlinear inverse problems. Geophysics **77**(1), 1–15 (2012)
3. Freund, Y., Schapire, R.E.: Experiments with a new boosting algorithm. In: *Machine Learning: Proceedings of the Thirteenth International Conference* (1996), pp. 325–332
4. Kumaran, R.: Gene expression changes across multiple regions of the Parkinson's disease brain. Geo Dataset GSE28894 (2013)
5. Rokach, L.: Ensemble-based classifiers. Artif. Intell. Rev. **33**(1), 1–39 (2010)
6. Sood, S.: A novel multi-tissue RNA diagnostic of healthy ageing relates to cognitive health status. Genome Biol. **16**, 185 (2015). PMID: 26343147
7. Zou, H., Hastie, T.: Regularization and variable selection via the elastic net. J. R. Stat. Soc. Ser. B Stat. Methodol. **67**(2), 301–320 (2005)

An Aortic Root Geometric Model, Based on Transesophageal Echocardiographic Image Sequences (TEE), for Biomechanical Simulation

Marcos Loureiro-Ga, Maria F. Garcia, Cesar Veiga, G. Fdez-Manin, Emilio Paredes, Victor Jimenez, Francisco Calvo-Iglesias, and Andrés Iñiguez

Abstract Aortic valve (AoV) stenosis is one of the most common valvular diseases. Assessing the aortic valve function could provide crucial information towards a better understanding of the disease, where numerical simulation will have an important role to play. The main scope of this work is to find an aortic root (AR) patient specific geometric model, which could be used for simulation purposes. Several models were followed to obtain an AR geometry implementing them in open source tools. Necessary parameters were obtained from 2D echo images. In order to test the obtained AR geometry, a finite element study was performed solving a fixed mesh fluid structure interaction (FSI) model. The fluid was supposed to be laminar and the tissues were modeled as St. Venant-Kirchhoff materials. Obtained results for the 1-way FSI study are compared with the published ones for structural and 2-way FSI studies showing similar results. An AR geometric reconstruction from clinic data is suited for numerical simulation.

M. Loureiro-Ga (✉)
Departamento de matemática aplicada II, Universidade de Vigo, 36310 Vigo, Spain

Instituto de Investigación Sanitaria Galicia Sur (IIS Galicia Sur) SERGAS-UVigo, Cardioloxía, Hospital Álvaro Cunqueiro,
36312 Vigo, Spain
e-mail: Marcos.Loureiro.Garcia@gmail.com

M.F. Garcia • C. Veiga • E. Paredes • V. Jimenez • F. Calvo-Iglesias • A. Iñiguez
Instituto de Investigación Sanitaria Galicia Sur (IIS Galicia Sur) SERGAS-UVigo, Cardioloxía, Hospital Álvaro Cunqueiro,
36312 Vigo, Spain
e-mail: Maria.Fernandez.Garcia2@sergas.es; Cesar.Veiga.Garcia@sergas.es; Emilio.Paredes.Galan@sergas.es; Victor.Alfonso.Jimenez.Diaz@sergas.es; Francisco.Calvo.Iglesias@sergas.es; Andres.Iniguez.Romo@sergas.es

G. Fdez-Manin
Departamento de matemática aplicada II, Universidade de Vigo, 36310 Vigo, Spain
e-mail: Manin@dmai.uvigo.es

M. Mateos, P. Alonso (eds.), *Computational Mathematics, Numerical Analysis and Applications*, SEMA SIMAI Springer Series 13,
DOI 10.1007/978-3-319-49631-3_15

1 Geometry Modeling

Measurements from 2D TEE are used as input for the equations in [2] in order to obtain the parameters for the geometric model of the aortic valve leaflets. Using those parameters, a closed geometry for the leaflets with different angles between each one, hence different size, is designed. In order to model the aortic sinuses, the procedure described in [3] is followed.

Several scripts were coded implementing the equations in order to obtain an automatized geometric model of the AR using open free software FreeCAD.

The geometry reconstruction starts by representing the main points, lines and curves of the aortic leaflets from the parameters previously calculated. Later, the main lines and curves for the sinuses of Valsalva are created together with surfaces from the exiting lines. A general thickness of 0.5 mm is considered for the leaflets and 2 mm for the exterior part in order to take into account the soft tissues surrounding the AR in vivo. Extrusions will be done together with solid from two surfaces in order to obtain solid elements. B-splines and ruled surfaces were used to create the above mentioned curves and surfaces. Finally, inter-leaflet triangles are reconstructed from the existing sinuses of Valsalva together with the final union of the different parts in one domain. The described process can be found in Fig. 1.

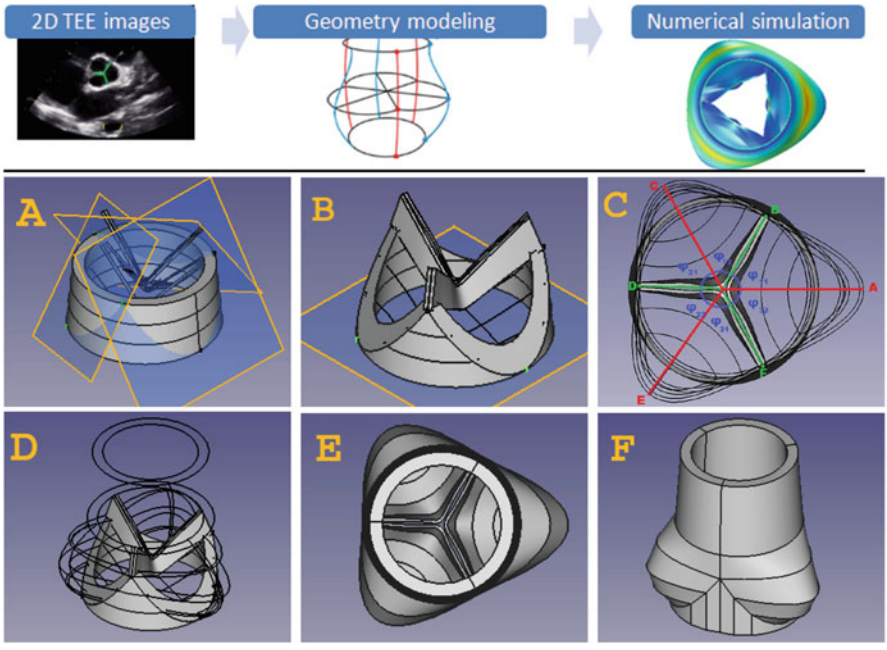


Fig. 1 Aortic root reconstruction process valid for numerical simulation. Measurements are obtained from 2D TEE images. Parameters for the geometric model are deduced using the equations. From (**a**) to (**f**) are shown the main steps followed in order to obtain the geometry

2 Mathematical Model and Numerical Simulation

The modelling of the blood flow behaviour in the AoV has to reflect two types of phenomenon that coexist: the blood flow, with any suitable model to represent its behaviour, and the AR wall displacement. In this section, the biomechanical models that allow to perform numerical simulation, on the geometry previously obtained, are presented. COMSOL MULTIPHYSICS is used and the domain is discretized with P2+P1 elements for the blood flow and P2 for the AR.

2.1 Blood Flow Model

Blood flow modeling is not straightforward due to the non-Newtonian nature of the blood since it is a suspension of cells and particles in plasma, which affect to its modeling in small vessels. In this paper, since the domain can be considered as a big vessel, blood flow is approximated by a Newtonian fluid as in the used bibliography [2–4]. Since the heat transfer between the blood and walls can be considered negligible and there is not a source term, it is an adiabatic flow.

Blood flow properties are $\rho = 1060\,\mathrm{kg/m^3}$ and $\mu = 0.004\,\mathrm{Pa}\cdot\mathrm{s}$. As a 1-way FSI problem is solved, the only information required from the fluid are the tensions, therefore a laminar and incompressible flow is supposed modelled by Navier-Stokes

$$\left.\begin{array}{c} div(\mathbf{v}) = \mathbf{0} \\ \rho\dfrac{\partial \mathbf{v}}{\partial t} + \rho(\mathbf{v}\cdot\nabla)\mathbf{v} - \nabla[-\mathbf{pI} + \boldsymbol{\mu}(\nabla\mathbf{v} + (\nabla\mathbf{v})^{\mathbf{T}})] = \mathbf{0} \end{array}\right\}. \tag{1}$$

2.2 Aortic Root Tissue Model

In order to model the AR tissue a 3D approach is done from the isotropic elastic point of view, with null volume forces. The tissue is under large strains and small deformations modelled as a non linear St. Venant-Kirchooff material [1]. The system evolution is supposed slow, the inertial terms can be despicable then a cuasi-static study is performed with equilibrium equation and material law behaviour given by,

$$\left.\begin{array}{c} div\,\sigma = \mathbf{0} \\ \sigma = \lambda tr(\mathbf{E}(\mathbf{u}))\mathbf{I} + 2\mu\mathbf{E}(\mathbf{u}) \end{array}\right\}, \tag{2}$$

where λ and μ are named Lamé parameters, σ the linearized part of the second Piola Kirchhoff stress tensor and $\mathbf{E}$ the Green St Venant tensor,

$$\mathbf{E} = \frac{1}{2}(\nabla \mathbf{u}^T + \nabla \mathbf{u} + \nabla \mathbf{u}^{\mathbf{T}} \nabla \mathbf{u}). \tag{3}$$

Material parameters are: $E = 10^6$ Pa and $\nu = 0.3$.

2.3 Fixed Mesh FSI

FSI models are the most realistic computational tools, however they are computationally expensive. The fixed mesh FSI, also called 1-way FSI, models situations where the displacements of the solid are assumed to be small enough for the geometry of the fluid domain to be considered as fixed during the interaction. The total force exerted on the solid boundary by the fluid is the negative of the reaction force on the fluid,

$$\mathbf{f} = -\mathbf{n} \cdot \mathbf{T} = \mathbf{n} \cdot \left[p\mathbf{I} + \mu(\nabla \mathbf{v} + (\nabla \mathbf{v})^T) \right] \tag{4}$$

where $\mathbf{T}$ is the fluid stress tensor and $\mathbf{n}$ the outward normal vector.

The one-way coupled models sequentially solve for the fluid flow, compute for each time step the load from Eq. (4), and then apply it in the solution for the solid displacement

$$\sigma_{solid} \cdot \mathbf{n} = \mathbf{T} \cdot \mathbf{n}. \tag{5}$$

2.4 Boundary Conditions

For the fluid domain, pressure boundary conditions are imposed. Obtained values from the free open software CircAdapt (http://www.circadapt.org/) are used in order to obtain the pressure in the left ventricle, $pLV(t)$ and in the aorta, $pAO(t)$. The solid is clamped on the inferior and superior faces and free in the rest.

3 Numerical Results and Conclusions

Several simulations with different meshes were done in order to choose the more appropriate mesh. It was chosen as optimum a 97,424 elements mesh and a simulation during two cardiac cycles was performed. Computations run in a server 2

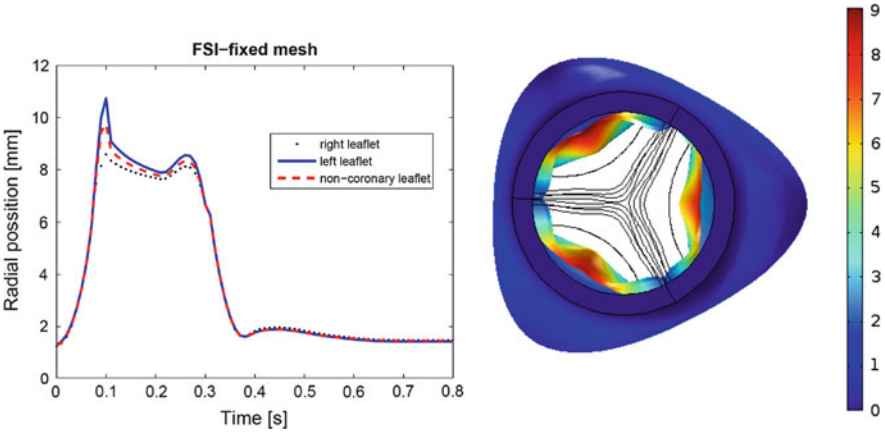


Fig. 2 *Left*: Time-dependent radial position of the nodulus of Arantius. Fixed mesh FSI results are similar to structural and moving mesh FSI analysis [4]. *Right*: total displacement in mm showing the proper opening of the aortic leaflets

Intel Xeon (2.60 GHz) using 25 threats. Computational time was of 8 h for the fluid and 4 h for the solid, hence the computational cost of FSI models is clear.

Volume average velocity and blood flowing through the inlet and outlet boundaries was computed. Blood flow velocity reached values of 1.2 m/s. Average von Mises, total displacement and nodule of Arantius radial position were also computed for three leaflets. Figure 2 compares our results with published ones in bibliography.

Approached results have a similar behaviour with the existing ones in bibliography. They can not be strictly compared because geometries are not exactly the same and pressure boundary conditions, which influence a lot the solution, are different.

The presented parametric model for AR geometric reconstruction from clinic data (including leaflets, inter-leaflet triangles and sinus of Valsalva) is suitable for numerical simulation. It is possible to develop a FSI fixed mesh model of the AR using the presented AR geometry, providing similar results as shown in literature.

The obtained results prove that the use of numerical simulation could be a valid and powerful tool that could be used in the future in clinic applications.

Acknowledgements This work was developed as an Industrial Mathematics master (M2i) thesis in collaboration with Instituto de Investigación Sanitaria Galicia Sur - Cardioloxía. The project was partially financed within the BIOCAPS project FP-7-REGPOT 2012-2013-1.

References

1. Ciarlet, P.: Élasticité Tridimensionnelle. Masson, Paris (1985)
2. Labrosse, M., Beller, C., Robicsek, F., Thubrikar, M.: Geometric modeling of functional trileaflet aortic valves: development and clinical applications. J. Biomech. **39**, 2665–2672 (2006)

3. Morganti, S., Valentini, A., Favalli, V., Serio, A., Gambarin, F.I., Vella, D., Mazzocchi, L., Massetti, M., Auricchio, F., Arbustini, E.: Aortic root 3D parametric morphological model from 2D-echo images. Comput. Biol. Med. **43**, 2196–2204 (2013)
4. Sturla, F., Votta, E., Stevanella, M., Conti, C., Redaelli, A.: Impact of modeling fluid-structure iteraction in the computational analysis of aortic root biomechanics. Med. Eng. Phys. **35**, 1721–1730 (2013)