MONOGRAFÍAS
DE LA
REAL ACADEMIA
DE CIENCIAS
Exactas
Físicas
Químicas y
Naturales
DE
ZARAGOZA

Nº 33

Proceedings of the MCalv6(5) Conference

A. Elipe, J.I. Montijano and L. Rández (Editors)

2010

Index

Preface

This *Monografía de la Real Academia de Ciencias de Zaragoza* collects the proceedings of the conference MCalv6(5) that took place in Zaragoza, September 7-9, 2009. (`http://iuma.unizar.es/mcal65/`).

The conference was organized to celebrate the 65th birthday of Manuel Calvo. The idea of having such a conference come from his Ph.D. students, who wanted to share Manuel's birthday with his many friends and colleagues.

Firstly, we would like to express our thanks to all speakers and contributors in the conference. We must say that when they were contacted to participate in this event, all of them immediately accepted. Manuel Calvo is, indeed, very popular among their colleagues.

The conference took place at the Instituto Universitario de Matemáticas y Aplicaciones (IUMA) of the University of Zaragoza and included many high quality talks on diverse topics, with speakers coming from Berlin, Leiden, Iowa, Gent, Florence, Valladolid, La Laguna, Alicante and Zaragoza. All of them, researchers of the first international level, are good friends of Manuel. More than 50 people attended the conference and many others that could not participate, transmitted us their congratulations to Manuel. A warm and friendly atmosphere characterized the meeting.

With this conference we wanted to acknowledge and to thank Manuel for all the excellent work he has being doing during the many years of his career. Manuel belongs to a group of people that had to start practically from scratch. Indeed, by the early seventies, there was no much research on Applied Mathematics in Spain, and in particular in Zaragoza. Manuel, was among the first ones in going abroad to study new theories, develop techniques and skills on Numerical Analysis, and then, coming back to create a new research group that soon acquired an international reputation, and all that with almost no means, neither financial support nor infrastructure.

The conference was organized not only because Manuel has done a good job in Applied Mathematics. With no doubt, the most important reason to organize the Conference in his honor is their human qualities. He is good teacher, generous, meticulous, careful, precise and, obviously, a good mathematician. Particularly, we would mention as a good quality his commitment to everything he does. Whatever he does, teaching, research or

administration, he has to do it well. A characteristic point of Manuel that we would like to emphasize is his curiosity. He is always willing to know new ideas, concepts, etc. He has passion for knowing and understanding new things. At any conference, after a talk when the chairman says "any question", we think "besides the ones from Manuel". But you can be sure that this is not due to vanity or because he likes to be the center of attention. It is just curiosity.

Manolo is a great master and also a good friend. We thank him for everything, and we are sure that all his colleagues, students and friends want to join us again in wishing him a long and fruitful scientific and personal life.

The Editors

Antonio Elipe
Juan Ignacio Montijano
Luis Rández

MCalv6(5)

Conference in Honour of the 65th Birthday of Prof. Manuel Calvo

Zaragoza, September 7-8, 2009

Main Speakers

L. Brugnano
J. M. Ferrándiz
M. Gasca
P. González-Vera
R. Grigorieff
L. Jay
J. Sanz-Serna
M. Spijker
G. Vanden Berghe

Organizing Committee
A. Elipe
J.I. Montijano
L. Rández

http://iuma.unizar.es/mcal65/
Instituto Universitario de Matemáticas y Aplicaciones

Universidad de Zaragoza

Patrocina: GOBIERNO DE ARAGON

Monografías de la Real Academia de Ciencias de Zaragoza **33**, 1–19, (2010).

Qualocation for periodic pseudodifferential operators: additional order convergence, an overview

R.D. Grigorieff

Institut für Mathematik, Technische Universität Berlin

Straße des 17. Juni 135, 10623 Berlin, Germany

Abstract

The aim of this note is to report on some new additional order convergence results for the qualocation method applied to periodic pseudodifferential operators using splines from $S_h^{r,M}$ as trial and from $S_h^{r',M}$ as test space. Here $S_h^{r,M}$ denotes the space of 1-periodic splines of order r and knot multiplicity $M \leq r$ on an equidistant mesh with mesh-size h.

1 Introduction

Additional order convergence (in brief: AOC) in discretization methods has attracted the interest of many mathematicians over decades. The AOC has the character of a gift. With a few changes in the original method which do not increase the overall numerical costs significantly a more accurate approximation is obtained. To be a little bit more detailed, in most cases the accuracy of a method will be described by a discretization parameter, say h, tending to zero with increasing approximation quality. The error between the exact solution u and the numerical solution u_h is measured, apart from a h-independent constant, by some power of h as $h \to 0$. One speaks of an AOC if another approximation $\tilde{u}_h$ can be calculated with comparably the same complexity or if by measuring the error in more subtile norms a higher convergence order is exhibited.

Perhaps the oldest and structural simplest AOC is provided by Richardson extrapolation. But as the mother of all AOC results the DeBoor & Swartz paper [1] on collocation at Gaussian points can be considered. And, naturally, a wide variety of AOC results can be found in the context of finite element methods (see for example the Lecture Note [20] and the overview article [10]).

In this paper we concentrate on AOC obtained for spline qualocation methods applied to elliptic periodic pseudodifferential operators (in brief: ψdos). Qualocation, introduced

by Sloan [15], is a Petrov-Galerkin method with quadrature as a compromise between the (full) Petrov-Galerkin and the collocation method: on the one hand it discretizes the inner product in the Petrov-Galerkin method making the numerical implementation easier, on the other hand it can use more mesh-points than in the collocation method thereby stabilizing it. In accordance, the word "qualocation" means quadrature modified collocation.

From the different principles invented for obtaining AOC we use here the principles of parameter selection for cancelling the leading error term and duality combined with negative norms.

Already when introducing the qualocation method in [15] and shortly after when analyzing it in greater generality in [3] AOC was in the center of interest. The analysis of qualocation was firstly developed for constant coefficient operators and smoothest splines in [15], [3], [18], where straight Fourier analysis could be applied, and later generalized to variable coefficient operators and multiple knot splines in [11], [19], [8] requiring a more sophisticated analysis.

In this paper we report on some recent AOC results obtained by the author and relate them to corresponding earlier results. In the first part the qualocation method is introduced and the principle convergence result is given. The approximation power of multiple knot splines in the Sobolev spaces H^s for $s \in \mathbb{R}$, which serve as trial and test spaces, are reviewed. In the last section it is shown that the conditions for additional order convergence from the third section hold true if the basic quadrature rule is symmetric and satisfies certain exactness properties thereby extending conditions given in [19].

2 The given problem

The given problem is of the form $Lu = f$, where L is a periodic ψdo and u and f are functions in certain Sobolev spaces. In this section we provide the definition of the function spaces and of ψdos together with some of their properties.

2.1 Periodic ψdos in the spaces $H^s(\mathbb{T})$

Denote by

$$\hat{v}(n) = \int_{\mathbb{T}} v(x) e^{-i2\pi nx}\, dx \quad \text{for } n \in \mathbb{Z}$$

the n-th complex Fourier coefficient of a 1-periodic distribution v and by $\mathbb{T} := \mathbb{R}\backslash\mathbb{Z}$ the one-dimensional torus of length 1. Then the ψdo L is defined by

$$L = L_0 + L_1, \tag{1}$$

where

$$L_0 v(x) := \sum_{n=-\infty}^{\infty} \sigma_0(x,n)\hat{v}(n)e^{i2\pi nx} \quad \text{for } x \in \mathbb{T}. \tag{2}$$

The symbol σ_0 has the form

$$\sigma_0(x,\xi) := a^+(x)|\xi|^\beta + a^-(x)\text{sign}(\xi)|\xi|^\beta \quad \text{for } x \in \mathbb{T} \text{ and } 0 \neq \xi \in \mathbb{R} \tag{3}$$

with coefficients a^+ and a^- in $C^\infty(\mathbb{T})$, where $\beta \in \mathbb{R}$ is the order of L_0. We assume σ_0 to be normalised by $\sigma_0(x,0) = 1$ for $x \in \mathbb{T}$. If a^- or a^+ vanishes the symbol σ_0 and the operator L_0 are said to be even or odd, respectively. L is assumed to be elliptic, i.e. $\sigma_0(x,\xi) \neq 0$ for $x \in \mathbb{T}$ and $|\xi| = 1$, and to have index $\kappa = 0$, where

$$\kappa := \frac{1}{2\pi}\left[\arg\frac{a^+(x)+a^-(x)}{a^+(x)-a^-(x)}\right]_0^1$$

is the winding number of the closed curve $(a^+ + a^-)/(a^+ - a^-)$ in the complex plane. It is known that then $L_0 : H^s \to H^{s-\beta}$ is a Fredholm operator with index 0 for all $s \in \mathbb{R}$, where $H^s \equiv H^s(\mathbb{T})$ is the usual Sobolev space of periodic distributions v equipped with the norm

$$\|v\|_s := \left(\sum_{n=-\infty}^{\infty} \langle n\rangle^{2s}\,|\hat{v}(n)|^2\right)^{1/2} \quad \text{with } \langle n\rangle := \begin{cases} 1 & \text{if } n = 0, \\ |n| & \text{if } n \neq 0. \end{cases} \tag{4}$$

It is at least assumed that L_1 maps $H^s \to H^{s-\beta+\delta}$ for some $\delta > 0$ and all $s \in \mathbb{R}$ and hence L is also Fredholm with index 0.

2.2 Examples for ψdos

Since the topic of the proceedings are mainly differential equations and not primarily ψdos it may be convenient for the readers to see two standard examples for the latter which we borrow from [16].

As first example, consider the boundary value problem

$$U_{XX} + U_{YY} = 0 \text{ in } \Omega \subset \mathbb{R}^2,\ \Omega \text{ bounded}, \quad U = F \text{ on } \Gamma := \partial\Omega,$$

where Γ is a smooth curve. One method to solve this problem is to express U as a single-layer potential with unknown (charge-)density W:

$$U(X) = \mathbb{V}W(X) := \frac{1}{\pi}\int_\Gamma \log\frac{1}{|X-Y|}W(Y)\,dY \quad \text{for } X \in \overline{\Omega}.$$

Parametrize Γ in the form $X = \gamma(x)$ for $x \in [0,1]$ to obtain

$$(\mathbb{V}W)(\gamma(x)) = \frac{1}{\pi}\int_0^1 \log\frac{1}{|\gamma(x)-\gamma(y)|}W(\gamma(y))|\gamma'(y)|\,dy$$

$$
\begin{aligned}
&= 2\int_0^1 \log\frac{1}{|2r\sin\pi(x-y)|}\frac{1}{2\pi}W(\gamma(y))|\gamma'(y)|\,dy \\
&\quad + 2\int_0^1 \log\frac{|2r\sin\pi(x-y)|}{|\gamma(x)-\gamma(y)|}\frac{1}{2\pi}W(\gamma(y))|\gamma'(y)|\,dy \\
&=: V_0u(x) + Ku(x),
\end{aligned}
$$

where

$$
u(x) := \frac{1}{2\pi}W(\gamma(x))|\gamma'(x)|. \tag{5}
$$

V_0u is the single-layer potential for a circle of radius r. For $\Phi_n := \exp(i2\pi nx)$ it can explicitely calculated that

$$
V_0\Phi_n = \begin{cases} \frac{1}{|n|}\Phi_n & \text{for } n \neq 0, \\ 1 & \text{for } n = 0. \end{cases}
$$

Thus, applying V_0 to a 1-periodic distribution

$$
u = \sum_{n\in\mathbb{Z}} \hat{u}(n)\Phi_n \in H^s
$$

yields the representation of V_0 as a ψdo,

$$
V_0u = \hat{u}(0) + \sum_{n\neq 0}\frac{1}{|n|}\hat{u}(n)\Phi_n.
$$

The symbol of V_0 is even and given by

$$
\sigma_0(x,\xi) = \frac{1}{|\xi|} \quad \text{for } \xi \neq 0,
$$

the order is $\beta = -1$. Evidently, V_0 maps $H^s \to H^{s+1} \equiv H^{s-\beta}$ boundedly for $s \in \mathbb{R}$.

More briefly, as second example the Hilbert transform is presented:

$$
\mathbb{S}U(X) := \frac{1}{i\pi}\int_\Gamma \frac{1}{Y-X}U(Y)\,dY,
$$

which can be transformed as in the first example to coordinates (x, y) yielding the principal part

$$
S_0u(x) := 2\int_0^1 \frac{\exp(i2\pi y)}{\exp(i2\pi y) - \exp(i2\pi x)}u(y)\,dy.
$$

The order of S_0 is $\beta = 0$, the symbol is odd and given by

$$
\sigma_0(x,\xi) = \operatorname{sign}(\xi) \quad \text{for } \xi \neq 0.
$$

3 The qualocation method

We consider the discretisation of the given problem by qualocation using splines with multiple knots on equidistant meshes as test and trial spaces. Let r, M, N with $1 \le M \le r$ be positive integers. We define the set of knots

$$\pi_h := \{x_j = jh, j = 0, \dots, N-1\}, \quad h \in \mathcal{H} := \{1/N, N \in \mathbb{N}\},$$

and denote by $S_h^{r,M}$ the space of 1-periodic splines of order r with M-fold breakpoints in π_h. $S_h^{r,M}$ is a subspace of C^{r-M-1} of dimension MN, where $C^k = C^k(\mathbb{T})$ is the space of 1-periodic k times continuously differentiable functions (with C^{-1} meaning piecewise continuity with jumps only at the knots in π_h). By $\mathcal{H}_1$ we denote a final section of the null-sequence of stepsizes $\mathcal{H}$, not necessarily the same at different occurencies.

Qualocation is based on a composite quadrature rule

$$Q_N f = h \sum_{k=0}^{N-1} \sum_{j=1}^{J} \omega_j f(x_{k,j}), \quad x_{k,j} := x_k + h\xi_j,$$

derived from the basic quadrature formula

$$Qf = \sum_{j=1}^{J} \omega_j f(\xi_j),$$

where the quadrature points $\{\xi_j\}$ and weights $\{\omega_j\}$ satisfy

$$0 \le \xi_1 < \xi_2 < \cdots < \xi_J < 1, \quad J \ge M, \quad \sum_{j=1}^{J} \omega_j = 1, \quad \omega_j > 0. \tag{6}$$

Associated with the quadrature rule we define an inner product

$$(v_h, w_h)_h := Q_N(v_h \overline{w}_h) \tag{7}$$

on the linear space W_h of mesh-functions v_h and w_h, which are functions on the set of mesh-points

$$\pi_h' := \{x_{k,j},\, k = 0, \dots, N-1,\, j = 1, \dots, J\}.$$

The inner product in (7) can be thought of as an approximation to

$$(v, w)_0 := \int_0^1 v(x)\overline{w}(x)\, dx \quad \text{for } v, w \in L^2(\mathbb{T}).$$

In the next section we give conditions for $(\cdot,\cdot)_h$ to be an inner product on $S_h^{r,M}$.

We choose splines of order r as trial space and splines of a possibly different order r' as test space. The qualocation method for approximately solving the equation $Lu = f$ is to find $u_h \in S_h^{r,M}$ such that

$$(Lu_h, z_h)_h = (f, z_h)_h \quad \text{for all } z_h \in S_h^{r',M}. \tag{8}$$

4 The spline space $S_h^{r,M}$ in qualocation

For the operator formulation of the qualocation equations the so-called qualocation projection $R_h : W_h \to S_h^{r,M}$ is needed which is defined by

$$(R_h v_h, \psi_h)_h = (v_h, \psi_h)_h \quad \text{for } \psi_h \in S_h^{r,M}.$$

It is not trivial that R_h is well-defined, i.e. that $(\cdot,\cdot)_h$ is an inner product on $S_h^{r,M}$, and in the following subsection we give criteria for this to be the case. Then in the next subsection we focus on the approximation power of R_h.

4.1 The qualocation projection R_h

The analysis of the approximation power and, more general, of the whole topic depends on Fourier techniques. An important role in this analysis plays a suitable spline basis. In the case of $S_h^{r,1}$, the space of smoothest splines of order r, a basis was found by Chandler & Sloan [3]:

$$\psi_\mu(x) := \sum_{j=1}^{N} \exp(i2\pi\mu x) b_j(x) \quad \text{for } \mu \in \Lambda_h := \left(-\frac{N}{2}, \frac{N}{2} \right] \cap \mathbb{Z},$$

where $\{b_j\}$ is the B-spline basis in $S_h^{r,1}$. A nice thing about the basis is that the qualocation equations for the principal part L_0 become diagonal if L_0 has constant coefficients.

The situation with M-fold knots, $M > 1$, is more delicate. In their collocation analysis McLean & Prößdorf [11] used the following characterization of splines.

Lemma 1 *$v \in S_h^{r,M}$ iff there exist trigonometric polynomials a_j such that*

$$m^r \hat{v}(m) = \sum_{j=0}^{M-1} m^j a_j(mh) \quad \textit{for } m \in \mathbb{Z}.$$

Working with this characterization makes the analysis uncomfortable. It was a step forward when a basis in $S_h^{r,M}$ was found in [7] which extends the one in [3]. Define

$$\begin{aligned}
\tilde{\Delta}_k(\xi, y) &:= \sum_{\ell \neq 0} \frac{\ell^{k-1}}{(y+\ell)^r} \Phi_\ell(\xi) \quad \text{for } |y| \le \frac{1}{2} \text{ and } \xi \in \mathbb{R}, \\
\Phi_\ell(\xi) &:= \exp(i2\pi\ell\xi) \quad \text{for } \ell \in \mathbb{Z} \text{ and } \xi \in \mathbb{R}, \\
\Delta_1(\xi, y) &:= 1 + y^r \tilde{\Delta}_1(\xi, y), \quad \Delta_k(\xi, y) := \tilde{\Delta}_k(\xi, y) \quad \text{for } k = 2, \ldots, M, \\
\psi_{k,\mu}(x) &:= \Phi_\mu(x) \Delta_k\Big(Nx, \frac{\mu}{N}\Big) \quad \text{for } k = 1, \ldots, M \text{ and } \mu \in \Lambda_h.
\end{aligned}$$

Then $\{\psi_{k,\mu}\}$ is a basis in $S_h^{r,M}$. The use of this basis makes the qualocation equations for the principal part L_0 block diagonal with blocks of size M if the coefficients are constant.

With the aid of $\{\psi_{k,\mu}\}$ it can be characterized whether $(\cdot,\cdot)_h$ is definite on $S_h^{r,M}$.

Proposition 1 *$(\cdot,\cdot)_h$ defines an inner product on $S_h^{r,M}$ iff the functions $\{\Delta_k(\cdot,y), k = 1,\dots,M\}$ restricted to the quadrature points $\{\xi_j, k = 1,\dots,J\}$ are linearly independent for $y = \mu/N$ and $\mu \in \Lambda_h$.*

We say that Condition (R) is satisfied if the condition in Proposition 1 holds for all $|y \le 1/2|$. It is known that Condition (R) holds in the following cases.

- [3]: $M = 1$

  (R) holds unless $J = 1$ and $\begin{cases} \xi_1 = \frac{1}{2} & \text{if } r \text{ is even,} \\ \xi_1 = 0 & \text{if } r \text{ is odd.} \end{cases}$

- [11]: $J = M = 2$

  $\xi_1 = 0, \xi_2 = \frac{1}{2}$: (R) holds iff r is odd,

  $\xi_1 = \varepsilon, \xi_2 = 1 - \varepsilon$ with $\varepsilon \in (0, \frac{1}{2})$: (R) fails if r is odd.

- [7], [12]: $J = M = 2$

  $\xi_1 = \varepsilon, \xi_2 = 1 - \varepsilon$ with $\varepsilon \in (0, \frac{1}{2})$: (R) holds iff r is even.

- [6]: $J = M = 3$

  $\xi_1 = 0, \xi_2 = \varepsilon, \xi_3 = 1 - \varepsilon$ with $\varepsilon \in (0, \frac{1}{2})$: (R) holds iff r is even,

  $\xi_1 = \varepsilon, \xi_2 = \frac{1}{2}, \xi_3 = 1 - \varepsilon$ with $\varepsilon \in (0, \frac{1}{2})$: (R) holds iff r is odd,

- [6]: for all J, M:

  (R) holds if $J > M$.

In the remaining part of the paper it is assumed that Condition (R) and Condition (R′) (this is Condition (R) with r replaced by r') hold.

4.2 Approximation power of R_h

The approximation power of the qualocation projection R_h proved in [7] is the content of the next proposition.

Proposition 2 *Let $0 \le s < r - M + \frac{1}{2}, s \le t \le r, t > \frac{1}{2}$. Then*

$$\|R_h v - v\|_s \le C h^{t-s} \|v\|_t \quad \text{for } v \in H^t.$$

An explanation for the given range of indices may be helpful.

- $s < r - M + \frac{1}{2}$: the limited smoothness of the spline $\psi \in S_h^{r,M}$ implies $\psi \in H^s$ for $s < r - M + \frac{1}{2}$ only,

- $t \le r$: the maximal t allowed, giving the highest error order, is determined by the order of the splines,

- $t > \frac{1}{2}$: the definition of $R_h v$ requires pointwise evaluation of v which is not well-defined for $t \le \frac{1}{2}$ since then $H^t \not\hookrightarrow C(\mathbb{T})$.

5 Principle error estimate

The application of L_0 to the spline basis leads to the functions (see [8])

$$\tilde{\Omega}_k(\xi, y; x) := \sum_{\ell \neq 0} \sigma_0(x, y+\ell) \frac{\ell^{k-1}}{(y+\ell)^r} \Phi_\ell(\xi) \quad \text{for } k = 1, \ldots, M,$$
$$\Omega_1(\xi, y; x) := 1 + (\sigma_0(x, y))^{-1} y^r \tilde{\Omega}_1(\xi, y; x) \quad \text{for } y \neq 0,$$
$$\Omega_k(\xi, y; x) := \tilde{\Omega}_k(\xi, y; x) \quad \text{for } k = 2, \ldots, M.$$

We omit the argument x if L_0 has constant coefficients. The stability of the qualocation method is connected with the ellipticity of the numerical symbol $D(y; x)$, which is a $M \times M$-matrix with elements

$$[D(y; x)]_{k,\ell} := Q\big(\Omega_\ell(\cdot, y; x), \Delta'_k(\cdot, y)\big), \quad Q(v, w) := \sum_{j=1}^{J} \omega_j (v\overline{w})(\xi_j).$$

The numerical symbol is encountered as the coefficient matrix in the linear system of qualocation equations if L_0 has constant coefficients. Ellipticity of D means that $D(y; x)$ is invertible for $|y| \leq 1/2$ and $x \in \mathbb{T}$.

Theorem 1 *Let L be elliptic and injective. Assume that D is elliptic and that*

$$\beta + M < r,\ s < r - M + \frac{1}{2},\ \beta + \frac{1}{2} < t,\ \beta \leq s \leq t \leq r.$$

Then the qualocation equations have a unique solution u_h for $h \in \mathcal{H}_1$ satisfying

$$\|u - u_h\|_s \leq C h^{t-s} \|u\|_t \quad \text{if } u \in H^t.$$

The condition $\beta + M < r$ ensures the absolute convergence of the series defining $\tilde{\Omega}_k$.

The theorem has been proved under varying assumptions.

- [15]: constant coefficients, $L \equiv L_0$, even symbol, $M = 1$,
- [3]: constant coefficients, even or odd symbol, $M = 1$,
- [18]: constant coefficients, symbol that may be neither even nor odd, $M = 1$,
- [11]: variable coefficients, collocation, multiple knots,
- [19]: variable coefficients, strongly and oddly elliptic L, $M = 1$,
- [8]: variable coefficients, multiple knots.

The proof for variable coefficient operators uses a localization technique. Such techniques are known from PDEs but are considerably more intrigued to apply for integral operators which themselves are non-local. The underlying abstract result is due to Prößdorf [13]. Basic tools for applying Prößdorf's result are the following superapproximation from [4] and commutator property from [8]. Both provide an AOC with order 1.

Proposition 3 *Let $g \in C^r(\mathbb{T})$ and $M < r, 0 \le s < r - M + \frac{1}{2}, t \le r - M$. Then*

$$\|(I - R_h)(gv_h)\|_s \le Ch^{1+t-s}\|g'\|_{r-1,\infty}\|v_h\|_t \quad \textit{for } v_h \in S_h^{r,M}.$$

Proposition 4 *Let $g \in C^r(\mathbb{T})$ and $M < r$, $0 \le s < r - M + \frac{1}{2}, \frac{1}{2} < t \le r$. Then*

$$\|R_h g(I - R_h)v\|_s \le Ch^{1+t-s}\|g'\|_{r-1,\infty}\|v\|_t \quad \textit{for } v \in H^t.$$

6 Additional order of convergence

The highest error order in the principle convergence theorem is

$$\|u - u_h\|_\beta \le Ch^{r-\beta}\|u\|_r \quad \text{if } u \in H^r.$$

For example, if L_0 is the single-layer potential, where $\beta = -1$, and choosing continuous linear splines as trial space, this means order $r = 2$ and $M = 1$, then

$$\|u - u_h\|_{-1} \le Ch^3\|u\|_2 \quad \text{if } u \in H^2.$$

This convergence order is disappointing when compared to the highest order negative norm error bound for the Galerkin solution u_h^G (see [9]),

$$\|u - u_h^G\|_{-3} \le Ch^5\|u\|_2 \quad \text{if } u \in H^2.$$

For the qualocation method progress to catch up with the order 5 was made by Sloan [15] who showed that with specially designed quadrature rules one can obtain the same error order,

$$\|u - u_h\|_{-3} \le Ch^5\|u\|_4 \quad \text{if } u \in H^4. \tag{9}$$

The estimate requires the higher regularity $u \in H^4$ compared to $u \in H^2$ for the Galerkin method. This disadvantage was overcome in [17] with the tolerant version of qualocation, where the inner product on the right-hand side is evaluated exactly.

6.1 Constant coefficient L_0

The AOC result (9) is proved by the principle of parameter selection for cancelling the leading error term. A key idea of the proof is the following. In [3] the asymptotic error expansion for the operator $L = L_\beta^+$ or $L = L_\beta^-$ is obtained by Fourier transform in the form

$$\hat{u}(\mu) - \hat{u}_h(\mu) = D\Big(\frac{\mu}{N}\Big)^{-1} E\Big(\frac{\mu}{N}\Big)\hat{u}(\mu) + \text{higher order terms in } \Big(\frac{\mu}{N}\Big)$$

for $\mu \in \Lambda_h \backslash \{0\}$, where the symbol of L_α^+ and L_α^- is $\sigma_0 = |\xi|^\alpha$ and $\sigma_0 = \operatorname{sign}\xi\,|\xi|^\alpha$, respectively, and

$$E(y) := \sum_{j=1}^{J} \omega_j \big(\Omega_1(\xi_j, y) - 1\big)\overline{\Delta_1'(\xi_j, y)} \quad \text{for } |y| \le \frac{1}{2}.$$

Note that due to the numerical ellipticity $\left|D(\frac{\mu}{N})^{-1}\right| \le C$. For any choice of the quadrature rule the function E behaves like

$$E(y) = \mathcal{O}(|y|^{r-\beta}) \quad \text{as } y \to 0.$$

The qualocation method (in the case $M = 1$ considered here) is said to have additional order $b > 0$ (see [15]) if

$$E(y) = \mathcal{O}(|y|^{r-\beta+b}) \quad \text{as } y \to 0. \tag{10}$$

The functions Ω_1 and Δ_1' are given if L_0 and the spline spaces are fixed. One can try to obtain an additional order by selecting the quadrature rule appropriately. Sloan showed that the choice

$$\xi_1 = 0,\ \xi_2 = \frac{1}{2},\ \omega_1 = \frac{3}{7},\ \omega_1 = \frac{4}{7}$$

combined with linear continuous splines as test and trial space gives $b = 2$ for the single-layer equation. Note that the quadrature points are from the trapezoidal rule but not the weights.

In the case of multiple knot splines the condition for additional order $b > 0$ from [5] is

$$\sum_{k=1}^{M} D(y)^{-1}_{1,k} Q\big(\tilde{\Omega}_1(\cdot, y), \Delta_k'(\cdot, y)\big) = \mathcal{O}(|y|^b) \quad \text{as } y \to 0, \tag{11}$$

which is in the case $M = 1$ and $b \le r'$ equivalent to Condition (10) and in the case $J = M$ to Condition (2.12) in [11].

The general AOC result for constant coefficient L_0 is stated in the following theorem.

Theorem 2 *Let $L = L_0 + K : H^\beta \to H^0$ be elliptic and injective, where L_0 has constant coefficients and K maps $H^q \to H^{q-\beta+b}$ boundedly for $q \in \mathbb{R}$. Assume that D is elliptic, that Condition (11) holds and that*

$$\beta + M < r,\ M \le r',\ s < r - M + \frac{1}{2},\ s \le t \le r,\ \beta - b \le s \le \beta < t - \frac{1}{2}.$$

Then the qualocation equations have a unique solution u_h for $h \in \mathcal{H}_1$ satisfying

$$\|u - u_h\|_s \le C h^{t-s} \|u\|_{t-s+\beta} \quad \text{if } u \in H^{t-s+\beta}.$$

The theorem has been proved in varying settings.

- [15]: even symbol, $L \equiv L_0$, $M = 1$, quadrature rule of Simpson type,
- [3]: even or odd symbol, $M = 1$,
- [11]: collocation, i.e. $M = J$, multiple knots,
- [18]: qualocation, symbol may be neither even nor odd, $M = 1$,
- [5]: qualocation, multiple knots.

6.2 Variable coefficients

The analysis for variable coefficient operators L_0 is technically considerably more involved than for constant coefficients. In the case of variable coefficients AOC has been proved for smoothest splines in [19]. The authors assume that $L = L_0 + L_1 + K$, where

$$L_1 = \sum_{i=1}^{b-1} \Big(a_i^+(x) L_{\beta-i}^+ + a_i^-(x) L_{\beta-i}^- \Big), \quad K : H^q \to H^{q-\beta+b+\nu} \text{ boundedly} \tag{12}$$

for $q \in \mathbb{R}$ and some $\nu > 1/2$. The basic assumptions in [19, Th. 4 and Th. 5] are conditions for the quadrature rule, which is supposed to be symmetric and to integrate certain functions exactly (see Lemma 4). In [5] multiple knot splines are considered for the same class (12) of ψdos and the following conditions for AOC in the spirit of [3] are given:

$$\big| Q\big(\tilde{\Omega}_k(\cdot, y), 1\big) \big| \le C|y|^b \text{ as } y \to 0 \text{ for } k = 1, \dots, M \tag{13}$$

with $b \in \mathbb{N}$ satisfying $\beta - s \le b \le \min(r', r - \beta)$, where $\tilde{\Omega}_k$ has to be taken for L_β^+ and L_β^-; additionally,

$$\big| Q\big(1, \tilde{\Delta}_1'(\cdot, y)\big) \big| \le C|y|^{r-\beta+b-r'} \text{ as } y \to 0, \tag{14}$$

$$\big| Q\big(1, \tilde{\Delta}_k'(\cdot, y)\big) \big| \le C|y|^{r-\beta} \text{ as } y \to 0 \text{ for } k = 2, \dots, M. \tag{15}$$

It may be helpful for interpreting the conditions in (13) - (15) to hint to the fact that $Q\big(\tilde{\Omega}_k(\cdot, y), 1\big)$ can be considered as the result of the quadrature rule Q applied to the integral $(\tilde{\Omega}_k(\cdot, y), 1)_0$, which vanishes as is immediately seen from the definition of $\tilde{\Omega}_k$. Also $(1, \tilde{\Delta}_k'(\cdot, y))_0 = 0$.

Theorem 3 *Let $L = L_0 + L_1 + K$ satisfy (12) with $\nu = 0$ and be elliptic and injective. Assume also that D is elliptic, that Conditions (13) - (15) hold and that*

$$\beta + M < r,\ M \le r',\ s < r - M + \frac{1}{2},\ s \le t \le r,\ s \le \beta < t - \frac{1}{2},$$

$$\beta - s \le b \le \min(r', r - \beta).$$

Then the qualocation equations have a unique solution u_h for $h \in \mathcal{H}_1$ satisfying

$$\|u - u_h\|_s \le Ch^{t-s}\|u\|_{t-s+\beta} \quad \textit{if } u \in H^{t-s+\beta}.$$

The operator L_1 in (12) is said to be even (odd) if $a_i^- = 0$ ($a_i^+ = 0$) for $i = 1, \ldots, b$.

Remark 1 *If L_0 and L_1 are both even or odd then it is sufficient that the qualocation method has strong additional order b of convergence to require (13) for L_β^+ or L_β^- only, respectively.*

6.3 Negative norm estimates are useful

A word on the significance of the negative norm estimates in the context of boundary integral equations may be in order (see [16]). As described in Subsection 2.2, for a given point $X_0 \in \Omega$ the solution of the boundary value problem

$$U_{XX} + U_{YY} = 0 \text{ in } \Omega \subset \mathbb{R}^2, \quad U = F \text{ on } \Gamma,$$

can be written in the form

$$U(X_0) = \frac{1}{\pi}\int_0^1 \frac{1}{\log|X_0 - \gamma(y)|} u(y)\, dy,$$

where u is from (5). The approximation

$$U_h(X_0) = \frac{1}{\pi}\int_0^1 \frac{1}{\log|X_0 - \gamma(y)|} u_h(y)\, dy$$

satisfies the error bound

$$\begin{aligned} |U(X_0) - U_h(X_0)| &= \frac{1}{\pi}(\log|X_0 - \gamma|, u - u_h)_0 \\ &\le C\|\log|X_0 - \gamma|\|_t \|u - u_h\|_{-t} \quad \text{for } t \in \mathbb{R}. \end{aligned}$$

Thus, the higher the order of $\|u - u_h\|_{-t}$ the higher the error order $(U - U_h)(X_0)$ since $\log|X_0 - \gamma| \in H^t$ for all $t \in \mathbb{R}$.

7 Symmetric quadrature rules

A basic quadrature rule Q satisfying the condition that if $\xi \in (0, \frac{1}{2})$ is a quadrature point then so is $(1 - \xi)$ with the same weight ω is called symmetric. In the case of smoothest splines in [18] and [19] exactness conditions for symmetric quadrature rules are given for AOC to hold. In this section we extend these conditions to multiple knot splines. We always assume that

$$\beta + M < r \quad \text{and} \quad M \le r'.$$

We need the following functions G_α for $\alpha > 0$ and $\xi \in (0,1)$ which have been studied in [2]:

$$G_\alpha(\xi) := 2\sum_{\ell=1}^{\infty} \frac{1}{\ell^\alpha} \cos 2\pi\ell\xi.$$

For symmetric quadrature rules the Conditions (13) - (15) can be further elaborated. Recall that

$$\tilde\Omega_k(\xi, y) = \sum_{\ell\neq 0} \sigma_0(y+\ell) \frac{\ell^{k-1}}{(y+\ell)^r} \Phi_\ell(\xi) \quad \text{for } k = 1, \dots, M,$$

where we consider these functions for the operators L_β^+ and L_β^-. As shown in [3], $\tilde\Omega_k(\xi,\cdot)$ has a Taylor expansion with respect to $y = 0$, where the coefficients for the real part are easily determined to be equal to

$$\begin{aligned} c_{k,m}(\xi) := \frac{1}{m!}\mathcal{R}e\frac{\partial^m \tilde\Omega_k(\xi,0)}{\partial y^m} &= \tbinom{\beta-r}{m} \sum_{\ell\neq 0} \sigma_0(\ell)\ell^{k-1-r-m} \cos 2\pi\ell\xi \\ &= \tbinom{\beta-r}{m} \sum_{\ell>0} \ell^{k-1-r-m}(\sigma_0(\ell) + \sigma_0(-\ell)(-1)^{k-1-r-m}) \cos 2\pi\ell\xi \end{aligned}$$

for $m \in \mathbb{N}_0$ and $\sigma_0(\xi) = |\xi|^\beta$ or $\sigma_0(\xi) = \operatorname{sign}\xi\,|\xi|^\beta$. It follows that

$$c_{k,m} = \begin{cases} G_{r-\beta-k+1+m} & \text{if } \sigma_0 \text{ and } r-k+1+m \text{ have like parity,} \\ 0 & \text{otherwise.} \end{cases} \tag{16}$$

In the next two lemmas we give the Conditions (13) - (15) another form.

Lemma 2 *If L_0 is even (odd) then Condition (13) is equivalent to*

$$\sum_{j=1}^{J} \omega_j G_{r-\beta+\ell}(\xi_j) = 0 \quad \textit{for even (odd) } \ell \in [-M+1, b-1] \tag{17}$$

if σ_0 and r have like (opposite) parity. If L_0 is neither even nor odd then (13) is equivalent to the equation in (17) for all even and odd $\ell \in [-M+1, b-1]$.

P r o o f . Since $\mathcal{I}m\,\tilde\Omega_k(1-\xi, y) = -\mathcal{I}m\,\tilde\Omega_k(\xi, y)$ and $\mathcal{I}m\,\tilde\Omega_k(0, y) = 0$ (needed in the case $\xi_1 = 0$) it follows by virtue of the symmetry of Q that $Q(\mathcal{I}m\,\tilde\Omega_k(\cdot, y), 1) = 0$ and Condition (13) is equivalent to

$$Q(\mathcal{R}e\,\tilde\Omega_k(\cdot, y), 1) = \mathcal{O}(|y|^b) \quad \text{as } y \to 0 \text{ for } k = 1, \dots, M.$$

This relation holds iff the coefficients of y^m in the Taylor series of $Q(\mathcal{R}e\,\tilde{\Omega}_k(\cdot,y),1)$, given by $Q(c_{k,m},1)$, vanish for $m<b$. If σ_0 and r have like parity then, in view of (16), this is equivalent to $-k+1+m$ to be even and $Q(G_{r-\beta-k+1+m},1)=0$ for $m=0,\ldots,b-1$. Since $k\in[1,M]$ the equivalence of (13) with (17) is proved. If σ_0 and r have opposite parity the proof is similar. The last assertion is then implied. □

Lemma 3 *Condition (14) is equivalent to*

$$\sum_{j=1}^{J}\omega_j G_\ell(\xi_j)=0 \quad \textit{for even } \ell\in[r',r-\beta+b-1] \tag{18}$$

and Condition (15) is void if $M=1$ and if $M>1$ equivalent to

$$\sum_{j=1}^{J}\omega_j G_\ell(\xi_j)=0 \quad \textit{for even } \ell\in[-M+1+r',r-\beta+r'-2]. \tag{19}$$

P r o o f . Note that $\tilde{\Delta}'_k$ is obtained as a special case of $\tilde{\Omega}_k$ for the (even) operator L^+_β with $\beta=0$ and r replaced by r'. Taking (16) into account it is seen that (14) is equivalent to $Q(1,G_{r'+m})=0$ for even $r'+m$ and $m=0,\ldots,r-\beta+b-r'-1$. Similarly, (15) is equivalent to $Q(1,G_{r'-k+1+m})=0$ for even $r'-k+1+m$ and $m=0,\ldots,r-\beta-1$ and the equivalence of (19) follows. □

Sufficient conditions for (17) - (19) can be derived by noting that G_α is for even α a multiple of the Bernoulli polynomial B_α (see [3]). From this observation the next corollary follows easily from Lemmas 2 and 3. In its formulation we use the notation of an extended symmetric quadrature formula Q. By this we mean a modification of Q, which is symmetric for periodic functions only, into a general symmetric formula $\tilde{Q}$. The modification is necessary only in the case that $\xi_1=0$. To obtain $\tilde{Q}$ the additional quadrature point $\xi_{J+1}:=1$ is introduced with weight $\omega_{J+1}:=\omega_1/2$ and the weight for $\xi_1=0$ is changed to be also equal to $\omega_1/2$.

Corollary 1 *Let σ_0 and r have like parity and $r-\beta$ be even or let σ_0 and r have opposite parity and $r-\beta$ be odd. Then the conditions (17) - (19) are satisfied if the extended symmetric quadrature rule Q has at least order $2q$ of exactness, where $q=[(r-\beta+b-1)/2]$ unless $M>1$ and $b<r'-1$, where $q=[(r-\beta+r'-2)/2]$. Here $[x]$ denotes truncation of x to the next integer not larger than x.*

In the case of a general operator L observe that by our index assumptions we have $r'\geq 1, -M+1+r'\geq 1$ and $r-\beta>0$ and the following corollary can be derived from Lemmas 2 and 3.

Corollary 2 *If the symmetric quadrature formula Q satifies*

$$\sum_{j=1}^{J} \omega_j B_\ell(\xi_j) = 0 \quad \textit{for even } \ell \in [2, r-\beta+b-1], \tag{20}$$

$$\sum_{j=1}^{J} \omega_j G_\ell(\xi_j) = 0 \quad \textit{for odd } \ell \in [r-\beta-M+1, r-\beta+b-1] \tag{21}$$

and, additionally, if $M > 1$ and $b < r' - 1$

$$\sum_{j=1}^{J} \omega_j B_\ell(\xi_j) = 0 \quad \textit{for even } \ell \in [r-\beta+b, r-\beta+r'-2] \tag{22}$$

then Conditions (17) - (19) hold true for general variable coefficient operators L.

If $M = 1$ these conditions coincide with (1.15) and (1.20) in [19].

In [18] a list of symmetric quadrature formulas with various exactness properties is provided. In the following table we collect those formulas which satisfy the conditions of Corollary 1 for certain choices of the parameters and, additionally, the formulas from Table 2. We keep the notation in [18]. A useful information for us is the first index indicating the number J.

M	$r-\beta$	b	r'	Formula
2	3	1	2	$G_{3,2,2}, L_{3,2,2}$
2	4	1	2	$G_{4,3,2}, L_{4,3,2}$
2	3	2	2,3	$G_{4,3,2}, L_{4,3,2}$
3	4	1	3	$G_{4,3,2}, L_{4,3,2}$
3	4	2	3	$G_{5,3,3}, L_{5,3,3}$

Table 1.— Quadrature formulas from [18] and Table 2 providing additional order b of convergence for general variable coefficient operators L

As an example how to determine the parameters of quadrature formulas like in Table 1 we derive Formula $G_{5,3,3}$ by an application of Corollary 2. With the parameters given in the last line in Table 1 Condition (20) is satisfied if all even polynomials of degree $\ell \in [2, 5]$ are integrated exactly (here we took into account that due to the normalization (6) constant functions are always integrated exactly). By applying the formulas to the polynomials $(\xi - 0.5)^2$ and $(\xi - 0.5)^4$ these two conditions take the form

$$\omega_1(2\xi_1 - 1)^2 + \omega_2(2\xi_2 - 1)^2 = 1/6,$$
$$\omega_1(2\xi_1 - 1)^4 + \omega_2(2\xi_2 - 1)^4 = 1/10,$$

where the symmetry relations $\omega_5 = \omega_1, \omega_4 = \omega_2, \xi_5 = 1 - \xi_1, \xi_4 = 1 - \xi_2, \xi_3 = 0.5$ were taken into account. The index ℓ in Condition (21) is odd and runs in $[2, 5]$ providing the two further equations

$$2\omega_1(G_n(\xi_1) - G_n(0.5)) + 2\omega_2(G_n(\xi_2) - G_n(0.5)) = G_n(0.5) \text{ for } n = 3, 5.$$

Condition (22) is void. Solving numerically the equations for the unknowns $\omega_1, \omega_2, \xi_1$ and ξ_2 yields the parameters for $G_{5,3,3}$ in Table 2. The parameters for $L_{5,3,3}$ are obtained similarly.

J	ξ_j	ω_j	Rule name
5	0.03675444410510	0.09796641612174	$G_{5,3,3}$
	0.20980173750308	0.24512752237399	
	0.5	0.31381212300853	
	0.79019826249692	0.24512752237399	
	0.96324555589490	0.09796641612174	
5	0.0	0.04767138349495	$L_{5,3,3}$
	0.09758560632523	0.17451387385978	
	0.34287284360121	0.30165043439274	
	0.65712715639879	0.30165043439274	
	0.90241439367477	0.17451387385978	

Table 2.— Quadrature formulas providing for $M = 3$ additional order $b = 2$ of convergence

Remark 2 *The stability of the formulas from [18] has been numerically checked there for strongly and oddly elliptic operators with integer $\beta \in [-1, 1]$. For some of the rules stability was proved analytically in [14].*

We conclude this section with some remarks concerning constant coefficient operators L_0 and the collocation method.

Lemma 4 *Let $M = 1$, $b \leq r'$ and D be elliptic. Then Condition (11) to hold for both L_β^+ and L_β^- is equivalent to*

$$\sum_{j=1}^{J} \omega_j G_{r-\beta+\ell}(\xi_j) = 0 \quad \textit{for } \ell = 0, \ldots, b-1, \tag{23}$$

and thus is identical with [18, Condition (4.13)] for AOC of order b.

The lemma follows from the observation that Condition (11) for $M = 1$ and $b \le r'$ is equivalent to

$$Q(\tilde{\Omega}_1(\cdot, y), 1) = \mathcal{O}(|y|^b) \quad \text{as } y \to 0,$$

which, as in Lemma 2, is equivalent to (23).

For collocation with double knot splines the following conditions for AOC have been given in [11].

Lemma 5 *Assume that the symbol (3) has constant coefficients satisfying $a^+ = 0$ or $a^- = 0$ and that $M = 2$. If the quadrature points are*

$$\xi_1 = 0, \xi_2 = \frac{1}{2} \quad \textit{and } \sigma_0 \textit{ and } r \textit{ have opposite parity,} \tag{24}$$

or

$$\xi_1 = \epsilon, \xi_2 = 1 - \epsilon \quad \textit{and } \sigma_0 \textit{ and } r \textit{ have like parity, where } G_{r-\beta}(\epsilon) = 0, \tag{25}$$

then (11) holds with $b = 1$ or $b = 2$, respectively.

It is shown in [3] that $G_{r-\beta}$ has a unique zero in $(0, 1/2)$.

P r o o f . For $J = M$ the collocation method is a special case of the qualocation method if Condition (R′) is satisfied. With the quadrature points in Lemma 5 and the choice $r' = 3$ or $r' = 2$ in the case of (24) or (25), respectively, (R′) has been proved in [7, Prop. 5.1 and 5.2].

In both cases the quadrature rules are symmetric. The numerical symbol D is elliptic, which follows from a slight generalization of [7, Prop. 5.1 and 5.2] in combination with [8, Lemma 3.1] (in the case of opposite parity also from [11, Lemma 5.1]). The conditions

$$Q\big(\tilde{\Omega}_1(\cdot, y), 1\big) = \mathcal{O}(|y|^b) \quad \text{and} \quad D_{1,2}^{-1}(y) Q\big(\tilde{\Omega}_1(\cdot, y), \Delta_2'(\cdot, y)\big) = \mathcal{O}(|y|^b) \tag{26}$$

are seen to be sufficient for (11), where for the first condition the boundedness of D^{-1} and the form $\Delta_1'(\cdot, y) = 1 + y^{r'} \tilde{\Delta}_1'(\cdot, y)$ with $r' (\ge 2) \ge b$ was taken into account. Consider the case of like parity. With the aid of (16) we conclude for the Taylor coefficients $c_{1,m}(\xi)$ of $\mathcal{R}e\, \tilde{\Omega}_1(\xi, \cdot)$ that $c_{1,0}(1 - \epsilon) = c_{1,0}(\epsilon) = G_{r-\beta}(\epsilon) = 0$. The relation (16) yields also $c_{1,1} = 0$. Consequently, $\mathcal{R}\mathrm{e}\, \tilde{\Omega}_1(\xi_1, y) = \mathcal{R}\mathrm{e}\, \tilde{\Omega}_1(\xi_2, y) = \mathcal{O}(|y|^2)$ and the first relation in (26) holds with $b = 2$. In the case of opposite parity, (16) yields $c_{1,0} = 0$ and, consequently, $\mathcal{R}\mathrm{e}\, \tilde{\Omega}_1(\xi, y) = \mathcal{O}(|y|)$. Thus the first relation in (26) holds with $b = 1$.

For the proof of the second relation in (26) first note that, due to

$$\mathcal{R}e\, \tilde{\Omega}_1(1 - \xi, \cdot) = \mathcal{R}e\, \tilde{\Omega}_1(\xi, \cdot), \quad \mathcal{I}m\, \tilde{\Omega}_1(1 - \xi, \cdot) = -\mathcal{I}m\, \tilde{\Omega}_1(\xi, \cdot), \quad \mathcal{I}m\, \tilde{\Omega}_1(0, \cdot) = 0,$$

the corresponding relations for Δ_2' and the symmetry of the quadrature formulas, we have

$$D_{1,2}^{-1}(y) Q\big(\tilde{\Omega}_1(\cdot, y), \Delta_2'(\cdot, y)\big) = D_{1,2}^{-1}(y) Q\Big(\mathcal{R}e\, \big(\tilde{\Omega}_1(\cdot, y) \overline{\Delta_2'(\cdot, y)}\big)\Big)$$

$$= D_{1,2}^{-1}(y)Q\big(\mathcal{R}e\,\tilde{\Omega}_1(\cdot,y)\mathcal{R}e\,\Delta_2'(\cdot,y)\big) + D_{1,2}^{-1}(y)Q\big(\mathcal{I}m\,\tilde{\Omega}_1(\cdot,y)\mathcal{I}m\,\Delta_2'(\cdot,y)\big). \tag{27}$$

¿From the first part of the proof follows that the first term in (27) has the correct order since $D_{1,2}^{-1}$ is bounded. For the second term note that in the case of opposite parity $\mathcal{I}m\,\tilde{\Omega}_1(\xi_1,0) = \mathcal{I}m\,\tilde{\Omega}_1(\xi_2,0) = 0$ (see (28)), which implies order $b = 1$ of the second term in (27). Consider the case of like parity. It is not difficult to check that for $k = 1, 2$

$$\mathcal{R}e\,\tilde{\Omega}_k(\cdot,-y) = (-1)^{k-1}\mathcal{R}e\,\tilde{\Omega}_k(\cdot,y), \quad \mathcal{I}m\,\tilde{\Omega}_k(\cdot,-y) = (-1)^k\mathcal{I}m\,\tilde{\Omega}_k(\cdot,y). \tag{28}$$

The functions Δ_k' satisfy the same relations since their symbol $\sigma_0 = |\xi|^0$ and $r' = 2$ have like parity. Then calculating the matrix element $D_{1,2}^{-1}$ with Cramer's rule it is seen to be odd. It follows also from (28) that $\mathcal{I}m\,\tilde{\Omega}_1(\xi,0) = 0$ and that $Q\big(\mathcal{I}m\,\tilde{\Omega}_1(\cdot,y)\mathcal{I}m\,\Delta_2'(\cdot,y)\big)$ is odd with respect to y. Thus the second term in (27) is even and vanishes for $y = 0$ implying the required order $b = 2$. □

Remark 3 *Due to the general assumption $\beta + M < r$, which is equally made in [18] (case $M = 1$), [11] and [5] it is not allowed in the case $M = 2$ to set $r - \beta = 2$ although the quantities in Lemmas 2 and 3 are well-defined for this choice and would lead to the same additional order $b = 1$ or $b = 2$ as in Lemma 5 for even or odd variable coefficient operators L (with weights $\omega_1 = 1/3, \omega_2 = 2/3$ for the rule (24)). The reasons for this restriction may be of technical nature.*

References

[1] C. DeBoor and B. Swartz (1973), Collocation at Gaussian points, SIAM J. Numer. Anal. **10**, 582–606.

[2] G. Brown, G.A. Chandler, I.H. Sloan and I.H. Wilson (1991), Properties of certain trigonometric series arising in numerical analysis, J. Math. Anal. and Applic. **162**, 371–380.

[3] G.A. Chandler and I.H. Sloan (1990), Spline qualocation methods for boundary integral equations, Numer. Math. **58**, 537–567.

[4] R.D. Grigorieff (2005), Superapproximation for projections on spline spaces, Numer. Math. **99**, 657–668.

[5] R.D. Grigorieff (2009), Additional order convergence in qualocation for elliptic boundary integral equations, to be published in J. Integral Equations Appl.

[6] R.D. Grigorieff (2010), Stability in interpolation with periodic multiple knot splines, in preparation.

[7] R.D. Grigorieff and I.H. Sloan (2005), Discrete orthogonal projections on multiple knot periodic splines, J. Approx. Th. **137**, 201–225.

[8] R.D. Grigorieff and I.H. Sloan (2006), Qualocation for boundary integral equations using splines with multiple knots, J. Integral Equations Appl. **18**, 117–140.

[9] G.C. Hsiao and W.L. Wendland (1981), The Aubin-Nitsche lemma for integral equations, J. Integral Equations **3**, 299–315.

[10] M. Křížek and P. Neittaanmäki (1987), On superconvergence techniques, Acta Applicandae Mathematicae **9**, 175–198.

[11] W. McLean and S. Prößdorf (1996), Boundary element collocation methods using splines with multiple knots, Numer. Math. **74**, 419–451.

[12] G. Plonka (1994), Periodic spline interpolation with shifted nodes, J. Approx. Th. **76**, 1–20.

[13] S. Prößdorf (1984), Ein Lokalisierungsprinzip in der Theorie der Spline-Approximationen und einige Anwendungen, Math. Nachr. **119**, 239–255.

[14] Schneider, C., Stability of qualocation methods for elliptic boundary integral equations, J. Integral Equations Appl. **15**, 203–216 (2003).

[15] I.H. Sloan (1988), A quadrature approach to improving the collocation method, Numer. Math. **54**, 41–56.

[16] I.H. Sloan (1988), Qualocation, J. Comp. Appl. Math. **79**, 461–478.

[17] I.H. Sloan and N. Tran (2001), The tolerant qualocation method for variable coefficient elliptic equations on curves, J. Integral Equations Appl. **13**, 73–98.

[18] I.H. Sloan and W.L. Wendland (1998), Qualocation methods for elliptic boundary integral equations, Numer. Math. **79**, 451–483.

[19] I.H. Sloan and W.L. Wendland (1999), Spline qualocation methods for variable-coefficient elliptic equations on curves, Numer. Math. **83**, 497–533.

[20] L.B. Wahlbin (1995), Superconvergence in Galerkin Finite Element Methods, Berlin: Springer-Verlag.

Monografías de la Real Academia de Ciencias de Zaragoza **33**, 21–38, (2010).

Efficient implementation of box-relaxation multigrid methods for the poroelasticity problem on semi-structured grids

Francisco José Gaspar, Francisco Javier Lisbona, and Carmen Rodrigo

Universidad de Zaragoza, Pedro Cerbuna 12, 50009, Zaragoza, Spain.

Abstract

We consider the numerical solution of a poroelasticity problem using a stabilized finite element method (FEM), based on the perturbation of the flow equation. Semi-structured triangular grids and stencil-based implementation of the linear FEM for displacements and pressure are used. An efficient procedure to construct the stencils associated with the basic diferential operators involved in the poroelasticity equations, using some reference stencils computed on a canonical hexagon, is presented. To solve the algebraic system of saddle point type, geometric multigrid methods based on box-relaxation are proposed which result to have a good performance for the considered problem. Numerical results are presented to show the behavior of the method.

1 Introduction

The theory of poroelasticity addresses the time dependent coupling between the deformation of a porous material and the fluid flow inside. Although this problem was first studied by Terzaghi in [18], its general statement was given by Biot in some papers, see [4, 5, 6]. Biot's consolidation models are used to study problems in a wide range of scientific disciplines, as geomechanics, hydrogeology, petrol engineering and biomechanics, for example. Here, we consider the quasi-static Biot model that sometimes is referred as the incompressible case model. The state of a poroelastic medium is characterized by the elastic displacements $\mathbf{u}$, and fluid pressure p at each point. It is well known that discretization by linear finite elements for both unknown fields results in an unstable method giving non-physical oscillations in the approximation of the pressure field. To overcome this trouble, we shall use a stabilized finite element scheme presented in [1], which permits us to use linear finite element spaces for both displacements and pressure, providing solutions without oscillations, independently of the chosen discretization parameters.

Finite element methods are usually considered to work with unstructured grids, due to its flexibility. These grids offer advantages with regard to their better fitting to complex geometries. However, an important issue in the finite element solution of PDE problems concerns the construction and storage of the large sparse system matrix. This is usually done by the process so-called "assembly", and due to the sparse character of the resulting matrix it is very important the way in which it is stored. Data structures commonly used to this purpose work with a system of indirect indexing to access only the non-zero entries of the matrix which sometimes leads to some performance difficulties. On the other hand, the data structures are much more efficient when working with structured grids. A good alternative which combines the advantages of both types of meshes is to work with semi-structured grids, that is, to consider an unstructured mesh as coarsest grid in order to fit well the domain geometry, and to apply regular refinements to its elements.

One of the most important aspects in the numerical solution of partial differential equations is the efficient solution of the corresponding large systems of equations arising from their discretization. Multigrid methods [7, 12, 19] are among the most powerful techniques for solving such type of algebraic systems, and they have become very popular among the scientific community. Geometric multigrid methods are characterized by employing a hierarchy of grids. We are interested in the use of semi-structured triangular grids, where a nested hierarchy of grids is obtained by dividing each triangle into four congruent ones, connecting the midpoints of their edges. These grids provide a suitable framework for the implementation of a geometric multigrid algorithm, permitting the use of stencil-based data structures, see [3], being necessary only a few stencils to represent the discrete operator. Besides, an efficient procedure to compute such stencils using canonical stencils associated with a reference hexagon is proposed, providing expressions which give the stencils corresponding to an arbitrary triangle.

The choice of a suitable smoother is an important feature for the design of an efficient geometric multigrid method, and even it requires special attention when one works with systems of PDEs because the smoother should smooth the error for all unknowns. Moreover, for saddle point problems, see [2], numerical experiments show that smoothing factors of standard collective point-wise relaxations are not satisfactory [19]. The poroelasticity problem is an example of such type of systems, and its resolution by multigrid on semi-structured grids is the aim of this paper. An overview of multigrid methods for discretizations on rectangular grids of saddle point problems is presented in [19], where box-relaxation appears as one of the most suitable smoothers for this kind of problems. It consists of decomposing the mesh into small subdomains and treating them separately, that is, all (or a part of) the equations corresponding to the points in each subdomain are solved simultaneously as a system. This class of smoothers was introduced by Vanka in [21], to solve the finite difference discretization on rectangular grids of the Navier-Stokes

equations. Since then, much literature can be found about the application of this type of smoothers, mainly in the field of Computational Fluid Dynamics (CFD) [13, 20]. There are less papers concerning to the performance of this relaxation in the context of Computational Solid Mechanics (CSM), see for example [22]. However, for discretizations of the poroelasticity problem on rectangular grids, it has been proved to obtain very good results with these smoothers. For instance, in [10] a box-relaxation is performed for a discretization of the problem on staggered rectangular grids. Hence, it seems a good idea to extend box-relaxation to regular triangular grids.

The outline of this paper is the following. In Section 2, the formulation of the poroelasticity problem, as well as its stabilized finite element discretization, will be introduced. In Section 3, an efficient stencil-based implementation on semi-structured grids of the stabilized FEM scheme is developed. Finally, Section 4 is devoted to introduce the proposed geometric multigrid algorithm, based on Vanka-type relaxation on triangular grids, and to present some numerical experiments illustrating the behavior of such smoothers for the poroelasticity problem.

2 Poroelasticity problem

We consider the quasi-static Biot's model for soil consolidation. The porous medium is assumed to be linearly elastic, homogeneous and isotropic, and to be saturated by an incompressible fluid. According to Biot's theory [4], the mathematical model of a consolidation process is given by the following system of equations:

$$\text{equilibrium equation:} \quad \operatorname{div} \boldsymbol{\sigma}' - \alpha \nabla p = \mathbf{g}, \quad \text{in } \Omega, \tag{1}$$

$$\text{constitutive equation:} \quad \boldsymbol{\sigma}' = \lambda \operatorname{tr}(\boldsymbol{\epsilon}) \mathbf{I} + 2\mu \boldsymbol{\epsilon}, \quad \text{in } \Omega, \tag{2}$$

$$\text{compatibility condition:} \quad \boldsymbol{\epsilon}(\mathbf{u}) = \frac{1}{2}(\nabla \mathbf{u} + \nabla \mathbf{u}^t), \quad \text{in } \Omega, \tag{3}$$

$$\text{Darcy's law:} \quad \mathbf{q} = -\frac{\kappa}{\eta} \nabla p, \quad \text{in } \Omega, \tag{4}$$

$$\text{continuity equation:} \quad \nabla \cdot \mathbf{q} + \alpha \frac{\partial}{\partial t}(\nabla \cdot \mathbf{u}) = f, \quad \text{in } \Omega, \tag{5}$$

where Ω is an open bounded region of $\mathbb{R}^n$, $n \leq 3$, with regular boundary Γ, λ and μ are the so-called Lamé coefficients, α is the Biot-Willis constant which we will assume to be equal to one, κ is the permeability of the porous medium, and η is the viscosity of the fluid. $\boldsymbol{I}$ represents the identity tensor, $\mathbf{u}$ is the displacement vector, p is the pore pressure, $\boldsymbol{\sigma}'$ and $\boldsymbol{\epsilon}$ are the effective stress and strain tensors for the porous medium, and $\mathbf{q}$ is the percolation velocity of the fluid relative to the soil, where we ignore gravity effects. The source terms in the right-hand side $\mathbf{g}$ and f represent a density of applied body forces and a forced fluid extraction or injection process, respectively.

To complete the formulation of the problem, appropriate boundary conditions must be included. For instance, we can consider

$$p = 0, \quad \boldsymbol{\sigma}' \, \mathbf{n} = \mathbf{t}, \quad \text{on } \Gamma_t,$$

$$\mathbf{u} = \mathbf{0}, \quad \frac{\kappa}{\eta}(\nabla p) \cdot \mathbf{n} = 0, \quad \text{on } \Gamma_c, \tag{6}$$

where $\mathbf{n}$ is the unit outward normal to the boundary and $\Gamma_t \cup \Gamma_c = \Gamma$, with Γ_t and Γ_c disjoint subsets of Γ. At the initial time, the following incompressibility condition

$$\nabla \cdot \mathbf{u}\,(\mathbf{x}, 0) = 0, \quad \mathbf{x} \in \Omega, \tag{7}$$

is fulfilled. To establish the variational formulation of the problem, the following function spaces $\mathcal{Q} = \{q \in H^1(\Omega) \,|\, q = 0 \text{ on } \Gamma_t\}$, and $\mathcal{U} = \{\mathbf{u} \in (H^1(\Omega))^n \,|\, \mathbf{u} = \mathbf{0} \text{ on } \Gamma_c\}$, are considered, where $H^1(\Omega)$ is the well-known subspace of square integrable scalar-valued functions with also square integrable first derivatives. Denoting by $(\,\cdot\,,\,\cdot\,)$ the usual inner product between square integrable functions, and by introducing the bilinear forms

$$a(\mathbf{u}, \mathbf{v}) = 2\mu \sum_{i,j=1}^{n} (\epsilon_{ij}(\mathbf{u}), \epsilon_{ij}(\mathbf{v})) + \lambda(\nabla \cdot \mathbf{u}, \nabla \cdot \mathbf{v}), \quad b(p, q) = \frac{\kappa}{\eta} \sum_{i=1}^{n} \left(\frac{\partial p}{\partial x_i}, \frac{\partial q}{\partial x_i} \right),$$

where $\epsilon_{ij}(\mathbf{u})$ are the entries of the effective strain tensor $\boldsymbol{\epsilon}(\mathbf{u}) = \dfrac{1}{2}(\nabla \mathbf{u} + \nabla \mathbf{u}^t)$, the variational formulation of problem (1)–(5) with boundary and initial conditions given in (6) and (7) reads:

For each $t \in (0, T]$, find $(\mathbf{u}(t), p(t)) \in \mathcal{U} \times \mathcal{Q}$ such that

$$a(\mathbf{u}(t), \mathbf{v}) + (\nabla p(t), \mathbf{v}) = (\mathbf{g}, \mathbf{v}) + (\mathbf{t}, \mathbf{v})_{\Gamma_t}, \quad \forall\, \mathbf{v} \in \mathcal{U}, \tag{8}$$

$$(\frac{\partial}{\partial t}(\nabla \cdot \mathbf{u}(t)), q) + b(p(t), q) = (f, q), \quad \forall q \in \mathcal{Q}, \tag{9}$$

with the initial condition $(\nabla \cdot \mathbf{u}(0), q) = 0, \forall\, q \in L^2(\Omega)$, and where

$$(\mathbf{t}, \mathbf{v})_{\Gamma_t} = \int_{\Gamma_t} \mathbf{t} \cdot \mathbf{v} \, \mathrm{d}\Gamma.$$

The most common way to solve poroelasticity problems is to use finite element methods, see for example [15]. However, standard finite element discretizations give satisfactory solutions only when the solution is smooth, since when sharp pressure gradients appear, these methods turn out to be unstable in the sense that strong non-physical oscillations appear in the approximation of the pressure. This oscillatory behavior is minimized with the use of FEM methods satisfying the LBB stability condition [9], although these oscillations are not completely eliminated, and therefore other stabilization techniques are

necessary, see [14, 16, 17] for example. In [1], using continuous piecewise linear approximation spaces for displacements and pressure, a new stabilization based on the perturbation of the flow equation is given, providing solutions without oscillations independently of the chosen discretization parameters. Here, this stabilized finite element scheme will be used.

Let us consider a triangulation $\mathcal{T}_h$ of Ω, which is assumed to satisfy the usual admissibility assumption. Let $S_h^1 \subset H^1(\Omega)$ be the C^0 piecewise linear polynomial finite element space. Let be $\mathcal{U}_h = \mathcal{U} \bigcap (S_h^1 \times S_h^1)$ and $\mathcal{Q}_h = \mathcal{Q} \bigcap S_h^1$. By considering these finite dimensional spaces, of dimensions n_u and n_p, respectively, and using an implicit time discretization, the following discrete formulation of the considered problem is obtained:

For a time-step $k \geq 1$, find $(\mathbf{u}_h^k, p_h^k) \in \mathcal{U}_h \times \mathcal{Q}_h$ such that

$$a(\mathbf{u}_h^k, \mathbf{v}_h) + (\nabla p_h^k, \mathbf{v}_h) = (\mathbf{g}^k, \mathbf{v}_h) + (\mathbf{t}, \mathbf{v}_h)_{\Gamma_t}, \quad \forall \mathbf{v}_h \in \mathcal{U}_h, \tag{10}$$

$$(\nabla \cdot \mathbf{u}_h^k, q_h) + \tau\, b(p_h^k, q_h) = (\nabla \cdot \mathbf{u}_h^{k-1}, q_h) + \tau\, (f^k, q_h) \quad \forall q_h \in \mathcal{Q}_h, \tag{11}$$

where τ is the time discretization parameter.

Let $\widetilde{\boldsymbol{\varphi}_i}$ be a vector nodal basis of $\mathcal{U}_h$, with all its components equal to φ_i, and φ_j a nodal basis of $\mathcal{Q}_h$. As consequence, the discrete approximations at time-step k can be written as

$$\mathbf{u}_h{}^k = \sum_{i=1}^{n_u} \mathbf{U}_i^k \varphi_i, \quad p_h^k = \sum_{j=1}^{n_p} P_j^k \varphi_j,$$

and the discrete formulation (10)-(11), can be expressed as a system of linear algebraic equations as follows

$$\begin{bmatrix} A & G \\ D & \tau\, B \end{bmatrix} \begin{bmatrix} \mathbf{U}^k \\ \mathbf{P}^k \end{bmatrix} = \begin{bmatrix} 0 & 0 \\ D & 0 \end{bmatrix} \begin{bmatrix} \mathbf{U}^{k-1} \\ \mathbf{P}^{k-1} \end{bmatrix} + \begin{bmatrix} \mathbf{G}^k \\ \tau\, \mathbf{F}^k \end{bmatrix}, \tag{12}$$

where $\mathbf{U}^k$ and $\mathbf{P}^k$ represent the unknown vectors in the $k^{\,th}$ time step: $(\mathbf{U}_1^k, \mathbf{U}_2^k, \ldots, \mathbf{U}_{n_u}^k)$ and $(P_1^k, P_2^k, \ldots, P_{n_p}^k)$, A is the elasticity matrix, B is the diffusive matrix multiplied by a coefficient κ/η, and G and D are the gradient and divergence matrices, respectively. $\mathbf{G}^k$ and $\mathbf{F}^k$ are the right hand side vectors in the $k^{\,th}$ time step, with components $\mathbf{G}_i^k = (\mathbf{g}^k, \widetilde{\boldsymbol{\varphi}_i}) + (\mathbf{t}, \widetilde{\boldsymbol{\varphi}_i})_{\Gamma_t}, i = 1, \ldots, n_u$ and $F_i^k = (f^k, \varphi_i), i = 1, \ldots, n_p$, respectively, and $D\mathbf{U}^0 = \mathbf{0}$.

As mentioned before,we consider the stabilized scheme presented in [1]. In such scheme, a term which arises from the discretization of the time derivative of the Laplacian of the pressure multiplied by a coefficient $\beta = h^2/4(\lambda + 2\mu)$, where h is the space discretization parameter, is added in the flow equation, and thus, the corresponding discrete

variational problem results in:

For $k \geq 1$, find $(\mathbf{u}_h^k, p_h^k) \in \mathcal{U}_h \times \mathcal{Q}_h$ such that

$$a(\mathbf{u}_h^k, \mathbf{v}_h) + (\nabla p_h^k, \mathbf{v}_h) = (\mathbf{g}^k, \mathbf{v}_h) + (\mathbf{t}, \mathbf{v}_h)_{\Gamma_t}, \quad \forall\, \mathbf{v}_h \in \mathcal{U}_h,$$
$$(\nabla \cdot \mathbf{u}_h^k, q_h) + (\tau + \beta')b(p_h^k, q_h) = \tau(f^k, q_h) + (\nabla \cdot \mathbf{u}_h^{k-1}, q_h) + \beta' b(p_h^{k-1}, q_h),\ \forall q_h \in \mathcal{Q}_h,$$

which in matrix-form reads as

$$\begin{bmatrix} A & G \\ D & (\tau+\beta')B \end{bmatrix} \begin{bmatrix} \mathbf{U}^k \\ \mathbf{P}^k \end{bmatrix} = \begin{bmatrix} 0 & 0 \\ D & \beta' B \end{bmatrix} \begin{bmatrix} \mathbf{U}^{k-1} \\ \mathbf{P}^{k-1} \end{bmatrix} + \begin{bmatrix} \mathbf{G}^k \\ \tau\, \mathbf{F}^k \end{bmatrix}, \tag{13}$$

where $\beta' = \beta \dfrac{\eta}{\kappa}$.

3 Implementation on semi-structured triangular grids

The stabilized finite element scheme introduced in Section 2 for problem (1)-(5) is considered on a particular triangulation of the domain related to a semi-structured grid obtained by local regular refinement of an input unstructured triangulation. The semi-structured character of the grid allows the use of low-cost memory storage of the discrete operator based on stencil formulation. An efficient procedure to construct these stencils by means of a reference hexagon is presented further on.
Let us denote $\mathcal{T}_0$ a coarse triangulation of Ω, which is assumed to be fine enough in order to fit the geometry of the domain. Once this coarse triangulation is given, each one of its triangles is divided into four congruent triangles connecting the midpoints of their edges, and this is repeated until a mesh $\mathcal{T}_f$ is obtained with the desired fine scale. This strategy generates a hierarchy of meshes, $\mathcal{T}_0 \subset \mathcal{T}_1 \subset \cdots \subset \mathcal{T}_f$.

For the implementation of the finite element method, we wish to store the coefficient matrix using a stencil–wise procedure, since a few types of stencils are enough to describe the discrete operator. For this, we distinguish three different types of points in the grid: interior nodes of a triangle of the coarsest grid, vertices of $\mathcal{T}_0$ and nodes lying on the edges of $\mathcal{T}_0$, see Figure 1. Depending on the location of a node in the grid, the way in which the discrete operator is described is different. A stencil form for interior points to each coarse triangle and different stencil forms for nodes lying on the edges of $\mathcal{T}_0$, are enough, since both types of points have a regular structure. However, in order to describe the discrete operator in the nodes of $\mathcal{T}_0$, which is unstructured, it is necessary to construct the stiffness matrix on the coarsest grid, by the usual assembly process, which will be scaled depending on the refinement level we are working with. In fact, two different data structures must be used, one of them totally unstructured, whereas the other, corresponding to the most

of the nodes, is a hierarchical structure, see Figure 1. This methodology, see [3], resembles the way of working with finite difference methods on block–structured grids.

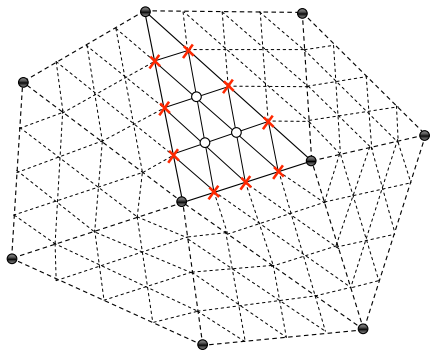


Figure 1.— Different kinds of nodes on a triangle of the coarsest grid: interior points (white circles), nodes lying on the edges (crosses), and vertexes of the unstructured grid (black circles).

Next, we concentrate on constructing the stencil associated with an interior point of a triangle $\mathcal{T}$ of the coarsest grid. To this end, we are going to define the regular grid arising inside this triangle. By considering a unitary basis of $\mathbb{R}^2$, $\{\mathbf{e'}_1, \mathbf{e'}_2\}$, fitting the geometry of the triangle, as we can see in Figure 2, we can define the grid for a refinement level $1 \leq \ell \leq f$ in the following way:

$$G_{\mathcal{T},\ell} = \{\mathbf{x} = (x,y)|_{\{\mathbf{e}'_i\}} \mid x = n\,h_1,\ y = m\,h_2,\ \ n = 0,\ldots,2^\ell,\ m = 0,\ldots,n\},$$

where $\mathbf{h} = (h_1, h_2)$ is the grid spacing in triangle $\mathcal{T}$, associated with the refinement level ℓ. Hence, a local numeration with double index can be fixed in each one of the triangles

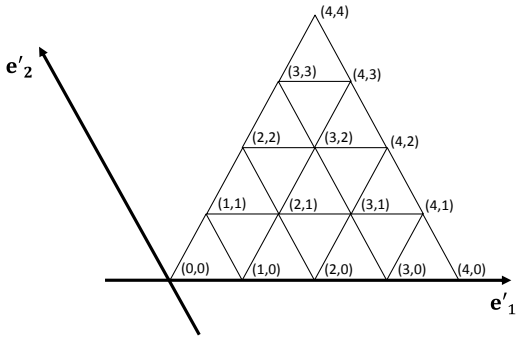


Figure 2.— New basis in $\mathbb{R}^2$ fitting the geometry of a triangle of the coarsest grid, and local numeration.

of the coarsest grid. For a refinement level ℓ, nodes are referred by (n,m), $n = 0,\ldots,2^\ell$, $m = 0,\ldots,n$, in such a way that the indexes of the vertices of the triangle are $(0,0)$,

$(2^\ell, 0)$, $(2^\ell, 2^\ell)$, as we can also observe in Figure 2 for $\ell = 2$. This way of numbering nodes is very convenient for identifying the neighboring nodes, which will be crucial when performing the geometric multigrid method.

For simplicity of presentation of the numerical scheme, we will consider only homogeneous boundary conditions. Spaces $\mathcal{U}$ and $\mathcal{Q}$ are the corresponding subspaces of functions in $(H^1(\Omega))^2$ and $H^1(\Omega)$, respectively, vanishing in the Dirichlet boundary, and the associated finite element spaces are built as in Section 2. Let us consider an interior node $\mathbf{x}_i$ to the grid $G_{\mathcal{T},\ell}$, associated with unknowns $\mathbf{u}_i^k$ and p_i^k. Since there is a biyective correspondence between the global and the local numeration, node $\mathbf{x}_i$ corresponds to an index (n, m), and therefore $\mathbf{u}_i^k$ and p_i^k can be denoted as $\mathbf{u}_\ell^k(\mathbf{x}_{n,m})$ and $p_\ell^k(\mathbf{x}_{n,m})$. Thus, the equations of system (13) corresponding to such interior point can be written in terms of discrete operators as follows:

$$A_\ell \mathbf{u}_\ell^k(\mathbf{x}_{n,m}) + G_\ell p_\ell^k(\mathbf{x}_{n,m}) = \mathbf{g}^k(\mathbf{x}_{n,m}), \tag{14}$$

$$D_\ell \mathbf{u}_\ell^k(\mathbf{x}_{n,m}) + (\tau + \beta) B_\ell p_\ell^k(\mathbf{x}_{n,m}) = D_\ell \mathbf{u}_\ell^{k-1}(\mathbf{x}_{n,m}) + \beta' B_\ell p_\ell^{k-1}(\mathbf{x}_{n,m}) + \tau f^k(\mathbf{x}_{n,m}),$$

where $\mathbf{u}_\ell^{k-1}(\mathbf{x}_{n,m})$ and $p_\ell^{k-1}(\mathbf{x}_{n,m})$ are known values, since represent the solution at previous time step, and A_ℓ, B_ℓ, G_ℓ, and D_ℓ denote the local discrete operators corresponding to an interior point of the considered triangle. As $\mathbf{u}_\ell^k(\mathbf{x}_{n,m}) = (u_\ell^k(\mathbf{x}_{n,m}), v_\ell^k(\mathbf{x}_{n,m}))^t$, operators A_ℓ and G_ℓ are vector discrete operators, which in stencil notation are given by

$$A_\ell = \begin{bmatrix} A_\ell^{11} & A_\ell^{12} \\ A_\ell^{21} & A_\ell^{22} \end{bmatrix}, \qquad G_\ell = \begin{bmatrix} G_\ell^x \\ G_\ell^y \end{bmatrix},$$

where

$$A_\ell^{ij} = \begin{bmatrix} 0 & a_{01}^{ij} & a_{11}^{ij} \\ a_{-10}^{ij} & a_{00}^{ij} & a_{10}^{ij} \\ a_{-1-1}^{ij} & a_{0-1}^{ij} & 0 \end{bmatrix}, \quad G_\ell^x = \begin{bmatrix} 0 & g_{01}^x & g_{11}^x \\ g_{-10}^x & g_{00}^x & g_{10}^x \\ g_{-1-1}^x & g_{0-1}^x & 0 \end{bmatrix}, \quad G_\ell^y = \begin{bmatrix} 0 & g_{01}^y & g_{11}^y \\ g_{-10}^y & g_{00}^y & g_{10}^y \\ g_{-1-1}^y & g_{0-1}^y & 0 \end{bmatrix}.$$

Analogously, D_ℓ is given by $\begin{bmatrix} D_\ell^x & D_\ell^y \end{bmatrix}$, where $D_\ell^x = G_\ell^x$ and $D_\ell^y = G_\ell^y$, whereas B_ℓ is a simple scalar discrete operator

$$B_\ell = \begin{bmatrix} 0 & b_{01} & b_{11} \\ b_{-10} & b_{00} & b_{10} \\ b_{-1-1} & b_{0-1} & 0 \end{bmatrix}.$$

Notice that each interior node is the center of a hexagon H, consisting of six congruent triangles, so the only unknowns appearing in the equations of node (n, m) are the corresponding to indexes $(n+1, m)$, $(n-1, m)$, $(n, m+1)$, $(n, m-1)$, $(n+1, m+1)$, $(n-1, m-1)$, see Figure 3. This allows us to write previous equations (14) in the following

way

$$\sum_{(\kappa_1,\kappa_2)\in\mathcal{I}} a^{11}_{\kappa_1,\kappa_2} u^k_\ell(\mathbf{x}_{n+\kappa_1,m+\kappa_2}) + \sum_{(\kappa_1,\kappa_2)\in\mathcal{I}} a^{12}_{\kappa_1,\kappa_2} v^k_\ell(\mathbf{x}_{n+\kappa_1,m+\kappa_2}) + \sum_{(\kappa_1,\kappa_2)\in\mathcal{I}} g^x_{\kappa_1,\kappa_2} p^k_\ell(\mathbf{x}_{n+\kappa_1,m+\kappa_2}) = \int_H g^k_1 \varphi_{n,m} \mathrm{d}\mathbf{x},$$

$$\sum_{(\kappa_1,\kappa_2)\in\mathcal{I}} a^{21}_{\kappa_1,\kappa_2} u^k_\ell(\mathbf{x}_{n+\kappa_1,m+\kappa_2}) + \sum_{(\kappa_1,\kappa_2)\in\mathcal{I}} a^{22}_{\kappa_1,\kappa_2} v^k_\ell(\mathbf{x}_{n+\kappa_1,m+\kappa_2}) + \sum_{(\kappa_1,\kappa_2)\in\mathcal{I}} g^y_{\kappa_1,\kappa_2} p^k_\ell(\mathbf{x}_{n+\kappa_1,m+\kappa_2}) = \int_H g^k_2 \varphi_{n,m} \mathrm{d}\mathbf{x},$$

$$\sum_{(\kappa_1,\kappa_2)\in\mathcal{I}} d^x_{\kappa_1,\kappa_2} u^k_\ell(\mathbf{x}_{n+\kappa_1,m+\kappa_2}) + \sum_{(\kappa_1,\kappa_2)\in\mathcal{I}} d^y_{\kappa_1,\kappa_2} v^k_\ell(\mathbf{x}_{n+\kappa_1,m+\kappa_2}) + (\tau+\beta) \sum_{(\kappa_1,\kappa_2)\in\mathcal{I}} b_{\kappa_1,\kappa_2} p^k_\ell(\mathbf{x}_{n+\kappa_1,m+\kappa_2}) =$$

$$\sum_{(\kappa_1,\kappa_2)\in\mathcal{I}} d^x_{\kappa_1,\kappa_2} u^{k-1}_\ell(\mathbf{x}_{n+\kappa_1,m+\kappa_2}) + \sum_{(\kappa_1,\kappa_2)\in\mathcal{I}} d^y_{\kappa_1,\kappa_2} v^{k-1}_\ell(\mathbf{x}_{n+\kappa_1,m+\kappa_2}) + \beta' \sum_{(\kappa_1,\kappa_2)\in\mathcal{I}} b_{\kappa_1,\kappa_2} p^{k-1}_\ell(\mathbf{x}_{n+\kappa_1,m+\kappa_2}) +$$

$$\tau \int_H f^k \varphi_{n,m} \mathrm{d}\mathbf{x}, \tag{15}$$

where $\mathcal{I}$ is the set $\mathcal{I} = \{(\kappa_1,\kappa_2) \,|\, \kappa_1,\kappa_2 = -1,0,1\} \subset \mathbb{Z}^2$.

In order to efficiently compute the previous stencils associated with discrete operators, we use a strategy similar to that used in the standard finite element assembly, taking in this case a reference hexagon. In order to illustrate such construction, we begin considering operator B_ℓ. The stencil form for operator B_ℓ reads

$$B_\ell = \frac{\kappa}{\eta} \cdot \begin{bmatrix} 0 & \displaystyle\int_{T_2\cup T_3} \nabla\varphi_{n,m+1}\cdot\nabla\varphi_{n,m}\,\mathrm{d}\mathbf{x} & \displaystyle\int_{T_1\cup T_2} \nabla\varphi_{n+1,m+1}\cdot\nabla\varphi_{n,m}\,\mathrm{d}\mathbf{x} \\ \displaystyle\int_{T_3\cup T_4} \nabla\varphi_{n-1,m}\cdot\nabla\varphi_{n,m}\,\mathrm{d}\mathbf{x} & \displaystyle\int_{\bigcup_{i=1}^6 T_i} \nabla\varphi_{n,m}\cdot\nabla\varphi_{n,m}\,\mathrm{d}\mathbf{x} & \displaystyle\int_{T_1\cup T_6} \nabla\varphi_{n+1,m}\cdot\nabla\varphi_{n,m}\,\mathrm{d}\mathbf{x} \\ \displaystyle\int_{T_4\cup T_5} \nabla\varphi_{n-1,m-1}\cdot\nabla\varphi_{n,m}\,\mathrm{d}\mathbf{x} & \displaystyle\int_{T_5\cup T_6} \nabla\varphi_{n,m-1}\cdot\nabla\varphi_{n,m}\,\mathrm{d}\mathbf{x} & 0 \end{bmatrix}, \tag{16}$$

where $T_i,\ i = 1,\ldots,6$ are the triangles composing the hexagon H around node (n,m), and the nodal basis functions $\varphi_{k,l}$ are referred to the local numeration, see Figure 3. In order to have an effective computation of this stencil we will use a reference hexagon $\hat{H}$ with center $\hat{\mathbf{x}}_{0,0} = (0,0)$ and vertices $\hat{\mathbf{x}}_{1,0} = (1,0)$, $\hat{\mathbf{x}}_{1,1} = (1,1)$, $\hat{\mathbf{x}}_{0,1} = (0,1)$, $\hat{\mathbf{x}}_{-1,0} = (-1,0)$, $\hat{\mathbf{x}}_{-1,-1} = (-1,-1)$, $\hat{\mathbf{x}}_{0,-1} = (0,-1)$, and an affine transformation F_H mapping hexagon $\hat{H}$ onto H, $\mathbf{x} = F_H(\hat{\mathbf{x}}) = \mathcal{B}_H\,\hat{\mathbf{x}} + b_H$ with

$$\mathcal{B}_H = \begin{pmatrix} x_{n+1,m} - x_{n,m} & x_{n+1,m+1} - x_{n+1,m} \\ y_{n+1,m} - y_{n,m} & y_{n+1,m+1} - y_{n+1,m} \end{pmatrix}, \quad b_H = \begin{pmatrix} x_{n,m} \\ y_{n,m} \end{pmatrix},$$

where $(x_{k,l}, y_{k,l})$ are the coordinates of the nodes $\mathbf{x}_{k,l}$. Note that matrix $\mathcal{B}_H$ is proportional with factor $2^{-\ell}$, where ℓ is the refinement level, to the matrix associated with the affine transformation between $\hat{T}_1$ (see Figure 3) and the current triangle of the input coarsest grid. With these definitions, we can translate the degrees of freedom and basis functions (denoted here by $\hat{\varphi}$) on the reference hexagon to degrees of freedom and basis functions

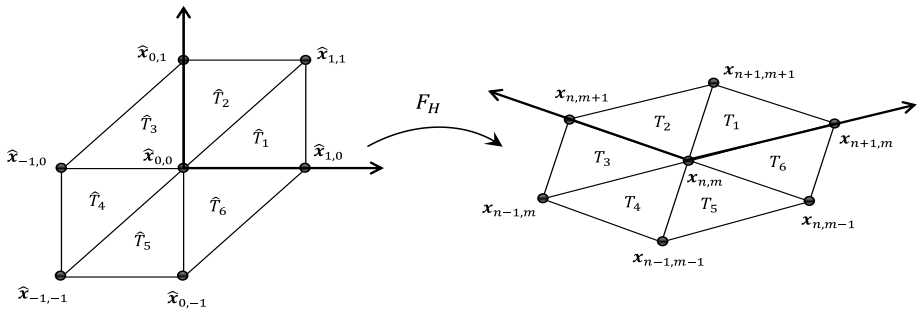


Figure 3.— Reference hexagon and corresponding affine transformation F_H.

on the hexagon H. In particular, we have

$$\hat{\varphi}_{k,l} = \varphi_{k,l} \circ F_H, \quad \nabla\hat{\varphi}_{k,l} = \mathcal{B}_H^t \nabla\varphi_{k,l} \circ F_H.$$

By applying the change of variable associated with the affine mapping, the entries of the stencil (16) yield the following expressions

$$b_{0,1} = |\det \mathcal{B}_H| \frac{\kappa}{\eta} \Big(\int_{\hat{T}_2} (\mathcal{B}_H^{-1})^t \nabla\hat{\varphi}_{0,1} \cdot (\mathcal{B}_H^{-1})^t \nabla\hat{\varphi}_{0,0} \mathrm{d}\hat{\mathbf{x}} + \int_{\hat{T}_3} (\mathcal{B}_H^{-1})^t \nabla\hat{\varphi}_{0,1} \cdot (\mathcal{B}_H^{-1})^t \nabla\hat{\varphi}_{0,0} \mathrm{d}\hat{\mathbf{x}} \Big),$$

$$b_{1,1} = |\det \mathcal{B}_H| \frac{\kappa}{\eta} \Big(\int_{\hat{T}_1} (\mathcal{B}_H^{-1})^t \nabla\hat{\varphi}_{1,1} \cdot (\mathcal{B}_H^{-1})^t \nabla\hat{\varphi}_{0,0} \mathrm{d}\hat{\mathbf{x}} + \int_{\hat{T}_2} (\mathcal{B}_H^{-1})^t \nabla\hat{\varphi}_{1,1} \cdot (\mathcal{B}_H^{-1})^t \nabla\hat{\varphi}_{0,0} \mathrm{d}\hat{\mathbf{x}} \Big),$$

$$b_{-1,0} = |\det \mathcal{B}_H| \frac{\kappa}{\eta} \Big(\int_{\hat{T}_3} (\mathcal{B}_H^{-1})^t \nabla\hat{\varphi}_{-1,0} \cdot (\mathcal{B}_H^{-1})^t \nabla\hat{\varphi}_{0,0} \mathrm{d}\hat{\mathbf{x}} + \int_{\hat{T}_4} (\mathcal{B}_H^{-1})^t \nabla\hat{\varphi}_{-1,0} \cdot (\mathcal{B}_H^{-1})^t \nabla\hat{\varphi}_{0,0} \mathrm{d}\hat{\mathbf{x}} \Big),$$

$$b_{0,0} = |\det \mathcal{B}_H| \frac{\kappa}{\eta} \Big(\sum_{i=1}^{6} \int_{\hat{T}_i} (\mathcal{B}_H^{-1})^t \nabla\hat{\varphi}_{0,0} \cdot (\mathcal{B}_H^{-1})^t \nabla\hat{\varphi}_{0,0} \mathrm{d}\hat{\mathbf{x}} \Big),$$

$$b_{1,0} = |\det \mathcal{B}_H| \frac{\kappa}{\eta} \Big(\int_{\hat{T}_1} (\mathcal{B}_H^{-1})^t \nabla\hat{\varphi}_{1,0} \cdot (\mathcal{B}_H^{-1})^t \nabla\hat{\varphi}_{0,0} \mathrm{d}\hat{\mathbf{x}} + \int_{\hat{T}_6} (\mathcal{B}_H^{-1})^t \nabla\hat{\varphi}_{1,0} \cdot (\mathcal{B}_H^{-1})^t \nabla\hat{\varphi}_{0,0} \mathrm{d}\hat{\mathbf{x}} \Big),$$

$$b_{-1,-1} = |\det \mathcal{B}_H| \frac{\kappa}{\eta} \Big(\int_{\hat{T}_4} (\mathcal{B}_H^{-1})^t \nabla\hat{\varphi}_{-1,-1} \cdot (\mathcal{B}_H^{-1})^t \nabla\hat{\varphi}_{0,0} \mathrm{d}\hat{\mathbf{x}} + \int_{\hat{T}_5} (\mathcal{B}_H^{-1})^t \nabla\hat{\varphi}_{-1,-1} \cdot (\mathcal{B}_H^{-1})^t \nabla\hat{\varphi}_{0,0} \mathrm{d}\hat{\mathbf{x}} \Big),$$

$$b_{0,-1} = |\det \mathcal{B}_H| \frac{\kappa}{\eta} \Big(\int_{\hat{T}_5} (\mathcal{B}_H^{-1})^t \nabla\hat{\varphi}_{0,-1} \cdot (\mathcal{B}_H^{-1})^t \nabla\hat{\varphi}_{0,0} \mathrm{d}\hat{\mathbf{x}} + \int_{\hat{T}_6} (\mathcal{B}_H^{-1})^t \nabla\hat{\varphi}_{0,-1} \cdot (\mathcal{B}_H^{-1})^t \nabla\hat{\varphi}_{0,0} \mathrm{d}\hat{\mathbf{x}} \Big).$$

Now, defining the 2×2 matrix $C_H = \mathcal{B}_H^{-1}(\mathcal{B}_H^{-1})^t$,

$$C_H = \begin{pmatrix} c_{11}^H & c_{12}^H \\ c_{12}^H & c_{22}^H \end{pmatrix},$$

the stencil (16) has the expression

$$B_\ell = |\det \mathcal{B}_H| \frac{\kappa}{\eta} \left(c_{11}^H \hat{S}_{xx} + 2\, c_{12}^H \hat{S}_{xy} + c_{22}^H \hat{S}_{yy} \right),$$

where

$$\hat{S}_{xx} = \begin{bmatrix} 0 & 0 & 0 \\ -1 & 2 & -1 \\ 0 & 0 & 0 \end{bmatrix}, \quad \hat{S}_{xy} = \hat{S}_{yx} = \frac{1}{2} \begin{bmatrix} 0 & 1 & -1 \\ 1 & -2 & 1 \\ -1 & 1 & 0 \end{bmatrix}, \quad \hat{S}_{yy} = \begin{bmatrix} 0 & -1 & 0 \\ 0 & 2 & 0 \\ 0 & -1 & 0 \end{bmatrix},$$

are the stencils associated with operators $-\partial_{xx}$, $-\partial_{xy}$, $-\partial_{yx}$, and $-\partial_{yy}$, in the reference hexagon.

Analogously, we can obtain similar expressions in function of these reference stencils, for discrete operators $A_\ell^{ij},\ i,j=1,2$, $G_\ell^x = D_\ell^x$, and $G_\ell^y = D_\ell^y$.
For the stencils corresponding to the elasticity operator A_ℓ, some of their coefficients are given by

$$\begin{aligned}
a_{01}^{11} &= \int_{T_2\cup T_3}\Big[(\lambda+2\mu)\frac{\partial\varphi_{n,m+1}}{\partial x}\frac{\partial\varphi_{n,m}}{\partial x}+\mu\frac{\partial\varphi_{n,m+1}}{\partial y}\frac{\partial\varphi_{n,m}}{\partial y}\Big]\,\mathrm{d}\mathbf{x},\\
a_{01}^{12} &= \int_{T_2\cup T_3}\Big[\lambda\frac{\partial\varphi_{n,m+1}}{\partial y}\frac{\partial\varphi_{n,m}}{\partial x}+\mu\frac{\partial\varphi_{n,m+1}}{\partial x}\frac{\partial\varphi_{n,m}}{\partial y}\Big]\,\mathrm{d}\mathbf{x},\\
a_{01}^{21} &= \int_{T_2\cup T_3}\Big[\lambda\frac{\partial\varphi_{n,m+1}}{\partial x}\frac{\partial\varphi_{n,m}}{\partial y}+\mu\frac{\partial\varphi_{n,m+1}}{\partial y}\frac{\partial\varphi_{n,m}}{\partial x}\Big]\,\mathrm{d}\mathbf{x},\\
a_{01}^{22} &= \int_{T_2\cup T_3}\Big[\mu\frac{\partial\varphi_{n,m+1}}{\partial x}\frac{\partial\varphi_{n,m}}{\partial x}+(\lambda+2\mu)\frac{\partial\varphi_{n,m+1}}{\partial y}\frac{\partial\varphi_{n,m}}{\partial y}\Big]\,\mathrm{d}\mathbf{x},
\end{aligned}$$

and the rest of them have analogous definitions. Using the change of variable previously introduced, and defining the inverse of the matrix of the transformation as:

$$\mathcal{B}_H^{-1} = \begin{pmatrix} b_{11}^H & b_{12}^H \\ b_{21}^H & b_{22}^H \end{pmatrix},$$

the four scalar stencils corresponding to A_ℓ can be written in terms of the reference stencils in the following way:

$$\begin{aligned}
A_\ell^{11} &= |\det\mathcal{B}_H|(((\lambda+2\mu)(b_{11}^H)^2+\mu(b_{12}^H)^2)\,\hat{S}_{xx}+(\mu(b_{22}^H)^2+(\lambda+2\mu)(b_{21}^H)^2)\,\hat{S}_{yy}\\
&\quad+((\lambda+2\mu)b_{11}^Hb_{21}^H+\mu b_{22}^Hb_{12}^H)\,(\hat{S}_{xy}+\hat{S}_{yx})),\\
A_\ell^{12} &= |\det\mathcal{B}_H|((\lambda+\mu)b_{11}^Hb_{12}^H\,\hat{S}_{xx}+(\lambda+\mu)b_{22}^Hb_{21}^H\,\hat{S}_{yy}+(\lambda b_{12}^Hb_{21}^H+\mu b_{22}^Hb_{11}^H)\,\hat{S}_{xy}\\
&\quad+(\lambda b_{22}^Hb_{11}^H+\mu b_{21}^Hb_{12}^H)\,\hat{S}_{yx}),\\
A_\ell^{21} &= |\det\mathcal{B}_H|((\lambda+\mu)b_{11}^Hb_{12}^H\,\hat{S}_{xx}+(\lambda+\mu)b_{22}^Hb_{21}^H\,\hat{S}_{yy}+(\lambda b_{22}^Hb_{11}^H+\mu b_{21}^Hb_{12}^H)\,\hat{S}_{xy}\\
&\quad+(\lambda b_{12}^Hb_{21}^H+\mu b_{22}^Hb_{11}^H)\,\hat{S}_{yx}),\\
A_\ell^{22} &= |\det\mathcal{B}_H|(((\lambda+2\mu)(b_{12}^H)^2+\mu(b_{11}^H)^2)\,\hat{S}_{xx}+(\mu(b_{21}^H)^2+(\lambda+2\mu)(b_{22}^H)^2)\,\hat{S}_{yy}\\
&\quad+((\lambda+2\mu)b_{22}^Hb_{12}^H+\mu b_{11}^Hb_{21}^H)\,(\hat{S}_{xy}+\hat{S}_{yx})),
\end{aligned}$$

Now, it is straightforward to see that the stencils corresponding to $G_\ell^x = D_\ell^x$ and $G_\ell^y = D_\ell^y$,

are given by

$$G_\ell^x = \begin{bmatrix} 0 & \displaystyle\int_{T_2\cup T_3} \frac{\partial \varphi_{n,m+1}}{\partial x}\,\varphi_{n,m}\,\mathrm{d}\mathbf{x} & \displaystyle\int_{T_1\cup T_2} \frac{\partial \varphi_{n+1,m+1}}{\partial x}\,\varphi_{n,m}\,\mathrm{d}\mathbf{x} \\ \displaystyle\int_{T_3\cup T_4} \frac{\partial \varphi_{n-1,m}}{\partial x}\,\varphi_{n,m}\,\mathrm{d}\mathbf{x} & \displaystyle\int_{\bigcup_{i=1}^6 T_i} \frac{\partial \varphi_{n,m}}{\partial x}\,\varphi_{n,m}\,\mathrm{d}\mathbf{x} & \displaystyle\int_{T_1\cup T_6} \frac{\partial \varphi_{n+1,m}}{\partial x}\,\varphi_{n,m}\,\mathrm{d}\mathbf{x} \\ \displaystyle\int_{T_4\cup T_5} \frac{\partial \varphi_{n-1,m-1}}{\partial x}\,\varphi_{n,m}\,\mathrm{d}\mathbf{x} & \displaystyle\int_{T_5\cup T_6} \frac{\partial \varphi_{n,m-1}}{\partial x}\,\varphi_{n,m}\,\mathrm{d}\mathbf{x} & 0 \end{bmatrix},$$

and

$$G_\ell^y = \begin{bmatrix} 0 & \displaystyle\int_{T_2\cup T_3} \frac{\partial \varphi_{n,m+1}}{\partial y}\,\varphi_{n,m}\,\mathrm{d}\mathbf{x} & \displaystyle\int_{T_1\cup T_2} \frac{\partial \varphi_{n+1,m+1}}{\partial y}\,\varphi_{n,m}\,\mathrm{d}\mathbf{x} \\ \displaystyle\int_{T_3\cup T_4} \frac{\partial \varphi_{n-1,m}}{\partial y}\,\varphi_{n,m}\,\mathrm{d}\mathbf{x} & \displaystyle\int_{\bigcup_{i=1}^6 T_i} \frac{\partial \varphi_{n,m}}{\partial y}\,\varphi_{n,m}\,\mathrm{d}\mathbf{x} & \displaystyle\int_{T_1\cup T_6} \frac{\partial \varphi_{n+1,m}}{\partial y}\,\varphi_{n,m}\,\mathrm{d}\mathbf{x} \\ \displaystyle\int_{T_4\cup T_5} \frac{\partial \varphi_{n-1,m-1}}{\partial y}\,\varphi_{n,m}\,\mathrm{d}\mathbf{x} & \displaystyle\int_{T_5\cup T_6} \frac{\partial \varphi_{n,m-1}}{\partial y}\,\varphi_{n,m}\,\mathrm{d}\mathbf{x} & 0 \end{bmatrix},$$

And following calculations similar to those for the previous operators, they can be written in terms of reference stencils in the following way

$$G_\ell^x = |\det \mathcal{B}_H| \left(b_{11}^H\, \hat{S}_x + b_{21}^H\, \hat{S}_y \right),$$
$$G_\ell^y = |\det \mathcal{B}_H| \left(b_{12}^H\, \hat{S}_x + b_{22}^H\, \hat{S}_y \right),$$

where

$$\hat{S}_x = \frac{1}{6} \begin{bmatrix} 0 & -1 & 1 \\ -2 & 0 & 2 \\ -1 & 1 & 0 \end{bmatrix}, \quad \hat{S}_y = \frac{1}{6} \begin{bmatrix} 0 & 2 & 1 \\ 1 & 0 & -1 \\ -1 & -2 & 0 \end{bmatrix},$$

are the stencils corresponding to operators ∂_x and ∂_y computed in the reference hexagon.

Finally, we normalize equations in (15) with the factor $|\det \mathcal{B}_H|$, and then the right–hand sides are approximations of $\mathbf{g}^k(\mathbf{x}_{n,m})$ and $f^k(\mathbf{x}_{n,m})$. Notice that these stencil forms of the discrete operators in function of the reference stencils give the stencil corresponding to an interior point of an arbitrary triangle with any geometry and any position in the plane.

With obvious modifications of the above process, it is possible to construct the stencil associated with the nodes located at the edges in $\mathcal{T}_0$ in terms of the basic stencils and the appropriate affine transfonmations.

4 Multigrid based on Vanka-type smoothers

The design of an efficient geometric multigrid method for a concrete problem depends strongly on the choice of adequate components of the algorithm. These components have to be chosen so that they efficiently interplay with each other in order to obtain a good

connection between the relaxation and the coarse-grid correction. In this paper, linear interpolation is chosen as the prolongation, and its adjoint as the restriction operator. The discrete operator on each mesh in the hierarchy results from the direct discretization of the system of partial differential equations on the corresponding grid. Due to the semi-structured character of the grid, we will use a block-wise multigrid algorithm, in which each triangle of the coarsest triangulation is treated as a different block with regard to the smoothing process. However, there are several points in the algorithm, where information from neighboring triangles must be transferred, and to facilitate this communication each triangle of the coarsest grid is augmented by an overlap-layer of so-called ghost nodes that surround it.

As commented in the introduction, standard smoothers, as simple point-wise gauss-seidel smoothers (with any ordering), are not appropriate for saddle point type problems. We shall see that box-relaxation can be taken as a suitable smoother to deal with poroelasticity problems. It consists of decomposing the grid into small overlapped subdomains and looping over all of them solving the system arising from the equations corresponding to the points in the subdomain. There are many variants of box-type smoothers, they can differ in the choice of the subdomains which are solved simultaneously, and in the way in which the local systems are solved. Also, the different subdomains can be visited in differente orderings, for example red-black or three-color ordering (see [11]), yielding to a wide variety of box-relaxations. Here, we only deal with a pair of them. Firstly, we consider a point-wise box Gauss-Seidel iterative algorithm, which consists of simultaneously updating all unknowns corresponding to the nodes located at the vertexes of an hexagon centered on a grid point, together with the unknowns at this point. This means that 21 unknowns corresponding to displacement and pressure unknowns (see left Figure 4) are relaxed simultaneously and therefore, a 21×21 system has to be solved for each box. In the variant considered here, the subdomains are visited in lexicographic order. The need of solving such systems makes these smoothers expensive. A simplified variant of them can be considered by only coupling the unknowns associated with a cell in the grid, that is, unknowns located at the three vertexes of each triangle, see right Figure 4. This implies to solve a 9×9 system on each triangular cell. This smoother is known as cell-wise box relaxation. In this paper, a variant of this cell-wise box smoother is considered, in particular a red-black cell-wise box-relaxation is used. It consists of looping the triangular cells of the grid in a checkerboard manner, that is, first the up-oriented triangles are updated, and the down-oriented ones are relaxed in the second partial relaxation step.

In order to see the suitability of box-smoothers for the considered problem, we solve the so-called poroelasticity footing problem on the computational rectangular domain depicted in left part of Figure 5, with dimensions $\Omega = [0, 1] \times [0, \sqrt{3}/2]$.

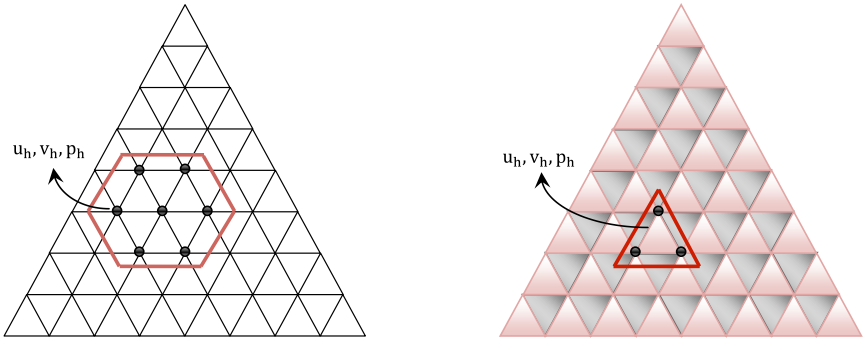


Figure 4.— Unknowns simultaneously updated in point-wise and cell-wise box Gauss-Seidel.

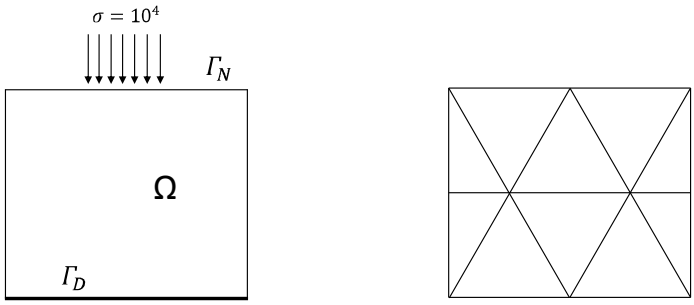


Figure 5.— Rectangular computational domain, and considered coarsest triangulation.

The body is assumed rigid at the bottom and a uniform load is applied in a strip of length 0.4, in the central part of the upper boundary. Besides, the whole contour is assumed free to drain. More concretely, the considered boundary conditions are the following

$$p = 0, \text{ on } \Gamma = \partial\Omega,$$
$$\boldsymbol{\sigma}' \mathbf{n} = \mathbf{t}, \text{ on } \Gamma_{t,1},$$
$$\boldsymbol{\sigma}' \mathbf{n} = \mathbf{0}, \text{ on } \Gamma_{t,2},$$
$$\mathbf{u} = \mathbf{0}, \text{ on } \Gamma_c = \Gamma \backslash \{\Gamma_{t,1} \cup \Gamma_{t,2}\},$$

where $\mathbf{t} = (0, -10^4)^t$, and

$$\Gamma_{t,1} = \{(x, y) \in \Gamma \,|\, y = \sqrt{3}/2,\ 0.3 \leq x \leq 0.7\},$$

$$\Gamma_{t,2} = \{(x, y) \in \Gamma \,|\, y = \sqrt{3}/2,\ 0 \leq x \leq 0.3 \text{ or } 0.7 \leq x \leq 1\},$$

and the material properties of the porous medium are given in Table 1. The considered time-step is $\tau = 10^{-2}$.

The coarsest triangulation, composed of 10 triangles, is also depicted in right part of Figure 5. From this grid, a regular refinement process is applied to each element of the

Property	Value	Unit
Young's modulus	3×10^4	N/m^2
Poisson's ratio	0.2	-
Permeability	10^{-10}	m^2
Fluid viscosity	10^{-3}	Pas

Table 1.— Material parameters for the considered poroelastic problem.

triangulation until to achieve a target grid with the desired fine scale to approximate the solution of the problem. For all the numerical experiments performed next, the stopping criterion per time step is that the absolute residual should be less than 10^{-6}.

First of all, a standard smoother is considered in order to see the convergence troubles that appear when trying to solve the system of poroelasticity with the corresponding multigrid method. In particular, a three-color Gauss-Seidel is considered, since in [11] this smoother resulted to have a better performance than other standard smoothers for the Poisson problem. Three color smoother consists of splitting grid nodes into three disjoint sets, with each set having a different color, and simultaneously updating all nodes of the same color. The good convergence properties displayed by this smoother for the Poisson problem are lost when it is used to solve a poroelasticity problem. This can be seen in Figure 6, where the history of the convergence of a multigrid algorithm based on three-color smoother, applied to the considered poroelasticity problem, is shown. An $F(2,1)$ cycle is used, and for different numbers of refinement levels it is shown that this smoother is not robust with respect to the space discretization parameter, since the number of iterations, necessary to reach the desired value for the residual, grows up as the number of refinement levels increase, yielding a significative deterioration of the convergence. Besides, even divergence can be seen for some refinement levels.

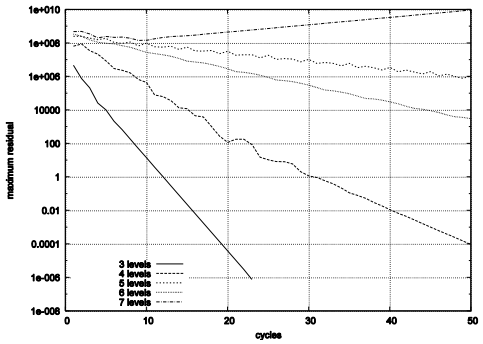


Figure 6.— History of the convergence of multigrid based on three-color smoother for poroelasticity.

This unsatisfactory convergence of the multigrid method, is improved by considering box-relaxation. Next, results for both box-smoothers previously introduced, are presented. In Figure 7, the obtained results for the convergence of the multigrid method based on cell-wise (right-hand panel) and point-wise box-smoothers (left-hand panel) are displayed. It is observed a very good behavior of these methods for the poroelasticity problem, and it can be seen that even the convergence improves when a finer grid is considered. Besides, a similar behavior of both smoothers is reported, being cheaper the cell-wise box-relaxation.

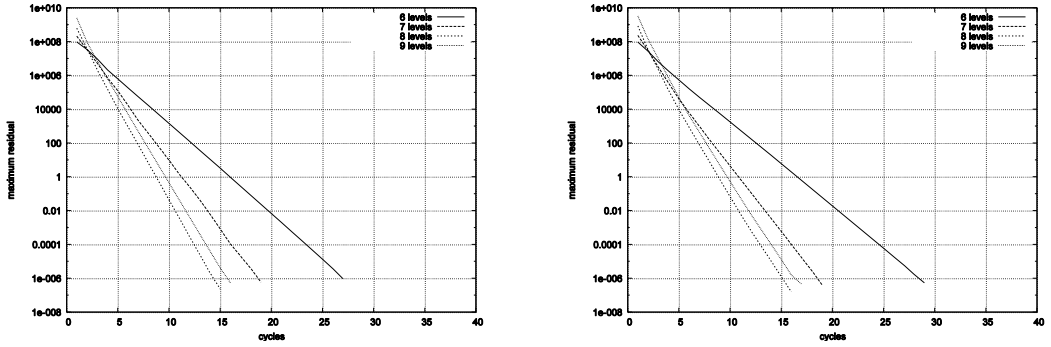


Figure 7.— Convergence of point-wise and cell-wise box-smoothers, respectively, for different numbers of refinement levels.

Finally, the behavior of $V-$ and $F-$cycles for cell-wise and point-wise box-smoothers is analyzed. In Figure 8, the convergence obtained with $V(2,1)$ and $F(2,1)-$ cycles is displayed for both smoothers. It is observed that $F-$ cycles provide good convergence, whereas $V-$cycles yield to divergence of the method. This is due to the poor coarse-grid correction which appears when stabilization terms are added to the equations. $F-$cycles, which invest more work on coarser grids, can overcome these troubles, but on the contrary $V-$cycles do not manage it. Moreover, an increase on the number of pre- and post-smoothing steps does not improve significantly the results. Some techniques to beat these problems, as residual overweighting and defect correction strategies, see [8], will be investigated in the future.

References

[1] Aguilar, G., Gaspar, F., Lisbona, F., Rodrigo, C.: 2008, "Numerical stabilization of Biot's consolidation model by a perturbation on the flow equation", *Int. J. Numer. Meth. Engng.* **75**, 1282–1300.

[2] Benzi, M., Golub, G.H., Liesen, J.: 2005, "Numerical solution of saddle point problems", *Acta Numerica* **14**, 1–137. Cambridge University Press, United Kingdom.

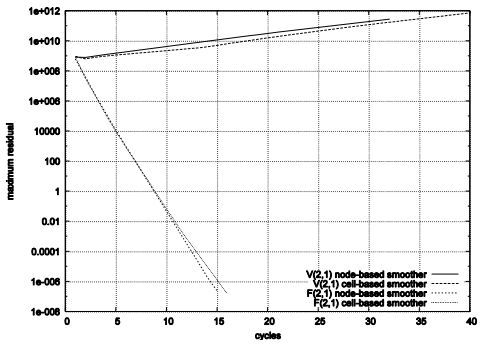


Figure 8.— Convergence of point-wise and cell-wise box-smoothers, respectively, for $F(2,1)$ and $V(2,1)$ cycles.

[3] Bergen, B., Gradl, T., Hülsemann, F., Rüde, U.: 2006, "A massively parallel multigrid method for finite elements". *Comput. Sci. Eng.* **8**, 56–62.

[4] Biot, M.: 1941, "General theory of three dimensional consolidation". *J. Appl. Phys.* **12**, 155–169.

[5] Biot M.: 1955, "Theory of elasticity and consolidation for a porous anisotropic solid". *J. Appl. Phys.* **33**:182–185.

[6] Biot M.: 1956 "General solutions of the equation of elasticity and consolidation for a porous material". *J. Appl. Mech.* **78**:91–96.

[7] Brandt, A.: 1977, "Multi-level adaptive solutions to boundary-value problems". *Math. Comput.* **31**, 333–390.

[8] Brandt, A., Yavneh, I.: 1993, "Accelerated multigrid convergence and high-Reynolds recirculating flows". *SIAM J. Sci. Comput.* **14**, 607–626.

[9] Brezzi, F.: 1974, "On the existence, uniqueness and approximation of saddle-point problems arising from Lagrange multipliers". *RAIRO Modelisation Mathematique et Analyse Numerique* **8**, 129–151.

[10] Gaspar, F.J., Lisbona, F.J., Oosterlee, C.W., Wienands, R.: 2004, "A systematic comparison of coupled and distributive smoothing in multigrid for the poroelasticity system". *Numer. Linear Algebra Appl.* **11**, 93–113.

[11] Gaspar, F.J., Gracia, J.L., Lisbona, F.J.: 2009, "Fourier analysis for multigrid methods on triangular grids". *SIAM J. Sci. Comput.* **31**, 2081–2102.

[12] Hackbusch, W.: 1985, *Multi-grid methods and applications*. Springer, Berlin.

[13] John, V., Tobiska, L.: 2000, "Numerical performance of smoothers in coupled multigrid methods for the parallel solution of the incompressible Navier-Stokes equations". *Int. J. Numer. Meth. Fluids* **33**, 453–473.

[14] Korsawe, J., Starke, G., Wang, W., Kolditz, O.: 2006, "Finite element analysis of poro-elastic consolidation in porous media: standard and mixed approaches.". *Computer Methods in Applied Mechanics and Engineering* **195**, 1096–1115.

[15] Lewis, R.W., Schrefler B.A.: 1998, *The Finite Element Method in the Static and Dynamic Deformation and Consolidation of Porous Media*. Wiley, New York.

[16] Mira, P., Pastor M., Li, T., Liu, X.: 2003, "A new stabilized enhanced strain element with equal order of interpolation for soil consolidation problems". *Computer Methods in Applied Mechanics and Engineering* **192**, 4257–4277.

[17] Pastor M., Li, T., Liu, X., Zienkiewicz, O.C.: 1999, "Stabilized low-order finite elements for failure and localization problems in undrained soils and foundations". *Computer Methods in Applied Mechanics and Engineering* **174**, 219–234.

[18] Terzaghi, K.: 1943, *Theoretical Soil Mechanics*. John Wiley, New York.

[19] Trottenberg, U., Oosterlee, C.W., Schüller, A.: 2001, *Multigrid*. Academic Press, New York.

[20] Turek, S.: 1999, *Efficient solvers for incompressible flow problems: an algorithmic and computational approach*. Springer, Berlin.

[21] Vanka, S.P.: 1986, "Block-implicit multigrid solution of Navier-Stokes equations in primitive variables". *J. Comput. Phys.* **65**, 138–158.

[22] Wobker, H., Turek, S.: 2009, "Numerical studies of Vanka-type smoothers in computational solid mechanics". *Adv. Appl. Math. Mech.* **1**, 29–55.

Monografías de la Real Academia de Ciencias de Zaragoza **33**, 39–53, (2010).

Non-negative Matrix Factorisation for Network Reordering

Clare M. Lee, Desmond J. Higham

Department of Mathematics and Statistics, University of Strathclyde, Glasgow G1 1XH.

and

Daniel Crowther, J. Keith Vass

Translational Medicine Research Collaboration, University of Dundee, Dundee DD1 5EH

Abstract

Non-negative matrix factorisation covers a variety of algorithms that attempt to represent a given, large, data matrix as a sum of low rank matrices with a prescribed sign pattern. There are intuitave advantages to this approach, but also theoretical and computational challenges. In this exploratory paper we investigate the use of non-negative matrix factorisation algorithms as a means to reorder the nodes in a large network. This gives a set of alternatives to the more traditional approach of using the singular value decomposition. We describe and implement a range of recently proposed algorithms and evaluate their performance on synthetically constructed test data and on a real data set arising in cancer research.

1 Introduction

Many large, complex networks contain hidden substructures that can be revealed using a range of post-processing algorithms. In particular, reordering the network nodes appropriately may help to summarise key properties by exposing significant clusters, or more generally sets of neighbours with similar features. This work focuses on the use of matrix factorisation methods to derive network reorderings.

Non-negative matrix factorisation (NMF) is a relatively new matrix computation tool that has been applied to problems in data compression and feature extraction [1, 2, 6]. We aim here to study the potential for NMF in network reordering. Our target applications concerns the behaviour of genes and proteins in cells. Microarray data produces large non-square matrices of information recording the behaviour of many genes across a small number of samples. This data is by its very nature non-negative. A common aim is then to cluster or order the genes/samples into groups where members behave similarly to each

other and differently to those in other groups. This allows us to find sets of genes whose behaviour distinguishes different sample types. As we are particularly interested in using the technique to classify clusters of samples from microarray data we typically talk about the rows of a matrix being genes and the columns as samples. Currently this is often done using the singular value decomposition. [3, 4].

In Section 2 we introduce non-negative matrix factorisation, before giving specific algorithms in Section 3. In this section we also describe how we use the factorisation for reordering. In Section 4 we test all the algorithms on synthetic data before looking at a real data set in Section 5.

2 Non-negative Matrix Factorisation

In general, given an input matrix A with non-negative entries, $a_{i,j} \geq 0$, NMF produces lower rank non-negative factors W and H, so that

$$A \approx WH, \tag{1}$$

with $A \in \mathbb{R}^{m \times n}$, $W \in \mathbb{R}^{m \times k}$ and $H \in \mathbb{R}^{k \times n}$, for $k \ll \min(m, n)$. We often express the two factors in terms of their rows or columns, that is, $W = [w_1, \ldots, w_k]$, and $H = [h_1, \ldots, h_k]^{\mathrm{T}}$. NMF techniques, which are relatively new in the context of network reordering, have the intuitive advantage of respecting the non-negative nature of the original data. It has been shown in other areas of research that the outer product of a column of the first factor w_i and a row of the second h_i may pick out a feature of the original data [6]. For network reordering this has the potential to generate bases of "eigen-genes" or "eigen-samples" that combine non-negatively to represent the expression of a particular gene across the samples or the gene expression for a particular sample type. In this setting the (i, j)th entry of $W_{i,j}$ is the level of expression of gene i in eigen-sample j, and $H_{i,j}$, the (i, j)th entry of H, is the importance of eigen-gene i in sample j.

Although NMFs take advantage of the non-negative nature of the original data, they have various theoretical and practical drawbacks that are yet to be fully explored in the network reordering context. Different NMF variations can produce very different results, and the iterative nature of the underlying computations makes them highly sensitive to the choice of target rank and initial starting guess.

3 Specific NMF Algorithms

There are many different NMF algorithms. We look at five variants [7].

Multiplicative Update Here both factors need to be initialised with starting guesses

W^0 and H^0, typically random matrices. The algorithm is then:

$$\left.\begin{aligned} H^{i+1} &= H^i . * \frac{{W^i}^{\mathrm{T}} A}{{W^i}^{\mathrm{T}} W^i H^i + 10^{-9}} \\ W^{i+1} &= W^i . * \frac{A {H^i}^{\mathrm{T}}}{W^i H^i {H^i}^{\mathrm{T}} + 10^{-9}} \end{aligned}\right\} \tag{2}$$

where $.*$ denotes point-wise multiplication.

Alternating Least Squares Only W^0 needs to be initialised. The algorithm is

$$\left.\begin{aligned} &\text{Solve for } H^{i+1} \text{ in } W^i H^{i+1} = A \\ &\text{Set all negative entries in } H^{i+1} \text{ to } 0 \\ &\text{Solve for } W^{i+1} \text{ in } W^{i+1} H^{i+1} = A \\ &\text{Set all negative entries in } W^{i+1} \text{ to } 0 \end{aligned}\right\} \tag{3}$$

Tri-factorisation Here A is factorised into three factors so that $A \approx WSH$ with $W \in \mathbb{R}^{m \times k}$ $S \in \mathbb{R}^{k \times \ell}$ and $H \in \mathbb{R}^{\ell \times n}$, with $k, \ell \ll \min(m, n)$. This allows different numbers of clusters in the rows and columns of the data. After initialising all the factors with random guesses the algorithm is then:

$$\left.\begin{aligned} H^{i+1} &= H^i . * \sqrt{\frac{{S^i}^{\mathrm{T}} {W^i}^{\mathrm{T}} A}{{S^i}^{\mathrm{T}} {W^i}^{\mathrm{T}} A {H^i}^{\mathrm{T}} H^i + 10^{-9}}} \\ W^{i+1} &= W^i . * \sqrt{\frac{A {H^i}^{\mathrm{T}} {S^i}^{\mathrm{T}}}{W^i {W^i}^{\mathrm{T}} A {H^i}^{\mathrm{T}} {S^i}^{\mathrm{T}} + 10^{-9}}} \\ S^{i+1} &= S^i . * \sqrt{\frac{{W^i}^{\mathrm{T}} A {H^i}^{\mathrm{T}}}{{W^i}^{\mathrm{T}} W^i S^i H^i {H^i}^{\mathrm{T}} + 10^{-9}}} \end{aligned}\right\} \tag{4}$$

Both the multiplicative update method and the alternating least squares algorithms are included in the MATLAB statistics toolbox[1]. There are also partial-non-negative factorisations, which allow an input matrix of mixed sign and produce factorisations where one of the two factors is non-negative. We will consider the case where $A \approx WH$ as before but with A and W allowed to have negative entries. Two examples of these are;

Semi-NMF Only H^0 is initialised with the algorithm being:

$$\left.\begin{aligned} W^{i+1} &= A {H^i}^{\mathrm{T}} (H^i {H^i}^{\mathrm{T}})^{-1} \\ H^{i+1} &= H^i . * \sqrt{\frac{({W^i}^{\mathrm{T}} A)^+ + ({W^i}^{\mathrm{T}} W^i)^- H^i}{({W^i}^{\mathrm{T}} A)^- + ({W^i}^{\mathrm{T}} W^i)^+ H^i}} \end{aligned}\right\} \tag{5}$$

[1] `http://www.mathworks.com/`

Convex-NMF In this factorisation the columns of W lie in the space spanned by the columns of A, i.e. $W = AR$ so $A \approx ARH$. The algorithm is;

$$\left.\begin{aligned} H^{i+1} &= H^i .* \sqrt{\frac{{R^i}^{\mathrm{T}}(A^{\mathrm{T}}A)^+ + {R^i}^{\mathrm{T}}(A^{\mathrm{T}}A)^- R^i H^i}{{R^i}^{\mathrm{T}}(A^{\mathrm{T}}A)^- + {R^i}^{\mathrm{T}}(A^{\mathrm{T}}A)^+ R^i H^i}} \\ R^{i+1} &= R^i .* \sqrt{\frac{(A^{\mathrm{T}}A)^+ {H^i}^{\mathrm{T}} + (A^{\mathrm{T}}A)^- R^i H^i {H^i}^{\mathrm{T}}}{(A^{\mathrm{T}}A)^- {H^i}^{\mathrm{T}} + (A^{\mathrm{T}}A)^+ R^i H^i {H^i}^{\mathrm{T}}}} \end{aligned}\right\} \tag{6}$$

where $(M)^+ = \dfrac{(|M| + M)}{2}$ and $(M)^- = \dfrac{(|M| - M)}{2}$, the positive and negative parts of the matrix.

In all cases the aim is that a row/column of the factorisation matrices W and H conveys information on a single feature of the data. For $k = 1$ all the algorithms produce the same result. In this case W is the first left singular vector of A and H is the first right singular vector. However, for $k > 1$ the algorithms all differ.

3.1 Permutation

We can show that one positive feature as of all these algorithms is that they are impervious to permutation of the original matrix. This means that the initial ordering of the nodes in the network has no influence on the final reordering produced by the algorithm.

This can be verified for all algorithms using the properties of permutation matrices. By way of example we show the calculations for the method of multiplicative update. Letting P_1 and P_2 denote arbitrary permutation matrices and starting with $\widetilde{W^0} = P_1 W^0$ and $\widetilde{H^0} = H^0 P_2^{\mathrm{T}}$, the factorisation of $P_1 A P_2^{\mathrm{T}}$ gives that;

$$\begin{aligned} \widetilde{W^1} &= P_1 W^0 .* \frac{P_1 A P_2^{\mathrm{T}} \left(H^0 P_2^{\mathrm{T}}\right)^{\mathrm{T}}}{P_1 W^0 H^0 P_2^{\mathrm{T}} \left(H^0 P_2^{\mathrm{T}}\right)^{\mathrm{T}} + 10^{-9}} \\ &= P_1 W^0 .* \frac{P_1 A {H^0}^{\mathrm{T}}}{P_1 W^0 H^0 {H^0}^{\mathrm{T}} + 10^{-9}} \\ &= P_1 \left(W^0 .* \frac{A {H^0}^{\mathrm{T}}}{W^0 H^0 {H^0}^{\mathrm{T}} + 10^{-9}} \right) \\ &= P_1 W^1 \end{aligned}$$

and

$$\begin{aligned}
\widetilde{H^1} &= H^0 P_2^{\mathrm{T}} . * \frac{\left(P_1 W^0\right)^{\mathrm{T}} P_1 A P_2^{\mathrm{T}}}{\left(P_1 W^0\right)^{\mathrm{T}} P_1 W^0 H^0 P_2^{\mathrm{T}} + 10^{-9}} \\
&= H^0 P_2^{\mathrm{T}} . * \frac{{W^0}^{\mathrm{T}} A P_2^{\mathrm{T}}}{{W^0}^{\mathrm{T}} W^0 H^0 P_2^{\mathrm{T}} + 10^{-9}} \\
&= \left(H^0 . * \frac{{W^0}^{\mathrm{T}} A}{{W^0}^{\mathrm{T}} W^0 H^0 + 10^{-9}} \right) P_2^{\mathrm{T}} \\
&= H^1 P_2^{\mathrm{T}} .
\end{aligned}$$

Therefore the result follows by induction.

3.2 Ordering and Clustering

Using ideas from [2] we put the rows of W and the columns of H into k clusters. Each row of W is placed into a cluster according to the most highly expressed column, e.g. row i of W belongs to cluster j if $W_{i,j}$ is the maximum over $W_{i,\{1,\ldots,k\}}$. Similarly for the columns of H, column j of H belongs to cluster i if $H_{i,j}$ is the maximum over $H_{\{1,\ldots,k\},j}$. Each cluster is then sorted independently by size of element in that column/row. So the rows of W in cluster j are sorted in order of the jth column of W, and the columns of H in cluster i are sorted in order of the ith row of H. The orderings are then put together into a single reordering vector with the reordered cluster 1 appearing before the reordered cluster 2 and so on. The MATLAB code for clustering and reordering W reads:

```
% finding the clusters
for i=1:m
    [b,clustW(i)]=max(W(i,:));
end

% sorting the clusters
for i=1:k
    k_ind=find(clustW==i);
    [b,tind]=sort(W(k_ind,i));
    indW(kind)=kind(tind);
end

% sorting the ordering so we have cluster 1
% then cluster 2 etc...
[d,cind]=sort(clustW);
indW=indW(cind);
clustW=d;
```

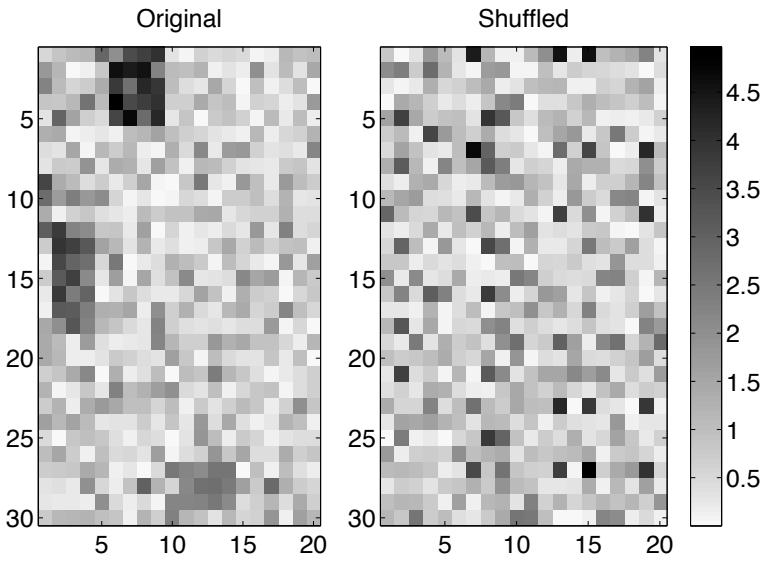


Figure 1.— The left panel has the original test data with three hidden blocks. The panel on the right shows this matrix with the rows and columns shuffled independently; this is the matrix that is factorised.

where `indW` contains the ordering indices for the rows and `clustW` has the reordered cluster numbers. The same method is used to sort the columns of H.

4 Performance on test data

The first test for the algorithms was whether they could find significant blocks placed in a random matrix. The test matrix, generated by a pseudo-random number generator, has three blocks with higher values, as shown in the left panel of Figure 1. The highest block is referred to as block 1 and the lowest is block 3. The rows and columns of this matrix are then shuffled independently to produce the picture on the right of Figure 1. This is the matrix the algorithms are tested on. In all the figures the darker colours represent the higher values in the matrix as shown by the colourbar in this figure. Each algorithm is run with 10 different random initial conditions and the factorisation that produces the the lowest approximation error in the Frobenius norm $\|A - WH\|_F$ is chosen to represent the factorisation method.

The results are shown in Figures 2–7. The first three panels of the top row show the matrix reshuffled using the relative sizes of the first, second and third rows of W and the first second and third columns of H respectively. The fourth shows them clustered and ordered as in Section 3.2. The bottom row shows the ordering and clusters in the clustered version with $*$ denoting the original block 1, $\circ$ block 2, and $+$ block 3; the remaining rows and columns are shown by dots. The vertical axis shows the cluster number with the reordered (as in Section 3.2) position on the horizontal axis. The graph on the left shows

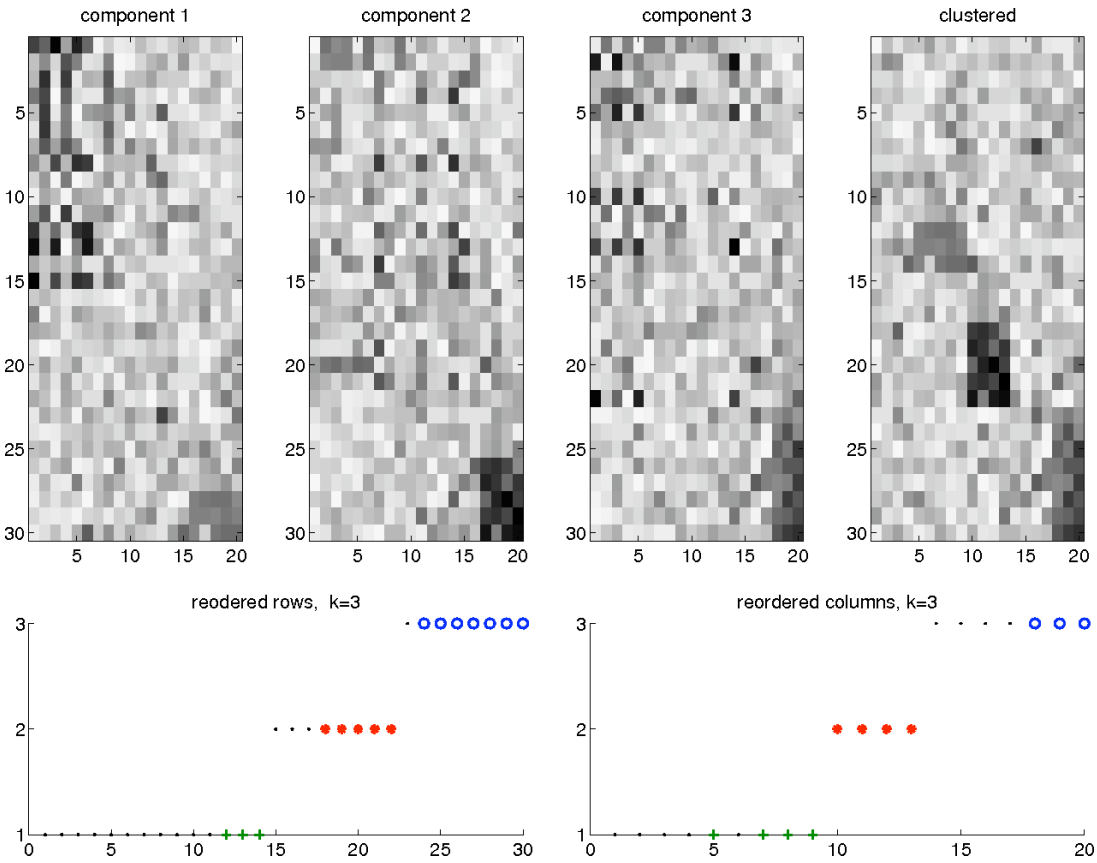


Figure 2.— Results using multiplicative update with $k = 3$.

the reordering of the rows and the graph on the right that for the columns.

Figure 2 shows the results using multiplicative update (2) for $k = 3$. We see that ordering purely on the magnitude of the first column of W and first row of H finds one of the blocks in the original matrix, as do the second and third components. Using the clustering and ordering given in Section 3.2 we see all the blocks appearing with one block in each cluster. For this algorithm we get that $\|A - WH\|_F = 0.55737$.

The results from the alternating least squares method (3) are in Figure 3 for $k = 3$. The algorithm converges to a rank one solution rather than a rank 3 solution. All but the first column of W and the first row of H are zeros. There is only one "cluster" found, but the row reordering nearly finds two of the three blocks. The column ordering is not as good. This algorithm is also slower than the multiplicative update, and the approximation isn't as close to the original matrix with $\|A - WH\|_F = 0.83243$.

With the tri-factorisation (4) the clusters are no longer necessarily in the same order in the rows and columns; this can be seen in Figure 4. However looking at the matrix S, shown in Figure 5, we can see that row cluster i corresponds to column cluster j if $S_{i,j} = \max_{p \in \{1,\ldots,\ell\}} S_{i,p}$. Similarly, column cluster j corresponds to row cluster i if $S_{i,j} = \max_{p \in \{1,\ldots,k\}} S_{p,j}$. If $k \neq \ell$ it is possible that there are row or column clusters

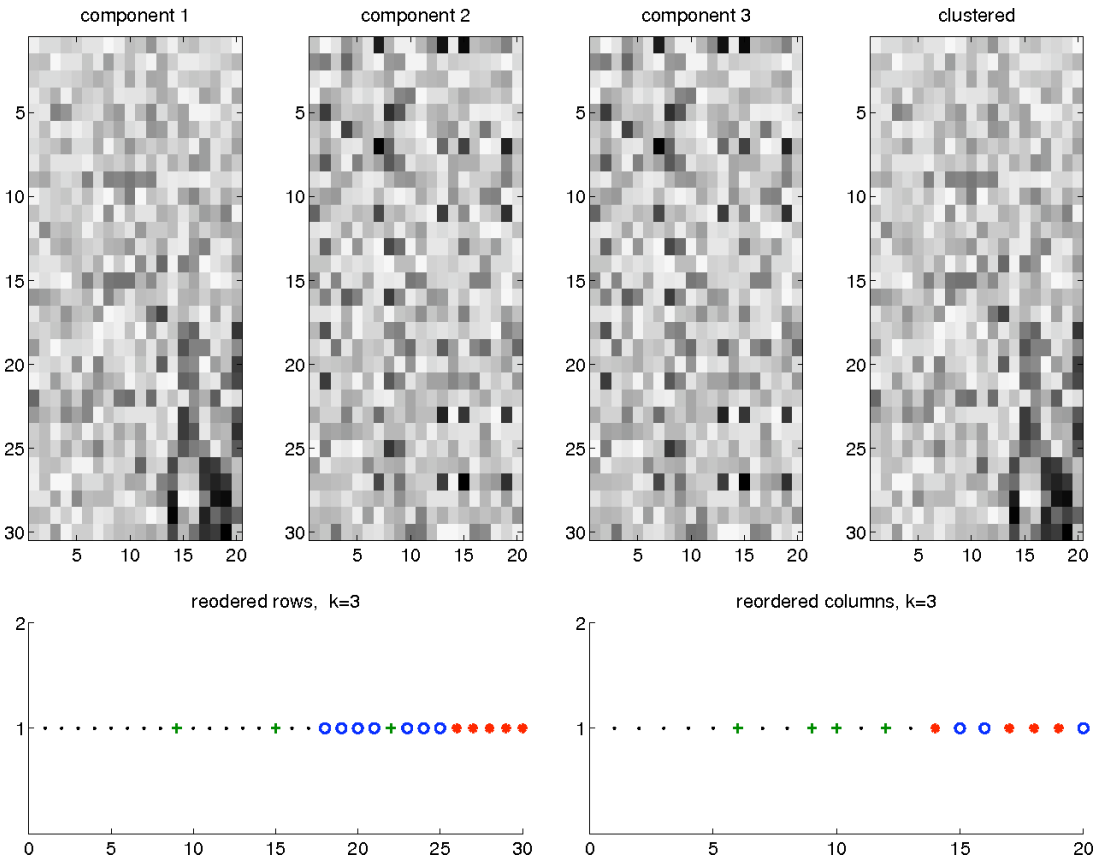


Figure 3.— Results using alternating least squares with $k = 3$.

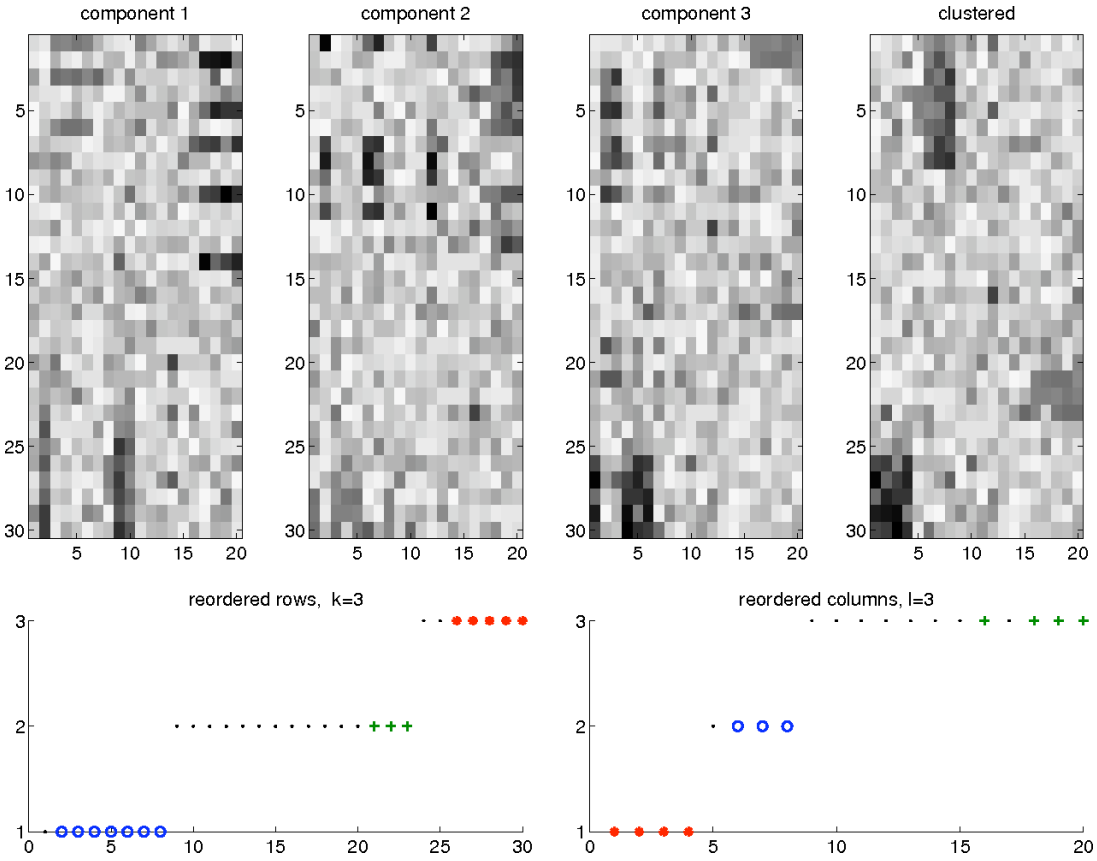


Figure 4.— Results using tri-factorisation with $k = 3$ and $l = 3$.

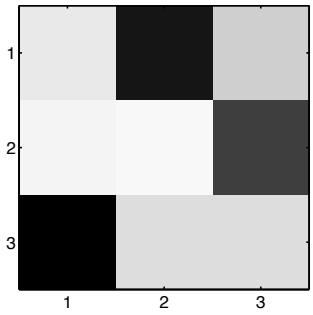


Figure 5.— A heat map of the middle factor S of the factorisation.

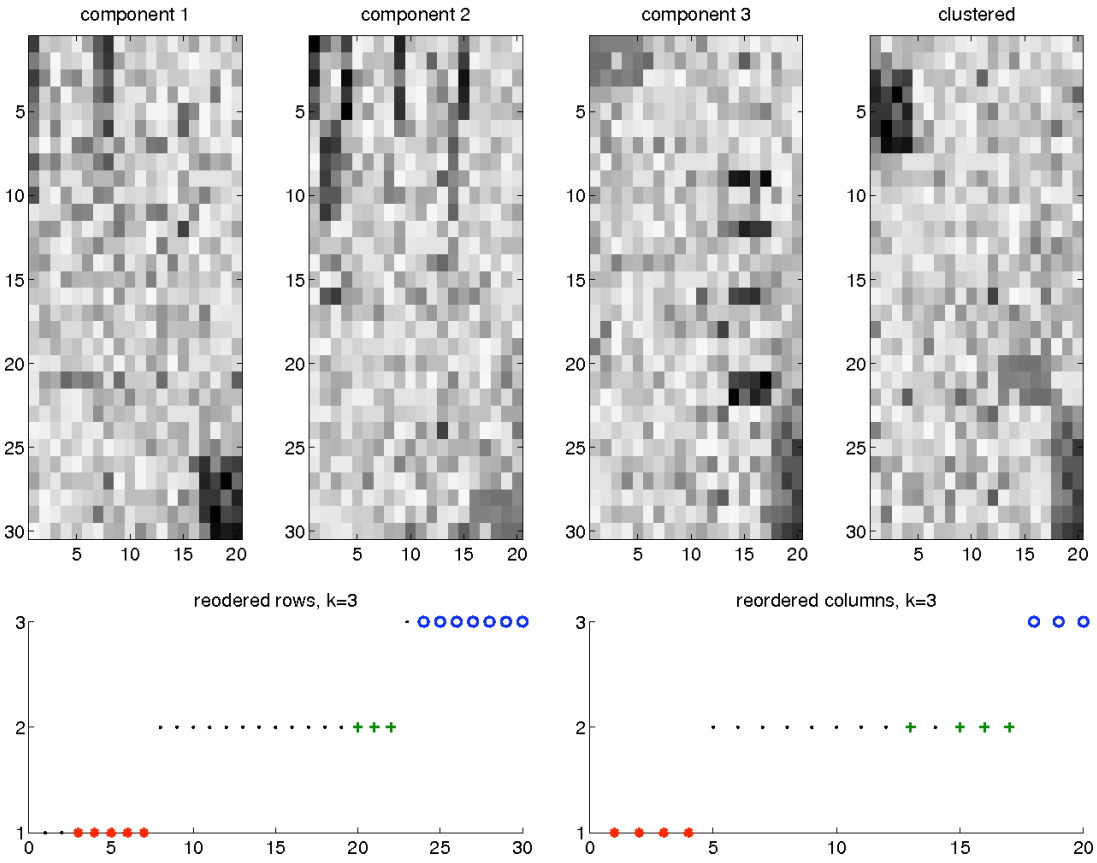


Figure 6.— Results using semi-NMF with $k = 3$.

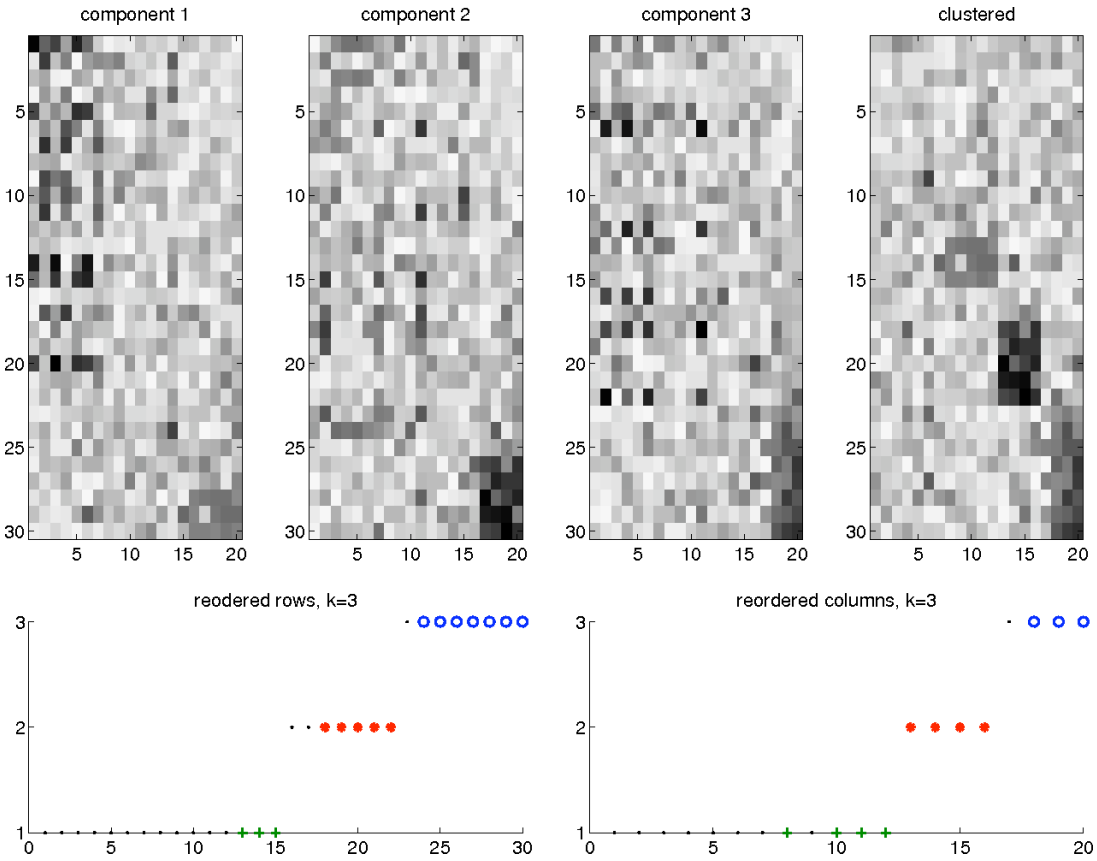


Figure 7.— Results using convex-NMF with $k = 3$.

that do not corrrespond to a cluster the other dimension, usually this row/column of S doesn't have any particularly large values. It is also possible that one row/column cluster corresponds to more than one different column/row clusters, these are still the highest values in S, usually noticeably larger than the other values. For this test example with $k = \ell = 3$ the results are very similar to those using multiplicative update, but with the row/column cluster shuffled. The algorithm also takes about the same time as the multiplicative update and $\|A - WSH\|_F = 0.57826$, similar to that of the multiplicative update. The major difference between the two algorithms is that here we can have different numbers of row clusters and column clusters. In some applications this can be a benefit, however, it does give another parameter that needs to be fixed before the factorisation can begin.

Semi-NMF (5), shown in Figure 6, is the fastest of the algorithms and also produces good results, with $\|A - WH\|_F = 0.55720$.

Initialising both R and H with random matrices the convex-NMF algorithm (6), shown in Figure 7, does as well as any of the other algorithms but it takes a lot longer to run with the error being on a par with the other algorithms at $\|A - WH\|_F = 0.56763$. It is considerably faster if $R^0 = g/(g' * g)$, where g is a random matrix, and the results are similar.

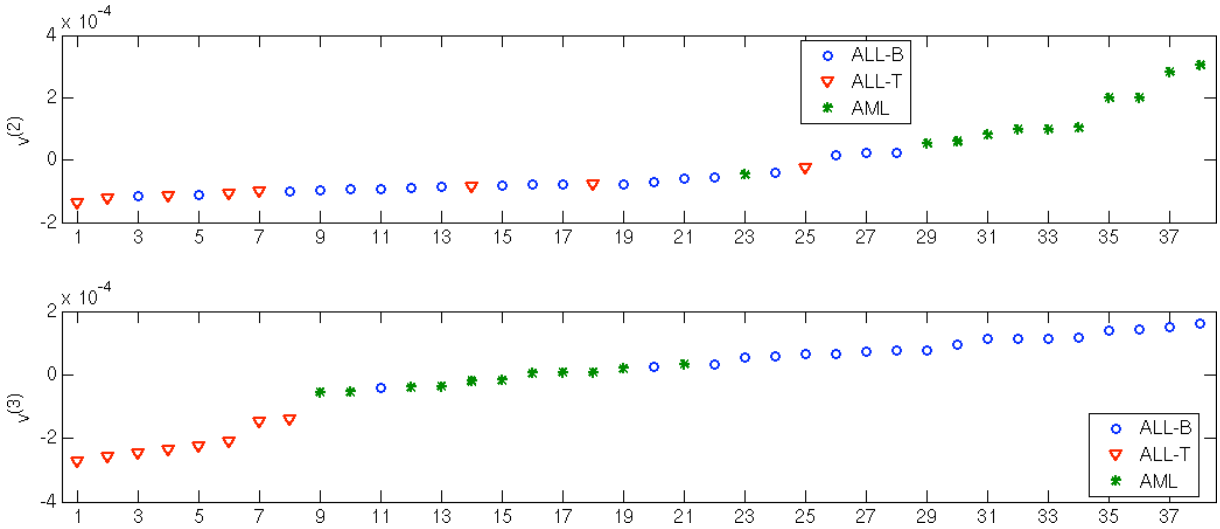


Figure 8.— The ordering as given by the singular value decomposition. The vertical axis has the value of the normalised second or third right singular vector in the top and bottom images respectively.

5 Performance on a cancer microarray data set

We have further tested the algorithms on microarray data from bone marrow samples from patients with acute leukaemia. In the data set the rows represent different genes and the columns are the samples. An individual entry is then the level of gene expression in that sample. The matrix is therefore long and thin with 5000 rows and only 38 columns. There are different types of leukaemia represented here: acute myelogenous leukaemia (AML) and acute lymphoblastic leukaemia (ALL). The second type is further divided into T and B cell subtypes. This data set has been widely used to test algorithms, see for example [2, 5]. It contains two samples that are misclassified by many methods, these are included in our analysis. This time the factorisation with the lowest approximation error is chosen from five different random initial conditions.

For comparison we first show the results achieved by using the singular value decomposition on a normalised data set, see [4] for details. These can be found in Figure 8; where we see the sorted index numbering along the horizontal axis with the value of the right singular vector on the vertical axis. The top graph shows the samples sorted using the second right singular vector, here we see that with one exception the AML type and the ALL types are split. The lower graph has the same information, but sorted by the third singular vector. This time the three different types are split with only two exceptions.

The results from the multiplicative update are in Figure 9. In this figure, and in those following, the value on the vertical axis is the cluster number. This time we see that choosing $k = 2$, as in the top graph, we correctly split the ALL and AML type with two exceptions. For $k = 3$ in the bottom graph we have split the three different leukaemia types again with two exceptions. As we cluster the results from the non-

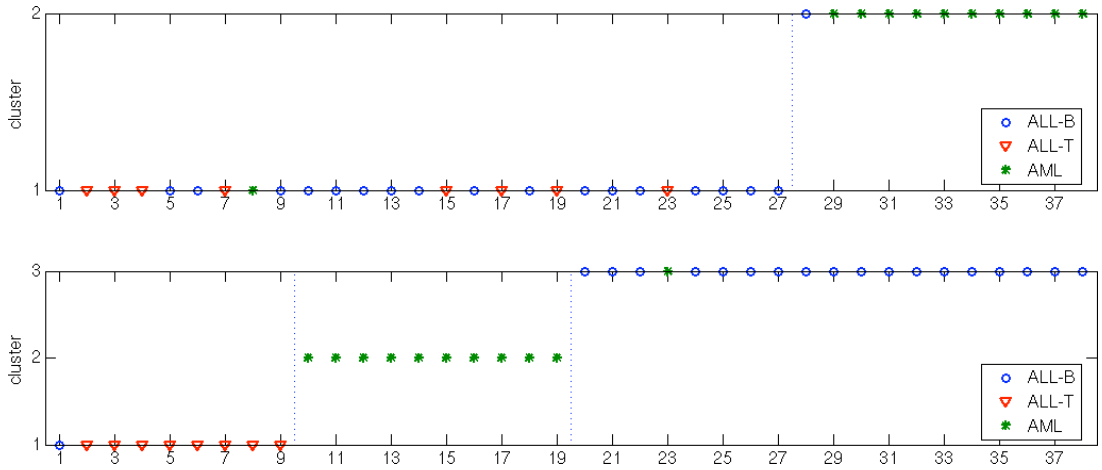


Figure 9.— The ordering as given by the multiplicative update. The vertical axis has the cluster number for $k = 2$, or $k = 3$ in the top and bottom images respectively.

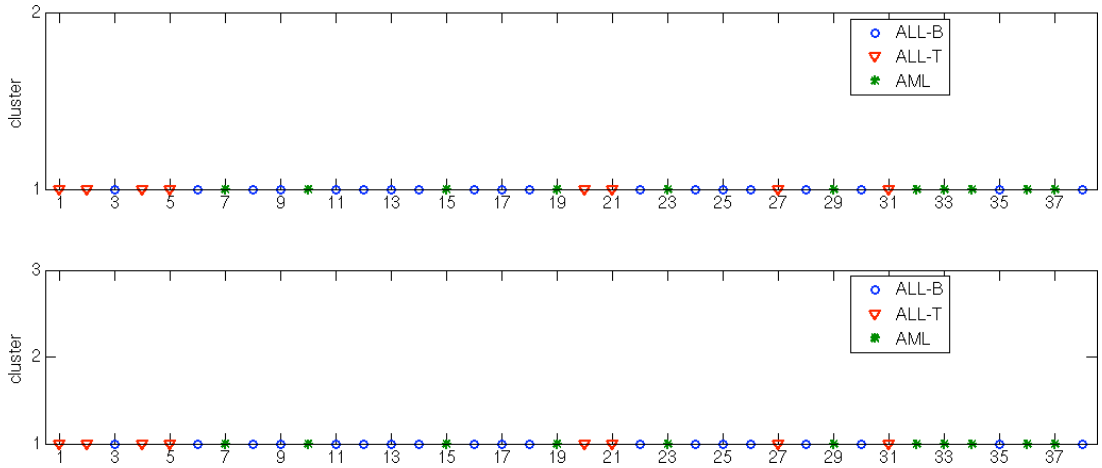


Figure 10.— The ordering as given by the alternating least squares algorithm. The vertical axis has the cluster number for $k = 2$, or $k = 3$ in the top and bottom images respectively.

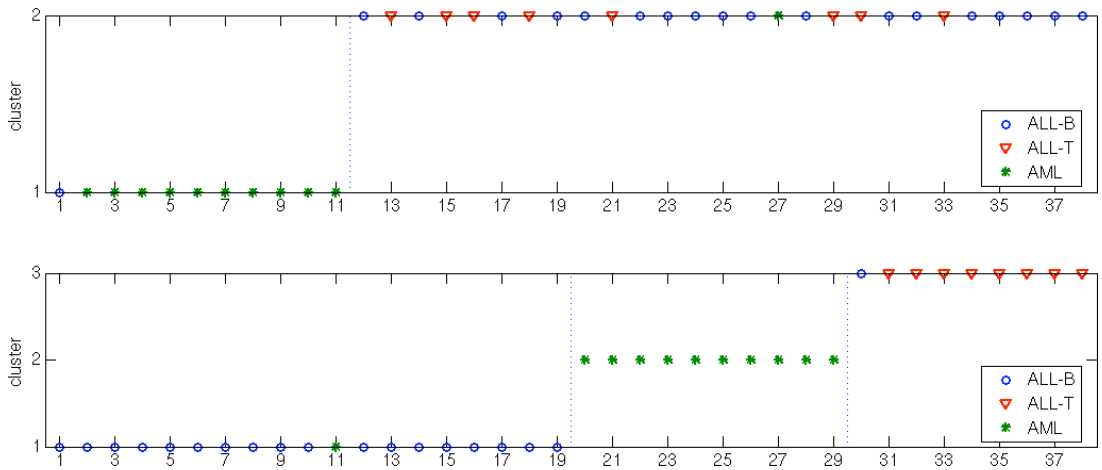


Figure 11.— The ordering as given by tri-factorisation. The vertical axis has the cluster number for $k = 2$, or $k = 3$ in the top and bottom images respectively.

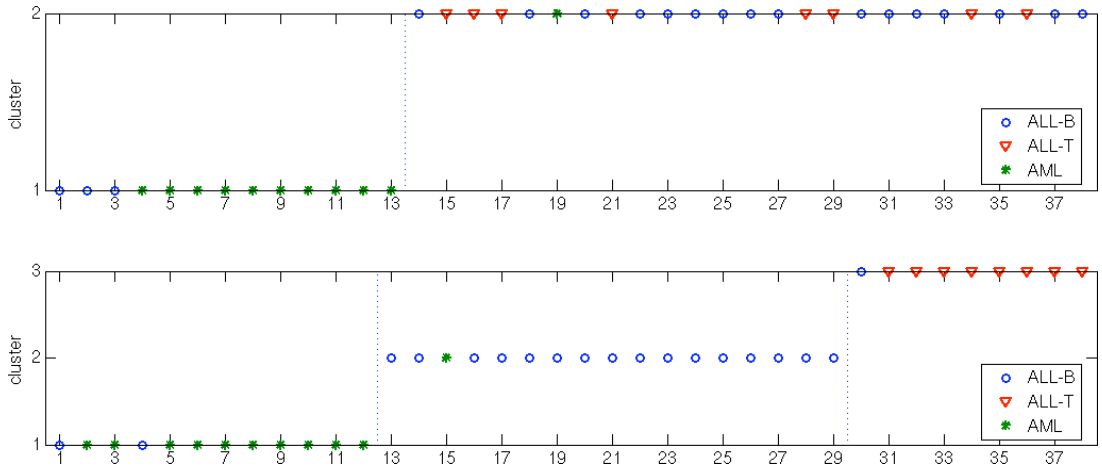


Figure 12.— The ordering as given by semi-NMF. The vertical axis has the cluster number for $k = 2$, or $k = 3$ in the top and bottom images respectively.

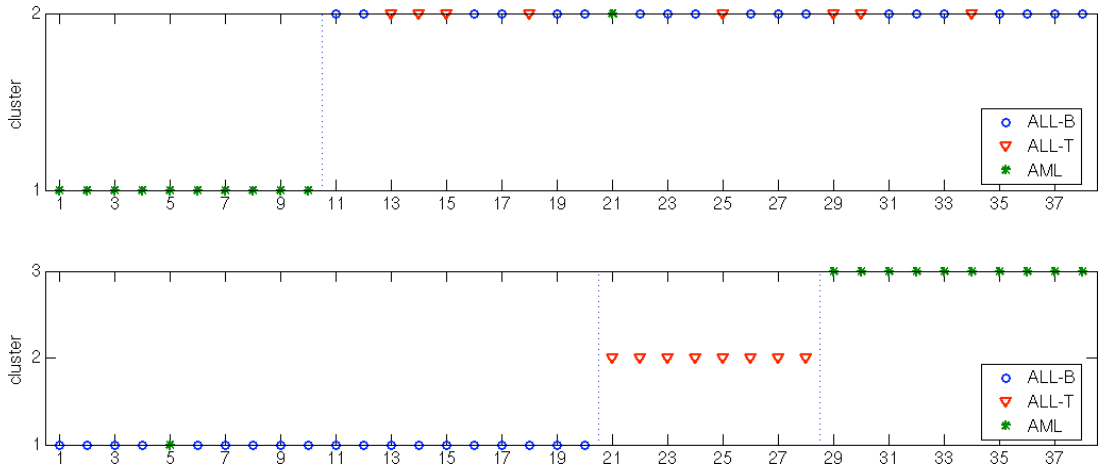


Figure 13.— The ordering as given by convex-NMF. The vertical axis has the cluster number for $k = 2$, or $k = 3$ in the top and bottom images respectively.

negative factorisations, it has the benefit of clear separation of the cell types unlike the singular value decomposition approach where the split is less clear.

The alternating least squares algorithm does a bad job of clustering or ordering the samples, just as it did with the test data in Section 4. The results are shown in Figure 10, where we see that the algorithm produces a rank one solution regardless of the choice of k, and even ordering on the size of the elements of the rank one matrix H does not split the data into the different groups.

As we have mentioned before, it is possible to have different numbers of row and column clusters when using tri-factorisation. However, for ease of comparison we present results for $k = l = 2$, or 3; see Figure 11. For both $k = l = 2$ and $k = l = 3$, we get very similar results to those from the multiplicative update method, though it takes longer. This method is of more benefit in situations where having different numbers of row and column clusters gives a clearer picture.

The semi-NMF algorithm is one of the quicker algorithms, though this test shows that it isn't as accurate in splitting the different leukaemia types. The results in Figure 12 show that with $k = 2$ in the top picture one AML type is placed amongst the ALL types, but three of the ALL types are also placed in the cluster dominated by the AML type. Increasing k to 3 doesn't increase the accuracy with four samples mis-assigned.

The convex-NMF algorithm, despite being able to factorise a mixed sign matrix into the product of a mixed sign and a non-negative matrix, produces clusters that split the different types. For $k = 2$ in the top row of Figure 13 shows that there is a single sample misplaced. When $k = 3$ as in the bottom graph there are only the two samples mis-assigned, these results are on a level with the multiplicative update, and the tri-factorisation.

6 Summary

In this paper we have investgated the use of non-negative matrix factorisation algorithms for reordering data sets and applied them to genetic microarray data. If we look solely at the ability to distinguish the different clusters it would appear that the multiplicative update, tri-factorisation and the convex-NMF algorithms are the best algorithms to use. However, since the tri-factorisation and convex-NMF are much slower, the multiplicative update appears to be the winner in most settings. The other two methods both have added features though, making them possibly more useful in other contexts.

The algorithms have both practical and theoretical challenges that need to be addressed. A basic issue is how to choose the rank of the desired factorisation. How do we choose a value of k (and ℓ) which gives full information from the data without producing misleading results? From a practical point of view it would be ideal to be able to evaluate the data set and then compute one factorisation. However this is probably unrealistic, a more feasible answer would be to find some statistic by which to compare the factorisations for different values of k. A further challenge is how to initialise the algorithms since they produce different results for different initial conditions. Currently we make multiple runs with different initial random matrices and use the factorisation that produces the lowest approximation error. With large data sets this can become costly and time consuming, so some investigation into how best to pick the initial guess would be beneficial. From a theoretical perspective, it would be very helpful to have analytical results that allow us to distinguish between the NMF variants in the network reordering application, and compare them with the more traditional singular value decomposition approach.

Acknowlegments

C.M. Lee and D.J. Higham are supported by EPSRC Grant EP/E049370/1

References

[1] Michael W. Berry, Murray Browne, Amy N. Langville, V. Paul Pauca, and Robert J. Plemmons. Algorithms and applications for approximate nonnegative matrix factorisation. *Comput. Stat. Data Anal*, 52:155–173, 2007.

[2] Jean-Philippe Brunet, Pablo Tamayo, Todd R. Golub, and Jill P. Mesirov. Metagenes and molecular pattern discovery using matrix factorisation. *Proc. Nat. Acad. Sci.*, 101(12):4164–4169, 2004.

[3] Desmond J. Higham, Gabriela Kalna, and Milla Kibble. Spectral clustering and its use in bioinformatics. *J. Comput. Appl. Math.*, 204:25–37, 2007.

[4] Desmond J. Higham, Gabriela Kalna, and J. Keith Vass. Analysis of the singular value decomposition as a tool for processing microarray expression data. In *Proceedings of ALGORITMY*, pages 250–259, Slovak Un. of Tech., 2005.

[5] Hyunsoo Kim and Haesun Park. Sparse non-negative matrix factorizations via alternating non-negativity-constrained least squares for microarray data analysis. *Bioinformatics*, 23(12):1495–1502, 2007.

[6] Daniel D. Lee and H. Sebastian Seung. Learning the parts of objects by non-negative matrix factorization. *Nature*, 401:788–791, 1999.

[7] Tao Li and Chris Ding. The relationship among various nonnegative matrix factorization methods for clustering. In *Proc. IEEE Int'l Conf. on Data Mining (ICDM'06)*, pages 362–371, 2006.

Monografías de la Real Academia de Ciencias de Zaragoza **33**, 55–70, (2010).

Fourth-order symplectic exponentially-fitted modified Runge-Kutta methods of the Gauss type: a review

G. Vanden Berghe and M. Van Daele

Vakgroep Toegepaste Wiskunde en Informatica, Universiteit Gent

Krijgslaan 281 - S9, B - 9000 Gent, Belgium

Dedicated to Manuel Calvo for his 65th birthday

Abstract

The construction of symmetric and symplectic exponentially-fitted Runge-Kutta methods for the numerical integration of Hamiltonian systems with oscillatory solutions is reconsidered. In previous papers fourth-order and sixth-order symplectic exponentially-fitted integrators of Gauss type, either with fixed or variable nodes, have been derived. In this paper new fourth-order integrators are constructed by making use of the six-step procedure of Ixaru and Vanden Berghe (*Exponential fitting*, Kluwer Academic Publishers, 2004). Numerical experiments for some oscillatory problems are presented and compared to the results obtained by previous methods.

MSC: 65L05,65L06

Keywords : Exponential fitting, symplecticness, RK-methods, Oscillatory Hamiltonian Systems

1 Introduction

The construction of Runge-Kutta (RK) methods for the numerical solution of ODEs, which have periodic or oscillating solutions has been considered extensively in the literature [1]-[12]. In this approach the available information on the solutions is used in order to derive more accurate and/or efficient algorithms than the general purpose algorithms for such type of problems. In [13] a particular six-step flow chart is proposed by which specific exponentially-fitted algorithms can be constructed. Up to now this procedure has

not yet been applied in all its aspects for the construction of symplectic RK methods of Gauss type.

In principle the derivation of exponentially-fitted (EF) RK methods consists in selecting the coefficients of the method such that it integrates exactly all functions of a particular given linear space, i.e. the set of functions

$$\{1, t, \ldots, t^K, \exp(\pm\lambda t), t\exp(\pm\lambda t), \ldots, t^P \exp(\pm\lambda t)\} \ , \tag{1}$$

where $\lambda \in \mathbb{C}$ is a prescribed frequency. In particular when $\lambda = i\omega, \omega \in \mathbb{R}$ the couple $\exp(\pm\lambda t)$ is replaced by $\sin(\omega t), \cos(\omega t)$. In all previous papers other set of functions have been introduced.

On the other hand, oscillatory problems arise in different fields of applied sciences such as celestial mechanics, astrophysics, chemistry, molecular dynamics and in many cases the modelling gives rise to Hamiltonian systems. It has been widely recognized by several authors [8, 12],[14]-[16] that symplectic integrators have some advantages for the preservation of qualitative properties of the flow over the standard integrators when they are applied to Hamiltonian systems. In this sense it may be appropriate to consider symplectic EFRK methods that preserve the structure of the original flow. In [12] the well-known theory of symplectic RK methods is extended to modified (i.e. by introducing additional parameters) EFRK methods, where the set of functions $\{\exp(\pm\lambda t)\}$ has been introduced, giving sufficient conditions on the coefficients of the method so that symplecticness for general Hamiltonian systems is preserved. Van de Vyver [12] was able to derive a two-stage fourth-order symplectic modified EFRK method of Gauss type with constant knot-points. Calvo *et al.* [2]-[4] have studied two-stage as well as three-stage methods. In their applications for fourth-order methods they consider pure EFRK methods. Their set of functions is the trigonometric polynomial one consisting essentially of the functions $\exp(\pm\lambda t)$ combined with $\exp(\pm 2\lambda t)$.They constructed fourth-order (two-stage case) methods of Gauss type with frequency dependent knot points. On the other hand Vanden Berghe *et al.* have constructed a two-stage EFRK method of fourth-order integrating the set of functions (1) with $(K = 2, P = 0)$ and $(K = 0, P = 1)$, but unfortunately these methods are not symplectic. In addition it has been pointed out in [14] that symmetric methods show a better long time behaviour than non-symmetric ones when applied to reversible differential systems.

In this paper we investigate the construction of two-stage (fourth-order) symmetric and symplectic modified EFRK methods which integrate exactly first-order differential systems whose solutions can be expressed as linear combinations of functions present in the set (1), but also give a review of previous work [2, 12]. Our purpose consists in deriving accurate and efficient modified EF geometric integrators based on the combination of the EF approach, followed from the sixth step flow chart [13], and symmetry and symplectic-

ness conditions. The paper is organized as follows. In Section 2 we present the notations and definitions used in the rest of the paper. In Section 3 we present the previously derived methods of order four. In Section 4 we derive a class of new two-stage symplectic modified EFRK integrators with frequency dependent nodes and based upon some properties of symplectic and symmetric methods also described in [4]. In Section 5 we present some numerical experiments for fourth-order methods with oscillatory Hamiltonian systems and we compare them with the results obtained by other symplectic (EF)RK Gauss integrators given in [2, 12, 14].

2 Notations and definitions

We consider initial value problems for first-order differential systems

$$y'(t) = f(t, y(t)), \quad y(t_0) = y_0 \in \mathbb{R}^m \ . \tag{2}$$

In case of Hamiltonian systems $m = 2d$ and there exits a scalar Hamiltonian function $H = H(t, y)$, so that $f(y) = -J\nabla_y H(t, y)$, where J is the $2d$-dimensional skew symmetric matrix

$$J = \begin{pmatrix} 0_d & I_d \\ -I_d & 0_d \end{pmatrix}, \quad J^{-1} = -J$$

and where $\nabla_y H(t, y)$ is the column vector of the derivatives of $H(t, y)$ with respect to the components of $y = (y_1, y_2, \ldots, y_{2d})^T$. The Hamiltonian system can then be written as

$$y'(t) = -J\nabla_y H(t, y(t)), \quad y(t_0) = y_0 \in \mathbb{R}^{2d} \ . \tag{3}$$

For each fixed t_0 the flow map of (2) will be denoted by $\phi_h : \mathbb{R}^m \to \mathbb{R}^m$ so that $\phi_h(y_0) = y(t_0 + h; t_0, y_0)$. In particular, in the case of Hamiltonian systems, ϕ_h is a symplectic map for all h in its domain of definition, i.e. the Jacobian matrix of $\phi_h(y_0)$ satisfies

$$\phi_h'(y_0) J \phi_h'(y_0)^T = J \ .$$

A desirable property of a numerical method ψ_h for the numerical integration of a Hamiltonian system is to preserve qualitative properties of the original flow ϕ_h such as the symplecticness, in addition to provide an accurate approximation of the exact ϕ_h.

Definition 2.1
A numerical method defined by the flow map ψ_h is called symplectic if for all Hamiltonian systems (3) it satisfies the condition

$$\psi_h'(y_0) J \psi_h'(y_0)^T = J \ . \tag{4}$$

One of the well-known examples of symplectic numerical methods is the s-stage RK Gauss methods which possess order $2s$. In this paper we shall deal with so-called (modified)

implicit RK-methods, introduced for the first time to obtain explicit EFRK methods [9] and re-used by Van de Vyver [12] for the construction of two-stage symplectic RK methods.

Definition 2.2
A s-stage modified RK method for solving the initial value problems (1) is a one step method defined by

$$y_1 = \psi_h(y_0) = y_0 + h\sum_{i=1}^{s} b_i f(t_0 + c_i h, Y_i) \ , \tag{5}$$

$$Y_i = \gamma_i y_0 + h\sum_{j=1}^{s} a_{ij} f(t_0 + c_j h, Y_j), \quad i = 1, \ldots, s \ , \tag{6}$$

where the real parameters c_i and b_i are respectively the nodes and the weights of the method. The parameters γ_i make the method modified with repect to the classical RK method, where $\gamma_i = 1, i = 1, \ldots, s$. The s-stage modified RK-method (5)-(6) can also be represented by means of its Butcher's tableau

$$\begin{array}{c|c|ccc} c_1 & \gamma_1 & a_{11} & \ldots & a_{1s} \\ c_2 & \gamma_2 & a_{21} & \ldots & a_{2s} \\ \vdots & \ldots & \vdots & \ddots & \vdots \\ c_s & \gamma_s & a_{s1} & \ldots & a_{ss} \\ \hline & & b_1 & \ldots & b_s \end{array} \tag{7}$$

or equivalently by the quartet (c, γ, A, b).
The conditions for a modified RK method to be symplectic have been obtained by Van de Vyver [12] and they are given in the following theorem.

Definition 2.3
A modified RK-method (5)-(6) for solving the Hamiltonian system (3) is symplectic if the following conditions are satisfied

$$m_{ij} \equiv b_i b_j - \frac{b_i}{\gamma_i} a_{ij} - \frac{b_j}{\gamma_j} a_{ji} = 0, \quad 1 \leq i, j \leq s \ . \tag{8}$$

In [2] it is shown that a modified RK-method not only preserves the linear invariants but also quadratic invariants if its coefficients satisfy conditions (8).

3 A review of previously constructed two-stage methods

In all applications we shall write down the results in terms of exponential or hyperbolic functions in order to make it easy for the reader to compare the formulae with previously published material.

3.1 The method of Van de Vyver [12]

Van de Vyver considers the modified RK method (7) with $s = 2$ and associates with the internal stages the following linear operators:

$$\mathcal{L}_i[h, \mathbf{a}]y(t) = y(t + c_i h) - \gamma_i y(t) - h \sum_{j=1}^{2} a_{ij} y'(t + c_j h), \quad i = 1, 2 \,, \tag{9}$$

and with the final stage the linear operator

$$\mathcal{L}[h, \mathbf{b}]y(t) = y(t + h) - y(t) - h \sum_{i=1}^{2} b_i y'(t + c_i h) \tag{10}$$

Requiring that the operators vanish for the functions $\exp(\pm\lambda t)$ with fixed nodes $c_i, i = 1, 2$ gives respectively rise to the following equations for the internal ($i = 1, 2$) and final stages

$$\begin{aligned} \cosh(c_i z) - \gamma_i - z(a_{i1} \sinh(c_1 z) + a_{i2} \sinh(c_2 z)) = 0 \\ \sinh(c_i z) - z(a_{i1} \cosh(c_1 z) + a_{i2} \cosh(c_2 z)) = 0 \end{aligned} \tag{11}$$

with $z = \lambda h$ and

$$\begin{aligned} \cosh(z) - 1 - z(b_1 \sinh(c_1 z) + b_2 \sinh(c_2 z)) = 0 \\ \sinh(z) - z(b_1 \cosh(c_1 z) + b_2 \cosh(c_2 z)) = 0 \end{aligned} \tag{12}$$

The equations (11) and (12) together with the symplecticity conditions

$$\begin{aligned} &b_1 \frac{a_{11}}{\gamma_1} + b_1 \frac{a_{11}}{\gamma_1} - b_1 b_1 = 0, \quad b_1 \frac{a_{12}}{\gamma_1} + b_2 \frac{a_{21}}{\gamma_2} - b_2 b_1 = 0, \\ &b_2 \frac{a_{22}}{\gamma_2} + b_2 \frac{a_{22}}{\gamma_2} - b_2 b_2 = 0, \end{aligned}$$

form a consistent non-linear system for the unknowns a_{ij}, b_i and γ_i. In order to obtain a fourth-order method the Gauss nodes are chosen, i.e. $c_{1,2} = \frac{1}{2} \pm \frac{\sqrt{3}}{6}$. The following solution was obtained:

$$\begin{aligned} a_{11} &= \frac{(\exp(z) - 1)(1 + E^2)}{z(\exp(z) + 1)(1 + E)^2}, & a_{12} &= \frac{2(\exp(z) - E^2)}{z(\exp(z) + 1)(1 + E)^2}, \\ a_{21} &= \frac{2(-1 + \exp(z)E^2)}{z(\exp(z) + 1)(1 + E)^2}, & a_{22} &= a_{11}, \\ \gamma_1 &= \frac{2 \exp(z/2)(1 + E + E^2 + E^3)}{\sqrt{E}(1 + E)^2(\exp(z) + 1)}, & \gamma_2 &= \gamma_1, \\ b_1 &= \frac{\exp(z) - 1}{z \exp(c_1 z)(1 + E)}, \quad b_2 = b_1, \\ &\text{with } E = \exp(z\sqrt{3}/3). \end{aligned} \tag{13}$$

The series expansions for these coefficients for small values of z are given by:

$$\begin{aligned}
b_1 &= \frac{1}{2}+\frac{1}{8640}z^4-\frac{1}{272160}z^6+\frac{13}{104509440}z^8-\frac{163}{38799129600}z^{10}+\dots\\
\gamma_1 &= 1-\frac{1}{288}z^4+\frac{1}{2160}z^6-\frac{881}{17418240}z^8+\frac{617}{117573120}z^{10}+\dots\\
a_{11} &= \frac{1}{4}-\frac{7}{8640}z^4+\frac{31}{272160}z^6-\frac{167}{13063680}z^8+\frac{1861}{1385683200}z^{10}+\dots\\
a_{12} &= -\frac{\sqrt{3}}{6}+\frac{1}{4}+\frac{\sqrt{3}}{216}z^2-(\frac{\sqrt{3}}{6480}+\frac{7}{8640})z^4+(\frac{17\sqrt{3}}{3265920}+\frac{31}{272160})z^6-\\
&\quad (\frac{31\sqrt{3}}{176359680}+\frac{167}{13063680})z^8+(\frac{691\sqrt{3}}{116397388800}+\frac{1861}{1385683200})z^{10}+\dots\\
a_{21} &= \frac{\sqrt{3}}{6}+\frac{1}{4}-\frac{\sqrt{3}}{216}z^2+(\frac{\sqrt{3}}{6480}-\frac{7}{8640})z^4+(-\frac{17\sqrt{3}}{3265920}+\frac{31}{272160})z^6\\
&\quad +(\frac{31\sqrt{3}}{176359680}-\frac{167}{13063680})z^8+(-\frac{691\sqrt{3}}{116397388800}+\frac{1861}{1385683200})z^{10}+\dots
\end{aligned}$$

Let us remark that these series are slowly converging and up to terms z^{22} have to be taken into account to reach an acceptable accuracy. It is also clear that in the limit $z \to 0$ the well-known classical fourth-order Gauss method is reproduced (see also (21)).

3.2 The method of Calvo et al. *[2]*

The method of Calvo *et al.* starts by considering two-stage methods with variable symmetric nodes $c_{1,2} = \frac{1}{2} \pm \theta(h,\lambda)$ such that all linear functionals(9) and (10) are exact for the set $\{1, \exp(\pm\lambda t)\}$. The requirement $\mathcal{L}_i[h,\mathbf{a}]1 = 0, i = 1,2$ implies that $\gamma_i = 1, i = 1,2$, meaning that classical RK are considered. The conditions $\mathcal{L}[h,\mathbf{b}]\exp(\pm\lambda t) = 0$ and $\mathcal{L}_i[h,\mathbf{a}]\exp(\pm\lambda t) = 0, i = 1,2$ results in a unique solution for the b_i's and a_{ij}'s, i.e.

$$\begin{aligned}
&b_1 = b_2 = \frac{\sinh(z/2)}{z\cosh(z\theta)}\\
&a_{11} = -\frac{\cosh(2z\theta)-\cosh(z(\theta+1/2))}{z\sinh(2z\theta)}, \quad a_{12} = -\frac{-1+\cosh(z(\theta-1/2))}{z\sinh(2z\theta)}\\
&a_{21} = \frac{-1+\cos(z(\theta+1/2))}{z\sinh(2z\theta)}, \quad a_{22} = \frac{\cosh(2z\theta)-\cosh(z(\theta-1/2))}{z\sinh(2z\theta)}
\end{aligned} \tag{14}$$

The symplecticness conditions (8) become here

$$\begin{aligned}
m_{11} &= b_1(2a_{11}-b_1) = 0\\
m_{22} &= b_1(2a_{22}-b_1) = 0\\
m_{12} &= m_{21} = b_1(b_1-a_{12}-a_{21}) = 0
\end{aligned} \tag{15}$$

The last condition of (15) is automatically satisfied in view of (14). The conditions m_{11} and m_{22} hold iff

$$\theta = \frac{1}{z}\operatorname{arccosh}\left(\frac{\cosh(z/2)+\sqrt{8+\cosh^2(z/2)}}{4}\right). \tag{16}$$

Further (14) and (16) imply that $\mathcal{L}[h,\mathbf{b}]\exp(\pm 2\lambda t) = 0$ automatically and therefore the final state is exact for the basis $\{1, \exp(\pm\lambda t), \exp(\pm 2\lambda t)\}$ or when $\lambda = i\omega$ for the trigonometric polynomial basis $\{1, \sin(\omega t), \cos(\omega t), \sin(2\omega t), \cos(2\omega t)\}$.
Also here it is worthwhile to give the series expansions:

$$b_1 = \tfrac{1}{2} - \tfrac{1}{2160}z^4 + \tfrac{1}{108864}z^6 + \tfrac{1}{2799360}z^8 - \tfrac{23}{1939956480}z^{10} + \dots$$

$$a_{11} = \tfrac{1}{4} - \tfrac{7}{8640}z^4 + \tfrac{31}{272160}z^6 - \tfrac{167}{13063680}z^8 + \tfrac{1861}{1385683200}z^{10} + \dots$$

$$a_{12} = (-\tfrac{\sqrt{3}}{6} + \tfrac{1}{4}) + \tfrac{\sqrt{3}}{432}z^2 + (-\tfrac{1}{4320} + \tfrac{13\sqrt{3}}{311040})z^4 + (-\tfrac{37\sqrt{3}}{17418240} + \tfrac{1}{217728})z^6 +$$
$$(-\tfrac{1121\sqrt{3}}{45148078080} + \tfrac{1}{5598720})z^8 + (\tfrac{355363\sqrt{3}}{178786389196800} - \tfrac{23}{3879912960})z^{10} + \dots$$

$$a_{21} = (\tfrac{\sqrt{3}}{6} + \tfrac{1}{4}) - \tfrac{\sqrt{3}}{432}z^2 - (\tfrac{1}{4320} + \tfrac{13\sqrt{3}}{311040})z^4 + (\tfrac{37\sqrt{3}}{17418240} + \tfrac{1}{217728})z^6 +$$
$$(\tfrac{1121\sqrt{3}}{45148078080} + \tfrac{1}{5598720})z^8 - (\tfrac{355363\sqrt{3}}{178786389196800} + \tfrac{23}{3879912960})z^{10} + \dots$$

$$a_{22} = \tfrac{1}{4} - \tfrac{1}{4320}z^4 + \tfrac{1}{217728}z^6 + \tfrac{1}{5598720}z^8 - \tfrac{23}{3879912960}z^{10} + \dots$$

$$\theta = \tfrac{\sqrt{3}}{6} + \tfrac{\sqrt{3}}{432}z^2 - \tfrac{\sqrt{3}}{311040}z^4 - \tfrac{17\sqrt{3}}{17418240}z^6 - \tfrac{61\sqrt{3}}{15049359360}z^8 + \tfrac{15073\sqrt{3}}{16253308108800}z^{10} + \dots$$

Let us remark here that these series are also slowly converging and up to terms z^{22} have to be taken into account to reach an acceptable accuracy.

4 New two-stage methods

It has been remarked by Hairer *et al.* [14] that symmetric numerical methods show a better long time behaviour than nonsymmetric ones when applied to reversible differential equations, as it is the case of conservative mechanical systems. In [3] it is observed that for modified RK methods whose coefficients are even functions of h the symmetry conditions are given by

$$c(h) + Sc(h) = e, \quad b(h) = Sb(h), \quad \gamma(h) = S\gamma(h), \quad SA(h) + A(h)S = \gamma(h)b^T(h)\ , \tag{17}$$

where

$$e = (1, ..., 1)^T \in \mathbb{R}^s \quad \text{and } S = (s_{ij}) \in \mathbb{R}^{s\times s} \quad \text{with } s_{ij} = \begin{cases} 1, & \text{if } i+j = s+1, \\ 0, & \text{if } i+j \neq s+1. \end{cases}$$

Since for symmetric EFRK methods the coefficients contain only even powers of h, the symmetry conditions can be written in a more convenient form by putting [3]

$$c(h) = \frac{1}{2}e + \theta(h), \quad A(h) = \frac{1}{2}\gamma(h)b^T(h) + \Lambda(h)\ , \tag{18}$$

where

$$d(h) = (\theta_1, \dots, \theta_s)^T \in \mathbb{R}^s \quad \text{and } \Lambda = (\lambda_{ij}) \in \mathbb{R}^{s\times s}\ .$$

Therefore, for a symmetric EFRK method whose coefficients a_{ij} are defined by

$$a_{ij} = \frac{1}{2}\gamma_i b_j + \lambda_{ij}, \quad 1 \leq i, j \leq s$$

the symplecticness conditions (8) reduce to

$$\mu_{ij} \equiv \frac{b_i}{\gamma_i}\lambda_{ij} + \frac{b_j}{\gamma_j}\lambda_{ji} = 0, \quad 1 \leq i, j, \leq s \ . \tag{19}$$

The idea of constructing symplectic EFRK taking into account the six-step procedure [13] is new. We briefly shall survey this procedure and suggest some adaptation in order to make the comparison with previous work more easy.

In step (i) we define the appropriate form of an operator related to the discussed problem. Each of the s internal stages (6) and the final stage (5) can be regarded as being a generalized linear multistep method on a nonequidistant grid; we can associated with each of them a linear operator (see (9) and (10)). We further construct the so-called moments which are for Gauss methods the expressions for $L_{i,j}(h, \mathbf{a}) = \mathcal{L}_i[h, \mathbf{a}]t^j, j = 0, \ldots, s-1$ and $L_i(h, \mathbf{b}) = \mathcal{L}[h, \mathbf{b}]t^j, j = 0, \ldots, 2s-1$ at $t = 0$, respectively, with $s = 2$.

In step (ii) the linear systems

$$L_{ij}(h, \mathbf{a}) = 0, \quad i = 1, \ldots, s, \quad j = 0, 1, \ldots, s-1 \ ,$$

$$L_i(h, \mathbf{b}) = 0, \quad i = 0, 1, \ldots, 2s-1 \ .$$

are solved to reproduce the classical Gauss RK collocation methods, showing the maximum number of functions which can be annihilated by each of the operators.

The steps (iii) and (iv) can be combined in the present context. First of all we have to define all reference sets of s and $2s$ functions which are appropriate for the internal and final stages respectively. These sets are in general hybrid sets of the following form

$$1, t, t^2, \ldots, t^K \text{ or } t^{K'}$$
$$\exp(\pm\lambda t), t\exp(\pm\lambda t), \ldots, t^P \exp(\pm\lambda t) \text{ or } t^{P'}\exp(\pm\lambda t)$$

where for the internal stages $K + 2P = s - 3$ and for the final stage $K' + 2P' = 2s - 3$. The set in which there is no classical component is identified by $K = -1$ and $K' = -1$, while the set in which there is no exponential fitting component is identified by $P = -1$ or $P' = -1$. It is important to note that such reference sets should contain all successive functions inbetween. Lacunary sets are in principle not allowed.

Once the sets chosen the operators (9)-(10) are applied to the members of the sets, in this particular case by taking into account the symmetry and the symplecticness conditions described above. The obtained independent expressions are put to zero and in step (v) the available linear systems are solved. The numerical values for $\lambda_{ij}(h)$, $b_i(h)$, $\gamma_i(h)$ and $\theta_i(h)$

are expressed for real values of λ (the pure exponential case) or for pure imaginary $\lambda = i\,\omega$ (oscillatory case). In order to make the comparison with previous work transparable we have opted to denote the results for real λ-values.

After the coefficients in the Butcher tableau have been filled in, the principal term of the local truncation error can be written down (step (vi)). Essentially, we know [11] that the algebraic order of the EFRK methods remains the same as the one of the classical Gauss method when this six-step procedure is followed, in other words the algebraic order is $\mathcal{O}(h^{2s})$, while the stage order is $\mathcal{O}(h^s)$. Explicit expressions for this local truncation error will not be discussed here.

Here we shall analyze in particular the construction of symmetric and symplectic EFRK Gauss methods with $s = 2$ stages whose coefficients are even functions of h. These EFRK methods have stage order 2 and algebraic order 4. From the symmetry conditions (17), taking into account (18) it follows that the nodes $c_j = c_j(h^2)$ and weights $b_j = b_j(h^2)$ satisfy

$$c_1 = \frac{1}{2} - \theta, \quad c_2 = \frac{1}{2} + \theta, \quad b_1 = b_2 \ ,$$

θ being a real parameter, and the coefficients $a_{ij} = a_{ij}(h^2)$ and $\gamma_i(h^2)$ satisfy:

$$a_{11} + a_{22} = \gamma_1 b_1, \quad a_{21} + a_{12} = \gamma_2 b_1 \ .$$

The symplecticness conditions (8) or (19) are equivalent to

$$a_{11} = \gamma_1 b_1/2, \quad \frac{a_{12}}{\gamma_1} + \frac{a_{21}}{\gamma_2} = b_1, \quad a_{22} = \gamma_2 b_2/2 \ ,$$

which results in

$$\gamma_1 = \gamma_2, \quad \lambda_{21} = -\lambda_{12} \ .$$

Taking into account the above relations the Butcher tableau can be expressed in terms of the unknowns $\theta, \gamma_1, \lambda_{12}$ and b_1 :

$$\begin{array}{c|c|cc} \frac{1}{2} - \theta & \gamma_1 & \dfrac{\gamma_1 b_1}{2} & \dfrac{\gamma_1 b_1}{2} + \lambda_{12} \\ \frac{1}{2} + \theta & \gamma_1 & \dfrac{\gamma_1 b_1}{2} - \lambda_{12} & \dfrac{\gamma_1 b_1}{2} \\ \hline & & b_1 & b_1 \end{array} \tag{20}$$

For the internal stages, the relation $K + 2P = -1$ results in the respective (K, P)-values:

- $(K = 1, P = -1)$ (the classical polynomial case with hybrid set $\{1, t\}$), and
- $(K = -1, P = 0)$ (the full exponential case with hybrid set $\{\exp(\lambda t), \exp(-\lambda t)\}$).

For the outer stage, we have $K' + 2P' = 1$, resulting in the respective (K', P')-values:

- $(K' = 3, P' = -1)$ (the classical polynomial case with hybrid set $\{1, t, t^2, t^3\}$),

- $(K' = 1, P' = 0)$ (mixed case with hybrid set $\{1, t, \exp(\pm\lambda t)\}$) and
- $(K' = -1, P' = 1)$(the full exponential case with hybrid set $\{\exp(\pm\lambda t), t\exp(\pm\lambda t)\}$).

The hybrid sets $(K = 1, P = -1)$ and $(K' = 3, P' = -1)$ are related to the polynomial case, giving rise to the well-known RK order conditions and to the fourth order Gauss method [17]

$$\begin{array}{c|c|cc} \frac{1}{2}-\frac{\sqrt{3}}{6} & 1 & \frac{1}{4} & \frac{1}{4}-\frac{\sqrt{3}}{6} \\ \frac{1}{2}+\frac{\sqrt{3}}{6} & 1 & \frac{1}{4}+\frac{\sqrt{3}}{6} & \frac{1}{4} \\ \hline & & \frac{1}{2} & \frac{1}{2} \end{array} \tag{21}$$

Let us remark that considering the $(K = 1, P = -1)$ set for the internal stages gives rise to $\gamma_1 = 1$, a value which is not compatible with the additional symmetry, symplecticity and order conditions imposed. Therefore in what follows we combine the $(K = -1, P = 0)$ case with either $(K' = 1, P' = 0)$ or $(K' = -1, P' = 1)$.

Case $(K' = 1, P' = 0)$

The operators (9) and (10) are applied to the functions present in the occurring hybrid sets, taking into account the structure of the Butcher tableau (20). Following equations arise with $z = \lambda h$:

$$2b_1 = 1 \tag{22}$$

$$2b_1\cosh(z/2)\cosh(\theta z) = \frac{\sinh(z)}{z} \tag{23}$$

$$\lambda_{12}\cosh(\theta z) = -\frac{\sinh(\theta z)}{z} \tag{24}$$

$$\lambda_{12}\sinh(\theta z) - \frac{\cosh(\theta z)}{z} = -\frac{\gamma_1}{z}\cosh(z/2) \tag{25}$$

resulting in the results

$$b_1 = 1/2, \qquad \theta = \frac{1}{z}\text{arccosh}\left(\frac{2\sinh(z/2)}{z}\right), \qquad \lambda_{12} = -\frac{\sinh(\theta z)}{z\cosh(\theta z)}$$
$$\gamma_1 = \frac{z}{\cosh(z/2)}\left(\frac{\sinh(\theta z)^2}{z\cosh(\theta z)} + \frac{\cosh(\theta z)}{z}\right) .$$

The series expansions for these coefficients for small values of z are given by

$$\theta = \sqrt{3}\left(\tfrac{1}{6} + \tfrac{1}{2160}z^2 - \tfrac{1}{403200}z^4 + \tfrac{1}{145152000}z^6 + \tfrac{533}{9656672256000}z^8 - \tfrac{2599}{2789705318400000}z^{10} + \ldots\right) ,$$
$$\lambda_{12} = \sqrt{3}\left(-\tfrac{1}{6} + \tfrac{1}{240}z^2 - \tfrac{137}{1209600}z^4 + \tfrac{143}{48384000}z^6 - \tfrac{81029}{1072963584000}z^8 + \tfrac{16036667}{8369115955200000}z^{10} + \ldots\right) ,$$
$$\gamma_1 = 1 - \tfrac{1}{360}z^4 + \tfrac{11}{30240}z^6 - \tfrac{71}{1814400}z^8 + \tfrac{241}{59875200}z^{10} + \cdots ,$$

showing that for $z \to 0$ the classical values are retrieved.

Case $(K' = -1, P' = 1)$

In this approach equations (23)-(25) remain unchanged and they deliver expressions for b_1, γ_1 and λ_{12} in terms of θ. Only (22) is replaced by

$$b_1(\cosh(\theta z)\,(2\cosh(z/2) + z\sinh(z/2)) + 2\theta z\cosh(z/2)\sinh(\theta z)) = \cosh(z) \qquad (26)$$

By combining (23) and (26) one obtains an equation in θ and z, i.e.:

$$\theta\sinh(z)\sinh(\theta z) = \cosh(\theta z)\left(\cosh(z) - \frac{\sinh(z)}{z} - \sinh^2(z/2)\right)$$

It is not anymore possible to write down an analytical solution for θ, but iteratively a series expansion can be derived. We give here those series expansions as obtained for the four unknowns

$\theta = \sqrt{3}\left(\frac{1}{6} + \frac{1}{1080}z^2 + \frac{13}{2721600}z^4 - \frac{1}{7776000}z^6 - \frac{1481}{1810626048000}z^8 + \frac{573509}{63552974284800000}z^{10} + \ldots\right)$,

$b_1 = \frac{1}{2} - \frac{1}{8640}z^4 + \frac{1}{1088640}z^6 + \frac{1}{44789760}z^8 - \frac{149}{775982592000}z^{10} + \ldots$

$\lambda_{12} = \sqrt{3}\left(-\frac{1}{6} + \frac{1}{270}z^2 - \frac{223}{2721600}z^4 + \frac{17}{9072000}z^6 - \frac{259513}{5431878144000}z^8 + \frac{9791387}{7944121785600000}z^{10} + \ldots\right)$,

$\gamma_1 = 1 - \frac{1}{480}z^4 + \frac{17}{60480}z^6 - \frac{2629}{87091200}z^8 + \frac{133603}{43110144000}z^{10} + \ldots$.

5 Numerical experiments

In this section we report on some numerical experiments where we test the effectiveness of the new and the previous [2, 12] (modified) Runge-Kutta methods when they are applied to the numerical solution of several differential systems. All the considered codes have the same qualitative properties for the Hamiltonian systems. In the figures we show the decimal logarithm of the maximum global error versus the number of steps required by each code in logarithmic scale. All computations were carried out in double precision and series expansions are used for the coefficients when $|z| < 0.1$.

Problem 1: Kepler's plane problem defined by the Hamiltonian function

$$H(p,q) = \frac{1}{2}(p_1^2 + p_2^2) - (q_1^2 + q_2^2)^{-1/2}\ ,$$

with the initial conditions $q_1(0) = 1 - e, q_2(0) = 0, p_1(0) = 0, p_2(0) = ((1+e)/(1-e))^{\frac{1}{2}}$, where $e, (0 \le e < 1)$ represents the eccentricity of the elliptic orbit. The exact solution of this IVP is a 2π-periodic elliptic orbit in the (q_1, q_2)-plane with semimajor axis 1, corresponding the starting point to the pericenter of this orbit. In the numerical

experiments presented here we have chosen the same values as in [4], i.e. $e = 0.001, \lambda = i\omega$ with $\omega = (q_1^2 + q_2^2)^{-\frac{3}{2}}$ and the integration is carried out on the interval $[0, 1000]$ with the steps $h = 1/2^m, m = 1, \ldots, 4$. The numerical behaviour of the global error in the solution is presented in figure 1. The results obtained by the four discussed methods (Calvo *et al.* (Calvo), Van de Vyver (Vyver), the new methods with $P = 0$ and $P = 1$) and the classical Gauss method (class.) are represented. The results for the four EFRK methods are approximately falling together. They are however more accurate than the results of the classical Gauss method of the same order.

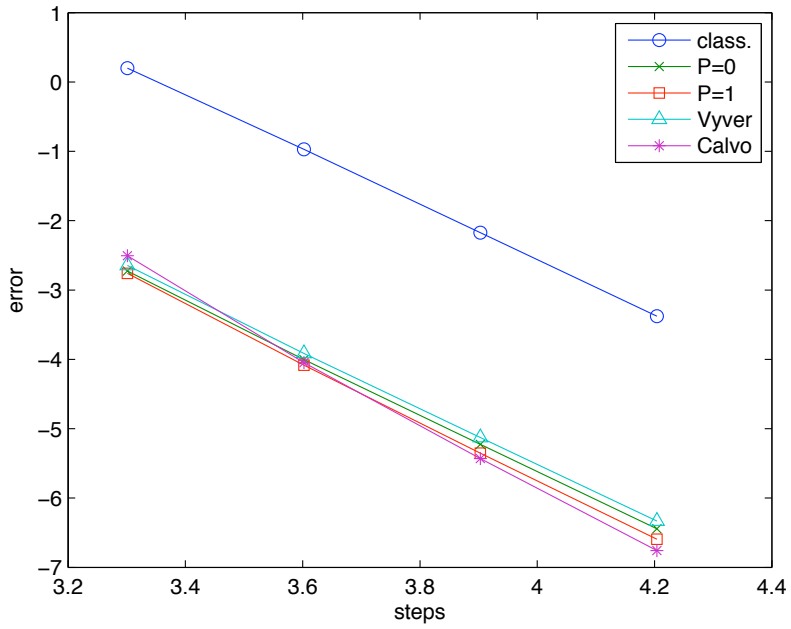


Figure 1.— Maximum global error in the solution of Problem 1.

Problem 2 A perturbed Kepler's problem defined by the Hamiltonian function

$$H(p, q) = \frac{1}{2}(p_1^2 + p_2^2) - \frac{1}{(q_1^2 + q_2^2)^{1/2}} - \frac{2\epsilon + \epsilon^2}{3(q_1^2 + q_2^2)^{3/2}} ,$$

with the initial conditions $q_1(0) = 1, q_2(0) = 0, p_1(0) = 0, p_2(0) = 1 + \epsilon$, where ϵ is a small positive parameter. The exact solution of this IVP is given by

$$q_1(t) = \cos(t + \epsilon t), \quad q_2(t) = \sin(t + \epsilon t), \quad p_i(t) = q_i'(t), i = 1, 2 .$$

As in [4] the numerical results are computed with the integration steps $h = 1/2^m, m = 1, \ldots, 4$. We take the parameter $\epsilon = 10^{-3}, \lambda = i\omega$ with $\omega = 1$ and the problem is integrated up to $t_{end} = 1000.$. The global error in the solution is presented in figure 2. The methods of Van de Vyver with the constant nodes gives the most accurate values. Our two new

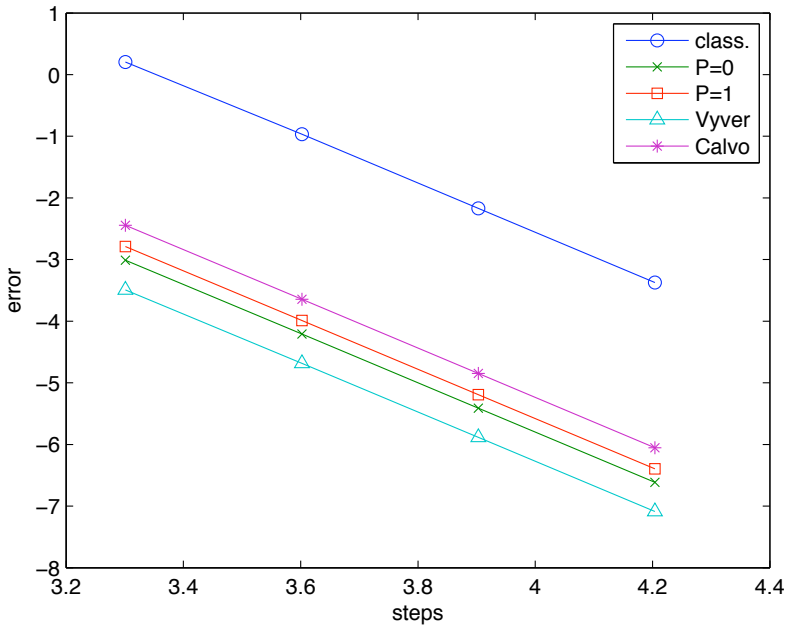


Figure 2.— Maximum global error in the solution of Problem 2.

symmetric methods are more accurate that the one of Calvo *et al.* All EFRK methods are more accurate than the classical Gauss method.

Problem 3 Euler's equations that describe the motion of a rigid body under no forces

$$\dot{q} = f(q) = ((\alpha - \beta)q_2 q_3, (1-\alpha)q_3 q_1, (\beta - 1)q_1 q_2)^T \ ,$$

with the initial values $q(0) = (0, 1, 1)^T$, and the parameter values $\alpha = 1 + \dfrac{1}{\sqrt{1.51}}$ and $\beta = 1 - \dfrac{0.51}{\sqrt{1.51}}$. The exact solution of this IVP is given by

$$q(t) = \left(\sqrt{1.51}\,\mathrm{sn}(t, 0.51), \mathrm{cn}(t, 0.51), \mathrm{dn}(t, 0.51)\right)^T \ ,$$

it is periodic with period $T = 7.45056320933095$, and $\mathrm{sn}, \mathrm{cn}, \mathrm{dn}$ stand for the elliptic Jacobi functions. Figure 3 shows the numerical results obtained for the global error computed with the interation steps $h = 1/2^m, m = 1, \ldots, 4$, on the interval $[0, 1000]$, and respective λ-values $\lambda = i2\pi/T$ (left) and $\lambda = i/2$ (right). In this problem the choice of the frequency is not so obvious and therefore the differentiation between the classical and the EF methods is not so pronounced. For $\lambda = i2\pi/T$ only the results of Calvo *et al.* are more accurate that the classical Gauss results. For $\lambda = i/2$ all EFRK results are falling together and are slightly more accurate than the classical results.

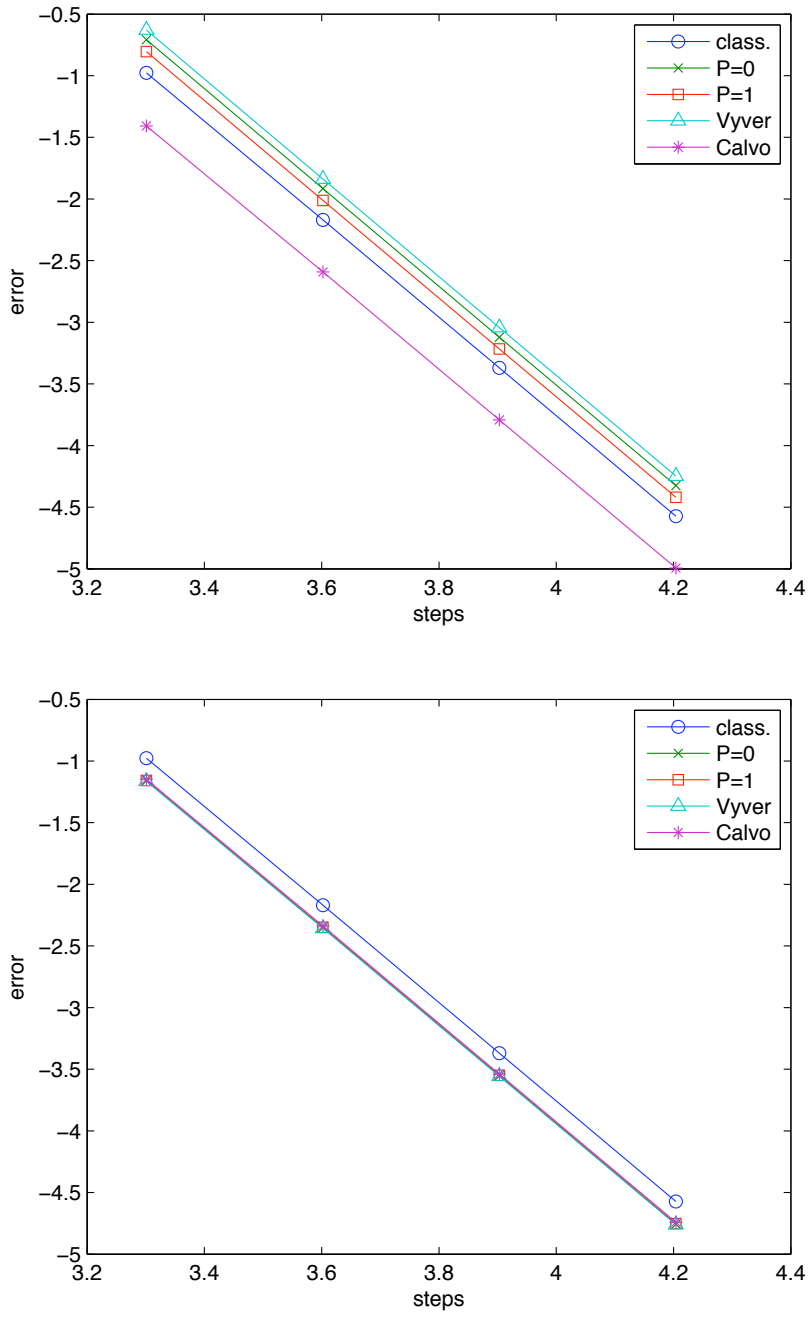


Figure 3.— Maximum global error in the solution of Problem 3. In the upper figure the results obtained with $\lambda = i2\pi/T$ are displayed. In the bottom figure the results obtained with $\lambda = i/2$ are shown.

6 Conclusions

In this paper another approach for constructing symmetric symplectic modified EFRK methods based upon the sixth-step procedure of [13] is presented. Two-stage fourth-order integrators of Gauss type which are symmetric and symplectic and which preserve linear and quadratic invariants have been derived. When the frequency used in the exponential fitting process is put to zero all considered integrators reduce to the classical Gauss integrator of the same order. Some numerical experiments show the utility of these new integrators for some oscillatory problems. The results obtained here are quite similar to the ones obtained in [2] and [12], but they differ in some of the details. The introduced method can be extended to EFRK with larger algebraic order.

References

[1] D. G. Bettis, Runge-Kutta algorithms for oscillatory problems, J. Appl. Math. Phys. (ZAMP) 30 (1979) 699–704.

[2] M. Calvo, J. M. Franco, J. I. Montijano, L. Rández, Structure preservation of exponentially fitted Runge-Kutta methods, Journ. Comp. Appl. Math. 218 (2008) 421-434.

[3] M. Calvo, J. M. Franco, J. I. Montijano, L. Rández, Sixth-order symmetric and symplectic exponentially fitted Runge-Kutta methods of the Gauss type, Comp. Phys. Commun. 178 (2008) 732–744.

[4] M. Calvo, J. M. Franco, J. I. Montijano, L. Rández, Sixth-order symmetric and symplectic exponentially fitted modified Runge-Kutta methods of the Gauss type, Journ. Comp. Appl. Math. 223 (2009) 387–398 .

[5] J. M. Franco, Runge-Kutta methods adapted to the numerical integration of oscillatory problems, Appl. Numer. Math. 50 (2004) 427-443.

[6] K. Ozawa, A functional fitting Runge-Kutta method with variable coefficients, Japan J. Indust. Appl. Math. 18 (2001) 107–130.

[7] T. E. Simos, An exponentially-fitted Runge-Kutta method for the numerical integration of initial-value problems with periodic or oscillating solutions, Comp. Phys. Commun. 115 (1998) 1–8.

[8] T. E. Simos, J. Vigo-Aguiar, Exponentially-fitted symplectic integrator, Phys. Rev. E67 (2003)1–7.

[9] G. Vanden Berghe, H. De Meyer, M. Van Daele, T. Van Hecke, Exponentially-fitted explicit Runge-Kutta methods, Comp. Phys. Commun. 123 (1999) 7–15.

[10] G. Vanden Berghe, H. De Meyer, M. Van Daele, T. Van Hecke, Exponentially-fitted Runge-Kutta methods, Journ. Comp. Appl. Math. 125 (2000)107–115.

[11] G. Vanden Berghe, M. Van Daele, H. Van de Vyver, Exponentially-fitted Runge-Kutta methods of collocation type: Fixed or variable knot points?, Journ. Comp. Appl. Math. 159 (2003) 217–239.

[12] H. Van de Vyver, A fourth order symplectic exponentially fitted integrator, Comp. Phys. Commun. 174 (2006) 255–262.

[13] L. Gr. Ixaru, G. Vanden Berghe, *Exponential Fitting*, Mathematics and its applications vol. 568, Kluwer Academic Publishers, 2004.

[14] E. Hairer, C. Lubich, G. Wanner, *Geometric Numerical Integration: Structure Preserving Algorithms for Ordinary Differential Equations*, Springer Verlag, Berlin 2002.

[15] J. M. Sanz-Serna, Symplectic integrators for Hamiltonian problems: An overview, Acta Numerica 1 (1992) 243–286.

[16] J. M. Sanz-Serna, M. P. Calvo, *Numerical Hamiltonian Problems*, Chapman and Hall, London 1994.

[17] E. Hairer, S. P. Nørsett, G. Wanner, *Solving Ordinary Differential Equations I, Nonstiff Problems*, Springer-Verlag Berlin, Heidelberg 1993.

Monografías de la Real Academia de Ciencias de Zaragoza **33**, 71–93, (2010).

Sobre la construcción de métodos Runge–Kutta–Nyström explícitos ajustados exponencialmente y de orden alto

J.M. Franco y I. Gómez

IUMA, Universidad de Zaragoza, CPS Ingenieros, Departamento de Matemática Aplicada, María de Luna 3, 50018 Zaragoza, Spain.

Abstract

Se analiza la construcción de métodos Runge–Kutta–Nyström (RKN) explícitos de orden elevado y ajustados exponencialmente (EF) para la resolución numérica de sistemas diferenciales con soluciones oscilatorias. Partiendo de dos métodos EFRKN básicos de referencia que son simétricos y simplécticos estudiamos dos procedimientos para construir métodos explícitos de orden alto. El primer procedimiento se basa en métodos de composición y permite construir métodos ajustados exponencialmente que son simétricos y simplécticos. El segundo procedimiento se basa en combinar distintos métodos para construir pares encajados de métodos paralelos ajustados exponencialmente que permiten su implementación en códigos a paso variable sin coste computacional añadido. Los experimentos numéricos realizados muestran el comportamiento cualitativo y la eficiencia numérica de los métodos construidos cuando se comparan con algunos métodos clásicos de la literatura científica.

Keywords: Métodos Runge–Kutta–Nyström; ajuste exponencial; métodos simétricos y simplécticos; métodos paralelos; sistemas diferenciales oscilatorios.

AMS classification: 65L05; 65L06; 65Y05

1 Introducción

En el presente trabajo abordamos la construcción de métodos Runge–Kutta–Nyström ajustados exponencialmente (EFRKN) de tipo explícito para la integración numérica de sistemas diferenciales oscilatorios. Los sistemas diferenciales oscilatorios son frecuentes en diferentes campos de las ciencias aplicadas tales como mecánica celeste, astrofísica, química, electrónica y dinámica de partículas entre otras (ver [1]). El diseño y la construcción de métodos numéricos para resolver sistemas diferenciales que poseen soluciones

periódicas u oscilantes ha sido considerado por diversos autores (ver [2–24] y las referencias citadas en estos artículos). El objetivo de estos métodos es utilizar la información disponible sobre las soluciones de los problemas correspondientes para construir algoritmos más precisos y eficientes que los algoritmos de propósito general para este tipo de problemas. Un trabajo pionero en esta materia se debe a Gautschi [14] en el cual se introdujeron métodos lineales multipaso ajustados exponencialmente para resolver sistemas diferenciales con soluciones oscilatorias. En cambio, el desarrollo de métodos RK o RKN ajustados exponencialmente (EFRK o EFRKN) se ha realizado más recientemente. Un survey detallado incluyendo una extensiva bibliografía sobre esta materia puede encontrarse en el libro de Ixaru and Vanden Berghe [16]. Un camino para construir métodos EFRK o EFRKN consiste en seleccionar los coeficientes del método de manera que integre exactamente un conjunto de funciones linealmente independientes que se eligen dependiendo de la naturaleza de las soluciones de los sistemas diferenciales que se pretenden resolver. Algunos resultados sobre la existencia y unicidad de solución para los coeficientes de un método EFRK han sido obtenidos por Ozawa [17, 18], y diversos autores [5, 10, 12, 13, 17, 19, 22] han construido métodos con coeficientes variables que integran exactamente sistemas diferenciales de primer o segundo orden cuyas soluciones pertenecen al espacio lineal generado por el conjunto de funciones $\{1, t, \ldots, t^k, \exp(\pm\lambda t), t\,\exp(\pm\lambda t), \ldots, t^p\,\exp(\pm\lambda t)\}$, donde $\lambda \in \mathbb{C}$ es una frecuencia determinada. En la práctica, estos métodos integran los sistemas diferenciales oscilatorios con mayor precisión que los métodos clásicos basados en funciones polinómicas.

Por otra parte, diversos autores (ver [2–4, 13, 25–27]) han comprobado que los integradores simplécticos obtienen superioridad numérica cuando se aplican a la resolución numérica de sistemas Hamiltonianos sobre intervalos de tiempo largos. Por lo tanto, para la clase de sistemas Hamiltonianos oscilatorios sería apropiado considerar métodos simplécticos ajustados exponencialmente que conserven la estructura del flujo original. Ejemplos de tales métodos se pueden encontrar en [3, 5, 6, 7] en los cuales se han construido métodos EFRK simplécticos de tipo implícito con dos y tres etapas y órdenes algebraicos cuatro y seis. Además, en [3] se ha extendido la teoría clásica de métodos RK simplécticos al caso de métodos EFRK modificados, obteniendo condiciones suficientes sobre los coeficientes de estos métodos que implican simplectitud para sistemas Hamiltonianos generales, y en [5] se ha analizado la preservación de propiedades por parte de los métodos EFRK modificados para el caso de sistemas diferenciales de primer orden.

La aparición de los ordenadores paralelos o multiprocesador ha dado lugar en las últimas décadas a un estudio intensivo para diseñar y construir nuevos integradores numéricos que utilicen las posibilidades de estas maquinas (ver por ejemplo [28–38] y las referencias citadas en ellos). En el caso de métodos RKN se han investigado diversas

clases de métodos explícitos de tipo predictor-corrector (PC) basados en métodos correctores RKN [30–34, 38]. Un objetivo común en los métodos PC iterados en paralelo consiste en reducir, para un orden de aproximación dado, el número f-evaluaciones secuenciales por paso utilizando procesadores paralelos. La principal desventaja de estos métodos es el alto coste de comunicación requerido entre los distintos procesadores lo que da lugar a una reducción significativa en la eficiencia cuando se implementan sobre algunas arquitecturas paralelas o cuando las f-evaluaciones no requieren mucho coste. Para evitar este inconveniente una alternativa consiste en considerar distintos métodos RKN explícitos de s_i-etapas ($i = 1, \ldots, k$) que se implementan sobre distintos procesadores y con las aproximaciones obtenidas se construye una combinación lineal de la forma

$$y_1 = \sum_{i=1}^{k} \omega_i \, y_1^{(i)}, \quad y_1' = \sum_{i=1}^{k} \omega_i \, y_1'^{(i)} \quad \text{donde} \quad \sum_{i=1}^{k} \omega_i = 1, \tag{1}$$

que proporciona las aproximaciones finales y_1 e y_1' en el instante $t_1 = t_0 + h$. Estos métodos no requieren de comunicación entre los distintos procesadores excepto en la última etapa en la cual los datos necesitan ser trasmitidos para calcular y_1 e y_1'. Métodos de la forma (1) han sido considerados por diversos autores [28, 35, 37] y aquí se utilizarán para construir métodos EFRKN explícitos de orden alto.

El objetivo de este trabajo de investigación es la construcción de métodos EFRKN explícitos de orden alto que integren exactamente sistemas diferenciales de segundo orden cuyas soluciones pertenecen al espacio lineal generado por el conjunto de funciones $\{\exp(\lambda t),\ \exp(-\lambda t)\}$, $\lambda \in \mathbb{C}$, ó $\{\sin(\omega t),\ \cos(\omega t)\}$ cuando $\lambda = i\omega$, $\omega \in \mathbb{R}$. La construcción de métodos EFRKN explícitos de orden bajo (hasta 4 ó 5) basada en las condiciones de orden para este tipo de métodos ya ha sido realizada (ver por ejemplo [12] y [13]). Pero para orden alto (≥ 6), las condiciones de orden de este tipo de métodos son complicadas de obtener y su resolución aún es más complicada de realizar debido a que ahora los coeficientes del método son funciones de un parámetro $z = \lambda h \in \mathbb{C}$, donde h es el paso de integración. Aquí, presentamos dos procedimientos alternativos que permiten la construcción de métodos EFRKN explícitos de orden alto. La organización del trabajo es la siguiente: La sección 2 se dedica a presentar los conceptos y resultados básicos así como la notación que utilizaremos a lo largo del documento. En la sección 3 presentamos un procedimiento basado en métodos de composición que permite construir métodos EFRKN explícitos de orden alto que son simétricos y simplécticos. En la sección 4 presentamos otro procedimiento basado en combinar distintos métodos EFRKN explícitos para construir pares encajados de métodos EFRKN paralelos de orden alto que permiten su implementación en códigos a paso variable. En la sección 5 presentamos algunos experimentos numéricos con un sistema diferencial oscilatorio que muestran el comportamiento cualitativo y la eficiencia numérica de los nuevos métodos EFRKN cuando se comparan con

algunos métodos clásicos de la literatura científica. Finalmente, la sección 6 la dedicamos a presentar algunas conclusiones.

2 Conceptos y resultados básicos

Consideramos problemas de valor inicial (IVPs) para sistemas diferenciales no stiff de segundo orden de la forma

$$y''(t) = f(t, y(t)), \quad y(t_0) = y_0, \quad y'(t_0) = y'_0, \tag{2}$$

suponiendo que $f : \mathbb{R} \times \mathbb{R}^m \to \mathbb{R}^m$ es suficientemente diferenciable, de manera que para todo $(t_0, y_0, y'_0) \in \mathbb{R} \times \mathbb{R}^m \times \mathbb{R}^m$, el IVP (2) tiene solución única $y(t) = y(t; t_0, y_0, y'_0)$ definida en algún entorno de t_0 con orden de derivación tan alto como sea necesario.

Un método RKN modificado de s etapas para la resolución numérica del IVP (2) es un método de un paso definido por las ecuaciones

$$\begin{aligned} Y_i &= \alpha_i\, y_0 + \gamma_i\, c_i\, h\, y'_0 + h^2 \sum_{j=1}^{s} a_{ij}\, f(t_0 + c_j h, Y_j), \qquad i = 1, \ldots, s, \\ y_1 &= \alpha_{s+1}\, y_0 + \gamma_{s+1}\, h\, y'_0 + h^2 \sum_{i=1}^{s} \bar{b}_i\, f(t_0 + c_i h, Y_i), \\ y'_1 &= y'_0 + h \sum_{i=1}^{s} b_i\, f(t_0 + c_i h, Y_i), \end{aligned} \tag{3}$$

donde h es el paso de integración, $Y_i \approx y(t_0 + c_i h)$ son las etapas internas del método e y_1 e y'_1 representan aproximaciones a $y(t_0 + h)$ e $y'(t_0 + h)$, respectivamente. Los parámetros reales c_i b_i y $\bar{b}_i$, $i = 1, \ldots, s$, se conocen como los nodos y los pesos del método y los parámetros α_i y γ_i se introducen (ver [12, 13]) para poder determinar métodos EFRKN explícitos. En el caso de métodos EFRKN los coeficientes α_i, γ_i, c_i b_i, $\bar{b}_i$ y a_{ij} dependen del paso de integración h. Además, cuando $\alpha_i = \gamma_i = 1$, $i = 1, \ldots, s+1$, el algoritmo (3) se reduce al de un método RKN. El método RKN modificado (3) se puede representar también mediante la tabla de coeficientes

$$\begin{array}{c|cc|c} \mathbf{c} & \alpha & \gamma & \mathbf{A} \\ \hline & \alpha_{s+1} & \gamma_{s+1} & \bar{\mathbf{b}}^T \\ \hline & & & \mathbf{b}^T \end{array} \quad = \quad \begin{array}{c|cc|ccc} c_1 & \alpha_1 & \gamma_1 & a_{11} & \cdots & a_{1s} \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ c_s & \alpha_s & \gamma_s & a_{s1} & \cdots & a_{ss} \\ \hline & \alpha_{s+1} & \gamma_{s+1} & \bar{b}_1 & \cdots & \bar{b}_s \\ \hline & & & b_1 & \cdots & b_s \end{array} \tag{4}$$

Un método RKN modificado tiene *orden* algebraico p si p es el mayor entero positivo de manera que el error de truncación local satisface

$$LE := \begin{pmatrix} y_1 \\ y_1' \end{pmatrix} - \begin{pmatrix} y(t_0+h) \\ y'(t_0+h) \end{pmatrix} = \mathcal{O}\left(h^{p+1}\right), \quad h \to 0, \tag{5}$$

y este error posee un desarrollo asintótico de la forma [41]

$$LE := \begin{pmatrix} y_1 \\ y_1' \end{pmatrix} - \begin{pmatrix} y(t_0+h) \\ y'(t_0+h) \end{pmatrix} = d_{p+1}(t_0)\, h^{p+1} + \cdots + d_{N+1}(t_0)\, h^{N+1} + \mathcal{O}\left(h^{N+2}\right), \tag{6}$$

siempre que la ecuación diferencial de (2) sea suficientemente diferenciable. Además, el desarrollo asintótico (6) implica que el error global tiene un desarrollo asintótico [41]

$$\Delta_n := \begin{pmatrix} y_n \\ y_n' \end{pmatrix} - \begin{pmatrix} y(x) \\ y'(x) \end{pmatrix} = e_p(x)\, h^p + \cdots + e_N(x)\, h^N + E_h(x)\, h^{N+1}, \tag{7}$$

donde $x = t_0 + nh$, $E_h(x)$ está acotado para $t_0 \le x \le t_{end}$ y $0 \le h \le h_0$, y los $e_j(x)$, $j \ge p$ son soluciones de los IVPs

$$e_j'(x) = L(x, y(x))\, e_j(x) + d_{j+1}(x), \quad e_j(t_0) = 0, \tag{8}$$

con

$$L(x, y(x)) = \begin{pmatrix} 0 & I \\ J_f(x) & 0 \end{pmatrix}, \quad J_f(x) := \frac{\partial f}{\partial y}(x, y(x)).$$

Para cada t_0 fijo, el flujo original del IVP (2) se denotará por ψ_h de manera que $\psi_h(y_0,\, y_0') = (y(t_0+h),\, y'(t_0+h))^T$, y el flujo numérico del método RKN modificado (3) se denotará por ϕ_h tal que $\phi_h(y_0,\, y_0') = (y_1,\, y_1')^T$. El flujo original del IVP (2) satisface la propiedad

$$(\psi_{-h} \circ \psi_h)\,(y_0,\, y_0') = (y_0,\, y_0')^T, \quad \text{para todo} \quad y_0,\, y_0' \in \mathbb{R}^m, \tag{9}$$

y cuando el flujo numérico ϕ_h también satisface la condición (9), al método de un paso se le llama simétrico. En general, para integraciones sobre intervalos de tiempo largos, los métodos numéricos simétricos muestran un comportamiento mejor que los no simétricos cuando se aplican a sistemas diferenciales reversibles, como es el caso de los sistemas mecánicos conservativos. Este hecho ha sido indicado por Hairer et al. [25] (ver Cap. V y XI), y estos autores han probado que para todo sistema diferencial cuya aplicación flujo es reversible, el flujo numérico de un método de un paso será también reversible si y solo si el método es simétrico. La clave para comprender el significado de esta propiedad es el concepto de método adjunto.

Definición 2.1 *El método adjunto ϕ_h^* de un método numérico ϕ_h es la aplicación inversa del método original con paso de integración opuesto $-h$, es decir, $\phi_h^* := \phi_{-h}^{-1}$. En otra palabras, $y_1 = \phi_h^*(y_0)$ esta definido implícitamente por $\phi_{-h}(y_1) = y_0$. Un método para el cual $\phi_h^* = \phi_h$ se llama simétrico.*

Una de las propiedades de los métodos simétricos es que su orden algebraico es par ($p = 2q$) y el desarrollo asintótico del error global (7) contiene solo potencias pares de h:

$$\Delta_n = e_{2q}(x)\, h^{2q} + e_{2q+2}(x)\, h^{2q+2} + \cdots + e_{2q+2r}(x)\, h^{2q+2r} + \cdots, \tag{10}$$

con $e_{2j}(t_0) = 0,\ \ j \geq q$.

En el caso de sistemas Hamiltonianos asociados a sistemas diferenciales de segundo orden (2) el campo vectorial $f(t, \mathbf{q})$ (cambiando $y \to \mathbf{q}$) esta definido mediante una función escalar (función potencial) $V = V(t, \mathbf{q}) : \mathbb{R} \times \mathbb{R}^m \to \mathbb{R}$, de manera que $f(t, \mathbf{q}) = -\mathbf{S}\ \nabla_{\mathbf{q}} V(t, \mathbf{q})$. Aquí $\mathbf{S}$ es una matriz m–dimensional regular y simétrica de coeficientes constantes y $\nabla_{\mathbf{q}} V(t, \mathbf{q})$ es el vector columna que contiene las derivadas de $V(t, \mathbf{q})$ con respecto a las componentes de $\mathbf{q} = (q_1, \ldots q_m)^T$. Así, el sistema Hamiltoniano queda definido mediante una función Hamiltoniana separable de la forma

$$H(t, \mathbf{p}, \mathbf{q}) = \frac{1}{2}\, \mathbf{p}^T\, \mathbf{S}\, \mathbf{p} + V(t, \mathbf{q}), \tag{11}$$

y puede escribirse de forma equivalente a (2) como

$$\mathbf{q}' = \mathbf{S}\, \mathbf{p}, \quad \mathbf{p}' = -\nabla_{\mathbf{q}} V(t, \mathbf{q}), \quad \mathbf{q}(t_0) = \mathbf{q}_0 = y_0, \quad \mathbf{p}(t_0) = \mathbf{p}_0 = \mathbf{S}^{-1}\, y_0'. \tag{12}$$

Para cada t_0 fijo la aplicación flujo $\psi_h(\mathbf{p}_0, \mathbf{q}_0)$ del sistema Hamiltoniano (12) es una transformación simpléctica para todo h en su dominio de definición (see [25, 26, 27]), es decir, la matriz Jacobiana de $\psi_h(\mathbf{p}_0, \mathbf{q}_0)$ satisface

$$\psi_h'(\mathbf{p}_0, \mathbf{q}_0)\, \mathbf{J}\, \psi_h'(\mathbf{p}_0, \mathbf{q}_0)^T = \mathbf{J}, \quad \forall\ t_0 \in \mathbb{R} \text{ and } (\mathbf{p}_0, \mathbf{q}_0) \in \mathbb{R}^{2m}, \tag{13}$$

y preserva la 2-forma diferencial $d\mathbf{p} \wedge d\mathbf{q} = dp_1 \wedge dq_1 + \cdots dp_m \wedge dq_m$:

$$d\mathbf{p} \wedge d\mathbf{q} = d\mathbf{p}_0 \wedge d\mathbf{q}_0, \quad \forall\ (\mathbf{p}_0, \mathbf{q}_0) \in \mathbb{R}^{2m}, \tag{14}$$

donde $\mathbf{J}$ es la matriz antisimétrica $2m$–dimensional

$$\mathbf{J} = \begin{pmatrix} 0_m & I_m \\ -I_m & 0_m \end{pmatrix}, \qquad \mathbf{J}^{-1} = -\mathbf{J}.$$

Una característica deseable en un método numérico ϕ_h para la resolución del sistema Hamiltoniano (12), además de proporcionar una aproximación precisa del flujo exacto ψ_h para un rango razonable de pasos de integración $h \in [0, h_0]$, es preservar propiedades cualitativas del flujo original ψ_h tales como la simplectitud dada por (13) ó (14).

Definición 2.2 *Un método numérico (3) definido por la aplicación flujo ϕ_h se llama simpléctico si para todo sistema Hamiltoniano (12) satisface la condición*

$$\phi_h'(\mathbf{p}_0, \mathbf{q}_0)\, \mathbf{J}\, \phi_h'(\mathbf{p}_0, \mathbf{q}_0)^T = \mathbf{J}, \quad \forall\ t_0 \in \mathbb{R} \text{ and } (\mathbf{p}_0, \mathbf{q}_0) \in \mathbb{R}^{2m}, \tag{15}$$

o la condición

$$d\mathbf{p}_1 \wedge d\mathbf{q}_1 = d\mathbf{p}_0 \wedge d\mathbf{q}_0, \quad \forall\ (\mathbf{p}_0, \mathbf{q}_0) \in \mathbb{R}^{2m}. \tag{16}$$

Uno de los más conocidos ejemplos de métodos RKN explícitos y simplécticos son los métodos de Störmer-Verlet [25, 27] los cuales poseen orden algebraico 2. Las condiciones para que un método RKN modificado y explícito sea simpléctico han sido obtenidas en [13].

A continuación introducimos algunos conceptos típicos en la terminología de métodos paralelos. Para un método RKN explícito, se define una *unidad de tiempo* secuencial como el tiempo requerido para una evaluación secuencial de la forma $f(t_n + c_i h, Y)$. En [36], los autores introducen el concepto de método r-paralelo y q-procesador en el contexto de los métodos RK que se puede extender con facilidad al caso de métodos RKN explícitos.

Definición 2.3 *Un método RKN explícito de s etapas es r-paralelo y q-procesador, si r y q son los enteros más pequeños para los cuales las s etapas internas del método pueden evaluarse en r unidades de tiempo secuenciales utilizando q procesadores.*

Otra forma equivalente de expresar esta definición es que se puede realizar una partición de la matriz A del método RKN (después de una permutación de las etapas) en r bloques o superetapas

$$A = \begin{pmatrix} 0 & & & & \\ A_{21} & 0 & & & \\ A_{31} & A_{32} & 0 & & \\ \vdots & \vdots & \ddots & \ddots & \\ A_{r1} & A_{r2} & \cdots & A_{r,r-1} & 0 \end{pmatrix}, \tag{17}$$

donde A_{ij} es una matriz de tamaño $\mu_i \times \mu_j$ y las derivadas segundas $f(t_0 + c_{\sigma+1} h, Y_{\sigma+1})$, $\cdots$, $f(t_0 + c_{\sigma+\mu_i} h, Y_{\sigma+\mu_i})$ de cada bloque se pueden calcular en paralelo sobre μ_i procesadores de forma que un paso del método RKN se calcula en r unidades de tiempo secuenciales ($q = \max\{\mu_i,\ 2 \le i \le r\}$).

Una cota sobre el orden algebraico alcanzable por un método RKN explícito r-paralelo y q-procesador se ha obtenido en [42] y viene recogida en el siguiente resultado:

Teorema 2.4 *El orden algebraico de un método RKN explícito r-paralelo y q-procesador cuya matriz A se puede partir como en (17) es como mucho $2r$, independientemente del número de procesadores q y del número de etapas s.*

Los métodos RKN explícitos con orden algebraico $p = 2r$ se denominan métodos *P-optimales*.

La idea de construir métodos numéricos que integren exactamente un conjunto de funciones linealmente independientes distintas de los polinomios ha sido propuesto por diversos autores (ver por ejemplo [9, 17, 18, 19, 22, 24]). Esta idea consiste en determinar los coeficientes del método de manera que éste integre exactamente, en el intervalo $[t_0,\ t_0+h]$, al conjunto de funciones escalares linealmente independientes

$$\mathcal{F} = \langle u_1(t), u_2(t), \ldots, u_r(t)\rangle, \quad r \leq s.$$

En el caso de métodos RKN modificados (3) que son exactos para las funciones del espacio lineal generado por $\mathcal{F}$, sus coeficientes quedan determinados por la solución de los sistemas lineales siguientes

$$\mathbf{b}^T\, u_k''(t\,\mathbf{e} + h\,\mathbf{c}) = \frac{u_k'(t+h) - u_k'(t)}{h}, \quad k = 1, \ldots, r, \tag{18}$$

$$\bar{\mathbf{b}}^T\, u_k''(t\,\mathbf{e} + h\,\mathbf{c}) = \frac{u_k(t+h) - \alpha_{s+1}\, u_k(t) - h\,\gamma_{s+1}\, u_k'(t)}{h^2}, \quad k = 1, \ldots, r, \tag{19}$$

$$\mathbf{A}\, u_k''(t\,\mathbf{e} + h\,\mathbf{c}) = \frac{u_k(t\,\mathbf{e} + h\,\mathbf{c}) - \alpha\, u_k(t) - h\,(\gamma\cdot\mathbf{c})\, u_k'(t)}{h^2}, \quad k = 1, \ldots, r, \tag{20}$$

donde $\mathbf{e} = (1, \ldots, 1)^T \in \mathbb{R}^s$, $\gamma\cdot\mathbf{c} = (\gamma_1\, c_1, \ldots, \gamma_s\, c_s)^T \in \mathbb{R}^s$, y para un vector $\mathbf{v} = (v_1, \ldots, v_s)^T \in \mathbb{R}^s$ y una función escalar g denotamos por $g(\mathbf{v})$ al vector real s-dimensional $(g(v_1), \ldots, g(v_s))^T$.

En particular, cuando $r = s$ los coeficientes $\mathbf{b} = (b_i)$, $\bar{\mathbf{b}} = (\bar{b}_i)$ y $\mathbf{A} = (a_{ij})$ definidos por los sistemas lineales (18)–(20) quedan determinados de forma única para todo $h > 0$, si la matriz

$$\begin{pmatrix} u_1''(t+c_1h) & \cdots & u_1''(t+c_sh) \\ \vdots & \cdots & \vdots \\ u_s''(t+c_1h) & \cdots & u_s''(t+c_sh) \end{pmatrix}, \tag{21}$$

es no singular, y estos coeficientes son h-dependientes.

El caso más habitual consiste en considerar funciones exponenciales y trigonométricas como conjunto de referencia: $\mathcal{F}_1 = \langle \exp(\lambda t),\ \exp(-\lambda t)\rangle$ o $\mathcal{F}_2 = \langle \sin(\omega t),\ \cos(\omega t)\rangle$. El caso trigonométrico $\mathcal{F}_2$ se obtiene de $\mathcal{F}_1$ con $\lambda = i\omega$. Para el conjunto de funciones de referencia $\mathcal{F}_1$ los sistemas lineales (18)–(20) se reducen a

$$\mathbf{b}^T \cosh(\mathbf{c}\, z) = \frac{\sinh(z)}{z}, \qquad \mathbf{b}^T \sinh(\mathbf{c}\, z) = \frac{\cosh(z) - 1}{z}, \tag{22}$$

$$\bar{\mathbf{b}}^T \cosh(\mathbf{c}\, z) = \frac{\cosh(z) - \alpha_{s+1}}{z^2}, \qquad \bar{\mathbf{b}}^T \sinh(\mathbf{c}\, z) = \frac{\sinh(z) - z\,\gamma_{s+1}}{z^2}, \tag{23}$$

$$\mathbf{A} \cosh(\mathbf{c}\, z) = \frac{\cosh(\mathbf{c}\, z) - \alpha}{z^2}, \qquad \mathbf{A} \sinh(\mathbf{c}\, z) = \frac{\sinh(\mathbf{c}\, z) - z\,(\gamma\cdot\mathbf{c})}{z^2}, \tag{24}$$

donde $z = \lambda h$.

Hasta el momento se han construido métodos EFRKN explícitos y simplécticos con orden algebraico hasta cuatro (ver [13]). En las secciones siguientes presentamos algunos procedimientos que permiten construir métodos EFRKN simétricos y simplécticos (SSEFRKN) y explícitos de orden alto, así como pares encajados de métodos EFRKN explícitos también de orden alto.

3 Métodos SSEFRKN explícitos de orden alto

En esta sección presentamos un procedimiento basado en métodos de composición que permite construir métodos EFRKN explícitos simétricos y simplécticos (métodos SSEFRKN explícitos) de orden alto. Este procedimiento, como en el caso de métodos RKN clásicos, se basa en partir de un método básico exponencialmente ajustado y explícito ϕ_h que sea simétrico y simpléctico, y mediante una composición simétrica de éste método de referencia construir integradores de orden alto.

3.1 Métodos básicos de referencia SSEFRKN

Primero construimos métodos básicos de referencia SSEFRKN que son extensiones de los métodos clásicos de Störmer-Verlet [27, 25] representados por las tablas de coeficientes

$$\begin{array}{c|cc} 1/2 & 0 \\ \hline & 1/2 \\ \hline & 1 \end{array} \qquad \begin{array}{c|cc} 0 & 0 & \\ 1 & 1/2 & 0 \\ \hline & 1/2 & 0 \\ \hline & 1/2 & 1/2 \end{array} \tag{25}$$

La versión exponencialmente ajustada del método SV_1 (EFSV_1) viene definida por las ecuaciones

$$Y_1 = \alpha_1\, y_0 + \frac{h}{2}\,\gamma_1\, y_0' \approx y(t_0 + \frac{h}{2}), \tag{26}$$

$$y_1 = y_0 + h\,\gamma_2\, y_0' + h^2\,\bar{b}_1\, f(t_0 + \frac{h}{2}, Y_1) \approx y(t_0 + h), \tag{27}$$

$$y_1' = y_0' + h\, b_1\, f(t_0 + \frac{h}{2}, Y_1) \approx y'(t_0 + h), \tag{28}$$

o la tabla de coeficientes

$$\begin{array}{c|ccc} 1/2 & \alpha_1 & \gamma_1 & 0 \\ \hline & 1 & \gamma_2 & \bar{b}_1 \\ \hline & & & b_1 \end{array} \tag{29}$$

Calculando el adjunto ϕ_h^* e imponiendo que $\phi_h^* = \phi_h$ (Definition 2.1), las condiciones para que el método (26)–(28) sea simétrico vienen dadas por

$$\bar{b}_1 = \frac{b_1\,\gamma_2}{2}, \qquad \gamma_1 = \alpha_1\,\gamma_2. \tag{30}$$

Si el método (26)–(28) se aplica al sistema Hamiltoniano (12), entonces

$$d\mathbf{p}_1 \wedge d\mathbf{q}_1 = d\mathbf{p}_0 \wedge d\mathbf{q}_0 + h^2 \left(\gamma_2\, b_1 - \bar{b}_1 - \frac{b_1\,\gamma_1}{2\,\alpha_1}\right) df_1 \wedge d\mathbf{p}_0, \tag{31}$$

con $f_1 = f(t_0 + h/2, Y_1)$, y el método será simpléctico (Definition 2.2) si satisface la condición

$$\bar{b}_1 = b_1 \left(\gamma_2 - \frac{\gamma_1}{2\,\alpha_1}\right). \tag{32}$$

Observar que las condiciones de simetría (30) implican la condición de simplectitud (32).

Si ahora imponemos que el método (26)–(28) sea exacto para toda función del espacio lineal $\mathcal{F}_1$, las condiciones (22)–(24) implican que

$$\alpha_1 = \cosh(z/2), \quad \gamma_1 = \frac{\sinh(z/2)}{z/2}, \quad \gamma_2 = \frac{2\,\sinh(z/2)}{z\,\cosh(z/2)}, \quad \bar{b}_1 = \frac{2\,\sinh^2(z/2)}{z^2\,\cosh(z/2)}, \quad b_1 = \gamma_1, \tag{33}$$

y los coeficientes (33) satisfacen las condiciones de simetría (30) y simplectitud (32). En consecuencia, el nuevo método EFSV_1 es simétrico, simpléctico y tiene orden algebraico 2.

La versión exponencialmente ajustada del método SV_2 (EFSV_2) viene definida por las ecuaciones

$$y_1 = y_0 + h\,\gamma_2\,y_0' + h^2\,\bar{b}_1\,f(t_0, y_0) \approx y(t_0 + h), \tag{34}$$

$$y_1' = y_0' + h\,\left(b_1\,f(t_0, y_0) + b_2\,f(t_0 + h, y_1)\right) \approx y'(t_0 + h), \tag{35}$$

o la tabla de coeficientes

$$\mathrm{EFSV}_2: \quad \begin{array}{c|cc|cc} 0 & 1 & 1 & 0 & \\ 1 & 1 & \gamma_2 & \bar{b}_1 & 0 \\ \hline & 1 & \gamma_2 & \bar{b}_1 & 0 \\ \hline & & & b_1 & b_2 \end{array} \tag{36}$$

Imponiendo que $\phi_h^* = \phi_h$, las condiciones para que el método (34)–(35) sea simétrico vienen dadas por

$$\bar{b}_1 = b_1\,\gamma_2, \qquad b_2 = b_1. \tag{37}$$

Si el método (34)–(35) se aplica al sistema Hamiltoniano (12), entonces

$$d\mathbf{p}_1 \wedge d\mathbf{q}_1 = d\mathbf{p}_0 \wedge d\mathbf{q}_0 + h^2 \left(\gamma_2\, b_1 - \bar{b}_1\right) df_0 \wedge d\mathbf{p}_0, \tag{38}$$

con $f_0 = f(t_0, y_0)$, y el método será simpléctico si satisface la condición

$$\bar{b}_1 = b_1\, \gamma_2. \tag{39}$$

Observar que las condiciones de simetría (37) implican la condición de simplectitud (39).

Si ahora imponemos que el método (34)–(35) sea exacto para toda función del espacio lineal $\mathcal{F}_1$, las condiciones (22)–(24) implican que

$$\gamma_2 = \frac{\sinh(z)}{z}, \quad \bar{b}_1 = \frac{\cosh(z)-1}{z^2}, \quad b_2 = b_1 = \frac{\sinh(z)}{z\,(1+\cosh(z))}, \tag{40}$$

y los coeficientes (40) satisfacen las condiciones de simetría (37) y simplectitud (39). En consecuencia, el nuevo método EFSV_2 es simétrico, simpléctico y tiene orden algebraico 2.

Notamos que cuando $z = 0$ los métodos EFSV_1 y EFSV_2 se reducen a los métodos clásicos SV_1 y SV_2. Además, en el caso trigonométrico ($\lambda = i\omega$) resulta que $z = i\nu$ con $\nu = \omega h$, y los coeficientes de los métodos EFSV_1 y EFSV_2 se obtienen teniendo en cuenta las relaciones $\cosh(i\nu) = \cos(\nu)$ y $\sinh(i\nu) = i\sin(\nu)$.

3.2 Métodos de composición exponencialmente ajustados

Dado un método básico de referencia ϕ_h y s números reales $\delta_1, \ldots, \delta_s$, un *método de composición* Φ_h viene definido por

$$\Phi_h = \phi_{\delta_1 h} \circ \phi_{\delta_2 h} \circ \cdots \circ \phi_{\delta_s h}, \qquad \sum_{i=1}^{s} \delta_i = 1, \tag{41}$$

donde $\phi_{\delta_i h}$ representa al método de referencia con paso de integración $\delta_i\, h$. Además, un método RKN modificado ϕ_h está ajustado al espacio lineal $\mathcal{F}$, si para toda función $u(t) \in \mathcal{F}$

$$\phi_h(u(t_0), u'(t_0)) = \begin{pmatrix} u(t_0+h) \\ u'(t_0+h) \end{pmatrix}. \tag{42}$$

Teorema 3.1 *Sea Φ_h un método de composición definido por (41)*

i) *Si el método de referencia ϕ_h está ajustado al espacio lineal $\mathcal{F}$, entonces el método de composición Φ_h también está ajustado a $\mathcal{F}$.*

ii) *Si el método de referencia ϕ_h es simétrico y los coeficientes δ_i satisfacen $\delta_i = \delta_{s+1-i}$, para todo i (composición simétrica), entonces el método de composición Φ_h es simétrico.*

iii) *Si el método de referencia ϕ_h es simpléctico, entonces el método de composición Φ_h también es simpléctico.*

Dem: i) Si ϕ_h satisface (42) para toda función $u(t) \in \mathcal{F}$, entonces se tiene que

$$\left(\phi_{\delta_1 h} \circ \phi_{\delta_2 h}\right)\left(u(t_0), u'(t_0)\right) = \phi_{\delta_1 h}(u(t_0 + \delta_2 h), u'(t_0 + \delta_2 h)) = \begin{pmatrix} u(t_0 + (\delta_1 + \delta_2)h) \\ u'(t_0 + (\delta_1 + \delta_2)h) \end{pmatrix}. \tag{43}$$

En consecuencia, para toda función $u(t) \in \mathcal{F}$

$$\Phi_h(u(t_0), u'(t_0)) = \begin{pmatrix} u(t_0 + (\sum_{i=1}^{s} \delta_i)h) \\ u'(t_0 + (\sum_{i=1}^{s} \delta_i)h) \end{pmatrix} = \begin{pmatrix} u(t_0 + h) \\ u'(t_0 + h) \end{pmatrix}. \tag{44}$$

ii) Φ_h es simétrico por ser una composición simétrica de métodos simétricos [25].

iii) Φ_h es simpléctico por ser una composición de métodos simplécticos (la composición de transformaciones simplécticas es una transformación simpléctica [25]). □

El Teorema 3.1 asegura que si el método EFRKN de referencia es simétrico y simpléctico, y los coeficientes del método de composición Φ_h son simétricos ($\delta_i = \delta_{s+1-i}$, para todo i), entonces el método Φ_h es un método EFRKN simétrico y simpléctico. En consecuencia, utilizando los métodos EFSV_1 y EFSV_2 como métodos de referencia se pueden construir métodos de composición EFRKN simétricos y simplécticos de orden alto.

Las condiciones de orden para métodos de composición con coeficientes simétricos pueden encontrarse en [25], así como los coeficientes δ_i. Acontinuación presentamos los coeficientes para métodos de orden 6 y 8.

- Orden 6 con $s = 9$ etapas

$$\begin{aligned} \delta_1 &= \delta_9 = 0.392161444007314139 \\ \delta_2 &= \delta_8 = 0.332599136789359438 \\ \delta_3 &= \delta_7 = -0.706246172557639359 \\ \delta_4 &= \delta_6 = 0.082213596293550800 \\ \delta_5 &= 0.798543990934829963 \end{aligned} \tag{45}$$

- Orden 8 con $s = 17$ etapas

$$\begin{aligned}
\delta_1 &= \delta_{17} = 0.130202483088890081 \\
\delta_2 &= \delta_{16} = 0.561162981775108384 \\
\delta_3 &= \delta_{15} = -0.389474962644847286 \\
\delta_4 &= \delta_{14} = 0.158841906555155601 \\
\delta_5 &= \delta_{13} = -0.395903894133237577 \\
\delta_6 &= \delta_{12} = 0.184539640978315707 \\
\delta_7 &= \delta_{11} = 0.258374387686322047 \\
\delta_8 &= \delta_{10} = 0.295011723609310299 \\
\delta_9 &= -0.605508533830034512
\end{aligned} \tag{46}$$

4 Métodos paralelos EFRKN explícitos de orden alto

Ahora investigamos la construcción de una familia de métodos paralelos EFRKN explícitos que consisten de k métodos EFRKN explícitos de s_i etapas $(i = 1, \ldots, k)$ que se implementan sobre distintos procesadores, y las aproximaciones obtenidas se combinan para obtener la solución numérica y_1 e y_1' en $t_1 = t_0 + h$. Utilizando la notación de Iserles and Norsset [35] se pueden representar en la forma

$$\begin{array}{c|cc|c}
\mathbf{c}_1 & \alpha^{(1)} & \gamma^{(1)} & \mathbf{A}_1 \\
\hline
 & \alpha_{s+1}^{(1)} & \gamma_{s+1}^{(1)} & \bar{\mathbf{b}}_1^T \\
\hline
 & & & \mathbf{b}_1^T
\end{array} \quad + \cdots + \quad
\begin{array}{c|cc|c}
\mathbf{c}_k & \alpha^{(k)} & \gamma^{(k)} & \mathbf{A}_k \\
\hline
 & \alpha_{s+1}^{(k)} & \gamma_{s+1}^{(k)} & \bar{\mathbf{b}}_k^T \\
\hline
 & & & \mathbf{b}_k^T
\end{array} \tag{47}$$

donde $\mathbf{c}_i$, $\mathbf{b}_i$, $\bar{\mathbf{b}}_i$, $\alpha^{(i)}$, $\gamma^{(i)}$, $\alpha_{s+1}^{(i)}$, $\gamma_{s+1}^{(i)}$ y $\mathbf{A}_i$ son los coeficientes de cada método EFRKN de s_i etapas que proporcionan las aproximaciones $y_1^{(i)}$ e $y_1'^{(i)}$ en $t_1 = t_0 + h$, y las aproximaciones finales se obtienen tomando una combinación lineal de ellos

$$y_1 = \sum_{i=1}^{k} \omega_i \, y_1^{(i)}, \quad y_1' = \sum_{i=1}^{k} \omega_i \, y_1'^{(i)}, \qquad \sum_{i=1}^{k} \omega_i = 1. \tag{48}$$

Los métodos (47)–(48) también se pueden representar mediante la tabla de Butcher

$$
\begin{array}{c|cc|cccc}
\mathbf{c}_1 & \alpha^{(1)} & \gamma^{(1)} & \mathbf{A}_1 & 0 & \cdots & 0 \\
\mathbf{c}_2 & \alpha^{(2)} & \gamma^{(2)} & 0 & \mathbf{A}_2 & \cdots & 0 \\
\vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\
\mathbf{c}_k & \alpha^{(k)} & \gamma^{(k)} & 0 & 0 & \ldots & \mathbf{A}_k \\
\hline
 & \sum_{i=1}^k \omega_i\, \alpha_{s+1}^{(i)} & \sum_{i=1}^k \omega_i\, \gamma_{s+1}^{(i)} & \omega_1\, \bar{\mathbf{b}}_1^T & \omega_2\, \bar{\mathbf{b}}_2^T & \ldots & \omega_k\, \bar{\mathbf{b}}_k^T \\
\hline
 & & & \omega_1\, \mathbf{b}_1^T & \omega_2\, \mathbf{b}_2^T & \cdots & \omega_k\, \mathbf{b}_k^T
\end{array}
\tag{49}
$$

4.1 Una familia de pares encajados de métodos EFRKN paralelos

Consideramos un método básico de referencia simétrico y de orden 2 (EFSV$_1$ o EFSV$_2$) que denotaremos por su aplicación flujo ϕ_h, y construimos un nuevo método $\Phi_h^{(s)}$ conectando s pasos de $\phi_{\Delta t}$ con tamaño $\Delta t = h/s$, es decir,

$$\Phi_h^{(s)} = \phi_{h/s} \circ \phi_{h/s} \circ \cdots \circ \phi_{h/s}, \quad (s-\text{veces}). \tag{50}$$

El nuevo método $\Phi_h^{(s)}$ también es simétrico (por ser una composición simétrica de métodos simétricos), tiene orden 2 y un desarrollo asintótico del error global (10) de la forma

$$\Delta_n^{(s)} = \frac{1}{s^2}\, e_2(x)\, h^2 + \frac{1}{s^4}\, e_4(x)\, h^4 + \cdots + \frac{1}{s^{2m}}\, e_{2m}(x)\, h^{2m} + \cdots \tag{51}$$

Dada una secuencia arbitraria $s_1, s_2, \ldots, s_k$ de enteros positivos satisfaciendo

$$0 < s_1 < s_2 < \cdots < s_k, \tag{52}$$

se puede construir el esquema k-procesador

$$\Phi_h = \sum_{i=1}^k \omega_i\, \Phi_h^{(s_i)}, \qquad \sum_{i=1}^k \omega_i = 1, \tag{53}$$

donde cada uno de los métodos $\Phi_h^{(s_i)}$ puede evaluarse en un procesador diferente ($\Phi_h^{(s_i)}$ se determina conectando s_i pasos de ϕ_h con tamaño h/s_i como in (50)), y cuyo error global admite un desarrollo asintótico de la forma

$$\Delta_n(\Phi_h) := \sum_{i=1}^k \omega_i\, \Delta_n^{(s_i)} = \sum_{i=1}^k \frac{\omega_i}{s_i^2}\, e_2(x)\, h^2 + \cdots + \sum_{i=1}^k \frac{\omega_i}{s_i^{2m}}\, e_{2m}(x)\, h^{2m} + \ldots, \tag{54}$$

En consecuencia, si los pesos ω_i se eligen satisfaciendo las condiciones

$$\sum_{i=1}^k \omega_i = 1, \qquad \sum_{i=1}^k \frac{\omega_i}{s_i^{2j}} = 0, \quad j = 1, \ldots, m-1, \tag{55}$$

entonces la expresión del error global (54) se reduce a

$$\Delta_n(\Phi_h) = \sum_{i=1}^{k} \frac{\omega_i}{s_i^{2m}}\, e_{2m}(x)\, h^{2m} + \mathcal{O}\left(h^{2m+2}\right). \tag{56}$$

Teniendo en cuenta que el error local coincide con el error global cometido en el primer paso ($LE(\Phi_h) := \Delta_1(\Phi_h)$) y el desarrollo asintótico $e_{2j}(t_0+h) = e_{2j}(t_0) + h\, e'_{2j}(t_0) + \mathcal{O}\left(h^2\right)$ con $e_{2j}(t_0) = 0$, se tiene que

$$LE(\Phi_h) = \sum_{i=1}^{k} \frac{\omega_i}{s_i^{2m}}\, e'_{2m}(t_0)\, h^{2m+1} + \mathcal{O}\left(h^{2m+2}\right), \tag{57}$$

y el método Φ_h definido en (53) con los pesos ω_i satisfaciendo las condiciones (55) tiene orden algebraico $p = 2m$.

A continuación veremos que si el esquema básico de referencia es de tipo EFRKN, entonces el esquema k-procesador (53) es un método EFRKN.

Teorema 4.1 *Si el método de referencia ϕ_h está ajustado al espacio lineal $\mathcal{F}$, entonces el método Φ_h definido en (53) también está ajustado a $\mathcal{F}$.*

Dem: Por el Teorema 3.1 el método de composición $\Phi_h^{(s_i)}$ está ajustado al espacio lineal $\mathcal{F}$. Por lo tanto, para toda función $u(t) \in \mathcal{F}$

$$\Phi_h(u(t_0), u'(t_0)) = \sum_{i=1}^{k} \omega_i\, \Phi_h^{(s_i)}(u(t_0), u'(t_0)) = \sum_{i=1}^{k} \omega_i \begin{pmatrix} u(t_0+h) \\ u'(t_0+h) \end{pmatrix} = \begin{pmatrix} u(t_0+h) \\ u'(t_0+h) \end{pmatrix}. \tag{58}$$

□

Entonces, para los métodos básicos de referencia EFSV_1 y EFSV_2 podemos escribir el siguiente resultado:

Teorema 4.2 *Si el método de referencia ϕ_h es un método EFRKN simétrico de orden 2 (*EFSV_1 ó EFSV_2*), entonces el método EFRKN definido en (53) tiene orden maximal $2k$ cuando los pesos estan dados por*

$$w_i = \begin{cases} 1, & \text{para } k = 1, \\ \dfrac{s_i^{2k-2}}{\displaystyle\prod_{\substack{j\neq i \\ j=1}}^{k} (s_i^2 - s_j^2)}, & \text{para } k \geq 2,\ i = 1, \ldots, k, \end{cases} \tag{59}$$

y el término principal del error local es

$$LE(\Phi_h) = \frac{(-1)^{k+1}}{s_1^2\, s_2^2 \cdots s_k^2}\, e'_{2k}(t_0)\, h^{2k+1} + \mathcal{O}\left(h^{2k+2}\right). \tag{60}$$

Dem: El orden maximal $2k$ se sigue de (55) con $m = k$ y los pesos (59) se obtienen resolviendo el sistema lineal de tipo Vandermonde resultante. Sustituyendo los pesos (59) en la expresión del error local (57) se obtiene (60). □

Observaciones.

1. Para $k \geq 4$, el teorema 4.2 proporciona un procedimiento para construir una familia de métodos EFRKN explícitos de orden elevado ($p \geq 8$) que requieren s_k f-evaluaciones secuenciales por paso.

2. El procedimiento (53) también permite la construcción de otro esquema paralelo

$$\Phi_h^* = \sum_{i=1}^{k-1} \omega_i^* \, \Phi_h^{(s_i)}, \tag{61}$$

con los pesos definidos por

$$w_i^* = \frac{s_i^{2k-4}}{\prod\limits_{\substack{j \neq i \\ j=1}}^{k-1} (s_i^2 - s_j^2)}, \quad i = 1, \ldots, k-1, \tag{62}$$

y orden algebraico $2k-2$. Por lo tanto, existe una familia de pares encajados de orden $2k(2k-2)$ que permite la implementación de los métodos paralelos EFRKN en códigos a paso variable sin coste adicional.

Entonces, considerando los métodos básicos de referencia EFSV_1 y EFSV_2 y seleccionando una secuencia (52) podemos calcular en paralelo sobre k procesadores

$$\Phi_h^{(s_i)}(y_0, y_0') = (y_1^{(s_i)}, y_1'^{(s_i)})^T, \qquad i = 1, \ldots, k,$$

(cada método $\Phi_h^{(s_i)}$ se evalua en un procesador) y después evaluar

$$(y_1, y_1')^T = \sum_{i=1}^{k} \omega_i \, \Phi_h^{(s_i)}(y_0, y_0'),$$
$$\text{Est(h)} = \sum_{i=1}^{k-1} (\omega_i - \omega_i^*) \, \Phi_h^{(s_i)}(y_0, y_0') + \omega_k \, \Phi_h^{(s_k)}(y_0, y_0'),$$

donde Est(h) representa una estimación del error local basada en el par encajado $2k(2k-2)$. En la Tabla I mostramos los pesos ω_i y ω_i^* para el caso de la secuencia $s_i = i$, $i = 1, \ldots, k$, con $k = 4, 5, 6$, y los métodos se denotarán por EFRKN$p\,(s_1, \ldots, s_k)$ donde p indica el orden.

TABLA I: Pesos ω_i y ω_i^* para los pares encajados $p\,(p-2)$ con $p = 8, 10, 12$

EFRKN8(1,2,3,4)		EFRKN10(1,2,3,4,5)		EFRKN12(1,2,3,4,5,6)	
w_i	w_i^*	w_i	w_i^*	w_i	w_i^*
$-\frac{1}{360}$	$\frac{1}{24}$	$\frac{1}{8640}$	$-\frac{1}{360}$	$-\frac{1}{302400}$	$\frac{1}{8640}$
$\frac{16}{45}$	$-\frac{16}{15}$	$-\frac{64}{945}$	$\frac{16}{45}$	$\frac{8}{945}$	$-\frac{64}{945}$
$-\frac{729}{280}$	$\frac{81}{40}$	$\frac{6561}{4480}$	$-\frac{729}{280}$	$-\frac{2187}{4480}$	$\frac{6561}{4480}$
$\frac{1024}{315}$		$-\frac{16384}{2835}$	$\frac{1024}{315}$	$\frac{65536}{14175}$	$-\frac{16384}{2835}$
		$\frac{390625}{72576}$		$-\frac{9765625}{798336}$	$\frac{390625}{72576}$
				$\frac{17496}{1925}$	

5 Experimentos numéricos

En esta sección presentamos algunos experimentos numéricos con un sistema diferencial oscilatorio para mostrar el comportamiento cualitativo y la eficiencia numérica de los nuevos métodos EFRKN construidos en las secciones 3 y 4 cuando se comparan con algunos métodos clásicos de la literatura científica. El criterio utilizado en las comparaciones numéricas se basa en calcular el máximo del error global en la solución (MGE = $\log_{10}(\max \|y(t_n) - y_n\|)$) o en el Hamiltoniano (MGEH = $\log_{10}(\max |H(0) - H_n|)$), y los cálculos se realizan en aritmética de doble precisión (16 dígitos).

El sistema diferencial oscilatorio considerado es el problema gravitacional de los dos cuerpos (problema plano de Kepler) definido mediante el Hamiltoniano

$$H(p,q) = \frac{1}{2}\left(p_1^2 + p_2^2\right) - \left(q_1^2 + q_2^2\right)^{-1/2},$$

con las condiciones iniciales $q_1(0) = 1-e,\ q_2(0) = 0,\ p_1(0) = 0,\ p_2(0) = ((1+e)/(1-e))^{1/2}$, donde e $(0 \le e < 1)$ representa la excentricidad de la órbita elíptica. La solución exacta de este PVI es una órbita elíptica 2π-periódica con semieje mayor 1 dada por

$$q_1(t) = \cos(u(t)) - e, \qquad q_2(t) = \sqrt{1-e^2}\,\sin(u(t)),$$

donde $u(t)$ es la solución de la ecuación de Kepler: $t = u(t) - e\,\sin(u(t))$. La integración se realiza en el intervalo $[0, 20]$ con los valores de los parámetros $e = 0.001$, $\lambda = i\omega$ con $\omega = (q_1^2 + q_2^2)^{-3/2}$ como en [3].

5.1 Comparaciones a paso fijo

Los nuevos métodos de composición EF de la sección 3 se han comparado con los métodos de composición simétricos de coeficientes constantes y órdenes 6 y 8 (SIM6 y SIM8) dados en [25]. La Figura 1 muestra el máximo del error global en la solución (MGE) con respecto al número de f-evaluaciones secuenciales requeridas por cada método para los pasos $h = 0.8/2^j$, $j \geq 0$.

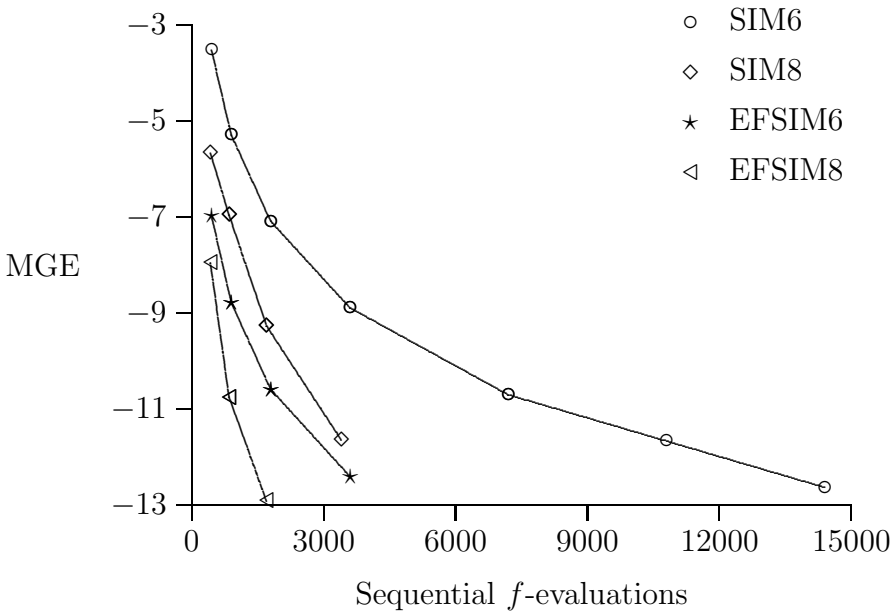


Figura 1: Curvas de eficiencia (a paso fijo).

La Figura 2 muestra el máximo del error global en el Hamiltoniano (MGEH) como una función del tiempo en los instantes finales $t_{end} = 10^j$, $j = 1, \ldots, 5$, con paso de integración $h = 0.8$.

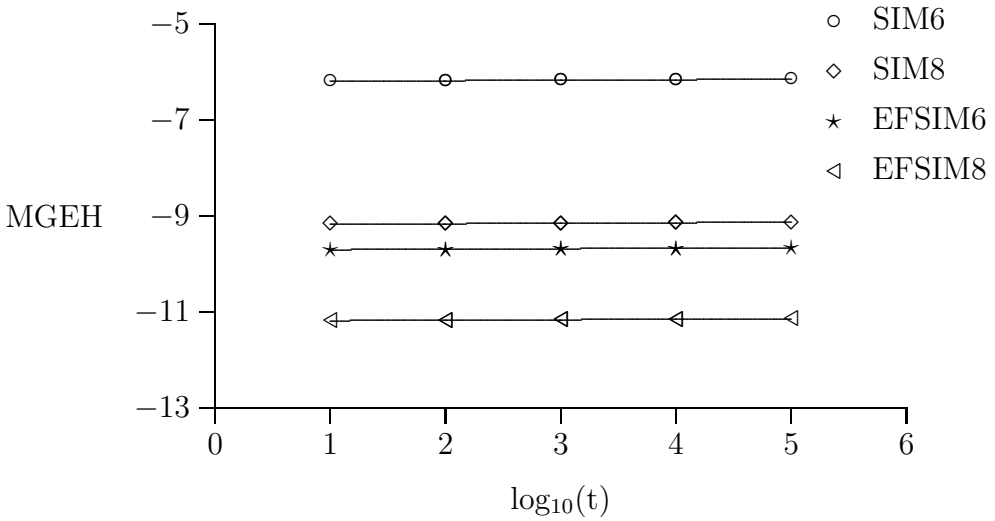


Figura 2: Error en el Hamiltoniano como una función del tiempo ($h = 0.8$).

De los resultados numéricos obtenidos (Figuras 1 y 2) se observa que los métodos de composición simétricos EF (EFSIM6 y EFSIM8) muestran un comportamiento más

eficiente que sus homólogos de coeficientes constantes (SIM6 y SIM8). Además, aunque los métodos simétricos y simplécticos no preservan el Hamiltoniano, la Figura 2 muestra que el error en la energía permanece prácticamente constante para el sistema Hamiltoniano oscilatorio considerado.

5.2 Comparaciones a paso variable

Los nuevos métodos paralelos EFRKN de la sección 4 se han comparado con el par encajado clásico RKN8(6) obtenido en [39]. En la Figura 3 se muestran las curvas de eficiencia de los métodos (máximo del error global en la solución (MGE) versus el coste computacional medido por el número de f-evaluaciones secuenciales requeridas por cada método) para tolerancias del error local de la forma $Tol = 10^{-j},\ j \geq 3$.

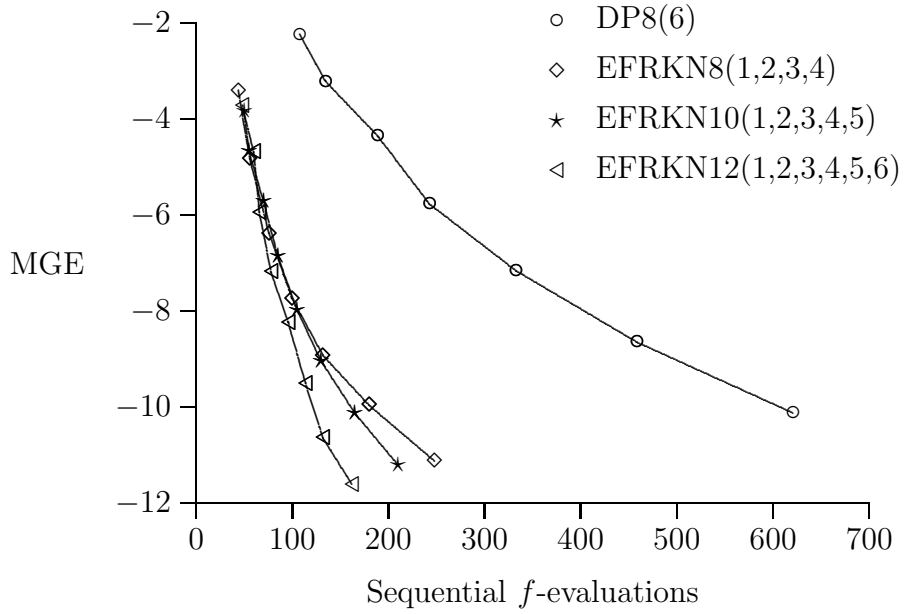


Figura 3: Curvas de eficiencia (a paso variable).

En el caso de códigos a paso variable, la Figura 3 muestra claramente que los esquemas paralelos de orden alto ajustados exponencialmente son más eficientes que el código clásico RKN8(6).

6 Conclusiones

En este trabajo se presentan dos procedimientos para construir métodos RKN explícitos ajustados exponencialmente de orden alto. El primer procedimiento se basa en la composición simétrica de métodos simétricos partiendo de un método básico de referencia que es ajustado exponencialmente, además de simétrico y simpléctico. El segundo procedimiento se basa en una combinación convexa de k métodos EFRKN explícitos de s_i etapas los cuales se construyen conectando s_i pasos de un método de referencia EFRKN

explícito y simétrico con longitud $\Delta t = h/s_i$. Considerando como métodos de referencia una versión ajustada exponencialmente de los métodos de Störmer-Verlet, se construye una familia de esquemas paralelos EFRKN explícitos en el formato de pares encajados con órdenes $2k(2k-2)$. Ambos procedimientos se muestran como alternativas de confianza frente a los códigos RKN clásicos para integrar sistemas diferenciales oscilatorios de segundo orden. Los experimentos numéricos realizados con un sistema Hamiltoniano oscilatorio muestran que los nuevos métodos EFRKN mejoran la eficiencia computacional obtenida con otros integradores RKN clásicos de la literatura científica.

Agradecimientos

El presente trabajo de investigación se ha financiado en parte mediante el Proyecto MTM2007-67530-C02-01 de la Dirección General de Investigación (Ministerio de Educación, Cultura y Deporte).

Referencias

[1] V.I. Arnold: *Mathematical Methods of Classical Mechanics*, Springer–Verlag, New York, 1989.

[2] T.E. Simos and J. Vigo–Aguiar: Exponentially fitted symplectic integrator, *Phys. Rev. E*, **67** (2003) 1–7.

[3] H. Van de Vyver: A fourth order symplectic exponentially fitted integrator, *Comput. Phys. Comm.*, **176** (2006) 255–262.

[4] J. Vigo-Aguiar, T.E. Simos and A. Tocino: An adapted symplectic integrator for Hamiltonian systems, *Int. J. Modern. Phys. C*, **12** (2001) 225–234.

[5] M. Calvo, J.M. Franco, J.I. Montijano and L. Rández: Structure preservation of Exponentially Fitted Runge–Kutta methods, *J. Comput. Appl. Math.*, **218** (2008) 421–434.

[6] M. Calvo, J.M. Franco, J.I. Montijano and L. Rández: Sixth-order symmetric and symplectic exponentially fitted Runge-Kutta methods of Gauss type, *J. Comput. Appl. Math.*, **223** (2009) 387–398.

[7] M. Calvo, J.M. Franco, J.I. Montijano and L. Rández: Sixth-order symmetric and symplectic exponentially fitted modified Runge-Kutta methods of Gauss type, *Comput. Phys. Comm.*, **178** (2008) 732–744.

[8] D.G. Bettis: Runge–Kutta Algorithms for Oscillatory Problems, *J. Appl. Math. Phys. (ZAMP)*, **30** (1979) 699–704.

[9] J.P. Coleman and S.C. Duxbury: Mixed collocation methods for $y'' = f(x, y)$, *J. Comput. Appl. Math.*, **126** (2000) 47–75.

[10] J.M. Franco: An embedded pair of exponentially fitted explicit Runge–Kutta methods, *J. Comput. Appl. Math.*, **149** (2002) 407–414.

[11] J.M. Franco: Runge–Kutta methods adapted to the numerical integration of oscillatory problems, *Appl. Numer. Math.*, **50** (2004) 427–443.

[12] J.M. Franco: Exponentially fitted explicit Runge–Kutta–Nyström methods, *J. Comput. Appl. Math.*, **167** (2004) 1–19.

[13] J.M. Franco: Exponentially fitted symplectic integrators of RKN type for solving oscillatory problems, *Comput. Phys. Comm.*, **177** (2007) 479–492.

[14] W. Gautschi: Numerical integration of ordinary differential equations based on trigonometric polynomials, *Numer. Math.* **3** (1961) 381–397. (2004) 427–443.

[15] N.S. Huang, R.B. Sidge and N.H. Cong: On functionally fitted Runge–Kutta methods, *BIT*, **46** (2006) 861–874.

[16] L. Gr. Ixaru and G. Vanden. Berghe: *Exponential Fitting*, Kluwer Academic Publishers, 2004.

[17] K. Ozawa: A functional fitting Runge-Kutta method with variable coefficients, *Japan J. Indust. Appl. Math.*, **18** (2001) 107–130.

[18] K. Ozawa: A functionally fitted three-stage explicit singly diagonally implicit Runge-Kutta method, *Japan J. Indust. Appl. Math.*, **22** (2005) 403–427.

[19] B. Paternoster: Runge–Kutta(–Nyström) methods for ODEs with periodic solutions based on trigonometric polynomials, *Appl. Numer. Math.*, **28** (1998) 401–412.

[20] T.E. Simos: An exponentially–fitted Runge–Kutta method for the numerical integration of initial–value problems with periodic or oscillating solutions, *Comput. Phys. Commun.*, **115** (1998) 1–8.

[21] G. Vanden Berghe, H. De Meyer, M. Van Daele and T. Van Hecke: Exponentially–fitted explicit Runge–Kutta methods, *Comput. Phys. Commun.*, **123** (1999) 7–15.

[22] G. Vanden Berghe, H. De Meyer, M. Van Daele and T. Van Hecke: Exponentially fitted Runge–Kutta methods, *J. Comput. Appl. Math.*, **125** (2000) 107–115.

[23] G. Vanden Berghe, L. Gr. Ixaru, H. De Meyer: Frequency determination and step-length control for exponentially fitted Runge–Kutta methods, *J. Comput. Appl. Math.*, **132** (2001), pp. 95-105.

[24] G. Vanden Berghe, M. Van Daele and H. Van de Vyver: Exponentially–Fitted Runge–Kutta methods of collocation type: fixed or variable knot points?, *J. Comput. Appl. Math.*, **159** (2003) 217–239.

[25] E. Hairer, C. Lubich and G. Wanner: *Geometric Numerical Integration: Structure Preserving algorithms for Ordinary Differential Equations*, Springer Verlag, Berlin, 2002.

[26] J.M. Sanz–Serna: Symplectic integrators for Hamiltonian problems: an overview, *Acta Numerica*, **1** (1992) 243–286.

[27] J.M. Sanz–Serna and M.P. Calvo: *Numerical Hamiltonian Problems*, Chapman and Hall, London, 1994.

[28] P. Bogacki: A family of parallel Runge-Kutta pairs, *Comput. Math. Appl.*, **31** (1996) 23–31.

[29] K. Burrage: *Parallel and Sequential Methods for Ordinary Differential Equations*, Oxford Sciences Publications, Oxford, 1995.

[30] N.H. Cong: Note on the performance of direct and indirect Runge-Kutta-Nyström methods, *J. Comput. Appl. Math.*, **45** (1993) 295–308.

[31] N.H. Cong: Explicit symmetric Runge-Kutta-Nyström methods for parallel computers, *Comput. Math. Appl.*, **31** (1996) 111–122.

[32] N.H. Cong: RKN-type parallel block PC methods with Lagrange-type predictors, *Comput. Math. Appl.*, **35** (1998) 45–57.

[33] N.H. Cong, K. Strehmel and R. Weiner: Runge-Kutta-Nyström-type parallel block predictor-corrector methods, *Adv. Comput. Math.*, **10** (1999) 115–133.

[34] N.H. Cong and N. Van Minh: Continuous parallel-iterated RKN-type PC methods for nonstiff IVPs, *Appl. Numer. Math.*, **57** (2007) 1097–1107.

[35] A. Iserles and S.P. Nørsett: On the theory of parallel Runge-Kutta methods, *IMA J. Numer. Anal.*, **10** (1990) 463-488.

[36] K.R. Jackson and S.P. Nørsett: The potential for parallelism in Runge-Kutta methods; Part I: RK formulas in standard form, *SIAM J. Numer. Anal.*, **32** (1995) 49–82.

[37] B. Paternoster: Order bound for a family of parallel Runge-Kutta-Nyström methods through computer algebra, *Comput. Math. Appl.*, **35** (1998) 107–119.

[38] B.P. Sommeijer: Explicit, high-order Runge-Kutta-Nyström methods for parallel computers, *Appl. Numer. Math.*, **13** (1993) 221–240.

[39] J.R. Dormand, M.E.A. El-Mikkawy and P.J. Prince: High-order embedded Runge-Kutta-Nyström Formulae. *IMA J. Numer. Anal.*, **7** (1987) 423-430

[40] M. El-Mikkawy and R. El-Desouky: A new optimized non FSAL embedded Runge-Kutta-Nyström algorithm of orders 6 and 4 in six stages, *Appl. Math. Comput.*, **145** (2003) 33–43.

[41] E. Hairer, S.P. Nørsett and G. Wanner: *Solving Ordinary Differential Equations I, Nonstiff Problems*, Springer–Verlag, Berlin, 1993.

[42] I. Gómez, J.M. Franco: Métodos Runge-Kutta-Nyström paralelos de tipo explícito para PVI de segundo orden, *Actas electrónicas del XXI-CEDYA/XI-CMA*, Ciudad Real (Spain), 21-25 de Septiembre de 2009.

Monografías de la Real Academia de Ciencias de Zaragoza **33**, 95–112, (2010).

Numerical comparisons between Gauss-Legendre methods and Hamiltonian BVMs defined over Gauss points*

Luigi Brugnano

Dipartimento di Matematica "U. Dini", Università di Firenze, Italy

and

Felice Iavernaro, Tiziana Susca

Dipartimento di Matematica, Università di Bari, Italy

Dedicated to Prof.Manuel Calvo, on the occasion of his 65th birthday.

Abstract

Hamiltonian Boundary Value Methods are a new class of energy preserving one step methods for the solution of polynomial Hamiltonian dynamical systems. They can be thought of as a generalization of collocation methods in that they may be defined by imposing a suitable set of *extended collocation conditions*. In particular, in the way they are described in this note, they are related to Gauss collocation methods with the difference that they are able to precisely conserve the Hamiltonian function in the case where this is a polynomial of any high degree in the momenta and in the generalized coordinates. A description of these new formulas is followed by a few test problems showing how, in many relevant situations, the precise conservation of the Hamiltonian is crucial to simulate on a computer the correct behavior of the theoretical solutions.

1 Introduction

Hamiltonian Boundary Value Methods (HBVMs) form a subclass of Boundary Value Methods (BVMs), whose main feature is that of precisely conserving the Hamiltonian function associated with a canonical Hamiltonian system

$$\begin{cases} \dot{y} = J\nabla H(y), \\ y(t_0) = y_0 \in \mathbb{R}^{2m}, \end{cases} \qquad J = \begin{pmatrix} 0 & I \\ -I & 0 \end{pmatrix} \in \mathbb{R}^{2m\times 2m}, \tag{1}$$

*Work developed within the project "Numerical methods and software for differential equations".

(I is the identity matrix of dimension m), in the case where such function is of polynomial type.

Two key ideas have permitted the realization of HBVMs: the definition of *discrete line integral* and what we called *extended collocation conditions*. The former, first introduced in [15, 16], represents the discrete counterpart of the line integral defined over conservative vector fields, while the second is a relaxation of the classical collocation conditions which assures the conservation of the energy along the numerical solution $\{y_n\}$ generated by the method itself.

Just as an initial clarification, we briefly show how this new approach to the problem reads when the classical Gauss collocation method is considered (see [18, Remark 2.1] for more details). Given a stepsize $h > 0$ and a set of s abscissae $c_1 < \cdots < c_s$ disposed according to a Gauss-Legendre distribution on $[0, 1]$, the Gauss method of order $2s$ is defined by means of the following polynomial collocation problem:

$$\begin{cases} \sigma(t_0) = y_0, \\ \dot\sigma(t_0 + c_i h) = J\nabla H(\sigma(t_0 + c_i h)), \quad i = 1, \dots, s. \end{cases} \tag{2}$$

As is well known, conditions (2) uniquely define a polynomial $\sigma(t)$ of degree s which is used to advance the solution by posing $y_1 = \sigma(t_0 + h)$, while the internal stages satisfy $Y_i = \sigma(t_0 + c_i h)$, $i = 1, \dots s$. The coefficients of the Butcher array and the weights are given by

$$b_j = \int_0^1 \ell_j(c)\mathrm{d}c, \qquad a_{ij} = \int_0^{c_i} \ell_j(c)\mathrm{d}c, \qquad \text{with} \;\; \ell_j(c) = \prod_{r\neq j} \frac{c - c_r}{c_j - c_r}.$$

The s-degree polynomial $\sigma(t)$ may be thought of as a path in the phase space linking the state vectors y_0 to y_1 and passing through the stages $\{Y_i\}$. Due to the conservative nature of the vector field, we have that

$$H(y_1) - H(y_0) = \int_\sigma \nabla H(y) \cdot \mathrm{d}y = h \int_0^1 \dot\sigma(t_0 + \tau h)^T \nabla H(\sigma(t_0 + \tau h))\mathrm{d}\tau. \tag{3}$$

Now, the above integral is exactly computed by the Gauss quadrature formula with abscissae $\{c_i\}$ and weights $\{b_i\}$ if the degree of the integrand is not greater than $2s-1$ which means that the degree of $H(y)$, say ν, must not exceed 2 (linear or quadratic Hamiltonians only). Under this assumption, taking into account the collocation conditions (2), we obtain

$$\begin{aligned} H(y_1) - H(y_0) &= h \textstyle\sum_{i=1}^s b_i (\dot\sigma(t_0 + c_i h))^T \nabla H(\sigma(t_0 + c_i h)) \\ &= -h \textstyle\sum_{i=1}^s b_i \nabla^T H(\sigma(t_0 + c_i h)) J \nabla H(\sigma(t_0 + c_i h)) = 0. \end{aligned} \tag{4}$$

Thus, by following a different route, we have obtained the classical result that the Gauss methods conserve quadratic Hamiltonian functions while fail to conserve polynomial Hamiltonian functions of higher degree.[1]

[1] This argument may be generalized to other classes of collocation methods.

The above example is the starting point of our approach: the *discrete line integral* is the first sum in (4), which turns out to vanish for quadratic Hamiltonians, due to the collocation conditions (2).

The next section reports a descriptive introduction to HBVMs with much emphasis to the key ideas they rely on. We refer the reader to the papers [3, 4, 18, 2, 5, 13, 14, 1] for the details about the basic theory and implementation of HBVMs, and to the monograph [6] as a reference for the theory of BVMs.

In Section 3 we report a number of test problems of some relevance in the literature, for which the precise conservation of the energy turns out to be a crucial feature for the correct reproduction of the long time behavior of the solutions. This will be testified by comparing HBVMs to the Gauss method which, by the way, is a symplectic integrator.

2 Hamiltonian Boundary Value Methods

In this section we introduce HBVMs by slightly elaborating the arguments in [3, 4, 5]. As was said above, the basic idea which HBVMs rely on is the so called discrete line integral, which is the discrete counterpart of the line integral associated with a conservative vector field. In more detail, starting from (3), we consider a polynomial, of degree at most s, such that

$$\sigma(t_0) = y_0, \qquad \sigma(t_0 + h) = y_1, \tag{5}$$

providing an approximation to the solution on the interval $[t_0, t_0 + h]$. We consider the following expansions,

$$\dot\sigma(t_0 + \tau h) = \sum_{j=1}^{s} P_j(\tau)\gamma_j, \qquad \sigma(t_0 + \tau h) = y_0 + h\sum_{j=1}^{s} \int_0^\tau P_j(x)\mathrm{d}x\,\gamma_j, \tag{6}$$

where the (vector) coefficients $\{\gamma_i\}$ are to be determined. We also assume that the polynomials $\{P_i\}$ constitute an orthonormal basis, on the interval $[0, 1]$, for the vector space Π_{s-1} of polynomials of degree at most $s - 1$, i.e.,

$$\int_0^1 P_i(\tau)P_j(\tau) = \delta_{ij}, \qquad i, j = 1, \ldots, s.$$

Such polynomials can be easily obtained by a suitable scaling of the shifted Legendre polynomials [5]. Substitution of the first expansion in (6) into the line integral in (3), which we require to vanish, then gives

$$\sum_{j=1}^{s} \gamma_j^T \int_0^1 P_j(\tau)\nabla H(\sigma(t_0 + \tau h))\mathrm{d}\tau = 0,$$

which is certainly satisfied by choosing

$$\gamma_j = \int_0^1 P_j(\tau) J \nabla H(\sigma(t_0 + \tau h)) \mathrm{d}\tau, \qquad j = 1, \ldots, s. \tag{7}$$

Multiplication of (7) by $h \int_0^c P_j(x)\mathrm{d}x$ and summation over j then gives, by virtue of the second expansion in (6),

$$\sigma(t_0 + ch) = y_0 + h \sum_{j=1}^{s} \int_0^c P_j(x)\mathrm{d}x \int_0^1 P_j(\tau) J \nabla H(\sigma(t_0 + \tau h))\mathrm{d}\tau, \qquad c \in [0, 1]. \tag{8}$$

Let us now assume that $H(y)$ is a polynomial of degree ν. Consequently, the integral appearing at the right-hand side in (8) can be exactly discretized by a Gaussian formula over k Gauss-Legendre abscissae $\{c_i\}$, which we shall consider hereafter, provided that

$$k \geq \frac{\nu s}{2}. \tag{9}$$

Let us denote by $\{\omega_i\}$ the weights of the quadrature formula, and set

$$y_i = \sigma(t_0 + c_i h), \quad a_{ij} = \int_0^{c_i} P_j(x)\mathrm{d}x, \qquad i = 1, \ldots, k, \quad j = 1, \ldots, s. \tag{10}$$

Consequently, (8) can be (exactly) discretized as:

$$y_i = y_0 + h \sum_{j=1}^{s} a_{ij} \sum_{\ell=1}^{k} \omega_\ell P_\ell(c_\ell) J \nabla H(y_\ell), \qquad i = 1, \ldots, k. \tag{11}$$

Definition 2.1 *The set of equations (11), to be solved for the unknowns $\{y_i\}$, defines an* HBVM with k steps and degree s, *in short HBVM(k, s).*

For such method, the following properties hold true [4]:

- it has order $2s$ for all $k \geq s$;
- it is symmetric and perfectly A-stable (i.e., its stability region coincides with the left-half complex plane, $\mathbb{C}^-$ [6]);
- for $k = s$, it reduces to the Gauss-Legendre method of order $2s$;
- it exactly preserves polynomial Hamiltonian functions of degree ν, provided that (9) holds true.

Remark 2.2 *The actual implementation of HBVM(k, s) can be seen to result in the solution of a system of (block) size s, whatever is the value of k considered [3, 5]. Consequently, if needed, large values of k can be easily considered.*

The arguments in the previous remark, allow us to consider the limit formula of (10)–(11), in the case where $H(y)$ is non-polynomial, as $k \to \infty$. Clearly such a limit is given by formula (8), which, according to [4], is named *HBVM*(∞, s) or ∞-*HBVM of degree* s.

However, we emphasize that formula (8) becomes an operative method only after that a suitable discretization of the inner integral is considered and, replacing the integral by a quadrature formula with k nodes, leads back to a HBVM(k, s) method.

One can easily argue that, since in the non polynomial case the quadrature formula can approximate the corresponding integral with an arbitrary accuracy, under suitable regularity assumptions for $H(y)$, a *practical* conservation of the energy may be obtained [4, 17]. The term "practical" means that, in many general situations, when k is high enough, the method makes no distinction between the function $H(y)$ and its polynomial approximation, being the latter in a neighborhood of size ε of the former, where ε denotes the machine precision.

We end this section by observing that, by differentiating both members of (8), one obtains

$$\dot\sigma(t_0 + ch) = \sum_{j=1}^{s} P_j(c) \int_0^1 P_j(\tau) J \nabla H(\sigma(t_0 + \tau h)) \mathrm{d}\tau, \qquad c \in [0, 1],$$

which at the points $\{c_i\}$ provides, assuming $H(y)$ to be a polynomial and k large enough:

$$\dot\sigma(t_0 + c_i h) = \sum_{j=1}^{s} P_j(c_i) \int_0^1 P_j(\tau) J \nabla H(\sigma(t_0 + \tau h)) \mathrm{d}\tau, \qquad i = 1, \ldots, k.$$

Such formulae (the former being the limit of the latter as $k \to \infty$) can be regarded as a kind of *extended collocation conditions* that generalize conditions (2), according to [18, Section 2] (see also [4]).

3 Numerical tests

We present a few numerical test highlighting the good behavior of HBVMs in the long-time simulation of Hamiltonian systems. A direct comparison of HBVMs with Gauss methods is reported in order to better emphasize the stability properties of the former methods even when compared to a well known class of symplectic formulae.[2]

The use of a large stepsize of integration is a prerogative in long-time simulation of an evolutionary problem but, in general, one is forced to reduce h under a critical threshold in order to guarantee the qualitative behavior of the theoretical solution to be well reproduced by the numerical solution. From this point of view, we show that HBVMs

[2] As was seen in the previous section, the choice of Gauss methods has also been dictated by the fact that they represent the *generating* formulae of HBVMs when we use a Gauss distribution of the abscissae, namely the Gauss method of order $2s$ coincides with HBVM(s,s).

allow the use of larger stepsizes than Gauss methods, which states that the conservation of the Hamiltonian function plays an important role in detecting the correct topological features of the solutions.

3.1 Sitnikov's problem

One of the main problems in Celestial Mechanics is to describe the motion of N point particles of positive mass $\{m_i\}$ moving under Newton's law of gravitation when we know their positions $\{q_i\}$ and momenta $\{p_i\}$ at a given time. Such a dynamical system, called the N-body problem, is in the form (1), with Hamiltonian

$$H(\boldsymbol{q},\boldsymbol{p}) = \frac{1}{2}\sum_{i=1}^{N}\frac{||p_i||_2^2}{m_i} - G\sum_{i=1}^{N} m_i \sum_{j=1}^{i-1}\frac{m_j}{||q_i - q_j||_2}, \tag{12}$$

with G the gravitational constant. While the two-body problem is completely solved in the sense that we can describe explicitly all its solutions (see, e.g., [12]), this is no more the case, for $N \geq 3$. Consequently, numerical simulation is of interest, in such a case.

The Sitnikov problem is a particular configuration of the 3-body dynamics. In this problem two bodies of equal mass (primaries) revolve about their center of mass, here placed at the origin, in elliptic orbits in the (x, y)-plane. A third, and much smaller body (planetoid), is placed on the z-axis with initial velocity parallel to this axis as well.

The third body is small enough that the two body dynamics of the primaries is not destroyed. Then, the motion of the third body will be restricted to the z-axis and oscillating around the origin but not necessarily periodic. In fact, this problem has been shown to exhibit a chaotic behavior when the eccentricity of the orbits of the primaries exceeds a critical value that, for the data set we have used, is $\bar{e} \simeq 0.725$ (see Figure 1).

We have solved the Kepler problem with Hamiltonian function (12) by the Gauss method of order 4 (HBVM(2,2)) and by HBVM(18,2) (order 4 and 18 steps), with the following set of parameters:

N	G	m_1	m_2	m_3	e	d	h	$t_{\max}$
3	1	1	1	10^{-5}	0.75	5	0.5	1500

where e is the eccentricity, d is the distance of the apocentres of the primaries (points at which the two bodies are the furthest), h is the time-step and $[0, t_{\max}]$ is the time integration interval. The eccentricity e and the distance d may be used to define the initial condition $\boldsymbol{y}_0 = [\boldsymbol{q}_0, \boldsymbol{p}_0]$ (see [19] for the details):

$$\begin{aligned}
\boldsymbol{q}_0 &= [-\tfrac{5}{2},\ 0,\ 0,\ \tfrac{5}{2},\ 0,\ 0,\ 0,\ 0,\ 10^{-9}],\\
\boldsymbol{p}_0 &= [0,\ -\tfrac{1}{20}\sqrt{10},\ 0,\ 0,\ \tfrac{1}{20}\sqrt{10},\ 0,\ 0,\ 0,\ \tfrac{1}{2}].
\end{aligned}$$

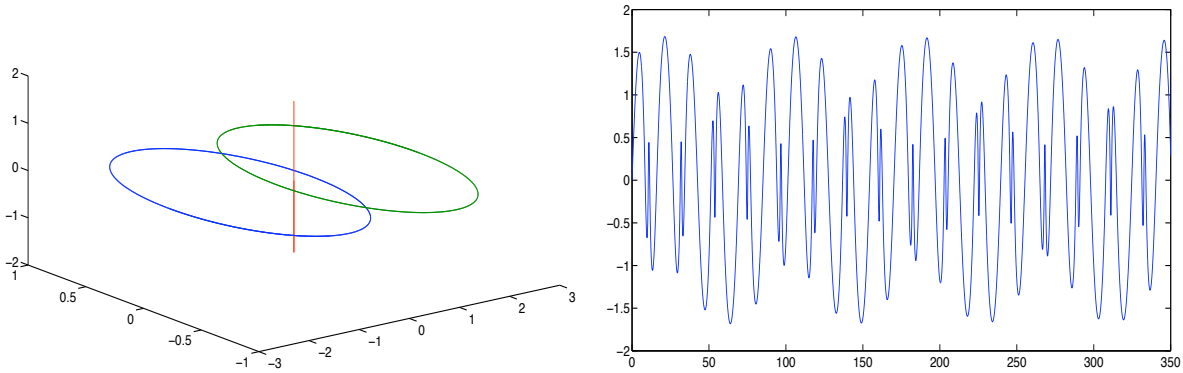


Figure 1.— The left picture displays the configuration of 3-bodies in the Sitnikov problem. To an eccentricity of the orbits of the primaries $e = 0.75$, there correspond bounded chaotic oscillations of the planetoid as is argued by looking at the space-time diagram in the right picture.

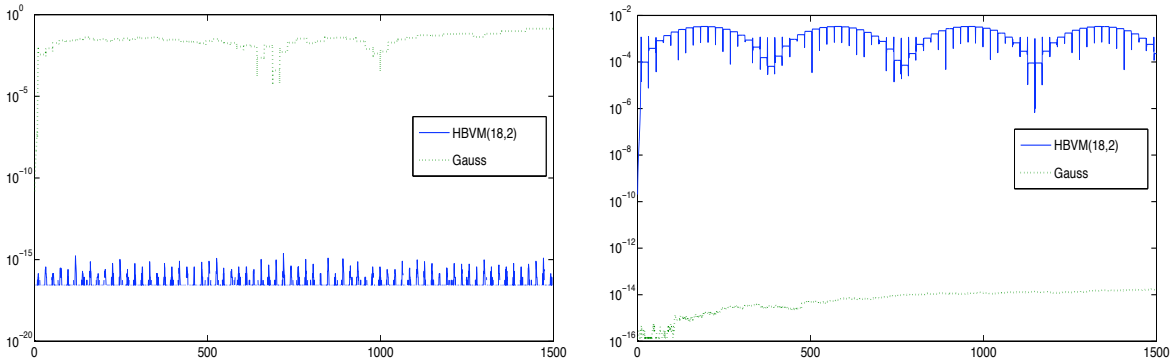


Figure 2.— Left picture: relative error $|H(y_n) - H(y_0)|/|H(y_0)|$ of the Hamiltonian function evaluated along the numerical solution of the HBVM(18,2) and the Gauss method. Right picture: relative error $|M(y_n) - M(y_0)|/|M(y_0)|$ of the angular momentum evaluated along the numerical solution of the HBVM(18,2) and the Gauss method.

First of all, we consider the two pictures in Figure 2 reporting the relative errors in the Hamiltonian function and in the angular momentum evaluated along the numerical solutions computed by the two methods. According to (9), we know that the HBVM(18,2) precisely conserves Hamiltonian polynomial functions of degree at most 18. This accuracy is high enough to guarantee that the nonlinear Hamiltonian function (12) is as well conserved up to the machine precision (see the left picture): from a geometrical point of view, this means that a local approximation of the level curves of (12) by a polynomial of degree 18 leads to a negligible error. The Gauss method exhibits a certain error in the Hamiltonian function while, being this formula symplectic, it precisely conserves the angular momentum, as is confirmed by looking at the right picture of Figure 2. From the same picture, one sees that the error in the numerical angular momentum associated with the HBVM(18,2) undergoes some bounded periodic-like oscillations.

Figures 3 and 4 show the numerical solution computed by the Gauss method and

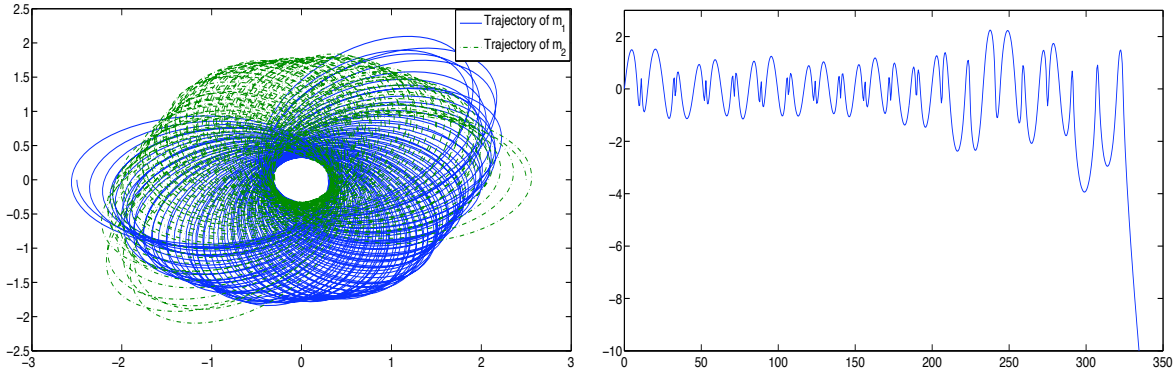


Figure 3.— The Sitnikov problem solved by the Gauss method of order 4, with stepsize $h = 0.5$, in the time interval $[0, 1500]$. The trajectories of the primaries in the (x, y)-plane (left picture) exhibit a very irregular behavior which causes the planetoid to eventually leave the system, as illustrated by the space-time diagram in the right picture.

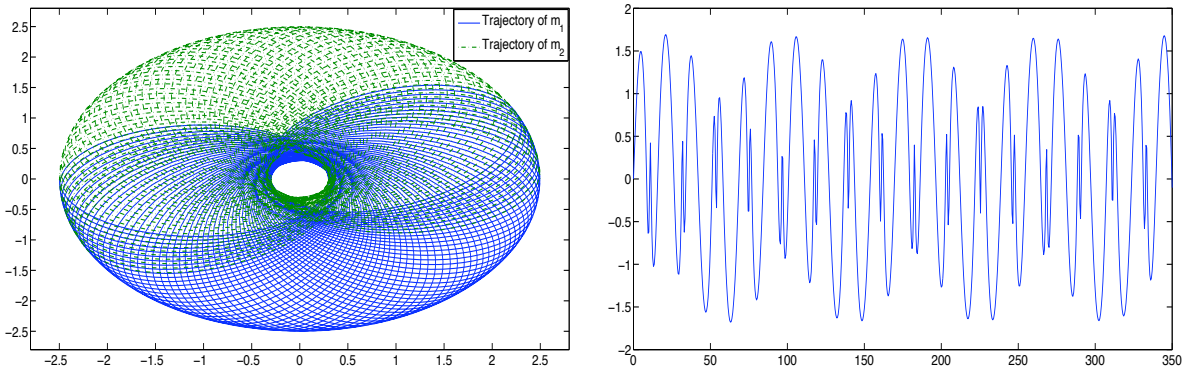


Figure 4.— The Sitnikov problem solved by the HBVM(18,2) method (order 4), with stepsize $h = 0.5$, in the time interval $[0, 1500]$. Left picture: the trajectories of the primaries are ellipse shape. The discretization introduces a fictitious uniform rotation of the (x, y)-plane which, however, does not alter the global symmetry of the system. Right picture: the space-time diagram of the planetoid on the z-axis displayed (for clearness) on the time interval $[0, 350]$ shows that, although a large value of the stepsize h has been used, the overall behavior of the dynamics is well reproduced (compare with the right picture of Figure 1).

HBVM(18,2), respectively. Since the methods leave the (x, y)-plane invariant for the motion of the primaries and the z-axis invariant for the motion of the planetoid, we have just reported the motion of the primaries in the (x, y)-phase plane (left pictures) and the space-time diagram of the planetoid (right picture).

We observe that, for the Gauss method, the orbits of the primaries are irregular in character so that the third body, after performing some oscillations around the origin, will eventually leave the system (see the right picture of Figure 3). On the contrary (left picture of Figure 4), the HBVM(18,2) generates a quite regular phase portrait. Due to the large stepsize h used, a sham rotation of the (x, y)-plane appears which, however, does

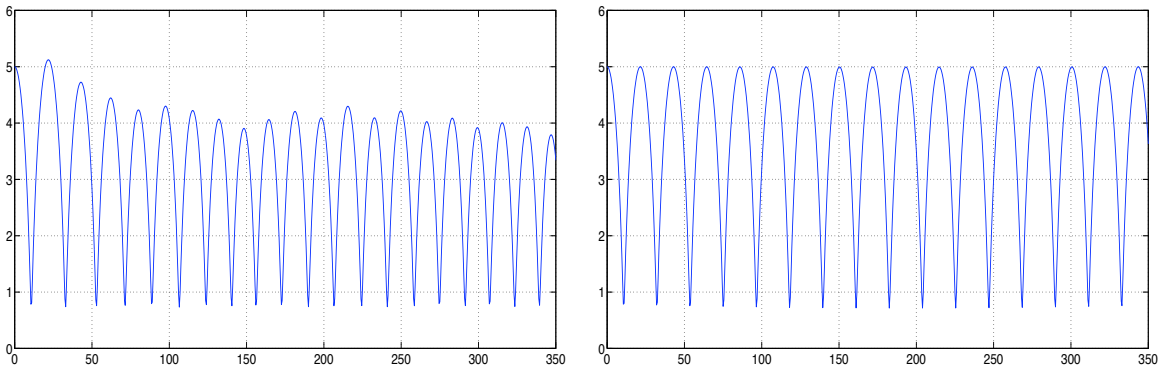


Figure 5.— Distance between the two primaries as a function of the time, related to the numerical solutions generated by the Gauss method (left picture) and HBVM(18,2) (right picture). The maxima correspond to the distance of apocentres. These are conserved by HBVM(18,2) while the Gauss method introduces patchy oscillations that destroy the overall symmetry of the system.

not destroy the global symmetry of the dynamics, as testified by the bounded oscillations of the planetoid (right picture of Figure 4) which look very similar to the reference ones in Figure 1. This aspect is also confirmed by the pictures in Figure 5, displaying the distance of the primaries as a function of the time. We see that the distance of the apocentres (corresponding to the maxima in the plots), as the two bodies wheel around the origin, are preserved by the HBVM(18,2) (right picture) while the same is not true for the Gauss method (left picture).

3.2 The Hénon-Heiles problem

The Hénon-Heiles equation originates from a problem in Celestial Mechanics describing the motion of a star under the action of a gravitational potential of a galaxy which is assumed time-independent and with an axis of symmetry (the z-axis) (see [11] and references therein). The main question related to this model was to state the existence of a third first integral, beside the total energy and the angular momentum.[3] By exploiting the symmetry of the system and the conservation of the angular momentum, Hénon and Heiles reduced from three (cylindrical coordinates) to two (planar coordinates) the degrees of freedom, thus showing that the problem was equivalent to the study of the motion of a particle in a plane subject to an arbitrary potential $U(q_1, q_2)$:

$$H(\boldsymbol{q}, \boldsymbol{p}) = \frac{1}{2}(p_1^2 + p_2^2) + U(q_1, q_2). \tag{13}$$

Since U in (13) has no symmetry in general, we cannot consider the angular momentum as an invariant anymore, so that the only known first integral is the total energy

[3] An analytical approach to the problem may be found in [10], where the author finds out a formal expansion of the third invariant.

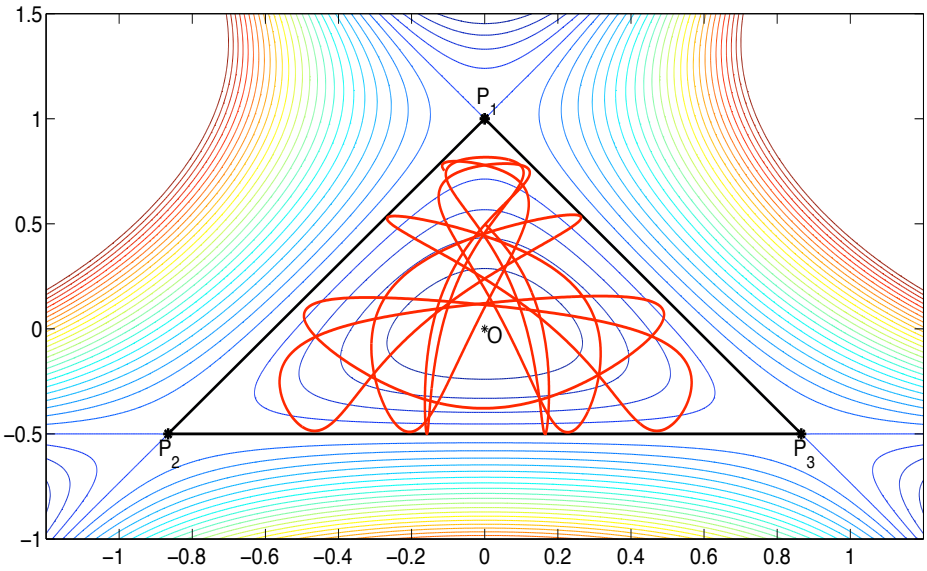


Figure 6.— Level curves of the potential $U(q_1, q_2)$ of the Hénon-Heiles problem (see (14)). The origin O is a stable equilibrium point, whose domain of stability contains the equilateral triangle having as vertices the saddle points P_1, P_2, and P_3, provided that the total energy does not exceed the value $\frac{1}{6}$. Inside the triangle an orbit $(q_1(t), q_2(t))$ is traced whose total energy is close (but lower than) $\frac{1}{6}$. The trajectory gets very close to the sides of the triangle, which makes the problem of conserving the total energy in the numerical solution an important feature to avoid instability when a large stepsize is used.

represented by (13) itself, and the question is whether or not a second integral does exist. Hénon and Heiles conducted a series of tests with the aim of giving a numerical evidence of the existence of such integral for moderate values of the energy H, and of the appearance of chaotic behavior when $H(\boldsymbol{q}, \boldsymbol{p})$ becomes larger than a critical value. In particular, for their experiments they choose

$$U(q_1, q_2) = \frac{1}{2}(q_1^2 + q_2^2) + q_1^2 q_2 - \frac{1}{3} q_2^3, \tag{14}$$

which makes the Hamiltonian function a polynomial of degree three.

When $U(q_1, q_2)$ approaches the value $\frac{1}{6}$, the level curves of U tend to an equilateral triangle, whose vertices are saddle points of U (see Figure 6). This vertices have coordinates $P_1 = (0, 1)$, $P_2 = (-\frac{\sqrt{3}}{2}, -\frac{1}{2})$ and $P_3 = (\frac{\sqrt{3}}{2}, -\frac{1}{2})$.

We consider an initial point $(\boldsymbol{q}_0, \boldsymbol{p}_0)$ such that $\boldsymbol{q}_0$ is inside the triangle $U \leq \frac{1}{6}$ and $H(\boldsymbol{q}_0, \boldsymbol{p}_0) < \frac{1}{6}$: then the orbit originating from $(\boldsymbol{q}_0, \boldsymbol{p}_0)$ will never abandon the triangle for any value of the time t. However, when $H(\boldsymbol{q}_0, \boldsymbol{p}_0)$ is chosen very close to $\frac{1}{6}$, a numerical method which does not preserve exactly the total energy could cause the (numerical) orbit to jump outside the triangle and possibly to diverge to infinity. This aspect is further emphasized when a large stepsize of integration is used, as is usually required in the long time simulation of a dynamical system.

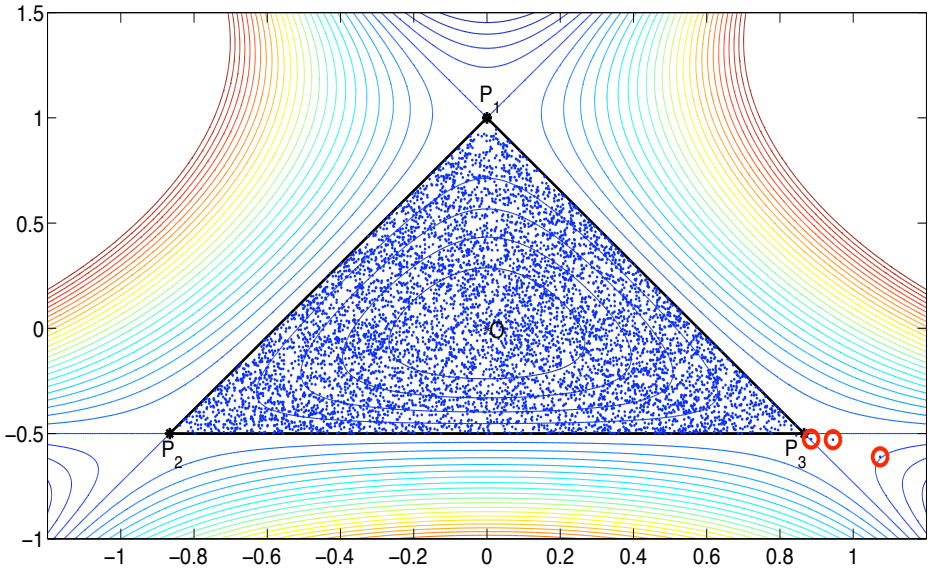


Figure 7.— The numerical trajectory in the (q_1, q_2)-plane computed by the Gauss method of order four with stepsize $h = 1$. The stable character of the continuous orbit is not correctly reproduced by the numerical method: after a time $t \simeq 7000$ the orbit escapes from the triangle (see the dots surrounded by small circles at the bottom right of the picture).

We have integrated problem (13) in the time interval $[0, 5 \cdot 10^4]$ with stepsize $h = 1$ by using the Gauss method of order four (HBVM(2,2)) and the HBVM(4,2) method which assures an exact conservation of the total energy.

Figures 7 and 8 show the numerical trajectories in the (q_1, q_2)-plane as dots that eventually will densely fill the triangle. The orbit generated by the Gauss method is plotted up to time $t \simeq 7000$, since it then escapes from the triangle, as highlighted by the three circles close to the saddle point P_3. In fact, as Figure 9 shows, the numerical Hamiltonian function associated with the Gauss method produces very irregular oscillations around the theoretical value (straight line) which eventually determine a loss of stability.

On the contrary, all the 50000 dots of the numerical trajectory computed by the HBVM(4,2) method are visible in Figure 8.

3.3 Computing the period annulus of a non-degenerate center of a polynomial Hamiltonian planar system.

Non-degenerate centers[4] of planar, in particular polynomial, Hamiltonian systems are extensively researched in the modern literature (see [9, 7, 22, 8] and references therein). The integration of such systems by means of HBVMs deserves a particular interest because, the degrees of freedom being one, the corresponding numerical solution is guaran-

[4] We recall that a center is an equilibrium point which is surrounded by periodic orbits. It is *non-degenerate* if the linearized vector field at this point has non-zero eigenvalues.

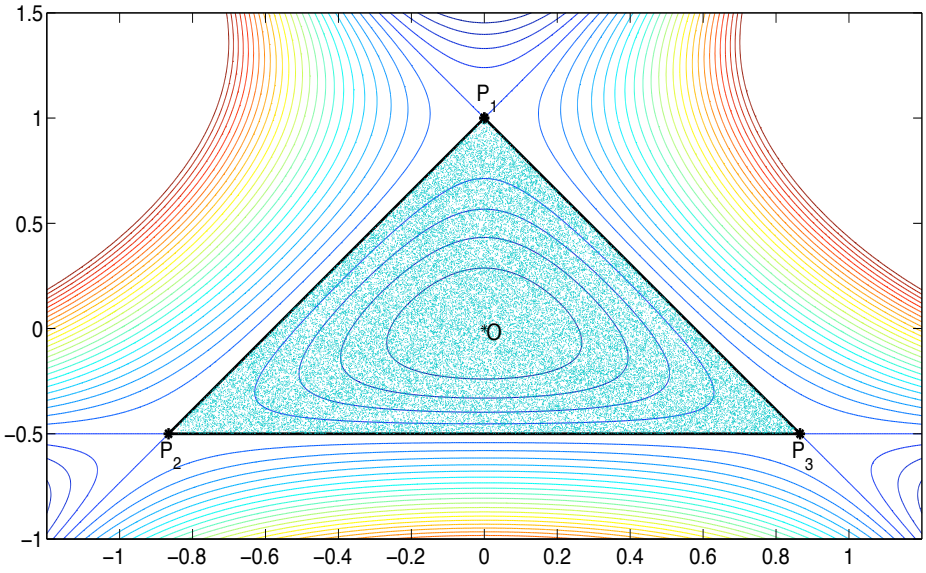


Figure 8.— The numerical trajectory in the (q_1, q_2)-plane computed by the HBVM(4,2) method with stepsize $h = 1$. Since this method precisely conserves the total energy of the system, the orbit is entirely contained in the triangle at all times.

teed to lie on the same level set $H(q,p) = H(q_0,p_0)$ as the theoretical orbit. Furthermore, if this latter consists of a closed orbit surrounding an equilibrium point (center), the numerical solution will (in general) fill densely the corresponding closed level curve, thus reproducing the very same phase portrait associated with the original continuous problem.

The region of marginal stability of a center P_0, is called the *period annulus of* P_0 and will be denoted by $\mathcal{P}$: it is the largest punctured neighborhood of the center consisting of only periodic orbits. The function which associates to any periodic orbit in $\mathcal{P}$ its period is called the *period function* of the center. Such function has been being intensively studied for many years: its behavior relates to problems of isochronicity,[5] monotonicity, bifurcation of its critical points, etc.

The aim of the present example is to consider one such system and try to reproduce numerically, as best as possible, the set $\partial\mathcal{P}$, that is the boundary of the period annulus $\mathcal{P}$. Let $H^* < +\infty$ be the value of the Hamiltonian function corresponding to any points on $\partial\mathcal{P}$.[6] The Hamiltonian function we consider here is the fifth-degree polynomial

$$H(p,q) = A(p) + B(p)q + C(p)q^2 + D(p)q^3, \tag{15}$$

where

$$A(p) = p^2(\tfrac{1}{2} + c_3 p + b_3 p^2 + a_3 p^3), \qquad B(p) = p^2(c_2 + b_2 p + a_2 p^2),$$
$$C(p) = \tfrac{1}{2} + c_1 p + b_1 p^2 + a_1 p^3, \qquad D(p) = c_0 + b_0 p + a_0 p^2,$$

[5] Namely, all the orbits surrounding the center P_0 share the same period.

[6] Here we assume that the center P_0 is non global: this is certainly true if $H(q,p)$ is a polynomial of odd degree.

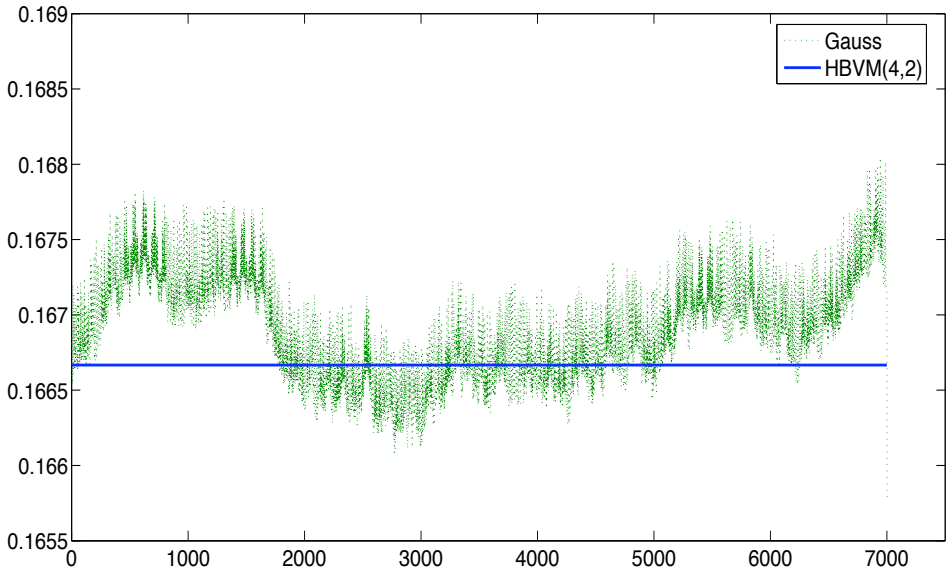


Figure 9.— Hamiltonian function evaluated along the numerical solution of the Gauss and HBVM(4,2) methods. The irregular oscillations introduced by the Gauss method will cause the associated numerical solution to eventually leave the stability region centered at the origin.

with $(a_0, a_1, a_2, a_3) \neq (0, 0, 0, 0)$.[7] Note that, since $H(q, p) = \frac{1}{2}(p^2 + q^2) + \text{h.o.t.}$, we can assume P_0 to be the origin $O = (0, 0)$.

The class of Hamiltonian systems defined by (15) has been proposed in [20] and [21].[8] Their main result was proving that the origin may not be an isochronous center [20] and, more specifically, that the period tends to infinity as $H(q_0, p_0) \nearrow H^*$, (q_0, p_0) being the initial condition associated with the differential system.

For our experiments, we have set the values of the coefficients $\{a_i\}$, $\{b_i\}$, and $\{c_i\}$ as follows:

$$\begin{array}{llll} a_0 = 0; & a_1 = 0; & a_2 = 1; & a_3 = 0; \\ b_0 = 0; & b_1 = 1; & b_2 = 0; & b_3 = 1; \\ c_0 = 0; & c_1 = 1; & c_2 = 1; & c_3 = 0. \end{array} \tag{16}$$

In such a case, besides the origin $P_0 = (0, 0)$, $H(q, p)$ admits the following real equilibrium points (up to the machine precision):

$$P_1 = (-6.879526475540134 \cdot 10^{-1},\ -5.206527058470621 \cdot 10^{-1}) \longrightarrow \text{saddle point};$$
$$P_2 = (-1.179582379893681,\ 1.756351969248087) \longrightarrow \text{saddle point}.$$

[7] Otherwise the degree of $H(q, p)$ becomes lower than 5.

[8] The authors showed that, without loss of generality, the form (15) may be associated to any polynomial Hamiltonian system of degree four and admitting a non-degenerate center, via a suitable change of coordinates.

Figure 10 reports the shape of the level curves of (15)–(16) in a region enclosing P_0 and P_1. We see that the limit closed orbit corresponding to $\partial\mathcal{P}$ is the one embracing P_0 and having P_1 as both ω-limit point and α-limit point[9] and, therefore, the value H^* may be computed with precision as

$$H^* = H(P_1) = 9.050199350868576 \cdot 10^{-2}. \tag{17}$$

Now suppose we do not know the value H^* in (17) (it will be used as a reference value) and that we want to reproduce the orbit covering $\partial\mathcal{P}$ by simply picking initial points (q_0, p_0) further and further away from the origin, and checking whether the numerical solution remains bounded over a long time.[10] More precisely, we will locate the limit cycle by means of a dichotomic search, according to the following algorithm:

step 1: find a point Q from which an orbit originates that does not embraces the critical point P_0 (that is $Q \notin \mathcal{P}$);

step 2: consider the segment joining P_0 to Q:

$$\gamma(c) = (1-c)P_0 + cP_1, \qquad c \in [0,1],$$

and set $c_0 = 0$ and $c_1 = 1$;

step 3: if $c_1 - c_0 <$ tol, STOP (tol is a specified tolerance);

step 4: set $c = \frac{c_0+c_1}{2}$ and solve numerically the Hamiltonian problem defined in (15), considering $\gamma(c)$ as initial condition, in the time interval $[0, hN]$ where $h > 0$ is the stepsize and N is a positive integer such that hN is large enough to give some information about the fate of the orbit originating from $\gamma(c)$.

step 5: if the numerical solution eventually depart from P_0, set $c_1 = c$, otherwise set $c_0 = c$, go to step 3;

The point $y_0 \equiv (q_0, p_0) = \gamma(c)$, where c is the value resulting after the execution of the above procedure, may be assumed as a point on $\partial\mathcal{P}$ within the specified tolerance tol. Detecting the limit cycle with high accuracy requires a huge number of simulations and therefore large run times, also taking into account the wide time intervals that must be used in order to inspect the asymptotic behavior of the numerical solution.[11] Consequently, it would be advisable to work with a relatively large stepsize h. We have set:

$$h = 1,\ 0.5, \quad N = 2500,\ 5000, \quad \text{tol} = 2^{-52} \text{ (i.e., the value of } \texttt{eps} \text{ in Matlab)}, \quad Q = (0,1),$$

to cover the integration interval $[0, 2500]$.

[9] That is, $\lim_{t\to\pm\infty}(q(t), p(t)) = P_1$ for any choice of $(q_0, p_0) \in \partial P$.

[10] Of course, we cannot assume $(q_0, p_0) = P_1$ since P_1 is an equilibrium point.

[11] Actually, by virtue of their conservation properties, HBVMs do not need to be integrated over a long time, even though here we do that for comparison purposes.

h	s	a point $y_0^{(s,s)} \in \partial\mathcal{P}$ computed by the Gauss method	$\frac{\lvert H(y_0^{(s,s)})-H^*\rvert}{H^*}$	a point $y_0^{(k,s)} \in \partial\mathcal{P}$ computed by HBVM(k,s)	$\frac{\lvert H(y_0^{(k,s)})-H^*\rvert}{H^*}$
1	2	$(0,\ 3.723580509957994\cdot 10^{-1})$	$2.15\cdot 10^{-2}$	$(0,\ 3.757055929263451\cdot 10^{-1})$	$7.66\cdot 10^{-16}$
	3	$(0,\ 3.748759009745006\cdot 10^{-1})$	$5.38\cdot 10^{-3}$	$(0,\ 3.757055929263451\cdot 10^{-1})$	$4.60\cdot 10^{-16}$
	4	$(0,\ 3.754691919292651\cdot 10^{-1})$	$1.53\cdot 10^{-3}$	$(0,\ 3.757055929263450\cdot 10^{-1})$	$1.22\cdot 10^{-15}$
	5	$(0,\ 3.756914213384024\cdot 10^{-1})$	$9.20\cdot 10^{-5}$	$(0,\ 3.757055929263451\cdot 10^{-1})$	$4.60\cdot 10^{-16}$
$\frac{1}{2}$	2	$(0,\ 3.756045691696934\cdot 10^{-1})$	$6.56\cdot 10^{-4}$	$(0,\ 3.757055929263451\cdot 10^{-1})$	$4.60\cdot 10^{-16}$
	3	$(0,\ 3.756828289241957\cdot 10^{-1})$	$1.47\cdot 10^{-4}$	$(0,\ 3.757055929263451\cdot 10^{-1})$	$4.60\cdot 10^{-16}$
	4	$(0,\ 3.757049796804918\cdot 10^{-1})$	$3.98\cdot 10^{-6}$	$(0,\ 3.757055929263451\cdot 10^{-1})$	$4.60\cdot 10^{-16}$
	5	$(0,\ 3.757055571549585\cdot 10^{-1})$	$2.32\cdot 10^{-7}$	$(0,\ 3.757055929263451\cdot 10^{-1})$	$4.60\cdot 10^{-16}$

Table 1.— A point y_0 on the boundary of the period annulus $\mathcal{P}$ is computed by the Gauss and HBVM methods of orders 4, 6, 8 and 10 (corresponding to $s = 2, 3, 4, 5$ respectively). By their very nature, if used with a sufficient number of silent stages, HBVMs produce a numerical orbit that precisely lie on the same level set $H(q,p) = H(q_0,p_0)$ as the theoretical one, therefore we see that HBVMs can locate the point y_0 with extreme precision, whatever the order and/or the stepsize used. On the contrary, Gauss methods produce a certain error that may be lowered by reducing the stepsize of integration h and/or by raising their order.

Table 1 compares the results obtained by using the Gauss (HBVM(s,s)) and HBVM(k,s) methods of orders 4, 6, 8 and 10 (therefore, since $s = 2, 3, 4, 5$, we must choose, according to (9), $k = 5, 8, 10, 13$, respectively, in order for the HBVM(k,s) to exactly conserve the Hamiltonian function). We have denoted by $y_0^{(k,s)}$ the point computed by the method HBVM(k,s), and reported the error $|H(y_0^{(k,s)}) - H^*|/H^*$ to estimate the accuracy with which each method computes the boundary of $\mathcal{P}$. As was expected, the accuracy in detecting the right boundary of the period annulus by means of HBVMs is of the same order as the machine precision whatever the order and stepsize used (indeed, the value of $y_0^{(k,s)}$ remains the same for all simulations). On the contrary, the Gauss methods produce a certain error which depends both on the stepsize and on the order used: increasing the accuracy would require a suitable reduction of the stepsize and/or a grow-up of the order. Figure 11 shows that even small oscillations of the numerical Hamiltonian function (left picture) could produce a noticeable irregularity of the numerical orbit in a neighborhood of the boundary of the period annulus (right picture). By their very nature, HBVMs succeed in detecting the set $\partial\mathcal{P}$ with an accuracy of the same order as the machine precision: the error in the Hamiltonian function is negligible (left picture) and the numerical orbit correctly passes through the saddle point P_1.

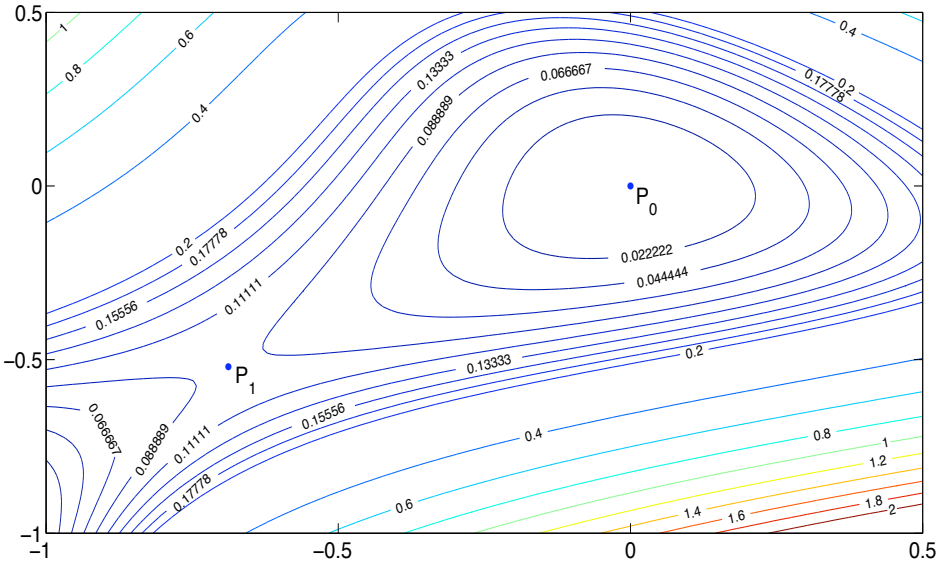


Figure 10.— Level curves of the Hamiltonian (15) in a region that embraces the center point P_0 and the saddle point P_1. Each level curve, corresponding to an orbit of the associated Hamiltonian system, is labeled by a number that indicates its elevation.

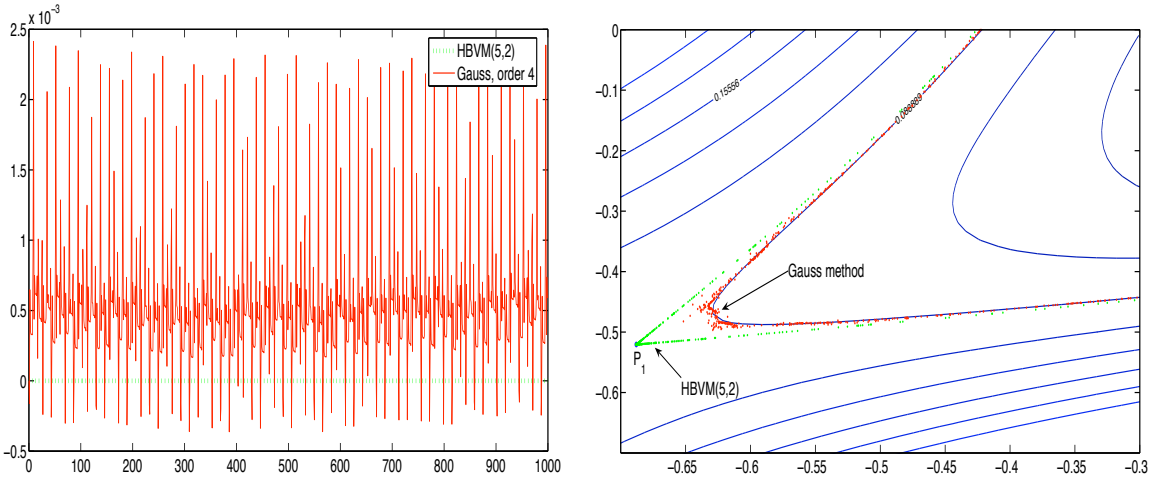


Figure 11.— Left picture: Error $H(q_n, p_n) - H(q_0, p_0)$ in the Hamiltonian function corresponding to the numerical solutions computed by the Gauss method of order 4 and HBVM(5,2) (order 4), with stepsize $h = 1$ and initial conditions $y_0^{(2,2)}$ and $y_0^{(5,2)}$ respectively. Right picture: a closeup of the two numerical orbits in a neighborhood of the saddle point P_1 reveals the difficulty of the Gauss method in detecting the boundary of the period annulus.

References

[1] L. Brugnano, F. Iavernaro, T. Susca. Hamiltonian BVMs (HBVMs): implementation details and applications. "Proceedings of ICNAAM 2009", *AIP Conf. Proc.* **1168** (2009) 723–726.

[2] L. Brugnano, F. Iavernaro, D. Trigiante. Hamiltonian BVMs (HBVMs): a family of 'drift free" methods for integrating polynomial Hamiltonian problems. "Proceedings of ICNAAM

2009", *AIP Conf. Proc.* **1168** (2009) 715–718.

[3] L. Brugnano, F. Iavernaro, D. Trigiante. Analisys of Hamiltonian Boundary Value Methods (HBVMs): a class of energy-preserving Runge-Kutta methods for the numerical solution of polynomial Hamiltonian dynamical systems. *BIT* (2009), submitted. (arXiv:0909.5659)

[4] L. Brugnano, F. Iavernaro, D. Trigiante. Hamiltonian Boundary Value Methods (Energy Preserving Discrete Line Integral Methods). *Jour. of Numer. Anal., Industr. and Appl. Math.* (2009) submitted. (arXiv:0910.3621)

[5] L. Brugnano, F. Iavernaro, D. Trigiante. Isospectral Property of Hamiltonian Boundary Value Methods (HBVMs) and their blended implementation. *BIT* (2010) submitted (arXiv:1002.1387).

[6] L. Brugnano, D. Trigiante. *Solving Differential Problems by Multistep Initial and Boundary Value Methods*, Gordon and Breach Science Publ., Amsterdam, 1998.

[7] C.J. Christopher and C.J. Devlin. Isochronous centers in planar polynomial systems. *SIAM J. Math. Anal.* **28** (1997) 162–177.

[8] A. Cima, A. Gasull and F. Mañosas. Period function for a class of Hamiltonian systems. *J. Differential Equations* **168** (no. 1) (2000) 180–199.

[9] F. Dumortier, J. Llibre and J.C. Artés. *Qualitative theory of planar differential systems. Universitext*. Springer-Verlag, Berlin, 2006.

[10] F. Gustavson. On constructing formal integrals of a Hamiltonian system near an equilibrium point. *Astron. J.* **71** (1966) 670–686.

[11] M. Hénon and C. Heiles. The Applicability of the Third Integral of Motion: Some Numerical Experiments. *Astron. J.* **69** (no. 1) (1964) 73–79.

[12] E. Hairer, C. Lubich, G. Wanner. *Geometric Numerical Integration. Structure-Preserving Algorithms for Ordinary Differential Equations, 2^{nd} ed.*, Springer, Berlin, 2006.

[13] F. Iavernaro, B. Pace. s-Stage Trapezoidal Methods for the Conservation of Hamiltonian Functions of Polynomial Type. *AIP Conf. Proc.* **936** (2007) 603–606.

[14] F. Iavernaro, B. Pace. Conservative Block-Boundary Value Methods for the Solution of Polynomial Hamiltonian Systems. *AIP Conf. Proc.* **1048** (2008) 888–891.

[15] F. Iavernaro, D. Trigiante. On some conservation properties of the Trapezoidal Method applied to Hamiltonian systems. *ICNAAM 2005 proceedings, T.E.Simos, G.Psihoyios, Ch.Tsitouras (Eds.)*. Wiley-VCH, Weinheim, 2005, pp. 254–257

[16] F. Iavernaro, D. Trigiante. Discrete conservative vector fields induced by the trapezoidal method. *J. Numer. Anal. Ind. Appl. Math.* **1** (2006) 113–130.

[17] F. Iavernaro, D. Trigiante. State-dependent symplecticity and area preserving numerical methods. *J. Comput. Appl. Math.* **205** no. 2 (2007) 814–825.

[18] F. Iavernaro, D. Trigiante. High-order symmetric schemes for the energy conservation of polynomial Hamiltonian problems. *J. Numer. Anal. Ind. Appl. Math.* **4**,1-2 (2009) 87–111.

[19] J.D. Mireles James. Celestial mechanics notes, Set 1: Introduction to the N-Body Problem. *Available at the url*: http://www.math.utexas.edu/users/jjames/celestMech

[20] X. Jarque, J. Villadelprat. Nonexistence of isochronous centers in planar polynomial Hamiltonian systems of degree four. *J. Differential Equations* **180**, no. 2 (2002) 334–373

[21] X. Jarque, J. Villadelprat. On the period function of centers in planar polynomial Hamiltonian systems of degree four. *Qual. Theory Dyn. Syst.* **3** (no. 1) (2002) 157–180.

[22] J. Llibre and G. Rodríguez. Configurations of limit cycles and planar polynomial vector fields. *J. Differential Equations* **198** (no. 2) (2004) 374–380.

Monografías de la Real Academia de Ciencias de Zaragoza **33**, 113–124, (2010).

On the convergence of generalized polar decompositions in Lie groups

Jordi P. García-Seguí and Fernando Casas

Institut de Matemàtiques i Aplicacions de Castelló (IMAC) and

Dept. de Matemàtiques, Universitat Jaume I, 12071 Castellón, Spain.

This paper is dedicated to Prof. Manuel Calvo on the occasion of his 65th anniversary

Abstract

We analyze the so-called generalized polar decomposition determined by an involutive automorphism in a Lie group. This concept generalizes the well known factorization of a matrix as the product of a positive semidefinite matrix and an orthogonal matrix in linear algebra. We provide a different constructive proof of the existence of such a decomposition in a neighborhood of the identity and obtain several explicit bounds on the convergence domain of the series defined each factor.

1 Introduction

The polar decomposition can be seen as the matrix analog of the polar form of a complex number $z = r\,\mathrm{e}^{i\theta}$, $r > 0$. If A is any $n \times n$ matrix, then there exists a unitary matrix U and a unique Hermitian positive semidefinite matrix H such that

$$A = H\,U.$$

Furthermore, if A is invertible, then H is positive definite and U is uniquely determined. If A is real, the matrix U is orthogonal and H is symmetric. It is well known that the factors H and U possess best approximation properties. Specifically, the polar factor U is the best unitary (orthogonal in the real case) approximant to A in any unitarily invariant norm, whereas H is a good Hermitian positive definite approximation to A when it is nonsingular and Hermitian, and $\frac{1}{2}(A + H)$ is a best Hermitian positive semidefinite approximation to A [8].

The polar decomposition has been generalized to abstract Lie groups and even to semigroups [11]. Such generalized polar decomposition has been found to be closely related with the concept of involutive automorphism and the subspace decomposition it induces. In this setting, the polar decomposition is equivalent to expressing a group element as the product of a term in a symmetric subspace and a term in a subgroup of the given Lie group.

More specifically, let G be a Lie group and $\sigma : G \longrightarrow G$ an involutive automorphism. By this we mean a one-to-one map such that $\sigma(xy) = \sigma(x)\sigma(y)$, $\sigma \neq \mathrm{id}$ and $\sigma^2 = \mathrm{id}$. Let G^σ denote the subgroup of G consisting of fixed points of σ, i.e., $G^\sigma = \{x \in G : \sigma(x) = x\}$ and G_σ the set of anti-fixed points of σ, $G_\sigma = \{x \in G : \sigma(x) = x^{-1}\}$. The set G_σ is not a group, but a symmetric space with the non-associative multiplication $x \cdot y \equiv xy^{-1}x$ [7]. Then, the generalized polar decomposition of $z \in G$ consists in writing

$$z = xy, \qquad x \in G_\sigma, \quad y \in G^\sigma. \tag{1}$$

Since σ induces an involutive automorphism $d\sigma$ on the Lie algebra $\mathfrak{g}$ corresponding to G as

$$d\sigma(X) \equiv \frac{d}{dt}\Big|_{t=0} \sigma(\exp(tX))$$

for all $X \in \mathfrak{g}$, then $\mathfrak{g}$ can be expressed as the direct sum

$$\mathfrak{g} = \mathfrak{p} \oplus \mathfrak{k}, \tag{2}$$

where $\mathfrak{k}$ corresponds to the set of fixed points of $d\sigma$ ($d\sigma(X) = X$) and $\mathfrak{p}$ to the set of anti-fixed points, $d\sigma(X) = -X$. The space $\mathfrak{k}$ is a subalgebra of $\mathfrak{g}$, whereas $\mathfrak{p}$ is a Lie triple system: $\mathfrak{p}$ is a vector space that is not closed under the commutator but under the double commutator, that is, $[X_1, [X_2, X_3]] \in \mathfrak{p}$ for $X_i \in \mathfrak{p}$, whereas $[X_1, X_2] \in \mathfrak{k}$ [7]. In general, the sets $\mathfrak{p}$ and $\mathfrak{k}$ verify the following commutation relations:

$$[\mathfrak{k}, \mathfrak{k}] \subseteq \mathfrak{k}, \qquad [\mathfrak{k}, \mathfrak{p}] \subseteq \mathfrak{p}, \qquad [\mathfrak{p}, \mathfrak{p}] \subseteq \mathfrak{k}.$$

As a result, every element $Z \in \mathfrak{g}$ can be uniquely written as

$$Z = P + K, \qquad P \in \mathfrak{p}, \quad K \in \mathfrak{k} \tag{3}$$

with

$$P = \frac{1}{2}(Z - d\sigma(Z)) \qquad \text{and} \qquad K = \frac{1}{2}(Z + d\sigma(Z)).$$

Moreover, if $K \in \mathfrak{k}$, then $\exp(tK) \in G^\sigma$, whereas $P \in \mathfrak{p}$ implies $\exp(tP) \in G_\sigma$.

As a well know example, let us consider the general linear group $\mathrm{GL}(n)$ of real $n \times n$ invertible matrices and the map $\sigma(x) = (x^{-1})^T$, which is an involutive automorphism. Now the set G_σ is the set of invertible symmetric matrices (a symmetric space), whereas

G^σ is the set of orthogonal matrices (which is a subgroup of $\mathrm{GL}(n)$). It can be checked that $d\sigma(X) = -X^T$, whence $\mathfrak{k}$ is the classical algebra of skew-symmetric matrices and $\mathfrak{p}$ is the set of symmetric matrices. In consequence, the decomposition (3) is nothing but the canonical decomposition of a matrix into its skew-symmetric and symmetric part: $P = (Z + Z^T)/2$, $K = (Z - Z^T)/2$.

In a series of papers [13, 15, 16], Zanna and his collaborators have analyzed the polar decomposition in a generic Lie group G. In particular, they provide a proof of its existence and uniqueness in a neighborhood of the identity $e \in G$, which can be established as the following theorem [13].

Theorem 1.1 *Let $z = \exp(tZ) \in G$, where $Z = P + K$ is the decomposition of Z in $\mathfrak{p} \oplus \mathfrak{k}$, i.e., $d\sigma(P) = -P$ and $d\sigma(K) = K$. Then, for sufficiently small values of t, the element z admits a unique generalized polar decomposition $z = xy$, where $x = \exp(X(t))$, $X(t) \in \mathfrak{p}$, and $y = \exp(Y(t))$ with $Y(t) \in \mathfrak{k}$.*

Moreover, they derive differential equations obeyed by $X(t)$ and $Y(t)$ and solve them perturbatively, thus constructing X and Y as a power series whose terms can be obtained by a recursive procedure. These recurrences are in turn used to prove the convergence of the series when $\mathfrak{g}$ is a Banach algebra. In this way, the function $X(t)$ is shown to be analytic in a sphere of radius

$$\rho = \frac{\delta}{2\alpha} \qquad \text{for some constant} \quad 0 < \delta < \pi \tag{4}$$

and $\alpha = \max\{\|P\|, \|K\|\}$ [13], although no specific value of δ is provided. On the other hand, the radius of convergence of the series $Y(t)$ is given implicitly as $\rho = \frac{r}{2\beta}$ [15], where $\beta = \max\{t\|Z\|, \|X(t)\|\}$ and r is related to the radius of convergence of the Baker–Campbell–Hausdorff (BCH) series. Notice that these estimates are all of a qualitative nature, whereas (at least up to our knowledge) no actual bounds for the convergence domain are found in the literature.

In this paper we try to fill this gap by first proposing new computationally well adapted recurrences for generating the series $X(t)$ and $Y(t)$. These recurrences are used to get numerical estimates on the convergence of the series $X(t)$ and also a bound on $\|X(t)\|$ itself, which is then used to establish the convergence of the series $Y(t)$. These results are supplemented with sharper numerical estimates obtained from the BCH series.

Although of theoretical nature, generalized polar decompositions in Lie groups have found interesting applications in numerical analysis, namely in connection with self-adjoint numerical integrators for differential equations [10] and the numerical approximation of the exponential of a matrix from a Lie algebra to a Lie group [16], especially in $\mathrm{SL}(n)$. From a more abstract point of view, they constitute a particular instance of the Atkinson factorization theorem for Rota–Baxter algebras [3, 5]. We believe that the convergence results provided here will be of interest in these different settings.

2 Recursion for the factor X

Our starting point is the factorization provided by Theorem 1.1

$$\mathrm{e}^{tZ} = \mathrm{e}^{X(t)}\,\mathrm{e}^{Y(t)}, \tag{5}$$

with $Z = P + K$. Differentiating (5) one arrives at the expression

$$\mathrm{e}^{-X}(Z - d\exp_X(X'))\,\mathrm{e}^{X} = d\exp_Y(Y'), \tag{6}$$

where

$$d\exp_Y(Y') \equiv \sum_{j=0}^{\infty} \frac{1}{(j+1)!}\mathrm{ad}_Y^j(Y') \in \mathfrak{k} \tag{7}$$

since $Y, Y' \in \mathfrak{k}$ and $\mathfrak{k}$ is a subalgebra of the Lie algebra $\mathfrak{g}$. Here ad_A stands for the adjoint operator of $A \in \mathfrak{g}$, which acts according to

$$\mathrm{ad}_A B = [A, B], \qquad \mathrm{ad}_A^j B = [A, \mathrm{ad}_A^{j-1} B], \qquad \mathrm{ad}_A^0 B = B, \qquad j \in \mathbb{N}, B \in \mathfrak{g}. \tag{8}$$

Notice that the left hand side of eq. (6) also belongs to $\mathfrak{k}$. We therefore analyze this term and separate the contribution in $\mathfrak{p}$, which has to be canceled.

First we note that

$$\begin{aligned} \mathrm{e}^{-X} Z\,\mathrm{e}^{X} \;\; = \;\; & -\sinh(u)(K) + \cosh(u)(P) && (\in \mathfrak{p}) \\ & +\cosh(u)(K) - \sinh(u)(P) && (\in \mathfrak{k}) \end{aligned}$$

where $u \equiv \mathrm{ad}_X$ and the functions involving u have to be understood as power series. On the other hand,

$$\begin{aligned} \mathrm{e}^{-X} d\exp_X(X')\,\mathrm{e}^{X} \;\; = \;\; & \frac{1}{u}\sinh(u)(X') && (\in \mathfrak{p}) \\ & +\frac{1}{u}(1 - \cosh(u))(X') && (\in \mathfrak{k}) \end{aligned}$$

In consequence,

$$-\sinh(u)(K) + \cosh(u)(P) - \frac{1}{u}\sinh(u)(X') = 0$$

whence, after some algebra, we arrive at the differential equation satisfied by X:

$$X' = -\mathrm{ad}_X K + \sum_{k=0}^{\infty} \frac{2^{2k} B_{2k}}{(2k)!}\mathrm{ad}_X^{2k} P, \qquad X(0) = 0, \tag{9}$$

with B_j denoting the Bernoulli numbers [1]. To solve equation (9), let us introduce a parameter $\epsilon > 0$ in Z and consider instead $\epsilon Z = \epsilon(K + P)$, i.e., the decomposition

$$\mathrm{e}^{t\,\epsilon\,Z} = \mathrm{e}^{X(\epsilon,t)}\,\mathrm{e}^{Y(\epsilon,t)}.$$

The corresponding equation satisfied by $X(\epsilon,t)$ is then

$$\frac{\partial X}{\partial t} = -\epsilon\,\mathrm{ad}_X K + \sum_{k=0}^{\infty} c_{2k}\mathrm{ad}_X^{2k}(\epsilon P), \qquad X(\epsilon,0)=0, \tag{10}$$

where, for simplicity, $c_{2k} = \frac{2^{2k}B_{2k}}{(2k)!}$. Now we try to determine the solution $X(\epsilon,t)$ perturbatively as an infinite series in ϵ,

$$X(\epsilon,t) = \sum_{n=1}^{\infty} \epsilon^n X_n(t). \tag{11}$$

To do that, first we substitute expression (11) into (10), thus obtaining for each terms up to order ϵ^n the expressions

$$\begin{aligned}
\frac{\partial}{\partial t}X(\epsilon,t) &= \sum_{j=1}^{n}\epsilon^j X_j'(t) + \mathcal{O}(\epsilon^n)\\
\mathrm{ad}_X(\epsilon K) &= \sum_{j=1}^{n-1}\epsilon^{j+1}\mathrm{ad}_{X_j}K + \mathcal{O}(\epsilon^{n+1})\\
\sum_{j=1}^{n-1} c_j\,\mathrm{ad}_X^j(\epsilon P) &= \sum_{l=2}^{n}\epsilon^l\sum_{j=1}^{l-1}c_j\sum_{\substack{k_1+\cdots+k_j=l-1\\ k_1\geq 1,\ldots,k_j\geq 1}}\mathrm{ad}_{X_{k_1}}\cdots\mathrm{ad}_{X_{k_j}}P + \mathcal{O}(\epsilon^{n+1}).
\end{aligned}$$

Then, by equating successive powers of ϵ, we get

$$\begin{aligned}
X_1' &= P\\
X_l' &= -\mathrm{ad}_{X_{l-1}}K + \sum_{j=2}^{l-1}c_j\sum_{\substack{k_1+\cdots+k_j=l-1\\ k_1\geq 1,\ldots,k_j\geq 1}}\mathrm{ad}_{X_{k_1}}\cdots\mathrm{ad}_{X_{k_j}}P, \qquad l\geq 2.
\end{aligned}$$

From the initial condition, it is clear that $X_l(0)=0$ for all $l\geq 1$, so that finally we arrive at the recursion

$$\begin{aligned}
X_1(t) &= tP \qquad (12)\\
X_l(t) &= -\int_0^t \mathrm{ad}_{X_{l-1}}K\,ds + \sum_{j=2}^{l-1}c_j\sum_{\substack{k_1+\cdots+k_j=l-1\\ k_1\geq 1,\ldots,k_j\geq 1}}\int_0^t \mathrm{ad}_{X_{k_1}}\cdots\mathrm{ad}_{X_{k_j}}P\,ds, \qquad l\geq 2.
\end{aligned}$$

If this recurrence is worked out explicitly, one gets for the first terms

$$\begin{aligned}
&X_2(t) = -\tfrac{t^2}{2}[P,K], \qquad X_3(t) = -\tfrac{t^3}{6}[K,[P,K]],\\
&X_4(t) = \tfrac{t^4}{24}([P,[P,[P,K]]] - [K,[K,[P,K]]])
\end{aligned}$$

3 Recursion for the factor Y

By considering the projection of equation (6) into $\mathfrak{k}$ we have

$$\cosh(u)(K) - \sinh(u)(P) + \frac{\cosh(u)-1}{u}(X') = d\exp_Y(Y'), \tag{13}$$

where, as before, $u \equiv \mathrm{ad}_X$. Inserting equation (9) into (13) results in

$$d\exp_Y(Y') = K + \frac{1-\cosh(u)}{\sinh(u)}(P).$$

Taking into account that

$$d\exp_Y^{-1}(Y') = \sum_{j=0}^{\infty} \frac{B_j}{j!}\mathrm{ad}_Y^j(Y')$$

and the power series of the function $(1-\cosh(u))/\sinh(u)$, we get finally

$$Y' = \sum_{j=0}^{\infty} \frac{B_j}{j!}\mathrm{ad}_Y^j \left(K - 2\sum_{k=2}^{\infty} \frac{(2^k-1)B_k}{k!}\mathrm{ad}_X^{k-1}(P) \right), \qquad Y(0) = 0. \tag{14}$$

Notice that solving for $Y(t)$ requires to previously compute $X(t)$. In spite of that, in the sequel we show that it is indeed possible to construct a power series for $Y(t)$ by recurrence. We proceed in a similar way as for the X factor: introduce the parameter $\epsilon > 0$ in Z and determine the successive terms in the expansion

$$Y(\epsilon, t) = \sum_{n=1}^{\infty} \epsilon^n Y_n(t) \tag{15}$$

by inserting it into the corresponding differential equation

$$\frac{\partial Y}{\partial t} = \epsilon\, d\exp_Y^{-1} D, \qquad Y(\epsilon, 0) = 0, \tag{16}$$

where

$$D \equiv K - 2\sum_{k=2}^{\infty} d_k\, \mathrm{ad}_X^{k-1}(P) \qquad \text{and} \qquad d_k = \frac{(2^k-1)B_k}{k!}.$$

It can be shown after some elementary algebra that the r.h.s. of equation (16) can be written as

$$\epsilon\, d\exp_Y^{-1} D = \epsilon K + A + B + C + \mathcal{O}(\epsilon^{n+1})$$

with

$$\begin{aligned}
A &= -2\sum_{l=2}^{n} \epsilon^l \sum_{j=1}^{l-1} d_{j+1} \sum_{\substack{k_1+\cdots+k_j=l-1 \\ k_1\geq 1,\ldots,k_j\geq 1}} \mathrm{ad}_{X_{k_1}} \cdots \mathrm{ad}_{X_{k_j}} P \\
B &= \sum_{l=2}^{n} \epsilon^l \sum_{j=1}^{l-1} \frac{B_j}{j!} \sum_{\substack{k_1+\cdots+k_j=l-1 \\ k_1\geq 1,\ldots,k_j\geq 1}} \mathrm{ad}_{Y_{k_1}} \cdots \mathrm{ad}_{Y_{k_j}} K \qquad (17) \\
C &= -2\sum_{l=3}^{n} \epsilon^l \sum_{j=2}^{l-1} \left(\sum_{m=1}^{j-1} \frac{B_m}{m!} \sum_{\substack{k_1+\cdots+k_m=j-1 \\ k_1\geq 1,\ldots,k_m\geq 1}} \mathrm{ad}_{Y_{k_1}} \cdots \mathrm{ad}_{Y_{k_m}} \right) \\
&\qquad \left(\sum_{p=1}^{l-j} d_{p+1} \sum_{\substack{r_1+\cdots+r_p=l-j \\ r_1\geq 1,\ldots,r_p\geq 1}} \mathrm{ad}_{X_{r_1}} \cdots \mathrm{ad}_{X_{r_p}} P \right)
\end{aligned}$$

Equating powers of ϵ leads one to the recursion

$$\begin{aligned} Y_1(t) &= t\,K \\ Y_n(t) &= \sum_{j=1}^{n-1} \frac{B_j}{j!} \sum_{\substack{k_1+\cdots+k_j=n-1 \\ k_1\geq 1,\ldots,k_j\geq 1}} \int_0^t \mathrm{ad}_{Y_{k_1}} \cdots \mathrm{ad}_{Y_{k_j}} K\, ds \\ &\quad -2\sum_{j=1}^{l-1} d_{j+1} \sum_{\substack{k_1+\cdots+k_j=n-1 \\ k_1\geq 1,\ldots,k_j\geq 1}} \int_0^t \mathrm{ad}_{X_{k_1}} \cdots \mathrm{ad}_{X_{k_j}} P\, ds \qquad (18) \\ &\quad -2\sum_{j=2}^{n-1} \int_0^t d\tau \left(\sum_{s=1}^{j-1} \frac{B_s}{s!} \sum_{\substack{k_1+\cdots+k_s=j-1 \\ k_1\geq 1,\ldots,k_s\geq 1}} \mathrm{ad}_{Y_{k_1}} \cdots \mathrm{ad}_{Y_{k_s}} \right) \\ &\quad \left(\sum_{p=1}^{n-j} d_{p+1} \sum_{\substack{r_1+\cdots+r_p=n-j \\ r_1\geq 1,\ldots,r_p\geq 1}} \mathrm{ad}_{X_{r_1}} \cdots \mathrm{ad}_{X_{r_p}} P \right) \qquad n \geq 2 \end{aligned}$$

which allows us to get the explicit expression of the first terms as

$$Y_2(t) = 0, \qquad Y_3(t) = -\frac{t^3}{12}[P,[P,K]], \qquad Y_4(t) = 0.$$

As a matter of fact, it is not difficult to prove that $Y(t)$ is and odd function of t, so that in general $Y_{2n}(t) = 0$ for all n. Notice that it is necessary to previously generate the terms X_i through recurrence (12) to obtain the series $Y(t)$ by (18). Although it is indeed possible to derive another recursion involving only terms Y_i, we have found the recursion (18) more convenient not only from a computational point of view (the implementation in a symbolic package is rather straightforward) but also for establishing explicit convergence domains for the series.

4 Convergence of the expansions

We next analyze the convergence of the previous series. For that purpose we assume that $\mathfrak{g}$ is a complete normed Lie algebra endowed with a norm compatible with associative multiplication, i.e., such that $\|AB\| \leq \|A\|\,\|B\|$ for all A, B in $\mathfrak{g}$. Then it is true that

$$\|[A,B]\| \leq 2\|A\|\,\|B\|.$$

First we consider the series

$$v(\varepsilon,t) = \sum_{j=1}^{\infty} \epsilon^j \|X_j(t)\|. \qquad (19)$$

From (12) it is clear that for $l \geq 2$

$$\|X_l(t)\| \leq 2\|K\| \int_0^t \|X_{l-1}\| ds + \|P\| \sum_{j=2}^{l-1} |c_j| 2^j \sum_{\substack{k_1+\cdots+k_j=l-1 \\ k_1\geq 1,\ldots,k_j\geq 1}} \int_0^t \|X_{k_1}(s)\| \cdots \|X_{k_j}(s)\| ds$$

and thus

$$\begin{aligned}\sum_{l=2}^{N}\epsilon^l\|X_l(t)\| &\leq \sum_{l=2}^{N}2\epsilon^l\|K\|\int_0^t\|X_{l-1}\|ds \\ &+ \epsilon\|P\|\sum_{j=1}^{N-1}|c_j|2^j\sum_{l=p}^{N-1}\varepsilon^l\sum_{\substack{k_1+\cdots+k_p=l\\ k_1\geq 1,\ldots,k_p\geq 1}}\int_0^t\|X_{k_1}(s)\|\cdots\|X_{k_j}\|ds,\end{aligned}$$

where we have interchanged the order of summation in the second term.

Let us denote $v_N(\epsilon,t)=\sum_{l=1}^{N}\epsilon^l\|X_l(t)\|$. Then it is easy to show that

$$(v_N(\epsilon,t))^p=\sum_{l=p}^{pN}\epsilon^l\sum_{\substack{k_1+\cdots+k_p=l\\ k_1\geq 1,\ldots,k_p\geq 1}}\|X_{k_1}\|\cdots\|X_{k_p}\|$$

so that, in the last inequality,

$$\sum_{l=p}^{N-1}\epsilon^l\sum_{\substack{k_1+\cdots+k_p=l\\ k_1\geq 1,\ldots,k_p\geq 1}}\|X_{k_1}\|\cdots\|X_{k_p}\|\leq(v_N(\epsilon,t))^p$$

and therefore

$$v_N(\epsilon,t) \leq 2\epsilon\|K\|\int_0^t v_{N-1}(\epsilon,s)ds+\epsilon\|P\|\sum_{j=0}^{N-1}|c_j|2^j\int_0^t v_N(\epsilon,s)^j\,ds.$$

Taking the limit $N\to\infty$ in the last expression we have

$$v(\epsilon,t)\leq 2\epsilon\|K\|\int_0^t v(\epsilon,s)ds+\epsilon\|P\|\int_0^t g(2v(\epsilon,s)))\,ds \tag{20}$$

since

$$\sum_{j=0}^{\infty}|c_j|\,(2x)^j=\sum_{j=0}^{\infty}\frac{|B_j|}{j!}(2x)^j=2+x(1-\cot x)\equiv g(x). \tag{21}$$

We proceed now as follows. Let us denote $k=\|K\|$ and $p=\|P\|$ and introduce the function $G(x)=\beta x+g(x)$, with $\beta=k/p\geq 0$. Then (20) can be written as

$$v(\epsilon,t)\leq\epsilon\,p\int_0^t G(2v(\epsilon,s))ds\equiv F(\epsilon,t).$$

In this way

$$\frac{\partial F(\epsilon,t)}{\partial t}=\epsilon\,p\,G(2v(\epsilon,t))\leq\epsilon\,p\,G(2F(\epsilon,t))$$

since G is a non-decreasing function on the domain $[0,\pi)$. In fact $G(z)$ is analytic for $|z|<\pi$ with positive coefficients in the power series and has no zeros in the ball $|z|<\pi$. The last inequality can be expressed as

$$\frac{\partial F(\epsilon,t)}{\partial t}\frac{1}{G(2F(\epsilon,t))}\leq\epsilon\,p$$

so that, by integrating, we get

$$H(2F(\epsilon,t)) \le 2\,\epsilon\,p\,t$$

where $H(t) \equiv \int_0^t \frac{1}{G(x)}dx$. Now $H(z)$ is also analytic in $|z| < \pi$ and $H'(z) = \frac{1}{G(z)} \neq 0$. Then $y = H(z)$ has an inverse function $z = H^{-1}(y)$ for y in the ball $|y| < H(\pi)$, which is also analytic there. In consequence,

$$v(\epsilon,t) \le F(\epsilon,t) \le \frac{1}{2}H^{-1}(2\,\epsilon\,p\,t)$$

for t such that $2\,\epsilon\,p\,t$ belongs to the domain of H^{-1}, i.e.,

$$2\,\epsilon\,p\,t < H(\pi) = \int_0^\pi \frac{1}{G(x)}dx \equiv \xi(\beta).$$

Therefore the series $X(t)$ is assured to be convergent for $0 \le t \le t_c$, with

$$t_c \equiv \frac{1}{2\,\epsilon\,p}\xi(\beta) = \frac{1}{2\,\epsilon\,p}\int_0^\pi \frac{1}{2+(1+\beta)x - x\cot x}dx. \tag{22}$$

If we take $\epsilon = 1$ in (19), then

$$v(\epsilon=1) = \|X(t)\| = \sum_{n=1}^{\infty}\|X_n(t)\| < \frac{1}{2}H^{-1}(\xi(\beta)) = \frac{\pi}{2} \tag{23}$$

in the convergence domain defined by (22).

For illustration, we collect next the values of $\xi(\beta)$ for several values of β:

β	0	1	10
$\xi(\beta)$	1.08687	0.83751	0.31228

If instead of using the norm ratio β we work with $\alpha = \max\{k,p\}$ then a similar argument shows that the series $\|X(t)\|$ is convergent for $0 \le t \le t_c$ with

$$t_c = \frac{\xi(1)}{2\alpha} \simeq \frac{0.83751}{2\alpha}. \tag{24}$$

Notice that we have obtained a numerical value for the constant δ in (4).

A enlarged convergence domain can indeed be established by means of the Baker–Campbell–Hausdorff (BCH) formula. As is well known, the BCH formula deals with the expansion of Z in $\mathrm{e}^{X_1}\,\mathrm{e}^{X_2} = \mathrm{e}^{Z}$ in terms of nested commutators of X_1 and X_2 when they are assumed to be non-commuting operators. Specifically,

$$Z = X_1 + X_2 + \sum_{n=2}^{\infty} G_n(X_1,X_2), \tag{25}$$

where $G_n(X_1,X_2)$ is a homogeneous Lie polynomial in X_1 and X_2 of grade n; in other words, G_n can be expressed in terms of X_1 and X_2 by addition, multiplication by rational

numbers and nested commutators. This result proves to be very useful in various fields of mathematics (theory of linear differential equations [12], Lie group theory [4], numerical analysis [6]) and theoretical physics (perturbation theory, transformation theory, Quantum Mechanics and Statistical Mechanics. In particular, in the theory of Lie groups, with this theorem one can explicitly write the operation of multiplication in a Lie group in canonical coordinates in terms of the Lie bracket operation in its algebra and also prove the existence of a local Lie group with a given Lie algebra [4].

The following theorem concerning the convergence of the BCH series has been proved (see [2]).

Theorem 4.1 *Let X_1 and X_2 be two bounded elements in a Hilbert space $\mathcal{H}$ with $\dim \mathcal{H} \geq 2$. Then the BCH formula in the form (25), i.e., expressed as a series of homogeneous Lie polynomials in X_1 and X_2, converges absolutely when $\|X_1\| + \|X_2\| < \pi$.*

Here the norm is taken as the 2-norm induced by the scalar product in $\mathcal{H}$. This result can be generalized, of course, to any number of non commuting operators $X_1, X_2, \ldots, X_q$. Specifically, the series

$$Z = \log(\mathrm{e}^{X_1}\, \mathrm{e}^{X_2}\, \cdots\, \mathrm{e}^{X_q}),$$

converges absolutely if $\|X_1\| + \|X_2\| + \cdots + \|X_q\| < \pi$.

We next show how this result can be used to get a sharper bound on the convergence domain of X. As usual, we set $z = \exp(tZ)$ with $Z = P + K$ the decomposition of Z into $\mathfrak{p} \oplus \mathfrak{k}$, and denote $w \equiv (\sigma(z))^{-1} = \sigma(z^{-1})$. Then it is true that $w = \exp(tW)$, with $W = P - K$. Now, since $\sigma(x) = x^{-1}$ and $\sigma(y) = y$ in the generalized polar decomposition $z = xy$, it is clear that

$$z\sigma(z)^{-1} = xy\sigma(xy)^{-1} = xyy^{-1}x = x^2$$

so that

$$\mathrm{e}^{2X(t)} = \mathrm{e}^{tZ}\, \mathrm{e}^{tW}. \tag{26}$$

As a matter of fact, it is possible to apply the algorithm proposed in [2] to generate the series $X(t) = \frac{1}{2}\log(\exp(tZ)\exp(tW))$ in an arbitrary generalized Hall basis of the free Lie algebra generated by P and K. Applying now Theorem 4.1 we conclude that the series $X(t)$ is convergent as long as $t(\|Z\| + \|W\|) < \pi$ or equivalently, when $0 \leq t < t_{bch}$, with

$$t_{bch} = \frac{\pi}{\|P + K\| + \|P - K\|}. \tag{27}$$

This estimate can be compared with (24) by taking into account that

$$\|P + K\| \leq p + k \leq 2\alpha, \qquad \|P - K\| \leq 2\alpha$$

where, as before, $\alpha = \max\{p, k\}$. In consequence, $t_{bch} > \pi/(4\alpha)$ and the convergence of $X(t)$ is guaranteed for

$$t \leq \frac{\pi}{4\alpha}.$$

Unfortunately, no bound for $\|X(t)\|$ in this domain can be obtained from the BCH series, contrarily to the estimate (23), valid when $t < t_c$.

Once $x = \exp(X(t))$ is known, one has $y = x^{-1}z$, or

$$\mathrm{e}^{Y} = \mathrm{e}^{-X(t)}\, \mathrm{e}^{tZ}.$$

Therefore, the series $Y(t)$ converges if $\|X(t)\| + t\|Z\| < \pi$. For $t < t_c$, we have shown that $\|X(t)\| < \pi/2$, so that one has convergence if $t\|Z\| < \pi/2$ or

$$t < \frac{\pi}{2\|P + K\|},$$

which is also greater than $\pi/(4\alpha)$. We then conclude that the generalized polar decomposition (5) exists with analytic functions $X(t)$ and $Y(t)$ at least for $t < t_c = \xi(1)/(2\alpha)$, whereas the series $X(t)$ is absolutely convergent for $0 \leq t < t_{bch}$.

Acknowledgements

This work has been partially supported by Ministerio de Ciencia e Innovación (Spain) under project MTM2007-61572 (co-financed by the ERDF of the European Union) and Fundació Bancaixa through project P1.1B2009-55. FC would like especially to thank the Instituto de Matemática Aplicada de la Universidad de Zaragoza (IUMA) for inviting him to participate in the workshop celebrating Manuel Calvo's birthday in September 2009.

References

[1] M. Abramowitz and I.A. Stegun. *Handbook of Mathematical Functions*. Dover, 1965.

[2] F. Casas and A. Murua. An efficient algorithm for computing the Baker–Campbell–Hausdorff series and some of its applications. *J. Math. Phys.* **50** (2009), 033513 (23 pages).

[3] K. Ebrahimi-Fard, J.M. Gracia-Bondía, and F. Patras. Rota–Baxter algebras and new combinatorial identities. *Lett. Math. Phys.* **81** (2007), 61–75.

[4] V.V. Gorbatsevich, A.L. Onishchik, and E.B. Vinberg. *Foundations of Lie Theory and Lie Transformation Groups*. Springer-Verlag, 1997.

[5] L. Guo. What is a Rota–Baxter algebra? *Notices of the AMS* **56** (2009), 1436–1437.

[6] E. Hairer, Ch. Lubich, and G. Wanner. *Geometric Numerical Integration. Structure-Preserving Algorithms for Ordinary Differential Equations*. Springer-Verlag, Second edition, 2006.

[7] S. Helgason. *Differential Geometry, Lie Groups, and Symmetric Spaces*. American Mathematical Society, 2001.

[8] N.J. Higham. Computing the polar decomposition. *SIAM J. Sci. Stat. Comput.* **7** (1986), 1160–1174.

[9] A. Iserles, H.Z. Munthe-Kaas, S.P. Nørsett, and A. Zanna. Lie-group methods. *Acta Numerica* **9** (2000), 215–365.

[10] S. Krogstad, H.Z. Munthe-Kaas, and A. Zanna. Generalized polar coordinates on Lie groups and numerical integrators. *Tech. Rep. 244, Department of Informatics, University of Bergen*, 2003.

[11] J.D. Lawson. Polar and Ol'shanskii decompositions. *J. Reine Angew. Math.* **448** (1994), 191–219.

[12] W. Magnus. On the exponential solution of differential equations for a linear operator. *Comm. Pure and Appl. Math.* **7** (1954), 649–673.

[13] H.Z. Munthe-Kaas, G.R.W. Quispel, and A. Zanna. Generalized polar decompositions on Lie groups with involutive automorphisms. *Found. Comput. Math.* **1** (2001), 297–324.

[14] V. S. Varadarajan. *Lie Groups, Lie Algebras, and Their Representations*. Springer-Verlag, 1984.

[15] A. Zanna. Recurrence relations and convergence theory of the generalized polar decomposition on Lie groups. *Math. Comp.* **73** (2004), 761–776.

[16] A. Zanna and H.Z. Munthe-Kaas. Generalized polar decompositions for the approximation of the matrix exponential. *SIAM J. Matrix Anal. Appl.* **23** (2001), 840–862.

Monografías de la Real Academia de Ciencias de Zaragoza **33**, 125–139, (2010).

Positivity properties for the classical fourth order Runge-Kutta method

I. Higueras

Departamento de Ingeniería Matemática e Informática

Universidad Pública de Navarra, 31006 Pamplona, Spain

Dedicated to Prof. Manuel Calvo, on the occasion of his 65th birthday

Abstract

Over the last few years a great effort has been done to develop Runge-Kutta (RK) methods that preserve properties such as monotonicity or contractivity for convex funtionals, or positivity. Provided that these properties hold for the explicit Euler scheme under certain stepsize restriction, it has been proved that these properties can also be maintained by some higher order RK methods under a modified stepsize. As this restriction includes the radius of absolute monotonicity of the RK scheme, strictly positive radius are required in order to obtain the desired properties with non trivial stepsizes. However, at least from the numerical positivity point of view, some authors have reported fairly good numerical results for some RK methods with zero radius, e.g. the classical fourth order four stages RK scheme. In this paper, we analyze this method and prove that, for some class of problems, it also preserves positivity. The study done strongly relies on the concept of region of absolute monotonicity for additive RK methods.

Keywords: Runge-Kutta, positivity, SSP, monotonicity, contractivity.

AMS subject classification: 65L06, 65L05, 65M20.

1 Introduction

Initial value problems for ordinary differential systems (ODEs)

$$\begin{aligned} \frac{d}{dt}u(t) &= f(t,u(t)) \qquad t \geq t_0 \\ u(t_0) &= u_0\,, \end{aligned} \tag{1}$$

arise directly in the modeling process of different phenomena, or after a method of lines approximation of evolutive partial differential equations. Quite often, the exact solution to (1) has certain property $\mathcal{P}$ (e.g., contractivity, monotonicity, positivity), usually with a physical meaning, which is relevant in the context where it appears. For example, we may have:

- Contractivity property: the solutions $u(t)$ and $\tilde{u}(t)$ satisfy

$$\|\tilde{u}(t) - u(t)\| \leq \|\tilde{u}(t_0) - u(t_0)\|\,, \qquad \text{for all } \; t \geq t_0\,, \tag{2}$$

  where $\|\cdot\|$ is a given convex function (norm, seminorm, entropy function, ...).

- Monotonicity property: the solution $u(t)$ satisfies

$$\|\tilde{u}(t)\| \leq \|u(t_0)\|\,, \qquad \text{for all } \; t \geq t_0\,, \tag{3}$$

  where again, $\|\cdot\|$ is a given convex function (norm, seminorm, entropy function, ...).

- Positivity property: if $u_0 \geq 0$, the solution $u(t)$ satisfies

$$u(t) \geq 0 \quad \text{for all} \quad t \geq t_0\,, \tag{4}$$

  where the inequalities should be understood component-wise.

In this situation, when the ODE (1) is solved numerically, it is natural to require the same qualitative property $\mathcal{P}$ to the numerical solution, $u_n \approx u(t_n)$. For this reason, when the exact solution to (1) satisfies whichever property (2)-(4), we will try to obtain, respectively,

$$\begin{aligned} &\|\tilde{u}_{n+1} - u_{n+1}\| \leq \|\tilde{u}_n - u_n\|\,, \\ &\|\tilde{u}_{n+1}\| \leq \|\tilde{u}_n\|\,, \\ &u_n \geq 0\,. \end{aligned} \tag{5}$$

Moreover, when a property $\mathcal{P}$ holds numerically, as the numerical solution depends on the stepsize h, a natural question is whether it holds for all step sizes $h > 0$, or it only holds under a step size restriction of the form $h \leq H$.

As a rule, when these issues are studied, there are four crucial aspects to consider:

i) How property $\mathcal{P}$ is obtained for the continuous problem (1).

ii) The class of problems $\mathcal{C}$ considered (e.g., linear, non linear).

iii) The type of function (*"norms"*) involved in property $\mathcal{P}$ (general convex functions, arbitrary norms, inner product norms, ...).

iv) The class of numerical methods used (Runge-Kutta, multistep, implicit or explicit schemes,...).

Depending on these aspects, different results can be obtained. Obviously, the most general conditions on problems, methods, norms,... will lead to more restricted results, whereas with more stringent conditions, sharper results will be obtained. Because of this, in order to get optimal results, it is important to analyze and determine the class of problems, the used *"norms"* and the methods we are dealing with.

Once the theoretical stepsize restrictions have been attained, it is mandatory to check their sharpness with numerical experiments on concrete problems. Although sometimes it is possible to construct a problem where the predicted and observed stepsize bounds fit, very often, for a wide class of problems, there is a great discrepancy between the effective stepsize restrictions and theoretical ones. This situation arises for example in the context of numerical positivity, studied for Runge-Kutta methods e.g. in [10, 11]. In this setting, some authors [12, 11] have reported good numerical results for schemes such that, according to the theory developed, they are not good. This the case for a widely used Runge-Kutta scheme: the fourth order four stages method (RK4). For this scheme, the radius of absolute monotonicity is trivial and therefore numerical positivity cannot be ensured. However, for many problems, RK4 method give fair good results.

In order to explain the favorable results observed, one should consider the possibility that the set of problems used to test the desired qualitative behavior belongs to a subclass $\tilde{\mathcal{C}}$ of the problems considered, $\tilde{\mathcal{C}} \subset \mathcal{C}$, and that the method used performs well on this class $\tilde{\mathcal{C}}$. This idea is not new and in the context of positivity, one could say that it is contained in the approach followed in [11] for linear and quasilinear problems; in fact, the reduction of initial values to a set of positive vectors done in [11] can be considered as a restriction of the class of problems.

In this paper we consider RK4 scheme and we will try to explain why it gives good results for certain class of problems. The approach followed differers from the one done in [11] in the sense that we do not impose any restriction on initial values but on the class of problems itself. On the other hand, we deal with non linear problems. The study done here strongly relies on the concept of region of absolute monotonicity for additive RK methods.

The rest of the paper is organized as follows. In sections 2 and 3 we introduce the methods used and we review the most relevant definitions and results concerning numerical

monotonicity. A simple example showing how RK4 scheme performs is given in section 4. A theoretical framework to explain this behavior is given in section 5. Next, these results are used for the example in section 4. The paper ends with some conclusions and forthcoming work.

2 Runge-Kutta and additive Runge-Kutta methods

A common class of one step methods to solve numerically (1) are the Runge-Kutta (RK) methods. An s-stages RK method is defined by an $s \times s$ real matrix $\mathcal{A}$ and a real vector $b \in \mathbb{R}^s$. From u_n, the numerical approximation of the solution $u(t)$ at $t = t_n$, we obtain u_{n+1}, the numerical approximation of the solution at $t_{n+1} = t_n + h$ from

$$u_{n+1} = u_n + \sum_{i=1}^{s} b_i \, f(t_n + c_i h, U_{ni}) \,, \tag{6}$$

where

$$U_{ni} = u_n + h \sum_{j=1}^{s} a_{ij} \, f(t_n + c_j h, U_{nj}) \,. \tag{7}$$

If the matrix $\mathcal{A}$ is strictly lower triangular, the method is explicit, otherwise the method is implicit. For nonlinear problems, implicit methods require the resolution of nonlinear systems of dimension $s \cdot m$, with m the dimension of the ODE system (1). Denoting the coefficients of the RK method by

$$\mathbb{A} = \begin{pmatrix} \mathcal{A} & 0 \\ b^t & 0 \end{pmatrix},$$

we can write (6)-(7) in compact form as

$$U = e \otimes u_n + h \, (\mathbb{A} \otimes I) \, F(U) \,, \tag{8}$$

where we have denoted by $e = (1, \ldots, 1)^t \in \mathbb{R}^{s+1}$, $U = (U_1^t, \ldots, U_s^t, u_{n+1}^t)^t \in \mathbb{R}^{(s+1)m}$, $F(U) = (f(U_1)^t, \ldots, f(U_s)^t, 0)^t \in \mathbb{R}^{(s+1)m}$, and similarly $\tilde{F}(U)$. The symbol $\otimes$ denotes the Kronecker product (see e.g. [2, Section 12.1])

$$A \otimes B = \begin{pmatrix} a_{11}B & \cdots & a_{1m}B \\ \vdots & & \vdots \\ a_{m1}B & \cdots & a_{mm}B \end{pmatrix}.$$

Explicit schemes are easy to implement but they are not adequate to solve stiff ODEs because they require small stepsizes; on the other hand, many implicit schemes do not suffer from these stepsize restrictions, but with them, one has to deal with the numerical

resolution (difficult sometimes) of nonlinear systems. However, many times, stiffness is only associated with a part of the problem; that is to say, the ODE can be written as

$$\frac{d}{dt}u(t) \;=\; f(t,u(t)) + \tilde{f}(t,u(t)) \qquad t \geq t_0 \tag{9}$$

where f contains the non stiff terms and $\tilde{f}$ contains the stiff ones. In this case, we can use IMplicit-EXplicit (IMEX) RK methods, where an explicit method is used for the non stiff terms, and an implicit one for the stiff part. In compact form, IMEX RK methods with coefficients $(\mathbb{A}, \tilde{\mathbb{A}})$, where $\mathbb{A}$ denotes the explicit method and $\tilde{\mathbb{A}}$ the implicit one, are given by

$$U = e \otimes u_n + h\,(\mathbb{A} \otimes I)\,F(U) + h\,(\tilde{\mathbb{A}} \otimes I)\,\tilde{F}(U)\,. \tag{10}$$

3 Monotonicity preserving methods (or Strong Stability Preserving methods)

Over the last years, a big effort has been done to develop methods such that monotonicity (contractivity, positivity) is preserved numerically. For RK methods, it is proven that these properties can be ensured under a stepsize restriction of the form

$$\Delta t \leq \tau_0 \cdot \mathcal{R}(\mathbb{A})\,. \tag{11}$$

In (11), τ_0 is a problem dependent parameter and $\mathcal{R}(\mathbb{A})$ is a method dependent parameter. To be more precise, τ_0 is a constant that ensures property $\mathcal{P}$ for the explicit Euler method, $u_{n+1} = u_n + \tau\, f(u_n)$, whenever $0 \leq \tau \leq \tau_0$, that is to say,

$$\begin{aligned}
&\|u_n + \tau\, f(u_n)\| \leq \|u_n\|\,, && \text{(monotonicity)}\\
&\|u_n - v_n + \tau\,(f(u_n) - f(v_n))\| \leq \|u_n - v_n\|\,, && \text{(contractivity)}\\
&u_n \geq 0 \quad \Longrightarrow \quad u_{n+1} = u_n + \tau\, f(u_n) \geq 0\,, && \text{(positivity)}
\end{aligned} \tag{12}$$

and $\mathcal{R}(\mathbb{A})$ is the radius of absolute monotonicity defined as follows.

Definition 3.1 *[13, Definition 2.4] An s-stage RK method with coefficients $\mathbb{A}$ is said to be absolutely monotonic at a given point $\xi \leq 0$ if $I - \xi\mathbb{A}$ is non singular, and*

$$(I - \xi\mathbb{A})^{-1}\mathbb{A} \geq 0\,, \qquad\qquad (I - \xi\mathbb{A})^{-1}e \geq 0\,, \tag{13}$$

where $e = (1, 1, \ldots, 1)^t \in \mathbb{R}^{s+1}$, and the vector inequalities are understood componentwise. Further, the method is said to be absolutely monotonic on a given set $\Omega \subset \mathbb{R}$ if it is absolutely monotonic at each $\xi \in \Omega$. The radius of absolute monotonicity $\mathcal{R}(\mathbb{A}))$ is defined by

$$\mathcal{R}(\mathbb{A}) = \sup\{\, r \,|\, r \geq 0 \text{ and } \mathbb{A} \text{ is absolutely monotonic on } [-r, 0]\,\}\,.$$

If there is no $r > 0$ such that $\mathbb{A}$ is absolutely monotonic on $[-r, 0]$, we set $\mathcal{R}(\mathbb{A}) = 0$.

As a result, if $\mathcal{R}(\mathbb{A}) = 0$, from (11) we obtain a trivial stepsize restriction. At this point we should remark that, as proven in [3], the stepsize restriction for monotonicity (11) for the class of problems (12) is optimal.

Observe that conditions (13) are

$$\begin{pmatrix} (I-\xi\mathcal{A})^{-1} & 0 \\ \xi\, b^t(I-\xi\mathcal{A})^{-1} & 1 \end{pmatrix} \begin{pmatrix} e \\ 1 \end{pmatrix} \geq 0\,, \qquad \begin{pmatrix} (I-\xi\mathcal{A})^{-1} & 0 \\ \xi\, b^t(I-\xi\mathcal{A})^{-1} & 1 \end{pmatrix} \begin{pmatrix} \mathcal{A} & 0 \\ b^t & 0 \end{pmatrix} \geq 0\,.$$

and hence absolute monotonicity at a given point ξ is equivalent to the following sign conditions:

$$\begin{aligned} \phi(\xi) &= 1 + \xi\, b^t(I-\xi\,\mathcal{A})^{-1}e \geq 0\,, \\ \mathcal{A}(\xi) &= \mathcal{A}(I-\xi\mathcal{A})^{-1} \geq 0\,, \\ b(\xi)^t &= b^t(I-\xi\mathcal{A})^{-1} \geq 0\,, \\ e(\xi) &= (I-\xi\mathcal{A})^{-1}e \geq 0\,, \end{aligned}$$

where now $e = (1, 1, \ldots, 1) \in \mathbb{R}^s$. Observe that $\phi(\xi)$ is the the stability function of the RK method.

For additive RK methods (10), the concept of radius of absolute monotonicity is extended to the region of absolute monotonicity.

Definition 3.2 *[9, Definition 2.3] An s-stage additive RK method $(\mathbb{A}, \tilde{\mathbb{A}})$ is said to be absolutely monotonic (a.m.) at a given point $(\xi, \tilde{\xi})$ with $\xi, \tilde{\xi} \leq 0$ if the matrix $I - \xi\mathbb{A} - \tilde{\xi}\tilde{\mathbb{A}}$ is invertible and*

$$\mathsf{A}(\xi, \tilde{\xi}) = (I - \xi\mathbb{A} - \tilde{\xi}\tilde{\mathbb{A}})^{-1}\mathbb{A} \geq 0\,, \tag{14}$$

$$\tilde{\mathsf{A}}(\xi, \tilde{\xi}) = (I - \xi\mathbb{A} - \tilde{\xi}\tilde{\mathbb{A}})^{-1}\tilde{\mathbb{A}} \geq 0\,, \tag{15}$$

$$\mathsf{e}(\xi, \tilde{\xi}) = (I - \xi\mathbb{A} - \tilde{\xi}\tilde{\mathbb{A}})^{-1}e \geq 0\,. \tag{16}$$

Further, the additive method is said to be absolutely monotonic on a given set $\Omega \in \mathbb{R}^2$ if it is absolutely monotonic at each $(\xi, \tilde{\xi}) \in \Omega$.

Observe that for RK we work in $\mathbb{R}$ but additive RK methods we have to work in $\mathbb{R}^2$. For this reason we define the region and the curve of absolute monotonicity as follows.

Definition 3.3 *[9, Definition 2.4] The region of absolute monotonicity, denoted by $\mathcal{R}(\mathbb{A}, \tilde{\mathbb{A}})$, is defined by*

$$\mathcal{R}(\mathbb{A}, \tilde{\mathbb{A}}) = \{\, (r, \tilde{r}) \,|\, r \geq 0\,, \tilde{r} \geq 0 \text{ and } (\mathbb{A}, \tilde{\mathbb{A}}) \text{ is a.m. on } [-r, 0] \times [-\tilde{r}, 0] \,\}\,.$$

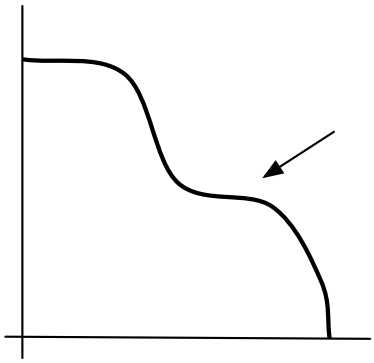


Figure 1.— Curve of absolute monotonicity

The curve of absolute monotonicity, denoted by $\partial\mathcal{R}(\mathbb{A},\tilde{\mathbb{A}})$*, is the frontier of the set* $\mathcal{R}(\mathbb{A},\tilde{\mathbb{A}})$ *excluding the coordinate axis (see figure 1). If there is no* $r>0$*,* $\tilde{r}>0$ *such that* $(\mathbb{A},\tilde{\mathbb{A}})$ *is absolutely monotonic on* $[-r,0]\times[-\tilde{r},0]$*, we set* $\partial\mathcal{R}(\mathbb{A},\tilde{\mathbb{A}})=(0,0)$*.*

For additive RK methods, it is assumed that numerical monotonicity holds when explicit Euler method is used for both functions f, $\tilde{f}$, i.e. there exists some fixed $\tau_0,\tilde{\tau}_0>0$ such that

$$\|u_n+\tau\,f(u_n)\|\le\|u_n\|\,,\qquad\qquad\|u_n+\tilde{\tau}\,\tilde{f}(u_n)\|\le\|u_n\|\,. \tag{17}$$

Under these assumptions, numerical monotonicity can be ensured for the additive RK method $(\mathbb{A},\tilde{\mathbb{A}})$ under the stepsize restriction

$$h\le\min\{r\,\tau_0,\,\tilde{r}\,\tilde{\tau}_0\}\,, \tag{18}$$

where r and $\tilde{r}$ are such that the point $(r,\tilde{r})\in\mathcal{R}(\mathbb{A},\tilde{\mathbb{A}})$ (see [9] for details). As it is proven in [19], stepsize restriction (18) is optimal for the class of problems (17).

Monotonicity properties of numerical schemes have also been deeply studied in the context of hyperbolic systems of conservation laws. In this setting, monotone schemes for the Total Variation (TV) seminorm are known as Total Variation Diminishing (TVD) or Strong Stability preserving methods (SSP). The class of ODEs considered in this context arise from a method of lines approximation of this class of partial differential equations, and a simple numerical example given in [5], shows that the use of non-SSP methods for the time discretization of these ODEs has the potential to produce an undesirable overshoot.

In the seminal paper [16], Shu & Osher consider SSP (or TVD) spatial discretizations such that

$$\|u_n+h\,f(u_n)\|_{TV}\le\|u_n\|_{TV}\,,\qquad h\le\Delta t_{FE}\,.$$

However, as the forward Euler method has the drawback of its low order of accuracy, higher order SSP methods are of great interest, and over the last few years a great effort

has been done to develop high order SSP methods ([16, 17, 6, 14, 15, 20], see [5, 18, 7] for reviews on this topic). It is important to point out that explicit RK methods in [16] are not written in the standard form (6)-(7), but as

$$\begin{aligned} u^{(1)} &= u_n \\ u^{(i)} &= \sum_{k=1}^{i-1} \left(\alpha_{ik} u^{(k)} + h\, \beta_{ik} f\left(u^{(k)}\right) \right) , \qquad i = 2, \ldots, s+1 \\ u_{n+1} &= u^{(s+1)} \end{aligned} \tag{19}$$

where $\alpha_{ik} \geq 0$ for all i, j, and $\sum_{k=1}^{i-1} \alpha_{ik} = 1$, $i = 2, \ldots, s+1$. It is also imposed that

$$\beta_{i,j} = 0 \qquad \text{whenever} \qquad \alpha_{ij} = 0\,. \tag{20}$$

It is straightforward to check that, if $\beta_{ij} \geq 0$, convex combinations of the forward Euler method are obtained in (19). In this case, the new method will also be strongly stable, with a modified step size restriction

$$h \leq c\, \Delta t_{FE}\,,$$

where the CFL coefficient c is given by

$$c = \min_{ik} \frac{\alpha_{ik}}{\beta_{ik}}\,. \tag{21}$$

Given a RK method in the Shu & Osher representation (19), if we denote by $\Lambda = (\alpha_{ij})$, $\Gamma = (\beta_{ij})$, it is not difficult to see that the Butcher matrix of the RK scheme is given by $\mathbb{A} = (I - \Lambda)^{-1}\Gamma$; with this notation, the sign conditions on α_{ij}, β_{ij} imply that $\Lambda \geq 0$, $\Gamma \geq 0$, the CFL coefficient in (21) is given by

$$\Lambda - c\,\Gamma \geq 0\,, \tag{22}$$

and condition (20) trivially follows from (22).

However, as many authors have pointed out, given a RK method $\mathbb{A}$, its representation Λ, Γ is not unique. For this reason, as the CFL coefficient (21) depends on the representation available, a problem of great interest and deeply studied in the SSP community has been how to obtain optimal representations. This problem was solved in [4, 8] where the connection between optimal Shu & Osher representations (19) and the radius of absolute monotonicity $\mathcal{R}(\mathbb{A})$ is given.

In the Shu & Osher representation, RK methods with $\mathcal{R}(\mathbb{A}) = 0$ require negative coefficients β_{ij}. This is the case the classical fourth order four stage RK method, whose Butcher coefficients are given by

$$
\begin{array}{c|cccc}
0 & 0 & & & \\
\frac{1}{2} & \frac{1}{2} & 0 & & \\
\frac{1}{2} & 0 & \frac{1}{2} & 0 & \\
1 & 0 & 0 & 1 & 0 \\
\hline
 & \frac{1}{6} & \frac{1}{3} & \frac{1}{3} & \frac{1}{6}
\end{array}
$$

In [16] when a negative value β_{ij} is required, function f is replaced by an associated operator $\tilde{f}$ corresponding to stepping backward in time. It is assumed that $\tilde{f}$ approximates the same spatial derivatives as f and

$$\|u_n - h\,\tilde{f}(u_n)\| \le \|u_n\|\,, \qquad \text{for } h \le \Delta t_{FE}\,, \tag{23}$$

where the stepsize restriction Δt_{FE} is the stepsize restriction needed to obtain monotonicity when the explicit Euler scheme is used for f in forward time,

$$\|u_n + h\,f(u_n)\| \le \|u_n\|\,, \qquad \text{for } h \le \Delta t_{FE}\,. \tag{24}$$

When negative values are required, the CFL coefficient with the Shu & Osher representations is computed from

$$c = \min_{ik} \frac{\alpha_{ik}}{|\,\beta_{ik}\,|}\,. \tag{25}$$

See [16] for details.

For example, the four stages RK scheme is written in [16] as

$$
\begin{aligned}
u^{(1)} &= u^{(0)} \\
u^{(2)} &= u^{(1)} + \tfrac{1}{2}hf(u^{(1)}) \\
u^{(3)} &= \tfrac{1}{2}u^{(1)} - \tfrac{1}{4}h\tilde{f}(u^{(1)}) + \tfrac{1}{2}u^{(2)} + \tfrac{1}{2}hf(u^{(2)}) \\
u^{(4)} &= \tfrac{6431}{80000}u^{(1)} - \tfrac{18769}{160000}h\tilde{f}(u^{(1)}) + \tfrac{18769}{80000}u^{(2)} - \tfrac{137}{400}h\tilde{f}(u^{(2)}) + \tfrac{137}{200}u^{(3)} + hf(u^{(3)}) \\
u^{(5)} &= \tfrac{1}{3}u^{(2)} + \tfrac{1}{6}hf(u^{(2)}) + \tfrac{1}{3}u^{(3)} + \tfrac{1}{3}u^{(4)} + \tfrac{1}{6}hf(u^{(4)})
\end{aligned}
\tag{26}
$$

In [8], Shu & Osher representations with negative coefficients were interpreted as perturbations of the original RK method $\mathbb{A}$ with a perturbation matrix $\tilde{\mathbb{A}}$. More precisely,

$$U = e \otimes u_n + h(\mathbb{A} \otimes I)F(U) + h(\tilde{\mathbb{A}} \otimes I)\left(F(U) - \tilde{F}(U)\right)\,. \tag{27}$$

We can separate the terms in f and $\tilde{f}$ and consider scheme (27) in additive form,

$$U = e \otimes u_n + h((\mathbb{A} + \tilde{\mathbb{A}}) \otimes I)F(U) - h(\tilde{\mathbb{A}} \otimes I)\tilde{F}(U)\,.$$

Observe that in this case we are assuming (23)-(24), and hence we have that $\tau_0 = \tilde{\tau}_0 = \Delta t_{FE}$. Applying the results for additive RK methods [9], we obtain monotonicity under the stepsize restriction (see (18))

$$h \leq r\, \Delta t_{FE}\,,$$

where r is such that $(r, r) \in \mathcal{R}(\mathbb{A} + \tilde{\mathbb{A}}, \tilde{\mathbb{A}})$.

In particular, for scheme (26), after some manipulations, we obtain that it is of the form (27) with coefficient matrices $(\mathbb{A}, \tilde{\mathbb{A}})$ given by

$$\mathbb{A} = \begin{pmatrix} 0 & & & & \\ \frac{1}{2} & 0 & & & \\ 0 & \frac{1}{2} & 0 & & \\ 0 & 0 & 1 & 0 & \\ \frac{1}{6} & \frac{1}{3} & \frac{1}{3} & \frac{1}{6} & 0 \end{pmatrix}, \qquad \tilde{\mathbb{A}} = \begin{pmatrix} 0 & & & & \\ 0 & 0 & & & \\ 1/4 & 0 & 0 & & \\ \frac{46169}{160000} & \frac{137}{400} & 0 & 0 & \\ \frac{28723}{160000} & \frac{137}{1200} & 0 & 0 & 0 \end{pmatrix}. \tag{28}$$

With the notation of additive RK method, the point $(0.685, 0.685) \in \mathcal{R}(\mathbb{A} + \tilde{\mathbb{A}}, \tilde{\mathbb{A}})$, and hence, for the perturbed RK method the CFL coefficient is 0.685.

The above examples show the potential of the study done in [9] for additive RK methods, and how we can transfer these results to different kind of schemes whenever they can be formally reinterpreted as additive RK methods. As we will see later on, these ideas can be used to explain why some non-SSP methods may perform well on certain classes of problems.

4 A simple example

As it has been pointed out above, the classical fourth order four stages RK scheme has $\mathcal{R}(\mathbb{A}) = 0$, and therefore monotonicity (or contractivity, positivity) cannot be ensured for the class of problems satisfying (12). However, in the context of positivity good results have been reported for this method [12]. In fact, it is not difficult to construct simple academic examples that give numerical positivity under nontrivial stepsizes.

Example 4.1 We consider the problem

$$y'(t) = y(t)\,(y(t) - 1)\,, \qquad y'(t_0) = y_0\,,$$

whose solution satisfies $y(t) \in [0, 1]$ whenever $y_0 \in [0, 1]$. It is easy to check that if $0 \leq y \leq 1$, we obtain that

$$0 \leq y + \tau\, y\,(y - 1) \leq 1 \qquad \text{for all } \; 0 \leq \tau \leq 1\,,$$

and hence this problem satisfies property (12) for $\tau_0 = 1$. Numerical positivity can be ensured for RK methods with coefficient matrix $\mathbb{A}$ under the stepsize restriction

$$h \leq \mathcal{R}(\mathbb{A})\,.$$

However, for RK4 scheme, after some computations, we obtain that $y_n \in [0,1]$ gives $y_{n+1} \in [0,1]$ under the non-trivial stepsize restriction $h \leq 1.2956$. □

In the next sections we give an explanation of this fact.

5 Main results

Given a RK method with coefficient matrix $\mathbb{A}$, we can formally reinterpret it as an additive RK method by splitting the coefficient matrix $\mathbb{A}$. For example, if we split $\mathbb{A}$ as $\mathbb{A} = \mathbb{A}_+ - \mathbb{A}_-$, with $\mathbb{A}$, $\tilde{\mathbb{A}} \geq 0$, the numerical scheme (8) can be written as

$$U = e \otimes u_n + h\,(\mathbb{A}_+ \otimes I)\,F(U) - h\,(\mathbb{A}_- \otimes I)\,F(U)\,, \tag{29}$$

that can be interpreted as an additive RK scheme with coefficients $\mathbb{A} = \mathbb{A}_+$, $\tilde{\mathbb{A}} = \mathbb{A}_-$ applied to the functions f and $-f$. In this case, following the ideas used in [9], we can rewrite the original RK method as

$$U = \mathsf{e}(-r,-\tilde{r}) \otimes u_n + (r\,\mathsf{A}(-r,-\tilde{r}) \otimes I)\left(U + \frac{h}{r}\,F(U)\right) + \left(\tilde{r}\,\tilde{\mathsf{A}}(-r,-\tilde{r}) \otimes I\right)\left(U - \frac{h}{\tilde{r}}\,F(U)\right)\,, \tag{30}$$

where $\mathsf{e}(-r,\tilde{r})$, $\mathsf{A}(-r,-\tilde{r})$ and $\tilde{\mathsf{A}}(-r,-\tilde{r})$ are given by (14)-(16), and $r, \tilde{r}$ are such that the matrix $I + r\,\mathbb{A} + \tilde{r}\,\tilde{\mathbb{A}}$ is invertible.

If the sign conditions (14)-(16) hold for $\mathsf{e}(-r,\tilde{r})$, $\mathsf{A}(-r,\tilde{r})$ and $\tilde{\mathsf{A}}(-r,\tilde{r})$, expression (30) is simply a convex combination of forward and backward Euler steps. Hence, imposing property $\mathcal{P}$ for explicit Euler steps for f and $-f$ with coefficients τ_+, τ_- respectively, we can obtain preservation of property $\mathcal{P}$ under a stepsize restriction of the form (18),

$$h \leq \min\left\{r\,\tau_+,\, \tilde{r}\,\tau_-\right\}\,, \tag{31}$$

where r and $\tilde{r}$ are such that the point $(r,\tilde{r}) \in \mathcal{R}(\mathbb{A}_+, \mathbb{A}_-)$.

From (31), the minimum value is obtained when $r\,\tau_+ = \tilde{r}\,\tau_-$. Hence, we can take $\tilde{r} = r\,\tau_+/\tau_-$ and compute the largest value r such that

$$\left(r,\, r\,\frac{\tau_+}{\tau_-}\right) \in \mathcal{R}(\mathbb{A}_+, \mathbb{A}_-)\,.$$

Proceeding in this way, (31) is $h \leq r\,\tau_+$. Observe that r depends on τ_+/τ_-, and hence, if we denote by $y = \tau_+/\tau_-$, we obtain the stepsize restriction

$$h \leq r(y)\,\tau_+\,. \tag{32}$$

We have finished if for each $y = \tau_+/\tau_-$ we are able to compute a value $r(y)$ and a splitting $\mathbb{A}_+$, $\mathbb{A}_-$, such that the point $(r(y),\, r(y)\, y)$ belongs to the region of absolute monotonicity of $(\mathbb{A}_+, \mathbb{A}_-)$, i.e.

$$(r(y),\, r(y)\, y) \in \mathcal{R}(\mathbb{A}_+, \mathbb{A}_-)\,.$$

Furthermore, by (32), we are interested in the largest value $r(y)$.

Observe that we have not modified the original scheme and hence, if the problem function f has the desired property for explicit Euler method not only in forward time but also in backward time, and the numerical method has splittings that lead to conditions (14)-(16), we can observe good results for larger stepsize restrictions.

With the notation of section 1, in the above analysis we are not considering the class

$$\mathcal{C} = \{ \text{ problems with property } \mathcal{P} \text{ for Euler steps in forward time } \}$$

but the subclass

$$\tilde{\mathcal{C}} = \{ \text{ problems with property } \mathcal{P} \text{ for Euler steps in forward and backward time } \}\,.$$

This idea can also been used for implicit-explicit Runge-Kutta methods [1].

We finish this section pointing out that this way of proceeding is closely related to the one followed by Shu & Osher in [16] when negative coefficients β_{ij} are required. The difference is that with our approach (29), due to the good properties of f, we do not need to use a different operator $\tilde{f}$.

6 A simple example (revisited)

We can now explain the good results obtained in section 4. The first step is to check if we have property $\mathcal{P}$ for $-f$. In this case, it is easy to check that

$$0 \le y - \tau\, y\, (y-1) \le 1 \qquad \text{for all } \; 0 \le \tau \le 1\,,$$

and hence we obtain $\tau_- = 1$.

The next step is to study the used scheme, RK4 in this case. Using a numerical optimization method, we have obtained for each y the largest value $r(y)$ such that there is a splitting $\mathbb{A}_+$, $\mathbb{A}_-$, with

$$(r(y),\, r(y)\, y) \in \mathcal{R}(\mathbb{A}_+, \mathbb{A}_-)\,.$$

In this process we have used the results in (see [8, p. 939]) that establish the compulsory nonzero elements in matrix $\mathbb{A}_-$ to obtain non trivial regions $\mathcal{R}(\mathbb{A}_+, \mathbb{A}_-)$, namely, the elements a_{31}, a_{41}, a_{51}, a_{42}, a_{52}. The values obtained are shown in figure 2.

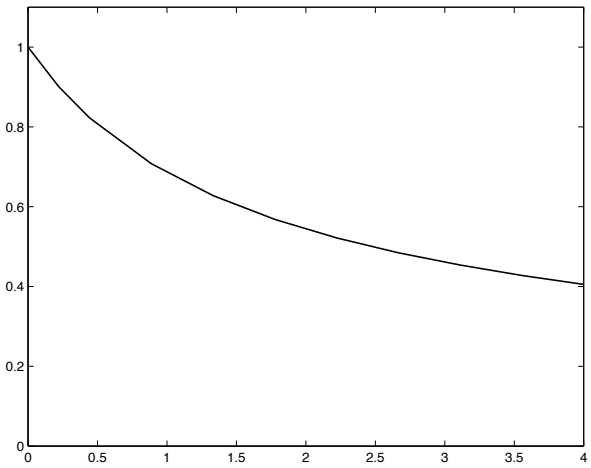


Figure 2.— Function $r(y)$ in (23), with $y = \tau_+/\tau_-$

In particular, we get $r(1) = 0.685$, and hence we obtain property $\mathcal{P}$ under the non-trivial stepsize restriction $h \leq 0.685$ (see (32)). Although this bound is not sharp for this problem, it is better than the trivial one obtained from $\mathcal{R}(\mathbb{A})$ (see (11)).

Observe that we have the same CFL coefficient obtained for the Shu & Osher representation (26). The splitting of the matrix $\mathbb{A}$ obtained for $y = 1$ is

$$\mathbb{A}_+ = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 \\ 0.5 & 0 & 0 & 0 & 0 \\ 0.2142 & 0.5 & 0 & 0 & 0 \\ 0.2640 & 0.3425 & 1. & 0 & 0 \\ 0.2295 & 0.3727 & 0.3333 & 0.1667 & 0 \end{pmatrix}, \qquad \mathbb{A}_- = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0.2142 & 0 & 0 & 0 & 0 \\ 0.2640 & 0.3425 & 0 & 0 & 0 \\ 0.0628 & 0.0394 & 0 & 0 & 0 \end{pmatrix}.$$

If we compare these matrices with the ones obtained with the perturbed method (28),

$$\mathbb{A}+\tilde{\mathbb{A}} = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 \\ 0.5 & 0 & 0 & 0 & 0 \\ 0.25 & 0.5 & 0 & 0 & 0 \\ 0.2886 & 0.3425 & 1. & 0 & 0 \\ 0.3462 & 0.4475 & 0.3333 & 0.1667 & 0. \end{pmatrix}, \qquad \tilde{\mathbb{A}} = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0.25 & 0 & 0 & 0 & 0 \\ 0.2886 & 0.3425 & 0 & 0 & 0 \\ 0.1795 & 0.1142 & 0 & 0 & 0 \end{pmatrix}.$$

we observe that they are different but some values are exactly the same. This fact shows that there is not uniqueness of splittings/perturbations for a given method to achieve the maximum stepsize restriction.

We should point out that for $y = 0$ we obtain $r(0) = 1$. However, as the value $y = 0$ corresponds to $\tau_+ = 0$, the stepsize restriction in (23) is trivial.

Finally, we would like to remark that for our problem we have $\tau_+ = \tau_- = 1$, but the analysis done is valid for other values. We simply have to compute the ratio $y = \tau_+/\tau_-$, and the corresponding value $r(y)$ to obtain the stepsize restricion (23).

7 Conclussions and forthcoming work

In this paper we have studied why, in the context of positivity, RK4 scheme gives good results for some problems. The results obtained strongly rely on the concept of region of absolute monotonicity for additive RK methods.

Although we have focused on positivity, the analysis done is valid for other properties $\mathcal{P}$. The basic requirement for the function problem f is the fulfillment of property $\mathcal{P}$ for explicit Euler method in forward and backward time.

We have centered on RK4 scheme, but the study can be also done for some other well known methods, both implicit and explicit.

Acknowledgements

The author acknowledge support from project MTM2008-00785

References

[1] R. Donat, I. Higueras, A. Martinez-Gavara, On stability issues for IMEX schemes applied to hyperbolic equations with stiff reaction terms. *Submitted*

[2] P. Lancaster, M. Tismenetsky, The theory of matrices, Academic Press, San Diego, CA, 1985.

[3] L. Ferracina and M.N. Spijker, Stepsize restrictions for the total-variation-diminishing property in general Runge-Kutta methods, SIAM J. Numer. Anal., 42 (2004), pp. 1073–1093.

[4] L. Ferracina and M.N. Spijker, An extension and analysis of the Shu-Osher representation of Runge-Kutta methods, Math. Comp., 74 (2005), pp. 201–219.

[5] S. Gottlieb, C.W. Shu, and E. Tadmor, Strong stability-preserving high order time discretization methods, SIAM Rev. 43 (2001), 89–112.

[6] S. Gottlieb, C.W. Shu, Total variation diminishing Runge-Kutta schemes, Math. Comp. 67 (1998), 73-85.

[7] I. Higueras, On strong stability preserving time discretization methods, J. Sci. Comput. 21 (2004), no. 2, 193–223.

[8] I. Higueras, Representations of Runge–Kutta methods and strong stability preserving methods, SIAM J. Numer. Anal. 43 (2005), no. 3, 924–948.

[9] I. Higueras, Strong stability for additive Runge-Kutta methods, SIAM J. Numer. Anal. 44 (2006), no. 4, 1735–1758.

[10] Z. Horváth, Positivity of Runge-Kutta and diagonally split Runge-Kutta methods , Appl. Numer. Math. 28 (1998), 309–326.

[11] Z. Horváth, On the positivity stepsize threshold of Runge-Kutta methods, Appl. Numer. Math. 53 (2005), 341–356.

[12] W. Hundsdorfer, B. Koren, M. van Loon, J.G. Verwer, A positive finite-difference advection scheme, J. Comput. Phys. 117 (1995) 3–46.

[13] J.F.B.M. Kraaijevanger, Contractivity of Runge-Kutta methods, BIT 31 (1991), 482–528.

[14] S.J. Ruuth and R.J. Spiteri, Two barriers on strong stability preserving time discretization methods, J. Sci. Comput., 17(2002), 211–220.

[15] S.J. Ruuth and R.J. Spiteri, High-order strong-stability-preserving Runge-Kutta methods with downwind-biased spatial discretizations, SIAM J. Numer. Anal. 42(2004), 974–996.

[16] C.W. Shu and S. Osher, Efficient implementation of essentially non-oscillatory shock-capturing schemes, J. of Comput. Phys. 77(1988), 439–471.

[17] C.W. Shu, Total variation diminishing time discretizations, SIAM J. Sci. Comput. 9(1988), 1073–1084.

[18] C.W. Shu, A survey of strong stability preserving high order time discretizations, Collected lectures on the preservation of stability under discretization. D. Estep and T. Tavener Editors. Proceedings in Applied Mathematics 109, SIAM 2002, 51–65.

[19] M.N. Spijker, Stepsize conditions for general monotonicity in numerical initial value problems, SIAM Journal on Numerical Analysis, 45 (2007), 1226–1245.

[20] R.J. Spiteri and S.J. Ruuth, A new class of optimal high order strong stability preserving time discretization methods, SIAM J. Numer. Anal., 40(2002), 469–491.

Monografías de la Real Academia de Ciencias de Zaragoza **33**, 141–154, (2010).

Extending convergence results of Runge–Kutta methods for stiff semi linear initial value problems

M. Calvo[1], S. Gonzalez-Pinto[2] and J.I. Montijano[1]

[1] Departamento Matemática Aplicada
Universidad de Zaragoza, 50009-Zaragoza, Spain

[2] Departamento Análisis Matemático
Universidad de La Laguna, 38271-La Laguna, Spain

Abstract

In this paper some convergence results for Runge–Kutta methods applied to semi–linear variable coefficients differential systems with the stiffness contained in the linear part and under some assumptions on the relative variation of the jacobian matrix are derived. Previous results on this subject given by the authors in BIT 40, 4 (2000), pp. 611–634, are generalised. In particular, it is shown that some non B–stable methods such as those of the Lobatto IIIA family and some DIRK methods that have been used in practical problems are convergent of order greater or equal than the stage order for this kind of problems. Some numerical examples are presented to illustrate the theory.

1 Introduction

The concepts of B–stability and B–convergence introduced to study the behaviour of Runge–Kutta (RK) methods for the numerical integration of stiff IVPs

$$y'(t) = f(t, y(t)),\ y(t_0) = y_0 \in \mathbb{R}^m,\ t \in I := [t_0, t_0 + T], \tag{1}$$

where $f(t, y)$ satisfies a one-sided Lipschitz condition with respect to y, have provided in the last two decades a well established B–theory that allows us to identify those RK methods that are suitable for this class of stiff systems [8], [10].

However, as remarked by Alexander [1] there exist some globally well behaved stiff problems (i.e. IVPs whose solution is asymptotically stable) that possess strongly positive one sided Lipschitz constants and therefore do not fit into the B–theory. In particular,

variable coefficient linear or semi linear problems with the stiffness contained in a non normal linear part are potential candidates to possess large one–sided Lipschitz constants (see [1, Sec. 1.1]).

Auzinger, Frank and Kirlinger [4] extended the B–theory to a class of stiff semi linear problems

$$y'(t) = f(t, y(t)) \equiv \widehat{J}(t)\ y(t) + g(t, y(t)), \quad t \in I, \quad y(t_0) = y_0 \in \mathbb{R}^m, \tag{2}$$

where $g(t, y)$ is Lipschitz continuous with respect to y and the logarithmic norm $\mu[S(t)\widehat{J}(t)S(t)^{-1}]$ is moderately sized for some smooth non singular matrix $S(t)$. These authors proved that any diagonally-stable and algebraically-stable RK method (A, b) with order p and stage order $q \leq p$, is B-convergent with order $\geq q$ for the class of semi linear problems (2). In the special case that $\widehat{J}(t) = J$ is a constant matrix, by requiring on the methods assumptions of linear-stability type, Burrage, Hundsdorfer and Verwer [5] obtained optimal convergence results with orders q (or $q + 1$), where q is the stage order of the method. Calvo, Montijano and Gonzalez-Pinto in [6], extended the convergence results of [5] to the class of semi-linear problems (2), with variation of $\widehat{J}(t)$ relatively bounded.

The purpose of this paper is to extend our results of [6] to the class of problems considered by Auzinger *et al.* [4], in the case that $J(t) = S(t)\widehat{J}(t)S(t)^{-1}$ varies on t in a relatively bounded form, when some smooth matrix $S(t)$ is considered. It will be seen that any A–stable RK method (A, b) with positive real part for each eigenvalue of A, is stable and convergent. Further, these properties also hold for some stiffly accurate A–stable methods having a first stage explicit, such as those of the Lobatto IIIA family and some DIRK methods, that have been used e.g. by Kennedy and Carpenter for convection–diffusion–reaction systems [11]. The paper is completed with some numerical experiments and comments about the convergence of some SDIRK methods. It must be also noticed that our results represent an improvement with regard to [4], in the case in which a bounded relative variation on $J(t)$ is assumed (see the assumption (H2) below). To keep the presentation within a reasonable length, we have omitted the proofs of the results, since they are based on an extension of those in [6]. An extended version of this paper that includes the proofs of the main results and more numerical experiments is given in [7].

2 Notations and basic assumptions

We assume that (2) possesses a unique smooth solution $y(t) = y(t; t_0, y_0),\ t \in I$, in the sense that for a positive integer p as large as required $\|y^{(j)}(t)\| \leq M_j = \mathcal{O}(1), \quad j = 0, \ldots, p + 1, \quad t \in I$, and $f(t, y)$ has continuous partial derivatives up to order p in some

cylinder $\mathcal{B}_\delta = \{(t,y);\ \|y - y(t)\| \le \delta, t \in I\}$ around the exact solution. The norm used is induced by some inner product with $\mu[\cdot]$ standing for the logarithmic norm associated to the induced norm and $\mathcal{O}(1)$ will mean any constant (or mapping) moderately sized independently of the stiffness.

We will consider semi-linear problems (2) under the following assumptions:

(H1) There exists a matrix $S(t) \in \mathbb{R}^{m,m}$ such that $J(t) := S(t)\widehat{J}(t)S(t)^{-1}$ satisfies $\mu[J(t)] \le 0$ and $S(t), S'(t), S^{-1}(t)$ are $\mathcal{O}(1)$ for $t \in I$.

(H2) There exist a constant $h^* > 0$ and a mapping $E_1(t, \Delta t)$, such that either

$$(i) \qquad J(t+\Delta t) - J(t) - \Delta t\ J(t)\ E_1(t,\Delta t) = \mathcal{O}(\Delta t),$$

or else

$$(ii) \qquad J(t+\Delta t) - J(t) - \Delta t\ E_1(t,\Delta t)\ J(t) = \mathcal{O}(\Delta t),$$

hold for all $t, t+\Delta t \in I$ with $|\Delta t| \le h^*$.

(H3) There exists a constant $\lambda_0 = \mathcal{O}(1)$ such that $\|g(t,y) - g(t,\tilde{y})\| \le \lambda_0 \|y - \tilde{y}\|$, for all $(t,y),\ (t,\tilde{y}) \in \mathcal{B}$.

The above assumptions imply that the stiffness of f is included in the linear term. Moreover the condition $\mu[J(t)] \le 0$ may be replaced by $\mu[J(t)] \le \nu = \mathcal{O}(1)$.

Observe that if $\widehat{J}(t)$ can be made diagonal by a smooth matrix $S(t)$ satisfying $\| S(t) \| \| S(t)^{-1} \| = \mathcal{O}(1)$, then the assumptions (H1)-(H2) are usually satisfied. This happens, for example, if the eigenvalues $\lambda_j(t)$ of $\widehat{J}(t)$ have a non–positive real part. The assumption (H2) is closely related to the relative Lipschitz condition used by Alexander in [1] as well as the assumption (a.1) introduced by van Dorsselaer and Spijker in [9] to study the convergence of Newton-type iterations in implicit RK methods. It has also been used by Calvo *et al.* [6] for the convergence analysis of Runge-Kutta methods by considering that the untransformed matrix $\widehat{J}(t)$ of (2) satisfies

$$\widehat{J}(t+\Delta t) - \widehat{J}(t) - \Delta t\ \widehat{J}(t)\ \widehat{E}_1(t,\Delta t) = \mathcal{O}(\Delta t),\ t \in I,\ \widehat{E}_1 = \mathcal{O}(1). \tag{3}$$

Assumptions similar to (H2) with $J(t)$ replaced by $\widehat{J}(t)$ have been also used by Strehmel and Weiner [14] and Schmitt [13] in connection with the analysis of stability and convergence of implicit Runge-Kutta methods and linearly implicit methods on time-dependant partial differential equations.

Remark 2.1 *For general non–linear problems (1), if $f(t,y)$ is analytic in a cylinder $\mathcal{B}_\delta$ around the exact solution $y = \varphi(t)$ of (1), then $f(t,y)$ can be written in the semilinear form (2) with $\hat{J}(t) = f_y(t,\varphi(t))$ and*

$$g(t,y) := f(t,\varphi(t)) - \hat{J}(t)\varphi(t) + \sum_{k\ge 2} \frac{1}{k!} f^{(k)}_{(t,\varphi(t))}(y - \varphi(t), \ldots^{(k}, y - \varphi(t)),$$

where $f^{(k)}_{(t,\varphi(t))}(u_1,\ldots,u_k)$ *denotes the k-Fréchet derivative of* f *with respect to* y *at* $(t,\varphi(t))$. *Now, if*

$$\|f^{(k)}_{(t,\varphi(t))}(u_1,\ldots,u_k)\| \leq \lambda_k \|u_1\|\cdots\|u_k\|,\ t\in[0,T],\ k=2,3,\ldots,$$

with $\|\lambda_k\| = \mathcal{O}(1)$, *then the assumption (H3) is fulfilled.*

For the numerical solution of (2) we consider an s-stage RK method specified by the Butcher matrices (A,b), $A=(a_{ij})\in\mathbb{R}^{s,s}$, $b=(b_i)\in\mathbb{R}^s$ and the knot vector $c=(c_i)=Ae$, $e=(1,\ldots,1)^T\in\mathbb{R}^s$. The step from $(t_0,y_0)\longrightarrow(t_1=t_0+h, y_1=Y_{s+1})$ is defined by the equations

$$Y_i = y_0 + h\sum_{j=1}^{s} a_{ij}\left[(S_j^{-1}J_jS_j)Y_j + g(\tau_j,Y_j)\right], \qquad (i=1,\ldots,s+1) \tag{4}$$

with $c_{s+1}=1, a_{s+1,j}=b_j$ and $\tau_i=t_0+c_ih,\ S_i=S(\tau_i),\ J_i=J(\tau_i)$.

The stage order q of the RK method is defined as $\min\{p_j; j=1,\ldots,s+1\}$ where the positive integers p_j are given by

$$\varepsilon_j \equiv y(t_0+c_jh) - y(t_0) - h\sum_{k=1}^{s} a_{jk}\, y'(t_0+c_kh) = \mathcal{O}(h^{p_j+1}), \quad j=1,\ldots,s+1.$$

with $c_{s+1}=1, a_{s+1,j}=b_j$. Note that by the smoothness of $y(t)$, the residual errors ε_i satisfy $\|\varepsilon_i\|\leq\mathcal{O}(h^{q+1})$.

To adapt the concepts of BSI-stability and BS-stability [8, Chaps. 5,7] to our class of problems, we consider the perturbed version of (4)

$$\widetilde{Y}_i = y_0 + h\sum_{j=1}^{s} a_{ij}\left[(S_j^{-1}J_jS_j)\widetilde{Y}_j + g(\tau_j,\widetilde{Y}_j)\right] + \eta_i, \quad (i=1,\ldots,s+1) \tag{5}$$

where $\eta_i\in\mathbb{R}^m$ are arbitrary perturbations.

In this situation, a RK method (A,b) is said to be BSI-stable if there exist two positive constants (independent of the stiffness) h^* and C_0 such that,

$$\max_{j=1,\ldots,s} \| Y_j - \widetilde{Y}_j \| \leq C_0 \sum_{i=1}^{s} \|\eta_i\|, \quad \text{whenever } h\in(0,h^*].$$

Further, the RK method (A,b) will be called BS–stable if

$$\|\widetilde{y}_1 - y_1\| = \|\widetilde{Y}_{s+1} - Y_{s+1}\| \leq C_1 \sum_{i=1}^{s+1} \|\eta_i\|, \quad C_1=\mathcal{O}(1).$$

We often require for the Runge-Kutta method (A,b) that each eigenvalue of its coefficient matrix A has a positive real part. This is equivalent to (see the assumption (M4) in [6])

$$(I-zA) \text{ is non singular for } \mathrm{Re}z\leq 0 \text{ and } \sup_{\mathrm{Re}z\leq 0} \|z(I-zA)^{-1}\|_2 < +\infty. \tag{6}$$

3 Main Results

In this section we state the main stability and convergence result whose proof has been omitted for the sake of brevity. These proofs are similar to the ones in previous paper of the authors [6], and can be seen in [7].

The first stability and convergence result is concerned with methods whose matrix A satisfies (6).

Theorem 3.1 *A Runge-Kutta method (A, b) satisfying (6),*

i) is BSI–stable and BS–stable for the class of stiff semi linear problems (2) under the assumptions (H1), (H2), (H3).

ii) If, moreover the method is A–stable, then it is convergent with order $\geq q$ (the stage order).

An important question is whether condition (6) is essential (apart from the A-stability) for convergence. We have found that there exist methods where A is a singular matrix with a special structure that are convergent for fixed step sizes or even on non uniform meshes, such that the number of given steps N multiplied by the maximum step-size is under some prefixed constant K. In particular, we have obtained positive convergence results for stiffly accurate methods, i.e., methods with $b^T = (a_{s1}, \ldots, a_{ss})$, with the first stage explicit and whose matrix A has the form

$$A = \begin{pmatrix} 0 & 0^T \\ a & \overline{A} \end{pmatrix} \in \mathbb{R}^{s,s}, \tag{7}$$

where the sub-matrix $\overline{A} \in \mathbb{R}^{(s-1),(s-1)}$ satisfies (6).

A convergence result for these stiffly accurate methods is given in the following:

Theorem 3.2 *Let (A, b) be a stiffly accurate RK method with a matrix A of type (7) and $\overline{A}$ satisfying (6), then*

i) If the method is A-stable, then it is convergent of order $\geq q$ (stage order) on uniform meshes for the class of stiff semi linear problems (2) under the assumptions (H1), (H2), (H3).

ii) In addition, it is also convergent with order $\geq q$ on special non uniform meshes $\{t_j\}_{j=0}^N$ provided that $N \max(t_j - t_{j-1}) \leq K$ with K independent of the grid.

Corollary 3.1 *The s-stage Lobatto IIIA method is convergent of order $q \geq s$, under the H-assumptions.*

Next result deals with AS– and ASI–stable methods. Recall (see e.g. [6]) that an s–stage RK method is said to be AS–stable (resp. ASI–stable) if $(I - zA)$ is a non singular matrix on Re $z \leq 0$ and

$$\sup_{\mathrm{Re} z \leq 0} \|zb^T(I - zA)^{-1}\|_2 < +\infty \quad \left(\text{resp.} \sup_{\mathrm{Re} z \leq 0} \|(I - zA)^{-1}\|_2 < +\infty \right). \tag{8}$$

For this case we have the weaker convergence (and internal stability) results,

Theorem 3.3 *An s-stage RK method (A, b) satisfying (8),*

i) is BSI–stable and BS–stable for the class of stiff semi linear problems (2) under the assumptions (H1), (H2-i), (H3), and (3).

ii) If, moreover the method is A–stable, then it is convergent with order not lesser than the stage order.

It must be remarked that if we replace in Theorem 3.2 the stiff-accuracy by the weaker assumption

$$b^T = d^T A, \quad \text{for some } d \neq e_j,\ j = 1, \ldots, s, \tag{9}$$

e_j denoting canonical vectors of $\mathbb{R}^s$, then the convergence statement in Theorem 3.2 does not necessarily hold as it will be clearly shown in the numerical experiments of the next section. However, under condition (9), the Theorem 3.3 holds true provided that the underlying method is also A-stable.

4 Numerical experiments

The aim of this section is to check the convergence behavior of several SDIRK methods on some variable coefficients stiff linear systems satisfying the H–assumptions. Two of these methods have been taken from the literature of stiff integrators and are not B–stable, but they satisfy the assumptions of either of the above theorems. The third one is a purposely chosen three–stage method with an A matrix of type (7) that is not stiffly accurate. They will be denoted by **SDIRKsX(p,ps)** where **s** is the number of stages, **X** is a reference to the author(s), **p** is the non stiff order and **ps** is the stage order. We have not included nonlinear Lipschitz continuous terms $g(t, y)$ in the problems presented here because they do not essentially modify the convergence behaviour of the methods. In our experiments we have used the Euclidean norm $\|\cdot\|_2$. The Runge-Kutta methods considered are:

- **SDIRK5HW(4,1)** is the five–stage SDIRK method given by Hairer–Wanner [10, p.100]. It has standard order four and stage–order one. It is stiffly accurate and L-stable. Since its coefficient matrix A satisfies (6), the method fulfils the assumptions of Theorems 3.1, 3.2 and 3.3.

- **SDIRK4A(3,2)** is the four–stage SDIRK method proposed by Alexander in [2, pp.2,6-8] and it is defined by

$$A = \begin{pmatrix} 0 & 0 & 0 & 0 \\ \gamma & \gamma & 0 & 0 \\ a_{31} & a_{32} & \gamma & 0 \\ b_1 & b_2 & b_3 & \gamma \end{pmatrix}, \qquad b = \begin{pmatrix} b_1 \\ b_2 \\ b_3 \\ \gamma \end{pmatrix}, \quad c = Ae,$$

with $\gamma = 0.43586652...$ and $c_3 = 1/2+\gamma/4$. Since it was required to have standard order three and stage order two, the remainder coefficients are univocally determined. It is L–stable, stiffly accurate and it satisfies (8). Hence, it meets the assumptions in Theorems 3.2 and 3.3.

- **SDIRK3(3,2)** is the three–stage SDIRK method defined by the coefficient matrices

$$A = \begin{pmatrix} 0 & 0 & 0 \\ \gamma & \gamma & 0 \\ a_{31} & a_{32} & \gamma \end{pmatrix}, \qquad b = \frac{1}{88} \begin{pmatrix} 52 - 72\gamma \\ 25 + 50\gamma \\ 11 + 22\gamma \end{pmatrix}, \tag{10}$$

with

$$a_{32} = \frac{2(1 - 2\sqrt{3})}{25}, \qquad a_{31} = c_3 - a_{32} - \gamma, \quad \gamma = \frac{3 + \sqrt{3}}{6}, \; c_3 = \frac{4}{5}.$$

This method is strongly A–stable ($R(\infty) = \gamma^{-1} - 2 = -0.732...$), has stage–order two and standard order three. It is not stiffly accurate but it satisfies (8), hence it fulfills the assumptions of Theorem 3.3, but not the ones of Theorem 3.2.

For the problems considered below, $\epsilon > 0$ is normally a small parameter, hence the stiffness of the problems, i.e. the Lipschitz condition on y for $f(t, y)$, is of order $\mathcal{O}(\epsilon^{-1})$.

Problem 1.- The two dimensional variable coefficient linear system

$$y'(t) = \widehat{J}(t)\,[y(t) - \varphi(t)] + \varphi'(t), \quad t \in [0, 4\pi], \quad y(0) = (1, 0)^T, \tag{11}$$

where $\varphi(t) = (\cos t, \sin t)^T$, $\widehat{J}(t) = S(t)^{-1}\Lambda S(t)$, $S(t) = PQ(t)$, and

$$\Lambda = \begin{pmatrix} -1 & 0 \\ 0 & -\epsilon^{-1} \end{pmatrix}, \; P = \begin{pmatrix} -1 & 0 \\ 1 & 1 \end{pmatrix}, \; Q(t) = \begin{pmatrix} 1 & \epsilon \\ e^{\sin t} & e^{\sin t} \end{pmatrix}. \tag{12}$$

The exact solution $y(t) = \varphi(t)$ is asymptotically stable and $\widehat{J}(t)$, that has as eigenvalues -1, and $-1/\varepsilon$, is highly non normal. In fact, its logarithmic norm behaves as ε^{-1} when $\varepsilon \to 0$. In this case $\widehat{J}(t)$ does not fulfill the assumptions of the B–theory. However, the assumption (H1), (H2-i) and (H3) are clearly satisfied.

It can be seen that for a matrix $\widehat{J}(t)$ of type $\widehat{J}(t) = S(t)^{-1}JS(t)$, with a non singular J, condition (3) holds if and only if $\parallel J^{-1}MJ - M \parallel$ can be bounded independently of the stiffness, where $M = (\Delta t)^{-1}(S(t)^{-1}S(t+\Delta t) - I)$. In this case we have that

$$J^{-1}MJ - M = \begin{pmatrix} 0 & -\nu(t) \\ -\nu(t) & 0 \end{pmatrix} \qquad \text{with} \quad \nu(t) = \frac{a(t+\Delta t) - a(t)}{a(t)\Delta t},$$

and therefore (3) is satisfied independently of ϵ.

Problem 2.- The variable coefficient linear problem similar to that one considered by Kreiss in [12],

$$y' = \widehat{J}(t)y \equiv S(t)^{-1}\ \Lambda\ S(t)\ y, \quad t \in [0, 4\pi], \tag{13}$$

where P and Λ are given by (12), $S(t) = P\Omega(t)$, and

$$\Omega(t) = \begin{pmatrix} \cos(t) & \sin(t) \\ -\sin(t) & \cos(t) \end{pmatrix}. \tag{14}$$

Since $y \to S(t)y$ transforms (13) into a constant coefficient linear system, the general solution of (13)-(14) can be written in the form

$$y(t) = S(t)^{-1} \begin{pmatrix} 1 & 1 \\ \lambda_+ & \lambda_- \end{pmatrix} \begin{pmatrix} C_+ e^{\lambda_+ t} \\ C_- e^{\lambda_- t} \end{pmatrix}, \tag{15}$$

with $\lambda_+ = -2\epsilon + \mathcal{O}(\epsilon^2)$ and $\lambda_- = -\epsilon^{-1} - 1 + \mathcal{O}(\epsilon)$. Moreover, all solutions tend quickly, after the initial transient layer, to the smooth stationary solution, which corresponds to the parameter $C_- = 0$. We have chosen $C_- = 0$ and $C_+ = 1$ for our numerical experiments. For the logarithmic norm, it can be shown that $\mu_2[\widehat{J}(t)] = \mathcal{O}(\epsilon^{-1}) \gg 1$. However, the assumptions (H1), (H2), (H3) are satisfied. In this case, the condition (3) is not accomplished.

Problem 3.- The linear system of partial differential equations of parabolic type,

$$\left.\begin{aligned} u_t &= a_{11}(t)u_{xx} + a_{12}(t)v_{xx} + r_1(x,t) \\ v_t &= a_{21}(t)u_{xx} + a_{22}(t)v_{xx} + r_2(x,t) \end{aligned}\right\} \tag{16}$$

where $u = u(x,t)$ and $v = v(x,t)$ represent the unknowns, the space variable x ranges in $[0,1]$, the functions $r_i(x,t),\ (i = 1,2)$ are given by

$$r_1(x,t) = 2\cos t - \phi(x)\sin t, \quad r_2(x,t) = -(2\sin t + \phi(x)\cos t), \quad \phi(x) = x(1-x).$$

and $a_{ij}(t)$ are defined by

$$A(t) = \begin{pmatrix} a_{11}(t) & a_{12}(t) \\ a_{21}(t) & a_{22}(t) \end{pmatrix} \equiv \Omega(t)(-\Lambda)\Omega(t)^{-1}, \tag{17}$$

where $\Omega(t)$ and Λ are given by (14) and (12) respectively.

Taking as initial–boundary conditions

$$u(x,0)=\phi(x),\ v(x,0)=0, \qquad u(0,t)=u(1,t)=0,\ v(0,t)=v(1,t)=0,$$

the exact solution of (16) is $u(x,t)=\phi(x)\cos t, \quad v(x,t)=-\phi(x)\sin t$, independent of the parameter ϵ (only $\epsilon>0$ gives stable solutions).

Taking a uniform grid $x_j=j\Delta x, j=0,1,\ldots,M+1$, in the spatial variable with $\Delta x=1/(M+1)$, and using second order centered differences, the semi discretization of (16) can be written as

$$\begin{aligned} u_j' &= a_{11}(t)\frac{u_{j-1}-2u_j+u_{j+1}}{(\Delta x)^2}+a_{12}(t)\frac{v_{j-1}-2v_j+v_{j+1}}{(\Delta x)^2}+r_1(x_j,t),\\ v_j' &= a_{21}(t)\frac{u_{j-1}-2u_j+u_{j+1}}{(\Delta x)^2}+a_{22}(t)\frac{v_{j-1}-2v_j+v_{j+1}}{(\Delta x)^2}+r_2(x_j,t),\\ & j=1,2,\ldots,M, \end{aligned} \tag{18}$$

where $u_j(t)=u(x_j,t)$ and $v_j(t)=v(x_j,t)$, $j=0,\ldots,M+1$. The initial conditions imply that $u_j(0)=\phi(x_j),\ v_j(0)=0,\ j=1,\ldots,M$. It must be observed that the discrete exact solutions of (18) and (16) are the same.

By introducing the matrix

$$W=\frac{1}{(\Delta x)^2}\begin{pmatrix} -2 & 1 & & & \\ 1 & -2 & 1 & & \\ & \ddots & \ddots & \ddots & \\ & & 1 & -2 & 1 \\ & & & 1 & -2 \end{pmatrix}\in\mathbb{R}^{M,M},$$

and the vectors

$$U(t)=\begin{pmatrix} u_1(t)\\ \vdots \\ u_M(t)\end{pmatrix}, \qquad V(t)=\begin{pmatrix} v_1(t)\\ \vdots \\ v_M(t)\end{pmatrix}, \qquad y(t)=\begin{pmatrix} U(t)\\ V(t)\end{pmatrix},$$

$$\widetilde{R}_1(t)=\begin{pmatrix} r_1(x_1,t)\\ \vdots \\ r_1(x_M,t)\end{pmatrix}, \quad \widetilde{R}_2(t)=\begin{pmatrix} r_2(x_1,t)\\ \vdots \\ r_2(x_M,t)\end{pmatrix}, \quad g(t)=\begin{pmatrix} \widetilde{R}_1(t)\\ \widetilde{R}_2(t)\end{pmatrix},$$

we get the initial value problem

$$y'(t)=\widehat{J}(t)y+g(t), \qquad y(0)=(\phi(x_1),\ldots,\phi(x_N),0,\ldots,0)^T, \tag{19}$$

with $\widehat{J}(t)=A(t)\otimes W$.

Since $\widehat{J}(t) = (\Omega(t) \otimes I)\ (-\Lambda \otimes W)(\Omega(t)^{-1} \otimes I)$, the assumption (H1) is satisfied for $J(t) = S(t)\widehat{J}(t)S(t)^{-1} = -\Lambda \otimes W$ because it is a constant matrix that fulfills $\mu_2[J(t)] \simeq -\pi^2$. Further (H2) and (H3) are clearly accomplished.

For the smoothness of the exact solution $y(t)$, it is enough to consider the weighted Euclidean norm $\| \cdot \| = M^{-1/2} \| \cdot \|_2$. Thus, it follows that $\| y^{(j)}(t) \| \leq \max_{x\in[0,1]} |\phi(x)| = 1/4, \quad j = 0, 1, 2, \ldots, \quad t \in \mathbb{R}$. Observe that for any matrix B, both norms are identical, i.e., $\| B \| = \| B \|_2$. In conclusion, this problem fulfills the H–assumptions. For our numerical experiments we have selected $M = 20$, so that the dimension of the linear system is 40. In this case, unlike of the problems 1 and 2, the stiffness comes from two sources, from the small parameter ϵ and from the discretization in space.

For each problem and method we have carried out integrations for $t \in [0, 4\pi]$ with fixed step sizes $h = 4\pi/N$ for $N = 100, 200, 400, 800, 1600$. Note that, since in all problems the initial conditions have been taken on a stationary (smooth) solution, a fixed step size strategy can be used in the integrations. We have computed the global error at the end point $GE(N) = \|y(t_N) - y_N\|$ and also the numerical order of convergence p_N defined by

$$p_N = \frac{1}{\log(2)}\ \log\left(\frac{GE(N)}{GE(2N)}\right).$$

Concerning the method SDIRK5HW(4,1) observe that Th. 3.1 implies that it is convergent for all problems with order greater or equal that the stage order $q = 1$. This can be checked in Table 1, where the global errors $GE(N)$ and numerical orders p_N obtained for the Problems 1 and 2, have been displayed for several values of the parameter ϵ. Similar results, not included here, were encountered for Problem 3. From these results we follow:

- $GE(N) \to 0$ when $N \to \infty$ for a wide range of values of ϵ, i.e., the method is convergent also for the considered stiff problems.

- For the non–stiff case $\epsilon = \mathcal{O}(1)$, the computed numerical orders agree with the classical order of convergence $p = 4$ as expected.

- For small ϵ-values the observed orders range, in most cases, between the stage order and the classical order. Hence, the lower order of convergence can not be improved in general.

- In Problem 1, for fixed step-sizes, the $GE(N)$ decreases when $\epsilon \to 0$. For this problem, it can be shown that the local error tends to zero when $\epsilon \to 0$ and this property is also reflected in the global error behaviour.

- It must be also remarked that the first–order of convergence above cannot be explained from the ϵ–theory (see Corollary 3.10, pp. 402-403 in [10, Chap. VI.3]),

since from that theory global errors of size $\mathcal{O}(h^4 + \epsilon h)$ are expected, and this is not the case as it can be observed from the results in Table 1, when moving either on rows (fixed h) or on columns (fixed ϵ).

Table 1.— Method SDIRK5HW(4,1) for Problems 1 and 2

Problem 1

	$\epsilon = 0.5$		$\epsilon = 10^{-2}$		$\epsilon = 10^{-4}$		$\epsilon = 10^{-6}$		$\epsilon = 10^{-8}$	
N	$GE(N)$	p_N	$GE(N)$	p_N	$GE(N)$	p_N	$GE(N)$	p_N	$GE(N)$	p_N
50	3.1e-5		1.3e-3		1.8e-5		3.7e-6		3.9e-6	
100	2.4e-6	3.68	3.9e-4	1.78	1.0e-5	0.78	2.0e-7	4.26	2.8e-7	3.80
200	1.7e-7	3.85	7.2e-5	2.44	5.2e-6	0.98	3.7e-8	2.41	1.9e-8	3.93
400	1.1e-8	3.93	8.9e-6	3.02	2.6e-6	1.02	2.5e-8	0.54	1.0e-9	4.21
800	7.1e-10	3.96	8.4e-7	3.40	1.2e-6	1.04	1.3e-8	0.95	7.1e-11	3.82
1600	4.5e-11	3.98	6.7e-8	3.65	5.8e-7	1.09	6.5e-9	1.00	6.1e-11	0.21

Problem 2

50	4.9e-9		1.1e-3		2.4e-4		2.7e-4		2.7e-4	
100	3.3e-10	3.91	2.2e-4	2.37	1.0e-5	4.55	1.8e-5	3.88	1.8e-5	3.87
200	2.1e-11	3.95	2.7e-5	2.99	9.6e-7	3.41	1.2e-6	3.95	1.3e-6	3.82
400	1.3e-12	3.97	2.8e-6	3.26	4.8e-7	1.01	7.2e-8	4.03	9.8e-8	3.73
800	8.5e-14	3.99	2.8e-7	3.36	1.3e-7	1.85	4.0e-9	4.17	4.0e-8	1.28
1600	5.3e-15	3.99	2.4e-8	3.52	3.2e-8	2.05	5.7e-10	2.81	1.3e-8	1.61

Numerical experiments with the method SDIRK4A(3,2) for the three problems are presented in Table 2. Note that Theorem 3.1 can not be applied to this method but it meets the assumptions of Theorem 3.2. From the displayed results, apart from convergence behaviour for all problems, it can be observed that

- For the non–stiff case $\epsilon = \mathcal{O}(1)$, the computed numerical orders agree with the classical order of convergence $p = 3$.
- For small ϵ the order reduction is not observed in Problems 1 and 2. However, for Problem 3 numerical orders lower than 3 are found, therefore the stage order $q = 2$, seems to be the guaranteed order of convergence.
- In all problems, for a fixed stepsize h, the $GE(N)$ seem to be non-dependent on ϵ. This fact is explained because the local errors are practically independent on ϵ.

Table 2.— Method SDIRK4A(3,2) for Problems 1, 2 and 3

Problem 1

	$\epsilon = 0.5$		$\epsilon = 10^{-2}$		$\epsilon = 10^{-4}$		$\epsilon = 10^{-6}$		$\epsilon = 10^{-8}$	
N	$GE(N)$	p_N	$GE(N)$	p_N	$GE(N)$	p_N	$GE(N)$	p_N	$GE(N)$	p_N
50	2.7e-4		3.7e-4		4.0e-4		4.0e-4		4.0e-4	
100	4.4e-5	2.63	4.6e-5	2.98	5.3e-5	2.90	5.3e-5	2.90	5.3e-5	2.90
200	6.4e-6	2.76	5.5e-6	3.08	6.9e-6	2.95	6.9e-6	2.95	6.9e-6	2.95
400	8.8e-7	2.87	6.2e-7	3.16	8.7e-7	2.98	8.8e-7	2.97	8.8e-7	2.97
800	1.2e-7	2.93	6.7e-8	3.21	1.1e-7	3.00	1.1e-7	2.99	1.1e-7	2.99
1600	1.5e-8	2.97	7.3e-9	3.20	1.4e-8	3.01	1.4e-8	2.99	1.4e-8	2.99

Problem 2

50	5.6e-8		6.8e-3		9.1e-3		9.1e-3		9.1e-3	
100	7.7e-9	2.87	8.6e-4	2.99	1.2e-3	2.95	1.2e-3	2.96	1.2e-3	2.96
200	1.0e-9	2.92	1.0e-4	3.11	1.5e-4	2.98	1.5e-4	2.98	1.5e-4	2.98
400	1.3e-10	2.96	9.7e-6	3.36	1.9e-5	2.99	1.9e-5	2.99	1.9e-5	2.99
800	1.7e-11	2.98	6.5e-7	3.89	2.4e-6	2.99	2.4e-6	2.99	2.4e-6	2.99
1600	2.1e-12	2.99	1.0e-8	6.03	3.0e-7	2.99	3.0e-7	3.00	3.0e-7	3.00

Problem 3

50	2.0e-5		9.8e-7		1.5e-7		1.5e-7		1.5e-7	
100	3.3e-6	2.58	2.4e-7	2.06	2.8e-8	2.48	2.7e-8	2.49	2.8e-8	2.47
200	5.0e-7	2.70	5.5e-8	2.11	7.0e-9	1.98	6.2e-9	1.98	6.6e-9	2.07
400	7.2e-8	2.81	1.2e-8	2.19	1.1e-9	2.63	1.1e-9	2.63	9.2e-10	2.85
800	9.7e-9	2.89	2.5e-9	2.28	1.6e-10	2.80	1.6e-10	2.84	1.5e-10	2.57
1600	1.3e-9	2.94	4.7e-10	2.38	2.3e-11	2.81	2.2e-11	2.85	1.8e-10	-0.24

Finally, the method SDIRK3(3,2) meets the assumptions of Theorem 3.3 and therefore it is convergent with order greater or equal than the stage order ($q = 2$) for Problem 1. This fact agrees with the numerical results displayed in Table 3, where a third–order convergence is observed for that problem. However, SDIRK3(3,2) satisfies neither the assumptions of Theorem 3.1 nor those of Theorem 3.2, hence the convergence on Problems 2 and 3 is not guaranteed. In fact, the numerical experiments in Table 3 (Problem 2) show that the method is not B-convergent on the whole class of Problems 2 with $0 < \epsilon \leq 1$.

Similar results of not B-convergence were encountered for Problem 3 when considering small values for ϵ.

5 Conclusions

New theoretical results on stability and convergence for stiff semi–linear problems that extend previous results on the subject [4], [6] have been derived. The new results support the use of some SDIRK formulas [2], [10], [11] that have been designed taking into account their linear stability properties and efficient implementation in practical codes. Numerical experiments show that the assumptions on the theorems and stiff orders can not be improved for the class of problems under consideration.

Table 3.— Method SDIRK3(3,2) for Problems 1 and 2

Problem 1

	$\epsilon = 0.5$		$\epsilon = 10^{-2}$		$\epsilon = 10^{-4}$		$\epsilon = 10^{-6}$		$\epsilon = 10^{-8}$	
N	$GE(N)$	p_N	$GE(N)$	p_N	$GE(N)$	p_N	$GE(N)$	p_N	$GE(N)$	p_N
50	6.9e-4		1.1e-3		1.1e-3		1.1e-3		1.1e-3	
100	1.1e-4	2.64	1.5e-4	2.87	1.6e-4	2.84	1.6e-4	2.84	1.6e-4	2.84
200	1.7e-5	2.67	2.0e-5	2.97	2.1e-5	2.91	2.1e-5	2.91	2.1e-5	2.91
400	2.5e-6	2.80	2.4e-6	3.05	2.7e-6	2.95	2.7e-6	2.95	2.7e-6	2.95
800	3.4e-7	2.89	2.7e-7	3.12	3.4e-7	2.98	3.5e-7	2.98	3.5e-7	2.98
1600	4.4e-8	2.94	3.0e-8	3.17	4.3e-8	2.99	4.4e-8	2.99	4.3e-8	2.99

Problem 2

50	2.0e-7		3.3e-2		4.8e+65		1.9e+163	–	1.8e+261	–
100	2.7e-8	2.87	3.4e-3	3.25	7.7e+59	–	3.0e+251	–	1.4e+303	–
200	3.5e-9	2.92	3.4e-4	3.35	6.0e+10	–	4.9e+303	–	2.4e+304	–
400	4.6e-10	2.95	3.0e-5	3.49	8.2e-04	–	7.8e+303	–	1.8e+304	–
800	5.8e-11	2.97	2.0e-6	3.90	5.7e-05	3.84	5.0e+154	–	6.3e+303	–
1600	7.3e-12	2.99	2.8e-8	6.14	4.0e-06	3.83	3.0e-004	–	1.2e+303	–

Acknowledgements

This work was supported by projects MTM2007-67530-C02-01 and MTM2007-67530-C02-02.

References

[1] R. K. Alexander, *Stability of Runge–Kutta methods for stiff ordinary differential equations*, SIAM J. Numer. Anal., 31, 4 (1994), pp. 1147–1168.

[2] R. K. Alexander, *Design and implementation of DIRK integrators for stiff systems*, Appl. Numer. Math. 46 (2003), pp. 1-17.

[3] W. Auzinger, R. Frank and G. Kirlinger, *A note on convergence concepts for Stiff Problems*, Computing 44 (1990), pp. 197–208.

[4] W. Auzinger, R. Frank and G. Kirlinger, *An extension of B–convergence for Runge–Kutta methods*, Appl. Numer. Math., 9 (1992), pp. 91–109.

[5] K. Burrage, W. H. Hundsdorfer and J. G. Verwer, *A study of B-convergence of Runge-Kutta methods*, Computing 36 (1986), pp. 17–34.

[6] M. Calvo, J. I. Montijano and S. Gonzalez-Pinto, *Runge–Kutta methods for the numerical solution of stiff semilinear initial value problems*, BIT 40, 4 (2000), pp. 611–639.

[7] M. Calvo, J. I. Montijano and S. Gonzalez-Pinto, *On the convergence of Runge-Kutta methods for stiff semi linear initial value problems*, Technical report, Dept. de Matemática aplicada, Univ. de Zaragoza (2005), http://pcmap.unizar.es/numerico/reports.

[8] K. Dekker and J. G. Verwer, *Stability of Runge-Kutta methods for stiff nonlinear differential equations*, North Holland, Amsterdam, 1984.

[9] J. L. M. van Dorsselaer van and M. N. Spijker, *The error committed by stopping the Newton iteration in the numerical solution of stiff initial value problems*, IMA J. Num. Anal., 14 (1994), pp. 183–209.

[10] E. Hairer and G. Wanner, *Solving Ordinary Differential Equations II: Stiff and Differential–Algebraic Problems*, Springer Verlag, Berlin, 1996.

[11] C. A. Kennedy and M. H. Carpenter, *Additive Runge-Kutta schemes for convection-diffusion-reaction equations*, Appl. Numer. Math., 44, 1–2 (2003), pp. 139–181.

[12] H.O. Kreiss, *Difference methods for stiff ordinary differential equations*, SIAM J. Numer. Anal., 15 (1978), pp. 21–58.

[13] B. A. Schmitt, *Stability of implicit Runge-Kutta methods for nonlinear stiff differential equations*, BIT, 28, (1988), pp. 884–897.

[14] K. Strehmel and R. Weiner, *B-convergence results for linearly implicit one step methods*, BIT, 27, (1987), pp. 264–281.

Monografías de la Real Academia de Ciencias de Zaragoza **33**, 155–163, (2010).

Modificaciones del método de Variación de los Parámetros. Aplicaciones en Astrodinámica.

R. Barrio y S. Serrano

Grupo de Mecánica Espacial. Dpto. Matemática Aplicada. IUMA
Universidad de Zaragoza.50009 Zaragoza. Spain.

Dedicado al Prof. Manuel Calvo Pinilla en conmemoración de su 65 cumpleaños.

Resumen

En los últimos años se ha dedicado mucho tiempo y esfuerzo a la integración numérica de problemas de Mecánica Celeste. En particular, se han desarrollado métodos numéricos eficientes basados en la preservación de propiedades geométricas asociadas a dichos problemas como los métodos simplécticos, reversibles, etc. Otra opción es la búsqueda de una formulación lo más adecuada posible para su integración numérica o analítica. Así surgieron las variables orbitales, el método de variación de las constantes, sistemas de variables redundantes, etc.

En el presente artículo pretendemos realizar una revisión del método de variación de las constantes, analizando y generalizando una modificación propuesta en la literatura: el método de Dziobek-Brouwer [6]. Finalmente, se presenta una comparación numérica tomando como problema el movimiento de un satélite artificial terrestre sujeto a la perturbación del potencial gravitatorio, la cual muestra las ventajas de estos métodos frente a la integración directa del problema en coordenadas cartesianas.

1 Introducción

Euler desarrolló el método de variación de los parámetros a mediados del siglo XVIII para estudiar las perturbaciones existentes entre Júpiter y Saturno, sin embargo, sus resultados no fueron del todo correctos debido a que no consideró el hecho de que todos los elementos orbitales variasen a la vez. Posteriormente, el método desarrollado por Euler fue perfeccionado por Lagrange, quien siguió considerando algunos elementos orbitales como constantes, lo que ocasionó errores en algunas de sus ecuaciones. En 1782, el

propio Lagrange corrigió y completó el método desarrollándolo en un trabajo sobre las perturbaciones de los cometas en órbitas elípticas y, algo más tarde, lo usaría también para el estudio del movimiento de los planetas. Este método sigue siendo empleado actualmente en numerosos estudios de carácter científico [1, 3, 4, 12, 13, 14].

En este trabajo estudiamos el método de variación de los parámetros y una modificación suya [3]. Así, en la sección 2 damos dos resultados teóricos que son la base del método clásico. En la sección 3 aplicamos dicho método sobre el problema de dos cuerpos. En la siguiente sección 4 damos el resultado que generaliza una modificación, inicialmente desarrollada para el problema de dos cuerpos, que combina el método de variación de los parámetros con el método de Encke [10]. Finalmente, en la sección 5, aplicamos los métodos sobre un problema concreto, un satélite artificial terrestre, para mostrar las ventajas de dichas formulaciones.

2 Variación de los parámetros

En esta sección introduciremos algunos resultados teóricos sobre los que se fundamenta el método [15], su demostración y un análisis más profundo de los mismos puede encontrarse en [3].

Teorema *Sean* $\boldsymbol{x}$ *e* $\boldsymbol{y}$ *las soluciones de*

$$\dot{\boldsymbol{x}} = \boldsymbol{F}_0(t, \boldsymbol{x}), \qquad \boldsymbol{x}(t_0) = \boldsymbol{x}_0 \tag{1}$$

$$\dot{\boldsymbol{y}} = \boldsymbol{F}_0(t, \boldsymbol{y}) + \boldsymbol{F}_p(t, \boldsymbol{y}), \qquad \boldsymbol{y}(t_0) = \boldsymbol{x}_0, \tag{2}$$

respectivamente, y supongamos que $\partial \boldsymbol{F}_0/\partial \boldsymbol{x}$ *y* $\partial \boldsymbol{F}_p/\partial \boldsymbol{y}$ *existen y son continuas. Además, supongamos que el primer sistema es integrable por cuadraturas y que su solución viene dada en función de un conjunto de n constantes de integración* $\boldsymbol{\alpha}_0 \in \mathbb{R}^n$*, es decir,*

$$\boldsymbol{x}(t) = \boldsymbol{f}(t; \boldsymbol{\alpha}_0).$$

Entonces, cualquier solución del sistema (2) viene dada por

$$\boldsymbol{y}(t) = \boldsymbol{f}(t; \boldsymbol{\alpha}(t)),$$

donde los parámetros $\boldsymbol{\alpha}(t)$ *y las constantes de integración* $\boldsymbol{\alpha}_0$ *están conectadas por*

$$\boldsymbol{\alpha}(t) = \boldsymbol{\alpha}_0 + \int_{t_0}^{t} \left(\frac{\partial \boldsymbol{y}}{\partial \boldsymbol{\alpha}}(s) \right)^{-1} \boldsymbol{F}_p(s, \boldsymbol{y}(s)) \mathrm{d}s.$$

Corolario *Sea* $(\boldsymbol{r}, \boldsymbol{v}) \in \mathbb{R}^{2n}$ *la solución de un sistema diferencial hamiltoniano integrable en el sentido de Liouville-Arnold dada por*

$$\dot{\boldsymbol{r}} = \boldsymbol{v}, \quad \dot{\boldsymbol{v}} = \frac{\partial U(\boldsymbol{r})}{\partial \boldsymbol{r}}, \quad (\boldsymbol{r}, \boldsymbol{v})(t_0) = (\boldsymbol{r}_0, \boldsymbol{v}_0),$$

con $U(\boldsymbol{r})$ la energía potencial. Además, denotamos por $(\boldsymbol{R},\boldsymbol{V}) \in \mathbb{R}^{2n}$ la solución del sistema perturbado

$$\dot{\boldsymbol{R}} = \boldsymbol{V}, \qquad \dot{\boldsymbol{V}} = \frac{\partial U(\boldsymbol{R})}{\partial \boldsymbol{R}} + \boldsymbol{F}_p(\boldsymbol{R},\boldsymbol{V}),$$

con las mismas condiciones iniciales $(\boldsymbol{R},\boldsymbol{V})(t_0) = (\boldsymbol{r}_0,\boldsymbol{v}_0)$. Entonces, las soluciones de los dos sistemas anteriores pueden ser obtenidas de forma implícita como

$$\boldsymbol{r}(t) = \boldsymbol{f}(t;\boldsymbol{\alpha}_0), \qquad \boldsymbol{v}(t) = \frac{\partial \boldsymbol{f}(t;\boldsymbol{\alpha}_0)}{\partial t} = \boldsymbol{g}(t;\boldsymbol{\alpha}_0),$$

$$\boldsymbol{R}(t) = \boldsymbol{f}(t;\boldsymbol{\alpha}(t)), \quad \boldsymbol{V}(t) = \frac{\partial \boldsymbol{f}(t;\boldsymbol{\alpha}(t))}{\partial t} = \boldsymbol{g}(t;\boldsymbol{\alpha}(t)),$$

donde los parámetros $\boldsymbol{\alpha}(t)$ y las constantes de integración $\boldsymbol{\alpha}_0$ están conectados por

$$\boldsymbol{\alpha}(t) = \boldsymbol{\alpha}_0 + \int_{t_0}^{t} \frac{\partial \boldsymbol{\alpha}}{\partial \boldsymbol{V}}(s) \cdot \boldsymbol{F}_p(\boldsymbol{R}(s),\boldsymbol{V}(s))\mathrm{d}s.$$

3 Variación de parámetros en el problema de dos cuerpos

Las ecuaciones del movimiento de dos cuerpos bajo una fuerza perturbadora $\boldsymbol{F}_p$ pueden formularse como un sistema de primer orden en la forma

$$\begin{cases} \dot{\boldsymbol{r}} = \boldsymbol{v}, \\ \dot{\boldsymbol{v}} = -\dfrac{\mu}{r^3}\boldsymbol{r} + \boldsymbol{F}_p(\boldsymbol{r},\boldsymbol{v}), \end{cases}$$

donde r hace referencia a la norma del vector $\boldsymbol{r}$ y μ a la constante de Gauss.

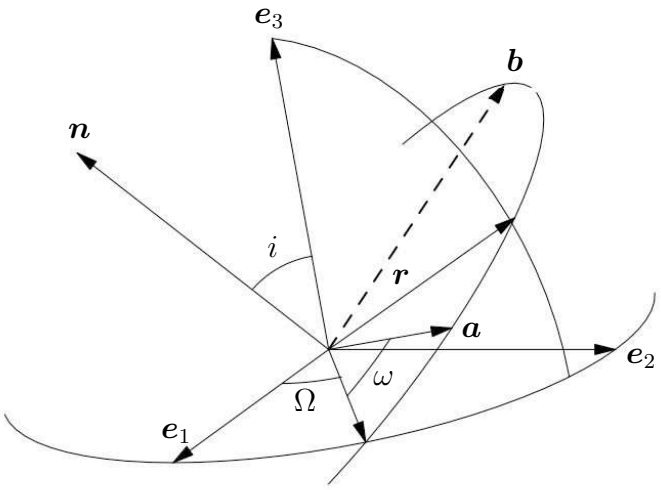


Figura 1.— Sistema apsidal y elementos orbitales clásicos.

Como se puede observar, dicho sistema corresponde al caso contemplado en el corolario de la sección anterior y, si consideramos el conjunto de elementos orbitales $\boldsymbol{\alpha} = (a, e, i, l_0, \omega, \Omega)$ (ver figura 1), para el caso $\boldsymbol{F}_p = 0$ existe una solución analítica que nos da la posición $\boldsymbol{r}_k = \boldsymbol{r}_k(t, \boldsymbol{\alpha}_0)$ y velocidad $\boldsymbol{v}_k = \boldsymbol{v}_k(t, \boldsymbol{\alpha}_0)$, en función de los elementos orbitales, que en este caso son constantes. Si $\boldsymbol{F}_p \neq 0$, el vector $\boldsymbol{\alpha}$ ya no es constante y para determinarlo tendremos que plantear el sistema dado por el corolario

$$\frac{\mathrm{d}\boldsymbol{\alpha}}{\mathrm{d}t} = \left(\frac{\partial \boldsymbol{\alpha}}{\partial \boldsymbol{v}}\right) \boldsymbol{F}_p,$$

que se corresponde, para este problema, con el clásico sistema de Gauss.

A la vista del sistema, vemos la necesidad de calcular la matriz de derivadas parciales que puede obtenerse como

$$\left(\frac{\partial \boldsymbol{\alpha}}{\partial \boldsymbol{v}}\right) = D \cdot \left(\frac{\partial \boldsymbol{r}}{\partial \boldsymbol{\alpha}}\right)^{\top},$$

$$\text{con} \quad D = \begin{pmatrix} 0 & d \\ -d^{\top} & 0 \end{pmatrix}, \quad \text{siendo} \quad d = \begin{pmatrix} \dfrac{2\sqrt{a}}{\sqrt{\mu}} & 0 & 0 \\ \dfrac{\eta^2}{\sqrt{\mu a}e} & \dfrac{-\eta}{\sqrt{\mu a}e} & 0 \\ 0 & \dfrac{\cot i}{\sqrt{\mu a}\eta} & \dfrac{-1}{\sqrt{\mu a}\eta \operatorname{sen} i} \end{pmatrix}$$

y

$$\frac{\partial \mathbf{r}}{\partial (a, e, l_0)} = M_p \cdot M_E, \qquad \frac{\partial \mathbf{r}}{\partial i} = M_p \cdot M_\omega \cdot V_p,$$

$$\frac{\partial \mathbf{r}}{\partial \omega} = M_p \cdot M_c \cdot V_p, \qquad \frac{\partial \mathbf{r}}{\partial \Omega} = M_c \cdot M_p \cdot V_p,$$

donde

$$M_E = \begin{pmatrix} (\cos E - e) + \dfrac{3}{2}\dfrac{a}{r} tn \sin E & -a(1 + \dfrac{a}{r}\sin^2 E) & -\dfrac{a^2}{r}\sin E \\ \eta \sin E - \dfrac{3}{2}\dfrac{an}{r} tn \cos E & \dfrac{a^2}{r\eta}\sin E(\cos E - e) & \dfrac{a^2\eta}{r}\cos E \\ 0 & 0 & 0 \end{pmatrix},$$

$$M_c = \begin{pmatrix} 0 & -1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}, \quad M_\omega = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ \sin\omega & \cos\omega & 0 \end{pmatrix},$$

y M_p es la matriz cuyas columnas son los vectores del sistema ortonormal estándar que define la orientación espacial de la órbita y que dependen únicamente de los elementos orbitales i, ω y Ω. Como es habitual, E denota la anomalía excéntrica, $\eta = \sqrt{1 - e^2}$ y $n = \sqrt{\mu/a^3}$ el movimiento medio.

4 Modificación del método

El siguiente resultado generaliza el método de Dziobek-Brouwer [2, 6, 7, 9] definido para el problema de dos cuerpos.

Teorema *Sean* $\boldsymbol{x} \in \mathbb{R}^n$ *e* $\boldsymbol{y} \in \mathbb{R}^n$ *las soluciones de*

$$\begin{aligned} \dot{\boldsymbol{x}} &= \boldsymbol{F}_0(t, \boldsymbol{x}), & \boldsymbol{x}(t_0) &= \boldsymbol{x}_0, \\ \dot{\boldsymbol{y}} &= \boldsymbol{F}_0(t, \boldsymbol{y}) + \boldsymbol{F}_p(t, \boldsymbol{y}), & \boldsymbol{y}(t_0) &= \boldsymbol{x}_0, \end{aligned}$$

respectivamente, y supongamos que el primer sistema es integrable por cuadraturas, siendo $\boldsymbol{\alpha}_0$ *un conjunto de* n *constantes de integración que definen la solución. Además, supongamos que* $\partial \boldsymbol{\alpha}_0 / \partial \boldsymbol{x}$ *es continua. Entonces las soluciones de los sistemas anteriores están conectadas por*

$$\boldsymbol{y}(t) = \boldsymbol{x}(t) + \frac{\partial \boldsymbol{x}(t)}{\partial \boldsymbol{\alpha}_0} \int_{t_0}^{t} \frac{\partial \boldsymbol{\alpha}_0}{\partial \boldsymbol{x}}(\tau, \boldsymbol{x}(\tau)) \cdot \boldsymbol{F}_p^*(\tau, \boldsymbol{x}(\tau), \boldsymbol{y}(\tau)) d\tau,$$

donde $\boldsymbol{F}_p^*$ *viene definida como*

$$\boldsymbol{F}_p^*(t, \boldsymbol{x}, \boldsymbol{y}) = \boldsymbol{F}_0(t, \boldsymbol{y}) - \boldsymbol{F}_0(t, \boldsymbol{x}) - \frac{\partial \boldsymbol{F}_0(t, \boldsymbol{x})}{\partial \boldsymbol{x}} \boldsymbol{s} + \boldsymbol{F}_p(t, \boldsymbol{y}),$$

siendo $\boldsymbol{s} = \boldsymbol{y} - \boldsymbol{x}$.

Aplicando dicho teorema sobre el problema de dos cuerpos llegamos a que la solución del sistema perturbado puede obtenerse mediante la expresión

$$\boldsymbol{r}(t) = \boldsymbol{r}_k(t) + \left(\frac{\partial \boldsymbol{r}_k}{\partial \boldsymbol{\alpha}} \right) \cdot \boldsymbol{K},$$

donde $\left(\dfrac{\partial \boldsymbol{r}_k}{\partial \boldsymbol{\alpha}} \right)$ se corresponde con las fórmulas aparecidas en la sección anterior pero evaluadas en los elementos orbitales $\boldsymbol{\alpha}_0$ (problema sin perturbar), mientras que $\boldsymbol{K}$ es la solución del sistema diferencial

$$\dot{\boldsymbol{K}} = \left(\frac{\partial \boldsymbol{\alpha}}{\partial \boldsymbol{v}}(\boldsymbol{r}_k, \boldsymbol{\alpha}_0) \right) \cdot \boldsymbol{F}_p^*(t, \boldsymbol{r}_k, \boldsymbol{v}_k, \boldsymbol{r}, \boldsymbol{v}), \tag{3}$$

con

$$\boldsymbol{F}_p^*(t, \boldsymbol{r}_k, \boldsymbol{v}_k, \boldsymbol{r}, \boldsymbol{v}) = -\mu \left(\frac{\boldsymbol{r}}{r^3} - \frac{\boldsymbol{r}_k}{r_k^3} \right) - \mu \frac{\partial (\boldsymbol{r}_k / r_k^3)}{\partial \boldsymbol{r}_k} \boldsymbol{s} + \boldsymbol{F}_p(\boldsymbol{r}, \dot{\boldsymbol{r}}).$$

Hay que observar que el nuevo sistema a integrar (3) tiene la misma expresión formal que el proporcionado por el método de variación de parámetros clásico, donde, además, si cada cierto tiempo se actualiza la órbita de referencia dada por el problema sin perturbar, las funciones $\boldsymbol{F}_p$ y $\boldsymbol{F}_p^*$ son muy similares. La diferencia fundamental en ambos métodos [2, 3, 8] es que el primer factor del segundo miembro ($\partial \boldsymbol{\alpha} / \partial \boldsymbol{v}$) se evalúa en la órbita

sin perturbar, donde los elementos orbitales no varían, por lo que muchos términos son constantes.

Un problema con el que nos podemos encontrar es la pérdida de dígitos significativos en el cálculo del término

$$\Delta \boldsymbol{F}_0 = -\mu \left(\frac{\boldsymbol{r}}{r^3} - \frac{\boldsymbol{r}_k}{r_k^3} \right) = -\frac{\mu}{r_k^3} \left[\left(\frac{r_k^3}{r^3} - 1 \right) \boldsymbol{r} + \boldsymbol{s} \right],$$

al ser la resta de dos expresiones de magnitudes similares. Para evitar dicho problema podemos hacer uso del siguiente resultado [5].

Lema *Sea* $\boldsymbol{a} = \boldsymbol{b} + \boldsymbol{c}$ *y* a, b, c *sus normas, entonces se tiene que*

$$\left(\frac{b^3}{a^3} - 1 \right) = (1+q)^{\frac{3}{2}} - 1 = q \frac{3 + 3q + q^2}{1 + (1+q)^{\frac{3}{2}}},$$

donde $q = \boldsymbol{c} \cdot (\boldsymbol{c} - 2\boldsymbol{a})/a^2$.

Así, el uso de la expresión

$$\Delta \boldsymbol{F}_0 = -\frac{\mu}{r_k^3} \left(q \frac{3 + 3q + q^2}{1 + (1+q)^{\frac{3}{2}}} \boldsymbol{r} + \boldsymbol{s} \right) \quad \text{con} \quad q = \frac{\boldsymbol{s} \cdot (\boldsymbol{s} - 2\boldsymbol{r})}{r^2},$$

evitará los problemas de redondeo comentados.

5 Tests numéricos

En esta sección presentamos diversos resultados numéricos aplicando las técnicas anteriormente descritas. Todos los tests numéricos han sido realizados en un ordenador Windows PC Pentium III-933Mhz usando `g77` (GNU `FORTRAN77`) y doble precisión. Las órbitas de referencia han sido calculadas usando una tolerancia (`tol` $= 10^{-30}$) en LF95 con precisión extendida.

Como problema test vamos a considerar el movimiento de un satélite artificial perturbado por el potencial terrestre dado por

$$V_T(r, \lambda, \varphi) = -\frac{\mu}{r} - \frac{\mu}{r} \sum_{l=2}^{n} \sum_{m=0}^{l} \left(\frac{R_T}{r} \right)^l P_l^m(\sin \varphi) \left(C_{lm} \cos m\lambda + S_{lm} \sin m\lambda \right)$$

donde (r, λ, φ) son las coordenadas esféricas, R_T el radio medio ecuatorial terrestre y $P_l^m(x)$ las funciones asociadas de Legendre. Los coeficientes C_{lm} y S_{lm} del potencial terrestre siguen el modelo de potencial terrestre GEM9&10 [16].

En primer lugar vamos a comparar los resultados obtenidos cuando un problema es integrado con el mismo integrador numérico pero aplicando tres formulaciones diferentes. La primera consiste en integrar directamente el problema en coordenadas cartesianas

(CART); la segunda se basa en el uso del método clásico de variación de los parámetros (VOP) y, finalmente, la tercera alternativa utiliza la modificación comentada en la sección anterior (VVOP). Como condiciones iniciales vamos a considerar el movimiento del satélite artificial GPSBII-02 (PRN 02) perturbado por el potencial terrestre dado por los cinco primeros harmónicos zonales del potencial terrestre. Las condiciones iniciales de la órbita (sacadas de la página web `http://celestrak.com`) de este satélite son:

$$a = 26559.212356 \text{ km.}, \quad e = 0.02334, \quad i = 53°4247,$$
$$\omega = 261°3417, \quad \Omega = 171°8804, \quad l_0 = 95°9617.$$

El integrador numérico en los tres métodos ha sido el RK DOPRI8(7) [11] a paso constante ya que la baja excentricidad de su órbita no plantea la necesidad de tomar paso variable.

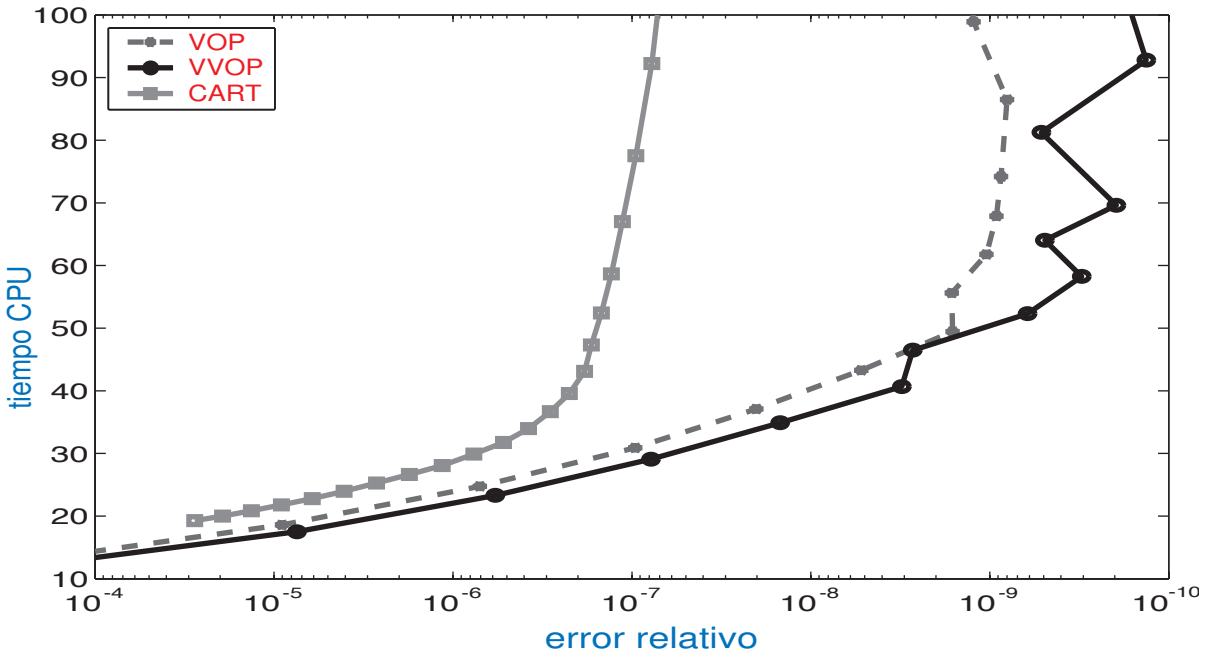


Figura 2.— Análisis de eficiencia de los métodos CART, VOP y VVOP. Tiempo final de integración $T = 10$ años.

El segundo miembro del sistema diferencial para el método CART es mucho mayor que para los otros dos métodos, lo que va a generar una mayor variación de las variables a integrar por el primero frente a VOP y VVOP. La diferencia de variación de estos dos últimos les permitirá usar pasos de integración mucho mayores que los empleados por CART para la misma precisión. En [17], los autores afirman erróneamente que, en general, dicha apreciación es falsa debido al hecho de que las mismas frecuencias aparecen en los tres métodos y que los términos de alta frecuencia controlan el paso de integración. Obviamente, el tamaño del paso tiene que ser inferior a la mayor frecuencia, sin embargo, el método CART necesitará pasos mucho más pequeños que dicha frecuencia, mientras que los otros dos métodos pueden trabajar con pasos de tamaño similar a la misma. De este modo, aunque la complejidad del método CART es inferior a la de los otros dos (en

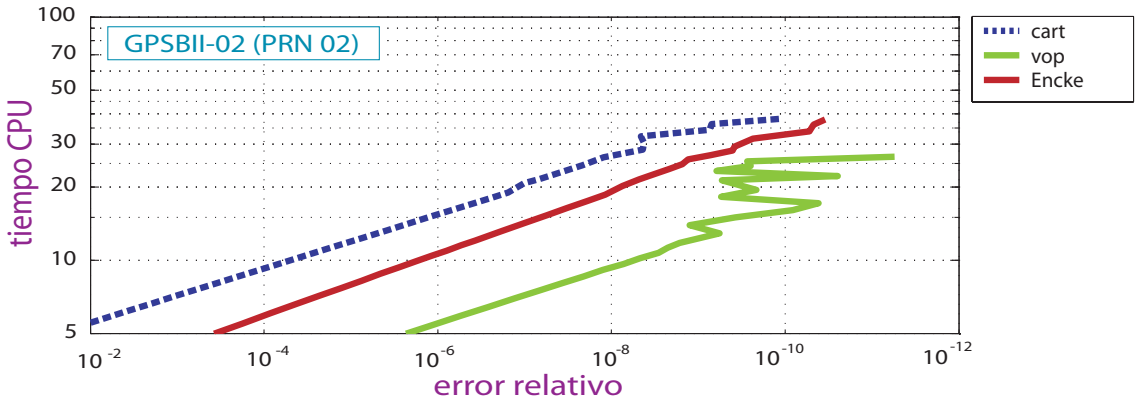


Figura 3.— *Error relativo* vs. *tiempo CPU* en la integración numérica usando el modelo completo 20×20 ($n = 20$) del potencial terrestre y las formulaciones CART, VOP-VVOP y Encke.

promedio, el cociente entre el tiempo de CPU de un paso para VVOP frente a CART es 2.56 y el de VOP frente a CART 2.72), los métodos VOP y VVOP son bastante más eficientes que CART, alcanzando los primeros mayor precisión y a un coste inferior de tiempo de CPU como se observa en la figura 2. Por otro lado, también se observa una ligera mejoría de VVOP frente al VOP clásico.

Por último presentamos en la figura 3 los resultados comparando las formulaciones CART, VOP-VVOP y el método de Encke. En este caso realizamos una comparación usando el modelo de potencial hasta orden 20 en términos zonales y teserales, es decir un modelo completo 20×20. De nuevo se observa que el método de variación de los parámetros presenta el mejor comportamiento de entre las distintas formulaciones analizadas.

Agradecimientos

Este trabajo ha sido financiado a través del proyecto de investigación MTM2009-10767.

Referencias

[1] H. Arakida and T. Fukushima, "Long-term integration error of Kustaanheimo-Stiefel regularized orbital motion I. Method of variation of parameters", Astronom. J. **121**, 1764–1767 (2001).

[2] R. Barrio, M. Palacios and A. Elipe, "Chebyshev collocation methods for fast orbit determination", Appl. Math. Comput. **99**, 195–207 (1999).

[3] R. Barrio and S. Serrano, "Modifications of the method of variation of parameters", Comput. Math. Appl. **51**, no. 3-4, 451–466 (2006).

[4] R. Barrio and S. Serrano, "Performance of perturbation methods on orbit prediction", Math. Comput. Modelling **48**, no. 3-4, 594–600 (2008).

[5] R. H. Battin, *An introduction to the mathematics and methods of astrodynamics* (AIAA Education Series, New York, 1987).

[6] R. Broucke, "Perturbations in rectangular coordinates by iteration", Celestial Mech. **1**, 110–126 (1969).

[7] D. Brouwer, "Integration of the equations of general planetary theory in rectangular coordinates", Astronom. J. **51**, 37–43 (1944).

[8] D. Brouwer and G. M. Clemence, *Methods of Celestial Mechanics* (Academic Press, New York, 1961).

[9] O. Dziobek, *Mathematical theories of planetary motions* (Dover Publications, New York, 1962).

[10] J. F. Encke, "Über die allgemeinen Störungen der Planeten", Berliner Astronomisches Jahrbuch für 1857, 319–397 (1854).

[11] E. Hairer, S. P. Nørsett and G. Wanner, *Solving ordinary differential equations, I. Nonstiff problems* (Springer-Verlag, Berlin, 1987).

[12] K. N. Jayasree and S. G. Deo, "Variation of parameters formula for the equation of Cooke and Wiener", Proc. Amer. Math. Soc. **112**, 75–80 (1991).

[13] S. K. Kaul and X. Liu, "Generalized variation of parameters and stability of impulsive systems", Nonlinear Anal. **40**, 295–307 (2000).

[14] G. L. Kraige and S. B. Skaar, "A variation of parameters approach to the arbitrarily torqued, asymmetric rigid body problem", J. Astronaut. Sci. **25**, 207–226 (1977).

[15] V. Lakshmikantham and S. G. Deo, *Method of variation of parameters for dynamic systems*, Series in Mathematical Analysis and Applications, 1, (Gordon and Breach Science Publishers, Amsterdam, 1998).

[16] F.J. Lerch, S.M. Klosko, R.E. Laubscher and C.A. Wagner, *Gravity Model Improvement using GEOS–3 (GEM9 & 10).* Goddard Space Flight Center, Greenbelt, Maryland (1977).

[17] W. I. Newman and M. Efroimsky, "The method of variation of constants and multiple time scales in orbital mechanics", Chaos **13**, 476–485 (2003).

Monografías de la Real Academia de Ciencias de Zaragoza **33**, 165–176, (2010).

Lagrangians of a non-mechanical type for second order Riccati and Abel equations

José F. Cariñena and Manuel F. Rañada

Departamento de Física Teórica and IUMA, Facultad de Ciencias

Universidad de Zaragoza, 50009 Zaragoza, Spain

Abstract

The Helmholtz approach to the inverse problem of the Lagrangian dynamics is studied first in the particular case of the second-order Riccati equation and then in the case of the second-order Abel equation. The existence of two alternative Lagrangian formulations is proved, both Lagrangians being of a non-natural class (neither potential nor kinetic term). These second-order Riccati and Abel equations are studied by means of their Darboux polynomials and Jacobi last multipliers. The existence of a family of constants of the motion is also discussed.

Keywords: Helmholtz conditions. Jacobi last multipliers. Second order Riccati and Abel equations. Alternative Lagrangians

AMS classification: 34A34; 34A26; 34C14; 37J05; 70H03

On the beginning of his scientific career Prof. Calvo was teaching Lagrangian mechanics for several years and beyond doubt he spent many hours thinking on the Inverse problem of mechanics. We report here several recent results on alternative Lagrangians for two interesting equations, the second order Riccati and Abel equations.

1 Introduction

In mathematical terms, the Newtonian approach to classical mechanics, constructed on the use of the second Newton Law, states that the behaviour of a mechanical system is governed by second-order differential equations. On the other side, the Lagrangian approach makes use of a variational formulation associated to the Hamilton's principle:

the motions of the system are those making extremal the action integral defined by the Lagrange function L of the system; therefore such trajectories are solutions of the corresponding set of Euler-Lagrange equations.

The inverse problem of Lagrangian dynamics is to obtain necessary and sufficient conditions for a system of second-order differential equations

$$\ddot{q}^i = F^i(q, \dot{q}), \quad i = 1, \ldots, n, \tag{1}$$

to be equivalent to the set of Euler-Lagrange equations of some regular Lagrangian function L. In other words, this amounts to look for a Lagrangian L such that

$$W_{ij}(\ddot{q}^j - F^j(q, \dot{q})) = W_{ij}\ddot{q}^j - \frac{\partial L}{\partial q^i} + \frac{\partial^2 L}{\partial q^j \, \partial \dot{q}^i}\dot{q}^j, \qquad i, j = 1, \ldots, n.$$

where the summation convention is used and W is the Hessian matrix with elements defined by

$$W_{ij} = \frac{\partial^2 L}{\partial \dot{q}^i \, \partial \dot{q}^j}, \quad i, j = 1, \ldots, n. \tag{2}$$

Such a function L only exists if some compatibility conditions hold; moreover, in some particular cases it can even exist several different (and non gauge equivalent) solutions. In these cases the alternative Lagrangians can be used to construct constants of the motion as was proved in [1] for the one-dimensional case and generalised in [2] for the multidimensional case (see also [3] for a geometric approach).

The aim of the paper is to show that these properties are related with a rather old result by Jacobi [4] which is called the theory of Jacobi Last Multiplier and to illustrate the theory with the determination of Lagrangians of non-mechanical type for second-order Riccati and Abel equations.

The organization of the paper is as follows: in section 2 we give a concise survey of the theory of the Inverse problem in mechanics and the Helmholtz conditions necessary for the existence of a Lagrangian. In section 3 we recall several notions of Darboux polynomials for polynomial vector fields and the relation with the theory of Jacobi Last Multiplier. The relevance of such multipliers in the search for Lagrangians for one-dimensional systems is shown in section 4 and the theory is illustrated with some particular examples. The method of Jacobi last ultipliers is used to determine alternative Lagrangians for the second order Riccati equation and the second order Abel equation.

2 The inverse problem and the Helmholtz conditions

From the geometric viewpoint, the system of equations (1) determines a $2n$-dimensional vector field Γ on the velocity phase space $\mathbb{R}^{2n} = \{(q^i, v^i) \mid i = 1, \ldots, n\}$ with $(v^i, F^i(q, v))$ as components, i.e.

$$\Gamma = v^i \frac{\partial}{\partial q^i} + F^i(q, v)\frac{\partial}{\partial v^i}, \tag{3}$$

the solutions of the system being the integral curves of Γ (the physical time t coincides with the parameter of the curves).

The solution of the Inverse Problem is given by a family of functions $g_{ij}(q,\dot{q})$ such that the equations

$$g_{ij}\left(\ddot{q}^j - F^j(q,\dot{q})\right) = 0, \qquad i,j = 1,\ldots,n,$$

become the Euler-Lagrange equations of some function L. The problem was studied long time ago by Helmholtz [5], who established the so-called Helmholtz conditions (see e.g. [6, 7, 8, 9]). In terms of the n functions F^i, these conditions can be presented as follows: there should exist a non-degenerate symmetric matrix valued function $g = [g_{ij}]$, i.e. a family of $n(n+1)/2$ functions $g_{ij} = g_{ij}(q,v)$, such that

i) $\det[g_{ij}] \neq 0$

ii) $\dfrac{\partial g_{ij}}{\partial v^k} = \dfrac{\partial g_{ik}}{\partial v^j}$

iii) $\Gamma(g_{ij}) + \dfrac{1}{2}g_{kj}\dfrac{\partial F^k}{\partial v^i} + \dfrac{1}{2}g_{ik}\dfrac{\partial F^k}{\partial v^j} = 0$

iv) $g_{ik}\left[\dfrac{\partial F^k}{\partial q^j} + \dfrac{1}{4}\dfrac{\partial F^k}{\partial v^l}\dfrac{\partial F^l}{\partial v^j} - \dfrac{1}{2}\Gamma\left(\dfrac{\partial F^k}{\partial v^j}\right)\right] = g_{jk}\left[\dfrac{\partial F^k}{\partial q^i} + \dfrac{1}{4}\dfrac{\partial F^k}{\partial v^l}\dfrac{\partial F^l}{\partial v^i} - \dfrac{1}{2}\Gamma\left(\dfrac{\partial F^k}{\partial v^i}\right)\right].$

These properties lead to the existence of a function L such that the g_{ij} take the form

$$g_{ij} = \frac{\partial^2 L}{\partial v^i\,\partial v^j}\,.$$

The regularity of L is a consequence of condition i). Although we are not concerned in this article with the abstract geometric formalism, at this point we make the observation that these conditions are related to the existence of a symplectic structure in the phase space.

These equations involve not only the known functions $F^i(q,v)$ but also the other functions $g_{ij}(q,v)$ whose form is completely unknown. Usually the problem is solved by using an ansatz on the components of the matrix $[g_{ij}]$. Moreover, these equations do not guarantee uniqueness: every set of $n(n+1)/2$ functions $g_{ij}(q,v)$ satisfying Eqs. i)-iv) leads to a particular Lagrangian. Thus, different sets of such functions $g^{(a)}_{ij}(q,v)$, $a = 1,\ldots,A$, give rise to different alternative Lagrangians $L^{(a)}$.

Hojman and Harleston proved [2] that if a system admits two alternative regular Lagrangians $L^{(a)}$, $a = 1,2$, then the traces of the powers of the product matrix $W_{21} = W_2^{-1}W_1$ are constants of motion which are obtained by a non-Noether procedure [3].

Consequently, the existence of alternative Lagrangians for a certain Lagrangian system also means the existence of a certain set of integrals of motion.

Two situations are particularly interesting. First, in the case of velocity-independent

forces, that is $F^i = F^i(q)$, the last three Helmholtz conditions reduce to

$$\begin{aligned} &\text{ii b)} \quad \frac{\partial g_{ij}}{\partial v^k} = \frac{\partial g_{ik}}{\partial v^j} \\ &\text{iii b)} \quad \Gamma(g_{ij}) = 0 \\ &\text{iv b)} \quad g_{ik}\left(\frac{\partial F^k}{\partial q^j}\right) = g_{jk}\left(\frac{\partial F^k}{\partial q^i}\right). \end{aligned}$$

We note that the equation iii b) means that the functions g_{ij} must be constants of the motion for the dynamics.

Secondly, the case of only one equation corresponding to a one degree of freedom system. In this case the matrix reduces to a function g and only one condition remains:

$$\Gamma(g) + g\,\frac{\partial F}{\partial v} = 0, \tag{4}$$

which will be shown to be that g is a Jacobi last multiplier, a concept we recall next.

3 Jacobi Last Multipliers

Given a vector field X in an oriented manifold (M, Ω), a function R such that $R\,i(X)\Omega$ is closed is said to be a Jacobi last multiplier (JLM) for X. Recall that the divergence of the vector field X (with respect to the volume form Ω) is defined by the relation

$$\mathcal{L}_X\Omega = (\operatorname{div} X)\,\Omega\,.$$

This means that R is a multiplier if and only if $R\,X$ is a divergence-less vector field and then

$$\mathcal{L}_{RX}\Omega = (\operatorname{div} RX)\,\Omega = [X(R) + R\operatorname{div}X]\,\Omega = 0\,,$$

and therefore we see that R is a last multiplier for X if and only if

$$X(R) + R\operatorname{div}X = 0\,. \tag{5}$$

Moreover, $f\,R$ is also a JLM iff f is a constant of the motion for X, because, for any function f,

$$X(f\,R) + f\,R\operatorname{div} X = (X f)\,R + f(X(R) + R\operatorname{div} X).$$

In the particular case of X being the vector field Γ given by (3) in the velocity phase space corresponding to the second order differential equation (1) with $n = 1$, as $\operatorname{div}\Gamma = \partial F/\partial v$, (5) becomes

$$v\,\frac{\partial R}{\partial q} + F(q,v)\,\frac{\partial R}{\partial v} + R(q,v)\,\frac{\partial F}{\partial v} = 0, \tag{6}$$

that when compared with (4) means that g is a JLM for Γ.

There is a method due to Darboux for finding JLM for polynomial vector fields. We recall that a polynomial function $\mathcal{D} : U \to \mathbb{R}$ is a Darboux polynomial for a polynomial

vector field X if there is a polynomial function f defined in U such that $X\mathcal{D} = f\mathcal{D}$ [10]. The function f is said to be the cofactor corresponding to such Darboux polynomial and the pair $(f, \mathcal{D})$ is called a Darboux pair.

The remarkable point is that if $\mathcal{D}_1, \ldots, \mathcal{D}_k$, are Darboux polynomials with corresponding cofactors f_i, $i = 1, \ldots, k$, one can look for multiplier factors of the form

$$R = \prod_{i=1}^{k} \mathcal{D}_i^{\nu_i} \tag{7}$$

and then

$$\frac{X(R)}{R} = \sum_{i=1}^{k} \nu_i \frac{X(\mathcal{D}_i)}{\mathcal{D}_i} = \sum_{i=1}^{k} \nu_i \, f_i \,,$$

and therefore, if the coefficients ν_i can be chosen such that

$$\sum_{i=1}^{k} \nu_i \, f_i = -\mathrm{div}\, X \tag{8}$$

holds, then we arrive to

$$\frac{X(R)}{R} = \sum_{i=1}^{k} \nu_i \, f_i = -\mathrm{div}\, X \,,$$

and consequently R is a Jacobi last multiplier for X.

4 Looking for Lagrangians from Jacobi last multipliers

The main point of the relation between Jacobi last multipliers and Lagrangian formulations for an autonomous second-order differential equation is given in the following theorem:

Theorem 1 *The normal form of the differential equation determining the solutions of the Euler-Lagrange equation defined by the regular Lagrangian function $L(y, v)$ admits the function*

$$R = \frac{\partial^2 L}{\partial v^2}. \tag{9}$$

as a Jacobi last multiplier. Conversely, if $R(y, v)$ is a multiplier function for a second order differential equation in normal form, then there exists a Lagrangian L for the system that is related to R by (9).

Proof: In fact, if there exists a Lagrangian function L, the function F is given by

$$F(q, v) = \frac{1}{R}\left(\frac{\partial L}{\partial q} - v\frac{\partial^2 L}{\partial q \partial v}\right) .$$

where the function R is given by (9). We can now see that such function R satisfy condition (4). In fact, it turns out to be

$$v\frac{\partial R}{\partial q} + \frac{\partial}{\partial v}\left(\frac{\partial L}{\partial q} - v\frac{\partial^2 L}{\partial q \partial v}\right) = v\frac{\partial^3 L}{\partial v^2 \partial q} + \frac{\partial^2 L}{\partial q \partial v} - \frac{\partial^2 L}{\partial q \partial v} - v\frac{\partial^3 L}{\partial v^2 \partial q} = 0 \,.$$

Conversely, let R be a function satisfying (6). Then, let L be a function such that condition (9) be satisfied, i.e. L is such that

$$\frac{\partial L}{\partial v} = \int^v R(q,\zeta)\, d\zeta + \phi_1(q)\,, \tag{10}$$

and then,

$$L(q,v) = \int^v dv' \int^{v'} R(q,\zeta)\, d\zeta + \phi_1(q)\, v + \phi_2(q)\,. \tag{11}$$

We can see that there exists a function $\phi_2(y)$ such that, for any choice of the function $\phi_1(y)$, the Euler Lagrange equation for such a Lagrangian function gives rise to the equation of motion. The function $\phi_2(y)$ is uniquely determined, while the other function is arbitrary because it corresponds to a gauge term. In fact, the Euler-Lagrange equation for (11), taking into account (10) and that

$$\frac{\partial L}{\partial q} = \int^v dv' \int^{v'} \left(\frac{\partial}{\partial q} R(q,\zeta)\right) d\zeta + \phi_1'(q) v + \phi_2'(q)\,,$$

turns out to be:

$$\int^v dv' \int^{v'} \left(\frac{\partial}{\partial q} R(q,\zeta)\right) d\zeta + \phi_1'(q) v + \phi_2'(q) = v \int^v \left(\frac{\partial}{\partial q} R(q,\zeta)\right) d\zeta + R(q,v) F(q,v) + \phi_1'(q) v\,.$$

Note that

$$\begin{aligned} &\frac{\partial}{\partial v}\left(v \int^v \left(\frac{\partial}{\partial q} R(q,\zeta)\right) d\zeta + R(q,v) F(q,v) - \int^v dv' \int^{v'} \left(\frac{\partial}{\partial q} R(q,\zeta)\right) d\zeta\right) \\ &\qquad = v \frac{\partial R}{\partial q}(q,v) + F(q,v) \frac{\partial R}{\partial v}(q,v) + R(q,v) \frac{\partial F}{\partial v}(q,v)\,, \end{aligned}$$

which vanishes because of the multiplier condition (5) for R and $X = \Gamma$. Then, the function $\phi_2(q)$ is uniquely determined, up to a constant, by

$$\phi_2'(q) = v \int^v \left(\frac{\partial}{\partial q} R(q,\zeta)\right) d\zeta + R(q,v) F(q,v) - \int^v dv' \int^{v'} \left(\frac{\partial}{\partial q} R(q,\zeta)\right) d\zeta. \tag{12}$$

The term $\phi_1(y)\, v$ is a gauge term which can be eliminated and then the expression (11) for the Lagrangian reduces to

$$L(y,v) = \int^v dv' \int^{v'} R(y,\zeta)\, d\zeta + \phi_2(y)\,. \tag{13}$$

Corollary 1 *If a system with one degree of freedom admits two different regular Lagrangians then the function f defined by*

$$f\, \frac{\partial^2 L_1}{\partial v^2} = \frac{\partial^2 L_2}{\partial v^2} \tag{14}$$

is a constant of the motion.

Proof: The function f is the quotient of two different Jacobi last multipliers.

The last result is usually attributed to Currie and Saletan [1], but actually data back to Jacobi's time. Conversely, if L_1 is an admissible Lagrangian and f is a constant of the motion, there is an alternative Lagrangian L_2 such that relation (14) holds.

Then, the inverse problem for one-dimensional systems reduces to find the function g which is a Jacobi last multiplier and L is obtained by integrating the function g two times with respect to velocities. The function L so obtained from g is unique up to addition of a gauge term.

As an instance, Jacobi derived in [11] the multiplier $R = e^{\varphi(y)}$ for a family of differential equation of the form

$$y'' + \frac{1}{2}\frac{d\varphi}{dy}\, y'^2 + \psi(y) = 0\,,$$

which includes as a particular instance the equation studied in [12]. In fact, one can see that such a function R satisfies (6), because now $\log R = \varphi(y)$ and $F(y,v) = -\frac{1}{2}\varphi'(y)\, v^2 - \psi(y)$, and then

$$\frac{\partial F}{\partial v} = -\frac{d\varphi}{dy}\, v = -\frac{d}{dx}\log R\,.$$

The corresponding Lagrangian is given, up to a gauge term, by

$$L(y,v) = \frac{1}{2}e^{\varphi(y)}v^2 + f(y)\,,$$

where f satisfies the following equation:

$$\frac{df}{dy} + e^{\varphi(y)}\psi(y) = 0\,,$$

because using (11) in the form (13) we find that

$$L = \int_0^v v' e^{\varphi(y)}\, dv' + \phi_2(y)$$

with ϕ_2 being determined by (12), i.e.

$$L(y,v) = \frac{1}{2}e^{\varphi(y)}v^2 + f(y)\,,$$

with $f(y)$ such that the previous equation holds.

For instance, the evolution equation

$$\ddot{x} = F(x,\dot{x}) = \frac{-a + \lambda\dot{x}^2}{1+\lambda\, x^2}\,, \quad a,\lambda \in \mathbb{R},$$

which corresponds to a nonlinear oscillator [13, 14], is a particular example for which

$$\frac{1}{2}\frac{d\varphi}{dx} = -\frac{\lambda\, x}{1+\lambda\, x^2}\,, \qquad \psi(x) = \frac{a\, x}{1+\lambda\, x^2}\,,$$

and therefore,

$$\varphi(x) = -\log(1+\lambda\, x^2)\,.$$

The Jacobi last multiplier is then

$$R = e^{\varphi(x)} = \frac{1}{1+\lambda\, x^2}\,,$$

and the corresponding Lagrangian turns out to be:

$$L(x,\dot{x}) = \frac{1}{2}\,\frac{\dot{x}^2}{1+\lambda\, x^2} - \frac{1}{2}\,\frac{a\, x^2}{1+\lambda\, x^2}\,.$$

5 Alternative Lagrangians for second-order Riccati and Abel equations

5.1 Second-order Riccati equation

Riccati differential equation is a nonlinear generalization of the inhomogeneous linear equation

$$\dot{x} = c_0(t) + c_1(t)\, x + c_2(t)\, x^2\,,$$

which is just the particular case corresponding to $c_2(t) \equiv 0$. It appears in the Lie reduction process when taking into account invariance under dilations, $u \mapsto \lambda\, u$, of second-order linear equations. The infinitesimal generator of such transformations is the Liouville vector field $\Delta = u\,\partial/\partial u$. Lie's recipe for order reduction of differential equations with symmetry consists on changing the dependent variable in such a way that $\Delta = \partial/\partial w$. More specifically, $u = e^w$, and under such change of dependent variable, as

$$\dot{u} = e^w\,\dot{w}\,, \qquad \ddot{u} = e^w(\dot{w}^2 + \ddot{w})\,.$$

Then the second-order linear differential equation $\ddot{u} + d_0\,\dot{u} + d_1\, u = 0$ becomes a Riccati differential equation for the function $\dot{w} = \dot{u}/u$:

$$\ddot{w} + \dot{w}^2 + d_0\,\dot{w} + d_1 = 0\,.$$

We call higher-order Riccati equations those appearing in a reduction process from a linear differential equations by using dilation invariance: the linear $(j+1)$-order differential equation $y^{(j+1)} = 0$ gives rise to a j-order Riccati equation. For instance the third order linear equation $y''' = 0$ defines the second order Riccati equation

$$\ddot{x} + 3\, x\,\dot{x} + x^3 = 0\,. \tag{15}$$

More specifically, the invariance under dilations of the differential equation $y^{(n)} = 0$, according to Lie recipe, suggests to look for a new variable z such that the dilation vector field $y\,\partial/\partial y$ becomes $\partial/\partial z$. Then $y = e^z$, up to an irrelevant factor.

It has been proved in [15] that the differential equation $y^{(n)} = 0$ becomes $R^{(n-1)}(x) = 0$ with $x = \dot{z}$, where $R^{(j)}(x)$ is defined in an iterative way by

$$R^{(j)}(x, \ldots, x^{(j)}) = \mathbb{D}^j x\,, \quad j = 0, 1, \ldots,$$

with

$$\mathbb{D} = \frac{d}{dt} + x\,.$$

It has been pointed out in [16] that the second-order Riccati equation (15) admits a Lagrangian formulation with the function

$$L = \frac{1}{\dot{x} + x^2}\ .$$

as a (non-standard) Lagrangian. Our aim in this section is to rederive this Lagrangian by means of the associated Darboux polynomials. The differential equation (15) has an associated system of differential equations

$$\left\{ \begin{array}{rcl} \dot{x} & = & v \\ \dot{v} & = & -3xv - x^3 \end{array} \right.$$

which determines the integral curves of the vector field

$$\Gamma^{(1)} = v\frac{\partial}{\partial x} - (3xv + x^3)\frac{\partial}{\partial v}\,.$$

We can look for a Darboux polynomial of the form

$$\mathcal{D}(x, v) = v + ax^2.$$

The condition $\Gamma^{(1)}\,\mathcal{D} = f\,\mathcal{D}$ implies first that f must be of the form $f(x, v) = (2a - 3)x$ and then that $-x = af = (2a - 3)ax$.

Therefore there are two solutions corresponding to the two roots of $2a^2 - 3a + 1 = 0$: either a must be equal to 1, and then the corresponding cofactor is $f_1 = -x$, or to $1/2$ with associated cofactor $f_{1/2} = -2x$.

In the first case, with $\mathcal{D}_1(x, v) = v + x^2$, we can choose in (8) $\nu_1 = -3$ because $\nu_1\, f_1 = 3x = -\mathrm{div}\,\Gamma^{(1)}$, and consequently we arrive to

$$R_1(x, v) = L_1^3(x, v) = \frac{1}{(v + x^2)^3}.$$

In the second case, we have $\mathcal{D}_2(x, v) = v + \frac{1}{2}x^2$, and we can choose in (8) $\nu_2 = -\frac{3}{2}$, so that we obtain

$$R_2 = \left(v + \frac{1}{2}x^2\right)^{-3/2}$$

as another Jacobi last multiplier.

¿From the first Jacobi last multiplier R_1 we obtain that the vector field $\Gamma^{(1)}$ is the Euler-Lagrange vector field of a Lagrangian L that just coincides with L_1. The alternative Lagrangian obtained from R_2 is:

$$L'(x, v) = \sqrt{v + \frac{1}{2}x^2}.$$

5.2 Second-order Abel equation

The Abel equation of first order,

$$\dot{x} = A_0(t) + A_1(t)x + A_2(t)x^2 + A_3(t)x^3,$$

is a generalisation of the Riccati equation (that appears as the particular case $A_3 = 0$).

In similarity with the Ricati case, let us define the differential operator

$$\mathbb{D}_A = \frac{d}{dt} + x^2(t)\,,$$

in such a way that iterating the action of $\mathbb{D}_A$ on x leads to the family of differential equations

$$\mathbb{D}_A^m x = 0\,, \quad m = 1, 2, 3, \ldots\,.$$

The three first equations in this hierarchy of higher-order Abel equations are given by $\mathbb{D}_A^0 x = 0$, $\mathbb{D}_A x = 0$, and $\mathbb{D}_A^2 x = 0$, with $\mathbb{D}_A^0 x$, $\mathbb{D}_A x$, and $\mathbb{D}_A^2 x$ given by

$$\begin{aligned}
\mathbb{D}_A^0 x &= x \\
\mathbb{D}_A x &= \left(\frac{d}{dt} + x^2\right) x = \dot{x} + x^3 \\
\mathbb{D}_A^2 x &= \left(\frac{d}{dt} + x^2\right)^2 x = \ddot{x} + 4x^2\dot{x} + x^5
\end{aligned}$$

The second-order Abel equation $\mathbb{D}_A^2 x = 0$ so obtained can be presented as a system of two first-order equations

$$\left\{ \begin{array}{rcl} \dfrac{dx}{dt} &=& v \\ \dfrac{dv}{dt} &=& -4\,x^2 v - x^5 \end{array} \right. \tag{16}$$

corresponding to the following vector field on the velocity phase space $\mathbb{R}^2$

$$\Gamma^{(2)} = v\,\frac{\partial}{\partial x} - \left(4\,x^2 v + x^5\right)\,\frac{\partial}{\partial v}\,.$$

In this case the polynomial $\mathcal{D}_1$ defined by

$$\mathcal{D}_1(x, v) = v + x^3$$

is a Darboux polynomial for $\Gamma^{(2)}$ with cofactor $-x^2$ since

$$\left(v\,\frac{\partial}{\partial x} - \left(4\,x^2 v + x^5\right)\,\frac{\partial}{\partial v}\right)(v + x^3) = -x^2(v + x^3)\,.$$

The divergence of the vector field $\Gamma^{(2)}$ is $-4x^2$, and then we see that there is a Jacobi last multiplier of the form

$$R = \mathcal{D}_1^{-4}\,.$$

Consequently, the Abel equation admits a Lagrangian description by means of a function L such that

$$\frac{\partial^2 L}{\partial v^2} = \frac{1}{(v+x^3)^4}\,,$$

from where we obtain the Lagrangian $L = L_A$ given by

$$L_A = \frac{1}{(v+x^3)^2}\,. \tag{17}$$

The polynomial $\mathcal{D}_2$ defined by

$$\mathcal{D}_2(x,v) = 3v + x^3$$

is a Darboux polynomial for $\Gamma^{(2)}$ with cofactor $-3x^2$, because

$$\left(v\,\frac{\partial}{\partial x} - \left(4\,x^2 v + x^5\right)\,\frac{\partial}{\partial v}\right)(3v+x^3) = 3\,x^2 v - 3(4\,x^2 v + x^5) = -3\,x^2(3v+x^3)\,,$$

and then we can find another Jacobi last multiplier of the form $\mathcal{D}_2^{\nu_2}$ with $\nu_2 = -4/3$. Therefore the Abel equation admits a Lagrangian description by means of a second function L such that

$$\frac{\partial^2 L}{\partial v^2} = (3\,v + x^3)^{-4/3}\,, \tag{18}$$

from where we obtain the Lagrangian $L = \widetilde{L}_A$ given by

$$\widetilde{L}_A = (3v + x^3)^{2/3}\,.$$

Acknowledgments

Partial financial support by research projects MTM2009-11.154 and E 24/1 (DGA) is acknowledged

References

[1] D.G. Currie and E.J. Saletan, "q-equivalent particle Hamiltonians. The classical one-dimensional case", J. Math. Phys. **7**, 967–974 (1966).

[2] S. Hojman and H. Harleston, "Equivalent Lagrangians: multidimensional case", J. Math. Phys. **22**, 1414–19 (1981).

[3] J.F. Cariñena and L.A. Ibort, "Non-Noether constants of motion", J. Phys. A: Math. Gen. **16**, 1–7 (1983).

[4] M. Jacobi, "Sur le principe du dernier multiplicateur et sur son usage comme nouveau principe général de mécanique", J. Math. Pures et Appl. **10**, 337-46 (1845).

[5] H. Helmholtz, "Uber die physikalische bedeutung des princips der kleinsten wirkung", J. Reine Angew. Math. **100**, 137-166 (1887).

[6] J. Lopuszanski, *The inverse variational problem in classical mechanics*, World Scientific Publishing, 1999.

[7] M. Crampin, "On the differential geometry of the Euler-Lagrange equations, and the inverse problem of Lagrangian dynamics", J. Phys. A: Math. Gen. **14**, 2567–2575 (1981).

[8] W. Sarlet, "The Helmholtz conditions revisited. A new approach to the inverse problem", J. Phys. A: Math. Gen. **15**, 1503–1517 (1982).

[9] G. Morandi, C. Ferrario, G. Lo Vecchio, G. Marmo and C. Rubano, "The inverse problem in the calculus of variations and the geometry of the tangent bundle", Phys. Rep. **188**, 147–284 (1990).

[10] G. Darboux, "Mémoire sur les équations différentielles algébriques du premier ordre et du premier degré", Bull. Sci. Math. (2) **2**, 60–96, 123–144, 151–200 (1878).

[11] C.G.J. Jacobi, "Theoria novi multiplicatoris systemati aequationum differentialium vulgarium applicandi", J. Reine Angew. Math. (Crelle J.) **29**, 213–279, 333–376 (1845).

[12] Z.E. Musielak, "Standard and non-standard Lagrangians for dissipative dynamical systems with variable coefficients", J. Phys. A: Math. Theor. **41**, 055205 (2008).

[13] P.M. Mathews and M. Lakshmanan, "On a unique nonlinear oscillator", Quart. Appl. Math. **32**, 215–218 (1974).

[14] J.F. Cariñena, M.F. Rañada, M. Santander and M. Senthilvelan, "A nonlinear oscillator with quasi-harmonic behaviour: two- and n-dimensional oscillators", Nonlinearity **17**, 1941–1963 (2004).

[15] J.F. Cariñena, P. Guha and M.F. Rañada, "A geometric approach to higher-order Riccati chain: Darboux polynomials and constants of the motion", Workshop on Higher Symmetries (Madrid, Spain, 2008) J. Phys. Conf. Ser. **175**, 012009 (2009).

[16] J.F. Cariñena, M.F. Rañada and M. Santander, "Lagrangian formalism for nonlinear second-order Riccati systems: one-dimensional integrability and two-dimensional superintegrability", J. Math. Phys. **46**, 062703 (2005).

[17] J.F. Cariñena, P. Guha and M.F. Rañada, "Higher-order Abel equations: Lagrangian formalism, first integrals and Darboux polynomials", Nonlinearity **22**, 2953–2269 (2009).

Monografías de la Real Academia de Ciencias de Zaragoza **33**, 177–196, (2010).

On the computation of symmetric Szegő-type quadrature formulas

A. Bultheel

Department of Computer Science, K.U.Leuven
Celestijnenlaan 200 A, B-3001 Heverlee, Belgium.

and

R. Cruz-Barroso, P. González-Vera, F. Perdomo-Pío

Department of Mathematical Analysis, La Laguna University
38271 La Laguna, Tenerife, Canary Islands, Spain

"*Dedicated to Prof. Manuel Calvo Pinilla for his 65-th birthday and as an acknowledgement of his decisive contribution to the development of Numerical Analysis in La Laguna University*".

Abstract

By $z = e^{i\theta}$ and $x = \cos\theta$, one may relate $x \in I = (-1, 1]$, with $\theta \in (-\pi, \pi]$ and a point z on the complex unit circle $\mathbb{T}$. Hence there is a connection between the integrals of 2π-periodic functions, integrals of functions over I and over $\mathbb{T}$. The well known Gauss quadratures approximate the integrals over I and their circle counterparts are the Szegő quadratures. When none, one or both endpoints of I are added to the usual Gauss nodes, one obtains the Gauss-type (Radau and Lobatto) quadratures. The circular counterparts are called Szegő-type quadratures. If the integrand and the weight function are symmetric for upper and lower half of $\mathbb{T}$, the choice of complex conjugate Szegő nodes with equal weights seems to be natural, and in that case, the Gauss nodes in I are just the projections of the Szegő nodes. Also the weights are related, and it becomes numerically interesting to compute the Szegő quadrature from the corresponding Gauss quadrature which reduces the computational cost considerably. Especially when the weights and nodes are computed via an eigenvalue problem, which for Gauss works with a tri-diagonal Jacobi matrix, but requires an upper Hessenberg matrix in the Szegő case.

Key words: Gauss-type quadrature formulas, Jacobi matrices, Szegő-type quadrature formulas, Hessenberg matrices, symmetric weight functions, Rogers-Szegő q-polynomials.

1 Introduction

Since the publication in 1989 of the paper [35] by W.B. Jones, O. Njåstad and W.J. Thron along with the recent works by B. Simon [39]-[42] among others (see also some of the most relevant contributions of the "Spanish Mathematical Community" on Orthogonal Polynomials, e.g. [1], [3]-[4], [10]-[12], [23] or [26]), the theory of orthogonal polynomials on the unit circle introduced by Szegő in [45] has become an interesting research topic both from a theoretical and from an applied point of view. In this respect, when dealing with the approximate calculation of a weighted integral of a 2π-periodic function or more generally a weighted integral over the unit circle, the so-called Szegő quadrature formulas introduced in [35] (see also [13], [29, Chapter 4], [30]-[31] and [44]) appear and represent the analog on the unit circle of the Gaussian Formulas. As it is known, a fundamental aspect of a family of quadrature rules is the efficient computation of its nodes and weights. Thus, the computation of the Gaussian formulas leads to an eigenvalue problem involving certain tri-diagonal (Jacobi) matrices meanwhile the Szegő formula can be efficiently computed in terms of an eigenvalue problem involving certain Hessenberg matrices [30]-[31] (see also [11] and [14] for an alternative approach).

In this paper, we will be mainly concerned with the computation of the Szegő formulas when both the weight function in the integral and the nodes in the quadrature rules satisfy symmetry properties. For this purpose, the well known connection between the theory of Orthogonal Polynomials on the unit circle and the real line will be used in order to drastically reduce the computational effort of such rules.

Thus, in order to make the paper self-contained it has been organized as follows: Sections 2 and 3 are dedicated to collect some preliminary results concerning the most relevant aspects of both Gaussian and Szegő formulas. In Section 4 the above symmetry properties are exposed and the characterization of the corresponding symmetric Szegő-type quadrature formulas deduced. The computation features are given in Section 5 meanwhile some numerical illustrative experiments are finally carried out in Section 6.

2 Preliminary results: Jacobi matrices and Gauss-type formulas

Given the integral,

$$I_\sigma(f) = \int_a^b f(x)\sigma(x)dx, \tag{1}$$

σ being a weight function on [a,b], by an n-point Gaussian formula $I_n(f) = \sum_{j=1}^n A_j f(x_j)$ for $I_\sigma(f)$ or σ we mean a quadrature formula so that $I_\sigma(P) = I_n(P)$ for any polynomial $P \in \mathcal{P}_{2n-1}$; in the sequel, $\mathcal{P}_k$ denotes the space of polynomials of degree less than or equal to k and $\mathcal{P}$ the space of all polynomials i.e., $\mathcal{P} = \cup_{k=0}^\infty \mathcal{P}_k$. A characterization of these rules is given in the following result (see e.g. [36, pp. 101-103] and [45, Theorem 3.4.2]),

Theorem 2.1 *Let $\{Q_k\}_{k=0}^\infty$ be the sequence of orthonormal polynomials for σ. Then, $I_n^\sigma(f) = \sum_{j=1}^n A_j f(x_j)$ is the n-point Gaussian formula for $I_\sigma(f)$, if and only if,*

1. *$\{x_j\}_{j=1}^n$ are the zeros of any orthogonal polynomial of degree n with respect to σ.*
2. *$A_j = \left(\sum_{k=0}^{n-1} |Q_k(x_j)|^2 \right)^{-1} > 0$, for all $j = 1, \ldots, n$ (Christoffel numbers).*

$I_n^\sigma(f)$ as given in Theorem 2.1 is *optimal* in the sense there exists $P \in \mathcal{P}_{2n}$ such that $I_n^\sigma(P) \neq I_\sigma(P)$.

On the other hand, efficient computation of the weights and nodes for $I_n^\sigma(f)$ has been carried out by means of the so-called Jacobi matrices associated with the three-term recurrence relation satisfied by the sequence $\{Q_k\}_{k=0}^\infty$. Indeed, it is known that it holds,

$$xQ_n(x) = a_{n+1}Q_{n+1}(x) + b_n Q_n(x) + a_n Q_{n-1}(x), \quad n \geq 0, \quad Q_{-1} \equiv 0,$$

so that by setting,

$$\mathcal{J} = \begin{pmatrix} b_0 & a_1 & 0 & 0 & \cdots \\ a_1 & b_1 & a_2 & 0 & \cdots \\ 0 & a_2 & b_2 & a_3 & \cdots \\ 0 & 0 & a_3 & b_3 & \cdots \\ \vdots & \vdots & \vdots & \vdots & \ddots \end{pmatrix} \tag{2}$$

then, the eigenvalues of the n-th truncation of the matrix $\mathcal{J}$ give us the set of nodes $\{x_j\}_{j=1}^n$ in $I_n^\sigma(f)$ and the square of the first component of the eigenvector of unit length corresponding to the eigenvalue x_j yields the weight A_j, for $j = 1, \ldots, n$. As seen, in a Gaussian formula no freedom is left to fix some nodes in advance so that the remaining nodes and weights can be chosen to produce quadrature formulas with similar features to the Gaussian ones, that is, positive weights and exactly integrating polynomials with as high degree as possible. This kind of quadratures which are of a great interest in the construction of methods to numerically solve differential and integral equations, have been studied in the last decades producing satisfactory results only in a few particular cases . Thus, in the simplest situation when $[a, b]$ is finite, say $[a, b] = [-1, 1]$, quadrature formulas with prescribed nodes at ± 1 can be constructed and exhibiting similar characteristics to the Gaussian ones. These are the so-called Gauss-Radau and Gauss-Lobatto formulas as

summarized in the following (for a more general situation see the recent papers [7] and [15]),

Theorem 2.2 *Given the integral $I_\sigma(f)$ in (1) and $r, s \in \{0, 1\}$, consider the n-point quadrature rule:*

$$I_n^{r,s}(f) = rA_n^+ f(1) + sA_n^- f(-1) + \sum_{j=1}^{n-r-s} A_j^{r,s} f(x_j^{r,s}).$$

Then $I_n^{r,s}(P) = I_\sigma(P)$, for all $P \in \mathcal{P}_{2n-1-r-s}$, if and only if,

1. *$I_n^{r,s}(P) = I_\sigma(P)$, for all $P \in \mathcal{P}_{n-1}$ (that is, it is of interpolatory type).*
2. *The nodes $\{x_j^{r,s}\}_{j=1}^{n-r-s}$ are the zeros of any orthogonal polynomial of degree $n-r-s$ with respect to the weight function $\sigma_{r,s}(x) = (1-x)^r(1+x)^s\sigma(x)$, $x \in [-1, 1]$. Furthermore, the weights A_n^+, A_n^- and $A_j^{r,s}$ for all $j = 1, \ldots, n-r-s$ are positive and it holds that,*

$$A_j^{r,s} = \frac{\tilde{A}_j^{r,s}}{(1-x_j^{r,s})^r(1+x_j^{r,s})^s}, \quad j = 1, \ldots, n-r-s,$$

$\{\tilde{A}_j^{r,s}\}_{j=1}^{n-r-s}$ being the Christoffel numbers for $\sigma_{r,s}$.

Thus,

1. As $r + s = 0$, $I_n^{0,0}$ is the n-point Gauss-formula.
2. As $r + s = 1$, $I_n^{1,0}$ and $I_n^{0,1}$ are the n-point Gauss-Radau formulas.
3. As $r + s = 2$, $I_n^{1,1}$ is the n-point Gauss-Lobatto formula.

Sometimes, we will refer to these quadratures as *Gauss-type formulas* so that they can be efficiently computed in terms of an eigenvalue problem involving Jacobi matrices. Indeed, let $\mathcal{J}$ be the Jacobi matrix associated with the weight function σ, set the Darboux transform $\tilde{\sigma}(x) = (x-\beta)\sigma(x)$ with $\beta \in \mathbb{R}$ such that $Q_n(\beta) \neq 0$ for all $n = 1, \ldots$, $\{Q_k\}_{k=0}^\infty$ being the sequence of orthonormal polynomials for σ and denote by $\tilde{\mathcal{J}}$ the Jacobi matrix associated with $\tilde{\sigma}$. Then, it holds that (see e.g. [4])

$$\tilde{\mathcal{J}} = UL + \beta I, \tag{3}$$

where I denotes the unit matrix and $\mathcal{J} - \beta I = LU$. That is, once we have obtained the LU decomposition of the known matrix $\mathcal{J} - \beta I$, (3) gives the Jacobi matrix $\tilde{\mathcal{J}}$ associated with $\tilde{\sigma}$.

When this is restricted to a finite section of the Jacobi matrix (2), it is equivalent with the eigenvalue techniques proposed by Gautschi and Golub (see [21]-[22], [28] and

also [7]). Indeed, if $\beta \in \{-1, 1\}$ and we want β to be a node of the quadrature, then we modify the last b_{n-1} and require that the corresponding $\hat{Q}_n = Q_n - \hat{b}_{n-1}Q_{n-1}$ has a zero in β, which leads to $\hat{b}_{n-1} = Q_n(\beta)/Q_{n-1}(\beta)$. By changing b_{n-1} into $b_{n-1} + \hat{b}_{n-1}$ we get a modified truncated Jacobi matrix $\hat{\mathcal{J}}_n$ which will deliver the nodes and weights of the Gauss-Radau formula like in the Gauss case. Similarly, one may consider $\hat{Q}_n = Q_n - \hat{b}_{n-1}Q_{n-1} - \hat{a}_{n-1}Q_{n-2}$ and solve for $\hat{a}_{n-1}$ and $\hat{b}_{n-1}$ by requiring that $\hat{Q}_{n-1}(\pm 1) = 0$, which leads to the system

$$\begin{pmatrix} Q_{n-1}(1) & Q_{n-2}(1) \\ Q_{n-1}(-1) & Q_{n-2}(-1) \end{pmatrix} \begin{pmatrix} \hat{b}_{n-1} \\ \hat{a}_{n-1} \end{pmatrix} = \begin{pmatrix} Q_n(1) \\ Q_n(-1) \end{pmatrix}.$$

Modifying the truncated Jacobi matrix by replacing (b_{n-1}, a_{n-1}) with $(b_{n-1} + \hat{b}_{n-1}, a_{n-1} + \hat{a}_{n-1})$ gives a matrix $\hat{\mathcal{J}}_n$ that provides the nodes and weights of the Gauss-Lobatto formula through its eigenvalue decomposition as in the classical Gauss case.

3 Integration of periodic functions

Suppose now we are concerned with the approximate calculation of the integral,

$$I_\omega(g) = \int_{-\pi}^{\pi} g(\theta)\omega(\theta)d\theta,$$

g and ω being 2π-periodic functions and ω a weight function on $[-\pi, \pi]$. Without loss of generality we will assume the normalization $\int_{-\pi}^{\pi} \omega(\theta)d\theta = 1$. For this purpose, we will use an n-point quadrature rule like,

$$I_n^\omega(g) = \sum_{j=1}^{n} \lambda_j g(\theta_j), \quad \{\theta_j\}_{j=1}^n \subset (-\pi, \pi], \quad \theta_j \neq \theta_k \text{ if } j \neq k,$$

but now imposing that $I_n^\omega(T) = I_\omega(T)$, for any trigonometric polynomial $T(\theta) = \sum_{k=0}^{N}(a_k \cos k\theta + b_k \sin k\theta)$ with as high degree N as possible. In this respect, it is known that $N \leq n - 1$ (see [36, pp. 73-74]) and that the case $N = n - 1$ gives rise to the quadrature formulas with the maximum trigonometric degree of precision which come characterized in terms of the so-called bi-orthogonal systems of trigonometric polynomials associated with ω (see [18] or [44] for further details). Alternatively, taking into account that any 2π-periodic function on $\mathbb{R}$ can be viewed as a function defined on the unit circle $\mathbb{T} = \{z \in \mathbb{C} : |z| = 1\}$ we could write,

$$I_\omega(g) = \int_{-\pi}^{\pi} g(e^{i\theta})\omega(\theta)d\theta,$$

to be approximated by,

$$I_n^\omega(g) = \sum_{j=1}^{n} \lambda_j f(z_j), \quad \{z_j\}_{j=1}^n \subset \mathbb{T}, \quad z_j \neq z_k \text{ if } j \neq k, \tag{4}$$

such that

$$I_n^{\omega}(L) = I_{\omega}(L), \text{ for all } L \in \Lambda_{-(n-1),n-1}, \tag{5}$$

where for p and q are integers with $p \leq q$, $\Lambda_{p,q} = \text{span}\{z^k : p \leq k \leq q\}$ and $\Lambda = \text{span}\{z^k : k \in \mathbb{Z}\}$. Here the functions in Λ are called Laurent polynomials so that if $T(\theta)$ is a trigonometric polynomial of degree m then one can write $T(\theta) = L(e^{i\theta})$ with $L \in \Lambda_{-(m-1),m-1}$. Moreover, for an ordinary polynomial $P_n(z)$ of exact degree n, we define its reverse or reciprocal as $P_n^*(z) = z^n\overline{P_n\,(1/\bar{z})}$. Concerning the construction and characterization of the quadrature rule (4) satisfying (5) one has the following (see [35] and [27]),

Theorem 3.1 *Set $I_{\omega}(g) = \int_{-\pi}^{\pi} g(e^{i\theta})\omega(\theta)d\theta$, $I_n^{\omega}(g) = \sum_{j=1}^{n} \lambda_j g(z_j)$ with $z_j \in \mathbb{T}$, $j = 1, \ldots, n$ and let $\{\varphi_k\}_{k=0}^{\infty}$ be the sequence of orthonormal (Szegő) polynomials for ω. Then $I_n^{\omega}(g) = I_{\omega}(g)$, for all $g \in \Lambda_{-(n-1),n-1}$, if and only if,*

1. *$\{z_j\}_{j=1}^{n}$ are the zeros of $B_n(z, \tau_n) = \varphi_n(z) + \tau_n\varphi_n^*(z)$ for some $\tau_n \in \mathbb{T}$,*
2. *$\lambda_j = \left(\sum_{k=0}^{n-1} |\varphi_k(z_j)|^2\right)^{-1} > 0$, for all $j = 1, \ldots, n$.*

$I_n^{\omega}(g)$ as given in Theorem 3.1 is called an n-point Szegő quadrature rule (see [35]) and represents the analog on the unit circle of the Gaussian formulas.

Szegő formulas are also *optimal* in the sense that there can not exist an n-point quadrature formula with nodes on $\mathbb{T}$ exactly integrating any Laurent polynomial either in $\Lambda_{-n,n-1}$ or in $\Lambda_{-(n-1),n}$. However, as mentioned in [46, Section 12], it can be proved that an n-point Szegő formula is exact in $\mathcal{L}_n \subset \Lambda$ such that $\dim(\mathcal{L}_n) = 2n$ and $\Lambda_{-(n-1),n-1} \subset \mathcal{L}_n$ (see [38]). Unlike the Gaussian rules, Szegő formulas are not uniquely determined because of the presence of the arbitrary parameter $\tau_n \in \mathbb{T}$. Thus, given $z_\alpha \in \mathbb{T}$, one can take $\tilde{\tau}_n \in \mathbb{T}$ such that $B_n(z_\alpha, \tilde{\tau}_n) = 0$ where $B_n(z, \tilde{\tau}_n) = \varphi_n(z) + \tilde{\tau}_n\varphi_n^*(z)$. Hence, $\tilde{\tau}_n = -\frac{\varphi_n(z_\alpha)}{\varphi_n^*(z_\alpha)} \in \mathbb{T}$ provides an n-point Szegő quadrature formula with a fixed node $z_\alpha \in \mathbb{T}$ in advance and called a Szegő-Radau quadrature rule.

On the other hand, if $\rho_n(z)$ denotes the monic Szegő polynomial of degree n, one can write (up to a multiplicative factor)

$$B_n(z, \tau_n) = \rho_n(z) + \tau_n\rho_n^*(z).$$

Now, from the recurrence relation for $\{\rho_k\}_{k=0}^{\infty}$ (see [25], [29], [45, Theorem 11.4.2] or [41, Theorem 1.5.2]),

$$\begin{pmatrix} \rho_{k+1}(z) \\ \rho_{k+1}^*(z) \end{pmatrix} = \begin{pmatrix} z & \delta_{k+1} \\ \overline{\delta_{k+1}}z & 1 \end{pmatrix} \begin{pmatrix} \rho_k(z) \\ \rho_k^*(z) \end{pmatrix}, \quad k = 0, 1, \ldots, \tag{6}$$

with $\rho_0(z)=\rho_0^*(z)=1$, $\delta_0=1$ and $\delta_k=\rho_k(0)\in\mathbb{D}$ for all $k\geq 1$ (Verblunsky parameters[1]), then for $\tau_n\in\mathbb{T}$ it follows that,

$$B_n(z,\tau_n)=\rho_n(z)+\tau_n\rho_n^*(z)=C_n[z\rho_{n-1}(z)+\tilde{\tau}_n\rho_{n-1}^*(z)],\quad C_n\neq 0 \text{ and } \tilde{\tau}_n\in\mathbb{T}.$$

Thus, to generate an n-point Szegő formula, we take $\tau_n\in\mathbb{T}$ and consider the zeros of $B_n(z,\tau_n)$ that essentially depends on the parameters $\delta_0,\delta_1,\ldots,\delta_{n-1}$ and τ_n. More precisely, define the matrix

$$H_n(\tau_n)=D_n^{-1/2}\begin{pmatrix} -\delta_1 & -\delta_2 & \cdots & -\delta_{n-1} & -\tau_n \\ \sigma_1^2 & -\overline{\delta_1}\delta_2 & \cdots & -\overline{\delta_1}\delta_{n-1} & -\overline{\delta_1}\tau_n \\ 0 & \sigma_2^2 & \cdots & \overline{\delta_2}\delta_{n-1} & -\overline{\delta_2}\tau_n \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & \sigma_{n-1}^2 & -\overline{\delta_{n-1}}\tau_n \end{pmatrix} D_n^{1/2}, \tag{7}$$

where $\sigma_k=\sqrt{1-|\delta_k|^2}\in(0,1]$, $k=1,2,\ldots n$ and $D_n=\operatorname{diag}[\gamma_0,\ldots,\gamma_{n-1}]\in\mathbb{R}^{n\times n}$ with $\gamma_0=1$, $\gamma_k=\gamma_{k-1}\sigma_k^2>0$, $k=1,\ldots,n-1$ and $\tau_n\in\mathbb{T}$. Under these conditions one has ([30]-[31]),

Theorem 3.2 *$H_n(\tau)$ given in (7) is an unreduced unitary upper Hessenberg matrix for all $\tau\in\mathbb{T}$, so that its eigenvalues $\{z_j\}_{j=1}^n$ which are distinct and of unit magnitude are the zeros of $B_n(z,\tau)=\rho_n(z)+\tau\rho_n^*(z)$ or equivalently, the nodes of the n-point Szegő formula for the parameter τ. Furthermore, the square of the first component of the eigenvector of unit length associated with z_j yields the weight λ_j.*

Finally, in a similar way as done when dealing with the estimation of $\int_{-1}^{+1}f(x)\sigma(x)dx$ so that the points ± 1 are taken as nodes in a quadrature formula (Gauss-Lobatto rule), suppose z_α and z_β on $\mathbb{T}$ such that $z_\alpha\neq z_\beta$ and take $n>2$. Then (see [5], [34]) there exist (and they can be easily computed) complex numbers $\tilde{\delta}_{n+1}\in\mathbb{D}$ and $\tilde{\tau}_n\in\mathbb{T}$ such that z_α and z_β are zeros of

$$\tilde{B}_n(z)=z\tilde{\rho}_{n-1}(z)+\tilde{\tau}_n\tilde{\rho}_{n-1}^*(z)\quad\text{with}\quad\tilde{\rho}_{n-1}^*(z)=z\rho_{n-2}(z)+\tilde{\delta}_{n-1}\rho_{n-2}^*(z). \tag{8}$$

Thus, if we denote by $z_1,\ldots,z_{n-2},z_\alpha$ and z_β the zeros of $\tilde{B}_n(z)$ we have that $z_j\in\mathbb{T}$, $z_j\neq z_k$ if $j\neq k$ and $z_j\notin\{z_\alpha,z_\beta\}$, $1\leq j,k\leq n-2$. Furthermore, there exist positive weights A,B and λ_j, $j=1,\ldots n-2$ such that,

$$\tilde{I}_n^\omega(g)=Ag(z_\alpha)+Bg(z_\beta)+\sum_{j=1}^{n-2}\lambda_j g(z_j)=I_\omega(g),\ \text{for all } g\in\Lambda_{-(n-2),n-2}. \tag{9}$$

[1]There are at least four other terms: Szegő, reflection, Schur and Geronimus parameters, see [41, Chapter 1.5].

$\tilde{I}_n^{\omega}(g)$ in (9), and that could be incidentally exact in $\Lambda_{-(n-1),n-1}$, provided that $\tilde{\delta}_{n-1} = \delta_{n-1}$, is called an n-point Szegő-Lobatto formula for ω with prescribed nodes z_α and z_β.

Szegő, Szegő-Radau and Szegő-Lobatto formulas will be sometimes referred as Szegő-type quadrature rules whose computation is the aim of this paper, under some special conditions on the weight ω as described on the next section.

4 Symmetric weight functions

In this section we will be concerned with Szegő-type rules associated with a symmetric weight function ω on $[-\pi, \pi]$, that is, $\omega(-\theta) = \omega(\theta)$, $\theta \in [-\pi, \pi]$. Setting (trigonometric moments)

$$\mu_k = \int_{-\pi}^{\pi} e^{-ik\theta}\omega(\theta)d\theta, \quad k = 0, \pm 1, \pm 2, \ldots \tag{10}$$

(recall that we are assuming that $\mu_0 = 1$) and considering the sequence $\{\delta_k\}_{k=0}^{\infty}$ of Verblunsky parameters, then it follows,

Lemma 4.1 *The following statements are all equivalent:*

1. *ω is a symmetric weight function on $[-\pi, \pi]$.*
2. *The Toeplitz matrices associated with ω are symmetric, i.e. $\mu_{-k} = \mu_k$ for all $k \in \mathbb{Z}$.*
3. *The trigonometric moments are real, i.e. $\mu_k \in \mathbb{R}$ for all $k \in \mathbb{Z}$.*
4. *The Verblunsky parameters δ_k lie in $(-1, 1)$ for all $k \geq 1$.*

As for the quadratures, it will be convenient to give the following

Definition 4.2 *Let ω be a symmetric weight function on $[-\pi, \pi]$. Then, an n-point Szegő formula $I_n^{\omega}(g) = \sum_{j=1}^{n} \lambda_j g(z_j)$ for $I_\omega(g)$ is said to be symmetric if the nodes are real or appear on $\mathbb{T}$ in complex conjugate pairs.*

Now we are concerned with the characterization and computation of symmetric Szegő formulas, if there exist. For this purpose, from Lemma 4.1 it should be taken into account that the sequence of monic Szegő polynomials $\{\rho_k\}_{k=0}^{\infty}$ has real coefficients and hence, it holds that (see e.g. [6]):

Proposition 4.3 *Let ω be a symmetric weight function on $[-\pi, \pi]$. Then,*

1. *An n-point Szegő formula $I_n^{\omega}(g) = \sum_{j=1}^{n} \lambda_j g(z_j)$ generated by $B_n(z, \tau_n) = \rho_n(z) + \tau_n \rho_n^*(z)$ is symmetric, if and only if, $\tau_n \in \{\pm 1\}$.*
2. *Let $I_n^{\omega}(g) = \sum_{j=1}^{n} \lambda_j g(z_j)$ be an n-point Szegő formula for $I_\omega(g)$ and suppose that $z_j = \overline{z_k}$ for some j and k, $1 \leq j, k \leq n$. Then, $\lambda_j = \lambda_k$.*

From Proposition 4.3, we see that when dealing with symmetric rules, their computation essentially reduces to one half. In this respect, computation will be carried out by passing to the interval [-1,1] having in mind the following,

Proposition 4.4 *ω is a symmetric weight function on $[-\pi,\pi]$, if and only if, there exists a weight function σ on $[-1,1]$ such that $\omega(\theta)=\sigma(\cos\theta)|\sin\theta|$. Furthermore, it holds*

$$\int_{-1}^{+1} f(x)\sigma(x)dx=\frac{1}{2}\int_{-\pi}^{\pi} g(e^{i\theta})\omega(\theta)d\theta,\quad g(e^{i\theta})=f\left(\frac{e^{i\theta}+e^{-i\theta}}{2}\right).$$

A connection between quadrature formulas for ω and σ on $[-\pi,\pi]$ and $[-1,1]$ respectively is shown in the following (see [6] and also [17]):

Proposition 4.5 *Take $r,s\in\{0,1\}$ and consider $n-r-s$ distinct nodes $\{x_j^{r,s}\}_{j=1}^{n-r-s}$ on $(-1,1)$ along with the n real numbers $A_+^{r,s}$, $A_-^{r,s}$ and $\{A_j^{r,s}\}_{j=1}^{n-r-s}$. Set $x_j^{r,s}=\cos\theta_j^{r,s}$, $\theta_j^{r,s}\in(0,\pi)$ and define $z_j^{r,s}=e^{i\theta_j^{r,s}}$, $z_{n-r-s+j}^{r,s}=\overline{z_j^{r,s}}$ and $\lambda_j^{r,s}=\lambda_{n+j-r-s}^{r,s}=A_j^{r,s}$, $1\le j\le n-r-s$. Then, the following statements are equivalent:*

1. *$I_{n;(r,s)}^{\sigma}(f)=rA_+^{r,s}f(1)+sA_-^{r,s}f(-1)+\sum_{j=1}^{n-r-s}A_j^{r,s}f(x_j^{r,s})=I_\sigma(f)$, for all $f\in\mathcal{P}_N$.*
2. *$I_{2n-r-s}^{\omega}(g)=2[rA_+^{r,s}g(1)+sA_-^{r,s}g(-1)]+\sum_{j=1}^{2(n-r-s)}\lambda_j^{r,s}g(z_j^{r,s})=I_\omega(g)$, for all $g\in\Lambda_{-N,N}$.*

Now, from Theorem 2.2 one sees that as $N=2n-1-r-s$ the following results in:

a) As $r=s=0$, then $N=2n-1$ and therefore $I_{n;(0,0)}^{\sigma}(f)=\sum_{j=1}^{n}A_j^{0,0}f(x_j^{0,0})$ coincides with the n-point Gaussian formula for σ yielding the following $2n$-point quadrature rule for ω:

$$I_{2n}^{\omega}(g)=\sum_{j=1}^{2n}\lambda_j^{0,0}g(z_j^{0,0})=I_\omega(g),\text{ for all }g\in\Lambda_{-(2n-1),2n-1},$$

which is clearly a $2n$-point symmetric Szegő rule. Hence, the nodes $\{z_j^{0,0}\}_{j=1}^{2n}$ are the zeros of $B_{2n}(z,\tau_{2n})=\rho_{2n}(z)+\tau_{2n}\rho_{2n}^*(z)$ with $\tau_{2n}\in\{\pm1\}$. Since $\rho_{2n}(-1)=\rho_{2n}^*(-1)$ it follows that $\tau_{2n}=1$ i.e. $\{z_j^{0,0}\}_{j=1}^{2n}$ are the zeros of $B_{2n}(z,1)=\rho_{2n}(z)+\rho_{2n}^*(z)$. In short, we have:

$$z_j^{0,0}=x_j^{0,0}+i\sqrt{1-(x_j^{0,0})^2}\ \text{ and }\ \lambda_j^{0,0}=A_j^{0,0}\ \text{ for all }\ j=1,\dots n,$$

where $\{x_j^{0,0}\}_{j=1}^{n}$ are the zeros of the n-th orthogonal polynomial for σ and $\{A_j^{0,0}\}_{j=1}^{n}$ the corresponding Christoffel numbers of order n for σ.

b) As $r=1$ and $s=0$, then $N=2n-2$ and $I_{n;(1,0)}^{\sigma}(f)=A_+^{1,0}f(1)+\sum_{j=1}^{n-1}A_j^{1,0}f(x_j^{1,0})$ represents the n-point Gauss-Radau formula for σ with a fixed node at $x=1$, giving rise to the following $(2n-1)$-point rule for ω:

$$I_{2n-1}^{\omega}(g)=2A_+^{1,0}g(1)+\sum_{j=1}^{n-1}\lambda_j^{1,0}[g(z_j^{1,0})+g(\overline{z_j^{1,0}})]=I_\omega(g),\text{ for all }g\in\Lambda_{-(2n-2),2n-2}.$$

Hence, we have again a $(2n-1)$-point symmetric Szegő formula for $I_\omega(g)$ whose nodes are the zeros of $B_{2n-1}(z,-1)=\rho_{2n-1}(z)-\rho^*_{2n-1}(z)$. Now it follows,

$$z_j^{1,0}=x_j^{1,0}+i\sqrt{1-(x_j^{1,0})^2}\ \text{ and }\ \lambda_j^{1,0}=A_j^{1,0}=\frac{\tilde{A}_j^{1,0}}{1-x_j^{1,0}}\ \text{ for all }\ j=1,\dots n-1,\ (11)$$

where $\{x_j^{1,0}\}_{j=1}^{n-1}$ are the zeros of the $(n-1)$-orthogonal polynomial for $\sigma_{1,0}(x)=(1-x)\sigma(x)$ and $\tilde{A}_j^{1,0}$, $j=1,\dots,n-1$ its corresponding Christoffel numbers of order $n-1$. Moreover, since $I_{2n-1}(1)=I_\omega(1)=1$ it follows,

$$A_+^{1,0}=\frac{1}{2}-\sum_{j=1}^{n-1}\lambda_j^{1,0}.$$

c) As $r=0$, $s=1$, then $N=2n-2$ and similarly to the previous case, $I^\sigma_{n;(0,1)}(f)=A_-^{0,1}f(1)+\sum_{j=1}^{n-1}\lambda_j^{0,1}f(x_j^{0,1})$ represents the n-point Gauss-Radau formula for σ with a fixed node at $x=-1$ and yielding

$$I^\omega_{2n-1}(g)=2A_-^{0,1}g(-1)+\sum_{j=1}^{n-1}\lambda_j^{0,1}[g(z_j^{0,1})+g(\overline{z_j^{0,1}})]=I_\omega(g),\ \text{for all } g\in\Lambda_{-(2n-2),2n-2},$$

that represents a $(2n-1)$-point symmetric Szegő formula for $I_\omega(g)$ whose nodes are the zeros $B_{2n-1}(z,1)=\rho_{2n-1}(z)+\rho^*_{2n-1}(z)$. Again, we have:

$$z_j^{0,1}=x_j^{0,1}+i\sqrt{1-(x_j^{0,1})^2}\ \text{ and }\ \lambda_j^{0,1}=A_j^{0,1}=\frac{\tilde{A}_j^{0,1}}{1+x_j^{0,1}}\ \text{ for all }\ j=1,\dots n-1,$$

where $\{x_j^{0,1}\}_{j=1}^{n-1}$ are the zeros of the $(n-1)$-th orthogonal polynomial for $\sigma_{0,1}(x)=(1+x)\sigma(x)$ and $\tilde{A}_j^{0,1}$, $j=1,\dots,n-1$ its corresponding Christoffel numbers of order $n-1$. In a similar way,

$$A_-^{1,0}=\frac{1}{2}-\sum_{j=1}^{n-1}\lambda_j^{0,1}.$$

d) As $r=s=1$, then $N=2n-3$ and we see that $I^\sigma_{n;(1,1)}(f)=A_+^{1,1}f(1)+A_-^{1,1}f(-1)+\sum_{j=1}^{n-2}A_j^{1,1}f(x_j^{1,1})$ represents the n-point Gauss-Lobatto formula for $I_\sigma(f)$ and giving rise to the following quadrature rule for $I_\omega(g)$:

$$I^\omega_{2n-2}(g)=2[A_+^{1,1}g(1)+A_-^{1,1}g(-1)]+\sum_{j=1}^{n-2}\lambda_j^{1,1}[g(z_j^{1,1})+g(\overline{z_j^{1,1}})].$$

Since $I^\omega_{2n-2}(g)=I_\omega(g)$, for all $g\in\Lambda_{-(2n-3),2n-3}$, we see that it represents again a $(2n-2)$-point symmetric Szegő formula and its nodes are the zeros of $B_{2n-2}(z,-1)=\rho_{2n-2}(z)-\rho^*_{2n-2}(z)$. Now,

$$z_j^{1,1}=x_j^{1,1}+i\sqrt{1-(x_j^{1,1})^2}\ \text{ and }\ \lambda_j^{1,1}=A_j^{1,1}=\frac{\tilde{A}_j^{1,1}}{1-(x_j^{1,1})^2}\ \text{ for all }\ j=1,\dots n-2,$$

where $\{x_j^{1,1}\}_{j=1}^{n-2}$ are the zeros of the $(n-2)$-th orthogonal polynomial for $\sigma_{1,1}(x) = (1-x^2)\sigma(x)$ and $\tilde{A}_j^{1,1}$, $j = 1, \ldots, n-2$ the corresponding Christoffel numbers of order $n-2$. As for the remaining weights $A_+^{1,1}$ and $A_-^{1,1}$ since the quadrature formula exactly integrates $g(z) = 1$ and $g(z) = z$, it follows a system of two equations in the unknowns $A_+^{1,1}$ and $A_-^{1,1}$ which can be explicitly solved, yielding

$$2A_+^{1,1} = \frac{1-\mu_{-1}}{2} - \sum_{j=1}^{n-2} \lambda_j^{1,1} \left(1 - \Re(z_j^{1,1})\right) \text{ and } 2A_-^{1,1} = \frac{1+\mu_{-1}}{2} - \sum_{j=1}^{n-2} \lambda_j^{1,1} \left(1 + \Re(z_j^{1,1})\right).$$

As a conclusion, we can say that when dealing with a symmetric weight function ω on $[-\pi, \pi]$ the computation of any n-point symmetric Szegő rule reduces to an eigenvalue problem for a Jacobi matrix of dimension $E[n/2]$ associated with the weight functions $\sigma_{r,s}(x) = (1-x)^r(1+x)^s\sigma(x)$ on $[-1, 1]$, where $E[x]$ denotes the integer part of x, and $r, s \in \{0, 1\}$ while σ is such that $\omega(\theta) = \sigma(\cos\theta)|\sin\theta|$.

Finally, let us analyze the computation of a symmetric Szegő-Lobatto formula, if there exists, when ω is symmetric and we have fixed in advance two nodes on $\mathbb{T}$, say z_α and z_β which are complex conjugate. As already seen, for $n > 2$ there exist positive numbers A, B and λ_j, $j = 1, \ldots, n-2$ along with $n-2$ distinct nodes $z_1, \ldots, z_n$ on $\mathbb{T}$ such that $z_j \notin \{z_\alpha, z_\beta\}$, $j = 1, \ldots, n-2$ and so that

$$\tilde{I}_n^\omega(g) = Ag(z_\alpha) + Bg(z_\beta) + \sum_{j=1}^{n-2} \lambda_j g(z_j) = I_\omega(g), \text{ for all } g \in \Lambda_{-(n-2),n-2} \quad \text{(Szegő-Lobatto formula).} \tag{12}$$

In this case and as shown in [34], the parameter $\tilde{\delta}_{n-1}$ in formula (8) can be taken real, i.e. $\tilde{\delta}_{n-1} \in (-1, 1)$ so that $\tilde{\tau}_n \in \{\pm 1\}$. Hence, the Szegő-Lobatto formula (12) is symmetric. Actually (12) is an n-point symmetric Szegő formula for a new symmetric weight function $\tilde{\omega}(\theta)$ whose first $n-1$ Verblunsky parameters are $\delta_1, \ldots, \delta_{n-2}$ and $\tilde{\delta}_{n-1}$. Therefore, computation reduces to the first situation but now replacing $\omega(\theta)$ by $\tilde{\omega}(\theta)$ and $\sigma(x)$ by $\tilde{\sigma}(x)$ such that $\tilde{\omega}(\theta) = \tilde{\sigma}(\cos(\theta))|\sin(\theta)|$.

For instance, once fixed z_α and z_β on $\mathbb{T}$ such that $z_\beta = \overline{z_\alpha}$, suppose n even, say $n = 2m$ and set $x_\alpha = \Re(z_\alpha)$. Since in formula (8), $\tilde{\delta}_{n-1}$ is real and $\tilde{\tau}_n \in \{\pm 1\}$, suppose that $\tilde{\tau}_n = 1$ i.e. the nodes of $\tilde{I}_{2m}^\omega(g)$ given by (12) are the zeros of $\tilde{B}_{2m}(z) = \tilde{B}_{2m}(z, 1) = z\rho_{2m-1}(z) + \tilde{\rho}_{2m-1}^*(z)$ which are real or appear in complex conjugate pairs. Since $\tilde{B}_{2m}(\pm 1) \neq 0$, we can write,

$$\tilde{I}_{2m}^\omega(g) = A[g(z_\alpha) + g(\overline{z_\alpha})] + \sum_{j=1}^{m-1} \lambda_j [g(z_j) + g(\overline{z_j})].$$

Setting $x_j = \Re(z_j)$, $j = 1, \ldots, m-1$ we see that $x_\alpha, x_1, \ldots, x_{m-1}$ are the zeros of the m-th orthogonal polynomial for $\tilde{\sigma}(x)$ and $A, \lambda_1, \ldots, \lambda_{m-1}$ the corresponding Christoffel numbers of order m. Furthermore, since in this case we know that $x_\alpha = \Re(z_\alpha)$ is an eigenvalue of

the Jacobi matrix, a deflation method could be conveniently used, having in mind that $A = \frac{1}{2} - \sum_{j=1}^{m-1} \lambda_j$. The other three remaining cases, that is, n even and $\tilde{\tau}_n = -1$ and n odd and $\tilde{\tau}_n = \pm 1$ can be treated in a similar way. In short, the computation of the symmetric Szegő-Lobatto formulas leads to the computation of the Gauss-type quadrature rules associated with the new weight function $\tilde{\sigma}(x)$ such that $\tilde{\omega}(\theta) = \tilde{\sigma}(\cos\theta)|\sin\theta|$, $\theta \in [-\pi, \pi]$.

5 The connection with Jacobi matrices

As seen in the previous section, given a symmetric weight function ω on $[-\pi, \pi]$, we can compute its Szegő-type quadrature formulas in terms of the Gauss-type rules for a weight function σ on $[-1, 1]$ such that $\omega(\theta) = \sigma(\cos\theta)|\sin\theta|$ so that to carry on an efficient computation we need the corresponding Jacobi matrices associated with σ. However, the initial available information that we have on the weight function ω are its trigonometric moments (10) and only in very few cases the Szegő polynomials are explicitly known. Here it should be recalled that the basic information to compute Szegő quadrature formulas are the Verblunsky parameters, $\delta_0 = 1$ and $\delta_k = \rho_k(0)$ for all $k = 1, 2, \ldots$. Thus, starting from the trigonometric moments, the Verblunsky parameters can be efficiently computed by means of Levinson algorithm, consisting in the implementation of the Szegő recurrence (6) (see [37]). An alternative procedure called split Levinson algorithm was derived in [19] when ω is symmetric. Indeed, the latter routine computes the corresponding para-orthogonal polynomials $B_n(z, \pm 1)$ (and hence, the Verblunsky parameters) with half the work of the computation saved. Also, a well known map from trigonometric moments to Verblunsky coefficients is Schur's algorithm (see e.g. [33]). Now the question is: given the Verblunsky parameters $\{\delta_k\}_{k=0}^{\infty}$ for ω, how can the Jacobi matrix for σ be computed? The answer can be found in the so-called Geronimus relations (see [24]).

Theorem 5.1 *Let ω be a symmetric weight function on $[-\pi, \pi]$ and σ the weight function on $[-1, 1]$ related to ω by $\omega(\theta) = \sigma(\cos\theta)|\sin\theta|$. Let $\{\delta_k\}_{k=0}^{\infty}$ be the sequence of Verblunsky parameters for ω and $\{a_n\}_{n=1}^{\infty}$ and $\{b_n\}_{n=0}^{\infty}$ be the coefficients of the Jacobi matrix (2) for σ. Then, the following holds:*

$$2a_n = \sqrt{(1-\delta_{2n})(1-\delta_{2n-1}^2)(1+\delta_{2n-2})}, \quad n \geq 1 \quad \text{and} \quad 2b_n = \delta_{2n-1}(1-\delta_{2n}) - \delta_{2n+1}(1+\delta_{2n}), \quad n \geq 0. \tag{13}$$

Example 5.2 *As an illustration of Theorem 5.1, let us consider the weight function ω on $[-\pi, \pi]$ associated with the Poisson Kernel, namely $\omega(\theta) = |z+\gamma|^{-2}$ where $\gamma \in (-1, 1)$ and $z = e^{i\theta}$. Since γ is real, ω is clearly an even function. In this case, it is well known (see e.g. [45, Theorem 11.2]) that $\rho_n(z) = z^{n-1}(z+\gamma)$ for all $n \geq 1$ and hence $\delta_0 = 1$, $\delta_1 = \gamma$, and $\delta_k = 0$ for all $k \geq 2$. Thus, from (13) it results that,*

$$a_1 = \frac{1}{2}\sqrt{2(1-\gamma^2)}, \quad b_0 = -\gamma \quad b_1 = \frac{\gamma}{2}, \quad a_k = 1 \quad \text{and} \quad b_k = 0, \text{ for all } k \geq 2.$$

In general, for our purposes the calculations can be arranged as follows. Suppose that the Verblunsky parameters $\{\delta_k\}_{k=0}^{\infty}$ for the symmetric weight function ω on $[-\pi,\pi]$ are known and set

$$\sigma_{r,s}(x) = (1-x)^r(1+x)^s\sigma(x), \quad r,s \in \{0,1\}, \quad x \in [-1,1],$$

where $\omega(\theta) = \sigma(\cos\theta)|\sin\theta|$.

Let $\mathcal{J}^{r,s}$ denote the Jacobi matrix associated with $\sigma_{r,s}(x)$. Thus, $\mathcal{J}^{0,0}$ is the Jacobi matrix for $\sigma(x)$ whose entries are directly given by Theorem 5.1. Set the LU decomposition $\mathcal{J}^{0,0} = L^{0,0}U^{0,0}$. Then, by (3) we have that $\mathcal{J}^{1,0} = U^{0,0}L^{0,0} + I$ and $\mathcal{J}^{0,1} = U^{0,0}L^{0,0} - I$. Finally, by considering the LU decomposition of $\mathcal{J}^{1,0}$, that is $\mathcal{J}^{1,0} = L^{1,0}U^{1,0}$, then $\mathcal{J}^{1,1} = U^{1,0}L^{1,0} - I$. Once the Jacobi matrices $\mathcal{J}^{r,s}$ have been determined, the computation of the symmetric Szegő-type quadratures for ω or equivalently the Gauss-type rules for σ is a straightforward task.

However, as for the computation of the Jacobi matrices $\mathcal{J}^{r,s}$, with $r,s \in \{0,1\}$, we might also think of the following alternative approach. Indeed, for $z = e^{i\theta}$ and $r,s \in \{0,1\}$, set

$$\omega_{r,s}(\theta) = \sigma_{r,s}(\cos\theta)|\sin\theta| = (1-\cos\theta)^r(1+\cos\theta)^s\omega(\theta) = \frac{1}{2^{r+s}}|z+r|^2|z+s|^2\omega(\theta),$$

and consider the trigonometric moments $\mu_k^{r,s} = \int_{-\pi}^{\pi} e^{-ik\theta}\omega_{r,s}(\theta)d\theta$. Then, it can be checked for all $k = 0,1,\ldots$ that

$$\mu_k^{r,s} = \frac{1}{2^{r+s}}\left\{[(r-s)^2+1+r^2s^2]\mu_k + [s(1+r^2) - r(1+s^2)](\mu_{k-1}+\mu_{k+1}) - rs(\mu_{k-2}+\mu_{k+2})\right\}. \tag{14}$$

Thus, starting from the trigonometric moments μ_k for ω we can compute the moments $\mu_k^{r,s}$ for $\omega_{r,s}$ by (14) and then from here, making use of the Levinson´s algorithm, the corresponding Verblunsky parameters $\delta_k^{r,s}$ for $\omega_{r,s}(\theta)$ can be computed. Finally, from Theorem 5.1 we deduce the Jacobi matrices $\mathcal{J}^{r,s}$. However, since the computations to generate the coefficients $\delta_k^{1,0}$ or $\delta_k^{0,1}$ can not be stored, in general, to compute $\delta_k^{1,1}$ this way seems to be much more expensive and longer. Even in the case where we dispose of the Verblunsky parameters $\delta_k^{0,0} = \delta_k$, $k = 0,1,\ldots$ for ω and although we can deduce an explicit relation between the sequences $\{\delta_k\}_0^{\infty}$ and $\{\delta_k^{r,s}\}_0^{\infty}$ with $r,s \in \{0,1\}$ (see [23]), this process involves so many calculations that it does not seem advisable. For instance, for all $n \geq 1$ it holds that,

$$\delta_n^{1,0} = \frac{\rho_{n+1}(1)\rho_n(1)}{k_n K_n(1,1)} - \delta_{n+1},$$

$\{\rho_k\}_{k=0}^{\infty}$ being the sequence of monic Szegő polynomial for ω, $k_n = \|\rho_n\|_\omega^2 = \prod_{j=1}^{n}(1-|\delta_j|^2)$ and where $K_n(x,y)$ denotes the reproducing kernel for $\mathcal{P}_n$ with respect to the inner product induced by ω, that is

$$K_n(x,y) = \sum_{j=0}^{n} \frac{\rho_j(x)\overline{\rho_j(y)}}{k_j}.$$

From the relation (see [41, pp. 57-58])

$$\rho_n(1) = \sqrt{k_n} \prod_{j=1}^{n} \sqrt{\frac{1+\delta_j}{1-\delta_j}}, \quad n \geq 1,$$

an explicit connection between $\{\delta_n^{1,0}\}_{n=0}^{\infty}$ and $\{\delta_n\}_{n=0}^{\infty}$ can be stated.

6 Numerical experiments involving Rogers-Szegő polynomials

As it is known, most of the examples about symmetric weight functions ω on $[-\pi, \pi]$ considered in the literature directly arise from weight functions σ on $[-1, 1]$ after making the change of variable $x = \cos\theta$ so that $\omega(\theta) = \sigma(\cos\theta)|\sin\theta|$. Thus, when some kind of information on the weight function σ in terms of moments or Jacobi parameters is available, the computation of the symmetric Szegő-type quadrature formulas for ω reduces to the computation of the Gauss-type quadrature for σ and very little has to be done. Hence, the interest appears when we have the usual available information on ω but little or no information on σ, apart from the above connection between both weight functions. This is the case of the symmetric weight function ω giving rise to the sequence of the so-called Rogers-Szegő polynomials, which will be used to illustrate the approach presented in the previous sections with some numerical experiments. This weight function is the "wrapped" Gaussian measure given by

$$\omega(\theta) = \omega(q;\theta) = \frac{1}{\sqrt{2\pi\log(1/q)}} \sum_{j=-\infty}^{+\infty} \exp\left(\frac{-(\theta - 2\pi j)^2}{2\log(1/q)}\right), \quad 0 < q < 1. \tag{15}$$

Properties of Rogers-Szegő polynomials, the family of orthogonal polynomials on $\mathbb{T}$ with respect to ω given by (15) have been recently studied; see e.g. [41, Chapter 1.6] and its references along with [16]. In spite of the rather special shape of ω, it is surprising we have explicit information about the familiar parameters characterizing ω. For instance, the sequence of trigonometric moments is given by (see e.g. [41, Chapter 1.6]) $\mu_n = q^{\frac{n^2}{2}}$ for all $n \geq 0$ and the Verblunsky parameters by

$$\delta_n = (-1)^n q^{\frac{n}{2}},\ n = 0, 1, \ldots. \tag{16}$$

Even more, the family of monic Rogers-Szegő polynomials is explicitly given by

$$\rho_n(z) = \sum_{j=0}^{n} (-1)^{n-j} \begin{bmatrix} n \\ j \end{bmatrix}_q q^{\frac{n-j}{2}} z^j,$$

where $\begin{bmatrix} n \\ j \end{bmatrix}_q = \frac{(n)_q}{(j)_q (n-j)_q}$ (q-binomial coefficient) with $(n)_q = (1-q)(1-q^2)\cdots(1-q^n)$. Observe that ω is clearly symmetric; hence there exists a weight function σ on $[-1, 1]$ such that $\omega(\theta) = \sigma(\cos\theta)|\sin\theta|$, but nothing is known about σ. The aim of this section is to

perform some illustrative numerical experiments concerning the computation of symmetric Szegő-type quadratures for ω given by (15) and making use of (16) along the results of Section 5. These quadratures are of a great interest when dealing with the computation of integrals of the form $\int_{-\infty}^{\infty} f(x)e^{-\gamma x^2}dx$, with $\gamma > 0$ and f a 2π-periodic function (see [16]).

From (16) it follows that the Hessenberg matrices defined in (7) for the weight function $\omega(\theta)$ given by (15) have the form

$$H_n(\tau_n) = D_n^{-1/2} \begin{bmatrix} q^{1/2} & -q & \cdots & (-1)^n q^{(n-1)/2} & -\tau_n \\ 1-q & q^{3/2} & \cdots & (-1)^{n+1} q^{n/2} & q^{1/2}\tau_n \\ 0 & 1-q^2 & \cdots & (-1)^n q^{(n+1)/2} & -q\tau_n \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & 1-q^{n-1} & (-1)^n q^{(n-1)/2}\tau_n \end{bmatrix} D_n^{1/2},$$

where $D_n = \mathrm{diag}[1, (1)_q, (2)_q, \ldots, (n-1)_q]$ and $\tau_n \in \mathbb{T}$. From Theorem 5.1, the coefficients of the Jacobi matrices for $\sigma(x)$ such that $\omega(\theta) = \sigma(\cos\theta)|\sin\theta|$ are given by,

$$a_n = \frac{1}{2}\sqrt{(1-q^n)(1-q^{2n-1})(1+q^{n-1})}, \quad n \geq 1 \quad \text{and} \quad b_n = \frac{1}{2}q^{n-\frac{1}{2}}\left(q^{n+1}+q^n+q-1\right), \quad n \geq 0.$$

Thus, we can compute the nodes and weights of the corresponding n-point symmetric Szegő rule for different values of n either via Hessenberg or via Jacobi matrices. A comparison has been made concerning the computational time in seconds required in the solution of both eigenvalue problems in MAPLE® 9.5[2] with 30 digits by using an standard routine and fixing $q = 0.2$ and $\tau_n = 1$. The results are displayed on Table 1 and they clearly show the advantage of Jacobi over Hessenberg. However, it should be recalled here that a whole variety of practical eigenvalue computation algorithms for unitary Hessenberg matrices has already been developed in the literature; see e.g. [2], [8]-[9], [20], [32] and [43].

Finally, let us recall that Szegő rules depend on a parameter $\tau \in \mathbb{T}$ so that when $\tau = \pm 1$, then symmetric formulas appear which can be efficiently computed via Jacobi matrices. If we take $\tau \in \mathbb{T}\backslash\{\pm 1\}$, the corresponding Szegő formulas are not symmetric anymore and its computation must be done by means of the Hessenberg matrices (7). With an illustrative character, an estimation of the integral $\int_{-\pi}^{\pi} g(\theta)\omega(\theta)d\theta$ with $\omega(\theta)$ given by (15) has been made by using n-point Szegő formulas with different values of τ. The absolute errors displayed in the tables below show that the symmetric rules ($\tau = 1$) produce similar results for the choices $\tau = \pm i$.

[2]MAPLE is a registered trademark of Waterloo Maple, Inc.

n	$Jacobi$	$Hessenberg$
100	0.016	0.0234
200	0.047	2.969
300	0.172	11.281
400	0.454	30.45
500	0.937	81.828
600	1.73	172

Table 1.— A comparision in seconds for the computation of an n point symmetric Szegő quadrature formula for the Rogers-Szegő weight function (15) with $q = 0, 2$ and $\tau_n = 1$, via Jacobi or Hessenberg matrices.

n	$\tau = 1$	$\tau = i$	$\tau = -i$
6	$7.8219774E-03$	$6.0089760E-03$	$6.0089760E-03$
8	$1.1307038E-03$	$7.9659815E-05$	$7.9659815E-04$
10	$1.2796254E-04$	$8.2581333E-05$	$8.2581333E-05$
12	$1.07083165E-05$	$6.3268992E-06$	$6.3268992E-06$

Table 2.— A comparision of the absolute errors in the computation of an n point Szegő quadrature formula for the Rogers-Szegő weight function (15) with $q = 0, 9$, $g(\theta) = (\cos\theta)^{19}$ and different values of τ.

Acknowledgments

This work is partially supported by Dirección General de Programas y Transferencia de Conocimiento, Ministerio de Ciencia e Innovación of Spain under grant MTM 2008-06671. A. Bultheel acknowledges financial support from the Belgian Network DYSCO (Dynamical Systems, Control, and Optimization), funded by the Interuniversity Attraction Poles Programme, initiated by the Belgian State, Science Policy Office. F. Perdomo-Pío has been partially supported by Grant of Agencia Canaria de Investigación, Innovación y Sociedad de la Información del Gobierno de Canarias.

References

[1] **M. Alfaro and J.M. Montaner**, *On five-diagonal Toeplitz matrices and orthogonal polynomials on the unit circle*, Numer. Algor. 10(1-2) (1995) 137-154.

[2] **G.S. Ammar, L. Reichel and D.C. Sorensen**, *An implementation of a divide and conquer algorithm for the unitary eigenproblem*, ACM Trans. Math. Software 18(3) (1992) 292-307.

n	$\tau = 1$	$\tau = i$	$\tau = -i$
6	$4.1917414E-06$	$1.8719070E-06$	$3.0695894E-06$
8	$1.2989586E-07$	$6.3949537E-08$	$8.1546108E-08$
10	$3.9410378E-09$	$2.0120779E-09$	$2.27096253E-09$
12	$1.1791645E-10$	$6.11319328E-11$	$6.4948880E-11$

Table 3.— A comparision of the absolute errors in the computation of an n point Szegő quadrature formula for the Rogers-Szegő weight function (15) with $q = 0,5$, $g(\theta) = (\sin\theta + 3)^{-1}$ and different values of τ.

n	$\tau = 1$	$\tau = i$	$\tau = -i$
6	$7.5024567E-06$	$3.8286286E-05$	$3.8340752E-05$
8	$2.1913111E-07$	$1.1277238E-06$	$1.1280409E-06$
10	$6.4403534E-09$	$3.3200865E-08$	$3.3202737E-08$
12	$1.8952515E-10$	$9.7736602E-10$	$9.7737363E-10$

Table 4.— A comparision of the absolute errors in the computation of an n point Szegő quadrature formula for the Rogers-Szegő weight function (15) with $q = 0,2$, $g(\theta) = \frac{\cos\theta}{\sin\theta+3}$ and different values of τ.

[3] **E. Berriochoa, A. Cachafeiro and F. Marcellán**, *Differential properties for Sobolev orthogonality on the unit circle*, J. Comp. Appl. Math. 133 (2001) 231-239.

[4] **M.I. Bueno and F. Marcellán**, *Darboux transformation and perturbation of linear functionals*, Lin. Alg. Appl. 384 (2004), 215-242.

[5] **A. Bultheel, L. Daruis and P. González-Vera**, *Quadrature formulas on the unit circle with prescribed nodes and maximal domain of validity*, J. Comp. Appl. Math. 231(2) (2009) 948-963.

[6] **A. Bultheel, L. Daruis and P. González-Vera**, *A connection between quadrature formulas on the unit circle and the interval* $[-1,1]$, Appl. Numer. Math. 132 (2001) 1-14.

[7] **A. Bultheel and R. Cruz-Barroso and M. Van Barel**, *On Gauss-type quadrature formulas with prescribed nodes anywhere on the real line*. Calcolo, to appear 2010, doi: 10.1007/s10092-009-0013-x.

[8] **A. Bunste-Gerstner and L. Elsner**, *Schur parameter pencils for the solution of the unitary eigenproblem*, Lin. Alg. Appl. (1992) 741-778.

[9] **A. Bunste-Gerstner and C. He**, *On a Sturm sequence of polynomials for unitary Hessenberg matrices*, SIAM J. Matrix Anal. Appl. 16(4) (1995) 1043-1055.

[10] **A. Cachafeiro and F. Marcellán**, *Orthogonal polynomials of Sobolev type on the unit circle*, J. Approx. Theory 78 (1994) 127-146.

[11] **M.J. Cantero, R. Cruz-Barroso and P. González-Vera**, *A matrix approach to the computation of quadrature formulas on the unit circle*, Appl. Num. Math. 58 (2008) 296-318.

[12] **M.J. Cantero, L. Moral and L. Velázquez**, *Five-diagonal matrices and zeros of orthogonal polynomials on the unit circle*, Lin. Alg. Appl. 362 (2003) 29-56.

[13] **R. Cruz-Barroso, L. Daruis, P. González-Vera, O. Njåstad**, *Sequences of orthogonal Laurent polynomials, biorthogonality and quadrature formulas on the unit circle*, J. Comp. Appl. Math. 200 (2007) 424-440.

[14] **R. Cruz-Barroso and S. Delvaux**, *Orthogonal Laurent polynomials on the unit circle and snake-shaped matrix factorizations*, J. Approx. Theory 161 (2009) 65-87.

[15] **R. Cruz-Barroso, P. González-Vera and F. Perdomo Pío**, *Orthogonality, interpolation and quadratures on the unit circle and the interval* $[-1, 1]$, J. Comp. Appl. Math. To appear 2010, doi: 10.1016/j.cam.2009.12.021.

[16] **R. Cruz-Barroso, P. González-Vera and F. Perdomo Pío**, *Quadrature formulas associated with Rogers-Szegő polynomials*, Comp. Math. Appl. 57 (2009) 308-323.

[17] **R. Cruz-Barroso, P. González-Vera and F. Perdomo Pío**, *Rational approximants associated with measures supported on the unit circle and the real line*, Appl. Num. Math. To appear 2010.

[18] **R. Cruz-Barroso, P. González-Vera and O. Njåstad**, *On bi-orthogonal systems of trigonometric functions and quadrature formulas for periodic integrands*, Numer. Algor. 44 (2007) 309-333.

[19] **P. Delsarte, Y. Genin**, *The Split Levinson Algorithm*, IEEE Trans Acoust. Spreech, Signal Proc. ASSP-34 (1986) 470-478.

[20] **S. Delvaux and M. Van Barel**, *Eigenvalue computation for unitary rank structured matrices*, J. Comp. Appl. Math. 213(1) (2008) 268-287.

[21] **W. Gautschi**, *Orthogonal Polynomials Computation and Approximation*, Oxford University Press, Oxford (2004).

[22] **W. Gautschi**, *Orthogonal polynomials, quadrature, and approximation: computational methods and software (in matlab)*, In: Orthogonal Polynomials and Special Functions. Lecture Notes in Mathematics 1883, Springer, Berlin (2006) 1-77.

[23] **L.E. Garza Gaona and F. Marcellán**, *Spectral transformations of measures supported on the unit circle and the Szegő transformation*, Numer. Algor. 49 (2008) 169-185.

[24] **Ya.L. Geronimus**, *On the trigonometric moment problem*, Ann. Math. 47(2) (1946) 742-761.

[25] **Ya.L. Geronimus**, *Polynomials orthogonal on a circle and their applications*, Amer. Math. Soc. Transl. 1, Providence, RI (1962) 1-78.

[26] **E. Godoy and F. Marcellán**, *Orthogonal polynomials on the unit circle: Distribution of zeros*, J. Comp. Appl. Math. 37(1-3) (1991) 195-208.

[27] **P. González-Vera, J.C. Santos-León and O. Njåstad**.- *Some results about numerical quadrature on the unit circle*, Adv. Comp. Math. 5 (1996) 297-328.

[28] **G. Golub**, *Some modified matrix eigenvalue problems*, SIAM Rev. 15(2) (1973) 318-334.

[29] **U. Grenander and G. Szegő**, *Toeplitz forms and their applications*, Univ. of California Press, Berkeley 1958, Chelsea, New-York, 2nd edition, 1984.

[30] **W.B. Gragg**, *Positive definite Toeplitz matrices, the Arnoldi process for isometric operators and Gaussian quadrature on the unit circle*, In E.S. Nicholaev Ed., Numer. Meth. Lin. Alg., Moscow University Press, Moscow (1982) 16-32 (in Russian).

[31] **W.B. Gragg**, *Positive definite Toeplitz matrices, the Arnoldi process for isometric operators and Gaussian quadrature on the unit circle*, J. Comp. Appl. Math. 46 (1993) 183-198. This is a slightly revised version of [30].

[32] **W.B. Gragg**, *The QR algorithm for unitary Hessenberg matrices*, J. Comp. Appl. Math. 16 (1986) 1-8.

[33] **C. Jagels and L. Reichel**, *On the construction of Szegő polynomials*, J. Comp. Appl. Math. 46 (1993) 241-254.

[34] **C. Jagels and L. Reichel**, *Szegő-Lobatto quadrature rules*, J. Comp. Appl. Math. 200 (2007) 116-126.

[35] **W.B. Jones, O. Njåstad and W.J. Thron**, *Moment theory, orthogonal polynomials, quadrature, and continued fractions associated with the unit circle*, Bull. London Math. Soc. 21 (1989) 113-152.

[36] **V.I. Krylov**, *Approximate Calculation of Integrals*, The MacMillan Company, New York, 1962.

[37] **N. Levinson**, *The Wiener RMS (root mean square) error criterion in filter design and prediction*, J.Math. Phys. 25 (1947) 261-278.

[38] **O. Njåstad and J.C. Santos-León**, *Domain of validity of Szegő quadrature formulas*, J. Comp. Appl. Math. 202(2) (2007) 440-449.

[39] **B. Simon**, *CMV matrices: Five years after*, J. Comp. Appl. Math. 208(1) (2007) 120-154.

[40] **B. Simon**, *OPUC on one foot*, Bull. Amer. Math. Soc. 42 (2005) 431-460.

[41] **B. Simon**, *Orthogonal Polynomials on the unit circle, Part 1: Classical Theory*, Amer. Math. Soc. Colloq. Publ. 54.1, Amer. Math. Soc., Providence, RI, 2005.

[42] **B. Simon**, *Orthogonal Polynomials on the Unit Circle, Part 2: Spectral Theory*, Amer. Math. Soc. Colloq. Publ. 54.2, Amer. Math. Soc., Providence, RI, 2005.

[43] **M. Stewart**, *An error analysis of a unitary Hessenberg QR algorithm*, SIAM J. Matrix Anal. Appl. 28(1) (2006) 40-67.

[44] **G. Szegő**, *On bi-orthogonal systems of trigonometric polynomials*, Magyar Tud. Akad. Kutató Int. Közl 8 (1963) 255-273.

[45] **G. Szegő**, *Orthogonal Polynomials*, Amer. Math. Soc. Coll. Publ. Vol 23, Amer. Math. Soc. Providence, R.I. 1975.

[46] **D. Watkins**, *Some perspectives on the eigenvalue problem*, SIAM Review 35(3) (1993) 430-471.

Monografías de la Real Academia de Ciencias de Zaragoza **33**, 197–208, (2010).

Multivariate polynomial interpolation: some new trends

J. M. Carnicer and M. Gasca

Instituto Universitario de Matemáticas y Aplicaciones (IUMA), Universidad de Zaragoza, Spain

Dedicated to Manuel Calvo with esteem and friendship in occasion of his 65th birthday

Abstract

In this paper we comment on some recent research in the field of multivariate polynomial interpolation with special emphasis in the influence of the relative position of the interpolation nodes to extend certain univariate techniques like simple Lagrange formulae, Aitken–Neville formulae and Lebesgue constants.

Keywords: Multivariate polynomial interpolation

MSC (Math. Sc. Class.) 41A05, 41A63, 65D05

1 Introduction

Univariate polynomial interpolation is a classical subject, with a long history and a rather complete theory. Its multivariate counterpart is much more complicated. Only very isolated papers, although due to important mathematicians as Kronecker or Jacobi, can be found before the beginning of the 20th Century.

Except for tensor product problems, which are obvious extensions of univariate problems, multivariate techniques have only been systematically considered in the second half of the 20th century due to the development of computers. Another reason for the interest of multivariate problems was the emergence of new mathematical methods, as finite element methods for solving partial differential equations, cubature formulae, etc. Several surveys on the history of the subject and its development have been written at the end of the last century (cf. [24, 25]). The purpose of this paper is to point out some advances in the field and show some new trends which have appeared in the last decade.

2 Constructing sets of nodes suitable for interpolation problems

The usual interpolating space in one variable is $P_n(\mathbb{R}) := \{p \in P(\mathbb{R}) : \deg p \leq n\}$, the space of polynomials of degree not greater than n. In contrast, there exist many

different choices for subspaces of $P(\mathbb{R}^d)$, the space of polynomials in d variables, for solving interpolation problems in the case $d > 1$, depending on the number and distribution of the interpolation points (also called nodes). In fact, it is necessary for the existence and uniqueness of the interpolant that the number of nodes equals the dimension of the interpolating space. The most common interpolating space is the space

$$P_n(\mathbb{R}^d) := \{p \in P(\mathbb{R}^d) : \deg p \leq n\}$$

of polynomials of total degree less than or equal to n. Another subspace of polynomials, specially used for rectangular grids, is the space

$$P_{n_1,\ldots,n_d}(\mathbb{R}^d) := \{p \in P(\mathbb{R}^d) : \deg_i p \leq n_i,\ i = 0, \ldots, d\}$$

where $\deg_i$ denotes the partial degree, that is, the degree with respect to the ith variable x_i. Generalizations of these approaches introducing polynomials whose directional degree, that is, degree along certain directions, is prescribed have also been used [12].

A set of nodes X is said to be correct for an interpolating space S, if the Lagrange interpolation problem on X has always a unique solution in S. An interpolating space S, $\dim S = N$, satisfies the Haar condition on a given domain D if any set of N nodes in D is correct for S. There are many spaces which satisfy this condition in one variable, in particular that of polynomials of degree not greater than N on any subinterval of the real line. However, except for the trivial case of problems with only one interpolation point, there exist no spaces in more than one variable satisfying the Haar condition on domains D containing an open set. Therefore the fact that a set of nodes is correct depends on the geometric distribution of the nodes. This is a remarkable difference with the univariate case and provides one of the main research subjects in multivariate interpolation.

Chung and Yao [21] identified the $P_n(\mathbb{R}^d)$-correct sets of nodes whose Lagrange polynomials can be factored as products of first degree polynomials. This geometric condition (usually called for brevity GC) describes distributions of nodes leading to simple Lagrange formulae. A GC_n-set X is a set with $\dim P_n(\mathbb{R}^d)$ nodes such that for each $x \in X$, there exist n hyperplanes containing all nodes but x. Chung and Yao provided two important examples of distributions of nodes satisfying their geometric condition: principal lattices and natural lattices. The geometric characterization can be easily used to check whether a given set is a GC_n set but it provides no suggestion about how to construct such sets. One of the research lines recently developed has focused on describing examples which generalize those provided by Chung and Yao. For the sake of simplicity we restrict ourselves to the bivariate case.

Principal lattices are distributions of points formed by the intersections of 3 pencils of equidistant parallel lines, $n+1$ lines each, in such a way that any node is the intersection of one line of each pencil. The standard example is the set of points $(i/n, j/n) : 0 \leq$

$i + j \leq n$, where $(i/n, j/n)$ is the intersection of the lines $x - i/n = 0$, $y - j/n = 0$, $x + y - (i + j)/n = 0$. Principal lattices were extended by Lee and Phillips to sets called 3-pencil lattices, allowing concurrent pencils of lines (parallel lines can be considered as a particular case of concurrence at infinity).

Jaklič et al. [30] have used a barycentric form as a useful tool to extend three-pencil lattices to triangulations covering polygonal domains. In this way, they construct continuous piecewise polynomials interpolating Lagrange data, analyzing the degrees of freedom in the selection of the nodes in each subtriangle. Multivariate extensions of these results have also been considered recently by the same authors.

In the last decade, the authors [14, 15] have extended the Lee-Phillips construction to lattices generated by cubic pencils. Cubic pencils are families of lines $ax + by + c = 0$ whose coefficients satisfy a cubic equation. An addition in the set of nonsingular lines Λ^* of a cubic pencil is introduced as a dualization of the addition of points of a cubic curve (a common tool in algebraic geometry). Three lines sum up to 0 if and only if they meet at a point which is not a vertex of the pencil. Usually the lines are parameterized in terms of an isomorphic classical group G,

$$L : t \in G \mapsto L(t) \in \Lambda^*,$$

so that $L(-t_1 - t_2)$ is the line in Λ^* concurrent with $L(t_1)$ and $L(t_2)$. For each $t_0, t_1, t_2 \in G$ with $t_0 + t_1 + t_2 = 0$ and $h \in G$, the set of points $X = \{x_{ijk} \mid i + j + k = n\}$, where

$$\{x_{ijk}\} = L(t_0 - (n - i)h) \cap L(t_1 + jh) \cap L(t_2 + kh), \quad i + j + k = n,$$

is a generalized principal lattice if the lines $L(t_r + ih)$, $i = 0, \ldots, n$, $r = 0, 1, 2$, are all distinct. This construction generalizes 3-pencil lattices. In fact, the product of the three linear pencils arising in the Lee-Phillips construction form a cubic pencil of lines

As an example, we might consider the cubic pencil formed by all lines

$$L(t) \equiv y = \tan(t/2)x - \sin(t), \quad t \in \mathbf{R}/2\pi\mathbf{Z}$$

tangent to a deltoid

$$x(t) = \cos t(\cos t + 1), \quad y(t) = \sin t(\cos t - 1).$$

Here the parameter group is $G = \mathbf{R}/2\pi\mathbf{Z}$. Figure 1 below illustrates an example of this construction. We observe that each of the three families $L(t_r + ih)$, $i = 0, \ldots, n$, do not belong to the same linear pencil, that is, they do not meet at a vertex.

An apparently more general situation was described in [14, 15] defining generalized principal lattices as distributions of points obtained from the intersections of three families of lines L_i^r, $i = 0, \ldots, n$, $r = 0, 1, 2$, not necessarily related to a cubic pencil. Later on, it

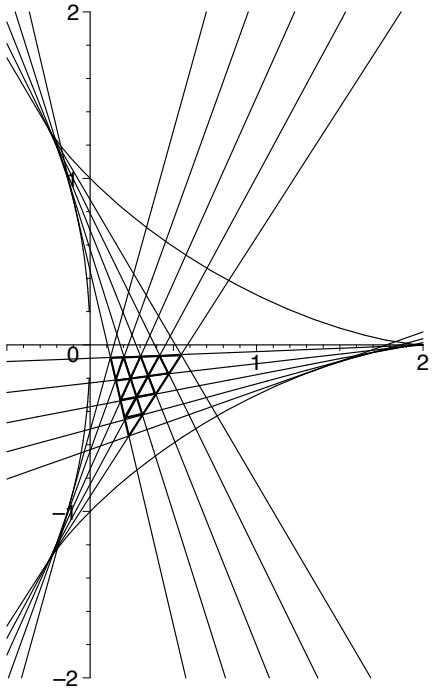


Figure 1.— A lattice generated by a cubic pencil

was proved in [20] that every bivariate generalized principal lattice can be obtained from a cubic pencil of lines. Starting with a canonical classification of cubic curves, the authors have classified in [16] all possible kinds of generalized principal lattices in two variables, up to projectivities.

An extension of these ideas to more than 2 variables was obtained in [18], showing some examples, and simultaneously pointing out the difficulties of getting a general construction.

The Aitken-Neville algorithm in one variable provides the solution of an interpolation problem of degree n on a set X of $n+1$ nodes by linear interpolation of the solutions of two subproblems of degree $n-1$ on n points of X. Multivariate extensions of the Aitken-Neville algorithm have been considered by several authors in the last half of the 20th century. In a recent paper [19] the relationship between multivariate Aitken-Neville algorithms and generalized principal lattices has been studied. Aitken-Neville sets in $\mathbb{R}^d$ were defined by Sauer and Xu [34] as distributions of points allowing a recursive interpolation formula that constructs an interpolating polynomial of degree n on $\binom{n+d}{d}$ nodes from the solutions of $d+1$ problems of degree $n-1$ with $\binom{n+d-1}{d}$ data each. The initial data provide the solutions of $\binom{n+d}{d}$ problems of degree 0 (1 data each). We construct with them the solutions of $\binom{n+d-1}{d}$ problems of degree 1 ($d+1$ data each) and so on, until getting the solution of the complete problem with $\binom{n+d}{d}$ data. The scheme extends the univariate Aitken-Neville algorithm. In [19] it has been shown that each

Aitken-Neville set satisfies the GC condition of Chung and Yao and that any generalized principal lattice is an Aitken-Neville set. As a consequence, an interpolation problem on a generalized principal lattice can be solved by an Aitken-Neville algorithm. Let us remark that interpolation problems on generalized principal lattices can also be solved by the Lagrange formula because they satisfy the GC condition.

In the plane, it has been proved in [19] that any Aitken-Neville set of degree $n > 2$ is a generalized principal lattice. However there are Aitken-Neville sets of degree 2 which are not generalized principal lattices.

Another approach to Aitken-Neville formulae for multivariate interpolation can be found in [17], where extensions to Aitken-Neville formulae (in the sense that an interpolant is obtained combining interpolants on smaller subsets of nodes) have been analyzed. The recursion formulae for Aitken-Neville sets introduced by Sauer and Xu [34, 19] are obtained as a consequence of the main result of [17], applied to the problem of obtaining an interpolant in $P_n(\mathbb{R}^d)$ in terms of subinterpolants in $P_{n-1}(\mathbb{R}^d)$.

A natural lattice of degree n in $\mathbb{R}^d$ is the set of all points obtained intersecting d hyperplanes among $n + d$ hyperplanes in general position in $\mathbb{R}^d$. The number of such intersections is $\binom{n+d}{d} = \dim P_n(\mathbb{R}^d)$. Natural lattices are adequate for total degree interpolation of degree n in $\mathbb{R}^d$ because they satisfy the GC_n condition. In the bivariate case, a natural lattice is the set of pairwise intersections among $n + 2$ lines in general position.

In one variable, Hermite problems can be seen as a limit case of Lagrange problems when nodes coalesce. Analogously, in [11], we have analyzed interpolation problems defined in the set of intersections of $n + 2$ distinct lines in any position, allowing multiple concurrences of lines at a point or parallel lines. This general situation gives rise to Hermite problems on subspaces of $P_n(\mathbb{R}^2)$ consisting of polynomials whose degree diminishes along directions corresponding to parallel or coincident lines among the lines defining the lattice.

The search of new distributions of interpolation nodes which form a correct set on a certain space of polynomials is a natural question in multivariate polynomial interpolation. In the last years, Bojanov and Xu [3] have studied bivariate Hermite problems, where the interpolant matches prescribed data consisting of function values and consecutive normal derivatives on a set of points placed on several circles centered at the origin. Lagrange interpolation is a particular case. For a given integer n, the interpolation nodes are the intersection points of $2[\frac{n+1}{2}] + 1$ rays from the origin with a set of concentric circles (here $[m]$ means the integer part of m). The circles can be repeated and, in this case, successive radial derivatives are provided as interpolation data. In [3] the number of circles, counting multiplicities, is $[\frac{n}{2}] + 1$. So, the total number of interpolation data is the dimension of $P_n(\mathbb{R}^2)$ and, if the rays are equidistant (i.e. the nodes on each circle are equidistant), the set of nodes is correct for $P_n(\mathbb{R}^2)$. The poisedness holds if the circles freely rotate. In

particular, Bojanov and Xu rediscovered in [3] a nice star-shaped example of a natural lattice, previously obtained by Hakopian [26].

Later on Bojanov and Xu [4] considered nonconcentric circles and Hakopian and Ismail [27] have extended the analysis to conic sections. Hakopian and Khalaf [28] have continued this work, proving that the poisedness of the data for the Bojanov-Xu problem [3] is equivalent to the unisolvence of certain $2[\frac{n+1}{2}]+1$ dimensional Lagrange interpolation problems. As a consequence, they prove that the Bojanov-Xu problem is poised not only for equidistant rays, but for a wide family of sets of rays satisfying some simple conditions.

3 Some conjectures on distributions of nodes suitable for interpolation problems

Gasca and Maeztu [22] considered interpolation nodes in $\mathbb{R}^2$ defined as intersections of lines and provided a method of constructing poised Lagrange and Hermite interpolation problems on apropriate subspaces of polynomials. The novelties of their approach consisted, on the one hand, in allowing multiple concurrences of lines (giving rise to derivatives as interpolation data) and, on the other hand, in solving the problem by means of a recurrence with a simple Newton-like formula. An extension to more than two variables was also suggested in that paper.

The simplest case arises when $n+1$ nodes lie on a line l_0, n nodes on another line l_1, none of them lying on l_0, $n-1$ points on another line l_2, none of them lying on $l_0 \cup l_1$, and so on. Then the Lagrange interpolation problem on these nodes is poised in the space $P_n(\mathbb{R}^2)$. This distribution of points has been rediscovered several times in the literature, apparently the first times by Berzolari [2] and much later by Radon [32]. Recently it has been referred to as the Berzolari-Radon distribution of points.

The Lagrange problem on the Berzolari-Radon distribution leads to a triangular system of equations and can be solved by a Newton formula. Obviously, not any lattice of this type verifies the geometric condition (GC) of Chung and Yao. On the one hand, the Berzolari-Radon lattices are straightforward to construct, and the corresponding interpolation problem can be solved with a simple Newton formula. On the other hand, for any given set of points in the plane, the geometric characterization can be checked and leads to a simple explicit Lagrange formula as mentioned in Section 2. However, the class of GC sets is not completely known. Some particular constructions like natural lattices, principal lattices and their generalizations are often used but other instances of GC sets are not so easy to construct and describe. It is a remarkable fact that all known GC sets are particular cases of the Berzolari-Radon construction.

A natural question arising in this context is whether or not any planar GC set is a Berzolari-Radon set. Bézout Theorem implies that no line of the plane can contain more

that $n+1$ nodes of a set correct for $P_n(\mathbb{R}^2)$. A conjecture launched in [22], known presently as the GM conjecture, states that, *if X is a planar GC set of order n, then $n+1$ points of X are collinear*. It is easy to see that, if a planar GC_n set has $n+1$ collinear points, then the set obtained removing those points is a GC_{n-1} set. Hence, if the GM conjecture is true, it can be shown by induction that any planar GC set of order n is a Berzolari-Radon set of order n. This is the reason why the GM conjecture has atracted much attention in the last twenty years. In spite of the relevance of the consequences of the geometric condition, the GM conjecture has only been proved up to degree $n = 4$ (see, for instance [10, 29]). However no counterexample has been found. The conjecture was reinforced in [13], where the authors proved that, if the GM conjecture holds for any degree, then there exist at least 3 lines containing $n + 1$ nodes of any planar GC_n set. The existence of at least 3 lines containing $n + 1$ nodes has been considered as a new conjecture, known as the CG conjecture, equivalent to the GM conjecture for points in the plane.

A multivariate version of the GM conjecture in $\mathbb{R}^d$ (the GM_d conjecture) was stated recently by de Boor [8]: there exist always a maximal hyperplane for any GC_n set in $\mathbb{R}^d$. A maximal hyperplane for a $P_n(\mathbb{R}^d)$-correct set X is any hyperplane containing exactly $\binom{n+d-1}{d-1}$ nodes. The name maximal is based on the fact that no hyperplane can contain more than $\binom{n+d-1}{d-1}$ nodes. The same author also launched a multivariate version of the CG conjecture: there exist at least $d+1$ maximal hyperplanes for any GC_n set in $\mathbb{R}^d$. This conjecture was disproved in [8], where a GC_2 set in $\mathbb{R}^3$ with only 3 maximal hyperplanes was described. This counterexample does not disprove the GM_d conjecture nor the planar CG conjecture. The search of new approaches to the GM conjecture by Hakopian, Jetter and Zimmerman has stimulated recent research on the number of maximal hyperplanes in multivariate GC sets. In a recent paper [1], it is shown that the GM conjecture holds for trivariate GC_2 sets.

4 The search of good interpolation nodes

The Lebesgue constant

$$\Lambda_n = \max_{x\in[a,b]} \sum_{i=0}^{n} \prod_{j\neq i} \frac{|x - x_j^n|}{|x_i^n - x_j^n|}$$

is the norm of the interpolation operator $L : C[a, b] \to C[a, b]$ associated to the Lagrange interpolation problem at nodes $x_0^n < \cdots < x_n^n$ in $[a, b]$ and measures in some sense the condition and stability of the interpolation process. By the Erdös-Brutman Theorem

$$\Lambda_n > \frac{2}{\pi} \log n + 0.5212$$

for any set of nodes $x_0^n < \cdots < x_n^n$ in $[a, b]$. This implies that Λ_n must diverge as $n \to \infty$, independently of the choice of nodes. The search of points for interpolation in P_n with

least possible Lebesgue constant is an interesting question without an explicit solution. The zeros of the Chebyshev polynomials are called the Chebyshev nodes and they are almost optimal in $[-1,1]$, in the sense that the Lebesgue constant has an asymptotic growth

$$\Lambda_n = \frac{2}{\pi}\log n + O(1), \quad n \to \infty.$$

The advantange of the Chebyshev nodes is that they have a simple explicit formula

$$x_i^n = -\cos\frac{(2i+1)\pi}{2(n+1)}, \quad i = 0,\ldots,n.$$

Another important choice of nodes are the Chebyshev-Lobatto nodes

$$x_k^n = -\cos\frac{kn}{n}, \quad k = 0,\ldots,n.$$

In general, the Lebesgue constant is expected to be low for all distributions of points $x_0^n < \cdots < x_n^n$ in $[-1,1]$ which tend to be uniformly distributed when $n \to \infty$ with respect to the Dubiner metric

$$d(x_1,x_2) = |\arccos x_2 - \arccos x_1|.$$

In several variables, there is no clear candidate for almost optimal points for general domains. In the last years, there have been new advances on the subject and new bivariate distributions of points have been proposed for the square and the circle.

In order to avoid the discussion of the different cases arising in total degree interpolation when the degree is even or odd, let us assume for the sake of simplicity that the degree $n = 2m$ is even. Let $\xi_k^n := \cos(k/n)$ denote Chebyshev-Lobatto nodes. Y. Xu [35] proposed the nodes

$$\begin{aligned}(x_{2i,2j+1}, y_{2i,2j+1}) &:= (\xi_{2i}^{2m}, \xi_{2j+1}^{2m}), \quad i = 0,\ldots,m, \quad j = 0,\ldots m-1,\\ (x_{2i+1,2j}, y_{2i,2j+1}) &:= (\xi_{2i+1}^{2m}, \xi_{2j}^{2m}), \quad i = 0,\ldots,m-1, \quad j = 0,\ldots,m,\end{aligned}$$

for interpolation on a subspace of $P_{2m}(\mathbb{R}^2)$ containing $P_{2m-1}(\mathbb{R}^2)$ on the square $[-1,1]^2$. Other authors (Caliari, de Marchi and Vianello [9]) have proposed the "Padua points" for total degree interpolation in $P_{2m}(\mathbb{R}^2)$

$$(x_i, y_j), \quad i = 0,\ldots,2m, \quad j = 0,\ldots,m,$$

where

$$x_i = \xi_i^{2m}, \quad i = 0,\ldots,2m,$$

and

$$y_j := \begin{cases} \xi_{2j}^{2m+1}, & \text{if } m \text{ is odd,} \\ \xi_{2j+1}^{2m+1}, & \text{if } m \text{ is even,} \end{cases} \quad j = 0,\ldots,m$$

with lower Lebesgue constant than the Xu points. Other bivariate and trivariate distributions of points with low Lebesgue constants have been proposed for the square, the circle and other simple bivariate domains.

The Xu points are equally spaced in the Dubiner metric

$$d((x_1, y_1), (x_2, y_2)) = \max(|\arccos x_2 - \arccos x_1|, |\arccos y_2 - \arccos y_1|).$$

A generalization of the Dubiner metric can be defined on any compact subset of $\mathbb{R}^n$. A conjecture stated in [9] says that: *nearly optimal points for polynomial interpolation on a compact set are asymptotically equidistributed with respect to the Dubiner metric.* The research on this subject is now very active and the reader is referred to recent papers by Bos, Caliari, de Marchi, Vianello, Xu among others.

5 Multivariate divided differences

Univariate divided differences can be Êdefined in different ways: as the coefficients of the Newton interpolation formula, as a certain linear functional vanishing on the space of polynomials of a given degree and by a recurrence relation, among others. From any of these definitions the other ones can be derived and also relevant properties in applications, such as error formulae in numerical quadrature and relations with B-spline functions. The extension of the concept to the multivariate case depends on the property that we want to preserve. In other words, the way in which multivariate divided differences are defined can lead to the loss of some common properties of the univariate ones. Generalizations of the divided differences to several variables have been recently proposed by Rabut [31], de Boor [7] and Sauer [33].

A general technique for multivariate polynomial interpolation on correct sets of points is based on constructing extensions of the Newton formula. The Newton approach can be described as the problem of constructing a basis of functions in the interpolation space such that the interpolation conditions give rise to a linear system whose coefficient matrix is lower triangular or block-lower triangular. Such a basis can be called a Newton basis. The space of multivariate polynomials have a graded structure and in order to exploit it, an additional condition for Newton bases of polynomials is usually required. The degree of the polynomials of a Newton basis must be increasing (or increasing by blocks) in some sense. So, extensions to the concept of degree might be necessary for deriving Newton formulae in more general situations. Most concepts of multivariate divided differences can be interpreted in terms of the construction of a suitable generalization of the Newton formula.

Acknowledgments

Partially supported by the Spanish Research Grant MTM2009-07315 and by Gobierno de Aragón.

References

[1] A. Apozyan, G. Avagyan and G. Ktryan, On the Gasca-Maeztu conjecture in $\mathbb{R}^3$. To appear in East Journal of Approximation.

[2] L. Berzolari, Sulla determinazione di una curva o di una superficie algebrica e su alcune questioni di postulazione, *Lomb. Ist. Rend.* **47** (1914), 556–564.

[3] B. Bojanov and Y. Xu. On a Hermite interpolation by polynomials of two variables. *SIAM J. Numer. Anal.* **39** (2002), 1780–1793.

[4] B. Bojanov and Y. Xu. On polynomial interpolation of two variables. *J. Approx. Theory* **120** (2003), 267–282.

[5] C. de Boor and A. Ron. On multivariate polynomial interpolation. *Construct. Approx.* **6** (1992), 287–302.

[6] C. de Boor and A. Ron. The least solution for the polynomial interpolation problem. *Math. Z.* **210** (1992), 347–378.

[7] C. de Boor. Divided differences. *Surveys in Approximation Theory* (electronic) **1** (2005), 46–69.

[8] C. de Boor. Multivariate polynomial interpolation: Conjectures concerning GC-sets. *Numer. Algorithms* **45** (2007), 113–125.

[9] M. Caliari, S. De Marchi and M. Vianello. Bivariate polynomial interpolation on the square at new nodal sets. *Applied Math. Comp.* **165** (2005), 261–274.

[10] J. M. Carnicer and M. Gasca, A conjecture on multivariate polynomial interpolation, *RACSAM, Rev. R. Acad. Cien. Serie A. Mat.* **95** (2001), 145–153.

[11] J. M. Carnicer and M. Gasca, A Newton approach to bivariate Hermite interpolation on generalized natural lattices. *RACSAM (Rev. R. Acad. Cien. Serie A. Mat.)* **96** (2002), 185–195.

[12] J. M. Carnicer and M. Gasca. Asymptotic conditions for degree diminution along prescribed lines. *Numer. Algorithms* **33** (2003), 183–192.

[13] J. Carnicer and M. Gasca. Classification of bivariate configurations wwith simple Lagrange interpolation formulae. *Advances in Comp. Math.* **20** (2004) ,5–16.

[14] J. Carnicer and M. Gasca. Generation of lattices of points for bivariate interpolation. *Numer. Algorithms* **39** (2005), 69–79.

[15] J. Carnicer and M. Gasca. Interpolation on lattices generated by cubic pencils. *Advances in Comp. Math.*, **24** (2006), 113–130.

[16] J. Carnicer and M. Gasca. Cubic pencils of lines and bivariate interpolation. *J. Comp. Appl. Math.*, **219** (2008), 370–382.

[17] J. Carnicer and M. Gasca. Aitken-Neville formulae for multivariate interpolation *Jaen Journal of Approximation* (2010), to appear.

[18] J. Carnicer, M. Gasca and T. Sauer. Interpolation lattices in several variables. *Numer. Math.*, **102** (2006), 559–581.

[19] J. Carnicer, M. Gasca and T. Sauer. Aitken-Neville sets, principal lattices and divided differences. *J. Approx. Theory* **156** (2009), 154–172

[20] J. Carnicer and C. Godés. Generalized principal lattices and cubic pencils. *Numer. Algorithms*, **44** (2007), 133–145.

[21] K. C. Chung and T.H. Yao. On lattices admitting unique Lagrange interpolation. *SIAM J. Num. Anal.*, **14** (1977), 735–743.

[22] M. Gasca and J.I. Maeztu. On Lagrange and Hermite interpolation in R^k. *Numer. Math.*, **39** (1982), 1–14.

[23] M. Gasca and V. Ramírez. Interpolation systems in R^k. *J. Approx. Theory*, **42** (1984), 36–51.

[24] M. Gasca and T. Sauer. On the history of multivariate polynomial interpolation. *J. Comput. Appl. Math.*, **122** (2000), 23–35.

[25] M. Gasca and T. Sauer. Multivariate polynomial interpolation. *Advances in Comp. Math.* **12** (2000), 377–410.

[26] H. A. Hakopian. Multivariate interpolation II of Lagrange and Hermite type. *Stud. Math* **80** (1984), 77–88.

[27] H. A. Hakopian and S. A. Ismail. On Bojanov-Xu interpolation on conic sections. *East J. Approx.* **9** (2003), 251–267.

[28] H. A. Hakopian and M. F. Khalaf. On the poisedness of Bojanov-Xu interpolation. *J. Approx. Theory* **135** (2005), 176–202.

[29] H. A. Hakopian, K. Jetter and G. Zimmermann, A new proof of the GascaMaeztu conjecture for $n = 4$ *Journal of Approximation Theory*, **159**, (2009), 224–242.

[30] G. Jaklič, J. Kozak, M. Krajnc, V. Vitrih and E. Žagar Three-pencil lattices on triangulations. *Numer. Algorithms* **45** (2007), 49–60.

[31] C. Rabut. Generalized divided differences. *SIAM J. Num. Anal.* **38** (2000), 1294–1311.

[32] J. Radon, Zur mechanischen Kubatur, *Monatsh. Math.* **52** (1948) 286–300.

[33] T. Sauer. Degree reducing polynomial interpolation, ideals and divided differences. In Albert Cohen, Jean-Louis Merrien and Larry L. Schumaker, eds., *Curve and surface fitting: Avignon 2006*, pp. 220–237, Mod. Methods Math., Nashboro Press, Brentwood, TN, 2007.

[34] T. Sauer and Yuesheng Xu. The Aitken-Neville scheme ins several variables In C.K. Chui, L.L. Schumaker and J. Stöckler eds., *Approximation Theory X*, pp. 353–366. Vanderbilt Univ. Press, 2002.

[35] Y. Xu, Lagrange interpolation on Chebyshev points of two variables, *Journal of Approximation Theory*, **87** (1996) 220–238.

Monografías de la Real Academia de Ciencias de Zaragoza **33**, 209–223, (2010).

On Sobolev type orthogonal polynomials with unbounded support: asymptotic properties

M. Alfaro[a], J. J. Moreno–Balcázar[b], A. Peña[a] and M. L. Rezola[a]

[a]Depto. de Matemáticas and IUMA. Universidad de Zaragoza (Spain).

[b]Depto. de Estadística y Matemática Aplicada and ICI. Universidad de Almería (Spain).

Abstract

In this expository paper we present a survey about asymptotic properties of Sobolev type orthogonal polynomials with unbounded support.

Key words: Sobolev orthogonal polynomials, Relative asymptotics; Mehler–Heine type formulas; zeros; Bessel functions.

2000MSC: 42C05, 33C45.

1 Introduction

The theory of orthogonal polynomials is a very interesting field in mathematics with important applications to numerical analysis, physics, probability, and statistics among other ones. Orthogonal polynomials are connected with topics like moment problems, mechanical quadratures, continued fractions, spectral methods, quantum mechanics and many other concepts.

Usually, in this theory, the orthogonality is considered with respect to a positive linear functional defined on the linear space of polynomials or, according to the Riesz representation theorem, with respect to a positive measure. Let μ be a finite positive Borel measure supported on an interval I in the real line, we say that the sequence of polynomials $\{P_n\}_{n\geq 0}$ is a sequence of orthogonal polynomials (o.p.) with respect to either the measure μ or the inner product $(f,g)=\int_I f\,g d\mu$ if, for all $n\geq 0$, $\deg P_n = n$ and

(i) $(P_n, P_m) = 0\,,\ n\neq m\,,$

(ii) $(P_n, P_n) > 0,\ n\geq 0.$

Along the paper such a kind of inner products will be called standard inner products. They have the following remarkable property: $(xp, q) = (p, xq)$, for all polynomials p, q. As a consequence, the corresponding standard orthogonal polynomials have nice properties such as the three–term recurrence relation, the summation formula, the interlacing properties of the zeros, etc. From a numerical point of view, a useful consequence is that a Gaussian mechanical quadrature formula has exact precision when we take as nodes the zeros of appropriate standard o.p..

Nonstandard inner products have also been considered in the literature. In particular, the so–called Sobolev inner products that are of the form

$$(f, g) = \int f\, g d\mu_0 + \sum_{i=1}^{r} \int f^{(i)}\, g^{(i)} d\mu_i,$$

where $\{\mu_i\}_{i=0}^{r}$ are finite positive Borel measures supported on the real line and the functions f and g belong to the Sobolev space:

$$W^{2,r}(\mu_0, \mu_1, \ldots, \mu_r) := \{f : \int |f|^2 d\mu_0 + \sum_{i=1}^{r} \int |f^{(i)}|^2 d\mu_i < +\infty\}.$$

Studied by the first time in the forties of the last century, the Sobolev orthogonal polynomials have been object of an increasing interest, approximately, in the last 20 years. Obviously, Sobolev inner products are nonstandard and therefore Sobolev o.p. loose the "good" properties of the standard o.p. However, it is interesting to study these "strange" polynomials that supply us with situations different from the standard ones: no three–term recurrence relation, zeros out of the convex hull of the support of the orthogonality measure including, some times, complex zeros, and so one.

Furthermore, some applications of the Sobolev orthogonality in the theory of standard o.p. are known, for instance, classical polynomials (Jacobi or Laguerre polynomials) with nonclassical parameters are not orthogonal in the usual sense but they are orthogonal with respect to Sobolev inner products (see among others [1] or [17]) and also, Sobolev o.p. in two real variables are solutions of some partial differential equations (see [9], [14], [19] or [24]).

In this paper we are concerned with the so–called Sobolev type (or discrete Sobolev) orthogonal polynomials, that is, polynomials orthogonal with respect to a Sobolev inner product in which $\{\mu_i\}_{i=1}^{r}$ are Dirac's deltas or, in general, discrete measures. More concretely, we consider an inner product of the form

$$(f, g) = \int f(x)g(x)d\mu(x) + \sum_{i=0}^{r} M_i f^{(i)}(c)g^{(i)}(c),$$

where μ is a finite positive Borel measure, $c \in \mathbb{R}$ and $M_i \geq 0$ for $i = 0, 1, ..., r$. In the sequel, we denote by $\{Q_{n,r}\}_{n\geq 0}$ the corresponding sequence of o.p. with the same leading coefficient as the standard o.p. with respect to μ.

More general products where cross–product terms appear in the discrete part (the non–diagonal case) have also been studied. But, recently in [18] the authors prove that every symmetric bilinear form can be reduced to a diagonal case, that is, an inner product without cross–product terms.

In some sense, Sobolev type o.p. are not so far than the standard o.p. since there is the possibility to transform the Sobolev type orthogonality into the standard quasi–orthogonality. As a consequence, several properties of the standard o.p. are partially recovered for the Sobolev type o.p.: they satisfy a $2r+3$ term recurrence relation (see, [13]) and have partial interlacing properties of the zeros ([2]).

Since the polynomial $Q_{n,r}$ is quasi–orthogonal of order $r+1$ with respect to the measure μ it can be expressed as a linear combination (with a fixed number of terms: $r+2$) of standard orthogonal polynomials R_n corresponding to the modified measure $(x-c)^{r+1}d\mu$, that is,

$$Q_{n,r}(x) = \sum_{j=0}^{r+1} a_n^j R_{n-j}(x)\,. \tag{1}$$

One of the topics in the theory is to compare the Sobolev type o.p. with the standard o.p. (with respect to μ) to investigate how the addition of the discrete part in the inner product influences the orthogonal system. Many formal results are known for the polynomials $Q_{n,r}$: recurrence relation, differential formulas, location of zeros, and so on. However, little is known about the asymptotic properties and most of the general results have been obtained when the support of μ ($\operatorname{supp}\mu$) is a bounded set. For instance, in [20], the authors assume that μ is a measure of bounded support for which the asymptotic behaviour of the corresponding o.p. is known; the most relevant class of this type is the Nevai class $M(0,1)$ of o.p. with appropriately converging recurrence coefficients. There, the relative asymptotics is studied when the mass point $c \notin \operatorname{supp}\mu$. The case $c \in \operatorname{supp}\mu$ has been considered in [22].

What happens is that in the bounded case, all the connexion coefficients a_n^j in (1) are bounded and the orthogonal polynomials R_n have an adequate finite ratio asymptotics: these two facts allow us to study each term of (1) separately, in order to get the relative asymptotics for $Q_{n,r}$ (see [20] and [22] where this technique is developed). However, the situation is quite different if we deal with the unbounded case because when we try to obtain the relative asymptotics with the techniques used for the bounded case and we take into account the ratio asymptotics for the polynomials R_n, we come across a serious problem. Indeed, we find that the idea that each term of (1) has a finite limit could not work now, in fact, as we will see later, for the Laguerre–Sobolev type o.p. each term of (1) tends to infinity, all of them being the same order, but with an alternating sign.

The aim of this paper is to describe the current state of the asymptotic properties for Sobolev type polynomials when $\operatorname{supp}\mu$ is unbounded. Mainly we will analyze the case

when μ is the Laguerre probability measure ($d\mu(x) = \frac{x^\alpha e^{-x}}{\Gamma(\alpha+1)}dx$ with $\alpha > -1$) and $c = 0$, that is, the Laguerre–Sobolev o.p. This choice for c is due to the fact that the point $x = 0$ is a singularity of the differential equation satisfied by the classical Laguerre polynomials.

Therefore, we will deal with classical Laguerre polynomials, that is, polynomials orthogonal with respect to the inner product

$$(p, q) = \frac{1}{\Gamma(\alpha+1)} \int_0^\infty p(x)q(x)\, x^\alpha e^{-x}\, dx\,, \quad \alpha > -1.$$

in the space of all polynomials with real coefficients. We will denote by L_n^α the nth Laguerre polynomial with $(-1)^n/n!$ as leading coefficient. Although many of the properties of Laguerre polynomials can be seen, for example, in the books of Chihara [10] and Szegő [23], we remind that the classical Laguerre polynomials with the normalization above quoted are defined by

$$L_n^\alpha(x) = \sum_{k=0}^{n} \binom{n+\alpha}{n-k} \frac{(-1)^k x^k}{k!}\,,$$

and their derivatives satisfy

$$(L_n^\alpha)^{(k)}(x) = (-1)^k L_{n-k}^{\alpha+k}(x)\,. \tag{2}$$

The evaluations at $x = 0$ of the polynomial L_n^α and its successive derivatives are given by

$$(L_n^\alpha)^{(k)}(0) = \frac{(-1)^k\, n!}{(n-k)!} \frac{\Gamma(\alpha+1)}{\Gamma(\alpha+k+1)} L_n^\alpha(0) = \frac{(-1)^k}{(n-k)!} \frac{\Gamma(n+\alpha+1)}{\Gamma(\alpha+k+1)}. \tag{3}$$

¿From Perron's formula in Szegő's book [23], the following asymptotic results can be deduced:

$$\frac{L_{n-1}^\alpha(x)}{L_n^\alpha(x)} \rightrightarrows 1, \quad x \in \mathbb{C} \setminus [0, \infty)\,, \tag{4}$$

$$\frac{n^{1/2} L_n^\alpha(x)}{L_n^{\alpha+1}(x)} \rightrightarrows \sqrt{-x}, \quad x \in \mathbb{C} \setminus [0, \infty)\,. \tag{5}$$

where the symbol $f_n(x) \rightrightarrows f(x)$, $x \in A$, denotes that the sequence $\{f_n\}$ converges to f uniformly on compact subsets of A.

Later on we will use the symbol $f(x) \sim g(x)$ $(x \to a)$ if $\lim\limits_{x\to a} \dfrac{f(x)}{g(x)} = 1$.

2 Laguerre–Sobolev type polynomials

¿From now on $\{Q_{n,r}\}_{n\geq 0}$ denotes the sequence of polynomials orthogonal with respect to an inner product of the form

$$(p, q)_r = \frac{1}{\Gamma(\alpha+1)} \int_0^\infty p(x)q(x)\, x^\alpha e^{-x}\, dx + \sum_{i=0}^{r} M_i p^{(i)}(0) q^{(i)}(0), \tag{6}$$

where $\alpha > -1$ and $M_i > 0$, $i = 0, \ldots, r$, with leading coefficient $(-1)^n/n!$. Notice that all the masses in the discrete part are positive.

Observe that, in fact, $(.,.)_r$ and $Q_{n,r}$ also depend on the parameter α but for simplicity we have omitted it in the notations.

These families of o.p. were considered by the first time by Koekoek and Meijer (see, among others, [15] and [16]) although no asymptotics were studied. The first asymptotic results for Laguerre–Sobolev type o.p. appear in [8]: exterior asymptotics, asymptotics on compact subsets of $(0, +\infty)$, exterior Plancherel–Rotach type asymptotics, Mehler–Heine type formulas and convergence of their zeros are obtained, but only for $r = 0$ and $r = 1$. Concerning the Mehler–Heine type formulas, with $r = 1$ and $M_0, M_1 > 0$ the authors found a behaviour pattern and they established a conjecture. A survey including these results can be seen in [21]. Some of these properties where proved for the non–diagonal case with $r = 1$ in [6] and [7], and later on in [11].

In all these papers the basic tool was the algebraic expression

$$Q_{n,1}(x) = B_0(n)L_n^\alpha(x) + B_1(n)xL_{n-1}^{\alpha+2}(x) + B_2(n)x^2L_{n-2}^{\alpha+4}(x)$$

where the coefficients $B_i(n)$ were given explicitly in [16].

In a discrete Laguerre–Sobolev inner product with an arbitrary number of terms, the problem is that we only have an algebraic expression given in [15], but not the explicit expression of the coefficients $B_i(n)$, of which we only know that they are a non trivial solution of a system with $r + 1$ equations and $r + 2$ unknowns.

Asymptotic properties of Sobolev orthogonal polynomials with respect to a general inner product as (6), that is, with an arbitrary number of masses, have been studied in [5] where, in particular, the conjecture established in [8] is proved to be true. In the sequel we summarize the results obtained there.

As we have already said, the interest lies in knowing the differences in the asymptotic behaviour between the Laguerre polynomials and the Sobolev polynomials $Q_{n,r}$. Intuitively one can imagine that these differences in the complex plane should be around the perturbation of the standard inner product involved in the Sobolev inner product, that is, around the origin and therefore we cannot expect that the addition of a finite number of masses to the inner product produces a modification in the global behaviour of the polynomials. A result which supports this intuition is Lemma 2 in [5] where it has been proved:

Lemma 1 *Let $Q_{n,r}$ be the polynomials orthogonal with respect to (6) with leading coefficients $(-1)^n/n!$. Then the following statements hold:*

(a) For $0 \leq k \leq r$,

$$Q_{n,r}^{(k)}(0) \sim \frac{C_{r,k}}{n^{\alpha+2k+1}}(L_n^\alpha)^{(k)}(0),$$

where $C_{r,k}$ is a nonzero real number independent of n.

(b) For $k \geq r+1$,

$$Q_{n,r}^{(k)}(0) \sim \frac{k!}{(k-(r+1))!}\frac{\Gamma(\alpha+k+1)}{\Gamma(\alpha+r+k+2)}(L_n^\alpha)^{(k)}(0)\,.$$

(c)

$$(Q_{n,r}, Q_{n,r})_r \sim \|L_n^\alpha\|^2\,.$$

Remark 1. Observe that both Laguerre and Laguerre–Sobolev type polynomials have asymptotically the same global size (from the point of view of the norm), while the size of the successive derivatives at the point $x = 0$ is affected by the discrete part of the inner product but only whenever the order of the derivatives corresponds to a positive mass.

Now we analyze two other asymptotics of the polynomials $Q_{n,r}$: the *relative asymptotics*, which assures that both families $Q_{n,r}$ and L_n^α are identical asymptotically on compact subsets of $\mathbb{C}\setminus[0,\infty)$, and the so–called Mehler–Heine type formula which shows how the presence of the masses in the inner product changes the asymptotic behaviour around the origin.

As we have mentioned before, for a discrete Sobolev inner product when $\operatorname{supp}\mu$ is bounded, a tool to obtain some results is the relation between the Sobolev orthogonality and the standard quasi–orthogonality.

Now, in our particular case, the sequence $\{Q_{n,r}\}_{n\geq 0}$ is quasi–orthogonal of order $r+1$ with respect to the Laguerre weight $x^{\alpha+r+1}e^{-x}$, that is,

$$\int_0^{+\infty} p(x)Q_{n,r}(x)x^{\alpha+r+1}e^{-x}dx = 0,$$

for every polynomial p with $\deg p \leq n-(r+1)-1$. Therefore, we have a *connexion formula* of the form

$$Q_{n,r}(x) = \sum_{j=0}^{r+1} a_{n,r}^j L_{n-j}^{\alpha+r+1}(x)\,, \quad a_{n,r}^0 = 1. \tag{7}$$

In order to deduce properties of Q_{nr} it is convenient to know the size of the *connexion coefficients* $a_{n,r}^j$. In [5], it has been introduced a fruitful and new technique which leads to determine their asymptotic behaviour.

Using (7), it can be obtained a new algebraic expression which relates $\frac{Q_{n,r}^{(k+1)}(0)}{(L_n^\alpha)^{(k+1)}(0)}$ to $\frac{Q_{n,r}^{(k)}(0)}{(L_n^\alpha)^{(k)}(0)}$ (see [5, Lemma 3]) and allows to prove:

Theorem 1 *Let $a_{n,r}^j$ be the connexion coefficients which appear in (7). Then, we have*

$$\lim_n a_{n,r}^j = (-1)^j \binom{r+1}{j}\,, \quad 0 \leq j \leq r+1\,. \tag{8}$$

As a token of the interest of this result we use it to deduce an asymptotics of the Laguerre–Sobolev polynomials on compact subsets of $(0,+\infty)$.

Proposition 1 *The sequence* $\{n^{-(2\alpha+2r+1)/4}Q_{n,r}\}_{n\geq 1}$ *is uniformly bounded on compact subsets of* $(0,+\infty)$.

Proof. The sequence $\{n^{-\alpha/2+1/4}L_n^\alpha\}_{n\geq 1}$ is uniformly bounded on compact subsets of $(0,+\infty)$ (see Theorem 8.22.1 in [23]), and then, for all $j=0,1,\ldots,r+1$, the sequences $\{n^{-(2\alpha+2r+1)/4}L_{n-j}^{\alpha+r+1}\}_{n\geq 1}$ are uniformly bounded on compact subsets of $(0,+\infty)$. ¿From (8) and the connexion formula the result follows. □

However, it is worth noticing that the knowledge of the asymptotic behaviour of the connexion coefficients is not enough to deduce other asymptotic properties. Indeed, concerning the relative asymptotics, from (7) we have

$$\frac{Q_{n,r}(x)}{L_n^\alpha(x)}=\sum_{j=0}^{r+1}a_{n,r}^j\frac{L_{n-j}^{\alpha+r+1}(x)}{L_n^\alpha(x)}.$$

Applying Theorem 1, and (4) and (5) each term in the above sum tends to infinity with the same order but with an alternating sign, that is,

$$a_{n,r}^j\frac{L_{n-j}^{\alpha+r+1}(x)}{L_n^\alpha(x)}\sim(-1)^j\binom{r+1}{j}\left(\frac{1}{\sqrt{-x}}\right)^{r+1}n^{\frac{r+1}{2}},$$

uniformly on compact subsets of $\mathbb{C}\setminus[0,\infty)$.

Since the techniques used in the bounded case do not work when $\operatorname{supp}\mu$ is an unbounded set we proceed in a different way to prove:

Theorem 2 *Let* $\{Q_{n,r}\}_{n\geq 0}$ *be the sequence of polynomials orthogonal with respect to the inner product (6) with* $(-1)^n/n!$ *as leading coefficient. Then, for* $k\geq 0$,

$$\lim_n\frac{Q_{n,r}^{(k)}(x)}{(L_n^\alpha)^{(k)}(x)}=1,$$

uniformly on compact subsets of $\mathbb{C}\setminus[0,\infty)$.

Proof. From the Fourier expansion of the polynomial Q_{n0} in terms of Laguerre polynomials, using Lemma 1, (3) and (4) the result follows for $k=0$. (For more details see Theorem 1 in [5]).

The functions $Q_{n,r}/L_n^\alpha$ are analytic in $\mathbb{C}\setminus[0,\infty)$ and $\frac{Q_{n,r}(x)}{L_n^\alpha(x)}\rightrightarrows 1$, $x\in\mathbb{C}\setminus[0,\infty)$, then $\left(\frac{Q_{n,r}}{L_n^\alpha}\right)'(x)\rightrightarrows 0\,,x\in\mathbb{C}\setminus[0,\infty)$. Therefore,

$$\left(\frac{Q_{n,r}'(x)}{(L_n^\alpha)'(x)}-\frac{Q_{n,r}(x)}{L_n^\alpha(x)}\right)\frac{(L_n^\alpha)'(x)}{L_n^\alpha(x)}\rightrightarrows 0\,,x\in\mathbb{C}\setminus[0,\infty).$$

¿From (2), (4) and (5), we get $\frac{(L_n^\alpha)'(x)}{L_n^\alpha(x)} \rightrightarrows \infty$ and then

$$\lim_n \frac{Q'_{n,r}(x)}{(L_n^\alpha)'(x)} = \lim_n \frac{Q_{n,r}(x)}{L_n^\alpha(x)} = 1$$

uniformly on compact subsets of $\mathbb{C} \setminus [0, \infty)$. So, the result holds for $k = 1$.

Using this technique, by an induction procedure, the result follows for all $k \geq 0$. □

Once we know that both sequences of orthogonal polynomials, $\{Q_{n,r}\}_{n\geq 0}$ and $\{L_n^\alpha\}_{n\geq 0}$, are asymptotically identical on compact subsets of $\mathbb{C}\setminus[0,\infty)$, we establish their differences.

To do this, we consider Mehler–Heine type formulas because they are nice tools to describe the polynomials around the origin. These kind of formulas are interesting twofold: they provide the scaled asymptotics for $Q_{n,r}$ on compact sets of the complex plane and they supply us with asymptotic information about the location of the zeros of these polynomials in terms of the zeros of other known special functions. More precisely, applying Hurwitz's Theorem in a straightforward way, the existence of an acceleration of the convergence of $r + 1$ zeros of these Sobolev polynomials towards the origin can be proved.

First of all, we recall the corresponding formula for the classical Laguerre polynomials, (see [23, Th.8.1.3]):

$$n^{-\alpha} L_n^\alpha \left(\frac{x}{n}\right) \rightrightarrows x^{-\alpha/2} J_\alpha(2\sqrt{x}),\ x \in \mathbb{C}\,, \tag{9}$$

where J_α is the Bessel function of the first kind of order α ($\alpha > -1$).

As it occurs in the study of the relative asymptotics, the Mehler–Heine type formulas cannot be deduced as a consequence of the connexion formula. Indeed, from (7) we have

$$n^{-\alpha} Q_{n,r} \left(\frac{x}{n}\right) = \sum_{i=0}^{r+1} a_{n,r}^i n^{-\alpha} L_{n-i}^{\alpha+r+1} \left(\frac{x}{n}\right).$$

and, applying Theorem 1 and (9), we have that each term tends to infinity with the same order but with an alternating sign.

Thus, to get the result for $\{Q_{n,r}\}_{n\geq 0}$, the problem should be focused on in a different way. An approach consists in to write the Taylor expansion of the polynomial $Q_{n,r}$

$$n^{-\alpha} Q_{n,r} \left(\frac{x}{n}\right) = \sum_{k=0}^{n} \frac{Q_{n,r}^{(k)}(0)}{(L_n^\alpha)^{(k)}(0)} \frac{(L_n^\alpha)^{(k)}(0)}{k!} \frac{x^k}{n^{\alpha+k}}\ ,$$

and to calculate the limit applying the Lebesgue's dominated convergence theorem. So, we need to prove that the ratios $Q_{n,r}^{(k)}(0)/(L_n^\alpha)^{(k)}(0)$ are uniformly bounded. It is clear that taking derivatives k times in (7) the connexion coefficients do not change. Then, it could be thought about the possibility to obtain this uniform bound from this formula. But again we come across the same problem, each term of $\displaystyle\sum_{i=0}^{r+1} a_{n,r}^i \frac{(L_{n-i}^{\alpha+r+1})^{(k)}(0)}{(L_n^\alpha)^{(k)}(0)}$, tends to infinity with order n^{r+1}, but with an alternating sign. To solve this problem, taking

into account the expression relating $Q_{n,r}^{(k+1)}(0)/(L_n^\alpha)^{(k+1)}(0)$ and $Q_{n,r}^{(k)}(0)/(L_n^\alpha)^{(k)}(0)$, (see [5, Lemma 3]), the neccessary uniform bound for the ratios could be derived (see [5, Lemma 4]). Then we have

Theorem 3 *Let $\{Q_{n,r}\}_{n\geq 0}$ be the sequence of polynomials orthogonal with respect to the inner product (6) with $(-1)^n/n!$ as leading coefficient. Then,*

$$\lim_n n^{-\alpha} Q_{n,r}\left(\frac{x}{n}\right) = (-1)^{r+1} x^{-\alpha/2} J_{\alpha+2r+2}(2\sqrt{x}),$$

uniformly on compact subsets of $\mathbb{C}$.

This result gives a positive answer to the conjecture posed in [8]. We would like to note that the approach is totally new and the techniques used in [5] to prove the above Theorem are not a simple generalization of the ones used in [8].

Next, we will show a remarkable difference between the zeros of L_n^α and the ones of $Q_{n,r}$ concerning the convergence acceleration to 0. First, we recall (see [23]) that the zeros of the Laguerre polynomials are real, simple and they are located in $(0,\infty)$. Denote by $(x_{n,k})_{k=1}^n$ the zeros of L_n^α in an increasing order, they satisfy the interlacing property $0 < x_{n+1,1} < x_{n,1} < x_{n+1,2} < \ldots$, and $x_{n,k} \underset{n}{\to} 0$ for each fixed k.

Let $(j_{\alpha,k})_{k\geq 1}$ be the positive zeros of the Bessel function J_α writing in an increasing order. Then, formula (9) and Hurwitz's theorem lead us to $n x_{n,k} \underset{n}{\to} j_{\alpha,k}$, $k \geq 1$, and therefore $x_{n,k} \sim \frac{C_k}{n}$, $k \geq 1$, where C_k is a positive constant depending on k.

Concerning the zeros of $Q_{n,r}$, standard arguments (see for instance [10]) allow to establish that $Q_{n,r}$ has at least $n-(r+1)$ zeros with odd multiplicity in $(0,+\infty)$, or equivalently $n-(r+1)$ changes of sign. Moreover, since $M_0 > 0$ and the mass point in the discrete part of the inner product belongs to the boundary of $(0,+\infty)$ then the number of zeros with odd multiplicity is at least $n-r$ (see [2]).

¿From Theorem 3 and Hurwitz's theorem, taking into account the multiplicity of 0 as a zero of the limit function in Theorem 3, we achieve

Corollary 1 *Let $(\xi_{n,k}^r)_{k=1}^n$ be the zeros of $Q_{n,r}$. Then*

$$n\,\xi_{n,k}^r \underset{n}{\to} 0, \quad 1 \leq k \leq r+1,$$

$$n\,\xi_{n,k}^r \underset{n}{\to} j_{\alpha+2r+2,k-r-1}, \quad k \geq r+2.$$

Remark 2. The presence of the positive masses M_i, $i = 0,\ldots,r$, in the inner product produces a convergence acceleration to 0 of $r+1$ zeros of the polynomials $Q_{n,r}$.

3 Laguerre–Sobolev inner products with holes

Until now, we have assume that all the masses M_i in the discrete part of the Sobolev inner product are positive. The possibility of some $M_i = 0$ has been also dealed in the literature. For instance, the case $M_0 = 0, M_1 > 0$ ([8]) and similar situations in the non–diagonal case ([7] and [11]) have been analyzed. Very recently, in [12], the authors study the particular case $M_i = 0$, $i = 0, \ldots, r-1$, for the Laguerre–Sobolev type polynomials. The results obtained in all these papers have been generalized in [5], where such a kind of inner products have been called Sobolev inner products with *holes*.

More concretely, we consider the inner product

$$(f,g)_{r,s} = \tag{10}$$
$$\frac{1}{\Gamma(\alpha+1)}\int_0^\infty f(x)g(x)x^\alpha e^{-x}dx + \sum_{i=0}^{r} M_i f^{(i)}(0)g^{(i)}(0) + M_s f^{(s)}(0)g^{(s)}(0),$$

where $s \geq r+2$ and $M_i > 0$ for $i = 0, \ldots, r$ and $i = s$.

Observe that we are concerned with inner products of the form

$$(p,q)_{r,s} = (p,q)_r + M_s p^{(s)}(0)q^{(s)}(0), \quad s \geq r+2,$$

where $M_s > 0$, and in $(.,.)_r$ all the masses are positive. That is, roughly speaking, there is a "hole" in the discrete part of the inner product $(.,.)_{r,s}$. We denote by $\{T_{n,r,s}\}_{n\geq 0}$ the sequence of polynomials orthogonal with respect to the inner product $(.,.)_{r,s}$ with leading coefficients $(-1)^n/n!$.

For this situation, the relative asymptotics and the Mehler-Heine type formulas have been established in [5]. We want to remark that this case has qualitative differences with respect to the case without holes. For example, concerning the convergence acceleration to 0 of the zeros of the polynomials, as we will below.

Arguing as in Lemma 1 it can be proved

Lemma 2 *Let $\{T_{n,r,s}\}_{n\geq 0}$ be the sequence of polynomials orthogonal with respect to the inner product (10) with $(-1)^n/n!$ as leading coefficient . Then the following statements hold:*

(a) For either $0 \leq k \leq r$ or $k = s$,

$$T^{(k)}_{n,r,s}(0) \sim \frac{C_{r,s,k}}{n^{\alpha+2k+1}}\,(L^\alpha_n)^{(k)}(0)\,,$$

where $C_{r,s,k}$ is a nonzero real number independent of n.

(b) For $k \geq r+1$ and $k \neq s$

$$T^{(k)}_{n,r,s}(0) \sim \frac{k!}{(k-(r+1))!}\frac{k-s}{\alpha+s+k+1}\frac{\Gamma(\alpha+k+1)}{\Gamma(\alpha+r+k+2)}\,(L^\alpha_n)^{(k)}(0)\,.$$

(c)

$$(T_{n,r,s}T_{n,r,s})_{r,s} \sim \|L_n^\alpha\|^2 .$$

Observe that, as in the complete case (without holes), the addition of the discrete part of the inner product modifies the size of the derivative of order k only when the corresponding mass M_k is positive.

Using this lemma the relative asymptotics for these orthogonal polynomials can be deduced:

Theorem 4 *Let $\{T_{n,r,s}\}_{n\geq 0}$ be the sequence of o.p. with respect to the inner product (10) with $(-1)^n/n!$ as leading coefficient . Then*

$$\frac{T_{n,r,s}(x)}{L_n^\alpha(x)} \rightrightarrows 1\,, \quad x \in \mathbb{C}\setminus[0,\infty).$$

The Mehler–Heine type formula adopts the form

Theorem 5 *Let $\{T_{n,r,s}\}_{n\geq 0}$ be the sequence of polynomials orthogonal with respect to the inner product (10) with $(-1)^n/n!$ as leading coefficient . Then,*

$$n^{-\alpha}T_{n,r,s}\left(\frac{x}{n}\right) \rightrightarrows (-1)^{r+1}x^{-\alpha/2} \tag{11}$$
$$\times\left[\frac{-(s-(r+1))}{\alpha+r+s+2}J_{\alpha+2r+2}(2\sqrt{x}) + \sum_{l=2}^{s-r+1}\lambda_i J_{\alpha+2r+2l}(2\sqrt{x})\right],\ x\in\mathbb{C}\,,$$

where λ_i are nonzero real numbers.

For the particular case $s = r+2$, i.e., when there is a hole of "length one", the above result generalizes the one obtained in [8]. Theorem 5 also generalizes the corresponding result in [12].

Now, we comment the acceleration of the convergence towards the origin of the zeros of the polynomials $T_{n,r,s}$. The quasi–orthogonality of order $s+1$ of the sequence $\{T_{n,r,s}\}_{n\geq 0}$ with respect to the positive weight $x^{\alpha+s+1}e^{-x}$ assures that $T_{n,r,s}$ has at least $n-(s+1)$ changes of sign in $(0,+\infty)$. However, in [2] the authors proved that the number of zeros in $(0,+\infty)$ does not depend on the order of the derivatives but on the number of terms in the discrete part of the inner product. So, $T_{n,r,s}$ has at least $n-(r+1)$ zeros with odd multiplicity in $(0,+\infty)$. Proceeding as in Corollary 1, we get:

Corollary 2 *Let $(\zeta_{n,k}^{r,s})_{k=1}^n$ be the zeros of $T_{n,r,s}$. Then*

$$n\,\zeta_{n,k}^{r,s} \underset{n}{\to} 0, \quad 1\leq k\leq r+1,$$

$$n\,\zeta_{n,k}^{r,s} \underset{n}{\to} j_{\alpha+2r+2,k-r-1}, \quad k\geq r+2.$$

Remark 3. We want to highlight that this result is in a way surprising since it does not depend on the number of terms in the discrete part, but on the position of the hole. So, despite the presence of the mass M_s, there only exists an acceleration of the convergence of $r+1$ zeros such as it occurs in the case of the inner products without holes. That is, the convergence acceleration to 0 of the zeros of the polynomials $Q_{n,r}$ and $T_{n,r,s}$ is the same and the addition of a mass M_s *after a hole* in the inner product does not affect the convergence acceleration to 0.

4 Generalized Hermite–Sobolev type polynomials

As a consequence of the previous results, asymptotic properties for the orthogonal polynomials $S^{\mu}_{n,r}$ associated with the inner product

$$(p,q) = \int_{\mathbb{R}} p(x)q(x)|x|^{2\mu}\, e^{-x^2} dx + \sum_{i=0}^{2r+1} M_i\, p^{(i)}(0)\, q^{(i)}(0), \tag{12}$$

with $\mu > -1/2$ and $M_i > 0$, $i = 0, \ldots, 2r+1$, can be established. We assume that the leading coefficient of $S^{\mu}_{n,r}$ is 2^n.

Remind that the polynomials H^{μ}_n orthogonal with respect to the weight $|x|^{2\mu}\, e^{-x^2}$ ($\mu > -1/2$) are called *generalized Hermite polynomials*. So, we are concerned with generalized Hermite–Sobolev type orthogonal polynomials.

Notice that in this case the polynomials $S^{\mu}_{n,r}$ are symmetric, that is, $S^{\mu}_{n,r}(-x) = (-1)^n\, S^{\mu}_{n,r}(x)$, and because of this symmetry, we can transform the inner product (12) into an inner product like (6) and so we can establish a simple relation between the polynomials $S^{\mu}_{n,r}$ and the polynomials $Q_{n,r}$ considered before. This technique is known as a symmetrization process. In fact, in [10] this process is considered for standard inner products associated with positive measures. The simplest case of this situation is the relation between Laguerre polynomials and Hermite polynomials, that is (see [10] or [23]), for $n \geq 0$,

$$H_{2n}(x) = (-1)^n\, 2^{2n}\, n!\, L_n^{-1/2}(x^2),\; H_{2n+1}(x) = (-1)^n\, 2^{2n+1}\, n!\, x L_n^{1/2}(x^2)\,.$$

Later in [3] the authors generalize the symmetrization process in the framework of Sobolev type orthogonal polynomials, (see Theorem 2 in [3]). Thus,

$$S^{\mu}_{2n,r}(x) = (-1)^n\, 2^{2n}\, n!\, Q^{\mu-1/2}_{n,r}(x^2),\; S^{\mu}_{2n+1,r}(x) = (-1)^n\, 2^{2n+1}\, n!\, x Q^{\mu+1/2}_{n,r}(x^2)$$

where $\{Q^{\mu-1/2}_{n,r}\}_{n\geq 0}$ (respectively, $\{Q^{\mu+1/2}_{n,r}\}_{n\geq 0}$) is the sequence of polynomials orthogonal with respect to an inner product like (6) with $\alpha = \mu - 1/2$ (respectively, $\alpha = \mu + 1/2$) and leading coefficient $(-1)^n/n!$.

Using this symmetrization process, the relative asymptotics and the Mehler–Heine type formulas for generalized Hermite-Sobolev type polynomials can be proved.

Proposition 2 *Let $\{S^{\mu}_{n,r}\}_{n\geq 0}$ be the sequence of polynomials orthogonal with respect to the inner product (12) with 2^n as leading coefficient . Then,*

(a)

$$\frac{S^{\mu}_{n,r}(x)}{H^{\mu}_{n}(x)} \rightrightarrows 1,\ x\in\mathbb{C}\setminus\mathbb{R}.$$

(b)

$$n^{-\mu+1/2}\,S^{\mu}_{2n,r}\left(\frac{x}{2\sqrt{n}}\right) \rightrightarrows (-1)^{r+1}\left(\frac{x}{2}\right)^{-\mu+1/2} J_{\mu+2r+3/2}(x),\ x\in\mathbb{C}$$

$$n^{-\mu+1/2}\,S^{\mu}_{2n+1,r}\left(\frac{x}{2\sqrt{n}}\right) \rightrightarrows (-1)^{r+1}\left(\frac{x}{2}\right)^{-\mu+1/2} J_{\mu+2r+5/2}(x),\ x\in\mathbb{C}.$$

Remark 4. These results generalize some of the results in [4] and solve the conjecture posed there.

Using a symmetrization process, relative asymptotics and Mehler–Heine type formulas for generalized Hermite–Sobolev polynomials with holes in the discrete part of the inner product can be deduced.

Finally, we hope this method can be used with other measures with unbounded support for which we have quite less explicit information about the corresponding orthogonal polynomials.

Acknowledgements

This work has been partially supported by MICINN of Spain under Grant MTM2009-12740-C03-03, FEDER funds (EU), and the Diputación General de Aragón, project E-64.

References

[1] M. Alfaro, M. Álvarez de Morales and M.L. Rezola, *Orthogonality of the Jacobi polynomials with negative parameters*, J. Comput. Appl. Math. **145** (2002), 379–386.

[2] M. Alfaro, G. López and M.L. Rezola, *Some properties of zeros of Sobolev-type orthogonal polynomials*, J. Comput. Appl. Math. **69** (1996), 171–179.

[3] M. Alfaro, F. Marcellán, H.G. Meijer and M.L. Rezola, *Symmetric orthogonal polynomials for Sobolev–type inner products*, J. Math. Anal. Appl. **184** (1994), 360–381.

[4] M. Alfaro, J.J. Moreno–Balcázar, A. Peña and M.L. Rezola, *Asymptotics for a generalization of Hermite polynomials*. Asymptotic Anal. **66** (2010), 103-117.

[5] M. Alfaro, J.J. Moreno–Balcázar, A. Peña and M.L. Rezola, *A new approach to the asymptotics for Sobolev orthogonal polynomials*. Submitted for publication, 2010. arXiv 1003.3336v1

[6] R. Álvarez–Nodarse and F. Marcellán, *A generalization of the classical Laguerre polynomials*, Rend.Circ. Mat. Palermo (2) **44** (1995), 315–329.

[7] R. Álvarez–Nodarse and F. Marcellán, *A generalization of the classical Laguerre polynomials: asymptotic properties and zeros*, Appl. Anal. **62** (1996), 349–366.

[8] R. Álvarez–Nodarse and J.J. Moreno–Balcázar, *Asymptotic properties of generalized Laguerre orthogonal polynomials*, Indag. Mathem., N.S. **15** (2004), 151–165.

[9] K.E. Atkinson and O. Hansen, *Solving the nonlinear Poisson equation on the unit disk*, J. Integral Eq. and Appl., **17** (2006), 223–241.

[10] T.S. Chihara, *An Introduction to Orthogonal Polynomials*. Gordon & Breach, New York, 1978.

[11] H. Dueñas and F. Marcellán, *Asymptotic behaviour of the Laguerre–Sobolev–type orthogonal polynomials. A nondiagonal case*, J. Comput. Appl. Math. (2009), doi:10.1016/j.cam.2009.07.055.

[12] H. Dueñas and F. Marcellán, *The Laguerre–Sobolev–Type Orthogonal Polynomials*, J. Approx. Theory **162** (2010), 421-440.

[13] D. Evans, L.L. Littlejohn, F. Marcellán, C. Markett and A. Ronveaux *On recurrence relations for Sobolev polynomials*, SIAM J. Math. Anal. **26** (1995), 446-467.

[14] L. Fernández, F. Marcellán, T.E. Pérez and M.A. Piñar *Recent trends on two variable orthogonal polynomials*, Contemp. Math. **509** (2010), 59-86.

[15] R. Koekoek, *Generalizations of Laguerre polynomials*, J. Math. Anal. Appl. **153** (1990), 576–590.

[16] R. Koekoek and H.G. Meijer, *A generalization of Laguerre polynomials*, SIAM J. Math. Anal. **24** (1993), 768–782.

[17] K.H. Kwon and L.L. Littlejohn, *The orthogonality of the Laguerre polynomials* $\{L_n^{(-k)}(x)\}$ *for a positive integer* k, Ann. Numer. Math. **2** (1995), 289–304.

[18] K.H. Kwon, L.L. Littlejohn and G.J. Yoon, *Ghost matrices and a characterization of symmetric Sobolev bilinears forms*, Linear Algebra Appl. **431** (2009), 104–119.

[19] J.K. Lee and L. L. Littlejohn *Sobolev orthogonal polynomials in two variables and second order partial differential equations*, J. Math. Anal. Appl. 322 (2006), 1001–1017.

[20] G. López, F. Marcellán and W. Van Assche, *Relative asymptotics for polynomials orthogonal with respect to a discrete Sobolev inner product*, Constr. Approx. **11** (1995), 107–137.

[21] F. Marcellán and J.J. Moreno–Balcázar, *Asymptotics and zeros of Sobolev orthogonal polynomials on unbounded supports*, Acta Appl. Math. **94** (2006), 163–192.

[22] I.A. Rocha, F. Marcellán and L. Salto, *Relative asymptotics and Fourier series of orthogonal polynomials with a discrete Sobolev inner product*, J. Approx. Theory **121** (2003), 336–356.

[23] G. Szegő, *Orthogonal Polynomials*, Amer. Math. Soc. Colloq. Publ. vol. **23**, Amer. Math. Soc., Providence R.I., 1975. Fourth Edition.

[24] Y. Xu, *A family of Sobolev orthogonal polynomials on the unit ball*, J. Approx. Theory, **138** (2006), 232–241.

Monografías de la Real Academia de Ciencias de Zaragoza **33**, 225–232, (2010).

A note on typical sections of isotropic convex bodies

David Alonso-Gutiérrez, Jesús Bastero, Julio Bernués

IUMA, Facultad de Ciencias, Universidad de Zaragoza, 50009 Zaragoza, Spain

Abstract

Let $K \subset \mathbb{R}^n$ be a centrally symmetric isotropic convex body. We prove that for random $F \in G_{n,k}$, and k slowly growing to infinity, the central section $|F^\perp \cap K|_{n-k}^{1/k}$ is almost constant. A simple approach using standard concentration of measure arguments is given.

1 Introduction and notation

Let $K \subset \mathbb{R}^n$ be a symmetric convex body. We say K is isotropic if it is of volume 1 and there exists a constant $L_K > 0$ called isotropy constant of K such that $L_K^2 = \int_K \langle x, \theta \rangle^2 \, dx, \forall \theta \in S^{n-1}$.

Since the works of [H], [B] or [MP] we know of the close relation between the isotropy constant and the size of the central sections of K. It is well known that for any $1 \leq k \leq n$ there exist $c_1(k), c_2(k) > 0$ such that for every subspace $F \in G_{n,k}$ (the Grassmann space)

$$\frac{c_1(k)}{L_K} \leq |F^\perp \cap K|_{n-k}^{1/k} \leq \frac{c_2(k)}{L_K}$$

where $|\cdot|_m$ is the Lebesgue measure in the appropiate m dimensional space.

Well known estimates (see [H], [MP] and [Kl]) imply $c_1(k) \geq c_1$ and $c_2(k) \leq c_2 k^{1/4}$, where $c_1, c_2 > 0$ are absolute numerical constants. These bounds are the best ones known to be valid for *every* subspace $F \in G_{n,k}$.

For random sections, much better estimates are possible. The following result was proved in [ABBP],

There exist absolute constants $c_1, c_2, c_3 > 0$ with the following property: If K is an isotropic convex body in $\mathbb{R}^n$ and $1 \leqslant k \leqslant \sqrt{n}$ then, the set of subspaces $F \in G_{n,k}$ such that

$$\frac{c_1}{L_K} \leqslant |K \cap F^\perp|_{n-k}^{1/k} \leqslant \frac{c_2}{L_K}$$

has Haar probability $\geq 1 - e^{-c_3 \frac{n}{k}}$

In [EK] the authors prove a version of the central limit theorem for convex bodies. Its proof uses the strong concentration behavior of the Euclidean norm on K, [Kl2], and a delicate study of the marginal distribution of some intermediate measures, namely the convolution of the uniform measure on K with an independent gaussian vector. As a consequence of it it is easy to check that

For $\varepsilon = \frac{1}{n^{c_1}}$, $k \leq n^{c_2}$ the set of subspaces $F \in G_{n,k}$ such

$$\frac{1-\varepsilon}{\sqrt{2\pi}L_K} \leq |K \cap F^\perp|_{n-k}^{1/k} \leq \frac{1+\varepsilon}{\sqrt{2\pi}L_K}$$

has Haar probability $\geq 1 - c_3 e^{-n^{c_4}}$.

These two results are different: the second one gives better constants ($\sim \frac{1}{\sqrt{2\pi}}$) but a worse dependence on k and on the estimate of the Haar probability.

In this note we use a simpler approach to the question. Our final result is weaker in k than the one deduced from [EK] and it provides better estimate of the Haar probability. But the main advantadge, we think, is that the arguments are simpler and the tools used are of independent interest: First we estimate Lipschitz constant of the section function $F \in G_{n,k} \to |F^\perp \cap K|_{n-k}$ (Proposition 2.3), for $k = 1$ this was proved in [ABP]. Then we apply the concentration of measure phenomenum on $G_{n,k}$ (equipped with the right distance (Proposition 2.2)). In this way we measure the closeness between the section function and its expectation. Finally, by expressing this expectation as a marginal, we related it to the marginal of a gaussian distribution. For that final step, we unavoidably use the concentration of the Euclidean norm on K, [Kl2] in the version stated in [BB]. Our result is

Theorem 2.8. *Let $K \subset \mathbb{R}^n$ isotropic. For all $\varepsilon > 0$, $1 \leq k \leq \frac{c\varepsilon \log n}{(\log\log n)^2}$, the set A of subspaces $F \in G_{n,k}$ such that*

$$\frac{1-\varepsilon}{\sqrt{2\pi}L_K} \leq |K \cap F^\perp|_{n-k}^{1/k} \leq \frac{1+\varepsilon}{\sqrt{2\pi}L_K} \tag{1.1}$$

holds, has probability $\mu(A) \geq 1 - c_1 e^{-c_2 n^{0.9}}$.

In $\mathbb{R}^n$, $|\cdot|$ denotes the Euclidean norm and B_2^n the Euclidean ball. For any k-dimensional subspace $F \subset \mathbb{R}^n$ we denote $S_F = S^{n-1} \cap F$ and by P_F the orthogonal projection onto F. $G_{n,k}$ is the grassmaniann space of all k dimensional subspaces of $\mathbb{R}^n$ and its Haar probability is denoted by μ. For any linear map T from $\mathbb{R}^n$, $\|T\|$ denotes the operator norm and $\|T\|_{HS} := \Big(\sum_{j=1}^n |T(e_j)|^2\Big)^{1/2}$, for (any) orthonormal basis (e_j) of $\mathbb{R}^n$, its Hilbert-Schmidt norm.

2 The result

In the first part we estimate the Lipschitz constant of the function $F \to |F^\perp \cap K|_{n-k}$ and also review concentration inequalities with respect to several natural distances on $G_{n,k}$. We start with the latter.

The following lemma constructs a suitable orthonormal basis for two subspaces E and F and will be very useful for our purposes

Lemma 2.1 ([GM], Lemma 4.1) *Let $E, F \in G_{n,k}$ such that $F^\perp \cap E = 0$. Then there exists $u_1, \dots u_k$ orthonormal basis of E such that the family $v_1, \dots v_k$ given by $v_j = \frac{P_F(u_j)}{|P_F(u_j)|}$ is an orthonormal basis of F. In particular, $\langle u_j, v_i \rangle = |P_F(u_j)|\ \delta_i^j$.*

The space $G_{n,k}$ appears in the literature equipped with a number of different distances. In the following Proposition, we estimate the equivalence constants between them. It is probably folklore but we include for the reader's convenience. The fact that one can move from one distance to another will be useful while computing the Lipschitz constant and also when considering the concentration phenomena on $G_{n,k}$.

The elements of the orthogonal group $O(n)$ will be denoted by $U = (u_1 \dots u_n)$ so the columns (u_i) form an orthonormal basis in $\mathbb{R}^n$.

Proposition 2.2 *For $E, F \in G_{n,k}$ we consider the following distances*

$d_0(E,F) = \max\{d(x, S_F) \mid x \in S_E\}$, *$d$ is the euclidean distance.*

$d_1(E,F) = \inf\{\varepsilon > 0 \mid S_E \subset S_F + \varepsilon B_2^n, S_F \subset S_E + \varepsilon B_2^n\}$

$d_2(E,F) = \inf\{\Big(\sum_{j=1}^k |u_j - v_j|^2\Big)^{1/2} E = \langle u_j\rangle_1^k,\ F = \langle v_j\rangle_1^k \text{ orthon. basis}\}$

$d_3(E,F) = \inf\{\Big(\sum_{j=1}^n |u_j - v_j|^2\Big)^{1/2} E = \langle u_j\rangle_1^k,\ F = \langle v_j\rangle_1^k \text{ orthon. basis}\}$

$d_4(E,F) = \|P_E - P_F\|_{HS}$

$d_5(E,F) = \inf\{\|U - V\|_{HS} \mid U, V \in O(n), E = \langle u_1 \dots u_k\rangle, F = \langle v_1 \dots v_k\rangle\}$

$d_6(E,F) = \|P_E - P_F\|$

Then, d_2, d_3, d_4, d_5 are equivalent with numerical equivalence constants, $d_0 = d_1$, $d_1 \le d_2 \le \sqrt{2k}\ d_1$ and $d_6 \le d_4 \le \sqrt{2k} d_6$.

$d_0 = d_1$: d_1 is the Hausdorff distance between S_E and S_F which also reads

$$d_1(E,F) = \max\Big\{\max_{x\in S_E} d(x, S_F), \max_{y \in S_F} d(y, S_E)\Big\}$$

so $d_0 \le d_1 \le \sqrt{2}$ and it is enough to check that the two inner maxima are equal.

If $E \cap F^\perp \neq 0$ then $d_0(E,F) = \sqrt{2}$. Suppose $E \cap F^\perp = 0$. For any $x \in S_E, y \in S_F$, $|x-y|^2 = 2 - 2\langle x, y\rangle = 2 - 2\langle P_F(x), y\rangle$. So, $d^2(x, S_F) = 2 - 2\sup_{y\in S_F}\langle P_F(x), y\rangle = 2 -$

$2|P_F(x)| = \big|x - \frac{P_F(x)}{|P_F(x)|}\big|^2$. Let $x_0 \in S_E$ that maximizes $d(x, S_F)$ on S_E or equivalently that minimizes $|P_F(x)|$. Denote $y_0 = \frac{P_F(x_0)}{|P_F(x_0)|}$ (observe $P_F(x_0) \neq 0$). By the arguments in [GM] Lemma 4.1, $P_F(x_0)$ is orthogonal to $E \cap x_0^\perp$ and so $P_E P_F(x_0)$ is parallel to x_0. Write $P_E(y_0) = \lambda x_0$. Then $\lambda = \langle P_E(y_0), x_0\rangle = \langle y_0, P_E(x_0)\rangle = |P_F(x_0)|$ and $\frac{P_E P_F(x_0)}{|P_E P_F(x_0)|} = x_0$. Therefore, $d(y_0, S_E) = d(x_0, S_F)$ and so $\max\{d(y, S_E) \mid y \in S_F\} \geq \max\{d(x, S_F) \mid x \in S_E\}$. Exchange E, F and equality follows.

$d_1 \leq d_2 \leq \sqrt{2k}\, d_1$: It is proved in [GM], Lemma 4.1.

$\frac{1}{\sqrt{2}} d_2 \leq d_4 \leq \sqrt{2}\, d_2$: Let $F^\perp \cap E := E_0$ and write the orthogonal decomposition $E = E_0 \oplus E_1$ with $E_1 \cap F^\perp = 0$. By Lemma 2.1, there exists an orthonormal basis in E_1, (u_j), such that $v_j = \frac{P_F(u_j)}{|P_F(u_j)|}$ is an orthonormal system in F. Now add vectors to complete an orthonormal basis in E (by adding vectors in E_0) and in F that we also denote as u_j and v_j. Trivially,

$$\|P_E - P_F\|_{HS}^2 \geq \sum_{j=1}^k |(P_E - P_F)(u_j)|^2$$

If $u_j \in E_1$ then, since $\langle u_j, v_j\rangle = |P_F(u_j)|$ (Lemma 2.1),

$$|(P_E - P_F)(u_j)|^2 = 1 - |P_F(u_j)|^2 \geq 1 - |P_F(u_j)| = \frac{1}{2}|u_j - v_j|^2$$

If $u_j \in E_0$ and $v_j \in F$ then $|(P_E - P_F)(u_j)|^2 = 1$. Also, since $\langle u_j, v_j\rangle = 0$ and so $|u_j - v_j|^2 = 2$.

For the second inequality, let $(u_j), (v_j)$ be orthonormal basis of $E, F \in G_{n,k}$ we write $P_E = \sum_{j=1}^k u_j \otimes u_j$ and $P_F = \sum_{i=1}^k v_i \otimes v_i$ and by definition

$$\|P_E - P_F\|_{HS}^2 = 2k - 2\sum_{i,j=1}^k \langle u_j, v_i\rangle^2 \leq 2\sum_{j=1}^k (1 - \langle u_j, v_j\rangle^2) \leq 2\sum_{j=1}^k |u_j - v_j|^2$$

since $1 - \langle u_j, v_j\rangle^2 \leq 2(1 - \langle u_j, v_j\rangle) = |u_j - v_j|^2$.

$d_2 \leq d_3 \leq \sqrt{5} d_2$: By definition $d_3^2(E, F) = d_2^2(E, F) + d_2^2(E^\perp, F^\perp)$. Now, $d_2^2(E^\perp, F^\perp) \leq 2d_4^2(E^\perp, F^\perp) = 2d_4^2(E, F) \leq 4d_2^2(E, F)$. With similar arguments one proves $d_2 \leq d_5 \leq 3d_2$.

$d_6 \leq d_4 \leq \sqrt{2k} d_6$: For T linear $\|T\| \leq \|T\|_{HS} \leq \sqrt{\dim(T(\mathbb{R}^n))}\|T\|$. □

Proposition 2.3 *Let $K \subset \mathbb{R}^n$ isotropic. The function given by $G_{n,k} \ni E \to |E^\perp \cap K|_{n-k}$ is Lipschitz and for all $E, F \in G_{n,k}$ we have the estimate*

$$\big|\, |E^\perp \cap K|_{n-k} - |F^\perp \cap K|_{n-k}\big| \leq \frac{(c\mathcal{L}_k)^{2k}}{L_K^k} \|P_E - P_F\|_{HS}$$

where $\mathcal{L}_k := \sup\{L_M \mid M \subset \mathbb{R}^k,$ convex body isotropic$\}$.

In order to prove it, one more lemma will be used. An equivalent version of it for $k = 1$ is due to Busemann.

Lemma 2.4 ([B]) *If $K \subset \mathbb{R}^n$ is a convex body and $E \in G_{n,k}$ then the function given by*

$$E^\perp \ni \theta \to \|\theta\| := \frac{|\theta|}{|K \cap E(\theta)|_{k+1}}$$

is a norm on $E^\perp$ where $E(\theta) = E \oplus \langle\theta\rangle$.

Proof of Proposition 2.3. Suppose $F^\perp \cap E = 0$ and let $E = \langle u_1 \dots u_k\rangle, F = \langle v_1 \dots v_k\rangle$ be the orthonormal basis in Lemma 2.1. Denote $E_0^\perp = E^\perp$, $E_j^\perp = v_1^\perp \cap \dots \cap v_j^\perp \cap u_{j+1}^\perp \cap \dots \cap u_k^\perp$ and $E_k^\perp = F^\perp$. Then

$$\big|\, |E^\perp \cap K|_{n-k} - |F^\perp \cap K|_{n-k}\big| \le \sum_{j=1}^k \big|\, |E_j^\perp \cap K|_{n-k} - |E_{j-1}^\perp \cap K|_{n-k}\big|$$

Let us estimate (say) the first summand. Set $\bar{E} = E^\perp \cap v_1^\perp = E_1^\perp \cap u_1^\perp$. Then, by Lemma 2.1, $E^\perp = \bar{E} \oplus P_{E^\perp}(v_1)$ and $E_1^\perp = \bar{E} \oplus P_{E_1^\perp}(u_1)$ so we can apply Lemma 2.4 to $\bar{E}$

$$\big|\, |E^\perp \cap K|_{n-k} - |E_1^\perp \cap K|_{n-k}\big| = \left| \frac{|P_{E^\perp}(v_1)|}{\|P_{E^\perp}(v_1)\|} - \frac{|P_{E_1^\perp}(u_1)|}{\|P_{E_1^\perp}(u_1)\|} \right|$$

and since $|P_{E_1}(u_1)| = |\langle u_1, v_1\rangle| = |P_E(v_1)|$ and the triangle inequality,

$$\left| \frac{|P_{E^\perp}(v_1)|}{\|P_{E^\perp}(v_1)\|} - \frac{|P_{E_1^\perp}(u_1)|}{\|P_{E_1^\perp}(u_1)\|} \right| \le \frac{|P_{E_1^\perp}(u_1)|}{\|P_{E_1^\perp}(u_1)\|\,\|P_{E^\perp}(v_1)\|} \|P_{E_1^\perp}(u_1) - P_{E^\perp}(v_1)\|$$

Finally, observe that $|P_{E_1^\perp}(u_1) - P_{E^\perp}(v_1)| = (1 - \langle u_1, v_1\rangle)|u_1 - v_1|$ and apply Hensley's estimate [H] to conclude with

$$\big|\, |E^\perp \cap K|_{n-k} - |E_1^\perp \cap K|_{n-k}\big| \le \frac{(1 - \langle u_1, v_1\rangle)}{(1 - \langle u_1, v_1\rangle^2)^{1/2}} |u_1 - v_1| \frac{(c\mathcal{L}_k)^{2k}}{L_K^k}$$

Since we can also suppose $\langle u_1, v_1\rangle \ge 0$, the first quotient above is bounded by 1. So,

$$\big|\, |E^\perp \cap K|_{n-k} - |F^\perp \cap K|_{n-k}\big| \le \sqrt{k}\Big(\sum_{j=1}^k |u_j - v_j|^2\Big)^{1/2} \frac{(c\mathcal{L}_k)^{2k}}{L_K^k}$$

By the proof of Proposition 2.2, $\Big(\sum_{j=1}^k |u_j - v_j|^2\Big)^{1/2} \le \sqrt{2}\|P_E - P_F\|_{HS}$. In the general case, if $F^\perp \cap E := E_0$ then we can write $E = E_0 \oplus E_1$ with $E_1 \cap F^\perp = 0$. Choose an orthonormal basis of E_0 and proceed as in the previous case. □

We recall the following celebrated result by M. Gromov and V. Milman, see for instance [MS].

Theorem 2.5 (Concentration of measure) *There exist absolute constants $c_1, c_2 > 0$ such that*

i) For every $A \subset G_{n,k}$ and every $\delta > 0$

$$\mu(A_\delta) \geq 1 - \frac{c_1}{\mu(A)} \exp\left(-c_2 \delta^2 n\right)$$

where $A_\delta = \{E \in G_{n,k};\ \exists\ F \in A, d_5(E,F) \leq \delta\}$

ii) For $f\colon G_{n,k} \to \mathbb{R}$ a Lipschitz function with Lipschitz constant σ, that is $|f(E) - f(F)| \leq \sigma d_5(E,F)$,

$$\mu\left\{E \in G_{n,k}; |f(E) - \mathbb{E}(f)| \leq a\right\} \geq 1 - c_1 \exp\left(-\frac{c_2 a^2 n}{\sigma^2}\right) \qquad \forall\ a > 0$$

Remark 2.6 *If $d, \tilde{d}$ are two distances on $G_{n,k}$ such that $d \leq M\tilde{d}$ for some $M > 0$ then a concentration inequality for $\tilde{d}$ with bound $c_1 \exp(-c_2\delta^2 n)$ implies one for d with bound $c_1 \exp\left(\frac{-c_2\delta^2 n}{M^2}\right)$. Similarly for Lipschitz functions. It is then possible to state concentration inequalities for the different distances (Proposition 2.2) on $G_{n,k}$.*

The last main ingredient is the concentration of $|\cdot|$ on K

Theorem 2.7 *[Kl2]. Let $K \subset \mathbb{R}^n$ be an isotropic convex body. Then,*

$$|\{x \in K : \big||x| - \sqrt{n}L_K\big| > t\sqrt{n}L_K\}|_n \leq c\exp(-Cn^\alpha t^\beta) \tag{2.2}$$

for all $0 \leq t \leq 1$ and $\alpha = 0.33, \beta = 3.33$.

It was proved by [So] (with sharp exponents α and β) for normalized unit balls of $\ell_p^n, 1 \leq p$ and in full generality in [Kl2].

As an application of the results we show the announced

Theorem 2.8 *Let $K \subset \mathbb{R}^n$ isotropic. For all $\varepsilon > 0$, $1 \leq k \leq \frac{\varepsilon \log n}{(\log\log n)^2}$, the set A of subspaces $E \in G_{n,k}$ such that*

$$\frac{1-\varepsilon}{\sqrt{2\pi}L_K} \leq |E^\perp \cap K|_{n-k}^{1/k} \leq \frac{1+\varepsilon}{\sqrt{2\pi}L_K}$$

holds, has probability $\mu(A) \geq 1 - c_1 \exp -c_2 n^{0.9}$

Consider the function $f\colon G_{n,k} \to \mathbb{R}$, $f(E) = |E^\perp \cap K|_{n-k}$. By Proposition 2.3 and Theorem 2.5 we have

$$\mu\left\{E \in G_{n,k}; |f(E) - \mathbb{E}(f)| \leq \varepsilon \mathbb{E}(f)\right\} \geq 1 - c_1 \exp\left(-\frac{c_2^k L_K^{2k} (\mathbb{E}(f))^2 \varepsilon^2 n}{(\mathcal{L}_k)^{2k}}\right)$$

On the other hand, denote (as in [BB]) $F_K(t,E) := \big|\{x \in K : |P_E(x)| \leq t\}\big|$, $t \geq 0$, the marginal measure on E of the euclidean ball of radius t and $\Gamma_K^k(t)$ the k-dimensional

Gaussian measure (centered with variance L_K^2) of $\{s \in \mathbb{R}^k : |s| \leq t\}$. Theorem 3.5 in [BB] and Theorem 2.7 readily imply

$$\left| \frac{\int_{G_{n,k}} F_K(t,E)\, d\mu(E)}{\Gamma_K^k(t)} - 1 \right| \leq \frac{c_1}{n^{0.09}} \qquad \forall\, t \geq 0$$

Taking limits as $t \to 0$ (see Corollary 3.6 in [BB]) yields

$$\left| \frac{\mathbb{E}\,(f)}{\frac{1}{(\sqrt{2\pi}L_K)^k}} - 1 \right| \leq \frac{c_1}{n^{0.09}} \Big(\leq \frac{\varepsilon}{3} \Big)$$

By the triangle inequality

$$\left| \frac{f(E)}{\frac{1}{(\sqrt{2\pi}L_K)^k}} - 1 \right| \leq \frac{\mathbb{E}\,(f)}{\frac{1}{(\sqrt{2\pi}L_K)^k}} \left| \frac{f(E)}{\mathbb{E}(f)} - 1 \right| + \left| \frac{\mathbb{E}\,(f)}{\frac{1}{(\sqrt{2\pi}L_K)^k}} - 1 \right|$$

So, if $\left| \dfrac{f(E)}{\mathbb{E}(f)} - 1 \right| \leq \dfrac{\varepsilon}{3}$, then $\left| \dfrac{f(E)}{\frac{1}{(\sqrt{2\pi}L_K)^k}} - 1 \right| \leq (1 + \dfrac{\varepsilon}{3})\dfrac{\varepsilon}{3} + \dfrac{\varepsilon}{3} \leq \varepsilon$ and conclude, using also $\mathcal{L}_k \leq ck^{1/4}$

$$\mu\big\{E \in G_{n,k};\, \big|f(E) - \frac{1}{(\sqrt{2\pi}L_K)^k}\big| \leq \frac{\varepsilon}{(\sqrt{2\pi}L_K)^k}\big\} \geq$$
$$\geq \mu\left\{E \in G_{n,k};\, |f(E) - \mathbb{E}\,(f)| \leq \frac{\varepsilon}{3}\mathbb{E}\,(f)\right\} \geq 1 - c_1 \exp\left(-\frac{c_2^k \varepsilon^2 n}{k^{k/2}}\right)$$

The hypothesis on k implies $\varepsilon \geq \frac{(\log\log n)^2}{\log n}$ and $k^{k/2} \ll n^{0,1}$, so

$$\mu\big\{E \in G_{n,k};\, \big|f(E) - \frac{1}{(\sqrt{2\pi}L_K)^k}\big| \leq \frac{\varepsilon}{(\sqrt{2\pi}L_K)^k}\big\} \geq 1 - c_1 \exp(-c_2 n^{0.9})$$

□

Acknowledgements

Partially supported by MTM2007-61446 and DGA E-64.

References

[ABBP] D. Alonso, J. Bastero, J. Bernués and G. Paouris, *High dimensional sections of isotropic convex bodies*, J. Math. Anal. Appl. **361** (2010), pp. 431-439.

[ABP] M. Anttila, K. Ball, I. Perissinaki, *The central limit theorem for convex bodies*, Trans. Amer. Math. Soc. **355** (2003), pp. 4723-4735.

[B] K. Ball, *Logarithmic concave functions and sections of convex sets in* $\mathbb{R}^n$, Studia Math. **88** (1988), pp. 69-84.

[BB] J. Bastero and J. Bernués, *Asymptotic behaviour of averages of k-dimensional marginals of measures on* $\mathbb{R}^n$, Studia Math. **190** (2009), pp. 1-31.

[EK] R. Eldan and B. Klartag, *Pointwise Estimates for Marginals of Convex Bodies*, Journal Functional Analysis **254** (2008), pp. 2275-2293.

[GM] A. Giannopoulos and V. Milman, *Mean width and diameter of proportional sections of a symmetric convex body*, J. Reine Angew. Math. **497** (1998), pp. 113-139.

[H] D. Hensley, *Slicing convex bodies, bounds of slice area in terms of the body's covariance*, Proc. Amer. Math. Soc. **79** (1980), pp. 619-625.

[Kl] B. Klartag, *On convex perturbations with a bounded isotropic constant*, Geom. and Funct. Anal. (GAFA) **16** (2006) 1274-1290.

[Kl2] B. Klartag, *Power-law estimates for the central limit theorem for convex sets*, Journal Functional Analysis **245** (2007), pp. 284-310.

[MP] V. Milman and A. Pajor, *Isotropic positions and inertia ellipsoids and zonoids of the unit ball of a normed n-dimensional space*, GAFA Seminar 87-89, Springer Lecture Notes in Math. **1376** (1989), pp. 64-104.

[MS] V. Milman and G. Schechtman, *Asymptotic theory of finite dimensional normed spaces*, Lecture Notes in Math. **1200**, Springer, (1986).

[So] S. Sodin, *Tail-sensitive Gaussian asymptotics for marginals of concentrated measures in high dimension*,GAFA Seminar,2004-05 Springer Lecture Notes in Math. **1910** (2007), pp. 271-295.

Monografías de la Real Academia de Ciencias de Zaragoza **33**, 233–240, (2010).

Racional identities in the Catalan triangle

Pedro J. Miana

Departamento of Matemáticas & I.U.M.A. Universidad de Zaragoza. Spain.

and

Natalia Romero

Departmento de Matemáticas y Computación. Universidad de La Rioja. Spain.

Dedicated to Dr. Manuel Calvo in his 65^{th} birthday

Abstract

In this paper we consider different racional identities in which appear the numbers $(B_{n,p})_{n,1\leq p\leq n}$ given by

$$B_{n,p} := \frac{p}{n}\binom{2n}{n-p},\ n,p\in\mathbb{N},\ p\leq n.$$

The set of numbers $(B_{n,p})_{n,1\leq p\leq n}$ is known as the Catalan triangle due to the Catalan numbers $(C_n)_{n\in\mathbb{N}}$,

$$C_n = \frac{1}{n+1}\binom{2n}{n},\qquad n\in\mathbb{N},$$

appear in the first column. These identities have been recently proved and some of them are connected with the dynamic behavior of certain iterative methods applied to quadratic polynomials. In the last section we conjeture some new identities which involve this family of numbers $(B_{n,p})_{n,1\leq p\leq n}$.

1 Introduction

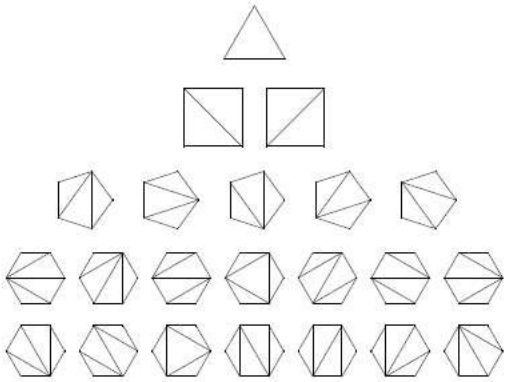
The Catalan number C_n is defined by the expression

$$C_n = \frac{1}{n+1}\binom{2n}{n},\qquad n\in\mathbb{N}.$$

The first ten values of C_n are $1, 2, 5, 14, 42, 132, 429, 1430, 4862, 16796$. Note that Catalan numbers have more than 165 different combinatorial interpretations, see for example [15, p. 219] and

http://www-math.mit.edu/∼rstan/ec/catadd.pdf

In particular, the number C_n is the solution to the Euler problem: how many different ways you can divide a convex polygon of $n+2$ sides in triangles using diagonals ([3]),



They also gives the number of binary bracketings of n letters (Catalan's problem) or the solution to the ballot problem [6].

In 1976, L. W. Shapiro introduced in [11], the following triangle of numbers

$n \setminus p$	1	2	3	4	5	6	...
1	1						
2	2	1					
3	5	4	1				
4	14	14	6	1			
5	42	48	27	8	1		
6	132	165	110	44	10	1	
...	...	...	...	...	...	...	...

(1)

which entries are given by

$$B_{n,p} := \frac{p}{n}\binom{2n}{n-p},\ n, p \in \mathbb{N},\ p \leq n.$$

These numbers $(B_{n,p})_{1 \leq n, n \in \mathbb{N}}$ also satisfy a recurrence relation,

$$B_{n,p} = B_{n-1,p-1} + 2B_{n-1,p} + B_{n-1,p+1}, \qquad p \geq 2.$$

Note that $B_{n,1} = C_n$ for $n \geq 1$.

Although the numbers $B_{n,k}$ are not as famous as Catalan numbers, they have also several applications (see [2, 11, 13] for more details). As a sample, we cite some of them:

- $B_{n,p}$ is the number of leaves at level $p+1$ in all ordered trees with $n+1$ edges
- $B_{n,p}$ is the number of walks of n steps, each in direction N, S, W or E, starting at the origin, remaining in the upper half-plane and ending at height p.
- $B_{n,p}$ denote the number of pairs of non-intersecting paths of length n and distance p (see the definitions in [11, p.84]).

In this short note, we present some of the main results given in [5, 8, 9] which involve $(B_{n,p})_{1\leq p\leq n, n\in\mathbb{N}}$. For example, we give the explicit expressions of the moments $(\Omega_m)_{m\geq 0}$,

$$\Omega_m(n) := \sum_{p=1}^{n} p^m B_{n,p}^2, \qquad n \in \mathbb{N}$$

for $1 \leq m \leq 7$ and general expressions for arbitrary m. Other formulae which appear in the dynamical study of certain iterative problem are also given. This collection of results have been considered and studied by other mathematicians, [1, 4, 12]. In the last section, we present two conjectures about new identities in which appear the numbers $(B_{n,p})_{1\leq p\leq n, n\in\mathbb{N}}$; the second one is connected with the values of some determinants associated to the triangle (1).

2 Main results

Different techniques are used in the proof of the following results: Chu-Vandermonde convolution formula; W-Z theory and Newton interpolation formula. Details of the power W-Z theory may be found in the monographic [10] and in [16].

Theorem 2.1 *[8] Let $n \in \mathbb{N}$. Then*

(i) $$\Omega_0(n) := \sum_{p=1}^{n} (B_{n,p})^2 = C_{2n-1}.$$

(ii) $$\Omega_2(n) := \sum_{p=1}^{n} p^2 (B_{n,p})^2 = \frac{(3n-2)n}{4n-3} C_{2n-1}.$$

(iii) $$\Omega_4(n) := \sum_{p=1}^{n} p^4 (B_{n,p})^2 = \frac{(15n^3-30n^2+16n-2)n}{(4n-3)(4n-5)} C_{2n-1}.$$

(iv) $$\Omega_6(n) := \sum_{p=1}^{n} p^6 (B_{n,p})^2 = \frac{(105n^5-420n^4+588n^3-356n^2+96n-10)n}{(4n-3)(4n-5)(4n-7)} C_{2n-1}.$$

Theorem 2.2 *[8] Let $n \in \mathbb{N}$. Then*

(i) $\Omega_1(n) := \sum_{p=1}^{n} p\,(B_{n,p})^2 = (n+1)C_n(2n-3)C_{n-2}.$

(ii) $\Omega_3(n) := \sum_{p=1}^{n} p^3(B_{n,p})^2 = (n+1)C_n n(2n-3)C_{n-2}.$

(iii) $\Omega_5(n) := \sum_{p=1}^{n} p^5(B_{n,p})^2 = (n+1)C_n n(3n^2-5n+1)C_{n-2}.$

(iv) $\Omega_7(n) := \sum_{p=1}^{n} p^7(B_{n,p})^2 = (n+1)C_n n(6n(n-1)^2-1)C_{n-2}.$

Remarks. Note that the polynomials which appear in the Theorem 2.1 and 2.2 do not belong to any known classical family. In the following theorem we give the moments of arbitrary order although a explicit expression is unknown.

Theorem 2.3 *[9] Let $n \in \mathbb{N}$. Then there exist P_{3m+1}, Q_{2m+2}, R_{3m-1} polynomials of integer coefficients and degree at least $3m+1$, $2m+2$ and $3m-1$ respectively such that*

$$\begin{aligned}
\Omega_{2m}(n) &= \frac{P_{3m+1}(n)}{\prod_{l=1}^{m}(4n-(2l+1))}C_{2n-1}, \qquad m \geq 0,\\
\Omega_{2m+1}(n) &= Q_{2m+2}(n+1)C_nC_{n-2}, \qquad m \leq 3,\\
\Omega_{2m+1}(n) &= \frac{R_{3m-1}(n)}{\prod_{l=1}^{m-3}(2n-(2l+3))}(n+1)C_nC_{n-2}, \qquad m \geq 4.
\end{aligned}$$

Theorem 2.4 *[5, 8] Let $n \in \mathbb{N}$, and $1 \leq i \leq n$. Then*

(i) $\sum_{p=1}^{i} B_{n,p}B_{n,n+p-i}(n+2p-i) = (n+1)C_n\binom{2(n-1)}{i-1}.$

(ii) $\sum_{p=1}^{i} B_{n,p}B_{n,n+p-i}(n+2p-i)^3 = (n+1)C_n\binom{2(n-1)}{i-1}\left(n^2+4n-2ni+i^2\right).$

3 An application to Newton-like iterative methods

The application of some iterative methods for solving nonlinear equations to a polynomial equation could give raise to rational iteration functions which dynamics are not well-known.

We present in the complex plane a study of the dynamical behavior of the following Newton-like methods

$$\begin{cases} z_{m+1} = R_n(z_m) = z_m - H_n(L_f(z_m))\dfrac{f(z_m)}{f'(z_m)}, \quad m \geq 0, \\ H_n(z) = \displaystyle\sum_{j=0}^{n} \frac{1}{2^j} C_j z^j, \quad n \geq 0, \quad L_f(z) = \frac{f(z)f''(z)}{f'(z)^2}, \end{cases} \tag{2}$$

which are written in terms of the Catalan numbers.

These methods give rise to rational functions defined in the extended complex plane, $\mathbb{C}_\infty = \mathbb{C} \cup \{\infty\}$. In particular, we prove that these rational root-finding algorithms are generally convergent for quadratic polynomials.

The idea of general convergence of a method for polynomials of a given degree was introduced by Smale [14] and McMullen [7] and it means that the method converges to a root for almost every starting point and for almost every polynomial of a given degree.

The conjugated rational map of R_n, $S_n := MR_nM^{-1}$, via the Möbius map $M(z) = (z-a)/(z-b)$, is given by

$$S_n(z) = z^{n+2}\frac{P_n(z)}{\hat{P}_n(z)}, \tag{3}$$

where $P_n(z) = \displaystyle\sum_{p=0}^{n} B_{n+1,p+1} z^p$ and $\hat{P}_n(z) = \displaystyle\sum_{p=0}^{n} B_{n+1,n+1-p} z^p$.

A rational map R, divides $\mathbb{C}_\infty$ in two subsets, that are known as *Fatou set* and *Julia set*. The Fatou set, denoted $\mathcal{F}(R)$ is defined as the set of points $z_0 \in \mathbb{C}_\infty$ such that the family of iterates R^n is a normal family in some neighborhood U_{z_0} of z_0. That is, every infinite sequence of R^n contains a subsequence R^{n_k} that converges locally uniformly on U_{z_0} to some continuous function $f \in \mathcal{C}(\mathbb{C}_\infty)$. Recall that $R^{n_k} \to f$ locally uniformly on U_{z_0} if for all $z \in U_{z_0}$, $R^{n_k} \to f$ uniformly on some neighborhood of z. The Julia set, $\mathcal{J}(R)$, is the complement of the Fatou set, $\mathcal{J}(R) = \mathbb{C}_\infty - \mathcal{F}(R)$.

Roughly speaking, the Fatou $\mathcal{F}(R)$ set includes the points whose orbits are predictable after iteration and the Julia set includes the points whose dynamical behaviour is complicated with independency of the number of iterations.

Applying item (i) of Theorem 2.4, we obtain

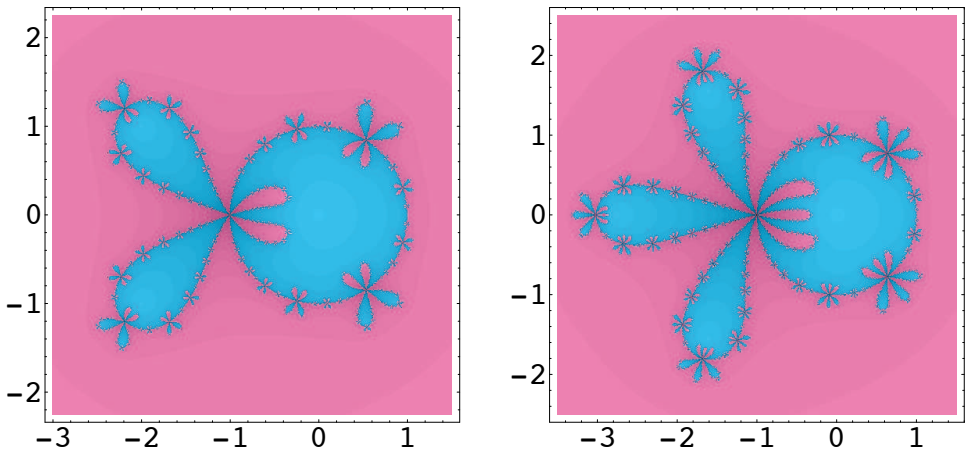
$$S_n'(z) = \frac{(n+2)C_{n+1}z^{n+1}(1+z)^{2n}}{\hat{P}_n(z)^2}.$$

Hence, we can describe the Fatou components associated to S_n, $n \geq 0$, and we can conclude that the rational map R_n is generally convergent for quadratic polynomials.

In fact, we have that the rational map $S_n(z)$, $(n \geq 0)$, defined in (3), has precisely two forward invariant Fatou components: a superattracting component where iterates converge to ∞ and a superattracting component where iterates converge to 0. The unit

circle $S^1(z) = \{z \in \mathbb{C};\ |z| = 1\}$ is forward invariant and it is contained in $\mathcal{J}(S_n)$ and moreover, $\mathfrak{m}(\mathcal{J}(S_n)) = 0$, where $\mathfrak{m}$ is the Lebesgue measure on $\mathbb{C}$.

Finally, we show the basins of attraction associated to the two roots of a quadratic polynomial $f(z) = (z-a)(a-b)$ when we apply S_2 and S_3. The basins of attraction clarify the structures of the universal Julia sets associated to the corresponding iterative methods R_2 and R_3.



Plot of the basins of attraction under S_2 and S_3 applied to quadratic polynomial $f(z) = (z-a)(z-b)$.

4 Two open problems

Now we come back to the triangle (1). Note that if we multiple the figures in the row n by the figures in the next row $n+1$, we obtain the Catalan number C_{2n}, for example

$$C_6 = 132 = 5 \cdot 14 + 4 \cdot 14 + 1 \cdot 6.$$

In fact, this result looks like true if we multiply two different rows: multiply the row n and $n+j$, we obtain the Catalan number C_{2n+j-1}. To check this conjeture, take the second and fifth rows and

$$C_6 = 132 = 42 \cdot 2 + 48 \cdot 1.$$

Then it is natural to conjeture that

$$C_{i+j-1} = \sum_{k=1}^{\min(i,j)} B_{i,k} B_{j,k}, \qquad i, j \geq 1.$$

From the triangle (1), we traslade the figures in each column p-th, $p-1$ steps to obtain

the new table,

$n \setminus p$	1	2	3	4	5	6	...
1	1	1	1	1	1	1	
2	2	4	6	8	10	12	
3	5	14	27	44	65	90	
4	14	48	110	208	350	544	
5	42	165	429	910	1700	2907	
6	132	572	1638	3808	7752	14364	
...	...	...	...	...	...	...	...

(4)

We denote by $(M_n)_{n\geq 1}$ the main minors of order n in the table (4); we obtain that

n	M_n
1	$1=2^0$
2	$2=2^1$
3	$8=2^3$
4	$64=2^6$
5	$1024=2^{10}$
6	$32768=2^{15}$

(5)

Taking into account (5), it is natural to conjeture that $M_n = 2^{\frac{n(n-1)}{2}}$ for $n \geq 1$.

Acknowledgements

The first author is partly supported by Projects MTM2007-61446, DGI-FEDER, and E-64, D.G. Aragón. The second author is partly supported by the Ministry of Science and Innovation MTM 2008-01952.

References

[1] X. Chen and W. Chu: *Moments on Catalan number.* J. Math. Anal. Appl., **349**, (2) (2009), 311–316.

[2] E. Deutsch and L. Shapiro: *A survey of the Fine numbers,* Discrete Math. **241** (2001), 241–265.

[3] H. G. Forder: *Some Problems in Combinatorics,*. Math. Gaz. **41** (1961), 199–201.

[4] V.J.W. Guo and J. Zeng: *Factors of binomial sums from Catalan triangle.* J. Number Theory, **130**, (1) (2010), 172–186.

[5] J. M. Gutierrez, M.A. Hernández, P.J. Miana, and N. Romero: *New identities in the Catalan triangle.* J. Math. Anal. Appl., **341**, (1) (2008), 52–61.

[6] P. Hilton and J. Pedersen: *Catalan numbers, their generalization and their uses,* Math. Intelligencer **13** (1991), 64–75.

[7] C. McMullen, Families of rational maps and iterative root-finding algorithms. Ann. Math., **Vol. 125**, (1987) 467–493.

[8] P.J. Miana and N. Romero: *Computer proofs of new identities in the Catalan triangle.* Biblioteca de la Revista Matemática Iberoamericana. Proc. of the "Segundas Jornadas de Teoría de Números", (2007), 203–208.

[9] P.J. Miana and N. Romero: *Moments of combinatorial and Catalan numbers,* to appear in J. Number Theory.

[10] M. Petkovsek, H. S. Wilf and D. Zeilberger: $A = B$. A. K. Peters Ltd., Wellesley, 1997. http://www.cis.upenn.edu/∼wilf/AeqB.html

[11] L. W. Shapiro: *A Catalan triangle,* Discrete Math. **14** (1976), 83–90.

[12] A. Slavík: *Identities with squares of binomial coefficients,* to appear in Ars Combinatoria.

[13] N. Sloane: http://www.research.att.com/∼njas/sequences/A039598

[14] S. Smale, On the efficiency of algorithms of analysis. Bull. AMS, **Vol. 13**, (1985) 87–121.

[15] R.P. Stanley: *Enumerative Combinatorics,* vol. 2, Cambridge University Press, 1999.

[16] H. Wilf and D. Zeilberger: *Rational functions certify combinatorial identities,* J. Amer. Math. Soc. **3** (1990), 147–158.

Monografías de la Real Academia de Ciencias de Zaragoza **33**, 241–254, (2010).

Groups with families of generalized normal subgroups

Leonid A. Kurdachenko

Department of Algebra, National University of Dnepropetrovsk
Gagarin prospect 72, Dnepropetrovsk 10, 49010, Ukraine

Javier Otal

Departamento de Matemáticas - IUMA, Universidad de Zaragoza
Pedro Cerbuna 12, 50009 Zaragoza, Spain

To Manolo with affection in his 65^{th} birthday

Introduction

In his treatment of the solvability of polynomial equations, *Évariste Galois* coined the term *group* and established a connection, now known as *Galois theory*, between the nascent Theory of Groups (formerly Theory of *Finite* Groups) and Field Theory, giving rise to one of the main historical sources of the Theory of Groups (being the others Number Theory and Geometry). In the well know Galois' correspondence mentioned above, Galois emphasized the fundamental role of some subgroups of the Galois group that are invariant under certain automorphisms, namely its normal subgroups. If G is a group (in its abstract form), we recall that a subgroup N of G is said to be a *normal* subgroup if it is invariant under *inner* automorphisms (G–invariant), that is $N^x := x^{-1}Nx = N$ for each $x \in N$. For example *every subgroup of an abelian group is normal*. In studying number fields (finite Galois extensions of the field $\mathbb{Q}$ of rational numbers), R. Dedekind [5] was able to determine the form of a non-abelian (finite) group with normal subgroups only (*Hamiltonian* groups), a result extended by R. Baer [1] to the general case.

Theorem 1 *Let G be a non-abelian group in which all subgroups are normal. Then $G = Q \times B \times D$, where Q is a copy of the quaternion of order 8, B is an elementary abelian 2–group and D is a periodic abelian group with all elements of odd order.*

Here, a group is said to be *periodic* if their elements have finite order, and *bounded* (or that has *finite exponent*) if these orders are bounded. Opposite to this, *a torsion-free group* is a group with no non-trivial elements of finite order.

We recall that a group G is called a *Dedekind group* if every subgroup of G is normal. By Theorem 1, the class of Dedekind groups is the union of the class of Hamiltonian groups and that of abelian groups.

It is well known that *being a normal subgroup* is not a transitive property, and to make that concept transitive, it is introduced the concept of subnormality. A subgroup H of a group G is said to be *subnormal* if there is a finite chain of intermediate subgroups

$$H = H_0 \leq H_1 \leq \cdots H_i \leq H_{i+1} \leq \cdots H_n = G$$

such that H_i is normal in H_{i+1} for every $0 \leq i \leq n-1$. This is a fairly generalization of the concept of normal subgroup. A natural extension of these concepts is that of an ascendant subgroup. H is said to be an *ascendant subgroup of* G if there exists an *ascending series* from H to G, that is, a chain of normal subgroups well-ordered by inclusion and indexed by the corresponding ordinal numbers

$$H = H_0 \leq H_1 \leq \ \cdots \ \leq H_\alpha \leq H_{\alpha+1} \leq \ \cdots \ \leq H_\gamma = G$$

with the additional stipulation that for each limit ordinal λ, H_λ is the union of all H_β, $\beta < \lambda$. The most easy way of realizing an ascending series is constructing a Prüfer group. If p is a prime, *the Prüfer p–group*

$$C_{p^\infty} = \langle x_1, \cdots, x_n \cdots \mid x_1^p = 1,\ x_n^p = x_{n-1} (n > 1) \rangle$$

is an infinite abelian group whose proper subgroups are all finite. In fact, the subgroups of C_{p^∞} are the terms of the ascending series

$$\langle 1 \rangle \leq C_1 \leq \cdots \leq C_n \leq \cdots \bigcup_{n \geq 1} C_n = C_{p^\infty},$$

where $C_n = \langle x_n \rangle$ for every $n \geq 1$. By the way, this an obvious example of a *locally finite group*, a group whose finitely generated subgroups are finite.

If $x, y \in G$, then $xy = yx(x^{-1}y^{-1}xy)$, and then the commutativity of x and y is measured by the so called *commutator* of x and y, namely $[x, y] := x^{-1}y^{-1}xy$, because we immediately have that $xy = yx \Leftrightarrow [x, y] = 1$. If $H, K \leq G$ and $S \subseteq G$, these considerations lead to the construction of the subgroups of G,

$$[H, K] = \langle [x, y] \mid x \in H, y \in K \rangle \text{ and } C_G(S) = \{x \in G \mid [x, y] = 1 \text{ for all } y \in S\}.$$

The most important cases are $[H, G]$ and $\zeta(G) = C_G(G)$, which are called *the commutator subgroup* of H by G and *the center* of G, respectively. Clearly G is abelian if and only if $[G, G] = \langle 1 \rangle$ if and only if $\zeta(G) = G$. Roughly speaking, we could say that we may construct generalizations of an abelian group making trivial commutators of higher weight

or stabilizing the natural pre-images of the subsequent centers. By definition, *the upper central series of* G is the ascending chain of subgroups

$$\langle 1\rangle = \zeta_0(G) \leq \zeta_1(G) \leq \cdots \zeta_\alpha(G) \leq \zeta_{\alpha+1}(G) \leq \cdots$$

given by $\zeta_{i+1}(G)/\zeta_i(G) = \zeta(G/\zeta_i(G)),\ i \geq 0$. Note that $\zeta_1(G) = Z(G)$. On the other hand *the lower central series of* G is the descending chain of subgroups

$$G = \gamma_1(G) \geq \gamma_2(G) \geq \cdots \geq \gamma_\alpha(G) \geq \gamma_{\alpha+1}(G) \geq \cdots$$

given by $\gamma_{i+1}(G) = [\gamma_i(G), G],\ i \geq 0$. Note that $\gamma_2(G) = [G, G]$. A group G is said to be *nilpotent* if there is some $c \geq 0$ satisfying one of the following equivalent conditions: (i) $\zeta_c(G) = G$; and (ii) $\gamma_{c+1}(G) = \langle 1\rangle$. More generally G is said to be *hypercentral* if there exists an ordinal α such that $\zeta_\alpha(G) = G$. Finally *the derived series of* G is the descending chain of subgroups

$$G = G^{(0} \geq \ \cdots G^{(n} \geq \cdots$$

given by $G^{(i+1} = [G^{(i}, G^{(i}],\ i \geq 0$. Here also $G^{(1} = [G, G]$. The group G is said to be *soluble* if there is some $d \geq 0$ such that $G^{(d} = \langle 1\rangle$. It is very easy to see that a nilpotent group is soluble although the converse is not true. Finite soluble groups were fundamental in the Galois' characterization of the solvability of polynomial equations by radicals.

The next result is standard inside the Theory of Groups and is similar to Theorem 1. It characterizes *nilpotent* groups in the finite case.

Theorem 2(W. Burnside) *For a finite group* G *the following conditions are equivalent:*

(1) G *is nilpotent;*

(2) Every subgroup of G *is subnormal; and*

(3) If H *is a proper subgroup of* G *then* H *is properly contained in its normalizer (the largest subgroup of* G *in which* H *is normal)* $N_G(H) = \{x \in G \mid H^x = H\}$.

It is worth mentioning that the implications (1) $\Rightarrow$ (2) $\Rightarrow$ (3) of Theorem 2 hold for arbitrary groups though the equivalence is false in general and gives rise to several classes of *generalized nilpotent groups*. The falsity holds for infinite groups as the following example shows. If G is a hypercentral group, it is very easy to show that every subgroup of G is ascendant. However if P is a Prüfer 2–group, we construct *the infinite dihedral group*,

$$D = \langle P, y \mid y^2 = 1, x^y = x^{-1} \text{ para todo } x \in P\rangle.$$

The group D is hypercentral but the subgroup $\langle y\rangle$ is not subnormal.

The aim of this survey paper is to review some families of subgroups that generalize normal subgroups as well as the classes of groups involved.

1 Subnormal subgroups

We begin stating a result that locates the groups under consideration.

Theorem 1.1 (A. I. Maltsev [16]). *A hypercental group is locally nilpotent, that is their finitely generated subgroups are nilpotent.*

A group G is said to satisfy *the normalizer condition* (or G *is an* N*–group*) if $H \neq N_G(H)$ for each proper subgroup H (see Theorem 2). Since every proper ascendant subgroup is properly contained in its normalizer, G is an N–group if and only if every subgroup of G is ascendant.

Theorem 1.2 (B. I. Plotkin [20]). *A group whose subgroups are ascendant is locally nilpotent.*

That is, an N–group is locally nilpotent. As we mentioned above, hypercentral groups are N–groups, but the converse is far from being true. In this setting one of the most celebrated examples in the Theory of Groups is given in the following result.

Theorem 1.3 (H. Heineken, I. J. Mohamed [11]). *There exists a p–group G, p a prime, satisfying the following properties:*

(1) G contains an elementary abelian normal p–subgroup A such that G/A is a Prüfer p–group;

(2) every proper subgroup of G is subnormal and nilpotent; and

(3) $\zeta(G) = \langle 1 \rangle$.

In relation with subnormal and ascendant subgroups of a group, some distinguished subgroups of the group can be constructed. That construction arises from the following results

Theorem 1.4 (R. Baer [2], K. W. Gruenberg [9]). *Let H and K be two finitely generated nilpotent subgroups of the group G. If H and K are subnormal (respectively ascendant), then so is $\langle H, K \rangle$.*

Let G be a group. Then the subgroup $B(G)$ generated by all subnormal cyclic subgroups of G is called *the Baer radical of* G, and the subgroup $Gr(G)$ generated by all ascendant cyclic subgroups of G is called *the Gruenberg radical of* G. Clearly, both subgroups are locally nilpotent normal subgroups of G. A group G is called *a Baer group* if $G = B(G)$ holds, and *a Gruenberg group* if $G = Gr(G)$.

We mention that every countable locally nilpotent group can be expressed as the union of an ascending chain of finitely generated nilpotent subgroups and therefore it is a Gruenberg group. But for uncountable groups it is not true, as the following result shows.

Theorem 1.5 (M. I. Kargapolov [13]). *There is a locally finite p–group that is not a Gruenberg group.*

Groups whose subgroups are subnormal were studied by many authors. In this area many interesting were obtained. We mention here only certain satisfactory structural results.

Theorem 1.6 (W. Möhres [18]). *A group whose subgroups are all subnormal is soluble.*

Theorem 1.7 (W. Möhres [17]). *A bounded group whose subgroups are all subnormal is nilpotent.*

Theorem 1.8 (W. Möhres [19]). *A hypercentral group whose subgroups are all subnormal is nilpotent.*

Theorem 1.9 (H. Smith [22]). *A torsion-free group in which all subgroups are subnormal is nilpotent.*

The last results are somewhat specific but we quote for their interest.

Theorem 1.10 (C. Casolo [3], H. Smith [22]). *Let G be a periodic group in which all subgroups are subnormal. If $\bigcap_\alpha \gamma_\alpha(G) = \langle 1 \rangle$, then G is nilpotent.*

Theorem 1.11 (C. Casolo [3]). *Let G be a periodic group in which all subgroups are subnormal. Then G contains a nilpotent normal subgroup H such that G/H is a divisible abelian group of finite special rank.*

2 Groups with many ascendant subgroups

The condition to be an ascendant subgroup is very wide than to be a subnormal subgroup. It is the main reason why the groups whose subgroups are all ascendant were not studied too well. There are many partial results about these groups, but in general its study is very difficult. There are quite a few general results on the structure of these groups. Some authors started consider groups in which the family of non-ascendant subgroups is not empty but it is very small. Some examples of this are

- S. N. Chernikov [4] who studied groups whose subgroups are either ascendant or finite.
- H. Heineken and L. A. Kurdachenko [10], who studied groups whose subgroups are either subnormal or finitely generated.
- H. Smith [24, 25], who studied groups whose subgroups are either subnormal or nilpotent.

as well as many others.

In studying the structure of groups whose subgroups belong to two types, there is an interesting approach that gives rise to obtain more information. Many important types of subgroups have their *antipodes*, i.e. subgroups that have diametrically opposite properties with respect to the original. For example, if H is a subgroup of G, then $H \leq N_G(H) \leq G$. If H is normal in G then $N_G(H) = G$. Therefore subgroups with the property $H = N_G(H)$ are in the antipodes to normal subgroups. These subgroups are called *self-normalizing*. As we mentioned above, subnormal subgroups and ascendant subgroups cannot be self-normalizing, so that we may conclude that self-normalizing subgroups are in the antipodes of subnormal and ascendant subgroups. Moreover we also apply Theorem 2 to deduce that *a nilpotent group has no proper self-normalizing subgroups*.

We also note that if H is a subgroup properly contained in its normalizer, that is $H \neq N_G(H)$, then $H^g = H$ for each $g \in N_G(H)$. If moreover $g \notin H$, then it is trivial that $g \notin \langle H, H^g \rangle$. A subgroup H of a group G is called *abnormal* if $g \in \langle H, H^g \rangle$ for every element $g \in G$. Therefore, we conclude that *a nilpotent group has no proper abnormal subgroups*, and we see that abnormal subgroups also are in the antipodes of normal, subnormal and ascendant subgroups. Thus in a certain sense we could say that the subgroups of a group that have some defining properties and those that have the antipodes with respect to these properties are located at the opposite ends of the group, while the other subgroups have some kind of mixed intermediate positions between these two ends. If a group G has few subgroups of mixed intermediate positions, it appears that the structure of G is more transparent. Therefore the following question can be naturally raised: *characterize groups whose subgroups have only a certain property and its antipode*. The first example appear considering nilpotent groups. Actually a nilpotent group only has subnormal subgroups and has neither abnormal subgroups nor self-normalizing subgroups. One of the first investigations carrying out this approach was the paper by A. Fattahi [8], where finite groups with normal and abnormal subgroups only were described. Later on, G. Ebert and S. Bauman [6] studied finite groups every subgroup of which is either subnormal or abnormal. Infinite groups with these properties and their generalizations were described by M. de Falco, L. A. Kurdachenko and I. Ya. Subbotin [7], and later L. A. Kurdachenko and H. Smith [15] studied groups whose subgroups are either subnormal or self-normalizing. We quote here the main results of these papers as well as latest results from which the previous are now a consequence.

Theorem 2.1 (L. A. Kurdachenko, J. Otal, A. Russo, G. Vincenzi [14])**.** *Let G be a locally finite group and suppose that G is not locally nilpotent. If every finitely generated non-ascendant subgroup of G is self-normalizing then there exist a prime p and a nilpotent normal subgroup A of G with no elements of order p such that the following conditions*

hold

(1) $G = AP$ *and* $A \cap P = \langle 1 \rangle$, *where* $P = \langle x \rangle$ *is a cyclic p-subgroup and* $C_P(A) = \langle g^p \rangle$;

(2) the commutator subgroup $[G, G] = A$; *and*

(3) P *is self-centralized, that is* $C_G(P) = P$.

Conversely, if the group G *satisfies the conditions (1)-(3), then every subgroup of* G *is either ascendant or self-normalizing.*

This result can be put in the usual form of these results.

Corollary 2.2. *Let* G *be a locally finite group and suppose that* G *is not locally nilpotent. Then every non-ascendant subgroup of* G *is self-normalizing if and only if there exist a prime* p *and a nilpotent normal subgroup* A *of* G *with no elements of order* p *such that the following conditions hold*

(1) $G = AP$ *and* $A \cap P = \langle 1 \rangle$, *where* $P = \langle x \rangle$ *is a cyclic p-subgroup and* $C_P(A) = \langle g^p \rangle$;

(2) the commutator subgroup $[G, G] = A$; *and*

(3) P *is self-centralized, that is* $C_G(P) = P$.

As we mentioned above we are able to obtain previous results.

Corollary 2.3 (L. A. Kurdachenko, H. Smith [15])**.** *Let* G *be a locally finite group and suppose that* G *is not locally nilpotent. Then every non-subnormal subgroup of* G *is self-normalizing if and only if there exist a prime* p *and a nilpotent normal subgroup* A *of* G *with no elements of order* p *such that the following conditions hold*

(1) $G = AP$ *and* $A \cap P = \langle 1 \rangle$, *where* $P = \langle x \rangle$ *is a cyclic p-subgroup and* $C_P(A) = \langle g^p \rangle$;

(2) the commutator subgroup $[G, G] = A$; *and*

(3) P *is self-centralized, that is* $C_G(P) = P$.

Corollary 2.4 (L. A. Kurdachenko, H. Smith [15])**.** *Let* G *be a locally finite group and suppose that* G *is not a Dedekind group. Then every non-normal subgroup of* G *is self-normalizing if and only if there exist a prime* p *and an abelian normal subgroup* A *of* G *with no elements of order* p *such that the following conditions hold*

(1) $G = AP$ *and* $A \cap P = \langle 1 \rangle$, *where* $P = \langle x \rangle$ *is a cyclic p-subgroup and* $C_P(A) = \langle g^p \rangle$;

(2) the commutator subgroup $[G, G] = A$;

(3) P *is self-centralized, that is* $C_G(P) = P$; *and*

(4) every subgroup of A is G–invariant.

Corollary 2.5. *Let G be a locally finite group and suppose that G is not locally nilpotent. Then every non-ascendant subgroup of G is abnormal if and only if there exist a prime p and a nilpotent normal subgroup A of G with no elements of order p such that the following conditions hold*

(1) $G = AP$ and $A \cap P = \langle 1 \rangle$, where $P = \langle x \rangle$ is a cyclic p-subgroup and $C_P(A) = \langle g^p \rangle$;

(2) the commutator subgroup $[G, G] = A$; and

(3) P is self-centralized, that is $C_G(P) = P$.

Corollary 2.6 (M. de Falco, L. A. Kurdachenko, I. Ya. Subbotin [7])**.** *Let G be a locally finite group and suppose that G is not locally nilpotent. Then every non-subnormal subgroup of G is abnormal if and only if there exist a prime p and a nilpotent normal subgroup A of G with no elements of order p such that the following conditions hold*

(1) $G = AP$ and $A \cap P = \langle 1 \rangle$, where $P = \langle x \rangle$ is a cyclic p-subgroup and $C_P(A) = \langle g^p \rangle$;

(2) the commutator subgroup $[G, G] = A$; and

(3) P is self-centralized, that is $C_G(P) = P$.

Corollary 2.7. *Let G be a locally finite group and suppose that G is not a Dedekind group. Then every non-normal subgroup of G is abnormal if and only if there exist a prime p and an abelian normal subgroup A of G with no elements of order p such that the following conditions hold*

(1) $G = AP$ and $A \cap P = \langle 1 \rangle$, where $P = \langle x \rangle$ is a cyclic p-subgroup and $C_P(A) = \langle g^p \rangle$;

(2) the commutator subgroup $[G, G] = A$;

(3) P is self-centralized, that is $C_G(P) = P$; and

(4) every subgroup of A is G–invariant.

For non-periodic groups, we have

Theorem 2.8 (L. A. Kurdachenko, J. Otal, A. Russo, G. Vincenzi [14])**.** *Let G be a group and suppose that every finitely generated subgroup is either ascendant or self-normalizing. If G is not periodic, then G is a Gruenberg group.*

Corollary 2.9. *Let G be a group whose subgroups are either ascendant or self-normalizing. If G is not periodic then G is a Gruenberg group.*

We apply our study to hyperabelian groups, a class of generalized soluble groups. We recall that a group G is said *hyperabelian* if there exists an ascending series $\{H_\alpha\}_{\alpha<\gamma}$ from the trivial subgroup $\langle 1\rangle = H_0$ to the whole group $G = H_\gamma$ such that $H_{\alpha+1}/H_\alpha$ is abelian for every ordinal α.

Theorem 2.10 (L. A. Kurdachenko, J. Otal, A. Russo, G. Vincenzi [14])**.** *Let G be a hyperabelian group whose subgroups are either ascendant or self-normalizing. If G is locally nilpotent, then every subgroup of G is ascendant.*

Corollary 2.11. *Let G be a hyperabelian group whose subgroups are either ascendant or self-normalizing. If G is not periodic, then G is locally nilpotent. In particular, every subgroup of G is ascendant.*

With some extra work we find out a little more.

Corollary 2.12 (L. A. Kurdachenko, H. Smith [15])**.** *Let G be a group whose subgroups are either subnormal or self-normalizing. If G is locally nilpotent, then every subgroup of G is subnormal.*

Proof. If G is finitely generated, then G is nilpotent, and the proof is over. Suppose that G has no a finite set of generators. Let $F \leq G$ be a finitely generated subgroup of G. Pick $x \notin F$. Then $\langle x, F\rangle$ is nilpotent and so $F \neq N_{\langle x,F\rangle}(F)$. Thus F is subnormal. Let

$$F = F_0 \trianglelefteq F_1 \trianglelefteq \cdots \trianglelefteq F_n = G$$

be a subnormal series of F in G, that is $F_n = F^G$, $F_{n-1} = F^{F_n}$, . . . , $F_1 = F^{F_2}$, where $X^Y = \langle x^y = y^{-1}xy \mid x \in X, y \in Y\rangle$. Then F_1 is the product of the nilpotent normal subgroups F^x, $x \in F_2$, and it is known that F_1 is hyperabelian. By Theorem 2.10, F_1 has no self-normalizing subgroups and thus every subgroup of F_1 is subnormal. By Theorem 1.6, F_1 is soluble. Now F_2 is the product of the soluble normal subgroups F_1^x, $x \in F_3$, and it is known that F_2 is hyperabelian. As above, we see that F_3 is hyperabelian. Proceeding in this way, after finitely many steps we see that G is hyperabelian. By Theorem 2.10, G has no self-normalizing subgroups, and hence every subgroup of G is subnormal. □

Corollary 2.13 (L. A. Kurdachenko, H. Smith [15])**.** *Let G be a group whose subgroups are either subnormal or self-normalizing. If G is not periodic, then every subgroup of G is subnormal. In particular, if G is torsion-free, then G is nilpotent.*

Corollary 2.14. *Let G be a group whose subgroups are either normal or self-normalizing. If G is not periodic, then G is abelian.*

Corollary 2.15. *Let G be a group whose subgroups are either subnormal or abnormal. If G is locally nilpotent, then every subgroup of G is subnormal.*

Corollary 2.16. *Let G be a group whose subgroups are either subnormal or abnormal. If G is not periodic, then every subgroup of G is subnormal.*

We mention that in [7] the latter was proved with the additional condition $G \neq [G, G]$.

Corollary 2.17. *A non-periodic group whose subgroups are either normal or abnormal is abelian.*

To finish this Section we mention the related result obtained in the paper [15].

Theorem 2.18 (L. A. Kurdachenko, H. Smith [15])**.** *Let G be a group whose subgroups are all subnormal. Suppose that there is $n \geq 1$ such that G is generated by elements of order at most n. Then G is nilpotent.*

3 Permutable subgroups

A subgroup H of a group G is said to be *permutable in G* (or quasi-normal in G), if $HK = KH$ for every subgroup K of G. This concept arises as a generalization of that of normal subgroup since it is immediate that a normal subgroup is permutable. The study of the properties of the permutable subgroups started a rather long time ago (see, for example [21]), where groups whose subgroups are all permutable were described. Before than giving that description we recall the following result that establish a certain connection among the concepts involved in this paper.

Theorem 3.1 (S. E. Stonehewer [26])**.** *A permutable subgroup of a group G is ascendant in G.*

In this case, by Theorems 3.1 and 1.2, G is locally nilpotent. Application of the results of a paper by K. Iwasawa [12] give us the following description.

Theorem 3.2. *Let G be a group whose subgroups are all permutable.*

(1) If G is periodic, then G can be expressed as a direct product

$$G = Dr_p G_p,$$

where G_p is the Sylow p–subgroup of G, and the following conditions holds:

(1A) if $p \neq 2$, then either G_p is abelian or $G_p = B_p\langle a_p\rangle$, where B_p is an abelian subgroup of exponent p^k, and there is a positive integer t such that $t = 1 + p^m$, for some $m \leq k \leq m + d$, where $p^d = |G_p/B_p|$, and $a_p^{-1} b a_p = b^t$ for all $b \in B_p$; and

(1B) if $p = 2$, then either G_p is a Dedekind group or $G_p = B_p\langle a_p\rangle$, where B_p is an abelian normal subgroup of exponent p^k, and there is a positive integer t such that $t = 1 + p^m$, where $p^d = |G_p/B_p|$, and $a_p^{-1}ba_p = b^t$ for all $b \in B_p$.

In both cases G_p is nilpotent, and bounded in the non-abelian case; and

(2) If G is not periodic, then

(2A) the set T consisting of all elements of G having finite order is a subgroup of G;

(2B) T and G/T are abelian;

(2C) every subgroup of T is G–invariant; and

(2D) if the abelian factor-group G/T has positive torsion-free rank, then G is abelian.

If G further is torsion-free, then G is abelian.

As a consequence of Theorem 2.1, we can now obtain the following result.

Theorem 3.3. *Let G be a locally finite group and suppose that G is not locally nilpotent. Then every non-permutable subgroup of G is self-normalizing if and only if there exist a prime p and an abelian normal subgroup A of G with no elements of order p such that the following conditions hold*

(1) $G = AP$ and $A \cap P = \langle 1\rangle$, where $P = \langle x\rangle$ is a cyclic p-subgroup and $C_P(A) = \langle g^p\rangle$;

(2) the commutator subgroup $[G, G] = A$;

(3) P is self-centralized, that is $C_G(P) = P$; and

(4) every subgroup of A is G–invariant.

Applying Theorem 2.10 and with some extra work, we are able to obtain.

Proposition 3.4. *Let G be a group whose subgroups are either permutable or self-normalizing. If G is locally nilpotent, then every subgroup of G is permutable.*

Proof. If G is finitely generated, then G is nilpotent, and the proof is over. Suppose that G has no a finite set of generators. Let $F \leq G$ be a finitely generated subgroup of G. Pick $x \notin F$. Then $\langle x, F\rangle$ is nilpotent and so $F \neq N_{\langle x,F\rangle}(F)$. Thus F is ascendant. Let

$$F = F_0 \trianglelefteq F_1 \trianglelefteq \cdots \trianglelefteq F_\alpha \leq F_{\alpha+1} \trianglelefteq \cdots F_\gamma = G$$

be an ascending series between F and G, and define $L_1 = \langle F^x \mid x \in F_2\rangle$. Any F^x is normal in F_1 if $x \in F_2$, and it readily follows that L_1 is hyperabelian. By Theorem 2.10, L_1 has no self-normalizing subgroups, and hence every subgroup of L_1 is permutable. By Theorem

3.2, L_1 is metabelian, that is an abelian extension of an abelian group. Proceeding in the same way we see that F_3 is metabelian, and applying transfinite induction we obtain that F^G is also metabelian. Hence G contains an abelian normal subgroup. Using transfinite induction again, we deduce that G itself is hyperabelian. By Theorem 2.10, G has no self-normalizing subgroups, and hence every subgroup of G is permutable, as required. □

Corollary 3.5. *Let G be a group whose subgroups are either permutable or self-normalizing. If G is not periodic, then every subgroup of G is permutable. If G further is torsion-free, then G is abelian.*

A subgroup H of a group G is said to be *contranormal* if $H^G = H$; it is clair that this concept defines subgroups that are some kind of antipodes of subnormal and normal subgroups. In the paper M. de Falco, L.A. Kurdachenko and I.Ya. Subbotin [7] the following description of groups whose subgroups are either subnormal or contranormal was obtained.

Theorem 3.6 (M. de Falco, L. A. Kurdachenko, I. Ya. Subbotin [7])**.** *Let G be a group such that $G \neq [G, G]$. Every non-subnormal subgroup of G is contranormal if and only one of the following holds.*

(1) Every subgroup of G is subnormal;

(2) G is a Baer group and has a normal subgroup H whose subgroups are subnormal such that G/H is a Prüfer p–group for some prime p; or

(3) $G = [G, G]P$, where $P = \langle g \rangle$ is a cyclic contranormal subgroup and there is a prime q such that every subgroup of $[G, G]\langle g^q \rangle$ is subnormal.

Acknowledgements

This research was supported by Proyecto MTM2007-60994 of Dirección General de Investigación del Ministerio de Educación (Spain).

References

[1] R. Baer, *Situation der Untergruppen und Struktur der Gruppe*, Heidelberg Akad. W. **2** (1933), 12–17.

[2] R. Baer, *Nilgruppen*, Math. Z. **62** (1955), 402–437.

[3] C. Casolo, *On the structure of groups with all subgroups subnormal*, J. Group Th. **5** (2002), 293–300.

[4] S. N. Chernikov, *On normalizer condition*, Math. Notes **3** (1968), 45–50.

[5] R. Dedekind, *Über Gruppen deren sämmtliche t. Normalteiler sind*, Math. Ann. **48** (1897), 548–561.

[6] G. Ebert, S. Bauman, *A note of subnormal and abnormal chains*, J. Algebra **36** (1975), 287–293.

[7] M. de Falco, L. A. Kurdachenko, I. Ya. Subbotin, *Groups with only abnormal and subnormal subgroups*, Atti Sem. Mat. Fis. Univ. Modena **46** (1988), 435–442.

[8] A. Fattahi, *Groups with only normal and abnormal subgroups*, J. Algebra **28** (1974), 15–19.

[9] K. W. Gruenberg, *The Engel elements of soluble groups*, Illinois J. Math. **3** (1959), 151–168.

[10] H. Heineken, L. A. Kurdachenko, *Groups with subnormality for all subgroups that are not finitely generated*, Ann. Mat. Pura Appl. IV Ser. **169** (1995), 203–232.

[11] H. Heineken, I. J. Mohamed, *A group with trivial centre satisfying the normalizer condition*, J. Algebra **10** (1968), 368–376.

[12] K. Iwasawa, *On the structure of infinite M–groups*, Jap. Journal Math. **18** (1943), 709–728.

[13] M. I. Kargapolov, *On generalized soluble groups*, Algebra and Logic **2** (1963), 19–28.

[14] L. A. Kurdachenko, J. Otal, A. Russo, G. Vincenzi, *Abnormal subgroups and Carter subgroups in some classes of infinite groups*, J. Algebra **297** (2006), 273–291.

[15] L. A. Kurdachenko, H. Smith, *Groups with all subgroups either subnormal or self-normalizing*, J. Pure and Applied Algebra **196** (2005), 271–278.

[16] A. I. Maltsev, *Torsion-free nilpotent groups*, Izv. AN SSSR, series Math. **13** (1949), 9–32.

[17] W. Möhres, *Auflösbare Gruppen mit endlichem Exponenten, deren Untergruppen alle subnormal sind II*, Rend. Sem. Mat. Univ. Padova **81** (1989), 269–287

[18] W. Möhres, *Gruppen, deren Untergruppen alle subnormal sind*, Arch. Math. **54** (1990), 232–235.

[19] W. Möhres, *Hyperzentrale Torsionsgruppen, deren Untergruppen alle subnormal sind*, Illinois J. Math. **35** (1991), 147–157.

[20] B. I. Plotkin, *To the theory of locally nilpotent groups*, Doklady AN SSSR **76** (1951), 639–641.

[21] R. Schmidt, Subgroups lattices of groups, Walter de Gruyter, Berlin, 1994.

[22] H. Smith, *Torsion-free groups with all subgroups subnormal*, Archiv Math. **76** (2001), 1–6.

[23] H. Smith, *Residually nilpotent groups with all subgroups subnormal*, J. Algebra **244** (2001), 845–850.

[24] H. Smith, *Torsion-free groups with all non-nilpotent subgroups subnormal*, Quaderni di Mat. **8** (2001), 297–308.

[25] H. Smith, *Groups with all non-nilpotent subgroups subnormal*, Quaderni di Mat. **8** (2001), 309–326.

[26] S. E. Stonehewer, *Permutable subgroups of infinite groups*, Math. Z. **125** (1972), 1–16.