S$\vec{e}$MA

BOLETÍN NÚMERO 45

Diciembre 2008

sumario

Editorial 5

Palabras del Presidente 7

Artículos 9

Some Liouville-type theorems for elliptic equations and their consequences, por A. Farina 9

Sobre la individualización de tratamientos en oncología radioterápica, por María Isabel Núñez et al. 35

Free software for numerical simulation in industrial problems, por G. Fdez-Manín, M. Meis y F. Varas 69

Splitting and composition methods in the numerical integration of differential equations, por S. Blanes, F. Casas y A. Murua 89

Educación Matemática 147

La ecuación de Riccati, por J.M. Amigó 147

Resúmenes de tesis doctorales 171

Resúmenes de libros 177

Noticias de S$\vec{e}$MA 179

Anuncios 185

Boletín de la Sociedad Española de Matemática Aplicada S$\vec{e}$MA

Página web de S$\vec{e}$MA
`http://www.sema.org.es/`

e-mail
`info@sema.org.es`

Dirección Editorial: Dpto. de Matemáticas. E.T.S.I. Industriales. Univ. de Castilla - La Mancha. Avda. de Camilo José Cela s/n. 13071. Ciudad Real. `boletin.sema@uclm.es`

ISSN 1575-9822.
Depósito Legal: AS-1442-2002.

Imprime: Gráficas Lope. C/ Laguna Grande, parc. 79, Políg. El Montalvo II 37008. Salamanca.

Diseño de portada: Ernesto Aranda
Ilustración de portada: Hiperboloides en el interior de la *Sagrada Familia* (A. Gaudí).

Estimados socios,

Aquí tenéis un nuevo número del boletín de la S$\vec{e}$MA, el último de 2008, en el que podréis encontrar información sobre las últimas noticias acaecidas en la Sociedad, de las que nos informa nuestro reelegido Presidente, Carlos Vázquez.

Queremos aprovechar también para recordaros que el próximo año celebramos en Ciudad Real el congreso de la Sociedad, el XXI CEDYA / XI CMA, en el que el Grupo Editor participa como parte de la organización. Es por ello que no podemos dejar pasar la oportunidad para animaros a asistir a este evento, del que se encuentra ya abierto el plazo de inscripción y presentación de comunicaciones. Podéis encontrar más información en la sección de Anuncios.

Como es habitual, en la sección de Artículos Científicos incluimos cuatro interesantes trabajos: uno sobre teoremas tipo Lioville para ecuaciones elípticas, otro sobre tratamientos en oncología radioterápica, un tercero sobre software libre para simulación numérica (estos dos últimos con un carácter más divulgativo) y el último sobre métodos numéricos para integración y ecuaciones diferenciales. Además, en la sección de Educación Matemática aparece un artículo de J.M. Amigó sobre La ecuación de Ricatti. Esperamos que los contenidos de este número sean de vuestro interés.

Sin más, nos despedimos con nuestros mejores deseos para el 2009.

Recibid un cordial saludo,

Grupo Editor
`boletin.sema@uclm.es`

Queridos socios de S$\vec{e}$MA, quiero aprovechar el último número del Boletín en 2008 para dirigirme a todos vosotros y referirme a algunas novedades relacionadas con S$\vec{e}$MA.

En primer lugar os informo de que, tras las elecciones celebradas el pasado mes de septiembre, han renovado sus cargos como miembros del Comité Ejecutivo: Rafael Bru (Universidad de Valencia), Inmaculada Higueras (Universidad de Navarra) y Pablo Pedregal (Universidad de Castilla-La Mancha), y yo mismo como Presidente. De este modo, todos los elegidos hemos renovado el compromiso de seguir trabajando desde nuestros cargos en beneficio de la Sociedad.

En 2008 la S$\vec{e}$MA ha apoyado un buen número de actividades, a cuyos organizadores agradecemos el buen trabajo realizado y el propiciar con ello la difusión de la Sociedad dentro de los distintos eventos.

También en 2008 se ha celebrado el XIII Escuela Jacques-Louis Lions Hispano Francesa sobre Simulación Numérica en Física e Ingeniería, organizado por el Departamento de Matemática Aplicada de la Universidad de Valladolid, con un aumento significativo de participantes respecto de ediciones anteriores. Se trataba de la primera edición planteada en colaboración con la SMAI francesa a la que agradecemos sus aportaciones. Desde S$\vec{e}$MA también enviamos nuestro sincero agradecimiento y felicitación a las personas más directamente involucradas en las tareas de organización del evento, coordinadas por Mari Paz Calvo, y que han posibilitado el éxito del mismo.

En 2009 S$\vec{e}$MA seguirá apoyando un buen número de actividades relacionadas con la Matemática Aplicada, a las que se dará publicidad por distintos medios y os animamos a asistir. En especial, esperamos vuestra participación en *nuestro congreso* CEDYA-CMA, que está siendo organizado con gran profesionalidad por un grupo del Departamento de Matemáticas de la Universidad de Castilla-La Mancha. Como en ediciones anteriores, la calidad y variedad de los conferenciantes invitados, las sesiones especiales, la posibilidad de escuchar y presentar comunicaciones en distintos temas de la Matemática Aplicada y los interesantes actos sociales previstos en Ciudad Real y Almagro, deberían convertir el XXI CEDYA-XI CMA en una cita inexcusable en nuestras agendas.

Con respecto a los Premios de S$\vec{e}$MA, ya publicamos en este número las bases de la convocatoria del "XII Premio al Joven Investigador", en las que cabe destacar el incremento de la dotación económica con respecto a las ediciones anteriores. Por otro lado, y esta es una de las novedades importantes, se ha optado por no convocar el premio a la Divulgación de la Matemática Aplicada y se ha creado el nuevo "Premio al mejor artículo del Boletín S$\vec{e}$MA", que premiará un artículo entre todos los publicados cada año en el Boletín. Las bases del nuevo premio también aparecen en este número del Boletín, en particular las que afectan a artículos aparecidos en 2008. Por otra parte, la sección de “Matemáticas e Industria”pasará a ser la de “Divulgación, Matemáticas e Industria”. Con ambas medidas pretendemos, por un lado, estimular la presencia de artículos de divulgación dentro de una sección propia y, por otro, premiar anualmente los trabajos de mayor calidad aparecidos en el Boletín.

Desde estas páginas quiero animaros a que le deis la mayor difusión posible a los premios de S$\vec{e}$MA y a que fomenteis el envío de artículos de calidad para su posible publicación en el Boletín.

Finalmente, como es habitual, insisto en reiteraros mi invitación a participar activamente en S$\vec{e}$MA, aportando ideas y posibles mejoras, con objetivo de contribuir al mejor funcionamiento de *nuestra* Sociedad.

Mis mejores deseos para el 2009.

Un abrazo,

Carlos Vázquez Cendón
Presidente de S$\vec{e}$MA
`carlosv@udc.es`

Bol. Soc. Esp. Mat. Apl.
n^o 45(2008), 9–34

SOME LIOUVILLE-TYPE THEOREMS FOR ELLIPTIC EQUATIONS AND THEIR CONSEQUENCES

ALBERTO FARINA

LAMFA, CNRS UMR 6140,
Université de Picardie Jules Verne
Faculté de Mathématiques et d'Informatique
33, rue Saint-Leu 80039 Amiens, France

alberto.farina@u-picardie.fr

Abstract

This paper is a short excursion into the world of the classification of solutions of partial differential equations of elliptic type defined on the entire Euclidean space $\mathbf{R}^N$ with a special emphasis on Liouville-type theorems. We have decided to present the results by considering some problems of physical and geometric interest whose solution depends on some Liouville-type theorem. Typical examples are a conjecture of De Giorgi concerning monotone solutions of the Allen-Cahn equation arising in phase transition and the celebrated Bernstein problem for minimal graphs. We discuss the sharpness of the considered results by means of examples and counterexamples and we give several references for further reading. This work is organized as follows:

1. Introduction
2. The linear case
3. The semilinear case
4. The quasilinear case

Key words: *Linear and nonlinear PDEs of elliptic type, Liouville-type theorems.*

AMS subject classifications: *35J15 35J60 35B05*

1 The linear case

A celebrated theorem in complex analysis states that

Theorem 2.1 *A bounded complex function $f : \mathbf{C} \to \mathbf{C}$ which is holomorphic on the entire complex plane is a constant function.*

Fecha de recepción: 21/07/2008. Aceptado (en forma revisada): 27/09/2008.

This theorem, known as *Liouville Theorem*, was first announced by J. Liouville [53,54] in a *Note in the Comptes Rendus de l' Académie des Sciences* (Paris, December 9, 1844) for the special case of a doubly-periodic function. Two weeks later, in another *Note in the Comptes Rendus* (Paris, December 23, 1844), A. Cauchy [11,12] published the first proof of the above stated theorem.

A proof of Theorem 2.1, usually based on Cauchy's integral formula, can be found in any book of complex analysis. A different way to prove the above theorem is to use the following result concerning real-valued harmonic functions defined on the entire Euclidean space $\mathbf{R}^N$.

Theorem 2.2 *Assume $N \geq 1$ and let $u \in C^2(\mathbf{R}^N)$ be a bounded solution of $-\Delta u = 0$ in $\mathbf{R}^N$. Then u is a constant function.*

To prove Theorem 2.1 we recall that the real and the imaginary parts of a holomorphic function satisfy the Cauchy-Riemann equations, therefore they are bounded real-valued harmonic functions on $\mathbf{R}^2$. The desired conclusion then follows by applying Theorem 2.2.

A simple proof of Theorem 2.2 can be obtained by the following classic properties of harmonic functions (see for instance [20,38]):

Theorem 2.3 (Mean value properties) *Assume $N \geq 1$ and let $u \in C^2(\Omega)$ be a solution of $-\Delta u = 0$ in an open set $\Omega \subset \mathbf{R}^N$, then for any open ball $B = B(x,r) \subset\subset \Omega$ we have:*

$$u(x) = \frac{1}{|B|}\int_B u(y)\,dy = \frac{1}{|\partial B|}\int_{\partial B} u(s)\,ds.$$

Here, $|B|$ denotes the Lebesgue measure of the N-dimensional ball B, while $|\partial B|$ denotes the surface measure of the boundary of B.

Proof of Theorem 2.2. The following simple and very elegant proof is due to E. Nelson [59]. For every $x \in \mathbf{R}^N$ and every $R > 0$ let $B(x,R)$ be the open ball centered at x and with radius R. The mean value formula gives

$$|u(x) - u(0)| = \frac{1}{|B(0,R)|}\Big|\int_{B(x,R)} u(y)\,dy - \int_{B(0,R)} u(y)\,dy\Big| =$$
$$\leq \frac{1}{|B(0,R)|}\int_{D(x,0,R)} |u(y)|\,dy \leq \left(\sup_{\mathbf{R}^N}|u|\right)\frac{|D(x,0,R)|}{|B(0,R)|},$$

where $D(x,0,R)$ is the symmetric difference of $B(x,R)$ and $B(0,R)$.
Since $\lim_{R\to+\infty}\frac{|D(x,0,R)|}{|B(0,R)|} = 0$ we obtain $u(x) = u(0)$. Thus, u is constant. ■

Theorem 2.2 admits an extension to harmonic functions which are bounded on one side (from above or from below). More precisely, we have:

Theorem 2.4 *Assume $N \geq 1$ and let $u \in C^2(\mathbf{R}^N)$ be a solution of $-\Delta u = 0$ in $\mathbf{R}^N$. If u is bounded on one side, then u must be constant.*

A simple proof of the above theorem is based on the following important result (which finds its origin in the work of C.G. Axel Harnack in 1887 [41]).

Theorem 2.5 (Harnack inequality) *Assume $N \geq 1$, $x_0 \in \mathbf{R}^N$ and $r > 0$. Let $u \in C^2((B(x_0, 4r))$ be a non-negative harmonic function. Then we have:*

$$\sup_{B(x_0,r)} u \leq 3^N \inf_{B(x_0,r)} u. \tag{1.1}$$

Proof. For any two points y and z belonging to $B(x_0, r)$, an application of the mean value theorem gives,

$$v(y) = \frac{1}{|B(y,r)|} \int_{B(y,r)} v(x)\, dx \leq \frac{1}{|B(y,r)|} \int_{B(z,3r)} v(x)\, dx$$

$$= \frac{|B(z,3r)|}{|B(y,r)|} \frac{1}{|B(z,3r)|} \int_{B(z,3r)} v(x)\, dx = 3^N v(z),$$

which immediately implies the desired inequality. ∎

Proof of Theorem 2.4. Up to consider $-u$ instead of u, we can assume that u is bounded from below. Set $m := \inf_{\mathbf{R}^N} u$ and $v = u - m$. Clearly v is a non-negative harmonic function with $\inf_{\mathbf{R}^N} v = 0$. For every $r > 0$, inequality (1.1) yields:

$$\sup_{B(0,r)} v \leq 3^N \inf_{B(0,r)} v.$$

By letting $r \to +\infty$, the right-hand side of the latter inequality tends to zero, hence $u \equiv m$ on $\mathbf{R}^N$. ∎

Theorem 2.4 says that any entire non-constant harmonic function must be unbounded both from below and from above. This property is not a prerogative of harmonic functions. Indeed, the Laplace operator Δ shares this property with a wide class of partial differential operators of elliptic type. As we shall see in the sequel, this is mainly due to the fact that the *Harnack inequality* (1.1) holds true for every second order uniformly elliptic operators both in divergence form and non-divergence form. More precisely we have:

Theorem 2.6 ([57]) *Let $N \geq 2$, $x_0 \in \mathbf{R}^N$, $r > 0$ and $A = (a_{hk})$ be a real symmetric matrix whose coefficients are measurable and bounded functions defined on $B(x_0, 4r) \subset \mathbf{R}^N$. Assume that there are $0 < \lambda \leq \Lambda < +\infty$ such that*

$$\forall\, x \in B(x_0, 4r), \quad \forall\, \xi \in \mathbf{R}^N \setminus \{0\}, \quad \lambda|\xi|^2 \leq \sum_{h,k=1}^{N} a_{hk}(x)\xi_h \xi_k \leq \Lambda|\xi|^2. \tag{1.2}$$

Let $u \in H^1(B(x_0, 4r))$ be a distribution solution of

$$\begin{cases} -div(A(x)\nabla u) = 0 & in \quad B(x_0, 4r), \\ u(x) \geq 0 & a.e. \quad in \quad B(x_0, 4r). \end{cases} \tag{1.3}$$

Then we have:

$$\text{ess sup}_{B(x_0,r)} u \leq C \, \text{ess inf}_{B(x_0,r)} u, \tag{1.4}$$

where $C = C(N, \Lambda/\lambda)$.

The analogous result for uniformly elliptic operators in non-divergence form is:

Theorem 2.7 ([46,47]) *Let $N \geq 2$, $x_0 \in \mathbf{R}^N$, $r > 0$ and $A = (a_{hk})$ be a real symmetric matrix whose coefficients are bounded and measurable functions defined on $B(x_0, 4r) \subset \mathbf{R}^N$. Assume that there are $0 < \lambda \leq \Lambda < +\infty$ such that*

$$\forall x \in B(x_0, 4r), \quad \forall \xi \in \mathbf{R}^N \setminus \{0\}, \qquad \lambda|\xi|^2 \leq \sum_{h,k=1}^{N} a_{hk}(x)\xi_h\xi_k \leq \Lambda|\xi|^2. \tag{1.5}$$

Let $u \in W^{2,N}(B(x_0, 4r))$ be a (strong) solution of

$$\begin{cases} -\sum_{h,k=1}^{N} a_{hk}(x) \frac{\partial^2 u}{\partial x_h \partial x_k} = 0 & a.e. \quad in \quad B(x_0, 4r), \\ \\ u(x) \geq 0 & in \quad B(x_0, 4r), \end{cases} \tag{1.6}$$

then we have:

$$\sup_{B(x_0,r)} u \leq C \inf_{B(x_0,r)} u, \tag{1.7}$$

where $C = C(N, \frac{\Lambda}{\lambda})$.

These results are cornerstones of the modern theory of linear PDEs. Theorem 2.6 was proved by J. Moser in 1961 while Theorem 2.7 is due to the work of N.V. Krylov and M.V. Safonov. For problem (1.6), with $N = 2$ and $u \in C^2(B(x_0; 4r))$, the Harnack inequality (1.7) was first proved by J. Serrin [65] in 1955.

It is beyond the scope of this work to present the proofs of the above classical theorems. The interested reader can consult the original works [57,46,47] and also [38]. The elegant proof given by J. Serrin [65] is based merely on the maximum principle.

Let us mention that Moser's result also implies that solutions of the linear uniformly elliptic problem (1.3) are Hölder-continuous. This celebrated result was first proved by E. De Giorgi [16] in 1957 and its discovery was crucial for the development of the higher-dimensional theory of quasilinear problems and also led to the solution of Hilbert's nineteenth problem [42].

Theorems 2.6 and 2.7 immediately imply the analogue of Theorem 2.4. Actually we have more:

Theorem 2.8 ([57,67]) *Let $N \geq 2$ and $A = (a_{hk})$ be a real symmetric matrix whose coefficients are bounded and measurable functions defined on $\mathbf{R}^N$. Assume that there are $0 < \lambda \leq \Lambda < +\infty$ such that*

$$\forall\, x \in \mathbf{R}^N, \quad \forall\, \xi \in \mathbf{R}^N \setminus \{0\}, \quad \lambda|\xi|^2 \leq \sum_{h,k=1}^{N} a_{hk}(x)\xi_h\xi_k \leq \Lambda|\xi|^2$$

and let $u \in H^1_{loc}(\mathbf{R}^N)$ be a distribution solution of

$$-\mathcal{L}u := -div(A(x)\nabla u) = 0 \qquad in \qquad \mathbf{R}^N. \tag{1.8}$$

Then:

i) Either u is constant or u must become both positively and negatively unbounded as $|x|$ tends to $+\infty$.

ii) If u is not constant then there exist $\alpha = \alpha(N, \frac{\Lambda}{\lambda}, u) > 0$ and $\beta = \beta(N, \frac{\Lambda}{\lambda}) > 0$ such that

$$M(R) \geq \alpha R^\beta, \qquad m(R) \leq -\alpha R^\beta, \qquad \forall\, R >> 1, \tag{1.9}$$

where, for every $t \geq 0$, we have set $M(t) = \sup_{|x|<t} u(x)$ and $m(t) = \inf_{|x|<t} u(x)$.

In particular, a non-constant solution of (1.8) *must go to infinity at least as fast as the power $|x|^\beta$, as $|x|$ tends to $+\infty$.*

For uniformly elliptic operators in non-divergence form, we have:

Theorem 2.9 *Let $N \geq 2$ and $A = (a_{hk})$ satisfy the assumptions of Theorem 2.8. Let $u \in W^{2,N}_{loc}(\mathbf{R}^N)$ be a solution of:*

$$-\mathcal{M}u := -\sum_{h,k=1}^{N} a_{hk}(x)\frac{\partial^2 u}{\partial x_h \partial x_k} = 0 \qquad a.e. \qquad in \qquad \mathbf{R}^N. \tag{1.10}$$

Then, the conclusions of Theorem 2.8 hold true.

Proof of Theorem 2.8. Proof of i) - Since we have at our disposal the Harnack inequality (1.4), the desired conclusion follows as in the proof of Theorem 2.4. ■

Proof of ii) - Assume that u is not constant. The function $M(R)$ is increasing and the function $m(R)$ is decreasing for $R > 0$. Furthermore, by step *i)* there exists $R_0 > 0$ such that $M(R) > 0$ and $m(R) < 0$ for $R \geq R_0$.

For every $t > 0$ the functions $M(4t) - u$ and $u - m(4t)$ are non-negative on the ball $B(0, 4t)$. Therefore, an application of Harnack inequality (1.4) yields:

$$M(4t) - m(t) \leq C[M(4t) - M(t)], \qquad \forall\, t > 0, \tag{1.11}$$

$$M(t) - m(4t) \leq C[m(t) - m(4t)], \qquad \forall\, t > 0, \tag{1.12}$$

where $C = C(N, \frac{\Lambda}{\lambda}) > 1$ (since $u \not\equiv$ const.) is the constant appearing in Theorem 2.6.

Adding the above inequalities we have,

$$M(4t) - m(4t) \geq \Big(\frac{C+1}{C-1}\Big)[M(t) - m(t)] \qquad \forall\, t > 0,$$

and by iterating we obtain:

$$M(4^n t) - m(4^n t) \geq \Big(\frac{C+1}{C-1}\Big)^n [M(t) - m(t)] \qquad \forall\, t > 0, \qquad \forall\, n \in \mathbf{N}^\star.$$

The last inequality, together with the monotonicity of the function $\omega(t) = M(t) - m(t)$ (the oscillation of u over the ball $B(0,t)$), immediately imply:

$$M(R) - m(R) \geq \gamma R^\beta \qquad \forall\, R \geq 4R_0, \tag{1.13}$$

with

$$\beta = \ln_4 \Big(\frac{C+1}{C-1}\Big) > 0, \qquad \gamma = \frac{M(R_0) - m(R_0)}{(4R_0)^\beta} > 0. \tag{1.14}$$

On the other hand, from (1.11) – (1.12) and (1.13) it follows that

$$(C-1)M(4t) \geq CM(t) - m(t) \geq \gamma t^\beta \qquad \forall\, t > 4R_0,$$

$$(C-1)m(4t) \leq -[M(t) - Cm(t)] \leq -\gamma t^\beta \qquad \forall\, t > 4R_0.$$

Hence, the desired conclusion (1.9) follows by taking $\alpha = \frac{\gamma}{4^\beta (C-1)} > 0$ and β as in (1.14). ■

The same proof also applies to the case of a uniformly elliptic operator in non-divergence form (one has only to use the Harnack inequality (1.7), instead of inequality (1.4), and the maximum principle of Alexandrov, Bakelman and Pucci [38] for solutions of class $W^{2,N}$) and for this reason we omit it.

Let $\mathcal{L}$ and $\mathcal{M}$ be the uniformly elliptic operators considered in Theorem 2.8 and Theorem 2.9, respectively. For every $\delta \geq 0$ we define the vector spaces:

$$\mathcal{H}_\delta(\mathcal{L}) := \{\, u \in H^1_{loc}(\mathbf{R}^N) : -\mathcal{L}u = 0,\ u(x) = O(|x|^\delta) \text{ as } |x| \to +\infty \,\},$$

$$\mathcal{H}_\delta(\mathcal{M}) := \{\, u \in W^{2,N}_{loc}(\mathbf{R}^N) : -\mathcal{M}u = 0,\ u(x) = O(|x|^\delta) \text{ as } |x| \to +\infty \,\}.$$

From Theorem 2.8 and Theorem 2.9, we know that there exists $\beta = \beta(N, \Lambda/\lambda) > 0$ such that

$$\forall\, \delta < \beta \qquad \dim \mathcal{H}_\delta(\mathcal{L}) = 1 \qquad (\text{resp.} \quad \dim \mathcal{H}_\delta(\mathcal{M}) = 1\,).$$

If one compares the above results with those proved for the special case of the Laplace operator, the following questions arise in a natural way:

Question A - Find the greatest exponent $\overline{\delta} \in (0, +\infty]$ such that

$$\forall\, \delta < \overline{\delta} \qquad \dim \mathcal{H}_\delta(\mathcal{L}) = 1 \qquad (\,\text{resp.} \quad \dim \mathcal{H}_\delta(\mathcal{M}) = 1\,).$$

Question B - Let $\delta > 0$ be a fixed constant. What is the dimension of $\mathcal{H}_\delta(\mathcal{L})$? Is it finite ? In the affirmative case, can one estimate $\dim\mathcal{H}_\delta(\mathcal{L})$ in terms of N and $\frac{\Lambda}{\lambda}$? Can one describe the structure of the vector space $\mathcal{H}_\delta(\mathcal{L})$? Same questions for an operator $\mathcal{M}$ in non-divergence form.

The value $\overline{\delta}$ gives the "least growth at infinity" of every non-constant solution of (1.8) (resp. (1.10)). The following examples, due to N. Meyers [55] and J. Serrin and H. Weinberger [67], show that, in general, $\overline{\delta}$ depends on both the dimension N and the ratio Λ/λ.

Example 2.10 ([55,67]) Let $N \geq 2$. For $\eta \in \mathbf{R}$ consider the real symmetric matrix $A = (a_{hk})$ with measurable and bounded coefficients given by:

$$a_{kh}(x) := (1+\eta^2)^{-\frac{1}{2}}\Big(\delta_{hk} + \eta^2 \frac{x_h x_k}{|x|^2}\Big), \qquad h,k = 1, ..., N.$$

The corresponding linear partial differential equation (1.8) is uniformly elliptic with $\lambda = (1+\eta^2)^{-\frac{1}{2}}$ and $\Lambda = (1+\eta^2)^{\frac{1}{2}}$. Furthermore, it admits the solution:

$$u(x) = x_1|x|^{-\theta}, \qquad \theta = \frac{1}{2}\{N - [(N-2)^2 + 4(N-1)(1+\eta^2)^{-1}]^{\frac{1}{2}}\} \in (0,1),$$

which goes to infinity at the rate $|x|^{1-\theta}$. Hence

$$\overline{\delta} \leq 1 - \theta = \frac{1}{2}\Big\{\Big[(N-2)^2 + 4(N-1)\Big(\frac{\Lambda}{\lambda}\Big)^{-1}\Big]^{\frac{1}{2}} - (N-2)\Big\} \in (0,1),$$

which goes to zero as the ratio Λ/λ goes to $+\infty$.

As far as we know, the most general result about questions A and B was proved (independently) in 1997 by T. Colding and W.P. Minicozzi [15] and by P. Li [51].

Theorem 2.11 ([15,51]) *Let $N \geq 2$ and $A = (a_{hk})$ be a real symmetric matrix whose coefficients are measurable and bounded functions defined on $\mathbf{R}^N$. Assume that there are $0 < \lambda \leq \Lambda < +\infty$ such that*

$$\forall\, x \in \mathbf{R}^N, \quad \forall\, \xi \in \mathbf{R}^N \setminus \{0\}, \qquad \lambda|\xi|^2 \leq \sum_{h,k=1}^{N} a_{hk}(x)\xi_h\xi_k \leq \Lambda|\xi|^2.$$

Then, $\dim \mathcal{H}_\delta(\mathcal{L})$ is finite for all $\delta > 0$. Furthermore, there exists a positive constant C, depending only on N and Λ/λ, such that

$$\forall\, \delta \geq 1 \qquad \dim\ \mathcal{H}_\delta(\mathcal{L}) \leq C\delta^{N-1}.$$

For the operator $\mathcal{M}$ we have:

Theorem 2.12 ([51]) *Let $N \geq 2$ and $A = (a_{hk})$ be a real symmetric matrix whose coefficients are measurable and bounded functions defined on $\mathbf{R}^N$. Assume that there are $0 < \lambda \leq \Lambda < +\infty$ such that*

$$\forall\, x \in \mathbf{R}^N, \quad \forall\, \xi \in \mathbf{R}^N \setminus \{0\}, \qquad \lambda|\xi|^2 \leq \sum_{h,k=1}^{N} a_{hk}(x)\xi_h\xi_k \leq \Lambda|\xi|^2.$$

Then, there exists a positive constant C, depending only on N and Λ/λ, such that

$$\forall\, \delta \geq 1 \qquad \dim \mathcal{H}_\delta(\mathcal{M}) \leq C\delta^{N-1}.$$

The proofs of these theorems rely on the *doubling property* of the N-dimensional Lebesgue measure $\mathcal{L}_N$:

$$\exists\, C_1 = C_1(N) > 0 : \forall\, x_0 \in \mathbf{R}^N, \forall\, r > 0 \quad \mathcal{L}_N(B(x_0, 2r)) \leq C_1 \mathcal{L}_N(B(x_0, r)),$$

and the *generalized mean value inequality* for non-negative subsolutions:

$$\exists\, C_2 = C_2(N, \Lambda/\lambda) > 0 \quad : \quad u(x_0) \leq \frac{C_2}{r^N} \int_{B(x_0,r)} u(x)\, dx,$$

for all u, for all $x_0 \in \mathbf{R}^N$ and for all $r > 0$ satisfying

$$u \geq 0, \quad -\mathcal{L}u \leq 0 \quad \text{on} \quad B(x_0, r).$$

The last property is a by-product of Moser's iteration method (resp. of Krylov-Safonov's analysis). See for instance [38] and [63]. Actually, the arguments used in [15] and [51] also work in a more general geometrical setting. For instance, they can be used to study Liouville-type properties for the Laplace-Beltrami operator of some complete Riemannian manifolds. For these topics we refer the interested reader to [15], [51] and [52], as well as the references therein.

2 The semilinear case

This section is devoted to prove some Liouville-type theorems for the semilinear Poisson equation

$$-\Delta u + f(u) = 0 \qquad \text{in} \quad \mathbf{R}^N, \tag{3.1}$$

where $f : \mathbf{R} \to \mathbf{R}$ is a continuous function. The above equation includes some models which are important for applications. It arises in the theory of superconductors and superfluids [39,50], in the theory of interfaces in both gasses and solids [2,62], in cosmology [10,36] and also in the study of various chemical models [19].

The first result we consider is the following:

Theorem 3.1 ([25]) *Assume $N \geq 1$ and let $f : \mathbf{R} \to \mathbf{R}$ be a continuous, non-decreasing function, with $f \not\equiv 0$. Let $u \in C^2(\mathbf{R}^N)$ be a solution of:*

$$\begin{cases} -\Delta u + f(u) = 0 & \text{in} \quad \mathbf{R}^N, \\ u(x) = o(|x|^2) & \text{as} \quad |x| \to +\infty. \end{cases}$$

Then u is a constant function, that is to say, $u \equiv c \in \mathbf{R}$, and $f(c) = 0$.

Proof. For every $x_0 \in \mathbf{R}^N$ and every $\epsilon > 0$ we define the function:

$$v(x) = u(x) - u(x_0) + 2\epsilon - \epsilon|x - x_0|^2, \qquad \forall\, x \in \mathbf{R}^N.$$

Since $v(x_0) > 0$ and $\lim_{|x|\to+\infty} v(x) = -\infty$, there exists a point $y_0 \in \mathbf{R}^N$ such that $v(y_0) = \max_{x\in\mathbf{R}^N} v(x) > 0$ and thus $\Delta v(y_0) \leq 0$. Therefore, we have:

$$u(x_0) - 2\epsilon < u(y_0), \qquad \Delta u(y_0) \leq 2\epsilon N$$

and the monotonicity of f implies:

$$f(u(x_0) - 2\epsilon) \leq f(u(y_0)) = \Delta u(y_0) \leq 2\epsilon N.$$

By letting $\epsilon \to 0$ in the latter inequality we obtain that $f(u(x_0)) \leq 0$. By a similar argument, using the function $w(x) = u(x) - u(x_0) - 2\epsilon + \epsilon|x - x_0|^2$ instead of v, we also have $f(u(x_0)) \geq 0$.

These facts implies:

$$\forall\, x \in \mathbf{R}^N, \qquad -\Delta u(x) = -f(u(x)) = 0$$

and in particular u is a harmonic function on $\mathbf{R}^N$ which cannot be surjective since otherwise we would have $f \equiv 0$, a contradiction. This yields immediately the desired conclusion by invoking Theorem 2.4. ■

Remarks 3.2

i) The above result is sharp, since the function $u(x) = |x|^2$ solves the equation (3.1) with $f = 2N$.

ii) The assumption $f \not\equiv 0$ cannot be removed, since the Laplace equation admits non-constant affine solutions on $\mathbf{R}^N$.

iii) Note that the monotonicity of f (i.e. f non-decreasing) cannot be dropped, since $u(x) = \sin(x_1)$ is a solution of (3.1) with $f(t) = -t$

iv) Under the more restrictive assumption $u(x) = o(|x|)$ as $|x| \to +\infty$, the above Theorem 3.1 (including the case $f \equiv 0$) was proved by J. Serrin [66] in 1972.

In 1978, E. De Giorgi posed the following striking question:

De Giorgi's conjecture ([18]) *Let $u \in C^2(\mathbf{R}^N)$ be a solution of*

$$-\Delta u + u^3 - u = 0, \tag{3.2}$$

such that

$$|u| \leq 1, \qquad \frac{\partial u}{\partial x_N} > 0$$

in the whole $\mathbf{R}^N$. Is it true that, for every $\lambda \in \mathbf{R}$, the sets $\{u = \lambda\}$ are hyperplanes, at least if $N \leq 8$?

Equivalently, De Giorgi's conjecture can be reformulated by saying that the considered solution u depends only on one variable, up to rotations (i.e. u is one-dimensional).

N. Ghoussoub and C. Gui [35] proved De Giorgi's conjecture for N=2 in 1998, while L. Ambrosio and X. Cabré [4] established the case $N = 3$ in 2000. De Giorgi's conjecture is still an open question for $4 \leq N \leq 8$.

Under the additional assumption that

$$\forall x^{'} \in \mathbf{R}^{N-1} \qquad \lim_{x_N \to \pm\infty} u(x^{'}, x_N) = \pm 1, \tag{3.3}$$

the conjecture was proved by O. Savin [64] in 2003 for $N \leq 8$ (see also [27–29] and [71] for more general results).

If one assumes that the limits in (3.3) are uniform with respect to $x' \in \mathbf{R}^{N-1}$ (but *without* any assumption on the monotonicity of u) then, for any $N \geq 2$, the solution is an increasing function which only depends on the variable x_N. This result was proved, independently and with different methods, by A. Farina [21,22] in 1999 (see also [24]), by H. Berestycki, F. Hamel and R. Monneau [6] in 2000 and by M. Barlow, R. Bass and C. Gui [5] in 2000.

For a complete account of the available results (old and new) concerning De Giorgi's conjecture we refer to the survey article [28].

In what follows we study De Giorgi's conjecture in dimension two, but in a more general context. To be more precise we study the extension/generalization (see Remarks 3.6) of the above classification problem in three different directions:

i) By *only* assuming $|\nabla u| > 0$ in $\mathbf{R}^2$ instead of the monotonicity hypothesis $\frac{\partial u}{\partial x_2} > 0$ in $\mathbf{R}^2$,

ii) By *only* assuming $|\nabla u| \in L^\infty(\mathbf{R}^2)$,

iii) By replacing the Allen-Cahn equation (3.2) by quasilinear equations of the form:

$$div\Big(a(|\nabla u|)\nabla u\Big) + f(u) = 0 \qquad \text{in} \quad \mathbf{R}^2,$$

where f is *any* locally Lipschitz-continuous function.

The function a will be assumed to satisfy the (minimal) structural assumptions:

$$a \in C^{1,1}_{loc}(0,+\infty), \qquad \lim_{t\to 0} ta(t) = 0, \qquad (ta(t))' > 0 \qquad \forall t > 0, \qquad (S_1)$$

$$\exists\, T' > 0, \quad \exists\, C = C(T') > 0 \quad : \quad (ta(t))' \leq Ct^{-2} \qquad \forall\, t \in (0, T']. \qquad (S_2)$$

Functions of the form:

$$a(t) = t^{\alpha}(\eta + t^2)^{\beta} \tag{3.4}$$

satisfy the above assumptions whenever $\eta > 0$, $\alpha > -1$ and $\beta \geq -(\alpha+1)/2$. For a suitable choice of the parameters in (3.4), we recover the m-Laplacian operator, with $1 < m < +\infty$, the mean curvature operator as well as some more general (singular or degenerate) operators satisfying non-standard growth conditions.

In the sequel we shall denote by $\arg(\nabla u)$ the argument of the gradient of u. This is a well-defined real-valued function since $|\nabla u| > 0$ in $\mathbf{R}^2$. With these notations we have:

Theorem 3.3 ([23]) *Let $f \in C^{0,1}_{loc}(\mathbf{R})$ and suppose that the structural assumptions $(S_1) - (S_2)$ are satisfied. Let $u \in C^1(\mathbf{R}^2)$ be a solution of*

$$\begin{cases} div(a(|\nabla u|)\nabla u) + f(u) = 0 & \text{in} \quad D'(\mathbf{R}^2), \\ |\nabla u| > 0 & \text{in} \quad \mathbf{R}^2, \\ \end{cases} \tag{3.5}$$

with $\nabla u \in L^{\infty}(\mathbf{R}^2)$. Assume that there exists $\delta < 1$ such that

$$|\arg(\nabla u)(x)| = O(\ln^{\delta}|x|), \qquad \text{as} \quad |x| \to +\infty. \tag{3.6}$$

Then u is one-dimensional and monotone, i.e. there exist $\nu \in \mathbf{S^1}$ and a function $g \in C^2(\mathbf{R})$ satisfying

$$u(x) = g(\nu \cdot x) \qquad \forall x \in \mathbf{R}^2, \qquad |g'(t)| > 0 \qquad \forall t \in \mathbf{R}.$$

If the solution u satisfies

$$\frac{\partial u}{\partial x_1} > 0 \quad \text{in} \quad \mathbf{R}^2,$$

then

$$\arg(\nabla u) = \arctan\left(\frac{u_2}{u_1}\right) \qquad \text{in} \quad \mathbf{R}^2,$$

where $u_j := \frac{\partial u}{\partial x_j}$, for $j = 1, 2$. Since in this case condition (3.6) is automatically satisfied, the above Theorem 3.3 immediately implies the following extension of De Giorgi's conjecture:

Corollary 3.4 ([23]) *Let $f \in C^{0,1}_{loc}(\mathbf{R})$ and suppose that the structural assumptions $(S_1) - (S_2)$ are satisfied. Let $u \in C^1(\mathbf{R}^2)$ be a solution of*

$$\begin{cases} div(a(|\nabla u|)\nabla u) + f(u) = 0 & \text{in} \quad D'(\mathbf{R}^2), \\ \dfrac{\partial u}{\partial x_1} > 0 & \text{in} \quad \mathbf{R}^2, \end{cases}$$

such that $\nabla u \in L^\infty(\mathbf{R}^2)$. Then, u is one-dimensional, i.e. there exist $\nu \in \mathbf{S^1}$ and a function $g \in C^2(\mathbf{R})$ such that

$$u(x) = g(\nu \cdot x) \qquad \forall x \in \mathbf{R}^2.$$

The proofs of the above results rely on the following Liouville-type theorem proved by D. Gilbarg and J. Serrin [37] in 1955 (see also [30] and [25]).

Theorem 3.5 ([37]) *Assume $0 < \delta < 1$ and let $B = (b_{hk})$ be a symmetric real matrix, whose entries are bounded locally Lipschitz-continuous functions defined on $\mathbf{R}^2$ and satisfying:*

$$\forall\, x \in \mathbf{R}^2, \qquad \forall \xi \in \mathbf{R}^2 \setminus \{0\}, \qquad \sum_{h,k=1}^{2} b_{hk}(x)\xi_h\xi_k > 0.$$

Then every distribution solution $u \in C^1(\mathbf{R}^2)$ of

$$\begin{cases} div(B(x)\nabla u) = 0 & in \quad \mathbf{R}^2, \\ u(x) = O(\ln^\delta(|x|)) & as \quad |x| \to +\infty, \end{cases} \tag{3.7}$$

is a constant function.

Proof of Theorem 3.3. We divide the proof in two steps.
Step 1 - We first prove that the solution u belongs to $C^2(\mathbf{R}^2)$ and satisfies

$$\begin{cases} div(\rho^2 A \nabla\theta) = 0 & \text{in} \quad \mathcal{D}'(\mathbf{R}^2), \\ div(A\nabla\rho) = \rho\,(-f'(u) + (A\nabla\theta)\nabla\theta) & \text{in} \quad \mathcal{D}'(\mathbf{R}^2), \end{cases} \tag{3.8}$$

where $\theta :=$ $\arg(\nabla u)$, $\rho := |\nabla u|$ and $A = (a_{hk})$ is the real symmetric matrix whose entries are $C^{0,1}_{loc}$ functions given by

$$a_{hk} := \frac{a'(|\nabla u|)}{|\nabla u|} u_h u_k + a(|\nabla u|)\delta_{hk}. \tag{3.9}$$

Since u has no critical points, the classical interior regularity results for quasilinear equations [70,48], immediately imply that any C^1 distribution-solution u of (3.5) is actually of class C^2. Therefore, the vector field $\frac{\nabla u}{|\nabla u|}$ is well-defined and belongs to $C^1(\mathbf{R}^2, \mathbf{S^1})$. Hence, there exists a $C^1(\mathbf{R}^2)$ function θ such that

$$\nabla u(x) = |\nabla u(x)|e^{i\theta(x)} := \rho(x)e^{i\theta(x)} \qquad \text{in} \qquad \mathbf{R}^2. \tag{3.10}$$

For $s = 1, 2$ we set $u_s := \frac{\partial u}{\partial x_s}$. Differentiating the equation in (3.5) yields:

$$div([a(|\nabla u|)\nabla u]_s) + f'(u)u_s = 0 \quad \text{in} \quad \mathcal{D}'(\mathbf{R}^2).$$

Since $a(|\nabla u|)\nabla u \in C^1$, a direct calculation gives

$$div(A(x)\nabla u_s) + f'(u)u_s = 0 \qquad \text{in} \qquad \mathcal{D}'(\mathbf{R}^2),$$

where A is the matrix whose entries are given by (3.9).

The complex function $z = u_1 + iu_2$ belongs to C^1 and satisfies the complex equation

$$div(A\nabla z) + f'(u)z = 0 \qquad \text{in} \qquad \mathcal{D}'(\mathbf{R}^2). \tag{3.11}$$

Inserting (3.10) in (3.11), we find:

$$\begin{aligned} -f'(u)\rho e^{i\theta} &= -f'(u)z = div(A\nabla z) = div(e^{i\theta}A\nabla\rho) + idiv(\rho e^{i\theta}A\nabla\theta)) \\ &= e^{i\theta}div(A\nabla\rho) + ie^{i\theta}(A\nabla\rho)\nabla\theta \\ &+ i\rho e^{i\theta}div(A\nabla\theta) + ie^{i\theta}(A\nabla\theta)\nabla\rho - e^{i\theta}\rho(A\nabla\theta)\nabla\theta \quad \text{in} \quad \mathcal{D}'(\mathbf{R}^2). \end{aligned}$$

Hence,

$$-f'(u)\rho = div(A\nabla\rho) - \rho(A\nabla\theta)\nabla\theta + 2i(A\nabla\rho)\nabla\theta + i\rho div(A\nabla\theta) \quad \text{in} \quad \mathcal{D}'(\mathbf{R}^2),$$

where we have used the symmetry of A.

Separating the imaginary and the real parts we obtain:

$$\begin{cases} \rho div(A\nabla\theta) + 2(A\nabla\rho)\nabla\theta = 0 & \text{in} \quad \mathcal{D}'(\mathbf{R}^2), \\ div(A\nabla\rho) - \rho(A\nabla\theta)\nabla\theta + \rho f'(u) = 0 & \text{in} \quad \mathcal{D}'(\mathbb{R}^2). \end{cases} \tag{3.12}$$

In particular, the second equation in (3.8) is established. To prove the first one we notice that

$$\begin{aligned} div(\rho^2 A\nabla\theta) &= \rho^2 div(A\nabla\theta) + 2\rho(A\nabla\theta)\nabla\rho \\ &= \rho\left(\rho div(A\nabla\theta) + 2(A\nabla\rho)\nabla\theta\right) \qquad \text{in} \quad \mathcal{D}'(\mathbf{R}^2). \end{aligned}$$

Thus, the claim follows from the first equation in (3.12).

Step 2 - End of the proof.

The assumption (S_2) yields the boundedness of the coefficients of the matrix $\rho^2 A$. An application of the Liouville-type Theorem 3.5 to the first equation in (3.8) implies that θ is a constant function. Thus $\nabla u(x) = |\nabla u(x)|e^{i\theta_0}$ in $\mathbf{R}^2$, for some $\theta_0 \in \mathbf{R}$. Setting $\tau = (-\sin(\theta_0), \cos(\theta_0))$ we have $\nabla u \cdot \tau = 0$ everywhere. This implies the desired conclusion with $\nu = (\cos(\theta_0), \sin(\theta_0))$. ■

We conclude this section with some remarks concerning the above results.

Remarks 3.6

i) We note that Theorem 3.3 and Corollary 3.4, besides the extension to locally Lipschitz-continuous nonlinearities, provide the following generalizations with respect to the question raised by De Giorgi: a) the assumption of monotonicity is weakened to $|\nabla u| > 0$ everywhere in $\mathbf{R}^2$, b) the solution u is not necessarily bounded. We only assume $\nabla u \in L^\infty(\mathbf{R}^2)$, which enable us to take into account some natural unbounded solutions. For instance, $u(x) = x_1$ is a one-dimensional monotone unbounded harmonic function, with bounded gradient, c) very general singular and degenerate operators can be taken into account.

ii) The assumption $\nabla u \in L^\infty(\mathbf{R}^2)$ is necessary in the above results. This is shown by the following example taken from [23]. The function $u(x,y) = x - \frac{y^2}{2}$ satisfies :

$$\begin{cases} \Delta u + 1 = 0 & \text{in} \quad \mathbf{R}^2, \\ \frac{\partial u}{\partial x} = 1 > 0 & \text{in} \quad \mathbf{R}^2, \\ |\nabla u| = \sqrt{1+y^2} \notin L^\infty(\mathbf{R}^2), \end{cases}$$

iii) For the case $N = 3$ we refer to [4,1] (see also [25]). Some more general results have been recently proved in [26–29] and [71].

iv) Theorem 3.5 does not extend to the higher dimensional case. To construct counter-examples, we first recall that

$$v(x) := \Big(\frac{\sqrt{N(N-2)}}{1+|x|^2} \Big)^{\frac{N-2}{2}}$$

is a positive solution of the Lane-Emden equation $-\Delta v = v^{\frac{N+2}{N-2}}$ in $\mathbf{R}^N$ for $N \geq 3$. Therefore the function $w(x) = e^{-v(x)}$ satisfies:

$$\Delta w = Vw, \qquad 0 < w < 1 \qquad \text{in} \quad \mathbf{R}^N, \tag{3.13}$$

where $V \in C^\infty(\mathbf{R}^N) \cap L^\infty(\mathbf{R}^N)$ is the positive potential

$$V = \Big[|\nabla v|^2 + v^{\frac{N+2}{N-2}} \Big] > 0 \qquad \text{in} \quad \mathbf{R}^N. \tag{3.14}$$

Next, we consider the function $u : \mathbf{R}^{N+1} \to \mathbf{R}$ defined by $u(x, x_{N+1}) = 2 + w(x)\cos(x_{N+1})$. A direct computation, using (3.13) – (3.14), shows that u is a positive, bounded and smooth solution of the linear *elliptic* equation in divergence form:

$$div(B(x)\nabla u) = 0 \qquad \text{in} \quad \mathbf{R}^{N+1},\ N \geq 3, \tag{3.15}$$

where $B := Diag(1, ..., 1, V)$ is a positive-definite diagonal matrix with bounded and smooth entries.

Thus, the function u provides a counter-example to Theorem 3.5 in any dimension *greater than or equal to 4* (note that $N \geq 3$ in the above construction).

v) The question of whether or not the conclusion of Theorem 3.5 holds true in dimension 3 remains open.

vi) For $N \geq 4$, the example given by equation (3.15) also shows that Theorem 2.8 does not hold true for second order operators in divergence form, if one only assumes *ellipticity.*

3 The quasi-linear case

In 1762, J.L. Lagrange [49] studied the problem of determining the graph of a smooth function $u = u(x, y)$ over a two-dimensional bounded open set Ω, having least area among all graphs that assume given values on the boundary of Ω.

Lagrange found that the function u must be a solution of the quasilinear partial differential equation:

$$- div\Big(\frac{\nabla u}{\sqrt{1 + |\nabla u|^2}}\Big) = 0 \qquad \text{in} \quad \Omega. \tag{4.1}$$

Because of the property of minimizing area, this equation has been called the *minimal surface equation.* On the other hand, from the viewpoint of geometry, the equation (4.1) says that the *mean curvature* of the graph of the function u (view as a non-parametric surface in $\mathbf{R}^3$) vanishes identically [60]. Owing to this fact it has become customary to call *minimal surface* any surface whose *mean curvature* is zero.

In 1915 S.N. Bernstein proved the following beautiful theorem, known as *Bernstein's Theorem*:

Theorem 4.1 ([7]) *Let $u \in C^2(\mathbf{R}^2)$ be a solution of the minimal surface equation* (4.1) *on the entire Euclidean plane $\mathbf{R}^2$. Then u is an affine function, i.e., the graph of u is a plane in $\mathbf{R}^3$.*

Bernstein's Theorem can be seen as a Liouville-type theorem although no assumptions are made on the growth of the solution u. However, there is a deeper relation between Bernstein's result and Liouville-type theorems. Indeed, S.N. Bernstein deduced its celebrated theorem from the following Liouville-type theorem for solutions of elliptic, not necessarily uniformly elliptic, equations in $\mathbf{R}^2$.

Theorem 4.2 ([7]) *Let $a, b, c : \mathbf{R}^2 \to \mathbf{R}$ be functions such that the matrix*

$$\begin{pmatrix} a(x,y) & b(x,y) \\ b(x,y) & c(x,y) \end{pmatrix} \qquad \textit{is positive definite for every } \ (x,y) \in \mathbb{R}^2. \tag{4.2}$$

Let $u \in C^2(\mathbf{R}^2)$ be a solution of:

$$\begin{cases} a(x,y)\dfrac{\partial^2 u}{\partial x^2}(x,y) + 2b(x,y)\dfrac{\partial^2 u}{\partial x \partial y}(x,y) + c(x,y)\dfrac{\partial^2 u}{\partial y^2}(x,y) = 0 & \textit{in} \quad \mathbf{R}^2, \\ u(x,y) = o(\sqrt{x^2+y^2}) & \textit{as} \quad \sqrt{x^2+y^2} \to +\infty. \end{cases} \tag{4.3}$$

Then u is a constant function.

Note that no continuity or boundedness assumptions are imposed to the functions a, b and c.

Proof of Theorem 4.2. From the ellipticity condition (4.2) and the equation satisfied by u we have:

$$0 \le a\Big(\frac{\partial^2 u}{\partial x^2}\Big)^2 + 2b\frac{\partial^2 u}{\partial x \partial y}\frac{\partial^2 u}{\partial x^2} + c\Big(\frac{\partial^2 u}{\partial x \partial y}\Big)^2 = -c\Big[\frac{\partial^2 u}{\partial x^2}\frac{\partial^2 u}{\partial y^2} - \Big(\frac{\partial^2 u}{\partial x \partial y}\Big)^2\Big],$$

$$0 \le a\Big(\frac{\partial^2 u}{\partial x \partial y}\Big)^2 + 2b\frac{\partial^2 u}{\partial x \partial y}\frac{\partial^2 u}{\partial y^2} + c\Big(\frac{\partial^2 u}{\partial y^2}\Big)^2 = -a\Big[\frac{\partial^2 u}{\partial x^2}\frac{\partial^2 u}{\partial y^2} - \Big(\frac{\partial^2 u}{\partial x \partial y}\Big)^2\Big].$$

Hence $\frac{\partial^2 u}{\partial x^2}\frac{\partial^2 u}{\partial y^2} - \Big(\frac{\partial^2 u}{\partial x \partial y}\Big)^2 \le 0$ everywhere in $\mathbf{R}^2$ and the equality holds only at points where $\frac{\partial^2 u}{\partial x^2} = \frac{\partial^2 u}{\partial y^2} = \frac{\partial^2 u}{\partial x \partial y} = 0$ (since the equation is elliptic). To conclude the proof, we invoke:

Theorem 4.3 ([7,44]) *Let $u \in C^2(\mathbf{R}^2)$ be a solution of:*

$$\frac{\partial^2 u}{\partial x^2}\frac{\partial^2 u}{\partial y^2} - \Big(\frac{\partial^2 u}{\partial x \partial y}\Big)^2 \le 0, \qquad \frac{\partial^2 u}{\partial x^2}\frac{\partial^2 u}{\partial y^2} - \Big(\frac{\partial^2 u}{\partial x \partial y}\Big)^2 \not\equiv 0 \qquad \text{in} \quad \mathbf{R}^2.$$

Then, $u = u(x,y)$ cannot be $o(\sqrt{x^2+y^2})$ as $\sqrt{x^2+y^2} \to +\infty$.

An application of Theorem 4.3 gives

$$\frac{\partial^2 u}{\partial x^2} = \frac{\partial^2 u}{\partial y^2} = \frac{\partial^2 u}{\partial x \partial y} = \frac{\partial^2 u}{\partial y \partial x} = 0$$

everywhere in $\mathbf{R}^2$. Thus, u is an affine function and the desired conclusion follows immediately, in view of the sub-linear growth of u. ∎

Now we are ready to prove Theorem 4.1.

Proof of Theorem 4.1. Any solution of the minimal surface equation is smooth (actually real-analytic) see for instance [40,58,61]. Then, a direct calculation shows that the smooth functions,

$$\psi_1 = \arctan\Big(\frac{\partial u}{\partial x}\Big), \qquad \psi_2 = \arctan\Big(\frac{\partial u}{\partial y}\Big)$$

are bounded solutions of the equation:

$$\Big(1+\Big(\frac{\partial u}{\partial y}\Big)^2\Big)\frac{\partial^2 v}{\partial x^2} - 2\frac{\partial u}{\partial x}\frac{\partial u}{\partial y}\frac{\partial^2 v}{\partial x\partial y} + \Big(1+\Big(\frac{\partial u}{\partial x}\Big)^2\Big)\frac{\partial^2 v}{\partial y^2} = 0 \qquad \text{in} \quad \mathbf{R}^2.$$

An application of Theorem 4.2, with $a(x,y) = \Big(1+\Big(\frac{\partial u}{\partial y}\Big)^2\Big)$, $b(x,y) = -\frac{\partial u}{\partial x}\frac{\partial u}{\partial y}$ and $c(x,y) = \Big(1+\Big(\frac{\partial u}{\partial x}\Big)^2\Big)$, then implies $\nabla u = Const.$

Therefore, u is an affine function. ■

Before going further we wish to make some remarks on the previous results.

Remarks 4.4

i) Bernstein's original proof of Theorem 4.3 contained a gap. This gap was discovered and fixed by E. Hopf in 1950 [44]. Almost at the same time, another proof of Theorem 4.3 was provided by E.J. Mickle [56]. We refer to [44] and [56] for a proof of Theorem 4.3.

ii) Both Theorem 4.2 and Theorem 4.3 are sharp. Indeed, the linear elliptic equation in (4.3) always admits the linear solution $u(x,y) = y$. On the other hand, the function $u(x,y) = x\tanh(y)$ satisfies $\frac{\partial^2 u}{\partial x^2}\frac{\partial^2 u}{\partial y^2} - \Big(\frac{\partial^2 u}{\partial x\partial y}\Big)^2 < 0$ everywhere in $\mathbf{R}^2$ and has linear growth at infinity.

iii) Theorem 4.2 does not extend to $\mathbf{R}^N$, for $N \geq 3$. A counterexample of this fact was built by E. Hopf [43] in 1929. See also [25].

Bernstein's Theorem stimulated a lot of research concerning the study of the higher dimensional minimal graph equation, i.e. the study of C^2 solutions of the equation:

$$-\,div\Big(\frac{\nabla u}{\sqrt{1+|\nabla u|^2}}\Big) = 0 \qquad \text{in} \quad \mathbf{R}^N, \quad N \geq 3. \tag{4.4}$$

One of the main open problems was whether Bernstein's Theorem generalizes to dimension $N \geq 3$, i.e., whether every solution of (4.4) is an affine function. E. De Giorgi proved Bernstein's Theorem for $N = 3$ in 1965 [17], F.J. Almgren proved it for $N = 4$ in 1966 [3], while the case $N \leq 7$ was established by J. Simons in 1968 [69]. Their proofs are not based on Liouville-type theorems but on the connection between minimal graphs defined over $\mathbf{R}^N$ and minimal hypercones in $\mathbf{R}^N$ (see for instance [40,68]). In 1969, E. Bombieri, E. De Giorgi and E. Giusti [8], settled Bernstein's problem proving the existence of a non-affine solution of the minimal surface equation (4.4) for any $N \geq 8$.

For more results about non-affine solutions of (4.4) we refer to the survey of L. Simon [68] (and the references therein). For further results about minimal surfaces we refer the interested reader to the beautiful survey of R. Osserman [60].

The rest of this section is devoted to the classification of solutions to quasilinear elliptic equations involving the mean curvature operator. To this end the next result, proved by J. Moser [57] in 1961, plays a crucial role.

Theorem 4.5 ([57]) *Assume that $N \geq 2$ and let $u \in C^2(\mathbf{R}^N)$ be a solution of:*

$$\begin{cases} -div\Big(\frac{\nabla u}{\sqrt{1+|\nabla u|^2}}\Big) = 0 \qquad in \quad \mathbf{R}^N, \\ \\ \nabla u \in L^\infty(\mathbf{R}^N). \end{cases} \tag{4.5}$$

Then u is an affine function, i.e. the graph of u is a hyperplane in $\mathbf{R}^{N+1}$.

Proof. As we have already said, any solution of the minimal surface equation is smooth; see for instance [40,58,61]. Hence, by differentiating the equation (4.5) with respect to x_j, we find that

$$-div(A(x)\nabla u_j) = 0 \qquad \text{in} \qquad \mathbf{R}^N, \tag{4.6}$$

where u_j denotes the function $\frac{\partial u}{\partial x_j}$, for any $j = 1, ..., N$ and $A = (a_{hk})$ is the real symmetric matrix whose entries are given by:

$$a_{hk} = a_{hk}(x) := \frac{\delta_{hk}}{(1+|\nabla u|^2)^{\frac{1}{2}}} - \frac{u_h u_k}{(1+|\nabla u|^2)^{\frac{3}{2}}}.$$

The smallest eigenvalue of the matrix A is $\frac{1}{(1+|\nabla u|^2)^{\frac{3}{2}}}$ while the largest one is $\frac{1}{(1+|\nabla u|^2)^{\frac{1}{2}}}$. The assumption $\nabla u \in L^\infty(\mathbf{R}^N)$ implies that the linear equation (4.6) is uniformly elliptic on the entire Euclidean space $\mathbf{R}^N$ with $\lambda = \frac{1}{(1+\|\nabla u\|^2_{L^\infty})^{\frac{3}{2}}}$ and $\Lambda = 1$.

Using once again the assumption $\nabla u \in L^\infty(\mathbf{R}^N)$ we infer, from Theorem 2.8, that every u_j is a constant function. This concludes the proof. ∎

Theorem 4.5 says that a non-affine minimal graph must have unbounded gradient. This suggests that more informations about the growth of the gradient of the solutions of (4.4) could be helpful to obtain further results of Bernstein type. The following deep result provides a universal interior gradient estimate for solutions to the minimal surfaces equation. It was proved by E. Bombieri, E. De Giorgi and M. Miranda [9] in 1969 (for the two-dimensional case this kind of estimate was first proved by R. Finn [31] in 1954).

Theorem 4.6 ([9]) *Let $N \geq 2$, $x_0 \in \mathbf{R}^N$ and $r > 0$. Let $u \in C^2(B(x_0, r))$ be a solution of:*

$$\begin{cases} -div\Big(\frac{\nabla u}{\sqrt{1+|\nabla u|^2}}\Big) = 0 & in \quad B(x_0, r), \\ u \geq 0 & in \quad B(x_0, r). \end{cases}$$

Then, we have the estimate:

$$|\nabla u(x_0)| \leq C_1\, e^{\Big[C_2 \frac{u(x_0)}{r}\Big]},$$

where C_1 and C_2 are constants depending only on N.

For the proof of the above theorem we refer the reader to the original work of E. Bombieri, E. De Giorgi and M. Miranda [9] (See also [45] and chapter 16 of [38]).

A combination of the above two results leads to the following extensions of the classical Bernstein's Theorem.

Theorem 4.7 ([9]) *Assume that $N \geq 2$ and let $u \in C^2(\mathbf{R}^N)$ be a solution of:*

$$-div\Big(\frac{\nabla u}{\sqrt{1+|\nabla u|^2}}\Big) = 0 \qquad in \qquad \mathbf{R}^N.$$

Then:

i) If u is bounded from below (above) then u is constant.

ii) If u satisfies,

$$u(x) \geq -K(|x|+1) \qquad \forall\, x \in \mathbf{R}^N, \tag{4.7}$$

for some positive constant K, then u is an affine function.

Proof. It is enough to prove the case ii). To this end we observe that, under the assumption (4.7), for every $x_0 \in \mathbf{R}^N$ and every $r > 0$ we have:

$$u(x_0) - \inf_{B(x_0,r)} u \leq u(x_0) + K[|x_0| + r + 1].$$

Therefore, an application of Theorem 4.6 to the function $u(x) - \inf_{B(x_0,r)} u$, on the ball $B(x_0, r)$ yields:

$$\forall\, x_0 \in \mathbf{R}^N, \quad \forall r > 0, \quad |\nabla u(x_0)| \leq C_1 \exp\Big[C_2 \frac{u(x_0) + K|x_0| + K}{r} + C_2 K\Big].$$

The latter implies $\|\nabla u\|_{L^\infty} \leq C_1\, e^{C_2 K} < +\infty$, and the claim follows from Theorem 4.5. ■

Next we turn to the study of the equation:

$$-div\Big(\frac{\nabla u}{\sqrt{1+|\nabla u|^2}}\Big)+f(u)=0 \qquad \text{in} \qquad \mathbf{R}^N, \tag{4.8}$$

where f is a continuous non-decreasing function. This equation includes some models which are important for applications. For instance, when $f \equiv H = const.$, the above equation (4.8) describes entire graphs (non-parametric hypersurfaces in $\mathbf{R}^{N+1}$) with constant mean curvature equal to H (see [14]) while, for $f(t)=at+b$, $a>0$, the equation (4.8) gives the well-known *capillary surface equation* [32–34]. More generally, the equation (4.8) says that the mean curvature of the graph of u (viewed as a non-parametric hypersurface in $\mathbf{R}^{N+1}$) is equal to $f(u(x))/N$ at every point $(x,u(x))$.

For this class of equations we have the following result:

Theorem 4.8 ([25]) *Assume that $N \geq 1$ and let $f:\mathbf{R}\to\mathbf{R}$ be a continuous, non-decreasing function, with $f \not\equiv 0$. Let $u \in C^2(\mathbf{R}^N)$ be a solution of* (4.8).

Then, u is a constant function, say $u \equiv c \in \mathbf{R}$, and $f(c)=0$.

Proof. Let $\{\rho_n\}_{n\in\mathbf{N}^\star}$ be a sequence of standard *mollifiers* and set $f_n = f * \rho_n$. Then, for every $n \in \mathbf{N}^\star$, f_n is a smooth and non-decreasing function on $\mathbf{R}$.

For every $\phi \in C_c^\infty(\mathbf{R}^N)$, $\phi \geq 0$ on $\mathbf{R}^N$, and every $n \in \mathbf{N}^\star$ we multiply the equation (4.8) by $|f_n(u)|^{N-1}f_n(u)\phi^{N+1}$ and we integrate by parts. This leads to:

$$\int f(u)|f_n(u)|^{N-1}f_n(u)\phi^{N+1} = -(N+1)\int \phi^N|f_n(u)|^{N-1}f_n(u)\frac{(\nabla u\cdot\nabla\phi)}{\sqrt{1+|\nabla u|^2}}$$

$$-N\int |f_n(u)|^{N-1}f_n'(u)\phi^{N+1}\frac{|\nabla u|^2}{\sqrt{1+|\nabla u|^2}} \leq (N+1)\int \phi^N|f_n(u)|^N|\nabla\phi|,$$

where we have used $f_n' \geq 0$. By letting $n \to +\infty$ we have:

$$\int |f(u)|^{N+1}\phi^{N+1} \leq (N+1)\int |f(u)|^N\phi^N|\nabla\phi|.$$

By the Schwarz-Hölder inequality we see that

$$\int |f(u)|^{N+1}\phi^{N+1} \leq (N+1)\Big[\int |f(u)|^{N+1}\phi^{N+1}\Big]^{\frac{N}{N+1}}\Big[\int |\nabla\phi|^{N+1}\Big]^{\frac{1}{N+1}}$$

and consequently

$$\int |f(u)|^{N+1}\phi^{N+1} \leq (N+1)^{N+1}\int |\nabla\phi|^{N+1}. \tag{4.9}$$

Now, for every $R>0$, we consider the function $\phi = \phi_R(x) = \varphi(|x|/R)$, where $\varphi \in C_c^\infty(\mathbf{R})$, $0 \leq \varphi \leq 1$ everywhere on $\mathbf{R}$, and

$$\varphi(t) = \begin{cases} 1 & \text{if} \quad |t| \le 1, \\ 0 & \text{if} \quad |t| \ge 2. \end{cases}$$

Inserting $\phi = \phi_R$ in (4.9) we obtain:

$$\int_{B(0,R)} |f(u)|^{N+1} \le C(N) \int_{B(0,2R)} \frac{1}{R^{N+1}} \to 0 \qquad \text{as} \qquad R \to +\infty.$$

which implies

$$\forall x \in \mathbf{R}^N \qquad f(u(x)) = 0.$$

Therefore, u must be bounded on at least one side (otherwise we would have $f \equiv 0$, a contradiction) and the conclusion follows from Theorem 4.7. ▪

Remarks 4.9

i) When $N \le 7$, one can drop the assumption $f \not\equiv 0$ in Theorem 4.8. In this case, the conclusion is the following: *u is an affine function* (as in Bernstein's Theorem). Indeed, the proof of the theorem yields $f(u) \equiv 0$ on $\mathbf{R}^N$. Consequently u is a minimal graph and hence an affine function, since $N \le 7$.

A similar result is not available for $N \ge 8$ since, in this case, non-affine entire minimal graphs do exist [8].

Theorem 4.8 improves an earlier result of S.Y. Cheng and S.T. Yau [13], established for $N = 2$ and for a non-negative and non-decreasing function f. Their method (of geometric nature) was entirely different from the one given here. It was based on the fact that, under the above assumptions on f, the graph of any solution of the considered equation is a parabolic Riemannian manifold in $\mathbf{R}^3$.

ii) The above result also says that an entire graph with constant mean curvature must be an entire minimal graph. This was first proved (by different methods) by S.S. Chern [14] in 1965.

iii) Note that the monotonicity assumption on f is necessary to obtain the desired conclusion in the above Theorem. Indeed, the non-affine function $u = \sqrt{1 + x_1^2}$ satisfies the equation (4.8) with $N \ge 1$ and

$$f(t) := \begin{cases} 1, & t \le 1, \\ \frac{1}{(2t^2-1)^{\frac{3}{2}}}, & t > 1. \end{cases}$$

It is easily checked that f is not monotonically increasing.

References

[1] Alberti G.; Ambrosio L.; Cabré X., *On a long-standing conjecture of E. De Giorgi: symmetry in 3D for general nonlinearities and a local*

minimality property. Special issue dedicated to Antonio Avantaggiati on the occasion of his 70th birthday. Acta Appl. Math. 65 (2001), no. 1–3, 9–33.

[2] Allen S.; Cahn J., *A microscopic theory for antiphase boundary motion and its application to antiphase domain coarsening.* Acta Metallurgica, 27, 1084–1095, 1979.

[3] Almgren F. J., Jr., *Some interior regularity theorems minimal surfaces and an extension of Bernstein's theorem.* Ann. Math. (2) 84, 1966, 277–292.

[4] Ambrosio L.; Cabré X., *Entire solutions of semilinear elliptic equations in $\mathbf{R}^3$ and a conjecture of De Giorgi.* J. Amer. Math. Soc. 13 (2000), no. 4, 725–739

[5] Barlow M.T.; Bass R.F.; Gui C., *The Liouville property and a conjecture of De Giorgi.* Comm. Pure Appl. Math. 53 (2000), no. 8, 1007–1038.

[6] Berestycki H.; Hamel F.; Monneau R., *One-dimensional symmetry of bounded entire solutions of some elliptic equations.* Duke Math. J. 103 (2000), no. 3, 375–396.

[7] Bernstein S., *Sur un théorème de géométrie et son application aux équations aux dérivées partielles du type elliptique.* Comm. Soc. Math. de Kharkov 2, 15, 38-45 (1915-1917); see also: Math. Z. 26 (1927), no. 1, 551–558.

[8] Bombieri E.; De Giorgi E.; Giusti E., *Minimal cones and the Bernstein problem.* Invent. Math. 7 1969 243–268.

[9] Bombieri E.; De Giorgi E.; Miranda M., *Una maggiorazione a priori relativa alle ipersuperfici minimali non parametriche.* Arch. Rational Mech. Anal. 32, 1969, 255–267.

[10] Carbou G., *Unicité et minimalité des solutions d'une équation de Ginzburg-Landau.* Annales Inst. H. Poincaré, Analyse Non linéaire, 12 (3), 305-318, 1995.

[11] Cauchy A., *Mémoires sur les fonctions complémentaires.* C. R. Acad. Sci. Paris, 19 (1844), 1377-1384; see also: Oeuvres complètes, Ier série, tome VIII, 378-383.

[12] Cauchy A.: C. R. Acad. Sci. Paris, 32 (1851), 452-454 (and Oeuvres complètes, Ier série, tome XI, 373-376).

[13] Cheng S.Y.; Yau S.T., *Differential equations on Riemannian manifolds and their geometric applications.* Comm. Pure Appl. Math. 28 (1975), no. 3, 333–354.

[14] Chern, S.S., *On the curvatures of a piece of hypersurface in Euclidean space.* Abh. Math. Sem. Univ. Hamburg 29, 1965, 77–91.

[15] Colding, T.H.; Minicozzi, W.P., II, *Harmonic functions on manifolds.* Ann. Math. (2) 146 (1997), no. 3, 725–747.

[16] De Giorgi E., *Sulla differenziabilità e l'analiticità delle estremali degli integrali multipli regolari.* Mem. Accad. Sci. Torino. Cl. Sci. Fis. Mat. Nat. (3) 3, 1957, 25–43.

[17] De Giorgi E., *Una estensione del teorema di Bernstein.* Ann. Scuola Norm. Sup. Pisa (3) 19, 1965, 79–85.

[18] De Giorgi E., *Convergence problems for functionals and operators.* Proceedings of the International Meeting on Recent Methods in Nonlinear Analysis (Rome, 1978), pp. 131–188, Pitagora, Bologna, 1979.

[19] Diaz J.I., *Nonlinear Partial Differential Equations and Free Boundaries.* Pitman Research Notes in Mathematics, 106, 1985.

[20] Evans Lawrence C., *Partial differential equations.* Graduate Studies in Mathematics, 19. American Mathematical Society, Providence, RI, 1998.

[21] Farina A., *Symmetry for solutions of semilinear elliptic equations in* $\mathbf{R}^N$ *and related conjectures.* Papers in memory of Ennio De Giorgi. Ricerche Mat. 48 (1999), suppl., 129–154.

[22] Farina A., *Symmetry for solutions of semilinear elliptic equations in* $\mathbf{R}^N$ *and related conjectures.* Atti Accad. Naz. Lincei Cl. Sci. Fis. Mat. Natur. Rend. Lincei (9) Mat. Appl. 10 (1999), no. 4, 255–265.

[23] Farina A., *One-dimensional symmetry for solutions of quasilinear equations in* $\mathbf{R}^2$. Boll. Unione Mat. Ital. Sez. B Artic. Ric. Mat. (8) 6 (2003), no. 3, 685–692.

[24] Farina A., *Rigidity and one-dimensional symmetry for semilinear elliptic equations in the whole of* $\mathbf{R}^N$ *and in half spaces.* Adv. Math. Sci. Appl. 13 (2003), no. 1, 65–82.

[25] Farina A., *Liouville-type theorems for elliptic problems.* In Handbook of Differential Equations - Stationary Partial Differential Equations, Ed. M. Chipot, Elsevier BV, 4 (2007), 60–116.

[26] Farina A.; Sciunzi B.; Valdinoci E., *Bernstein and De Giorgi type problems: new results via a geometric approach.* Ann. Scuola Norm. Sup. Pisa Cl. Sci. (to appear) 2008.

[27] Farina A.; Valdinoci E., *Geometry of quasiminimal phase transitions.* Calc. Var. Partial Differential Equations, 33 (1), 1–35, 2008.

[28] Farina A.; Valdinoci E., *The state of the art for a conjecture of De Giorgi and related problems.* To appear in " Reaction-Diffusion Systems and Viscosity Solutions " World Scientific, 2008.

[29] Farina A.; Valdinoci E., *1D symmetry for solutions of semilinear and quasilinear elliptic equations.* Preprint 2008. http://www.math.utexas.edu/mp_arc.

[30] Finn R., *Sur quelques généralisations du théorème de Picard.* C. R. Acad. Sci. Paris 235, (1952), 596–598.

[31] Finn R., *On equations of minimal surface type.* Ann. Math. (2) 60, (1954), 397–416.

[32] Finn R., *New estimates for equations of minimal surface type.* Arch. Rational Mech. Anal. 14, 1963, 337–375.

[33] Finn R., *Remarks relevant to minimal surfaces, and to surfaces of prescribed mean curvature.* J. Analyse Math. 14, 1965, 139–160.

[34] Finn R., *On the Laplace formula and the meniscus height for a capillary surface.* With addenda by the author. Z. Angew. Math. Mech. 61 (1981), no. 3, 165–173, 175–177.

[35] Ghoussoub N.; Gui C., *On a conjecture of De Giorgi and some related problems.* Math. Ann. 311 (1998), no. 3, 481–491.

[36] Gibbons G.W.; Townsend P.K., *Bogomolnyi equation for intesecting domain walls.* Phys. Rev. Lett., 83 (9), 1727–1730, Aug. 1999.

[37] Gilbarg D.; Serrin J., *On isolated singularities of solutions of second order elliptic differential equations.* J. Analyse Math. 4 (1955/56), 309–340.

[38] Gilbarg D.; Trudinger N.S., *Elliptic partial differential equations of second order.* Reprint of the 1998 edition. Classics in Mathematics. Springer-Verlag, Berlin, 2001.

[39] Ginzburg V.L.; Pitaevskii L.P., *On the theory of superfluidity.* Soviet Physics, JETP, 34 (7) 858–861 (1240–1245 Z. Eksper. Teoret. Fiz.), 1958.

[40] Giusti E., *Minimal surfaces and functions of bounded variation.* Monographs in Mathematics, 80. Birkhäuser Verlag, Basel, 1984.

[41] Harnack A., *Die Grundlagen der Theorie des logarithmischen Potentiales und der eindeutigen Potentialfunktion in der Ebene.* V.G.Teubner, Leipzig, 1887.

[42] Hilbert D., *Mathematical problems.* Bull. Amer. Math. Soc. 37 (2000), 407–436. Reprinted from Bull. Amer. Math. Soc. 8 (July 1902), 437–479. Originally published as *Mathematische Probleme. Vortrag, gehalten auf dem internationalen Mathematike-Congress zu Paris 1900*, Gött. Nachr. 1900, 253–297, Vandenhoeck and Ruprecht, Göttingen. Translated for the *Bulletin*, with the author's permission, by Dr. Mary Winston Newson, 1902.

[43] Hopf E., *Bemerkungen zu einem Satze von S. Bernstein aus der Theorie der elliptischen Differentialgleichungen.* Math. Z. 29 (1929), no. 1, 744–745.

[44] Hopf E., *On S. Bernstein's theorem on surfaces $z(x, y)$ of nonpositive curvature.* Proc. Amer. Math. Soc. 1, (1950). 80–85.

[45] Korevaar N., *An easy proof of the interior gradient bound for solutions to the prescribed mean curvature equation.* Nonlinear functional analysis and its applications, Part 2 (Berkeley, Calif., 1983), 81–89, Proc. Sympos. Pure Math., 45, Part 2, Amer. Math. Soc., Providence, RI, 1986.

[46] Krylov N. V., *Nonlinear elliptic and parabolic equations of the second order.* Translated from the Russian by P. L. Buzytsky [P. L. Buzytskiĭ]. Mathematics and its Applications (Soviet Series), 7. D. Reidel Publishing Co., Dordrecht, 1987.

[47] Krylov N. V.; Safonov M. V., *A property of the solutions of parabolic equations with measurable coefficients.* Izv. Akad. Nauk SSSR Ser. Mat. 44 (1980), no. 1, 161–175, 239.

[48] Ladyzhenskaya O.A.; Uraltseva N.N., *Linear and quasilinear elliptic equations.* Translated from the Russian by Scripta Technica, Inc. Translation editor: Leon Ehrenpreis Academic Press, New York-London, 1968.

[49] Lagrange J.L., *Essai d'une nouvelle méthode pour déterminer les maxima et les minima des formules intégrales indéfinies.* Miscellanea Taurinensia, t. II, Torino, 1762.

[50] Landau L.D., *Collected papers of L.D. Landau, Edited and with an introduction by D. ter Haar.* Second printing. Gordon and Breach Science Publisher, New York, 1967.

[51] Li P., *Harmonic sections of polynomial growth.* Math. Res. Lett. 4 (1997), no. 1, 35–44.

[52] Li P., *Curvature and function theory on Riemannian manifolds.* Surveys in differential geometry, 375–432, Surv. Differ. Geom., VII, Int. Press, Somerville, MA, 2000.

[53] Liouville J., . C. R. Acad. Sci. Paris, 19 (1844), 1262.

[54] Liouville J., . J. Math. Pures et Appl., 20, (1855) 201–208.

[55] Meyers N.G., *An L^p-estimate for the gradient of solutions of second order elliptic divergence equations.* Ann. Scuola Norm. Sup. Pisa (3) 17, 1963, 189–206.

[56] Mickle E.J., *A remark on a theorem of Serge Bernstein.* Proc. Amer. Math. Soc. 1, (1950), 86–89.

[57] Moser J., *On Harnack's theorem for elliptic differential equations.* Comm. Pure Appl. Math. 14, 1961, 577–591.

[58] Müntz Ch. H., *Die Lösung des Plateauschen Problems über konvexen Bereichen.* Math. Ann. 94 (1925), no. 1, 53–96.

[59] Nelson E., *A proof of Liouville's theorem.* Proc. Amer. Math. Soc. 12, 1961, 995.

[60] Osserman R., *A survey of minimal surfaces.* Second edition. Dover Publications, Inc., New York, 1986.

[61] Rado T., *Über den analytischen Charakter der Minimalflächen.* Math. Z. 24 (1926), no. 1, 321–327.

[62] Rowlinson J.S., *Translation of J.D. van der Waals "The thermodynamic theory of capillarity under the hypothesis of a continuous variation of density".* J. Statist. Phys. 20 (2), 197–244, 1979

[63] Saloff-Coste L., *Aspects of Sobolev-type inequalities.* London Mathematical Society Lecture Note Series, 289. Cambridge University Press, Cambridge, 2002.

[64] Savin V.O., *Phase Transitions: Regularity of Flat Level Sets.* Ph. D Thesis at the University of Texas at Austin, 2003.

[65] Serrin J.*On the Harnack inequality for linear elliptic equations.* J. Analyse Math. 4 (1955/56), 292–308.

[66] Serrin J., *Entire solutions of nonlinear Poisson equations.* Proc. London. Math. Soc. (3) 24, 1972, 348–366.

[67] Serrin J.; Weinberger H.F., *Isolated singularities of solutions of linear elliptic equations.* Amer. J. Math. 88, 1966, 258–272.

[68] Simon L., *The minimal surface equation.* Geometry, V, 239–272, Encyclopaedia Math. Sci., 90, Springer, Berlin, 1997.

[69] Simons J., *Minimal varieties in riemannian manifolds.* Ann. Math. (2) 88, 1968, 62–105.

[70] Tolksdorf P., *Regularity for a more general class of quasilinear elliptic equations.* J. Differential Equations 51 (1984), no. 1, 126–150.

[71] Valdinoci E.; Sciunzi B.; Savin V.O., *Flat level set regularity of p-Laplace phase transitions.* Mem. Amer. Math. Soc. 182 (2006), no. 858.

Bol. Soc. Esp. Mat. Apl.
n^{o}45(2008), 35–68

SOBRE LA INDIVIDUALIZACIÓN DE TRATAMIENTOS EN ONCOLOGíA RADIOTERÁPICA

MARíA ISABEL NÚÑEZ[1,2], ESCARLATA LÓPEZ[1,3], BEATRIZ PINAR[4,5], DAMIÁN GUIRADO[6], ROSARIO GUERRERO[1,3], ROSARIO DEL MORAL[1,3], MERCEDES VILLALOBOS[1,2], PEDRO C. LARA[4,5], JOSÉ MARIANO RUIZ DE ALMODÓVAR[1,6,7]

[1]Instituto de Biopatología y Medicina Regenerativa, Centro de Investigación Biomédica, Universidad de Granada

[2]Departamento de Radiología y Medicina Física, Universidad de Granada

[3]Servicio de Radioterapia, Hospital Universitario Virgen de las Nieves, Granada

[4]Servicio de Oncología Radioterápica, Hospital Dr. Negrín, Gran Canaria

[5]Instituto Canario de Investigación del Cáncer, Gran Canaria

[6]Hospital Universitario San Cecilio, Granada

[7]J.M. Ruiz de Almodóvar. Centro de Investigación Biomédica. Parque Tecnológico de Ciencias de la Salud. 18100, Armilla, Granada. España

`jmrdar@ugr.es` (J.M. Ruiz de Almodóvar)

Resumen

Lo que más preocupa a los profesionales de la medicina son los resultados de los procedimientos diagnósticos y terapéuticos que empleamos y proponemos. Nos gustaría que el diagnóstico fuese perfecto; esto es: que el método utilizado para diagnosticar la enfermedad fuese absolutamente fiable, sensible y específico y que la terapia fuese segura y eficaz en todos los casos. Obviamente esto no es así y entonces hablamos de acierto y de error, del cociente riesgo/beneficio, de la individualización terapéutica y de la predicción de la respuesta. Y sabemos que en medicina, acierto, lo probable, y error, lo inesperado, son resultados que difieren en su frecuencia más que en su esencia. Lo anterior nos obliga al estudio de las probabilidades de uno y otro signo y a la investigación conducente al desarrollo de métodos que nos indiquen ciertas regularidades que podamos pensar como leyes de nuestra ciencia y que, utilizadas mediante su expresión matemática, nos permitan predecir y así faciliten nuestras decisiones diagnósticas, o terapéuticas, y nos den seguridad científica. En este trabajo vamos a hablar de cáncer. Más precisamente, describiremos nuestras aproximaciones metodológicas, y matemáticas, al problema de la predicción de la respuesta de los pacientes al tratamiento con radioterapia.

Palabras clave: *modelos matemáticos en oncología, individualización de tratamientos, radioterapia*

Clasificación por materias AMS: *92C50 92B05 62P10 90C90*

1 Introducción

Es claro que cada paciente es diferente; por ello su tratamiento debería también ser diferente; esto es, cada paciente necesita un tratamiento específicamente ajustado a sus características y al pronóstico de la enfermedad que le aqueja. Pronóstico en cáncer significa opinión fundada sobre la probabilidad de curación, o de supervivencia con enfermedad, en cada paciente concreto; también mediante el pronóstico se busca una aproximación a la probabilidad de las complicaciones que el paciente puede tener como consecuencia del tratamiento instaurado.

Para la mayor parte de las enfermedades neoplásicas,[1] el pronóstico de la enfermedad es función de un reducido número de variables. Y aunque la elección de estas variables está apoyada en un amplio consenso médico y se supone que a cada paciente se le aplica el tratamiento que se considera más adecuado para conseguir la curación del mismo, el número de fracasos terapéuticos que resultan constituye un problema médico de singular importancia. Estas variables (dicotómicas, discretas o continuas) toman valores que se obtienen de la exploración clínica, del estudio de extensión de la enfermedad y de los análisis que se realizan sobre muestras biológicas procedentes del paciente, incluyendo el espécimen tumoral.

En la actualidad el tratamiento de los enfermos oncológicos se decide en base al tamaño del tumor, el estado de los linfáticos locorregionales, la presencia o ausencia de enfermedad distal, el tipo histológico y el estado de bienestar del paciente. Conocidos los valores necesarios se procede a clasificar (estadificar) los pacientes en categorías clínicas bien definidas. Esta clasificación por estadios [1] sirve para que el médico disponga de una aproximación general al pronóstico de la enfermedad que padece la persona a la que está atendiendo, para que proponga el tratamiento que considere más adecuado y sobre todo para ofrecer al paciente la información necesaria para decidir, y consentir, cómo quiere ser tratado.

Sin embargo, los datos que cuantifican el pronóstico, de gran importancia para médicos y pacientes, son imprecisos e insuficientes. En efecto, cuando se estudian, a largo plazo, los resultados de la terapéutica en grupos de enfermos clasificados en el mismo estadio, se encuentra una variabilidad de la respuesta que es imposible de predecir. Esto es, dentro de cada estadio y para cada paciente concreto, desconocemos cuál será el resultado terapéutico; tampoco

Fecha de recepción: 16/07/2008. Aceptado (en forma revisada): 09/09/2008.

[1] Entendemos por *tumores, blastomas o neoplasias* las neoformaciones de uno o varios tipos celulares, sin capacidad de alcanzar una forma definitiva en su desarrollo y que reproducen más o menos la estructura del tejido u órgano del que se han formado. Así mismo poseen la capacidad de persistir una vez establecidos, siendo generalmente de crecimiento progresivo e indefinido. No desempeñan ninguna actividad útil para el organismo, sino que actúan independiente de él y regulados por leyes propias.

podemos predecir en qué casos aparecerán las consecuencias adversas derivadas del tratamiento, ni en qué pacientes esas consecuencias serán de carácter grave. Si los tumores que no responden (o lo hacen pobremente) pudieran ser identificados, se podrían proponer cambios en la terapéutica, buscando incrementar el porcentaje de éxitos. De igual forma, la identificación precisa de los pacientes que padecerán complicaciones graves asociadas al tratamiento permitiría evitar o reducir este riesgo.

Lo anterior es así porque tanto el estudio de extensión de las enfermedades neoplásicas, como la predicción de las probabilidades de control tumoral o de complicaciones tras la terapéutica, se basan en técnicas de diagnóstico que son imperfectas; por ello, no siendo posible establecer un pronóstico exacto para cada individuo, la investigación en este terreno es de enorme interés [2, 3, 4, 5, 6].

En este trabajo resumimos nuestra experiencia en el estudio de la frecuencia, y la gravedad, de los efectos adversos producidos por la radiación en la piel de las pacientes tratadas con radioterapia por cáncer de mama. Para ello, a partir del seguimiento clínico de las pacientes y de los resultados de un *test de radiosensibilidad* (basado en el análisis cuantitativo de la distribución de fragmentos de ADN rotos por la radiación), hemos investigado la significación biológica de los procedimientos desarrollados. En particular, hemos analizado la correlación de los resultados del test con la *morbilidad* asociada a la radioterapia.[2]

2 Probabilidad de curación en radioterapia

El éxito de la radioterapia se basa en erradicar las células tumorales existentes en el volumen terapéutico. Uno de los modelos matemáticos más utilizados para describir la relación entre la probabilidad de control tumoral PCT y la dosis D es la ecuación logística [7]:

$$PCT(D) = \left[1 + \left(\frac{D_{50}}{D}\right)^{4\gamma}\right]^{-1}, \tag{1}$$

donde

1. D_{50} es la dosis necesaria para conseguir controlar el 50 % de los tumores tratados;

2. D es el valor de la dosis total administrada. En la gráfica de la Figura 1, D varía de 0 a 100 Gy;[3]

3. γ es el máximo valor del gradiente dosis-respuesta normalizado [7].

[2] La morbilidad asociada a la administración de un tratamiento es una cuantificación de la importancia de los efectos secundarios producidos directamente relacionados (causados) con la terapéutica prescrita.

[3] 2 Gy es la fracción usual de la dosis que se administra cada día en radioterapia oncológica; 1 Gy equivale a 1 J/Kg; la dosis en radioterapia se administra en fracciones diarias, 5 días por semana, durante períodos que suelen durar de 5 a 7 semanas.

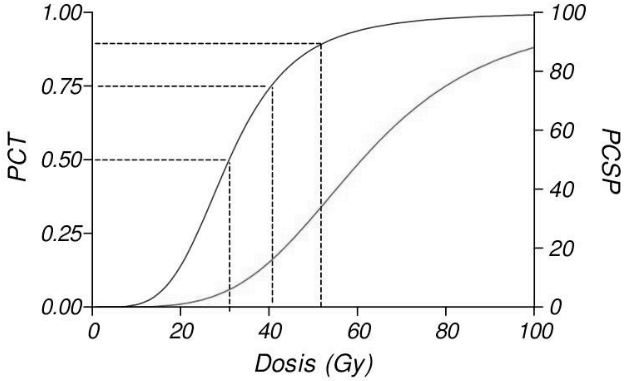


Figura 1: Representación gráfica idealizada de la probabilidad de control tumoral (*PCT*, curva superior negra) y de la probabilidad de padecer complicaciones severas en la piel (*PCSP*, curva inferior roja) asociadas a la radioterapia por cáncer de mama, en función de la dosis (en unidades Gy).

La representación gráfica de PCT, suponiendo que $D_{50} = 40$ Gy y $\gamma = 3$, se ofrece en la Figura 1. En esta Figura también se ha incluido la probabilidad de inducir complicaciones severas en la piel incluida dentro del volumen sometido a tratamiento, $PCSP$. En la curva correspondiente a esta probabilidad se han utilizado los valores $D_{50} = 45$ Gy, $\gamma = 6$.

La Figura 1 demuestra que el éxito en radioterapia — la probabilidad de controlar el tumor — depende de la dosis administrada; pero el incremento de la dosis aumenta también la probabilidad de inducir complicaciones severas en los tejidos normales irradiados conjuntamente con el tumor. La incidencia de complicaciones, o su gravedad, puede llegar a ser clínicamente inaceptable. Dicho de otra forma: lo que limita la dosis a administrar en un paciente con cáncer es la tolerancia de los tejidos normales a la radiación [8].

Aunque el planteamiento precedente es muy simple, conduce de entrada a varias cuestiones de interés: ¿Qué estrategias pueden establecerse para "maximizar" $PCT(D)$ y "minimizar" $PCSP(D)$? ¿Tiene sentido la búsqueda de equilibrios ?

3 Radiosensibilidad celular

Entenderemos la *radiosensibilidad celular* como una medida de la respuesta de las células a la radiación ionizante. Sabemos que tanto la radiosensibilidad de las células de los tumores, como la radiosensibilidad de las células de los tejidos sanos incluidos en el volumen sometido a tratamiento, son variables.

Una manera sencilla de cuantificar la radiosensibilidad celular consiste en medir la capacidad de la radiación para impedir que células procedentes del tumor, o de los tejidos sanos, sean capaces de formar colonias cuando se las cultiva adecuadamente. A este procedimiento analítico se le llama *ensayo clonogénico* y básicamente consiste en cultivar, en un medio apropiado, una suspensión de las células del tumor, o de los tejidos sanos, y medir su capacidad para formar colonias de células descendientes [9].

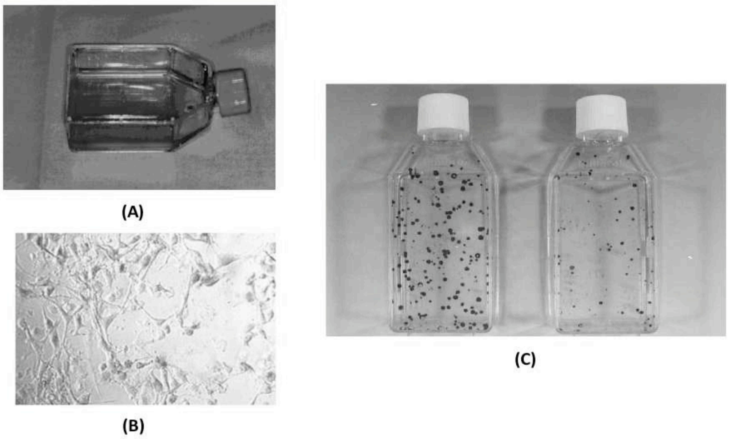


Figura 2: Ensayo clonogénico. Procedentes del tumor, o de tejidos normales, se siembran en un frasco de cultivo celular (A), el número de células que se considere adecuado. Las células disponen de medio nutriente y de las condiciones ambientales necesarias para crecer en el cultivo. En (B) se muestra una imagen de microscopía por contraste de fase, obtenida tras 6 días de cultivo. Las células crecen adheridas a la superficie del frasco y se mantienen agrupadas en torno a la progenitora. Al cabo de tres semanas de cultivo se elimina el medio nutriente, se lavan las células y se tiñen con una disolución de colorante (en este caso azul de metileno). Las colonias formadas son visibles a simple vista como puntos azules.

En el ejemplo de la Figura 2 se preparó una suspensión de células tumorales (A), que se cultivaron hasta alcanzar un número apropiado de descendientes. Después se dividió el cultivo en dos partes: células-control y células tratadas (con una dosis de 6 Gy). Los dos frascos se cultivaron en las mismas condiciones y durante el mismo tiempo. Después del período de incubación necesario, las colonias formadas se tiñeron (con una disolución de azul de metileno) y se contaron. En el frasco correspondiente a las células-control, situado a la izquierda en la Figura 2 (C), se habían sembrado 100 células; 3 semanas después del inicio del experimento se pueden contar 80 colonias; por tanto, podemos decir que la eficiencia del sembrado fue $Ef_C = 80/100 = 0{,}8$. En el frasco de células tratadas se sembraron 400 células y tres semanas después se contaron 40 colonias; en consecuencia, la eficiencia del sembrado de las células tratadas (frasco situado a la derecha en la Figura 1 (C)) fue más baja: $Ef_T = 40/400 = 0{,}1$.

La *fracción de supervivencia* de las células tratadas se calcula mediante la operación siguiente:

$$F_S = \frac{Ef_T}{Ef_C} = \frac{0{,}1}{0{,}8} = 0{,}125.$$

La fracción de supervivencia se puede interpretar como el porcentaje de células que sobreviven frente a la dosis y es una medida de la radiosensibilidad celular para ese valor de la dosis.

4 La curva de supervivencia celular

La *curva de supervivencia* es la gráfica que resulta al representar los valores de la fracción de supervivencia (que, como se ha dicho, permite cuantificar la radiosensibilidad celular) en función de la dosis.

De los resultados de las curvas de supervivencia celular obtenidas para numerosos tipos de células [10, 11], se ha extraído información muy valiosa sobre los efectos biológicos de las radiaciones. La Figura 3 corresponde a datos de supervivencia celular tras radiación en 5 líneas celulares tumorales humanas. En la Figura 3-a los datos han sido representados en escala lineal. En esa gráfica se aprecia un descenso de los valores de supervivencia que es muy pronunciado en la región de bajas dosis, para después descender más lentamente y aproximarse de manera asintótica a cero cuando la dosis aumenta considerablemente. En la Figura 3-b la representación gráfica se ha realizado en coordenadas semi-logarítmicas.

Hay dos razones que aconsejan representar las curvas de supervivencia en coordenadas semi-logarítmicas:

1. Si la muerte celular ocasionada por la radiación es un suceso estocástico, entonces esperamos que la representación semi-logarítmica de la supervivencia celular frente a la dosis sea (aproximadamente) una línea recta [10, 15].

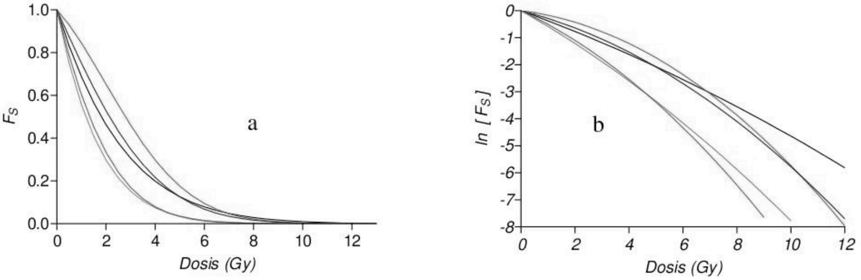


Figura 3: Representación de la curva de supervivencia celular frente a la dosis: (a) escala lineal y (b) escala semi-logarítmica. F_S representa la fracción de supervivencia.

2. La escala semi-logarítmica permite ver y comparar más fácilmente los efectos de la radiación a bajos niveles de dosis.

5 Un modelo matemático de supervivencia celular

Los estudios en radiobiología han avanzado considerablemente gracias a los modelos matemáticos. El modelado y los métodos matemáticos también son necesarios cuando se quieren relacionar los estudios experimentales con la respuesta de las personas enfermas al tratamiento que se les prescribe [12, 13, 14]. Esto es, cuando se trata de aplicar lo aprendido y los procedimientos desarrollados en el laboratorio a la práctica clínica.

Cuando los datos de supervivencia de las células de mamíferos irradiadas *in vitro* se representan en coordenadas semi-logarítmicas en función de la dosis, suelen tener una forma curvilínea (véase la Figura 3 b). Los datos obtenidos utilizando células tumorales humanas se ajustan bien a un modelo matemático en el que la fracción de supervivencia F_S se relaciona con la dosis D como sigue:

$$F_S(D) = e^{-\alpha D - \beta D^2}. \tag{2}$$

En la expresión anterior, aparecen dos coeficientes α (componente lineal) y β (componente cuadrática). Una manera alternativa de escribir (2) es la siguiente:

$$-\frac{\log F_S(D)}{D} = \alpha + \beta D. \tag{3}$$

Si representamos el primer miembro de (3) frente a la dosis D, obtenemos una línea recta (Figura 4) cuya ordenada para $D = 0$ y cuya pendiente son, respectivamente, α y β:

$$\alpha = -\frac{\log F_S(D)}{D}\bigg|_{D=0}, \quad \beta = \frac{d}{dD}\left(-\frac{\log F_S(D)}{D}\right).$$

En la región de dosis bajas, la forma de la curva de supervivencia está dominada por la componente lineal [15]. Teniendo en cuenta (3) y que la fracción de dosis con la que diariamente se trata cada paciente es de aproximadamente 2 Gy, se suele aceptar que una buena aproximación de la ordenada inicial α es

$$\alpha \simeq -\frac{\log F_S(D)}{D}\bigg|_{D=2\,\mathrm{Gy}}.$$

Tanto α como β son coeficientes de considerable importancia en radioterapia oncológica; de hecho la componente lineal de la curva de supervivencia refleja la producción de lesiones irreparables que conducen a la muerte de la célula [16]. Por tal motivo, se dice que α es el *coeficiente de lesión letal*.

El coeficiente β se interpreta como una medida de la probabilidad de muerte celular secundaria, debida a la existencia de lesiones que, producidas simultánea o sucesivamente, pueden ser reparadas o, por el contrario, interaccionar para dar origen a una lesión letal. Así, la reparación y la fijación de lesiones se pueden entender como mecanismos independientes que compiten entre sí. De acuerdo con la terminología propuesta por Curtis [17], las lesiones de este tipo se denominan *potencialmente letales*. Por tanto, β es el *coeficiente de lesión potencialmente letal*.

6 Radiosensibilidad de las células de los tumores humanos

Los estudios radiobiológicos realizados en la década de los 80 [11, 18] permitieron sugerir la hipótesis de que la fracción de supervivencia celular $F_S(D_{\min})$ medida *in vitro* tras la irradiación de las células con una dosis $D_{\min} = 2$ Gy, podría ser útil para predecir la respuesta de un tumor a la radioterapia.

Se suele decir que $F_S(D_{\min})$ es un valor representativo de la *radiosensibilidad celular intrínseca.*

Denominaremos *efecto global* a la fracción de supervivencia celular que resulta tras la administración de N dosis diarias, de 2 Gy cada una, a un conjunto de células. Está dado por

$$E = \left(F_S(D_{\min})\right)^N. \tag{4}$$

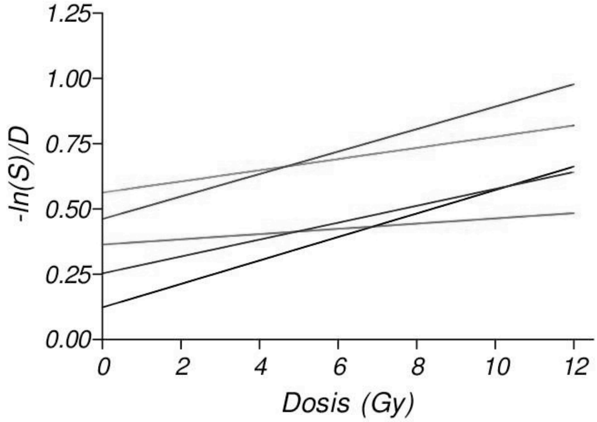


Figura 4: Modelo lineal-cuadrático de curva de supervivencia. Las curvas (rectas) dependen de dos coeficientes: el coeficiente lineal α (ordenada para dosis cero) y el coeficiente cuadrático β (la pendiente).

La variabilidad en la supervivencia a 2 Gy, esto es, en la radiosensibilidad celular intrínseca, puede ser suficiente para explicar los resultados (éxitos o fracasos) de la radioterapia oncológica [11]. En la Figura 5 se ha supuesto que la fracción de supervivencia varía de 0,1 a 0,6. Si aceptamos que el 50 % de la probabilidad de control tumoral se alcanza cuando la supervivencia de las células del tumor se reduce a 1, la Figura 5 deja claro que tal valor se alcanza para números de dosis distintos, dependiendo de la radiosensibilidad que se considere.

7 La predicción de la respuesta tumoral y tisular a la radiación

Supongamos ahora que los resultados de la terapéutica (probabilidad de control tumoral y probabilidad de padecer complicaciones asociadas al tratamiento) dependen de la radiosensibilidad de las células del tumor y de la radiosensibilidad de las células de los tejidos sanos incluidos en el volumen de irradiación. Dicho de otra forma, que se curarán más fácilmente los tumores cuyas células sean más radiosensibles y aparecerán menos complicaciones secundarias de carácter grave en las personas cuyos tejidos sean más resistentes

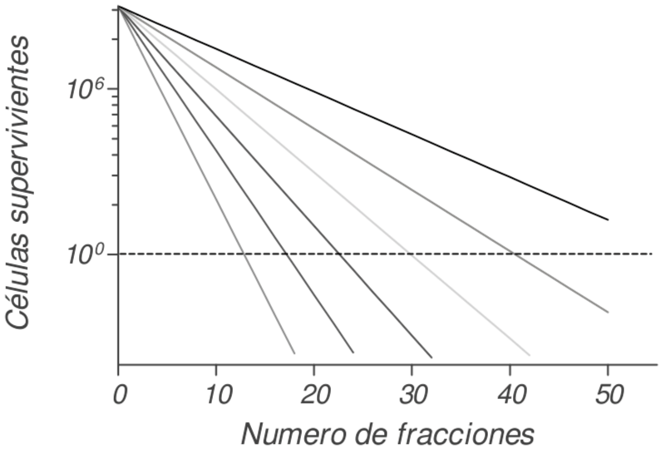


Figura 5: Probabilidad de supervivencia celular en función de la radiosensibilidad celular intrínseca (medida como $F_S(D_{\min})$) y del número de dosis administradas N. La línea horizontal punteada marca el nivel hipotético de supervivencia para una probabilidad de curación del 50 %.

(tolerantes) a la radiación.

La práctica clínica nos ha permitido saber que los efectos secundarios asociados a la radioterapia varían entre indetectables, o mínimos, y tan extraordinariamente severos que ponen en riesgo la vida de la persona y obligan a la suspensión del tratamiento [19, 20, 21]. Por otra parte, numerosas observaciones clínicas han demostrado que esa variabilidad en la respuesta tisular no puede atribuirse exclusivamente a factores de carácter físico tales como la dosimetría, el fraccionamiento, o la planificación de la radioterapia [22].[4]

[4] El daño producido por la radiación sobre las células ocasiona una perturbación del equilibrio tisular (el equilibrio de los tejidos del organismo) a la que los tejidos deben responder. La respuesta consta de reacciones de dos tipos:

a) Reacción precoz: en ella se incluyen los síntomas clínicos derivados del efecto tóxico inmediato de la radiación. En la piel, estos síntomas aparecen pocos días después del inicio del tratamiento y entre ellos y de menor a mayor gravedad, está el eritema (enrojecimiento de la piel), la descamación y la ulceración. Las lesiones agudas curan con facilidad.

b) Reacción tardía: Se trata de la manifestación de toxicidad que aparece varios meses después del término del tratamiento. El riesgo de que este daño colateral acontezca perdura

Los aspectos biológicos relacionados con el estudio cuantitativo de la radiosensibilidad intrínseca de las células de los tejidos sanos y con la respuesta tisular a la radiación en pacientes sometidos a radioterapia son de enorme interés clínico; véase [2].

La hipótesis de trabajo puede formularse así:

> *"La respuesta de los tejidos a la radiación está determinada por una componente genética; por tanto, es razonable sugerir que la radiosensibilidad de las células en cultivo es un reflejo de la constitución genética de los individuos de los que derivan. El estudio de la respuesta celular al estrés inducido por el tratamiento genotóxico*[5] *puede permitir la identificación de aquellos pacientes con mayor riesgo de padecer complicaciones de carácter grave, como efecto secundario del tratamiento antitumoral*".

Si esta hipótesis quedara definitivamente probada, la selección de enfermos en base a estimadores de pronóstico de control y complicación permitiría ofrecerles opciones terapéuticas individualizadas orientadas a la reducción de los índices globales de morbilidad y al aumento, si ello fuera posible, de la probabilidad de control de las enfermedades neoplásicas [12, 20, 23, 24].

A pesar del esfuerzo de muchos investigadores clínicos y básicos y de la enorme cantidad de información de la que se dispone, aún no es posible definir cuál es "exactamente" la mejor opción terapéutica para un individuo en particular. Por ello el progreso en el tratamiento del cáncer sigue necesitando:

1. Del estudio de extensión del tumor, de sus células y de las células de los tejidos sanos del paciente.

2. De la acertada decisión terapéutica y de su cuidadosa ejecución.

3. Del seguimiento de los pacientes y la valoración clínica de los resultados del tratamiento, incluyendo un informe detallado de la morbilidad asociada.

4. Del análisis de correlación entre las variables que surjan de unos y otros estudios.

5. Finalmente, de la aportación de modelos teóricos que nos ayuden a comprender para poder predecir.

8 Un método para cuantificar el daño inicial radioinducido en el ADN

Se cree que la radiación ionizante mata a las células eucariotas a través del daño que induce sobre la estructura del ADN. Trabajos previos nos han permitido

durante toda la vida de la persona tratada y su gravedad también es variable. Este factor de riesgo es el limitante principal de la magnitud de la dosis de radiación que puede administrarse a un paciente.

[5] Se trata del agente físico o químico que produce la destrucción celular a través del daño que induce en el ADN genómico.

saber que la supervivencia de células tumorales, tras irradiación, se correlaciona con la cantidad de rupturas dobles de la cadena de ADN ocasionados por unidad de dosis (Gy) y unidad de ADN (la unidad de ADN comúnmente aceptada equivale a 200 millones de pares de bases, 200 Mpb) [3, 8, 25, 26].

Cuando las células se irradian a la temperatura de 0^o C, las cadenas de ADN rotas por la radiación permanecen separadas. Si en estas condiciones se somete el ADN a electroforesis en gel utilizando la técnica de *campo pulsante,* es posible separar los fragmentos de ADN en función de su tamaño. El gel puede ser calibrado midiendo las distancias que recorren fragmentos de ADN de tamaño conocido (p. ej. *Saccharomyces pombe* y *Saccharomyces cerevisiae*).

Basándonos en este hecho, hemos desarrollado un método para calcular el número de rupturas dobles de cadena del ADN por unidad de dosis que está basado en el modelo publicado previamente [27, 28] que utiliza la distribución de Poisson para describir la frecuencia de tamaños de los fragmentos de ADN para un número de rupturas dobles de cadena conocido.

Para obtener datos numéricos de la distribución de tamaño de los fragmentos de ADN, una vez desarrollada la electroforesis en el ADN, se tiñó el gel con el colorante fluorescente Bromuro de Etidio y se visualizó mediante iluminación ultravioleta (véase Figuras 6 y 8). Es posible cuantificar la intensidad de la señal fluorescente mediante el programa comercial de análisis de imagen “Visilog”.

Sea S el tamaño medio del cromosoma y designemos con el símbolo x los distintos tamaños de fragmentos de ADN. Representando en abscisas los valores de x/S (tamaños relativos) frente a la fracción de ADN extraída del pozo, se obtiene la distribución de la señal de fluorescencia para cada dosis. Al incremento de la dosis le corresponde un incremento de la cantidad de ADN extraído y, en consecuencia, un aumento de la intensidad registrada (Figura 8).

La distribución de los tamaños de los fragmentos de ADN rotos por la radiación obedece a la ley

$$F(x) = \frac{\mu}{S} G(x)\, e^{-\mu x/S}, \tag{5}$$

donde $F(x)$ es la intensidad de la señal de fluorescencia registrada, μ es el número de rupturas dobles en la cadena de ADN y $G(x)$ viene dada por la siguiente expresión:

$$G(x) = x\left[2 + \frac{\mu(S-x)}{S}\right]. \tag{6}$$

Las intensidades de señal $F_1(x)$ y $F_2(x)$ correspondientes a dos dosis D_1 y D_2 (con $D_2 > D_1$) serán:

$$F_1(x) = \frac{\mu_1}{S} G_1(x) e^{-\mu_1 x/S} \tag{7}$$

y

$$F_2(x) = \frac{\mu_2}{S} G_2(x) e^{-\mu_2 x/S}, \tag{8}$$

donde μ_1 y μ_2 son los números de rupturas dobles de cadenas de ADN para las respectivas dosis.

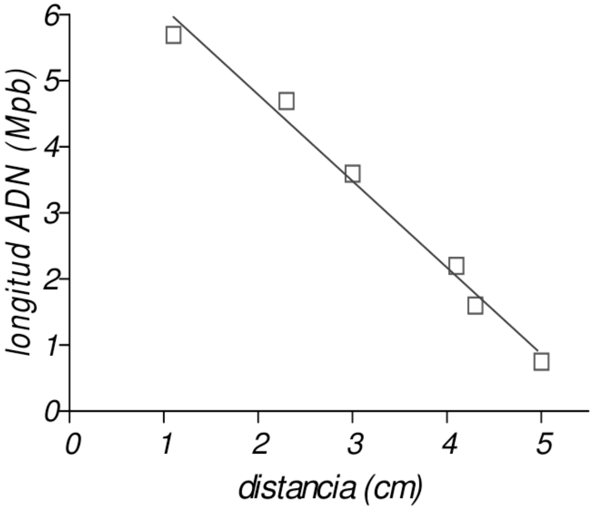


Figura 6: Calibración del tamaño de los fragmentos de ADN en función de la distancia recorrida en el gel de electroforesis de campo pulsante por los cromosomas de levaduras (*Saccharomyces pombe* y *Saccharomyces cerevisiae*) de longitud (Mpb) conocida.

Ahora, dividiendo $F_1(x)$ por $F_2(x)$ tenemos:

$$F_R(x) = \frac{F_1(x)}{F_2(x)} = Ae^{(\mu_2-\mu_1)x/S}, \tag{9}$$

donde

$$A = \frac{\mu_1(2S + \mu_1(S-x))}{\mu_2(2S + \mu_2(S-x))}. \tag{10}$$

Si, en la ecuación (10), x es mucho menor que S, podemos simplificar A para obtener

$$A_0 = \frac{\mu_1(2+\mu_1)}{\mu_2(2+\mu_2)}, \tag{11}$$

con lo que podemos escribir:

$$F_R(x) \simeq A_0 e^{(\mu_2-\mu_1)x/S} \tag{12}$$

y tomando logaritmos

$$\log F_R(x) \simeq \log A_0 + (\mu_2 - \mu_1)\frac{x}{S}. \tag{13}$$

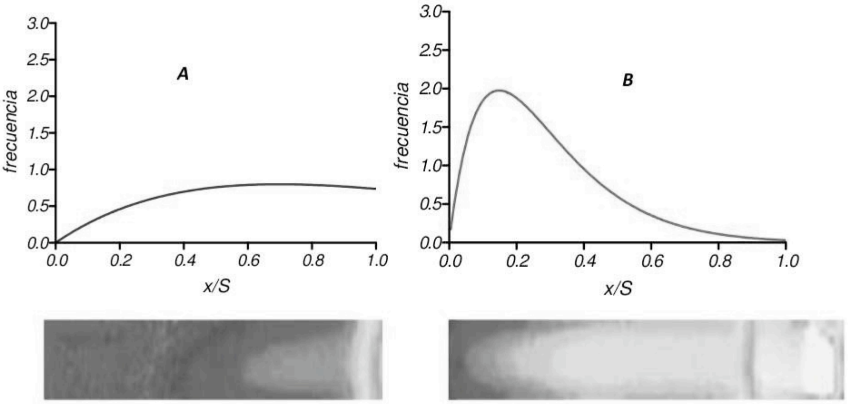


Figura 7: Distribución de tamaños relativos de ADN (x es el tamaño del fragmento; S es el tamaño medio del cromosoma) tras irradiación de dos nuestras celulares a diferentes dosis. La dosis empleada en el experimento representado en A es inferior a la empleada en el experimento representado en B. Bajo cada gráfica se ha incluido la imagen de migración de los fragmentos de ADN visualizada mediante tinción con bromuro de etidio e iluminación ultravioleta.

La representación gráfica de esta aproximación de $\log F_R$ da una línea recta cuya pendiente B es la diferencia entre las rupturas dobles de cadena de ADN producidas por la dosis 2 y la dosis 1, i. e. $B = \mu_2 - \mu_1$, y de cuya ordenada para $x = 0$ se puede deducir A_0 — ecuación (11) — que es también función de μ_1 y μ_2. Los valores de B y A_0 se obtienen tras el ajuste de los resultados experimentales a la relación (13). De esta forma obtenemos un sistema de dos ecuaciones con dos incógnitas para μ_1 y μ_2 cuya solución es:

$$\mu_1 = \frac{-(2 - A_0 - 2A_0B) + \sqrt{(2 - A_0 - 2A_0B)^2 + 4A_0B(1 - A_0)(2 + B)}}{2(1 + A_0)} \tag{14}$$

y

$$\mu_2 = B + \mu_1 \tag{15}$$

En un experimento dirigido al cálculo del daño inducido por la radiación sobre una determinada muestra celular, se debe incluir el análisis de fragmentos

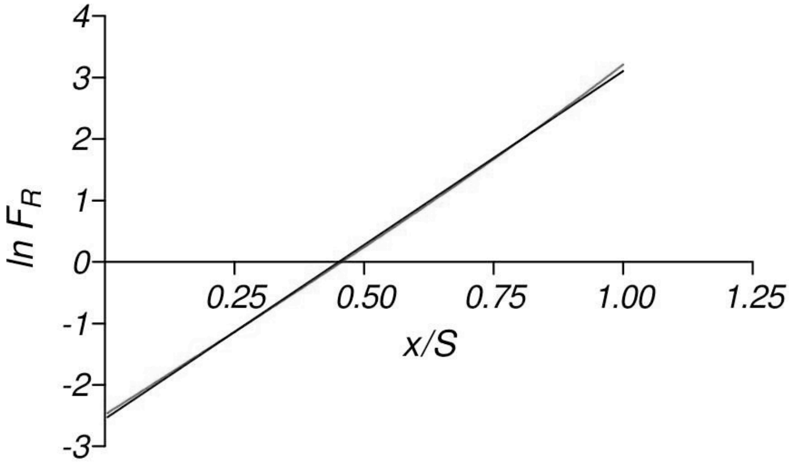


Figura 8: Utilizando los datos de la Figura 7, se obtine aquí la representación gráfica de (13). De la pendiente y de la ordenada para $x = 0$ se pueden deducir los valores de μ_1 y μ_2.

de ADN producidos tras la irradiación de las células a diferentes dosis. En nuestro caso, se han utilizado de 8 a 10 valores distintos de dosis, comprendidos dentro del rango de 0 a 45 Gy). De la comparación de cada dos parejas de distribuciones resulta un valor del parámetro μ para cada una de las dosis. Se obtienen así 8 ó 10 valores de μ por dosis.

La representación del valor medio de μ frente a la correspondiente dosis da una línea recta (Figura 9) cuya pendiente, calculada por regresión lineal, proporciona una estimación del promedio de las rupturas dobles de cadena de ADN producidas sobre la muestra celular por unidad de dosis y unidad de ADN.

De nuestros experimentos se puede deducir lo siguiente:

1. La relación entre el número de rupturas dobles de la cadena de ADN y la dosis se ajusta a una ecuación lineal al menos hasta los 300 Gy de dosis total [25].

2. La presencia de oxígeno influye claramente sobre la cantidad de daño inicial radioinducido [25] potenciando el efecto de la radiación. Este resultado coincide con lo que se ha descrito en estudios de supervivencia celular tras irradiación.

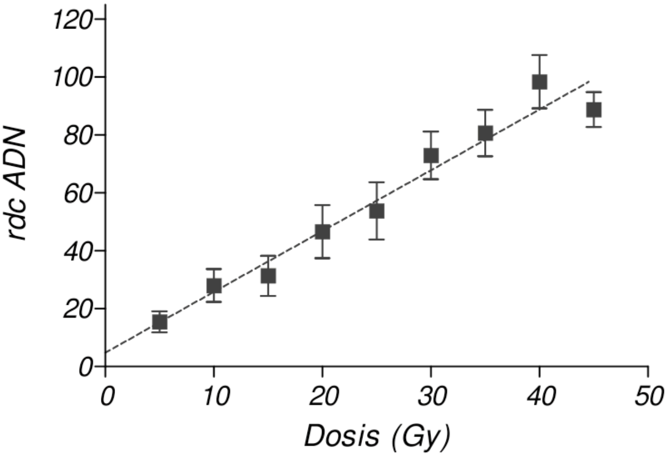


Figura 9: Daño molecular — rupturas dobles de la cadena de ADN — inducido sobre el ADN de la línea celular MCF-7. La pendiente de la recta de regresión obtenida ($x = 2.1$ rdc/Gy/200 Mpb) es una medida de la radiosensibilidad celular.

3. Entre la radiosensibilidad celular intrínseca, medida como supervivencia a 2 Gy, $F_S(D_{min})$, y la cantidad de daño inicial, medida como número de rupturas dobles de cadenas de ADN por unidad de dosis, existe una clara relación de proporcionalidad [9, 26].

Los resultados anteriores han permitido interpretar el número de rupturas dobles de la cadena de ADN referidas a la unidad de dosis y unidad de ADN como una nueva medida de la radiosensibilidad celular intrínseca.

9 El problema clínico de las consecuencias adversas de la radioterapia

La radioterapia contribuye al tratamiento oncológico de un amplísimo número de pacientes con cáncer y más de 800.000 pacientes europeos cada año se benefician de los procedimientos radiológicos. Inevitablemente, en el volumen corporal sometido a irradiación junto al tumor se incluye cierta cantidad de tejidos peritumorales sanos. El continuo avance del conocimiento, de la instrumentación y de las técnicas de dosimetría y planificación en radioterapia, han permitido que la proporción de pacientes con complicaciones severas haya disminuido considerablemente. A pesar de ello la morbilidad asociada a la radioterapia es un efecto colateral impredecible, progresivo e irreversible que incide sobre el 5 % de los pacientes tratados ($\approx$ 40.000 pacientes por año en Europa), que ocasiona el deterioro de la calidad de vida de los pacientes afectos y que tiene un costo asistencial importante. Se cree que los efectos secundarios de la radioterapia tienen una componente de predisposición genética determinante pero la identificación de cuáles son los pacientes sobre los que gravita el mayor riesgo de padecer complicaciones severas tras radioterapia es hoy un problema por resolver [14, 19, 21, 22].

En el momento actual, para cada tipo de tumor, se prescribe la dosis con la que empíricamente se han conseguido los mejores resultados. Es así que a todos los pacientes se les administra la misma dosis, en la confianza de que con ello se consigue la máxima probabilidad de control de la enfermedad local, acompañada de la mínima incidencia de complicaciones severas. La existencia de variabilidad en las reacciones tóxicas observadas hace que la máxima dosis utilizable en tratamientos convencionales se determine empíricamente, para conseguir que la proporción de pacientes que manifiesten reacciones tóxicas de carácter severo quede limitada a una cifra inferior al 5 % [23]. Si estos pacientes pudieran ser identificados por adelantado, se les ofrecería un tratamiento alternativo que tendría como objetivo evitar la reacción adversa. Al mismo tiempo, en el resto de pacientes supuestamente más *radioresistentes*, se prescribiría una dosis total superior capaz de aumentar la probabilidad de control tumoral [23]; recuérdese (1).

Por tanto, los beneficios de disponer de un procedimiento válido para identificar los pacientes en base a su radiosensibilidad y prescribir la dosis en base a este conocimiento, serían de considerable importancia [29]. Este planteamiento abre con claridad una puerta a la incorporación de técnicas matemáticas.

10 Reacciones agudas y tardías tras radioterapia

El objetivo de la radioterapia es reducir a cero la supervivencia de las células tumorales con capacidad clonogénica.[6] La supervivencia de éstas en el tumor lleva a la recurrencia tumoral. Por otra parte, para minimizar los efectos adversos en los tejidos peritumorales sanos, hay que respetar la supervivencia, y la función, del mayor número posible de las células de los tejidos normales incluidos en el volumen terapéutico. Esta doble condición determinante a la vez de probabilidad de curación y de probabilidad de complicaciones severas, explica los límites de la dosis a administrar.

La respuesta diferencial a la radiación de los tejidos corporales está en relación directa con su ritmo proliferativo: manifiestan una respuesta tóxica

[6] Una célula clonogénica (o célula *stem*) es aquélla que, por sucesivas divisiones mitóticas en cultivo, es capaz de dar origen a un clon de células descendientes.

precoz (aguda) los tejidos altamente proliferativos (por ejemplo, el epitelio del tubo digestivo, la piel o la médula ósea); el resto, es decir, la mayor parte de los tejidos normales de los seres humanos, no manifiestan cambios morfológicos precoces ni alteraciones funcionales agudas tras ser tratados con las dosis de radiación habituales. Pueden manifestar, sin embargo, lesiones tardías, y es sobre todo la gravedad de este efecto adverso lo que limita la dosis de radioterapia que es posible aplicar.

Las lesiones ocasionadas por la radiación sobre los tejidos sanos, pueden curar gracias a dos procesos fundamentales: (a) uno de base molecular, que hace referencia a los mecanismos de reparación de las lesiones inducidas sobre el ADN y (b) otro de base celular, en el que se incluye la regeneración tisular derivada de la proliferación de las células troncales que sobreviven en el tejido sano irradiado, con su viabilidad y su potencial de diferenciación indemnes.

Al término de la radioterapia ciertas células de los tejidos irradiados pueden quedar altamente modificadas, constituyendo un foco de estimulación crónica que dé origen a la morbilidad tardía asociada al tratamiento [30], con independencia de cuáles sean las características proliferativas del tejido en cuestión.

De lo anterior se deduce que de la cantidad de daño inicial, de la eficacia de los mecanismos de reparación y del número, movilidad y plasticidad de las células troncales supervivientes, dependerá la importancia de los efectos tóxicos asociados al tratamiento con radiación.

11 Radiosensibilidad de las células de los tejidos sanos

Si la hipótesis de trabajo formulada en la Sección 7 fuese cierta, también debería ser cierto que las células normales, tomadas de distintos tejidos de la misma persona, dan resultados de radiosensibilidad equivalentes.

Para aclarar esta cuestión, hemos estudiado el daño inicial producido por la radiación sobre el ADN en dos tipos de células normales, linfocitos y células epidérmicas, tomadas simultáneamente de la misma persona en una observación pareada [8]. Del análisis estadístico de los datos apareados se puede concluir que tanto en los linfocitos como en las células de la piel existe una amplia variación en los valores de radiosensibilidad medidos en unas y otras células. Además, entre los valores de radiosensibilidad obtenidos para los linfocitos R_{S_L} y para las células epidérmicas R_{S_E}, existe una relación lineal $R_{S_E} = mR_{S_L} + n$ de pendiente $m = 0{,}9 \pm 0{,}13$, $n = 0{,}35 \pm 0{,}3$, coeficiente de correlación $R^2 = 0{,}694$ y significación estadística[7] entre ambas $P < 0{,}001$.

Este resultado apoya la hipótesis de que la radiosensibilidad de las distintas células pertenecientes a los tejidos normales del cuerpo humano es similar, circunstancia que debe ser simple reflejo de la identidad genómica de las distintas células y de la similar conformación de la cromatina que contienen [8].

[7] Aquí, P es la probabilidad de que la relación entre los valores de radiosensibilidad de los linfocitos y de las células epidérmicas, pertenecientes a una misma persona, no sea estadísticamente significativa.

En base a lo anterior y por razones de facilidad en la toma de la muestra, hemos elegido los linfocitos como modelo para medir la radiosensibilidad de las células de los tejidos normales de los pacientes sometidos a radioterapia.

12 La radiosensibilidad *in vitro* como test predictivo de respuesta a la radiación

Una vez elegido el modelo celular tipo sobre el que estudiar la radiosensibilidad de los tejidos normales a la radiación [31, 32, 33, 34], debemos, en primer lugar, probar que existe una real y significativa variación en los niveles de daño inicial inducido por la radiación sobre linfocitos tomados de distintos pacientes; después estudiaremos la distribución de los valores encontrados y finalmente avanzaremos las posibilidades de aplicación del test en la predicción del grado de lesión adversa que la radiación podría producir en cada paciente.

En trabajos previos de nuestro grupo de investigación hemos incluido un amplio grupo de mujeres tratadas de cáncer de mama [29, 35]. En todas ellas hemos medido la radiosensibilidad de sus linfocitos (ver Figura 9). De nuestros trabajos podemos concluir que la radiosensibilidad medida *in vitro* sobre las células humanas es una variable x cuyas características específicas son:

1. El valor de x se obtiene como promedio del daño inicial sobre el ADN, expresado como número de rupturas dobles de cadenas de ADN por unidad de dosis (Gy) y unidad de ADN (200 Mbp).

2. Células con diferente radiosensibilidad sufren diferentes cantidades de daño inicial en su ADN.

Naturalmente, los valores que se obtienen para x son siempre positivos. Los valores negativos carecen de significado biológico.

Muchas de las mujeres inicialmente incluidas en el estudio fueron sometidas a radioterapia tras cirugía. En ellas, el último día de tratamiento se realizó un análisis sistemático de los efectos agudos que la radiación había producido sobre su piel. De las reacciones tóxicas observadas se tiene constancia documental (una imagen fotográfica tomada en las mismas condiciones de distancia e iluminación, en todos los casos). Los efectos adversos agudos fueron clínicamente valorados de acuerdo con la clasificación del *Radiation Therapy Oncology Group* (RTOG). Los datos de toxicidad encontrados permiten categorizar las reacciones tóxicas en cinco grupos diferentes [36] que, de menor a mayor severidad, son:

- ***ARR*:** pacientes *altamente radioresistentes,* que son aquéllas que no manifiestan ningún efecto tóxico al término del tratamiento.

- ***MRR*:** pacientes *moderadamente radioresistentes,* que manifiestan al término del tratamiento un ligero eritema en la zona sometida a radioterapia.

- ***PRS*:** pacientes con *radiosensibilidad promediada,* que presentan eritema intenso y alguna zona con descamación.

- ***MRS*****:** pacientes *moderadamente radiosensibles,* en los que la descamación es manifiesta y pueden presentar áreas de descamación húmeda.

- ***ARS***: pacientes *altamente radiosensibles,* en las que la descamación húmeda es intensa y además existen áreas de ulceración y/o necrosis en la piel.

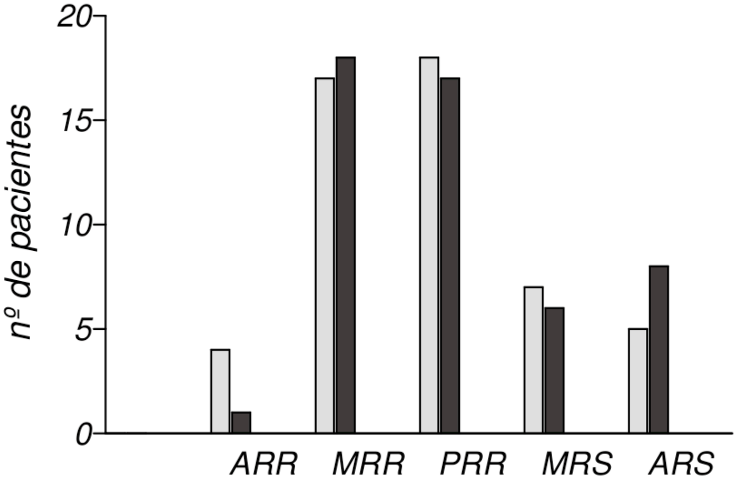


Figura 10: Resultados de la valoración clínica de la toxicidad aguda inducida por la radiación en la piel de mujeres tratadas con radioterapia por cáncer de mama tras mastectomía — gris claro—, o cirugía conservadora — gris oscuro— (véase la explicación en el texto).

El histograma de la Figura 10 muestra los resultados de la valoración de la toxicidad aguda, tras radioterapia, encontrada en este subconjunto de pacientes. Para estudiar la relación entre los resultados de la radiosensibilidad medida en los linfocitos y los efectos tóxicos observados en la piel, hemos realizado la comparación de los valores de radiosensibilidad *in vitro* en dos grupos de pacientes que, de acuerdo con la clasificación de Burnet [23], son:

1. Aquéllas en las que se manifiestan efectos adversos compatibles con lo que clínicamente se entiende como *toxicidad tolerable* (las pacientes incluidas en los grupos *ARR, MRR, PRS* y *MRS* en el histograma de la Figura 10).

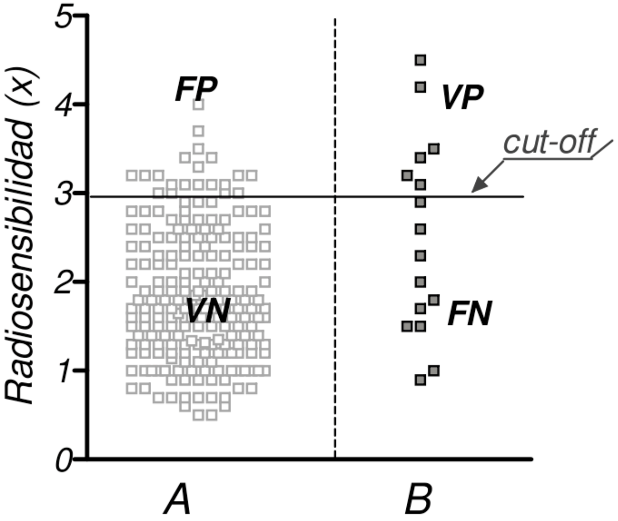


Figura 11: Representación de los valores individuales de radiosensibilidad en dos grupos de pacientes clasificados por la gravedad de sus reacciones agudas tóxicas tras radioterapia: A) personas tratadas que muestran efectos adversos tolerables y B) pacientes con reacciones tóxicas severas.

2. Las que, durante la radioterapia o a su término, manifiestan efectos que son clasificados como severos (pacientes altamente radiosensibles ARS en la Figura 10).

De la representación gráfica de los resultados obtenidos (Figura 11) se desprende que las distribuciones de los valores de radiosensibilidad, correspondientes a uno y otro grupo, se superponen en gran medida. Como consecuencia, el número de predicciones correctas de personas altamente radiosensibles (Verdaderos Positivos: VP en grupo B) y el número de errores (Falsos positivos: FP en grupo A) dependen del umbral de discriminación que se elija.

Un parámetro numérico útil para evaluar estadísticamente la fiabilidad de este test diagnóstico es la medida del área Z que queda bajo la curva ROC (Figura 12)[8] El valor numérico de Z es indicativo de la potencia del test para

[8] La curva ROC es la que resulta al unir los puntos (E', S) correspondientes a los distintos valores de corte (umbrales de decisión o *cut-off*). Aquí, $E' = 100 - E$, E es la especificidad del test y S es la sensibilidad del mismo. Los valores de E' y S son, respectivamente, las

discriminar entre las distribuciones de los grupos comparados. En nuestro caso este valor ha resultado ser: $Z = 0{,}675 \pm 0{,}072$ (Figura 12), lo que indica que el test, en las condiciones que ha sido aplicado, posee escasa o moderada capacidad discriminatoria y no permite identificar con seguridad a los pacientes sobre los que gravita el mayor riesgo de padecer efectos agudos de carácter grave [29].

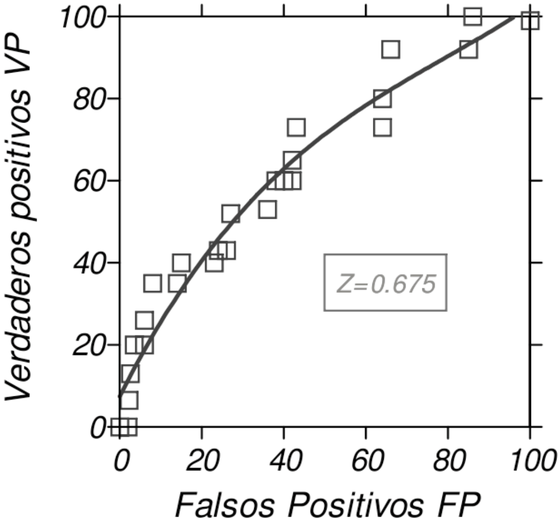


Figura 12: A partir de los datos representados en la Figura 11 se obtiene la correspondiente curva ROC para evaluar la fiabilidad del test de radiosensibilidad Z.

13 ¿ Cuál puede ser la ganancia terapéutica basada en la individualización de la dosis ?

A pesar de la ausencia de pruebas concluyentes de existencia de correlación del parámetro de radiosensibilidad estimado *in vitro* (a partir del número de rupturas dobles de cadenas de ADN) y las manifestaciones precoces de daño en los tejidos sanos, del estudio de la distribución de los valores de radiosensibilidad se puede extraer información valiosa. En efecto, a partir de esta distribución estadística es posible predecir cuál sería la ganancia de un programa de tratamiento basado en la individualización de la dosis a administrar a cada

proporciones de Falsos y Verdaderos Positivos (en %).

paciente en función de su radiosensibilidad [37].

Trabajos anteriores de nuestro grupo han permitido demostrar que el daño inducido por la radiación sobre el ADN, esto es, el número de rupturas dobles R de la cadena de ADN, crece de manera lineal con la dosis D (en unidades Gy), tal y como se indica en la Figura 9:

$$R = xD. \tag{16}$$

La pendiente x caracteriza esta relación y puede ser interpretada como el estimador de radiosensibilidad del ensayo. Si existe una relación entre el nivel de daño y el efecto final, existirán también un valor de referencia para el daño tolerable, R_r, y, por tanto, un valor de referencia para el parámetro de radiosensibilidad, x_r, que permitirá dividir a la población de pacientes en dos conjuntos, según la respuesta que esperamos sea mayor o menor que la de referencia.

Si D_T es la dosis total preestablecida (determinada por las caracterís ticas del tumor y del tratamiento establecido), x_r está dado por

$$x_r = \frac{R_r}{D_T}\,.$$

Para aquellos individuos con $x < x_r$, podemos plantearnos un aumento de la dosis, de tal manera que se alcance el nivel de rupturas dobles de referencia:

$$\Delta D = D_T\left(\frac{x_r}{x} - 1\right), \tag{17}$$

esperando de este modo aumentar la probabilidad de control tumoral alcanzando unos efectos secundarios preestablecidos y tolerables (aunque mayores que los padecidos por el paciente en el caso de no realizarse el ensayo y el consiguiente aumento de su dosis).

En la Figura 13 se muestra la distribución de rupturas dobles de cadena por unidad de dosis y de longitud de ADN (rdc/Gy/200 Mpb) para los linfocitos de una población de 226 mujeres con cáncer de mama. El histograma puede ajustarse adecuadamente mediante una distribución log-normal [1, 4]:

$$P(x) = \frac{1}{x\sigma\sqrt{2\pi}}\exp\left[-\frac{(\log x - \mu)^2}{2\sigma^2}\right], \qquad x > 0, \tag{18}$$

de parámetros: $\mu = 0{,}42 \pm 0{,}03$ y $\sigma = 0{,}52 \pm 0{,}02$.

A partir de esta distribución del parámetro de radiosensibilidad, proponemos el siguiente programa de individualización de las dosis, que depende de la elección de x_r y de un valor máximo x_m del parámetro de radiosensibilidad:

1. Si el valor del parámetro de radiosensibilidad es mayor o igual que x_m, el riesgo de efectos severos en el tejido normal es alto. Para estos pacientes debe hacerse una vigilancia estrecha durante el tratamiento o buscarse un programa terapéutico alternativo.

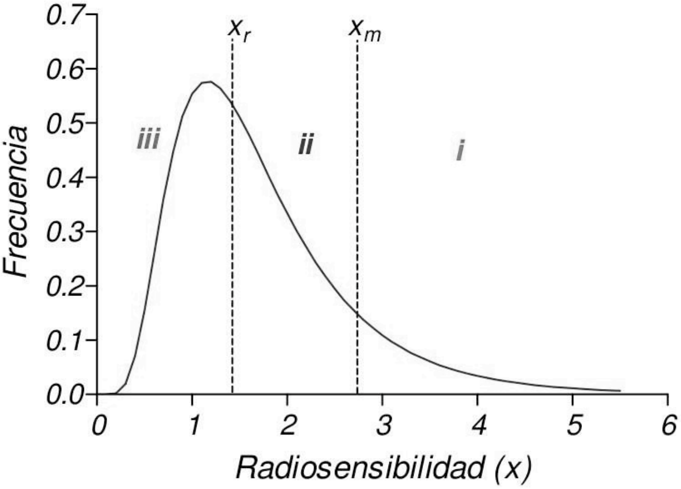


Figura 13: . Distribución log-normal de los valores de radiosensibilidad en linfocitos tomados de pacientes con cáncer de mama. En la simulación de resultados de la individualización terapéutica se considerarán las regiones i, ii y iii, (véase en el texto).

2. Para valores de x entre x_r y x_m, no se hacen cambios en el tratamiento previsto.
3. Finalmente, para los pacientes con un parámetro de radiosensibilidad menor que x_r, puede plantearse un aumento de la dosis de acuerdo con la ecuación (17); no obstante, por precaución, hemos limitado este aumento al 20 % de la dosis total de referencia:

$$\Delta D = 0{,}2\, D_T \left(\frac{x_r}{x} - 1\right).$$

La división de pacientes en estos tres grupos se muestra también en la Figura 13. Naturalmente, sería esencial determinar con precisión x_r y x_m, pero esto exige un conocimiento clínico y biológico mayor del que poseemos y, por tanto, hemos de conformarnos con tanteos razonables. Más abajo, para los cálculos realizados, tomaremos un sólo conjunto de valores, pero en [37] pueden verse los resultados de diferentes elecciones.

Puesto que tendremos que evaluar la probabilidad de control tumoral de los pacientes, hemos de elegir un modelo concreto que nos permita calcularla y comparar, como es nuestro propósito, las probabilidades de control antes y después de la existencia de un programa de individualización.

Como modelo para el cálculo de la probabilidad de control tumoral hemos elegido el modelo logístico que se indicó en (1) (las razones pueden consultarse en [37]). Utilizando esta ecuación es posible calcular el incremento de probabilidad de control tumoral que podría esperarse si se aplicara un programa como el propuesto mediante un procedimiento analítico [37], pero el problema parece enunciado para emplear el método de Monte Carlo: tenemos una distribución de probabilidad conocida para el parámetro de radiosensibilidad y podemos elegir distribuciones normales para los parámetros del modelo de probabilidad de control en la población de pacientes.

Basta, pues, "sortear" los parámetros asociados a cada paciente (x, D_{50} y γ); calcular la probabilidad de control que cabe esperar en un esquema de radioterapia con una dosis total prefijada D_T; después, obtener la nueva dosis total y la probabilidad de control a que da lugar de acuerdo con el procedimiento descrito y, por último, la diferencia entre las probabilidades de control nueva y antigua. En la Figura 14 puede verse un sencillo diagrama de flujo del programa que guía los cálculos propuestos.

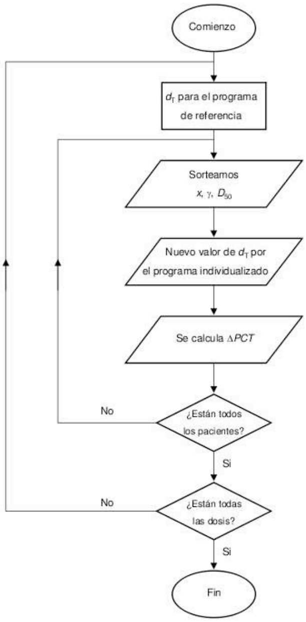

Comienzo
d_T para el programa de referencia
Sorteamos x, γ, D_{50}
Nuevo valor de d_T por el programa individualizado
Se calcula ΔPCT
No
¿Están todos los pacientes?
Si
No
¿Están todas las dosis?
Si
Fin

Figura 14: Diagrama de flujo en el que se esquematizan los pasos necesarios para obtener la probabilidad de control tumoral ΔPCT en función de la dosis total administrada D_T, el parámetro γ y el valor de D_{50}.

D_{50}(Gy), γ	$\Delta(PCT)$(%) Valor medio	Porcentaje de pacientes con $\Delta(PCT) > 0\,\%$	Porcentaje de pacientes con $\Delta(PCT) > 10\,\%$	Porcentaje de pacientes con $\Delta(PCT) > 50\,\%$
40 ± 5, 3 ± 1	5.0 ± 0.1	50.6 ± 0.2	17.7 ± 0.1	0.30 ± 0.03
40 ± 5, 5 ± 1	3.4 ± 0.1	35.6 ± 0.2	9.2 ± 0.1	0.88 ± 0.04
50 ± 5, 3 ± 1	17.0 ± 0.3	57.1 ± 0.2	48.9 ± 0.2	7.47 ± 0.08
50 ± 5, 5 ± 1	21.7 ± 0.4	56.2 ± 0.3	45.5 ± 0.3	19.7 ± 0.2

Cuadro 1: Datos numéricos del incremento de *PCT* (Monte Carlo) tras individualización terapéutica basada en el conocimiento de la radiosensibilidad.

Para obtener resultados concretos tenemos todavía que aclarar algunos detalles. En principio, parece razonable tomar como x_r un valor igual al de la media de la distribución; sin embargo, debido a la incertidumbre en el proceso de determinación, adoptamos un criterio conservador (que supondrá no aumentar las dosis para muchos de los pacientes):

$$x_r = \bar{x} - 3\sigma = 1{,}56 \text{ rdc/Gy/200 Mpb.}$$

Por otra parte, tenemos que elegir un valor para x_m, de manera que, para los pacientes que lo superen, se busquen programas terapéuticos alternativos. Aquí tomamos $x_m = 2{,}96$ rdc/Gy/200 Mpb, que supone un 10 % de la población (para comparar con los resultados de otras elecciones, véase [37]).

La simulación se realiza con 10 series de 5.000 pacientes y para varias elecciones de los parámetros del modelo logístico de *PCT* que corresponden a la media y la desviación típica de la distribución normal que se emplea en las simulaciones. Los resultados se muestran en la Tabla 1 y en la Figura 15. Ésta última representa la proporción acumulada de los pacientes que alcanzan un determinado aumento en la *PCT*. Las incertidumbres corresponden a un intervalo de confianza del 95 %, aunque en la Figura no pueden apreciarse por ser menores que los símbolos correspondientes a los puntos.

Como se observa, el aumento de *PCT* depende mucho de la curva de dosis-respuesta de cada tumor en particular. En ciertos casos este aumento es grande y en promedio vemos que el 40 % del total de pacientes puede alcanzar más de un 10 % del mismo; para $D_{50} = 50$ Gy y $\gamma = 5$, un 4 % de los pacientes multiplican por 10 su probabilidad de control local. Aunque la información global que proporciona la simulación es semejante a la producida por métodos analíticos, la simulación permite examinar la respuesta de cada paciente en particular y no sólo los valores medios. Esto es importante por su consistencia con el propósito de la individualización. También, justifica el interés del método de Monte Carlo en este tipo de problemas, en los que todavía no es tan frecuente su uso como parecen indicar la sencillez y el carácter intuitivo del desarrollo que nos ha llevado hasta aquí.

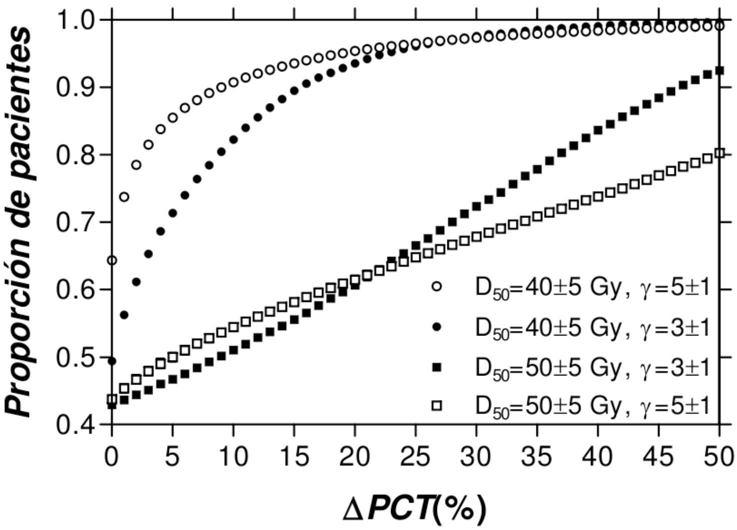


Figura 15: Resultados del cálculo del incremento de PCT obtenido por simulación (Monte Carlo) con diez series de 5.000 pacientes para varias elecciones de los parámetros (γ y D_{50}) del modelo logístico de PCT.

14 Estudio de correlación entre radiosensibilidad *in vitro* y efectos adversos

El primer párrafo de la hipótesis que planteábamos en la Sección 7 decía:

"La respuesta de los tejidos a la radiación está determinada por una componente genética; por tanto, es razonable sugerir que la radiosensibilidad de las células en cultivo es un reflejo de la constitución genética de los individuos de los que derivan."

Para comprobar la validez de esta hipótesis, a lo largo del desarrollo de varios proyectos de investigación hemos obtenido datos de radiosensibilidad sobre los linfocitos tomados de más de 250 pacientes afectas de cáncer de mama, hemos estudiado la frecuencia y la gravedad de los efectos adversos precoces y la gravedad y la duración del intervalo de tiempo entre el fin del tratamiento y la manifestación de efectos tóxicos tardíos en muchas de las pacientes incluidas en el estudio [29, 30, 35, 36].

Mediante la observación y el seguimiento clínico de las pacientes, hemos obtenido datos suficientes para clasificar la respuesta de la piel a la radioterapia y precisar la gravedad de la misma tanto en el período de respuesta precoz como tras el tiempo necesario para que la valoración clínica indique la magnitud de la respuesta tardía [30, 35].

Las complicaciones que más frecuentemente aparecen tras radioterapia por cáncer de mama son: eritema, descamación y descamación húmeda. En relación a los efectos secundarios agudos de la radioterapia, de acuerdo con el sistema de clasificación de la RTOG [38], hemos encontrado que aproximadamente el 13 % de las pacientes corresponden a lo que Burnet [23] denomina altamente radiosensibles. Los efectos adversos agudos suelen curar con rapidez sin necesitar de un tratamiento específico. Un numeroso grupo de estas mujeres han sido seguidas durante al menos 7 años y los efectos tardíos en la piel irradiada fueron medidos a lo largo de 7 meses (entre Diciembre de 2003 y Junio de 2004 [30, 36]). Los efectos tardíos progresan con el tiempo, no es posible detener su evolución, ni hacerlos regresar.

De nuestros resultados se deduce que no hay relación entre la gravedad de los efectos precoces y las manifestaciones de daño tardío observadas en la misma persona [30]. Tampoco nos ha sido posible demostrar la existencia de ningún tipo de relación entre el valor de radiosensibilidad *in vitro* medido como rupturas dobles de cadenas de ADN y la respuesta, precoz o tardía [30], evaluada clínicamente por observación de la piel incluida en el campo de tratamiento. Sin embargo, parece claro que existe una clara relación entre la severidad de las reacciones y la cantidad de tejido normal incluido en el volumen terapéutico.

Unos pocos días tras el término del tratamiento, las células incluidas en el volumen terapéutico sólo pueden actuar de tres maneras: i) crecer y dividirse (esta es la base de la curación de las lesiones agudas); ii) permanecer vivas sin entrar en proceso de división y iii) sobrevivir durante largo tiempo con importantes cambios genómicos y/o fenotípicos para desaparecer muy lentamente por apoptosis, o autofagia, y convertirse en un foco de estimulación crónica del sistema inmunológico que, de una manera aleatoria, a través de una función de probabilidad dependiente del tiempo, dé origen a las reacciones adversas tardías observables [30].

Quizás los resultados más esperanzadores de nuestro trabajo trabajo deriven del estudio de un grupo de pacientes afectas de carcinoma avanzado de mama que fueron tratadas exclusivamente con radioterapia radical (con fines curativos). Los objetivos del estudio fueron: i) cuantificar prospectivamente la importancia (grado) y el tiempo de inicio de los efectos adversos en la piel de las pacientes tratadas de acuerdo con un protocolo de hiperfraccionamiento, con escalación de dosis, de radioterapia y ii) valorar si el test de radiosensibilidad puede ser usado como factor predictivo de probabilidad de padecer reacciones adversas tardías de carácter grave en los tejidos normales incluidos en el volumen de tratamiento [35].

Para explorar la posibilidad de identificar variaciones en la radiosensibilidad de los linfocitos y relacionarlas con los cambios en el riesgo de morbilidad asociada a la radioterapia, se ajustaron los datos actuariales de probabilidad de efectos tardíos en función del tiempo al modelo mono-exponencial decreciente [39] usando la ecuación:

$$P(t) = 100\, e^{-k(t-t_{lag})}, \tag{19}$$

donde $P(t)$ es la probabilidad actuarial (en %) de estar libre de complicaciones, k es la pendiente de la línea que se obtiene de la representación, en forma semilogarítmica, de los valores experimentales, t es el tiempo trascurrido tras el tratamiento y t_{lag} es el intervalo ("lag") de tiempo necesario para que aparezca la primera complicación en el grupo de pacientes tratadas. El valor de t_{lag} también se obtiene del ajuste de los valores experimentales a la curva (19).

La Figura 16 **A** recoge los datos de seguimiento de todo el grupo. En ordenadas se representa la probabilidad de que las pacientes estén libres de padecer efectos tóxicos tardíos de máxima gravedad en función del tiempo tras el tratamiento. Los parámetros correspondientes al ajuste son: $r^2 = 0{,}948$; $P = 0{,}0001$; $k = 1{,}7 \pm 0{,}1\,\%$ por mes; $t_{1/2} = 42{,}0 \pm 2{,}0$ meses y $t_{lag} = 10{,}2 \pm 2{,}3$ meses. Aquí, r es el coeficiente de correlación, P es la significación estadística del ajuste y $t_{1/2}$ es el tiempo en el que la mitad de las pacientes tratadas desarrollan efectos tardíos severos en la piel, tras el tratamiento con radioterapia.

En el subgrupo de pacientes tratadas con la dosis de 81.6 Gy, estratificadas por el tamaño de la mama irradiada, hemos encontrado que existe relación entre la probabilidad de desarrollar efectos tardíos severos y el volumen de irradiación. Este resultado confirma nuestra observación anterior (Figura 16 **B**). Los parámetros correspondientes al ajuste son:

(a) Tamaño pequeño: $r^2 = 0{,}828$; $P = 0{,}0001$, $k = 1{,}9 \pm 0{,}2\,\%$ por mes; $t_{1/2} = 44{,}3 \pm 2{,}5$ meses; $t_{lag} = 19{,}3 \pm 3{,}5$ meses.

(b) Tamaño grande: $r^2 = 0{,}964$; $P = 0{,}0001$, $k = 1{,}6 \pm 0{,}2\,\%$ por mes; $t_{1/2} = 35{,}3 \pm 3{,}2$ meses, $t_{lag} = 6{,}3 \pm 3{,}2$ meses.

Entre las dos líneas de la Figura 16 **B** no aparecen diferencias significativas cuando se comparan las pendientes, pero sí cuando se comparan los valores de t_{lag}.

Los parámetros para el subgrupo de pacientes tratados con 81.6 Gy, estratificados por los valores de radiosensibilidad intrínseca (Fig. 16***C***), fueron:

1. Baja radiosensibilidad: $r^2 = 0{,}957$; $P = 0{,}0001$, $k = 2{,}4 \pm 0{,}2\,\%$ por mes; $t_{1/2} = 28{,}3 \pm 0{,}1$ meses; $t_{lag} = 16{,}0 \pm 2{,}7$ meses.
2. Alta radiosensibilidad: $r^2 = 0{,}902$; $P = 0{,}0001$, $k = 5{,}1 \pm 0{,}1\,\%$ por mes; $t_{1/2} = 13{,}4 \pm 0{,}1$ meses; $t_{lag} = 11{,}2 \pm 4{,}1$ meses.

Estos resultados sugieren que el riesgo de desarrollar efectos adversos severos depende de la dosis, del volumen de tejido irradiado y de la radiosensibilidad individual. El cociente entre las pendientes de pacientes con valores bajos y altos de radiosensibilidad proporciona un coeficiente al que hemos denominado [35] *factor de dosis biológica significativa* f_{DBS}:

$$f_{DBS} = \frac{k_a}{k_b} = \frac{5{,}1}{2{,}4} = 2{,}1. \tag{20}$$

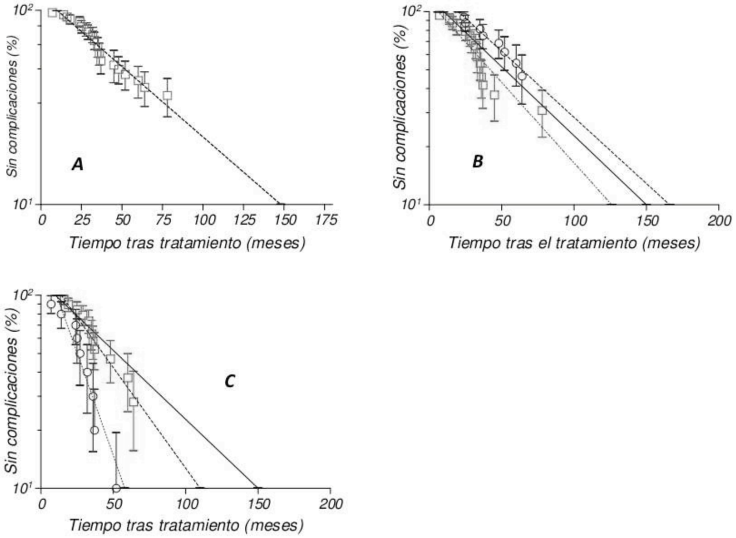


Figura 16: Ajuste de los datos actuariales de probabilidad de supervivencia sin efectos tóxicos tardíos en función del tiempo a la ecuación (19). En **A** se incluye la totalidad de los casos. En **B** los casos se subdividen en función del volumen de irradiación. Las líneas tienen la misma pendiente y difieren en el valor de t_{lag}. En **C** los casos se subdividen en función de la radiosensibilidad. Las líneas muestran diferencias en las pendientes.

Así, para la misma dosis y el mismo protocolo (dosis/fraccionamiento), el riesgo de desarrollar efectos adversos severos es doble en el grupo de pacientes con valores de radiosensibilidad superiores a 1.69 rdc/Gy/200 Mpb que en el grupo de pacientes con valores de radiosensibilidad inferiores a ese límite.

El coeficiente f_{DBS} que aquí proponemos es dependiente de tres variables: (a) la dosis total prescrita D_T; (b) el volumen de tejido incluido en el campo de irradiación V y (c) la radiosensibilidad intrínseca de las células de los tejidos normales x.

Para terminar, mencionaremos algunas conclusiones a las que nos han llevado los argumentos y experiencias que preceden:

1. La estimación cuantitativa del daño inicial inducido por la radiación sobre el ADN es un indicador biológico de la energía transferida al genoma.

2. La probabilidad de morbilidad tardía asociada a la radioterapia está directamente relacionada con este factor, al menos cuando la dosis

empleada en el tratamiento es elevada.

3. A día de hoy, el test de radiosensibilidad que hemos desarrollado está todavía lejos de poder ser usado con fiabilidad en radioterapia clínica.

Agradecimientos

A todos los que a lo largo del tiempo nos acompañaron en el desarrollo de esta línea de investigación. A los que aún siguen creyendo en este quehacer y continúan. A los que recientemente se han incorporado al grupo y, con su interés, mueven el nuestro. A los que dejaron de creer. A todos y a todas, gracias. A las enfermas de cáncer de mama que prestaron su conformidad para ser incluidas en el estudio. A Renato Álvarez-Nodarse por su interés y por su excelente trabajo como editor del manuscrito y también a Enrique Fernández Cara por su magnífico trabajo como editor final del manuscrito y por todos sus comentarios e indicaciones. Estos trabajos han sido posibles gracias a la financiación recibida a través del Programa de Acciones Integradas Hispano-Británicas (DGICYT Ref. HB-250, de la European Commission, Science Research Development, Biomedicine and Health Research Biomed 2, BMH4-CT96-0638) y del Ministerio de Educación y Ciencia a través de los proyectos SAF95/0778, PM97-0185, SAF 2001-3533 y SAF 2004-00889.

Referencias

[1] Singletary, S.E., et al., Staging system for breast cancer: revisions for the 6th edition of the AJCC Cancer Staging Manual. Surg Clin North Am, 2003, 83(4) p. 803-19.

[2] West, C.M., et al., Molecular markers predicting radiotherapy response: report and recommendations from an International Atomic Energy Agency technical meeting. Int J Radiat Oncol Biol Phys, 2005, 62(5) p. 1264-73.

[3] McMillan, T.J. and J.H. Peacock, Molecular determinants of radiosensitivity in mammalian cells. Int J Radiat Biol, 1994, 65(1) p. 49-55.

[4] Peacock, J.H., et al., Initial damage or repair as the major determinant of cellular radiosensitivity? Int J Radiat Biol, 1989, 56(5) p. 543-7.

[5] Robnett, T.J., et al., Factors predicting severe radiation pneumonitis in patients receiving definitive chemoradiation for lung cancer. Int J Radiat Oncol Biol Phys, 2000, 48(1) p. 89-94.

[6] West, C.M., R.M. Elliott, and N.G. Burnet, The genomics revolution and radiotherapy. Clin Oncol (R Coll Radiol), 2007, 19(6) p. 470-80.

[7] Kallman, P., A. Agren, and A. Brahme, Tumour and normal tissue responses to fractionated non-uniform dose delivery. Int J Radiat Biol, 1992, 62(2) p. 249-62.

[8] Nuñez, M.I., et al., DNA damage and prediction of radiation response in lymphocytes and epidermal skin human cells. Int J Cancer, 1998, 76(3) p. 354-61.

[9] Ruiz de Almodovar, J.M., et al., Dose-rate effect for DNA damage induced by ionizing radiation in human tumor cells. Radiat Res, 1994, 138(1 Suppl) p. S93-6.

[10] Steel, G.G., et al., The dose-rate effect in human tumour cells. Radiother Oncol, 1987, 9(4) p. 299-310.

[11] Steel, G.G. and J.H. Peacock, Why are some human tumours more radiosensitive than others? Radiother Oncol, 1989, 15(1) p. 63-72.

[12] Burnet, N.G., et al., Normal tissue radiosensitivity–how important is it? Clin Oncol (R Coll Radiol), 1996, 8(1) p. 25-34.

[13] Brock, W.A., et al., Fibroblast radiosensitivity versus acute and late normal skin responses in patients treated for breast cancer. Int J Radiat Oncol Biol Phys, 1995, 32(5) p. 1371-9.

[14] Ataman, O.U., et al., Audit of effectiveness of routine follow-up clinics after radiotherapy for cancer: a report of the REACT working group of ESTRO. Radiother Oncol, 2004, 73(2) p. 237-49.

[15] Peacock, J.H., et al., The nature of the initial slope of radiation cell survival curves. BJR Suppl, 1992, 24 p. 57-60.

[16] Peacock, J.H., et al., The intrinsic alpha/beta ratio for human tumour cells: is it a constant? Int J Radiat Biol, 1992, 61(4) p. 479-87.

[17] Curtis, S.B., Lethal and potentially lethal lesions induced by radiation–a unified repair model. Radiat Res, 1986, 106(2) p. 252-70.

[18] Fertil, B. and E.P. Malaise, Intrinsic radiosensitivity of human cell lines is correlated with radioresponsiveness of human tumors: analysis of 101 published survival curves. Int J Radiat Oncol Biol Phys, 1985, 11(9) p. 1699-707.

[19] Turesson, I., Individual variation and dose dependency in the progression rate of skin telangiectasia. Int J Radiat Oncol Biol Phys, 1990, 19(6) p. 1569-74.

[20] Burnet, N.G., et al., The relationship between cellular radiation sensitivity and tissue response may provide the basis for individualising radiotherapy schedules. Radiother Oncol, 1994, 33(3) p. 228-38.

[21] Turesson, I., et al., Prognostic factors for acute and late skin reactions in radiotherapy patients. Int J Radiat Oncol Biol Phys, 1996, 36(5) p. 1065-75.

[22] Safwat, A., et al., Deterministic rather than stochastic factors explain most of the variation in the expression of skin telangiectasia after radiotherapy. Int J Radiat Oncol Biol Phys, 2002, 52(1) p. 198-204.

[23] Burnet, N.G., et al., Describing patients' normal tissue reactions: concerning the possibility of individualising radiotherapy dose prescriptions based on potential predictive assays of normal tissue radiosensitivity. Steering Committee of the BioMed2 European Union Concerted Action Programme on the Development of Predictive Tests of Normal Tissue Response to Radiation Therapy. Int J Cancer, 1998, 79(6) p. 606-13.

[24] Qvarnstrom, O.F., et al., DNA double strand break quantification in skin biopsies. Radiother Oncol, 2004, 72(3) p. 311-7.

[25] Ruiz de Almodovar, J.M., et al., A comparison of methods for calculating DNA double-strand break induction frequency in mammalian cells by pulsed-field gel electrophoresis. Int J Radiat Biol, 1994, 65(6) p. 641-9.

[26] Ruiz de Almodovar, J.M., et al., Initial radiation-induced DNA damage in human tumour cell lines: a correlation with intrinsic cellular radiosensitivity. Br J Cancer, 1994, 69(3) p. 457-62.

[27] Contopoulou, C.R., V.E. Cook, and R.K. Mortimer, Analysis of DNA double strand breakage and repair using orthogonal field alternation gel electrophoresis. Yeast, 1987, 3(2) p. 71-6.

[28] Cook, V.E. and R.K. Mortimer, A quantitative model of DNA fragments generated by ionizing radiation, and possible experimental applications. Radiat Res, 1991, 125(1) p. 102-6.

[29] Mariano Ruiz de Almodovar, J., et al., Individualization of radiotherapy in breast cancer patients: possible usefulness of a DNA damage assay to measure normal cell radiosensitivity. Radiother Oncol, 2002, 62(3) p. 327-33.

[30] Lopez, E., et al., Early and late skin reactions to radiotherapy for breast cancer and their correlation with radiation-induced DNA damage in lymphocytes. Breast Cancer Res, 2005, 7(5) p. R690-8.

[31] Dickson, J., et al., Relationship between residual radiation-induced DNA double-strand breaks in cultured fibroblasts and late radiation reactions: a comparison of training and validation cohorts of breast cancer patients. Radiother Oncol, 2002, 62(3) p. 321-6.

[32] Loncaster, J.A., et al., Prediction of radiotherapy outcome using dynamic contrast enhanced MRI of carcinoma of the cervix. Int J Radiat Oncol Biol Phys, 2002, 54(3) p. 759-67.

[33] Routledge, J.A., et al., Evaluation of the LENT-SOMA scales for the prospective assessment of treatment morbidity in cervical carcinoma. Int J Radiat Oncol Biol Phys, 2003, 56(2) p. 502-10.

[34] West, C.M., et al., Lymphocyte radiosensitivity is a significant prognostic factor for morbidity in carcinoma of the cervix. Int J Radiat Oncol Biol Phys, 2001, 51(1) p. 10-5.

[35] Pinar, B., et al., Radiation-induced DNA damage as a predictor of long-term toxicity in locally advanced breast cancer patients treated with high-dose hyperfractionated radical radiotherapy. Radiat Res, 2007, 168(4) p. 415-22.

[36] Lopez, E., et al., Breast cancer acute radiotherapy morbidity evaluated by different scoring systems. Breast Cancer Res Treat, 2002, 73(2) p. 127-34.

[37] Guirado, D. and J.M. Ruiz de Almodovar, Prediction of normal tissue response and individualization of doses in radiotherapy. Phys Med Biol, 2003, 48(19) p. 3213-23.

[38] Cox, J.D., J. Stetz, and T.F. Pajak, Toxicity criteria of the Radiation Therapy Oncology Group (RTOG) and the European Organization for Research and Treatment of Cancer (EORTC). Int J Radiat Oncol Biol Phys, 1995, 31(5) p. 1341-6.

[39] Jung, H., et al., Quantification of late complications after radiation therapy. Radiother Oncol, 2001, 61(3) p. 233-46.

Bol. Soc. Esp. Mat. Apl.
n^{o}45(2008), 69–87

FREE SOFTWARE FOR NUMERICAL SIMULATION IN INDUSTRIAL PROBLEMS

G. FDEZ–MANíN, M. MEIS AND F. VARAS

Dpto. Matemática Aplicada II, Universidad de Vigo, Spain

manin@dma.uvigo.es marcos@dma.uvigo.es curro@dma.uvigo.es

Abstract

In this paper, some general ideas about free software and some specific comments (including some reflections concerning the potential benefits) on scientific computing free software are introduced, and a review of relevant free software codes in the domain of computer–aided engineering (CAE) is done. In particular, codes related to computer–aided design (CAD), mesh generation, data visualization, structural and termomechanical analysis, computational fluid dynamics (CFD) and multiphysics are reviewed.

Key words: *Free Software, Open Source Software, Scientific Computing, Numerical Simulation, Computer–Aided Engineering (CAE)*
AMS subject classifications: *65C20 68N30 68U07 68U20*

1 Some reflections on free software

1.1 What is free software?

Roughly speaking, *free software* is a kind of software that can be used, studied, and modified without restriction, and which can be copied and redistributed in modified or unmodified form either without restriction, or with minimal restrictions only to ensure that further recipients can also do these things.

According to the Free Software Foundation (see `http://www.fsf.org`), free software must guarantee that you, as a user of this software, have four freedoms:

- The freedom to run the program, for any purpose (freedom 0).

Fecha de recepción: 24/11/2008

A previous version of this presentation was distributed in the lecture notes of the XIII Spanish–French School Jacques–Louis Lions on Numerical Simulation in Physics and Engineering, held in Valladolid on 15–19 September 2008

- The freedom to study how the program works, and adapt it to your needs (freedom 1). Access to the source code is a precondition for this.
- The freedom to redistribute copies so you can help your neighbor (freedom 2).
- The freedom to improve the program, and release your improvements to the public, so that the whole community benefits (freedom 3). Access to the source code is a precondition for this.

A quite similar definition is adopted by the Open Source Initiative in the definition of *open source software* (see `http://www.opensource.org/`), that can be roughly defined as a kind of software for which the source code is made available under a copyright license that permits users to use, change, and improve the software, and to redistribute it in modified or unmodified form. In particular, the first three criteria stablished by the Open Source Definition read

1. Free Redistribution. The license shall not restrict any party from selling or giving away the software as a component of an aggregate software distribution containing programs from several different sources. The license shall not require a royalty or other fee for such sale.
2. Source Code. The program must include source code, and must allow distribution in source code as well as compiled form. Where some form of a product is not distributed with source code, there must be a well-publicized means of obtaining the source code for no more than a reasonable reproduction cost, preferably downloading via the Internet without charge.
3. Derived Works. The license must allow modifications and derived works, and must allow them to be distributed under the same terms as the license of the original software.

There is no very fundamental (technical) difference between these two definitions, the main disagreement being in the stress on the key benefit from this software development model: in the case of the Free Software Foundation the most important rationale for using free software is freedom, whereas for the Open Source Initiative this claim is the technical superiority of software developed under peer reviewed and transparent processes. In the remainder we will use the term *free software* instead of *open source software* since the use of the former one is much more spread.

On the other hand, any software that does not grant the above mentioned rights to the user will be called propietary software. In particular, according to this principle, *freeware* in binary form, software not allowing further distribution or software freely available exclusively for educational or non–profit research can not be considered free software (they are simply unusual forms of proprietary software).

Frequently, some confusion arises around the adjetive *free* in the expression *free software*. Usually, it is explained that *free* must be understood as in *free speech*, and not as in *free beer*, because, as stated above, free software is a matter of liberty not price: one may have paid money to get copies of free software, or he/she may have obtained copies at no charge. Fortunately, this ambiguity concerning the adjetive *free* is absent in French (*logiciel libre*) and Spanish (*software libre*) terms.

1.2 Some words on free software licencing

Even if it could be striking at first sight, most of free software is copyrighted. This is done through the licenses under which this software is released, that are intended to clearly state both the rights granted to the users and the limitations of the responsibility of the authors. At the same time, these licenses introduce some restrictions to assure due recognition to the authors, or to preserve the rights of future users.

It must be pointed out that software with the source code in the public domain represents one exception in this policy. Nevertheless, under the Berne Convention, every written work is automatically copyrighted and, as a consequence, a source code can not be considered in the public domain unless legal steps are taken to enforce this attribute (this makes software actually in the public domain rather uncommon).

There is a large number of licenses to distribute free software (see [1] and [2] for a survey on free software licensing). This *license proliferation* could lead to some compatibility problems when different free software codes are to be merged into a single one. In order to avoid these situations, both the Free Software Foundation (`http://www.fsf.org/licensing/licenses/`) and the Open Source Initiative (`http://www.opensource.org/licenses`) provide information about compatibility issues (as the same time, they recommend the use of some particular licenses).

Concerning the differences among the available licenses, the most important one is the distinction between the so called *permissive licenses* and *copyleft licenses*. Licences in the first group do not include any restriction for derivative works (for instance, free software released under any license of this type can be merged into proprietary software) whereas copyleft licenses impose restrictions on derivative works that are intended to preserve the rights of future users (in particular, derivative works must be distributed under the same conditions than the original work).

The most frequently used free software licenses are:

- GNU General Public License (GNU GPL)
- GNU Lesser General Public License (GNU LGPL)
- BSD style licenses

The GNU General Public License (`http://www.gnu.org/copyleft/`) is the most widely used copyleft license (and also the license adopted by the largest part of free software projects). The BSD (Berkeley Software Distribution) style licenses (`http://www.opensource.org/licenses/bsd-license.php`) are a family of permissive (and simple) licenses. On the other hand, the GNU Lesser General Public License is a specific license for libraries. This license is intended to be used in cases when linking (of the library) with propietary software is to be allowed.

At the same time, there are some specific licenses for software documentation. The most important one is the GNU Free Documentation License (GNU FDL). Under this license, the author grants the right to make copies of the document, to modify the document and to distribute it. Derivative works must also be licensed under the Free Documentation License.

In some sense, the use of the term *free software* constitutes an abuse of language. The use of the expression *software released under a free software license* should be a (much) more accurate way to refer to this kind of software. The reason is that *free* and *non–free* are not at all attributes of a piece of software but properties of a software license. In fact, any software can be released under a dual licensing (and this is the case for many software projects): for instance, a code could be relesead as free software under a GNU General Public License and, at the same time, the authors could offer to sell licenses to merge this code into a proprietary software (which is not allowed under a GPL license).

1.3 Free software in scientific computing

Progress of science strongly relies on the fact that scientific knowledge is not owned by the researchers themselves. Scientific findings and discoveries instead are expected to be published or communicated whatsoever so that other researchers could use them, granting due recognitizion to the authors of the original work.

At the same time, scientific research proceeds in a very open manner, where the publication of research findings is preceded by a peer review process. All of the assumptions, calculations and experiments that lead to the results are scrutinized before the findings are accepted by journals for publication. In fact, granting access to these details is essential in order to make possible that other researchers could extend or adapt the original work.

In the field of scientific computing (and, more precisely, in the numerical simulation) it seems quite natural to apply the principles above not only to algorithms but also to software but, in practice, the usual approach is quite different.

Concerning numerical algorithms and their analysis, researchers use to take ideas and results from other people as the basis for their own research, then they extend or adapt these results and finally they communicate their original achievements to the community (probably encouraging other researchers to use them).

Nevertheless, they rarely use the same approach for the software developed according to the corresponding algorithms. It is indeed unusual for scientific computing researchers to develop software absolutely from scratch but, in contrast, it is a very common approach to develop a code from in–house pieces of software (previously written by themselves) and a few low–level external libraries.

It must be pointed out that one can hardly imagine the equivalent scenario (to the approach of software development described above) in the case of numerical algorithms (a researcher constructing a complex algorithm without making use of any known numerical method) or their numerical analysis (a researcher studying the properties of a numerical method exclusively from in–house convergence results and ignoring theoretical contributions coming from any other people).

On the other hand, since software development is highly time–consuming, it must be made by a skilled programmer and the development of software itself is not (frequently) a recognized merit in the (applied mathematics) academia, researchers are little encouraged to face very challenging projects of industrial interest.

Alternatively, the usual scientific research principles could be applied to the software development. Under this alternative model, a first step would consist in a review of the state–of–the–art through existing (free) software, then researchers would adapt/extend this software and, finally, they would release the derivative works (encouraging other researchers to use the new software). There would be some foreseen benefits associated to this approach. First, a state–of–the–art numerical simulation software (developed by the whole scientific community) should be freely available to any researcher. Also researchers productivity would be highly improved (programming workload would be strongly alleviated since it should be reduced to the programming of the new features). As a result, those challenging industrial projects would be much more affordable and innovation could be strongly boosted.

Scientific computing software has always been present in the history of free software [3],[4]. All together, the very origin of free software (the seminal work of Richard M. Stallman at the beginning of the eighties in the MIT's Artificial Intelligence Laboratory) is not far from scientific computing, there are examples of scientific computing free software since the early times (as, for instance, `FreeFEM` family predecessors `MacFem` and `PCfem` written by O. Pironneau in the mid-eighties), and (free software) TEX and LATEX compilers or `LAPACK` libraries are part of the daily life of any member of the scientific computing community.

Concerning scientific computing free software development, there are a number of different actors. There are still, as in the origins, individuals who write free software. But at the same time, teams from universities, research institutes and companies are now involved in writing scientific free software.

Development of free software computer-aided engineering (CAE) tools in France is a good example of the diversity of sources of free sofware, as we can find free software CAE tools developed by universities or higher education centers (Laboratoire Jean Kuntzmann Grenoble, Université Blaise Pascal, INSA Toulouse, Université Paris VI, among others), research institutions (INRIA, CEA) and companies (Electricité de France, SDTools, CEDRAT).

On the other hand, there are two alternative models in free software development [5]:

- the cathedral model: following this model, software is built like cathedrals, carefully crafted under the strict supervision of one (or very few) individual, with no beta version to be released before its time

- the bazaar model: a highly decentralized one, used very sucessfully in the development of Linux, with a large number of developers in which the software is early and often released

Quite curiously, up to now, almost every scientific computing free software falls in the first category. Of course, scientific computing software has some specific features, but one wonders about the possibilities of speeding up scientific free software development by adopting a bazaar model (with a large number of individual developers incorporating/improving pieces of software).

Finally, even if diversity could be a value in the free software movement, there seem to be good reasons in considering the possibility to interconnect some projects. As it will be mentioned, this is already being done for a CAD tool (`Salome`) and a FEA code (`Code_Aster`) in order to develop a complete CAE environment (`Salome-Meca`).

2 A review of free software CAE tools

Inside the (very) large field of scientific computing free software, and from the perspective of the numerical simulation of problems of industrial interest, we will focus on CAE–related tools. As a consequence, lower level software (such as linear/non linear system solvers, optimization software, eigensolvers, etc.) will no be considered.

There are many web pages reviewing or listing scientific computing free software. The Free Software Foundation `http://directory.fsf.org/` has a (very) short list of free software in this field. More complete compilations are mantained by individuals or teams as, for instance, those appearing in the following links:

- `http://norma.mas.ecp.fr/wikimas/ScientificComputingSoftware`

  contains a scientific computing software list mantained by Florian De Vuyst (Laboratoire de Mathématiques Appliquées aux Systèmes, École Centrale Paris, France).

- `http://www.ann.jussieu.fr/~lehyaric/freesoft/free.htm`

  presents a list of free numerical analysis software compiled by Antoine Le Hyaric (Laboratoire Jacques-Louis Lions, at the Université Paris VI, France).

- `http://homepage.usask.ca/~ijm451/finite/fe_resources/`

is a page written by Roger Young (Industrial Research Ltd, New Zealand) and Ian MacPhedran (Engineering Computer Center, University of Saskatchewan, Canada) that describes finite element resources in the Internet, including free software

- `http://www.cfd-online.com`

  is an on-line center for Computational Fluid Dynamics that has many links to CFD and visualization software

Additionally, the free software contest *Les Trophées du Libre* has stablished in 2007 a category for scientific software and its web page (`www.trophees-du-libre.org/`) contains information on some scientific computing free software codes.

In the following subsection, we will present examples of free software tools in three domains:

- CAD, mesh generation and visualization tools
- dedicated CAE tools
- multiphysics CAE tools

2.1 CAD, meshing and visualization tools

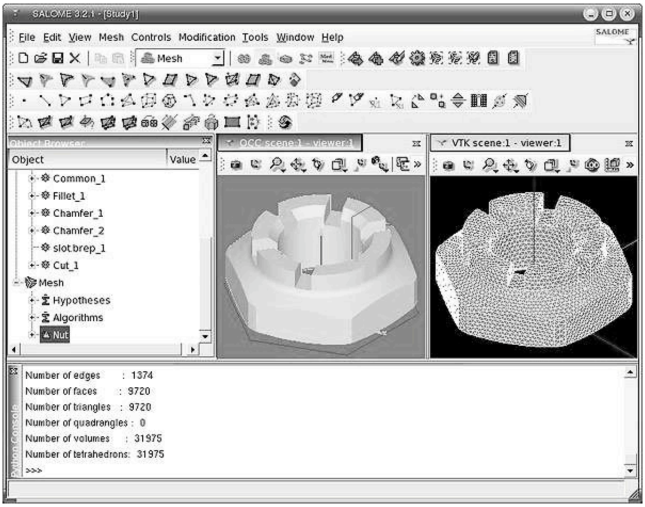
Table 1 lists some relevant CAD, meshing and visualization tools available as free software.

Name	URL	Comments
`Gmsh`	`wwww.geuz.org/gmsh`	2D/3D meshing tool
`MayaVi2`	`mayavi.sourceforge.net`	data visualizer
`Netgen`	`www.hpfem.jku.at/netgen`	3D meshing tool
`ParaView`	`www.paraview.org`	data visualizer
`Salome`	`wwww.salome-platform.org`	2D/3D CAD

Table 1: List of some CAD, meshing and visualization tools

Concerning CAD tools, only one code (`Salome`) has been including in the list. There exist another free software CAD tools (such as `BRLCAD`, available at `www.brlcad.org`, or `gCAD3D`, available at `www.gcad3d.org`), but their kernels are based on constructive solid geometry (CSG) techniques, or their graphical interfaces are not well developed and they are poorly documented.

`Salome`, instead, is a very complete CAD system. The program is developed by a large consortium of companies (Electricité de France R&D, Bureau Veritas, Open Cascade, Principia R&D and CEDRAT) and private and public laboratories (EADS CCR, CEA, Laboratoire d'Informatique Paris VI and Laboratoire d'Electrotechnique de Grenoble). It is released under GNU GPL license.



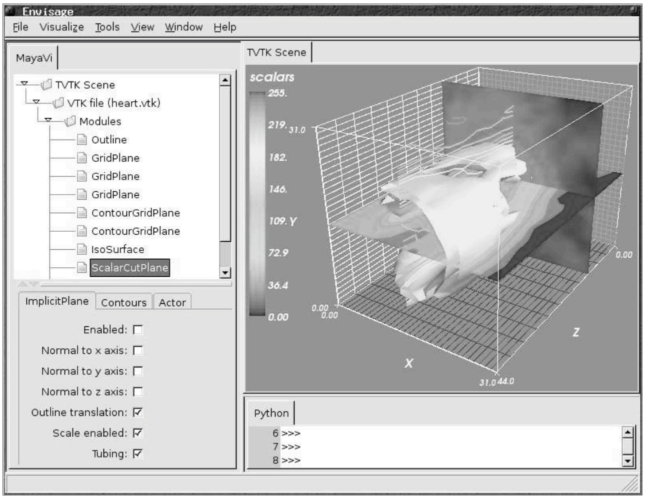
```
Number of edges       :  1374
Number of faces       :  9720
Number of triangles   :  9720
Number of quadrangles :  0
Number of volumes     :  31975
Number of tetrahedrons:  31975
>>>
```

Figure 1: Screenshot of `Salome` software.

`Salome` is intended to provide a generic platform for pre- and post-processing in numerical simulations. It provides some modelling and meshing utilities (such as geometry reparation and mesh quatility control) and is well integrated with some free meshing tools (as `Mefisto` and `Netgen`) and finite element analysis (FEA) tools (such as the solid mechanics analysis program `Code_Aster`, presented below and integrated with `Salome` in the suite `Salome-Meca`; see `www.caelinux.com` for more information). Finally, `Salome` allows Python scripting, which results in a very interesting feature since Python provides a universal language for scripting in scientific applications and it is very well suited (through numpy and scipy packages) for (large) scientific computing problems.

Concernig mesh generators, two free software codes have been selected: `Gmsh` (developed by Ch. Geuzaine, Université de Liège, Belgium, and J.F. Remacle, Université catholique de Louvain, Belgium, and released under the GNU GPL license) and `Netgen` (developed by J. Schöberl, H. Gerstmayr and R. Gaisbauer, Jonahhes Kepler University, Austria, and released under the GNU LGPL license). Although both include simple CAD tools to generate meshes from scratch (the second one, at the same time, can be easily integrated in `Salome`), they are intended to be used importing geometry files from CAD systems in STEP, IGES or BREP formats. In addition, `Gmsh` allows parametric input through its own scripting language.

Two visualization tools have been included in table 1: `MayaVi2` (deleloped by Entought Scientific Computing Solutions, USA, and released under a BSD-style license) and `Paraview` (developed through the collaboration of the companies Kitware and CSimSoft with the Sandia National Laboratories, Los Alamos National Laboratories and the Army Research Laboratories in USA, and



Enabled: □
Normal to x axis: □
Normal to y axis: □
Normal to z axis: □
Outline translation: ☑
Scale enabled: ☑
Tubing: ☑

```
6 >>>
7 >>>
8 >>>
```

Figure 2: Screenshot of Mayavi2 software.

released under a BSD-style license). They both use the Visualization Toolkit (VTK) and provide easy–to–use graphical user interfaces as well as they allow scripting using Python language. Paraview was specifically developed to deal with parallel solvers of large scale problems, so that it has a good support for distributed computing. Mayavi2 in turn can be used as a library and easily embedded into other applications.

There are some other free software visualization tools such as Mayavi (mayavi.sourceforge.net, the precursor of Mayavi2), VisIt (wci.llnl.gov/codes/visit/, a code similar to Paraview) and OpenDX (www.opendx.org/, a more general scientific visualization tool).

An interesting additional tool, needed in distributed computing, is a mesh partition utility. There is (at least) one such tool released as free software: the program Scotch developed by F. Pellegrini, Université de Bordeaux, France, and available at www.labri.fr/perso/pelegrin/scotch. Scotch is a software package and libraries for graph, mesh and hypergraph partitioning, static mapping, and parallel and sequential sparse matrix block ordering, released under the CeCILL-C license (a free software license adapted to French law). Scotch has also a library that provides compatibility with the popular mesh partitioning tool Metis, a family of programs for partitioning graphs and hypergraphs and computing fill–reducing orderings of sparse matrices (see glaros.dtc.umn.edu/gkhome/views/metis). Metis itself is not a free software as it can be freely used but redistributed only under certain conditions.

2.2 Dedicated CAE tools

In this subsection, and for the sake of brevity, we will concentrate on structural analysis and computational fluid dynamics (CFD) tools. Anyway, there are free software tools in many other fields (such as acoustics, electromagnetics, metheorology/forecasting, oceanography, molecular dynamics or atomic modelling) that are omitted in this short review.

Table 2 contains a list of free software codes for numerical simulation in solid mechanics, including structural, thermomechanical and dynamical analyses. In the first (loose) category, *structural analysis*, we have included those codes mainly designed to undertake static or vibration analysis of structures: `Calculix` (developed by G. Dhondt and K. Wittig, from MTU Aero Engines, Germany, and released under the GNU GPL license), `FELyX` (developed by EVEN Evolutionary Engineering, Switzerland, and also released under the GNU GPL license) and `WARP3D` (developed by the Department of Civil Engineering, University of Illinois at Urbana-Champaign, USA, and released under the GNU GPL license). `Calculix` and `FELyX` could be considered as libraries to manipulate the finite element method (FEM) matrices arising from the discretization of the corresponding structures, and able to solve some structural problems. They also provide support for importing geometries and meshes from some external codes (including several widely used propietary codes). `WARP3D` is a specialized code in static and dynamic analysis of fracture problems (able to deal with finite strain formulations, crack growth modelling or stress intensity factors) with parallel capabilities.

Name	URL	Comments
`Calculix`	`www.calculix.de`	structural anal.
`Code_Aster`	`www.code-aster.org`	thermomech.
`FELyX`	`felyx.sourceforge.net`	structural anal.
`Impact`	`impact.sourceforge.net`	dynamic anal.
`Tahoe`	`tahoe.ca.sandia.gov`	thermomech.
`WARP3D`	`cern49.cee.uiuc.edu/cfm/warp3d.html`	structural anal.

Table 2: List of structural analysis tools

In the (equally loose) group *thermomechanical analysis*, two codes capable to carry out complex analyses in solid mechanics, including, for instance, thermal coupling, plasticity or contact problems have been included: `Code_Aster` and `Tahoe`.

`Code_Aster` (developed as an in–house software in Electricité de France and released under the GNU GPL license since 2001) is a very complete end user program, ready to afford real industrial problems, which is presently being integrated with the CAD software `Salome` into a complete CAD/CAE platform known as `Salome-Meca`. In addition to termomechanical computations (including fracture, damage or fatigue, and with a wide range of constitutive laws) `Code_Aster` can also perform some multiphysics analysis (including drying and hydratation, metallurgy analysis or fluid–structure interactions). The program also includes a large library of elements (including X-FEM techniques) and allows Python scripting.

In turn, Tahoe (developed by the Sandia National Laboratory, USA, and released under a BSD-style license) is a research-oriented, open source platform for the development of numerical methods and material models for the analysis of complex problems in solid mechanics (including fracture or failure, interfacial adhesion and debonding, shear banding, length-scale dependent elasticity and plasticity, and deformation in small-scale structures). The code includes meshfree simulation tools with particle methods and supports parallel execution. Unfortunately, some parts of the documentation are not updated.

Finally, Impact (developed by J. Forssell, Sweden, and many others, and released under the GNU GPL license) is a small (Java) implementation of an explicit dynamic finite element program (based on the ideas of the well–known Dyna3D code developed by J.O. Hallquist at the Lawrence Livermore National Laboratories) to solve problems involving large deformations and high velocities (the kind of problems arising, for instance, in crashes or metal forge). Impact also incorporates some pre- and post–processing utilities as well as some parallel capabilities. On the other hand, the program is poorly documented, there is only a beta version and the project activity is quite low.

The table 3 shows some CFD codes based both on finite element and finite volume schemes.

Name	URL	Comments
Code_Saturne	www.code-saturne.org	general purpose
FEATFlow	www.featflow.de	incompr. NS
Gerris	gfs.sourceforge.net	incompr. Euler/NS
OpenFOAM	www.opencfd.co.uk/openfoam	general purpose

Table 3: List of CFD tools

Code_Saturne (developed by Electricité de France as an in-house program and released as free software under the GNU GPL license in 2007) is a powerful general purpose industrial CFD code capable to afford combustion, multiphase flow, radiation phenomena or magnetohydrodynamics. Code_Saturne has also parallel capacities and allows importation/exportation operations with many other codes (such as I-DEAS, Gmsh, Gambit, Salome or EnSight). Code_Saturne can be easily linked to Code_Aster (notably through Salome) to solve fluid-structure interaction problems.

OpenFOAM (developed by the software and consulting company OpenCFD, UK, and released under the GNU GPL license) is a finite volume toolbox to treat complex fluid problems. The code includes different models of turbulence and solvers for compressible flow, multiphase flow, and combustion. It also includes pre- and post-processing capabilities and is able to perform parallel computations.

On the other hand, FEATFlow and Gerris are more specific codes. FEATFlow (developed by S. Turek and co–workers, Universität Dortmund, Germany, and released under a BSD-style license) is a finite element incompressible Navier Stokes solver. FEATFlow is planned to be superseded by FEAST, a more general finite element solver (see www.feast.uni-dortmund.de for more information). Gerris (developed by S. Popinet, National Institute of Water and Atmospheric Research, New Zealand, and released under the GNU GPL license) is a finite volume solver for incompressible Euler and Navier-Stokes equations (as well as coupled advection–diffusion transport equations). The code also includes some mesh generation and adaptive mesh refinement capabilities. Both codes, FEATFlow and Gerris, have parallel capabilities, which are necessary to afford realistic 3D computations in industrial problems.

2.3 Multiphysics CAE tools

A list of some free software codes with multiphysics simulation capabilities is presented in table 4. In spite of the looseness of the clasification, we have tried to divide existing codes in three categories:

- end user programs, with a collection of solvers and the possibility to easily enlarge this collection with additional ones
- finite element (FE) environments, with a dedicated language or suitable macros to describe (variational) formulations at a high level
- libraries of FE utilities, requiring to write some middle (or low) level code to solve a specific problem

It must be stressed that due to the flexibility of the finite element method solving multiphysics problems (as compared with other techniques) the review has been limited to finite element codes. This means that other interesting tools not based on finite element techniques (for instance the finite volume based code FiPy; see www.ctcms.nist.gov/fipy) have been disregarded.

According to that list, only one code (Elmer) can be considered as an end user program. For this reason, the features of this code will be discused in some detail in the next section.

FEniCS, FreeFEM family (including FreeFEM++ and FreeFEM3D) and Rheolef, are cataloged as finite element environments since they provide a dedicated language or some kind of objet–oriented programming support to describe the set of partial differential equations to be solved.

FreeFEM family (developed by O. Pironneau and co–workers, Université Paris VI, France, and distributed under the GNU GPL license) and Rheolef (developed by P. Saramito and co–workers, Laboratoire Jean Kuntzmann, Grenoble, and also released with the GNU GPL license), codes provides a language that allows an easy definition of the (variational) formulation of the problem to be solved and a direct control on the discretization scheme. As a consequence, they constitute excellent tools is research and teaching. However,

Name	URL	Comments
ALBERTA	www.alberta-fem.de	FE library
Deal.II	www.dealii.org	FE library
Elmer	www.csc.fi/english/pages/elmer	end user program
FEniCS	www.fenics.org	FE environment
FreeFEM	www.freefem.org	FE environment
GetDP	geuz.org/getdp	FE library
GetFEM++	home.gna.org/getfem	FE library
LibMesh	libmesh.sourceforge.net	FE library
OFELI	www.ofeli.net	FE library
OOFEM	www.oofem.org	FE library
Rheolef	ljk.imag.fr/membres/Pierre.Saramito	FE environment

Table 4: List of multiphysics codes

application to industrial problems on complex (3D) geometries could be very tricky (see, for instance, [6]).

The FEniCS project appears as a very promising one with a really challenging goal: the programming of a complete set of tools to automate the solution of partial differential equations (see [7] concerning different approaches to the automatization of the finite element method). The development team involves a large number of laboratories, including University of Chicago, Argonne National Laboratory, Delft University of Technology, Royal Institute of Technology KT, Simula Research Laboratory, Finite Element Center, Texas Tech University, and University of Cambridge. Some parts are released under the GNU GPL license whereas other parts are distibuted with the GNU LGPL license. The development so far has been focused on core components programming (in particular, libraries and compilers of variational forms) whereas completion of easy–to–use interfaces have received significantly less attention. In addition, the documentation is (al least at this moment) very incomplete.

The remaining codes require writing some middle level code to define the problem to be solved and the numerical discretization to be used.

The statement above is not exactly true for the code GetDP (developed by P. Dular and C. Geuzaine, Université de Liège, Belgium, and distributed under the GNU GPL license), which in some sense is similar to FreeFEM family or Rheolef. Incidentally, it can also be remarked that this solver is easily linked to the mesh generation code Gmsh.

On the other hand, some other codes, as the library GetFEM++ (developed by Y. Renard, INSA Lyon, France, and J. Pommier, INSA Toulouse, France, released under the GNU LGPL license) provide interfaces to Python to facilitate the access to the finite element library. GetFEM++ tries at the same time to provide a flexible framework to define the numerical scheme used to discretize the problem.

ALBERTA (developed by A. Schmidt, Universität Bremen, Germany, and co–workers and released under the GNU GPL license) is described as an adaptive hierarchical finite element toolbox providing hierarchical meshes, routines for mesh adaptation, and the complete administration of finite element spaces and the corresponding degrees of freedom during mesh adaption.

In a similar manner, LibMesh (developed at The University of Texas at Austin, USA, with contributions from Technische Universität Hamburg-Harburg, Germany, Sandia National Laboratories, USA, and NASA Lyndon B. Johnson Space Center, USA, released under the GNU LGPL license) is a finite element library (based on object oriented programming) providing support for adaptive mesh refinement computations in parallel. Deal.II (developed by W. Bangerth, Texas A&M University, USA, R. Hartmann, German Aerospace Center (DLR), Germany and G.Kanschat, Universität Heidelberg, Germany, and released under the Q Public License, which is a scarcely used licence of non–permissive type) is also a general purpose finite element library using object oriented programming and including adaptive techniques.

On the other hand, OFELI (developed by R. Touzani, Université Blaise Pascal, France, and released under the GNU GPL licence) is a framework of C++ classes for the development of finite element programs. Quite similarly, OOFEM (developed by B. Patzak and some others and distributed with the GNU GPL license) is an Object Oriented Programming finite element library with some application modules (such as modules for structural mechanics or incompressible flows)

It must be observed that the activity level in the development of some of the listed codes can be rather low. For instance, the latest versions of Rheolef and GetDP codes were released in 2006.

3 Elmer software suite

Elmer is a multiphysics free sofware developed by the CSC (the Finish IT Center for Science, a non–profit company providing IT support and resources for academia, research institutes and companies which is administered by the Finish Ministry of Education) since 1995. Elmer was released under the GNU GPL license in 2005.

Originally designed as a CFD code, Elmer evolved into a general purpose finite element package, able to deal with a set of partial differential equations which may be coupled in a generic manner. Elmer includes physical models of fluid dynamics, structural mechanics, electromagnetics, heat transfer and acoustics, among others. Elmer solvers library can be enlarged since user defined equation solvers can be (easily) added. At the same time, in order to tackle large scale problems, Elmer provides support for distributed computing (this support can also be used to take advantage of the multicore capabilities of current personal computer microprocessors).

Although Elmer software has limited mesh generation and visualization capabilities, it provides a good support for importing and exporting external format files as it will be indicated below.

Elmer software has a complete and updated documentation available at the web page of the project (www.csc.fi/english/pages/elmer). The set of documents include detailed tutorials. At the same time, a good support is provided through the discussion mailing list. This list is intended to be used to ask questions about compiling, running simulations, giving bug reports or feature wishes, and questions are very quickly answered by the developers or (less frequently) by advanced users.

3.1 Elmer modules

The core of the Elmer software suite is composed by the following programs:

- Mesh manipulation utility ElmerGrid

  Apart from generating very simple structured meshes, ElmerGrid is designed to be used for mesh manipulation operations as importing meshes generated by external programs or performing mesh partitions for parallel Elmer runs. Mesh formats that can be imported in Elmer include file formats of Gmsh, COMSOL Multiphysics, ANSYS, Abaqus and Ideas, among others. In addition, some codes (as Netgen and GiD) can save mesh files in native Elmer format.

- Processor ElmerSolver

  ElmerSolver, the finite element solver of the suite, uses as input a mesh in native Elmer format and an (ASCII) input file that describes in a quite organized manner the problem to be solved and the parameters to be used in the numerical solution.

At the same time, the Elmer software suite include some graphical environments

- Graphical preprocessor ElmerFront

  ElmerFront is a graphical interface intended to provide an easy (but limited) access to the solver for simple problems.

- Postprocessor ElmerPost

  The postprocessor ElmerPost is not intended to be a powerful postprocessing tool but simply to fulfill some basic postprocessing needs. Instead, ElmerSolver is able to generate output VTK files to be processed by external data visualization tools (as MayaVi, MayaVi2 or Paraview).

Not actively developed during some time, ElmerFront and ElmerPost have been (very) recently superseded by ElmerGUI. In contrast to its predecessors, ElmerGUI is a customizable GUI aimed to provide more complete access to Elmer features. For instance, ElmerGUI

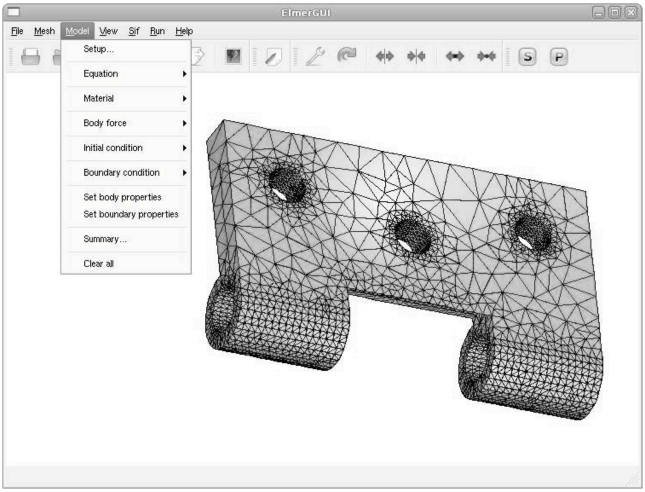


Figure 3: Screenshot of `ElmerGUI`.

- includes some (basic) CAD operations through Open CASCADE Technology (see `www.opencascade.org`)
- allows mesh import and basic mesh manipulation operations through `ElmerGrid`
- allows other more complex mesh manipulation operations since plugins can be compiled for some meshing codes as `Netgen` (and non–free codes as `TetGen`, released under the MIT-License)
- has a material library support to specify material properties in the solvers
- provides a friendly postprocessing VTK-based tool
- can control parallel computation specifying the number of processors to be used

3.2 Overview of ELMER numerical capabilities

`ElmerSolver` includes a large set of programmed solvers. The list contains solvers for many classical models from fluid dynamics (as Navier-Stokes -including free surface problems-, heat transfer -including phase change and simple radiation models- or transport models associated to advection-diffusion equations), electromagnetics (electrostatics, magnetostatics, and induction or eddy currents problems), structural analysis (linear elasticity, elastic plates) and acoustics (Helmholtz equations, including BEM solvers). `ElmerSolver` also includes less standard solvers as electrokinetics or quantum mechanics solvers as well as support for level–set methods, arbitrary Lagrangian-Eulerian (ALE) formulations, mesh adaptation or operator splitting.

`ElmerSolver` can easily handle coupling bewtween two (or more) models. In particular, most part of standard couplings between classical models (such as bouyancy forces or Joule heating) are already included in the corresponding solvers. Other coupling terms in existing solvers sould be programmed by user defined functions (that can be easily invoked by `ElmerSolver`). On the other hand, coefficients of any solver can be made dependent on the solution of any other solver. Its dependency should be described through a dedicated language called MATC (alternatively, it can also be described by a compiled function).

Coupled models (both in steady and time–dependent problems) in `ElmerSolver` are solved through segregated iterations. Even if this (Picard) strategy could imply sometimes slow convergence in steady problems (when compared with Newton iteration on the global problem), it also appears as the only effective possibility in large scale problems solved with distributed computing.

New solvers can be easily added either written from scratch (some examples are included in the software documentation) or, more likely, adapting them from the existing ones. In any case, only high–level operations must be programmed since the `Elmer` library provide all the low–level subroutines.

Concerning the numerical techniques implemented in the code

- `ElmerSolver` has a library of finite elements that includes (arbitrary) higher–order finite elements (with side, face and internal modes)

- stabilization techniques in `ElmerSolver` include Streamline Upwind Petrov–Galerkin (SUPG) stabilization, discontinuous Galerkin and residual–free bubbles

- `ElmerSolver` performs time integration with Backward Differentiation Formula (BDF) or Crank-Nicolson scheme (the Bossak method can also be used for second order problems). Adaptive time–steping is available for BDF schemes.

- concerning linear systems solvers, `ElmerSolver` includes

  - direct solver based on the unsymmetric multifrontal method (`ElmerSolver` incorporates `UMFPACK` package)
  - Krylov iterative solvers, including conjugate gradient, biconjugate gradient stabilized and generalized minimal residual (GMRES)
  - algebraic and geometric multigrid methods

  Krylov solvers preconditioning can be done with classical iterative methods, incomplete LU factorization and algebraic or geometric multigrid. Also, some tools are included to perform block preconditioning. Unfortunately, for the time being, multigrid methods can not be used as preconditioners in parallel computations. In any case, external solvers can also be invoked from `Elmer`, so that other external multigrid preconditioners can be used. Scalable preconditioners in `Hypre` (a library for solving large, sparse linear systems of equations on massively parallel computers, see `acts.nersc.gov/hypre`) could be an interesting alternative.

- to solve (global) nonlinear systems arising in coupled problems `ElmerSolver` uses, as already said, Picard iterations; inside each solver, instead, Newton is usually programed. In any case, to gain some robustness, Newton iteration can be switched to Picard iteration at the first iterations and some (fixed) dumping can be incorporated to the Newton iteration. Of course, this algorithm can be significatively improved by modifying the choice of the descent direction and incorporating a line search strategy but this strategy must be programmed (in the corresponding solver).
- some matrix manipulation utilities (mainly concerning boundary conditions and nodal loads) are included in `ElmerSolver`

4 Final remarks

Some conclusions about the development of scientific free software and the use of numerical simulation free software in industrial applications can be drawn from the analysis above:

- there exists a huge potential in the numerical analysis community to build free state–of–the–art scientific computing software; to unfold this potential, it suffices to extend to software the same *full disclosure* principles the community uses for algoritms;
- development of free scientific software through a *bazaar model* seems to be an almost unexplored but promising software development model;
- free software model in scientific computing provides tools to largely enhance researchers productivity, as a large body of software can be available to easily implement new ideas and algorithms on it;
- after a review of avaliable numerical simulation free software, one can conclude that this category of free software is now mature enough to tackle industrial problems (in this sense, the `Elmer` software suite consitutes an excellent example of this maturity);

Acknowledgements

This work was partially funded by the Centro de Supercomputación de Galicia (CESGA) by means of the research contract “Software libre para a simulación numérica na empresa” aimed to review free software projects related to numerical simulation of industrial problems. The project team, led by

Alfredo Bermúdez de Castro, also included (in addition to the authors and in alphabetical order) Ana Ferreiro, José Antonio García, María Cristina Naya, Peregrina Quintela, Diana Rivas, María Teresa Sánchez and Carlos Vázquez. Some parts of this presentation have largely benefited from the final report of this project.

The research of the authors is also partially supported by the Spanish Gouvernment under Grants BIA2006-14244 (G.F.-M.) and MTM2007-67596-C02-02 and CSD2006-00032 (M.M. and F.V).

The final form of this document has also benefited from some sugestions of J. Rafael Rodríguez-Galván who read a first draft of the manuscript.

References

[1] A. M. St. Laurent, *Understanding Open Source and Free Software Licensing*, O'Reilly, Sebastopol, 2004.

[2] S. L. Chen, *Free/Open Source Software. Licensing*, United Nations Development Programme – Asia-Pacific Development Information Programme (UNDP-APDIP) & Elsevier, New Delhi, 2006.

[3] C. DiBona, S. Ockman, M. Stone (Eds.), *Open Sources. Voices from the Open Source Revolution*, O'Reilly, Sebastopol, 1999.

[4] R. M. Stallman, *Free Software, Free Society: Selected Essays of Richard M. Stallman*, Free Software Foundation, Boston, 2002.

[5] E. R. Raymond, *The Cathedral and the Bazaar*, O'Reilly, Sebastopol, 1999.

[6] S. Del Pino, B. Maury, *2D/3D Turbine Simulations with FreeFEM* in *Numerical Analysis and Scientific Computing for PDEs and their Challenging Applications*, J. Haataja, R. Stenberg, J. Periaux, P. Raback and P. Neittaanmaki (Eds.), CIMNE, Barcelona, 2007.

[7] A. Logg, *Automating the Finite Element Method*, Archives of Computational Methods in Engineering, **14(2)** (2007), 93–138.

Bol. Soc. Esp. Mat. Apl.
n^{o}45(2008), 89–145

SPLITTING AND COMPOSITION METHODS IN THE NUMERICAL INTEGRATION OF DIFFERENTIAL EQUATIONS

SERGIO BLANES[1] FERNANDO CASAS[2] AND ANDER MURUA[3]

[1]Instituto de Matemática Multidisciplinar, Universidad Politécnica de Valencia,
E-46022 Valencia, Spain.
[2]Departament de Matemàtiques, Universitat Jaume I, E-12071 Castellón, Spain.
[3]Konputazio Zientziak eta A.A. saila, Informatika Fakultatea, EHU/UPV,
Donostia/San Sebastián, Spain.

serblaza@imm.upv.es Fernando.Casas@uji.es Ander.Murua@ehu.es

Abstract

We provide a comprehensive survey of splitting and composition methods for the numerical integration of ordinary differential equations (ODEs). Splitting methods constitute an appropriate choice when the vector field associated with the ODE can be decomposed into several pieces and each of them is integrable. This class of integrators are explicit, simple to implement and preserve structural properties of the system. In consequence, they are specially useful in geometric numerical integration. In addition, the numerical solution obtained by splitting schemes can be seen as the exact solution to a perturbed system of ODEs possessing the same geometric properties as the original system. This backward error interpretation has direct implications for the qualitative behavior of the numerical solution as well as for the error propagation along time. Closely connected with splitting integrators are composition methods. We analyze the order conditions required by a method to achieve a given order and summarize the different families of schemes one can find in the literature. Finally, we illustrate the main features of splitting and composition methods on several numerical examples arising from applications.

1 Introduction by examples

The basic idea of splitting methods for the time integration of ordinary differential equations (ODEs) can be formulated as follows. Given the initial value problem

$$x' = f(x), \qquad x_0 = x(0) \in \mathbb{R}^D \tag{1}$$

Fecha de recepción: 29/11/2008

with $f : \mathbb{R}^D \longrightarrow \mathbb{R}^D$ and solution $\varphi_t(x_0)$, let us suppose that f can be expressed as $f = \sum_{i=1}^m f^{[i]}$ for certain functions $f^{[i]} : \mathbb{R}^D \longrightarrow \mathbb{R}^D$, in such a way that the equations

$$x' = f^{[i]}(x), \qquad x_0 = x(0) \in \mathbb{R}^D, \qquad i = 1, \dots, m \tag{2}$$

can be integrated exactly, with solutions $x(h) = \varphi_h^{[i]}(x_0)$ at $t = h$, the time step. Then, by combining these solutions as

$$\chi_h = \varphi_h^{[m]} \circ \cdots \circ \varphi_h^{[2]} \circ \varphi_h^{[1]} \tag{3}$$

and expanding χ into Taylor series, one finds that $\chi_h(x_0) = \varphi_h(x_0) + \mathcal{O}(h^2)$, so that χ_h provides a first-order approximation to the exact solution. As we will see, it is possible to get higher order approximations by introducing more maps with additional coefficients, $\varphi_{a_{ij}h}^{[i]}$, in the previous composition (3).

One thus may say that splitting methods involve three steps: (i) choosing the set of functions $f^{[i]}$ such that $f = \sum_i f^{[i]}$; (ii) solving either exactly or approximately each equation $x' = f^{[i]}(x)$; and (iii) combining these solutions to construct an approximation for (1). One obvious requirement is that the equations $x' = f^{[i]}(x)$ should be simpler to integrate than the original system.

Some of the advantages that splitting methods possess can be summarized as follows:

- They are usually simple to implement.
- They are, in general, explicit.
- Their storage requirements are quite modest. The algorithms are sequential and the solutions at intermediate stages are stored in the solution vectors. This property can be of great interest when they are applied to partial differential equations (PDEs) previously semidiscretized.
- There exist in the literature a large number of specific methods tailored for different structures.
- They preserve structural properties of the exact solution, thus conferring to the numerical scheme a qualitative superiority with respect to other standard integrators, especially when long time intervals are considered. Examples of these structural features are symplecticity, volume preservation, time-symmetry and conservation of first integrals. In this sense, splitting methods constitute an important class of *geometric numerical integrators*.

Let us give more details on this last item. Traditionally, the goal of numerical integration of ODEs consists in computing the solution to the initial value problem (1) at time $t_N = Nh$ with a global error $\|x_N - x(t_N)\|$ smaller than a prescribed tolerance and as efficiently as possible. To do that one chooses the class of method (one-step, multistep, extrapolation, etc.), the order (fixed or adaptive) and the time step (constant or variable). In contrast, with a geometric numerical integrator one typically fix a (not necessarily small) time step and compute solutions for very long times for several initial conditions, in order to get an approximate phase portrait of the system. It turns out that, although the global error of each trajectory may be large, the phase portrait is in some sense close to that of the original system.

The aim of geometric numerical integration is thus to reproduce the qualitative features of the solution of the differential equation which is being discretised, in particular its geometric properties [17, 41]. The motivation for developing such structure-preserving algorithms arises independently in areas of research as diverse as celestial mechanics, molecular dynamics, control theory, particle accelerators physics, and numerical analysis [41, 44, 57, 58, 49]. Although diverse, the systems appearing in these areas have one important common feature. They all preserve some underlying geometric structure which influences the qualitative nature of the phenomena they produce. In the field of geometric numerical integration these properties are built into the numerical method, which gives the method an improved qualitative behaviour, but also allows for a significantly more accurate long-time integration than with general-purpose methods. In addition to the construction of new numerical algorithms, an important aspect of geometric integration is the explanation of the relationship between preservation of the geometric properties of the scheme and the observed favorable error propagation in long-time integration.

Before proceeding further, let us introduce at this point some splitting methods and illustrate them on simple examples.

Example 1: Symplectic Euler and leapfrog schemes. Suppose we have a Hamiltonian system of the form $H(q,p) = T(p) + V(q)$, where $q \in \mathbb{R}^d$ are the canonical coordinates, $p \in \mathbb{R}^d$ are the conjugate momenta, T represents the kinetic energy and V is the potential energy. Then the equations of motion read [36]

$$q' = T_p(p), \qquad p' = -V_q(q), \tag{4}$$

where T_p and V_q denote the vectors of partial derivatives. Equations (4) can be formulated as (1) with $x = (q,p)^T$, $f(x) = (T_p, -V_q)^T = J\nabla H(x)$ and $D = 2d$. Here J denotes the $2d \times 2d$ canonical symplectic matrix

$$J = \begin{pmatrix} 0 & I_d \\ -I_d & 0 \end{pmatrix}$$

and I_d stands for the d-dimensional identity matrix. In this case the exact flow φ_t is symplectic [1]. The simple Euler method applied to this system provides the following first order approximation for a time step h:

$$\begin{array}{rcl} q_{n+1} & = & q_n + hT_p(p_n) \\ p_{n+1} & = & p_n - hV_q(q_n). \end{array} \tag{5}$$

On the other hand, if we consider H as the sum of two Hamiltonians, the first one depending only on p and the second only on q, the corresponding Hamilton equations

$$\begin{array}{rcl} q' & = & T_p(p) \\ p' & = & 0 \end{array} \qquad \text{and} \qquad \begin{array}{rcl} q' & = & 0 \\ p' & = & -V_q(q) \end{array} \tag{6}$$

with initial condition (q_0, p_0) can be readily solved to yield

$$\varphi_t^{[T]} : \begin{array}{rcl} q(t) & = & q_0 + t\,T_p(p_0) \\ p(t) & = & p_0 \end{array} \qquad \text{and} \qquad \varphi_t^{[V]} : \begin{array}{rcl} q(t) & = & q_0 \\ p(t) & = & p_0 - t\,V_q(q_0), \end{array} \tag{7}$$

respectively. Composing the time $t = h$ flow $\varphi_h^{[V]}$ (from initial condition (q_n, p_n)) followed by $\varphi_h^{[T]}$, gives the scheme

$$\chi_h \equiv \varphi_h^{[T]} \circ \varphi_h^{[V]} : \begin{array}{rcl} p_{n+1} & = & p_n - h\,V_q(q_n) \\ q_{n+1} & = & q_n + h\,T_p(p_{n+1}). \end{array} \tag{8}$$

Since it is a composition of the flows of two Hamiltonian systems and in addition the composition of symplectic maps is again symplectic, χ_h is a symplectic integrator, which can be called appropriately the *symplectic Euler method.* It is of course also possible to compose the maps in the opposite order, $\varphi_h^{[V]} \circ \varphi_h^{[T]}$, thus obtaining another first order symplectic Euler scheme:

$$\chi_h^* \equiv \varphi_h^{[V]} \circ \varphi_h^{[T]} : \begin{array}{rcl} q_{n+1} & = & q_n + h\,T_p(p_n) \\ p_{n+1} & = & p_n - h\,V_q(q_{n+1}). \end{array} \tag{9}$$

One says that (9) is the *adjoint* of χ_h. Yet another possibility consists in using a 'symmetric' version

$$\mathcal{S}_h^{[2]} \equiv \varphi_{h/2}^{[V]} \circ \varphi_h^{[T]} \circ \varphi_{h/2}^{[V]}, \tag{10}$$

which is known as the Strang splitting [77], the leapfrog or the Störmer–Verlet method [85], depending on the context where it is used. Observe that $\mathcal{S}_h^{[2]} = \chi_{h/2} \circ \chi_{h/2}^*$ and it is also symplectic and second order.

Example 2: Harmonic oscillator. Let us consider now the Hamiltonian function $H(q, p) = \frac{1}{2}(p^2 + q^2)$, where now $q, p \in \mathbb{R}$. Then the corresponding equations (4) are linear and can be written as

$$x' \equiv \begin{pmatrix} q' \\ p' \end{pmatrix} = \Bigg[\underbrace{\begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}}_{A} + \underbrace{\begin{pmatrix} 0 & 0 \\ -1 & 0 \end{pmatrix}}_{B} \Bigg] \begin{pmatrix} q \\ p \end{pmatrix} = (A + B)\,x. \tag{11}$$

This system has periodic solutions for which the energy H is conserved. In addition, it is area preserving and time reversible. The numerical solution obtained by the Euler scheme (5) reads

$$\begin{pmatrix} q_{n+1} \\ p_{n+1} \end{pmatrix} = \begin{pmatrix} 1 & h \\ -h & 1 \end{pmatrix} \begin{pmatrix} q_n \\ p_n \end{pmatrix}, \tag{12}$$

whereas the symplectic Euler method (9) leads to

$$\left(\begin{array}{c} q_{n+1} \\ p_{n+1} \end{array} \right) = \left(\begin{array}{cc} 1 & h \\ -h & 1-h^2 \end{array} \right) \left(\begin{array}{c} q_n \\ p_n \end{array} \right) = \mathrm{e}^{hB}\mathrm{e}^{hA} \left(\begin{array}{c} q_n \\ p_n \end{array} \right). \tag{13}$$

Both render first order approximations to the exact solution, which can be expressed as $x(t) = \mathrm{e}^{h(A+B)}x_0$, but there are important differences between them. First, the map (13) is area preserving (because it is symplectic), in contrast with (12). Second, the approximation obtained by the symplectic Euler scheme verifies

$$\frac{1}{2}(p_{n+1}^2 + hp_{n+1}q_{n+1} + q_{n+1}^2) = \frac{1}{2}(p_n^2 + hp_nq_n + q_n^2).$$

Third, it can be shown that (13) *is* the exact solution at $t = h$ of the *perturbed* Hamiltonian system

$$\begin{aligned} \tilde{H}(q,p,h) &= \frac{2\arcsin(h/2)}{h\sqrt{4-h^2}}(p^2 + h\,p\,q + q^2) \qquad (14) \\ &= \frac{1}{2}(p^2+q^2) + h\left(\frac{1}{2}\,p\,q + \frac{1}{12}h(p^2+q^2) + \cdots\right). \end{aligned}$$

In other words, the numerical approximation (13), which is only of first order for the exact trajectories of the Hamiltonian $H(q,p) = \frac{1}{2}(p^2+q^2)$, is in fact the exact solution of the perturbed Hamiltonian (14).

How these features manifest in actual simulations? To illustrate this point we take initial conditions $(q_0,p_0) = (4,0)$ and integrate with a time step $h = 0.1$. Figure 1 shows the first five numerical approximations obtained by the Euler method (12) and the symplectic Euler scheme (13) in the left panel, and the results for the first 100 steps in the right panel. It is clear that for one time step there are not significant differences between the standard Euler and the symplectic Euler methods, but the picture is completely different for longer integrations, where the superiority of the splitting symplectic method is evident. Note that the numerical solution it provides evolves on a slightly perturbed ellipse.

Example 3: The 2-body problem (Kepler problem). The motion of two bodies attracting each other through the gravitational law can be described by

$$q_i'' = -\frac{q_i}{(q_1^2+q_2^2)^{3/2}}, \qquad i = 1,2 \tag{15}$$

in conveniently normalized coordinates in the plane of motion. This system has a number of characteristic geometric properties. First, equations (15) can be derived from the Hamiltonian function

$$H(q,p) = T(p) + V(q) = \frac{1}{2}(p_1^2+p_2^2) - \frac{1}{r}, \qquad r = \sqrt{q_1^2+q_2^2}. \tag{16}$$

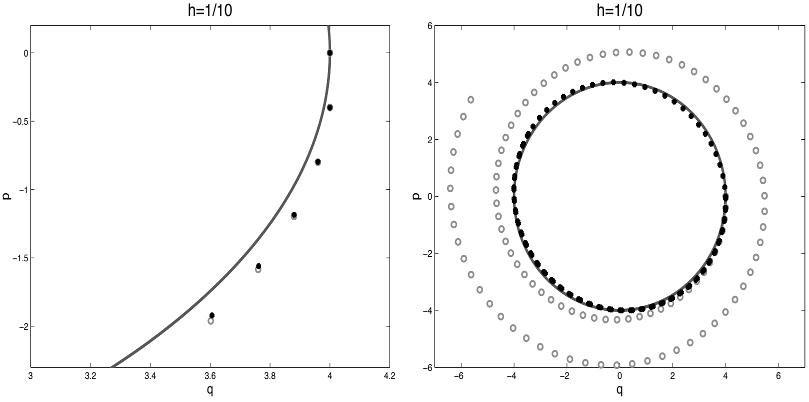


Figure 1: Numerical integration of the harmonic oscillator using the Euler method (white circles) and the symplectic Euler method (black circles) with initial condition $(q_0, p_0) = (4, 0)$ and time step $h = \frac{1}{10}$. The left panel shows the results for the first 5 steps, whereas the right panel shows the results for the first 100 steps. The exact solution corresponds to the solid line.

Second, it is invariant under continuous translations in time and rotations in space, and thus both the Hamiltonian and the angular momentum $L = q_1p_2 - q_2p_1$ are preserved. In addition, the so-called Laplace–Runge–Lenz vector is also constant along their solutions, due to the fact that the symmetry group of this problem is the group of four-dimensional real proper rotations SO(4) [36].

For the numerical integration of this problem we choose as initial value

$$q_1(0) = 1 - e, \quad q_2(0) = 0, \quad p_1(0) = 0, \quad p_2(0) = \sqrt{\frac{1+e}{1-e}}, \tag{17}$$

where $0 \leq e < 1$ represents the eccentricity of the orbit. In this case the total energy is $H = H_0 = -1/2$, the period of the solution is 2π, the initial condition corresponds to the pericenter and the major semiaxis of the ellipse is 1.

Figure 2 shows some numerical solutions obtained with schemes (5) and (8) for the initial conditions (17) with eccentricity $e = 0.6$. The left panel shows the results for the integration of 3 periods with time step $h = \frac{1}{100}$. As in the previous example, the explicit Euler method provides an approximate orbit that spirals outwards, whereas the symplectic Euler scheme merely distorts the ellipse, but also exhibits a precession effect. To better illustrate this effect, we repeat the experiment for a longer interval (15 periods) and a larger time step ($h = \frac{1}{20}$) in the right panel. The explanation of these phenomena can be formulated as follows. On the one hand, the symplectic Euler method exactly conserves the angular momentum. On the other hand, the numerical solution it provides can be seen as the exact solution of a slightly perturbed Kepler problem, and thus SO(4) is no longer the symmetry group of the problem, so that the Laplace–Runge–Lenz vector is not preserved and the trajectories are not closed anymore.

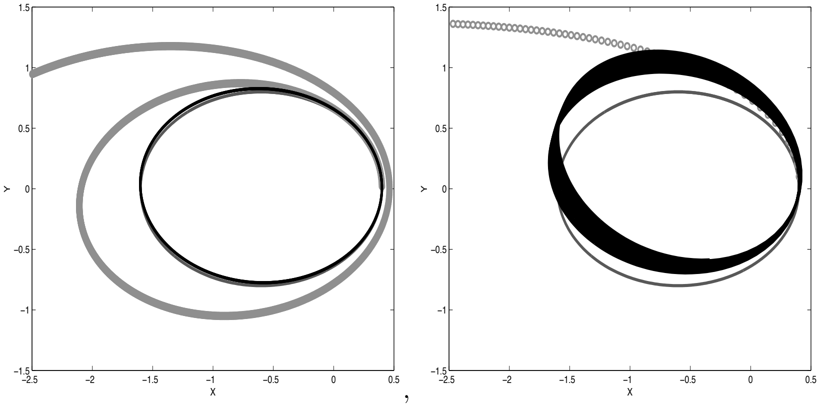


Figure 2: Numerical integration of the 2-body problem using the Euler method (white circles) and the symplectic Euper scheme (black circles) for the initial conditions (17) with eccentricity $e = 0.6$. The left panel shows the results for $h = \frac{1}{100}$ and the first 3 periods and the right panel shows the results for $h = \frac{1}{20}$ and the first 15 periods.

Next we check how the error in the preservation of energy and the global error in position propagates with time. For comparison, we also include the results obtained with a Runge–Kutta method of order 2 (Heun's method) and the leapfrog scheme (10). We now consider $e = 1/5$ and integrate for 500 periods. The step size is $h = \frac{2\pi}{1500}$ in all cases, except for the Heun method which uses $h = \frac{2\pi}{750}$ instead. In this way all methods require the same number of force evaluations (Heun's method computes twice the force per step). The corresponding results are shown in Figure 3 in a log-log scale. Notice that the average error in energy does not grow for the split symplectic methods and the error in positions grows only linearly with time, in contrast with Euler and Heun schemes. The Störmer–Verlet integrator provides more accurate results than the Heun method with the same computational cost.

A collection of (additional) examples. Splitting methods constitute an important tool in different areas of science. In addition to Hamiltonian systems, they can be successfully applied in the numerical study of Poisson systems, systems possessing integrals of the motion (such as energy and angular momentum) and systems with (continuous, discrete and time-reversal) symmetries. As a matter of fact, splitting methods have been designed (often independently) and extensively used in fields as distant as molecular dynamics, simulation of storage rings in particle accelerators, celestial mechanics and

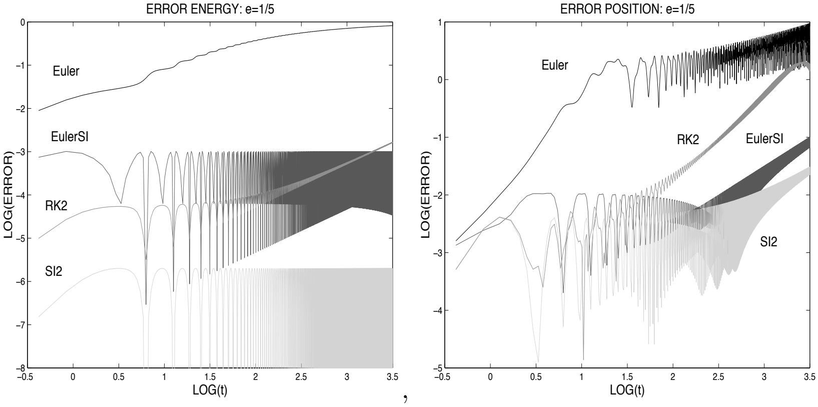


Figure 3: Error growth in energy and position for the Kepler problem with $e = 1/5$ and 500 periods achieved by the first order symplectic Euler (EulerSI) and the second order Störmer–Verlet integrator (SI2). For comparison, we also include the first order Euler and the second order Heun (RK2) methods. The time step is adjusted in such a way that all methods use 1500 force evaluations per period.

quantum physics simulations. To see why this is so, next we collect a number of differential equations which appear in different contexts ranging from Celestial Mechanics to electromagnetism and Quantum Mechanics. These examples also try to illustrate the fact that very often one particular equation can be split into different ways and the most appropriate methods may depend on the split chosen.

We (arbitrarily) classify our examples into three different categories.

1. Hamiltonian systems.

   (a) Generalized harmonic oscillator ($M, N \in \mathbb{R}^{d\times d}$):

$$H = \frac{1}{2} p^T M p + \frac{1}{2} q^T N q. \tag{18}$$

   (b) Hénon–Heiles Hamiltonian [43]:

$$H = \frac{1}{2}(p_1^2 + p_2^2) + \frac{1}{2}(q_1^2 + q_2^2) + q_1^2 q_2 - \frac{1}{3} q_2^3. \tag{19}$$

   (c) Perturbed Kepler problem. It models the dynamics of a satellite moving into the gravitational field produced by a slightly oblate planet:

$$H = \frac{1}{2}(p_1^2 + p_2^2) - \frac{1}{r} - \frac{\varepsilon}{2r^3}\left(1 - \frac{3q_1^2}{r^2}\right) \tag{20}$$

   where ε is typically a small parameter. When $\varepsilon = 0$, the 2-body problem (16) is recovered.

(d) The gravitational N-body problem ($q_i, p_i \in \mathbb{R}^3,\ i = 1, \dots, N$):

$$H = \frac{1}{2} \sum_{i=1}^{N} \frac{1}{2m_i} p_i^T p_i - G \sum_{i=2}^{N} \sum_{j=1}^{i-1} \frac{m_i m_j}{\|q_i - q_j\|}. \tag{21}$$

(e) The motion of a charged particle in a constant magnetic field perturbed by k electrostatic plane waves [21]:

$$H(q,p,t) = \frac{1}{2}p^2 + \frac{1}{2}q^2 + \varepsilon \sum_{i=1}^{k} \cos(q - \omega_i t). \tag{22}$$

2. More general dynamical systems.

(a) The Volterra–Lotka problem [41],

$$\frac{d}{dt}\begin{bmatrix} u \\ v \end{bmatrix} = \begin{bmatrix} -2 & 0 \\ 0 & 1 \end{bmatrix}\begin{bmatrix} u \\ v \end{bmatrix} + \begin{bmatrix} uv \\ -uv \end{bmatrix} = \begin{bmatrix} u(v-2) \\ 0 \end{bmatrix} + \begin{bmatrix} 0 \\ v(1-u) \end{bmatrix}, \tag{23}$$

with first integral $I(u,v) = \log u - u + 2\log v - v$.

(b) The Lorenz system [52, 39] (split into linear and non-linear parts):

$$\frac{d}{dt}\begin{bmatrix} x \\ y \\ z \end{bmatrix} = \begin{bmatrix} -\sigma & \sigma & 0 \\ r & -1 & 0 \\ 0 & 0 & b \end{bmatrix}\begin{bmatrix} x \\ y \\ z \end{bmatrix} + \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & -x \\ 0 & x & 0 \end{bmatrix}\begin{bmatrix} x \\ y \\ z \end{bmatrix}. \tag{24}$$

Here $\sigma, r, b > 0$ are constant. The values considered in [52] were $\sigma = 10$, $r = 28$ and $b = 8/3$.

(c) The ABC-flow ($f = f_A + f_B + f_C$, but other splits are also possible [57]):

$$\frac{d}{dt}(x,y,z) = A(0, \sin x, \cos x) + B(\cos y, 0, \sin y) + C(\sin z, \cos z, 0). \tag{25}$$

3. Evolutionary PDEs.

Although we are mainly concerned here with splitting methods applied to ODEs, it turns out that they can also be used in the numerical integration of certain partial differential equations. Specifically, a number of PDEs relevant in the applications, after an appropriate space discretization, lead to a system of ODEs which can be subsequently solved numerically by splitting methods. Among these equations, the following are worth to be mentioned.

(a) The Schrödinger equation ($\hbar = 1$):

$$i\frac{\partial}{\partial t}\Psi(x,t) = \left(-\frac{1}{2m}\nabla^2 + V(x)\right)\Psi(x,t). \tag{26}$$

(b) The Gross–Pitaevskii equation [68]:

$$i\frac{\partial}{\partial t}\Psi(x,t) = \left(-\frac{1}{2m}\nabla^2 + V(x) + \alpha|\Psi(x,t)|^2\right)\Psi(x,t) \qquad (27)$$

(c) The Maxwell equations

$$\frac{\partial}{\partial t}\mathbf{B} = -\frac{1}{\mu}\nabla\times\mathbf{E}, \qquad \frac{\partial}{\partial t}\mathbf{E} = \frac{1}{\varepsilon}\nabla\times\mathbf{B}, \qquad (28)$$

where $\mathbf{E}(x,t)$, $\mathbf{B}(x,t)$ are the electric and magnetic field vectors, $\mu(x)$ is the the permeability and $\varepsilon(x)$ is the permittivity.

As we stated before, splitting methods form a subclass of geometric numerical integrators for various types of ODEs. The reason is easy to grasp from the examples analyzed before. Suppose that the flow of the original differential equation (1) forms a particular group of diffeomorphisms (in the case of Hamiltonian system, the group of symplectic maps). If f is conveniently split as $f = \sum_i f^{[i]}$ (step (i) in the construction process of a splitting scheme) and the flows corresponding to each $f^{[i]}$ also belong to the same group of diffeomorphisms in such a way that they can be explicitly obtained (step (ii)), then, by composing these flows (step (iii)) we get an approximation in the group, thus inheriting geometric properties of the exact solution. These considerations also hold (with some modifications) when the exact flow forms a semigroup or a symmetric space.

With respect to steps (i) and (ii) before, several comments are in order. First, whereas for certain classes of ODEs, the splitting can be constructed systematically for any f, in other cases no general procedure is known, and thus one has to proceed on a case by case basis. Second, sometimes a standard splitting is possible for a certain f, but there exist other possible choices leading to more efficient schemes (we will see some examples in section 8). Third, whereas the original system possesses several geometric properties which are interesting to preserve by the numerical scheme, different splittings preserve different properties and it is not always possible to find one splitting preserving all of them. These aspects have been analyzed in detail in [57], where a classification of ODEs and general guidelines to find suitable splittings in each case is provided. Here, by contrast, we will concentrate on the third step of any splitting method: given a particular splitting, we will show how to combine the flows of the pieces $f^{[i]}$ to get efficient higher order approximations. In any case, the reader is referred to the excellent review paper [57] and the monographs [41, 49], where these and other issues, mainly in connection with geometric numerical integration, are thoroughly examined.

2 Splitting and composition methods

2.1 Increasing the order of an integrator by composition

It is well known that numerical integrators of arbitrarily high order can be obtained by composition of a basic integrator of low order. Consider for instance the leapfrog scheme (10), which is a second-order integrator $\mathcal{S}^{[2]} : \mathbb{R}^{2d} \to \mathbb{R}^{2d}$. Then, a 4th order integrator $\mathcal{S}^{[4]} : \mathbb{R}^{2d} \to \mathbb{R}^{2d}$ can be obtained as

$$\mathcal{S}_h^{[4]} = \mathcal{S}_{\alpha h}^{[2]} \circ \mathcal{S}_{\beta h}^{[2]} \circ \mathcal{S}_{\alpha h}^{[2]}, \qquad \text{with} \quad \alpha = \frac{1}{2 - 2^{1/3}}, \qquad \beta = 1 - 2\alpha. \tag{29}$$

More generally, if one recursively defines $\mathcal{S}^{[2k+2]} : \mathbb{R}^{2d} \to \mathbb{R}^{2d}$ for $k = 1, 2, \ldots$ as

$$\mathcal{S}_h^{[2k+2]} = \mathcal{S}_{\alpha h}^{[2k]} \circ \mathcal{S}_{\beta h}^{[2k]} \circ \mathcal{S}_{\alpha h}^{[2k]}, \tag{30}$$

with

$$\alpha = \frac{1}{2 - 2^{1/(2k+1)}}, \qquad \beta = 1 - 2\alpha, \tag{31}$$

then the schemes $\mathcal{S}_h^{[2k]}$ are of order $2k$ ($k \geq 1$) [78, 88]. We will prove later on this assertion. At this point it is useful to introduce the notion of adjoint of a given integrator ψ_h [73]. By definition, this is the method ψ_h^* such that $\psi_h^* = \psi_{-h}^{-1}$. A method that is equal to its own adjoint is called self-adjoint or (time-)*symmetric*. In this case, $\psi_{-h} \circ \psi_h = \text{id}$. Since the leapfrog scheme (10) can be rewritten as

$$\mathcal{S}_h^{[2]} = \chi_{h/2} \circ \chi_{h/2}^*, \tag{32}$$

where χ_h is the symplectic Euler method (8), then $\mathcal{S}_h^{[2]}$ is certainly time-symmetric. Actually, given any basic first order integrator $\chi_h : \mathbb{R}^D \to \mathbb{R}^D$ for the ODE system (1), the composition (32) is a time-symmetric method of order 2, and the other way around: any self-adjoint second order integrator can be expressed as (32). Furthermore, the integrators $\mathcal{S}_h^{[2k]}$ ($k = 1, 2, \ldots$) recursively defined by (30)–(31) are time-symmetric methods of order $2k$. In particular, if $f(x)$ in the ODE (1) is split as

$$f(x) = \sum_{i=1}^{m} f^{[i]}(x) \tag{33}$$

then, time-symmetric integrators $\mathcal{S}_h^{[2k]}$ of order $2k$ can be constructed in this way by considering the basic first order integrator

$$\chi_h = \varphi_h^{[m]} \circ \cdots \circ \varphi_h^{[2]} \circ \varphi_h^{[1]} \tag{34}$$

and its adjoint

$$\chi_h^* = \chi_{-h}^{-1} = \varphi_h^{[1]} \circ \varphi_h^{[2]} \circ \cdots \circ \varphi_h^{[m]}.$$

This general procedure of constructing geometric integrators of arbitrarily high order, although simple, presents some drawbacks. In particular, the resulting methods require a large number of evaluations and usually have large truncation errors.

As we will show, efficient schemes can be built by considering more general composition integrators. First observe that the $(2k)$th order integrators $\mathcal{S}_h^{[2k]}$ can be rewritten in the form

$$\psi_h = \chi_{\alpha_{2s}h} \circ \chi^*_{\alpha_{2s-1}h} \circ \cdots \circ \chi_{\alpha_2 h} \circ \chi^*_{\alpha_1 h} \tag{35}$$

with $s = 3^{k-1}$ and some fixed coefficients $(\alpha_1, \ldots, \alpha_{2s}) \in \mathbb{R}^{2s}$. Then the idea is to consider composition integrators of the form (35) with appropriately chosen coefficients $(\alpha_1, \ldots, \alpha_{2s}) \in \mathbb{R}^{2s}$.

In the particular case where the ODE (1) is split in two parts $f = f^{[a]} + f^{[b]}$ and $\chi_h = \varphi_h^{[b]} \circ \varphi_h^{[a]}$, one can trivially check that the composition integrator (35) can be rewritten as

$$\psi_h = \varphi^{[b]}_{b_{s+1}h} \circ \varphi^{[a]}_{a_s h} \circ \varphi^{[b]}_{b_s h} \circ \cdots \circ \varphi^{[b]}_{b_2 h} \circ \varphi^{[a]}_{a_1 h} \circ \varphi^{[b]}_{b_1 h}, \tag{36}$$

where $b_1 = \alpha_1$ and for $j = 1, \ldots, s$,

$$a_j = \alpha_{2j-1} + \alpha_{2j}, \qquad b_{j+1} = \alpha_{2j} + \alpha_{2j+1} \tag{37}$$

(with $\alpha_{2s+1} = 0$). Conversely, any integrator of the form (36) satisfying that $\sum_{i=1}^{s} a_i = \sum_{i=1}^{s+1} b_i$ can be expressed in the form (35) with $\chi_h = \varphi_h^{[b]} \circ \varphi_h^{[a]}$. For later reference, we state the following result, due to McLachlan [54].

Theorem 1 *The integrator (36) is of order r for ODEs of the form (1) with $f : \mathbb{R}^D \longrightarrow \mathbb{R}^D$ arbitrarily split as $f = f^{[a]} + f^{[b]}$ if and only if the integrator (35) (with coefficients α_j obtained from (37)) is of order r for arbitrary consistent integrators χ_h.*

2.2 Integrators and series of differential operators

Before proceeding further with the analysis, let us relate generic numerical integrators with formal series of differential equations. This relationship will allow one to formulate in a rather simple way the conditions to be satisfied by an integration scheme to achieve a given order of consistency.

First of all, let us recall that an integrator $\psi_h : \mathbb{R}^D \to \mathbb{R}^D$ for the system (1) is said to be of order r if for all $x \in \mathbb{R}^D$

$$\psi_h(x) = \varphi_h(x) + \mathcal{O}(h^{r+1}) \tag{38}$$

as $h \to 0$, where φ_h is the h-flow of the ODE (1).

It is well known that, for any smooth function $g : \mathbb{R}^D \to \mathbb{R}$, it formally holds that [66]

$$g(\varphi_h(x)) = g(x) + \sum_{n \geq 1} \frac{1}{n!} F^n[g](x) = \exp(hF)[g](x),$$

where F is the Lie derivative associated to the ODE system (1), i.e., the first order linear differential operator F acting on functions in $C^\infty(\mathbb{R}^D,\mathbb{R})$ as follows. For each $g \in C^\infty(\mathbb{R}^D,\mathbb{R})$ and each $x=(x_1,\ldots,x_D)\in\mathbb{R}^D$

$$F[g](x) = \sum_{j=1}^{D} f_j(x)\frac{\partial g}{\partial x_j}(x), \tag{39}$$

where $f(x) = (f_1(x),\ldots,f_D(x))^T$. Motivated by this fact, we consider for a basic integrator $\chi_h : \mathbb{R}^D \to \mathbb{R}^D$, the linear differential operators X_n $(n \geq 1)$ acting on smooth functions $g \in C^\infty(\mathbb{R}^D,\mathbb{R})$ as follows:

$$X_n[g](x) = \frac{1}{n!}\frac{d^n}{dh^n}\, g(\chi_h(x))|_{h=0}\,, \tag{40}$$

so that formally $g(\chi_h(x)) = X(h)[g](x)$, where

$$X(h) = I + \sum_{n\geq 1} h^n X_n, \tag{41}$$

and I denotes the identity operator. Thus, the integrator χ_h is of order r if

$$X_n = \frac{1}{n!}F^n, \qquad 1 \leq n \leq r.$$

Alternatively, one can consider the series of vector fields

$$Y(h) = \sum_{n\geq 1} h^n Y_n = \log(X(h)) = \sum_{m\geq 1} \frac{(-1)^{m+1}}{m}\left(hX_1 + h^2X_2 + \cdots\right)^m,$$

that is,

$$Y_n = \sum_{m\geq 1}^{n} \frac{(-1)^{m+1}}{m} \sum_{j_1+\cdots+j_m=n} X_{j_1}\cdots X_{j_m},$$

so that $X(h) = \exp(Y(h))$, and formally, $g(\chi_h(x)) = \exp(Y(h))[g](x)$. Clearly, the basic integrator is of order r if

$$Y_1 = F, \qquad Y_n = 0 \quad \text{for} \quad 2 \leq n \leq r.$$

For the adjoint integrator $\chi_h^* = \chi_{-h}^{-1}$, one obviously gets $g(\chi_h^*(x)) = \mathrm{e}^{-Y(-h)}[g](x)$. This shows that χ_h is time-symmetric if and only if $Y(h) = hY_1 + h^3Y_3 + \cdots$, and in particular, that time-symmetric methods are of even order.

It is possible now to check that the symmetric integrators $\mathcal{S}_h^{[2k]}$ given by (30)–(31) are actually schemes of order $2k$ provided that $\mathcal{S}_h^{[2]}$ is a symmetric second order integrator. Consider the series of differential operators

$$F^{[2k]}(h) = hF + h^{2k+1}F_{2k+1}^{[2k]} + h^{2k+3}F_{2k+3}^{[2k]} + \cdots$$

such that $g(\mathcal{S}_h^{[2k]}(x)) = \exp(F^{[2k]}(h))[g](x)$. Then one clearly has

$$\exp(F^{[2k+2]}(h)) = \exp(F^{[2k]}(\alpha h)) \exp(F^{[2k]}(\beta h)) \exp(F^{[2k]}(\alpha h))$$

which implies

$$F^{[2k+2]}(h) = h\,(2\alpha + \beta)\,F + h^{2k+1}\,(2\alpha^{2k+1} + \beta^{2k+1})\,F_{2k+1}^{[2k]} + \mathcal{O}(h^{2k+3}),$$

and thus $\mathcal{S}_h^{2k+2}$ is of order $2k+2$ provided that $\mathcal{S}_h^{2k}$ is of order $2k$ and α and β satisfy the equations

$$2\alpha + \beta = 1, \qquad 2\alpha^{2k+1} + \beta^{2k+1} = 0,$$

whose unique real solution is given by (31).

In the general case, for the composition method (35) we have

$$g(\psi_h(x)) = \Psi(h)[g](x),$$

where $\Psi(h) = I + h\Psi_1 + h^2\Psi_2 + \cdots$ is a series of differential operators satisfying

$$\Psi(h) = X(-\alpha_1 h)^{-1} X(\alpha_2 h) \cdots X(-\alpha_{2s-1} h)^{-1} X(\alpha_{2s} h), \tag{42}$$

where the series $X(h)$ is given by (40)–(41), and

$$X(h)^{-1} = I + \sum_{m\geq 1} (-1)^{m+1} \left(hX_1 + h^2 X_2 + \cdots\right)^m. \tag{43}$$

Thus, the order of a composition integrator of the form (35) can be checked by comparing the series of differential operators $\Psi(h)$ with the series $\exp(hF)$ associated to the flow of the system (1). That is, the integrator (35) is of order r if

$$\Psi_n = \frac{1}{n!} F^n, \qquad 1 \leq n \leq r. \tag{44}$$

Instead of using (42) to obtain the terms Ψ_n of the series $\Psi(h)$, one can equivalently consider the formal equality

$$\Psi(h) = \mathrm{e}^{-Y(-h\alpha_1)}\, \mathrm{e}^{Y(h\alpha_2)} \cdots \mathrm{e}^{-Y(-h\alpha_{2s-1})}\, \mathrm{e}^{Y(h\alpha_{2s})}, \tag{45}$$

to obtain the series expansion of $\log(\Psi(h)) = \sum_{n\geq 1} h^n F_n$, so that rth order compositions methods can also be characterized by the conditions

$$F_1 = F, \qquad F_n = 0 \quad \text{for} \quad 2 \leq n \leq r. \tag{46}$$

As for the splitting integrator (36) when the ODE (1) is split in two parts,

$$f(x) = f^{[a]}(x) + f^{[b]}(x), \tag{47}$$

let $F^{[a]}$ and $F^{[b]}$ be the Lie derivatives corresponding to $f^{[a]}$ and $f^{[b]}$ respectively, that is,

$$F^{[a]}[g](x) = \sum_{j=1}^{D} f_j^{[a]}(x) \frac{\partial g}{\partial x_j}(x), \qquad F^{[b]}[g](x) = \sum_{j=1}^{D} f_j^{[b]}(x) \frac{\partial g}{\partial x_j}(x) \tag{48}$$

for each $g \in C^\infty(\mathbb{R}^D, \mathbb{R})$ and each $x \in \mathbb{R}^D$. Then, the series $\Psi(h)$ of differential operators associated to the integrator ψ_h in (36) can be formally written as

$$\Psi(h) = \mathrm{e}^{b_1 h F^{[b]}} \, \mathrm{e}^{a_1 h F^{[a]}} \cdots \mathrm{e}^{b_s h F^{[b]}} \, \mathrm{e}^{a_s h F^{[a]}} \, \mathrm{e}^{b_{s+1} h F^{[b]}}. \tag{49}$$

3 Order conditions of splitting and composition methods

There are several procedures to get the order conditions for the coefficients of splitting and composition methods of a given order. These are, generally speaking, large systems of polynomial equations in the coefficients which are obtained from equations (46). Perhaps the two most popular are (i) the expansion of the series $\log(\Psi(h)) = \sum_{n \geq 1} h^n F_n$ of vector fields by applying recursively the Baker–Campbell–Hausdorff (BCH) formula [88], and (ii) a generalization of the theory of rooted trees used in the theory of Runge–Kutta methods, which allows one to get an equivalent set of simpler order conditions in a systematic way [63] (see also [41]). In this section we first summarize briefly how to get these equations with the BCH formula, and then we present a novel approach, related to that in [63], but based on Lyndon words instead of rooted trees.

3.1 Order conditions via BCH formula

As is well known, if X and Y are two non-commuting operators, the BCH formula establishes that formally, $\mathrm{e}^X \mathrm{e}^Y = \mathrm{e}^Z$, where Z belongs to the Lie algebra $\mathcal{L}(X, Y)$ generated by X and Y with the commutator $[X, Y] = XY - YX$ as Lie bracket [84]. Moreover,

$$Z = \log(\mathrm{e}^X \, \mathrm{e}^Y) = X + Y + \sum_{m=2}^{\infty} Z_m, \tag{50}$$

with $Z_m(X, Y)$ a homogeneous Lie polynomial in X and Y of degree m, i.e., it is a $\mathbb{Q}$-linear combination of commutators of the form $[V_1, [V_2, \ldots, [V_{m-1}, V_m] \ldots]]$ with $V_i \in \{X, Y\}$ for $1 \leq i \leq m$. The first terms read

$$\begin{aligned} Z_2 &= \frac{1}{2}[X, Y] \\ Z_3 &= -\frac{1}{12}[[X, Y], X] + \frac{1}{12}[[X, Y], Y] \\ Z_4 &= \frac{1}{24}[[[X, Y], Y], X], \end{aligned}$$

and explicit expressions up to $m = 20$ have been recently computed in an arbitrary generalized Hall basis of $\mathcal{L}(X, Y)$ [22].

The procedure to get the order conditions for the composition method (35) with this approach can be summarized as follows. First, consider the series of differential operators $\Psi(h)$ associated to the integrator (35), expressed as a product of exponentials of vector fields, i.e., equation (45). Then, apply repeatedly the BCH formula to get the series expansion $\log(\Psi(h)) = \sum_{n\geq 1} h^n F_n$. In this way, one gets

$$\begin{aligned} \log(\Psi(h)) &= hw_1Y_1 + h^2w_2Y_2 + h^3(w_3Y_3 + w_{12}[Y_1,Y_2]) \\ &\quad +h^4(w_4Y_4 + w_{13}[Y_1,Y_3] + w_{112}[Y_1,[Y_1,Y_2]]) + \mathcal{O}(h^5) \end{aligned} \tag{51}$$

where the $w_{j_1\cdots j_m}$ are polynomials of degree $n = j_1 + \cdots + j_m$ in the parameters $\alpha_1,\ldots,\alpha_{2s}$. The first such polynomials are

$$w_1 = \sum_{i=1}^{2s} \alpha_i, \qquad w_2 = \sum_{i=1}^{2s}(-1)^i\alpha_i^2, \qquad w_3 = \sum_{i=1}^{2s}\alpha_i^3. \tag{52}$$

In general, the expressions of the polynomials $w_{j_1\cdots j_m}$ in (51) obtained by repeated application of the BCH formula are rather cumbersome.

The order conditions for the composition integrator (35) are then obtained by imposing equations (46) to guarantee that the scheme has order $r \geq 1$. Thus, the order conditions are $w_1 = 1$, and $w_{j_1\cdots j_m} = 0$ whenever $2 \leq j_1+\cdots+j_m \leq r$.

One can proceed similarly to get the order conditions of the splitting scheme (36) in terms of the coefficients a_i, b_i: Consider the series of differential operators $\Psi(h)$ associated to the integrator (36) expressed as (49); then, apply repeatedly the BCH formula to get the series expansion $\log(\Psi(h)) = \sum_{n\geq 1} h^n F_n$, so that the order conditions will be obtained by imposing equations (46) to guarantee order $r \geq 1$. It can be seen that the following holds for $\log(\Psi(h))$,

$$\begin{aligned} \log(\Psi(h)) &= h(v_aF^{[a]} + v_bF^{[b]}) + h^2v_{ab}F^{[ab]} + h^3(v_{abb}F^{[abb]} + v_{aba}F^{[aba]}) \\ &\quad +h^4(v_{abbb}F^{[abbb]} + v_{abba}F^{[abba]} + v_{abaa}F^{[abaa]}) + \mathcal{O}(h^5), \end{aligned} \tag{53}$$

where

$$\begin{aligned} F^{[ab]} &= [F^{[a]},F^{[b]}], \quad F^{[abb]} = [F^{[ab]},F^{[b]}], \quad F^{[aba]} = [F^{[ab]},F^{[a]}], \\ F^{[abbb]} &= [F^{[abb]},F^{[b]}], \quad F^{[abba]} = [F^{[abb]},F^{[a]}], \quad F^{[abaa]} = [F^{[aba]},F^{[a]}], \end{aligned}$$

and $v_a, v_b, v_{ab}, v_{abb}, v_{aba}, v_{abbb},\ldots$ are polynomials in the parameters a_i, b_i of the splitting scheme (36). In particular, one gets

$$\begin{aligned} v_a &= \sum_{i=1}^{s} a_i, \qquad v_b = \sum_{i=1}^{s+1} b_i, \qquad v_{ab} = \frac{1}{2}v_av_b - \sum_{1\leq i\leq j\leq s} b_ia_j, \qquad (54) \\ 2v_{aba} &= -\frac{1}{6}v_a^2v_b + \sum_{1\leq i<j\leq k\leq s} a_ib_ja_k, \qquad 2v_{abb} = \frac{1}{6}v_av_b^2 - \sum_{1\leq i\leq j<k\leq s+1} b_ia_jb_k. \end{aligned}$$

From (53), we see that a characterization of the order of the splitting scheme (36) can be obtained by considering $v_a = v_b = 1$ and $v_{ab} = v_{abb} = v_{aba} = \cdots = 0$ up to polynomials of that form of the required order. The set of order conditions thus obtained will be independent in the general case if the vector fields $F^{[a]}, F^{[b]}, F^{[ab]}, F^{[abb]}, F^{[aba]}$ considered in (53) correspond to a basis of the free Lie algebra on the alphabet $\{a, b\}$. Notice that in (53) we have considered a Hall basis (the classical basis of P. Hall) associated to the Hall words $a, b, ab, abb, aba, abbb, abba, abaa, \cdots$ [69]. The coefficients v_w in (53) corresponding to each Hall word w can be systematically obtained using the results in [62] in terms of rooted trees and iterated integrals. An efficient algorithm (based on the results in [62]) of the BCH formula and related calculations that allows one to obtain (53) up to terms of arbitrarily high degree is presented in [22].

3.2 A set of independent order conditions

We next present a set of order conditions for composition integrators (35) derived in [25].

From (41)–(43), it follows that

$$\Psi(h) = I + \sum_{n \geq 1} h^n \sum_{j_1 + \cdots + j_r = n} u_{j_1 \cdots j_r}(\alpha_1, \ldots, \alpha_{2s})\, X_{j_1} \cdots X_{j_r}, \tag{55}$$

for some polynomial functions $u_{j_1 \cdots j_r}$ of the parameters $\alpha_1, \ldots, \alpha_{2s}$ of the method. We next introduce some notation in order to explicitly write these polynomials. For each positive integer j, we write $j^* = j - 1$ if j is even, and $j^* = j$ if j is odd. Finally, for each pair (i, j) of positive integers, we write $\alpha_j^{(i)} = (-1)^{j(i-1)}(\alpha_j)^i$. That is, $\alpha_j^{(i)} = (\alpha_j)^i$ if j is even or i is odd, and $\alpha_j^{(i)} = -(\alpha_j)^i$ if j is odd and i is even. Now, it is not difficult to check that, for each multi-index $(i_1, \ldots, i_m)$ of length $m \geq 1$ and $(\alpha_1, \ldots, \alpha_{2s}) \in \mathbb{R}^{2s}$,

$$u_{i_1 \cdots i_m}(\alpha_1, \ldots, \alpha_{2s}) = \sum_{1 \leq j_1 \leq j_1^* \leq j_2 \leq \cdots \leq j_{m-1} \leq j_{m-1}^* \leq j_m \leq 2s} \alpha_{j_1}^{(i_1)} \cdots \alpha_{j_m}^{(i_m)}. \tag{56}$$

Obviously, each $u_{i_1 \cdots i_m}$ can be seen as a real-valued function defined on the set

$$\{(\alpha_1, \ldots, \alpha_{2s}) \in \mathbb{R}^{2s} \ : \ s \geq 1\}. \tag{57}$$

Observe that each $u_{i_1 \cdots i_m}(\alpha_1, \ldots, \alpha_{2s})$ is a polynomial of degree $n = i_1 + \cdots + i_m$ in the variables $\alpha_1, \ldots, \alpha_{2s}$.

Now, the order conditions of the composition scheme (35) can be obtained by comparing the series (55) with $\exp(hF)$, that is, (44). Since $X_1 = F$, as the basic integrator χ_h is assumed to be of order 1, we have that the method is of order r if for each multi-index $(i_1, \ldots, i_m)$ with $i_1 + \cdots + i_m = n \leq r$,

$$u_{i_1 \cdots i_m}(\alpha_1, \ldots, \alpha_{2s}) = \left\{ \begin{array}{lcl} \frac{1}{n!} & \text{if} & (i_1, \ldots, i_m) = (1, \ldots, 1), \\ 0 & \text{otherwise.} & \end{array} \right. \tag{58}$$

However, such order conditions are not independent. For instance, it can be checked that

$$u_{11} = \frac{1}{2}(u_1^2 + u_2), \qquad u_{21} = -u_{12} + u_3 + u_1 u_2, \qquad u_{111} = \frac{1}{6}u_1^3 + \frac{1}{2}u_{12} + \frac{1}{3}u_3,$$

which implies that the order conditions (58) for u_{11}, u_{12}, u_{111} are fulfilled provided that the conditions for u_1, u_2, u_3, u_{12} hold.

A set of independent order conditions can be obtained as follows. Consider the lexicographical order $<$ (i.e., the order used when ordering words in the dictionary) on the set of multi-indices. A multi-index $(i_1, \ldots, i_m)$ is a Lyndon multi-index if $(i_1, \ldots, i_k) < (i_{k+1}, \ldots, i_m)$ for each $1 \leq k < m$. For each $n \geq 1$, we denote as L_n the set of functions $u_{i_1 \cdots i_m}$ such that $(i_1, \ldots, i_m)$ is a Lyndon multi-index satisfying that $i_1 + \cdots + i_m = n$. The first sets L_n are

$$\begin{aligned} L_1 &= \{u_1\}, \quad L_2 = \{u_2\}, \quad L_3 = \{u_{12}, u_3\}, \quad L_4 = \{u_{112}, u_{13}, u_4\}, \\ L_5 &= \{u_{1112}, u_{113}, u_{122}, u_{14}, u_{23}, u_5\}. \end{aligned}$$

In particular, we have

$$\begin{aligned} u_1(\alpha_1, \ldots, \alpha_{2s}) &= \sum_{j=1}^{2s} \alpha_j, \\ u_2(\alpha_1, \ldots, \alpha_{2s}) &= \sum_{j=1}^{2s} (-1)^j \alpha_j^2, \\ u_3(\alpha_1, \ldots, \alpha_{2s}) &= \sum_{j=1}^{2s} \alpha_j^3, \\ u_{12}(\alpha_1, \ldots, \alpha_{2s}) &= \sum_{j_2=1}^{2s} (-1)^{j_2} \alpha_{j_2}^2 \sum_{j_1=1}^{j_2^*} \alpha_{j_1}, \\ u_4(\alpha_1, \ldots, \alpha_{2s}) &= \sum_{j=1}^{2s} (-1)^j \alpha_j^4, \\ u_{13}(\alpha_1, \ldots, \alpha_{2s}) &= \sum_{j_2=1}^{2s} \alpha_{j_2}^3 \sum_{j_1=1}^{j_2^*} \alpha_{j_1}, \\ u_{112}(\alpha_1, \ldots, \alpha_{2s}) &= \sum_{j_3=1}^{2s} (-1)^{j_3} \alpha_{j_3}^2 \sum_{j_2=1}^{j_3^*} \alpha_{j_2} \sum_{j_1}^{j_2^*} \alpha_{j_1}, \end{aligned}$$

and so on.

We can finally state the following result [25]: Given $(\alpha_1, \ldots, \alpha_{2s})$, the integrator (35) is of order r for arbitrary ODE systems (1) and arbitrary consistent integrators χ_h if and only if $\alpha_1 + \cdots + \alpha_{2s} = 1$ (i.e. $u_1(\alpha_1, \ldots, \alpha_{2s}) = 1$) and

$$\forall\, u \in \bigcup_{n \geq 2}^{r} L_n, \qquad u(\alpha_1, \ldots, \alpha_{2s}) = 0. \tag{59}$$

Furthermore, such order conditions are independent to each other if arbitrary sequences $(\alpha_1, \ldots, \alpha_{2s})$ of coefficients of the method are considered.

3.3 Order conditions of compositions methods with symmetry

The order conditions are simplified for $(2s)$-tuplas $(\alpha_1, \dots, \alpha_{2s})$ such that

$$\alpha_{2s-j+1} = \alpha_j, \qquad \text{for all } j. \tag{60}$$

It is easy to check that the simplifying assumption (60) implies that the composition integrator (35) is time-symmetric (i.e., $\psi_h^* = \psi_h$). In that case, only the conditions for $u \in L_n$ with odd n remain independent.

The order conditions can be alternatively simplified by requiring that

$$\alpha_{2j} = \alpha_{2j-1}, \quad \forall\, j, \tag{61}$$

in which case, only the conditions for Lyndon multi-indices $(i_1, \dots, i_m)$ with odd $i_1, \dots, i_m$ are required. The simplifying assumption (61) means that the composition integrator (35) can be rewritten as

$$\psi_h = \mathcal{S}_{h\beta_s}^{[2]} \circ \cdots \circ \mathcal{S}_{h\beta_1}^{[2]}, \tag{62}$$

where $\beta_j = 2\alpha_{2j}$ and $\mathcal{S}_h^{[2]}$ is the self-adjoint second order integrator $\mathcal{S}_h^{[2]} = \chi_{h/2} \circ \chi_{h/2}^*$.

The order conditions are thus considerably reduced if one considers composition methods satisfying both assumptions (60)–(61), that is, methods of the form (62) satisfying that

$$\beta_j = \beta_{s-j+1}, \quad \forall\, j. \tag{63}$$

Schemes of this form can be dubbed as symmetric compositions of symmetric schemes. For instance, for order $r \geq 6$ one has the conditions

$$\sum_{j=1}^{2s} \alpha_j = 1, \qquad \sum_{j=1}^{2s} \alpha_j^3 = 0, \qquad \sum_{j=1}^{2s} \alpha_j^5 = 0, \qquad \sum_{j_3=1}^{2s} \alpha_{j_3}^3 \sum_{j_2=1}^{j_3^*} \alpha_{j_2} \sum_{j_1}^{j_2^*} \alpha_{j_1} = 0$$

in terms of the α_i coefficients (the actual expressions in terms of β_i are slightly more involved). In Table 1 we display for each $k \geq 1$ the number n_k of Lyndon multi-indices $(i_1, \dots, i_m)$ with $i_1 + \cdots + i_m = k$, and the number m_k of Lyndon multi-indices $(i_1, \dots, i_m)$ with $i_1 + \cdots + i_m = k$ and odd indices $i_1, \dots, i_m$. Thus, the number of independent conditions to guarantee that the general composition integrator (35) is at least of order r is $n_1 + \cdots + n_r$, while in the case of the composition (62) based on a symmetric second order integrator $\mathcal{S}_h^{[2]}$ (or equivalently, a composition integrator (35) with the additional symmetry condition (61)), the number of independent order conditions is $m_1 + \cdots + m_r$. If time-symmetry is imposed in the method (35) (resp. (62)) by the additional assumption (60) (resp. (63)), then there are $n_1 + n_3 + \cdots + n_{2l-1}$ (resp. $m_1 + m_3 + \cdots + m_{2l-1}$) independent conditions that guarantee order at least $r = 2l$.

k	1	2	3	4	5	6	7	8	9	10	11
n_k	1	1	2	3	6	9	18	30	56	99	186
m_k	1	0	1	1	2	2	4	5	8	11	17

Table 1: The numbers n_k and m_k of independent order conditions for general composition methods (35) and for compositions (62) of a basic time-symmetric method, respectively.

3.4 Relation among different sets of order conditions of composition methods

In [63], a set of independent necessary and sufficient order conditions is given using labelled rooted trees (see also [41]). A family of sets $\{\mathcal{T}_n\}_{n=1,2,\ldots}$ of functions defined on the set (57) is identified such that the integrator (35) is of order r if and only if $\alpha_1 + \cdots + \alpha_{2s} = 1$ and

$$\forall\, u \in \bigcup_{n\geq 2}^{r} \mathcal{T}_n, \qquad u(\alpha_1, \ldots, \alpha_{2s}) = 0. \tag{64}$$

Each $u(\alpha_1, \ldots, \alpha_{2s})$ with $u \in \mathcal{T}_n$ is (as in the case where $u \in \mathcal{L}_n$), a polynomial of homogeneous degree n. In particular,

$$\mathcal{T}_1 \;=\; \{u_1\}, \quad \mathcal{T}_2 = \{u_2\}, \quad \mathcal{T}_3 = \{u_{21}, u_3\}, \quad \mathcal{T}_4 = \{v_{211}, u_{31}, u_4\},$$

where the functions of the form $u_{i_1\cdots i_m}$ are defined in (56), and

$$v_{211}(\alpha_1, \ldots, \alpha_{2s}) \;=\; \sum_{j_2=1}^{2s} (-1)^{j_2} \alpha_{j_2}^2 \left(\sum_{j_1=1}^{j_2^*} \alpha_{j_1} \right)^2 .$$

As shown in [25], the order conditions (64) are equivalent to the conditions (59), as both $\cup_{n\geq 1} L_n$ and $\cup_{n\geq 1}\mathcal{T}_n$ generate the same graded algebra $\mathcal{H} = \bigoplus_{n\geq 1} \mathcal{H}_n$ of functions on the set (57) (for each $u \in \mathcal{H}_n$, $u(\alpha_1, \ldots, \alpha_{2s})$ is a polynomial of homogeneous degree n, actually, a linear combination of polynomials $u_{i_1\cdots i_m}$ of homegeneous degre $n = i_1 + \cdots + i_m$). For instance, it can be seen that

$$v_{211} = 2u_{211} - u_{22} = 2(u_{112} - u_{13} + u_1 u_{12} + u_3 u_1) + u_1^2 u_2 + \frac{1}{2}(u_4 - u_2^2).$$

Finding an independent set of order conditions for composition integrators is equivalent to finding a set of functions of homogeneous degree that generate the algebra $\mathcal{H}$ (see [25]) for more details.

Of course, the functions $w_{i_1\cdots i_m}$ in (51) obtained when deriving the order conditions of composition integrators by repeated use of the BCH formula also belong to the same algebra of functions. For instance, $w_n = u_n$, and $w_{12} = u_{12} - u_3 - u_1 u_2$.

Recall that Theorem 1 characterizes the order conditions of splitting integrators of the form (36), where the ODE (1) is split in two parts (47), in terms of the order conditions of composition integrators (35). Actually, the polynomials $v_a, v_b, v_{ba}, v_{baa}, v_{bab}, v_{baaa}, \dots$ (on the parameters a_i, b_i) in (53) can be rewritten as linear combinations of the polynomials (on the parameters α_i) in (56) provided that (37) and $v_a = v_b = u_1$ hold. In particular, it can be seen that

$$\begin{aligned}
v_{ab} &= \frac{u_2}{2}, \\
v_{abb} &= \frac{1}{12}\left(-u_3 - 3u_{12} + 3u_{21}\right), \\
v_{aba} &= \frac{1}{12}\left(u_3 - 3u_{12} + 3u_{21}\right), \\
v_{abbb} &= \frac{1}{12}\left(u_{22} - u_{31} + u_{112} - 2u_{121} + u_{211}\right), \\
v_{abba} &= \frac{1}{24}\left(-u_4 - 2u_{13} + 4u_{22} - 2u_{31} + 4u_{112} - 8u_{121} + 4u_{211}\right), \\
v_{abaa} &= \frac{1}{12}\left(-u_{13} + u_{22} + u_{112} - 2u_{121} + u_{211}\right).
\end{aligned}$$

3.5 Negative time steps

It has been noticed that some of the coefficients in splitting schemes (36) are negative when the order $r \geq 3$. In other words, the methods always involve stepping backwards in time. This constitutes a problem when the differential equation is defined in a semigroup, as arises sometimes in applications, since then the method can only be conditionally stable [57]. Also schemes with negative coefficients may not be well-posed when applied to PDEs involving unbounded operators.

The existence of backward fractional time steps in this class of methods is unavoidable, as shown in [35, 75, 79]. In fact, it can be established in an elementary way by virtue of the relationship between the order conditions of schemes (36) and (35) stated in Theorem 1 [4]: Any splitting method of the form (36) that has order $r \geq 3$ neccesarily must fullfil the condition

$$u_3(\alpha_1, \dots, \alpha_{2s}) = \sum_{i=1}^{2s} \alpha_i^3 = \sum_{i=1}^{s} (\alpha_{2i-1}^3 + \alpha_{2i}^3) = 0, \tag{65}$$

with coefficients α_j obtained from the relations (37). Since, for all $x, y \in \mathbb{R}$, it is true that $x^3 + y^3 < 0$ implies $x + y < 0$, then there must exist some $i \in \{1, \dots, s\}$ in the sum of (65) such that

$$\alpha_{2i-1}^3 + \alpha_{2i}^3 < 0 \qquad \text{and thus} \qquad \alpha_{2i-1} + \alpha_{2i} = a_i < 0.$$

Obviously, one can also write (by taking $\alpha_0 = 0$)

$$u_3(\alpha_1, \dots, \alpha_{2s}) = \sum_{i=0}^{2s+1} \alpha_i^3 = \sum_{i=1}^{s+1} (\alpha_{2i-1}^3 + \alpha_{2i-2}^3) = 0$$

just by grouping terms in a different way, and thus, by repeating the argument, there must exist some $j \in \{1, \ldots, s+1\}$ such that

$$\alpha_{2j-1} + \alpha_{2j-2} = b_j < 0.$$

This proof shows clearly the origin of the existence of backward time steps: the equation $u_3 = 0$ can be satisfied only if at least one a_i and one b_i are negative. According to this conclusion, any splitting method of the form (36) verifying the order condition $u_3 = 0$ has necessarily some negative coefficient a_i and also some negative b_i.

3.6 Near-integrable systems

In Hamiltonian dynamics one often encounters systems whose Hamiltonian function H is a small perturbation of an exactly integrable Hamiltonian H_0, that is $H = H_0 + \varepsilon H_1$ with $\varepsilon \ll 1$. The perturbed Kepler problem (20) belongs to this category of near-integrable Hamiltonian systems. The gravitational N-body problem (21), when using Jacobi coordinates, also falls within this class of problems. In that case, H_0 represents the Keplerian motion and εH_1 the mutual perturbations of the bodies on one another [86].

More generally, let us consider an ODE system

$$x' = f^{[a]}(x) + \varepsilon f^{[b]}(x), \tag{66}$$

containing a small parameter $|\varepsilon| \ll 1$. If the exact h-flows $\varphi_h^{[a]}$ and $\varphi_h^{[b]}$ of $x' = f^{[a]}(x)$ and $x' = \varepsilon\, f^{[b]}(x)$ respectively can be efficiently computed, then a scheme ψ_h of the form (36) can perform particularly well provided that the coefficients a_i, b_i are appropriately chosen. To see this, consider the Lie derivatives (48) of $f^{[a]}$ and $f^{[b]}$ respectively, so that the corresponding series $\Psi(h)$ (49) of differential operators associated to the scheme (36) becomes

$$\Psi(h) = \mathrm{e}^{b_1 h \varepsilon F^{[b]}}\, \mathrm{e}^{a_1 h F^{[a]}} \cdots \mathrm{e}^{b_s h \varepsilon F^{[b]}}\, \mathrm{e}^{a_s h F^{[a]}}\, \mathrm{e}^{b_{s+1} h \varepsilon F^{[b]}}.$$

Successive application of the BCH formula then leads to (53) with $F^{[b]}$ replaced by $\varepsilon F^{[b]}$, that is

$$\begin{aligned}\log(\Psi(h)) \;&=\; h v_a F^{[a]} + \varepsilon(h v_b F^{[b]} + h^2 v_{ab} F^{[ab]} + h^3 v_{aba} F^{[aba]} + h^4 v_{abaa} F^{[abaa]}) \\ &\quad + \varepsilon^2 (h^3 v_{abb} F^{[abb]} + h^4 v_{abba} F^{[abba]}) + \varepsilon^3 h^4 v_{abbb} F^{[abbb]} + \mathcal{O}(\varepsilon h^5).\end{aligned}$$

In practical applications one usually has $\varepsilon \ll h$ (or at least $\varepsilon \approx h$), so that one is mainly interested in eliminating error terms with small powers of ε. For instance, if the coefficients a_i, b_i of the splitting methods are chosen in such a way that

$$v_a = 1 = v_b, \quad v_{ab} = v_{aba} = v_{abaa} = v_{abb} = 0,$$

then

$$\log(\Psi(h)) - hF = \mathcal{O}(\varepsilon h^5 + \varepsilon^2 h^4),$$

where $F = F^{[a]} + \varepsilon F^{[b]}$. More generally, one is interested in designing methods such that [12]

$$\log(\Psi(h)) - hF = \mathcal{O}(\varepsilon h^{s_1+1} + \varepsilon^2 h^{s_2+1} + \varepsilon^3 h^{s_3+1} + \cdots + \varepsilon^m h^{s_m+1}). \tag{67}$$

Observe that s_1 is the order of consistency the method would have in the limit $\varepsilon \to 0$. It is relatively easy to eliminate errors of order εh^k because there is only one such term for each order k, namely $h^k \varepsilon\, v_{aba\cdots a} F^{[aba\cdots a]}$ (with $F^{[aba\cdots a]} = [[\cdots[[F^{[a]}, F^{[b]}], F^{[a]}]\ldots], F^{[a]}])$.

If one is interested in designing methods that approximate the exact solution up to higher powers of ε, more terms have to be considered. In particular, there are $\lfloor \frac{1}{2}(k-1) \rfloor$ terms of order $\mathcal{O}(\varepsilon^2 h^k)$ and $\lfloor \frac{1}{6}(k-1)(k-2) \rfloor$ terms of order $\mathcal{O}(\varepsilon^3 h^k)$ [53].

3.7 Runge–Kutta–Nyström methods

Suppose now that one is interested in integrating numerically second-order ODE systems of the form

$$y'' = g(y), \tag{68}$$

where $y \in \mathbb{R}^d$ and $g : \mathbb{R}^d \longrightarrow \mathbb{R}^d$. In this case it is still possible to use schemes (36) applied to the equivalent first-order ODE system. More specifically, introducing the new variables $x = (y, v)$, with $v' = y$, and the maps $f^{[a]} : \mathbb{R}^{2d} \to \mathbb{R}^{2d}$ and $f^{[b]} : \mathbb{R}^{2d} \to \mathbb{R}^{2d}$ defined as

$$f^{[a]}(y, v) = (v, 0), \qquad f^{[b]}(y, v) = (0, g(y)), \tag{69}$$

equation (68) can be rewritten as $x' = f^{[a]}(x) + f^{[b]}(x)$. Then, the splitting scheme (36) can be efficiently implemented, as the exact h-flows $\varphi_h^{[a]}$ and $\varphi_h^{[b]}$ of $x' = f^{[a]}(x)$ and $x' = f^{[b]}(x)$ are simply given by

$$\begin{array}{rcl} \varphi_h^{[a]}(y, v) & = & (y + hv, v), \\ \varphi_h^{[b]}(y, v) & = & (y, v + hg(y)). \end{array} \tag{70}$$

It is not difficult to check that the splitting schemes of the form (36) are particular instances of Runge–Kutta–Nyström (RKN) methods (see for instance [41]).

One of the most important applications of this class of schemes is the study of Hamiltonian systems of the form $H(q, p) = T(p) + V(q)$, where the kinetic energy $T(p)$ is quadratic in the momenta p, i.e., $T(p) = \frac{1}{2} p^T M p$ for a symmetric square constant matrix M, and $V(q)$ is the potential. In that case, the corresponding Hamiltonian system can be written in the form (68) with $y = q$, $y' = v = Mp$, and $g(y) = -\nabla V(y)$.

Although a splitting integrator (36) designed for arbitrary ODE systems $x' = f(x)$ split into two parts (47) will perform well when applied to a second order ODE system of the form (68) with the splitting (69), much more efficient methods can be designed in that case [56, 19]. The main point here is that in the

k	2	3	4	5	6	7	8	9	10	11
n_k	1	2	3	6	9	18	30	56	99	186
l_k	1	2	2	4	5	10	14	25	39	69

Table 2: The numbers l_k of independent order conditions (at order k) for splitting methods in the RKN case compared to the numbers n_k of conditions in the general case.

case (69) (we will refer to that as the RKN case), $[[[F^{[b]}, F^{[a]}], F^{[b]}], F^{[b]}] = 0$ identically. This is equivalent to $F^{[babb]} = 0$ in (53), which introduces some linear dependencies among higher order terms in the expansion of $\log(\Psi(h))$ (see [59] for a detailed study). This means that the characterization given in Theorem 1 for a splitting integrator (36) to be of order r (for $r \geq 4$) is no longer applicable if one restricts to the case (69). In Table 2, the number of necessary and sufficient independent order conditions for a splitting method (36) to be of order r in the RKN case is compared to the general case: For arbitrary systems split into two parts (47), there are $2 + n_2 + \cdots + n_r$ independent conditions (including the two consistency conditions $v_a = v_b = 1$), while in the RKN case, the number of independent order conditions is $2 + l_2 + \cdots + l_r$. Unfortunately, up to order three the order conditions are the same in both cases and then, the results for negative time steps still apply.

Since the reduction in the number of order conditions is due to the fact that $f^{[a]}(y, v)$ is linear in v, it is immediate to see that the methods obtained in this way also apply to the more general problem $f^{[a]}(y, v) = (w_1(x)v + w_2(y), w_3(x)v + w_4(y))$, which includes the system

$$y'' = My' + g_1(y) + g_2(y). \tag{71}$$

Here splitting RKN methods are useful if the reduced problem $y'' = My' + g_1(y)$ (i.e., $f^{[a]}(y, v) = (v, Mv + g_1(y))$) is easily solvable. For Hamiltonian systems, this generalization corresponds to $H(q, p) = T(q, p) + V(q)$, where $T(q, p) = \frac{1}{2}p^T M(q)p + f^T(q)p + W(q)$. Obviously, the exact solution for $T(q, p)$ is only known for some particular cases, e.g. if T corresponds to the Kepler problem in (20) or (21) or to the harmonic oscillator in (19) or (22).

It is interesting to note that in quantum mechanics the kinetic and potential energy verify analogue commutator rules to classical mechanics and then RKN methods can also be used. If applied, for instance to problems (26) and (27), one should keep in mind that in the resulting composition method $\varphi_h^{[a]}$ must correspond to the kinetic part.

We should also remark that, if the Hamiltonian function $H(p, q) = \frac{1}{2}p^T Mp + V(q)$ is such that in addition $V(q) = \frac{1}{2}q^T Nq$ (i.e., it corresponds to the generalized harmonic oscillator (18)), then the number of order conditions reduces drastically: it is not difficult to see that there is only one independent condition to increase the order from $r = 2k - 1$ to $r = 2k$, and only two to increase the order from $r = 2k$ to $r = 2k + 1$ (see [9] for more details). We will return later to this system and take profit of its special features to design specially adapted splitting methods.

4 Additional techniques to reduce the number of order conditions

4.1 Methods with modified potentials

The splitting method (36) can be generalized by composing the exact flows of other vector fields in addition to $F^{[a]}$ and $F^{[b]}$, provided that they lie on the Lie algebra generated by $F^{[a]}$ and $F^{[b]}$. For instance, one could consider compositions that, in addition to $\varphi_h^{[a]}$ and $\varphi_h^{[b]}$, use the h-flow $\varphi_h^{[abb]}$ of the vector field $F^{[abb]} = [[F^{[a]}, F^{[b]}], F^{[b]}]$. To illustrate this fact, consider the composition

$$\psi_h = \varphi_{h/6}^{[b]} \circ \varphi_{h/2}^{[a]} \circ \varphi_{h/3}^{[b]} \circ \varphi_{-h/72}^{[abb]} \circ \varphi_{h/3}^{[b]} \circ \varphi_{h/2}^{[a]} \circ \varphi_{h/6}^{[b]}. \tag{72}$$

The scheme (72), constructed in [26, 47], is of order four. Indeed, by repeated application of the BCH formula to

$$\Psi(h) = \mathrm{e}^{\frac{h}{6}F^{[b]}} \mathrm{e}^{\frac{h}{2}F^{[a]}} \mathrm{e}^{\frac{h}{3}F^{[b]}} \mathrm{e}^{-\frac{h}{72}F^{[abb]}} \mathrm{e}^{\frac{h}{3}F^{[b]}} \mathrm{e}^{\frac{h}{2}F^{[a]}} \mathrm{e}^{\frac{h}{6}F^{[b]}}$$

one can check that $\Psi(h) = \mathrm{e}^{h(F^{[a]}+F^{[b]})+\mathcal{O}(h^5)}$.

Recall that, although the h-flows $\varphi_h^{[a]}$ and $\varphi_h^{[b]}$ of the vector fields $F^{[a]}$ and $F^{[b]}$ are by assumption computed easily, this is not necessarily the case for the h-flow $\varphi_h^{[abb]}$ of the vector field $F^{[abb]}$. However, in the RKN case (68) considered in Subsection 3.7, where $f^{[a]}$ and $f^{[b]}$ are of the form (69), the h-flow of $F^{[abb]}$ is of the form

$$\varphi_h^{[abb]}(y, v) = (y,\, v + h^3 g^{[3]}(y)), \qquad \text{where} \quad g^{[3]}(y) = 2g'(y)g(y).$$

This shows in addition that $\varphi_h^{[b]}$ and $\varphi_h^{[abb]}$ commute, and that in particular, for arbitrary $b_j, c_j \in \mathbb{R}$,

$$\varphi_{b_j\,h}^{[b]} \circ \varphi_{c_j\,h}^{[abb]} \circ \varphi_{b_j\,h}^{[b]}(y, v) = (y,\, v + 2hb_j\, g(y) + h^3 c_j\, g'(y)g(y)), \tag{73}$$

which is precisely the h-flow of the vector field $2b_j\, F^{[b]} + c_j h^2 F^{[abb]}$. It thus makes sense to construct methods defined as compositions of $\varphi_{a_j h}^{[a]}$ and maps of the form (73) for $j = 1, \ldots, s$.

For Hamiltonian systems $H(p, q) = T(p) + V(q)$ with quadratic kinetic energy $T(p) = \frac{1}{2}p^T M p$, the vector field $F^{[abb]} = [[F^{[a]}, F^{[b]}], F^{[b]}]$ is the vector field associated to the Hamiltonian function $(\nabla V)^T \nabla V$, which only depends on the position vector q. Thus, (73) is just the h-flow of the system with Hamiltonian function

$$2b_j\, V(q) + c_j\, h^2 (\nabla V(q))^T \nabla V(q), \tag{74}$$

which reduces to the potential $V(q)$ of the system for $b_j = 1/2$ and $c_j = 0$. This explains the term ‘splitting methods with modified potentials’ used in the recent literature [51, 71, 87] to refer to splitting methods obtained by composing the h-flows of T and modified potentials of the form (74).

This procedure can be generalized by considering "modified potentials" of higher degree in h. In particular, the flow $\varphi_h^{[abbab]}$ of the vector field

$$F^{[abbab]} = [[[F^{[a]}, F^{[b]}], F^{[b]}], [F^{[a]}, F^{[b]}]], \tag{75}$$

is of the form $\varphi_h^{[abbab]}(y, v) = (y, v + h^5 g^{[5]}(y))$, and similarly for the vector fields

$$\begin{aligned} F^{[abbabab]} &= [[[[F^{[a]}, F^{[b]}], F^{[b]}], [F^{[a]}, F^{[b]}]], [F^{[a]}, F^{[b]}]], \\ F^{[abbaabb]} &= [[[[F^{[a]}, F^{[b]}], F^{[b]}], F^{[a]}], [[F^{[a]}, F^{[b]}], F^{[b]}]], \end{aligned} \tag{76}$$

with $h^5 g^{[5]}(y)$ replaced by $h^7 g^{[7,1]}(y)$ and $h^7 g^{[7,2]}(y)$ respectively. The functions $g^{[5]}, g^{[7,1]}, g^{[7,2]}$ can be written in terms of g and its partial derivatives (see [13] for more details).

In some applications, the simultaneous evaluation of $g(y)$, $g^{[3]}(y)$, $g^{[5]}(y)$, $g^{[7,1]}(y)$ and $g^{[7,2]}(y)$ is not substantially more expensive in terms of computational cost than the evaluation of $g(y)$ alone. In that case, by replacing in the scheme (36) each $\varphi_{b_i h}^{[b]}$ by the h-flow of

$$\begin{aligned} C_h(b_i, c_i, d_i, e_{i1}, e_{i2}) \equiv\ & b_i\, F^{[b]} + h^2 c_i\, F^{[abb]} + h^4 d_i\, F^{[abbab]} \\ & + h^6 (e_{i,1} F^{[abbabab]} + e_{i,2}\, F^{[abbaabb]}) \end{aligned} \tag{77}$$

additional free parameters are introduced to the scheme without increasing too much the computational cost, which allows the construction of more efficient integrators.

Of course, this can be further generalized by considering more general nested commmutators of $F^{[a]}$ and $F^{[b]}$ that gives rise to "modified potentials". In that case, higher degree commutators afected by higher powers of h should be added in (77).

Notice that in this case the coefficients a_i, b_i have not to satisfy all the order conditions at order $r \geq 3$ and then, the results for negative time steps do not apply in this case. As a result, schemes with positive coefficients do exist. In this case, negative coefficients appear in methods of order six [27].

4.2 Methods with processing

Recently, the processing technique has been used to find composition methods requiring less evaluations than conventional schemes of order r. The idea consists in enhancing an integrator ψ_h (the *kernel*) with a parametric map $\pi_h : \mathbb{R}^D \longrightarrow \mathbb{R}^D$ (the *post-processor*) as

$$\hat{\psi}_h = \pi_h \circ \psi_h \circ \pi_h^{-1}. \tag{78}$$

Application of n steps of the new (and hopefully better) integrator $\hat{\psi}_h$ leads to

$$\hat{\psi}_h^n = \pi_h \circ \psi_h^n \circ \pi_h^{-1},$$

which can be considered as a h-dependent change of coordinates in phase space. Observe that processing is advantageous if $\hat{\psi}_h$ is a more accurate method than ψ_h and, either the cost of π_h is negligible or frequent output is not required, since in that case, it provides the accuracy of $\hat{\psi}_h$ at essentially the cost of the least accurate method ψ_h.

The simplest example of a processed integrator is provided in fact by the Störmer–Verlet method. As a consequence of the group property of the exact flow, we have

$$\begin{aligned} \mathcal{S}_h^{[2]} &= \varphi_{h/2}^{[a]} \circ \varphi_h^{[b]} \circ \varphi_{h/2}^{[a]} = \varphi_{h/2}^{[a]} \circ \varphi_h^{[b]} \circ \varphi_h^{[a]} \circ \varphi_{-h}^{[a]} \circ \varphi_{h/2}^{[a]} \\ &= \varphi_{h/2}^{[a]} \circ \chi_h \circ \varphi_{-h/2}^{[a]} = \pi_h \circ \chi_h \circ \pi_h^{-1} \end{aligned} \tag{79}$$

with $\pi_h = \varphi_{h/2}^{[a]}$ and the symplectic Euler method $\chi_h = \varphi_h^{[b]} \circ \varphi_h^{[a]}$. Hence, applying the basic integrator $\chi_h = \varphi_h^{[b]} \circ \varphi_h^{[a]}$ with processing yields a second order of approximation.

Although initially proposed for Runge–Kutta methods [18], the processing technique has proved its usefulness mainly in the context of geometric numerical integration [41], where constant step-sizes are widely employed.

We say that the method ψ_h is of *effective order* r if a post-processor π_h exists for which $\hat{\psi}_h$ is of (conventional) order r [18], that is,

$$\pi_h \circ \psi_h \circ \pi_h^{-1} = \varphi_h + \mathcal{O}(h^{r+1}).$$

Hence, as the previous example shows, the basic splitting $\varphi_h^{[b]} \circ \varphi_h^{[a]}$ is of effective order 2. Obviously, a method of order r is also of effective order r (taking $\pi_h = \mathrm{id}$) or higher, but the converse is not true in general.

The analysis of the order conditions of the method $\hat{\psi}_h$ shows that many of them can be satisfied by π_h, so that ψ_h must fulfill a much reduced set of restrictions [6, 11]. For instance, if the kernel is defined as (35) with a basic first order integrator χ_h and the post-processor is similarly defined as

$$\pi_h = \chi_{\gamma_{2m} h} \circ \chi^*_{\gamma_{2s-1} h} \circ \cdots \circ \chi_{\gamma_2 h} \circ \chi^*_{\gamma_1 h} \tag{80}$$

then, conditions

$$u_1(\alpha) = 1, \quad u_2(\alpha) = u_3(\alpha) = u_4(\alpha) = 0 \tag{81}$$

guarantee that the kernel ψ_h is of effective order four. If in addition the post-processor (80) satisfies

$$u_1(\gamma) = 0, \quad u_2(\gamma) = u_{12}(\alpha), \quad u_3(\gamma) = u_{13}(\alpha), \quad u_{12}(\gamma) = u_{112}(\alpha),$$

then the processed integrator (78) has conventional order four. Here, we use the notation $\alpha = (\alpha_1, \ldots, \alpha_{2s})$ and $\gamma = (\gamma_1, \ldots, \gamma_{2m})$ for the coefficients of the kernel and the post-processor respectively. If in addition the following conditions are fulfilled by the coefficients of the kernel,

$$u_5(\alpha) = u_{23}(\alpha) = 0, \qquad 2u_{122}(\alpha) + u_{14}(\alpha) + u_{12}(\alpha)^2 = 0,$$

then the kernel ψ_h has at least effective order five. In that case, the processed method (78) achieves conventional order five if in addition, the equalities

$$\begin{aligned} u_4(\gamma) &= u_{14}(\alpha), \quad u_{13}(\gamma) = u_{113}(\alpha), \\ u_{112}(\gamma) &= u_{1112}(\alpha) + \frac{1}{2}u_{12}(\alpha)^2 - \frac{1}{2}u_{112}(\alpha) - \frac{1}{6}u_{12}(\alpha) \end{aligned}$$

hold for the coefficients of the post-processor (80).

Thus, the number and complexity of the conditions to be verified by the coefficients α_j of a kernel of the form (35) is notably reduced. Highly efficient processed composition methods that take advantage of that have been proposed [11, 57]. Nevertheless, when both the kernel ψ_h and the post-processor π_h are constructed as a composition of the form (35) (or (36)), the use of the resulting processed scheme is not recommended in situations where intermediate results are required at each step. Indeed, the total number of compositions per step in a processed method (78) of that form is typically higher than for a non-processed method of comparable accuracy.

To overcome this drawback, in [6] a technique has been developed for obtaining approximations to the post-processor at virtually cost free and without loss of accuracy. The key idea is to replace π_h by a new map $\tilde{\pi}_h \simeq \pi_h$ obtained from the intermediate stages in the computation of ψ_h. The post-processor π_h can safely be replaced by an approximation $\tilde{\pi}_h$, since the error introduced by the cheap approximation $\tilde{\pi}_h$ is of a purely local nature [6] (it is not propagated along the evolution, contrarily to the error in π_h^{-1}).

In [6], a general study of the number of independent effective order order conditions versus the number of conventional order conditions is presented. In particular, it is shown that, in the case of kernels of the form (35), the number of conditions to increase the effective order of the kernel from $k > 1$ (resp. $k = 1$) to $k + 1$ is $n_{k+1} - n_k$ (resp. $n_2 - n_1 + 1$), where each n_k is the cardinal of L_k, that is, the number Lyndon multi-indices of degree k. Thus, whereas the total number of independent conditions to achieve conventional order r is $n_1 + \cdots + n_r$, only $1 + n_r$ conditions have to be imposed to the kernel for effective order r. If the kernel (35) is time-symmetric (i.e., if its coefficients satisfy (60)), then there are $N_r = \sum_{i=1}^{q} n_{2j-1}$ independent conditions for order $r = 2q$, and $N_r^* = n_1 + \sum_{i=1}^{q-1}(n_{2j+1} - n_{2j})$ conditions for effective order $r = 2q$. A similar situation occurs for the total numbers M_r and M_r^* of conventional and effective order conditions of symmetric kernels of the form (62) with (63) (where the n_k are replaced by the number m_k in Table 1). That also happens to be true for symmetric kernels of the form (36), both in the general case (which is essentially equivalent to the case of kernels of the form (35)) and in the RKN case considered in Subsection 3.7. In Table 3, the total number of conditions for conventional order $r = 2q$ for symmetric kernels is compared with the total number of effective order conditions in three kinds of integrators: (i) (N_r, N_r^*) for composition (35) of a basic first order integrator and its adjoint, (ii) (M_r, M_r^*) for compositions (62) of a symmetric second order basic integrator, (iii) (L_r, L_r^*) for splitting integrators in the RKN case.

r	2	4	6	8	10	12
N_r	1	3	9	27	83	269
N_r^*	1	2	5	14	40	127
M_r	1	2	4	8	16	33
M_r^*	1	2	3	5	8	14
L_r	2	4	8	18	43	112
L_r^*	2	3	5	10	21	51

Table 3: Number of conventional and effective order conditions for symmetric kernels: (i) (N_r, N_r^*) for composition integrators (35), (ii) (M_r, M_r^*) for compositions (62) of symmetric second order basic integrator, (iii) (L_r, L_r^*) for splitting integrators in the RKN case.

5 A collection of splitting methods

As we have mentioned before, splitting methods have found application in many different areas of science during the last decades. It is therefore not surprising that there is a large number of different schemes available in the literature. Sometimes, even the same method has been rediscovered several times in different contexts. Our aim in this section is to offer the reader a comprehensive overview of the existing methods, by classifying them into different families and giving the appropriate references where the corresponding coefficients can be found.

At this point it is important to remark that the efficiency of a method is measured by taking into account the computational cost required to achieve a given accuracy (we do not take into account the important property of the stability of the methods). For instance, given several methods of order r with different computational cost (usually measured as the number of stages or evaluations of the functions involved), the most efficient method does not necessarily correspond to the cheapest method. The extra cost of some methods can be compensated by an improvement in the accuracy obtained.

We next present a short review indicating the splitting methods which have been published in the literature at different orders, with different number of stages and for several families of problems.

Symmetric compositions of symmetric methods. As we pointed out in section 2.1, although by applying recursively the composition (30)-(31) it is possible to increase the order, the resulting methods are computationally expensive. To reduce the number of evaluations the more general composition (62) may be considered to achieve a given order r. If we choose symmetric compositions ($\beta_{s+1-i} = \beta_i$), then half of the parameters of the method are fixed, but the order conditions at even orders are automatically satisfied. In other words, the parameters of a (non-symmetric) method of order $r = 2k$ have to solve a system of $\sum_{i=1}^{2k} m_i$ equations (see Table 3), whereas for a symmetric composition $e_r = m_2 + m_4 + \cdots + m_{2k}$ order conditions are automatically satisfied if the order conditions at odd orders are fulfilled. In this way, only $M_r = m_1 + m_3 + \cdots + m_{2k-1}$ independent order conditions need to be imposed in the case of symmetric compositions. Due to this fact, the number of conditions to be solved (which is typically the bottleneck in the numerical search of methods) is reduced considerably when imposing symmetry. Furthermore, since $m_{2i-1} < m_{2i}$ then $M_r < e_r$ and symmetric compositions, in addition to having more favourable geometric properties (due to the time-symmetric property), usually require smaller number of stages than their non-symmetric counterparts. Taking into account the number M_r (resp. M_r^*) of independent conditions to achieve conventional order r (resp. effective order r) from Table 3, it is possible to determine the minimum number $s_r = 2M_r - 1$ of stages of the integrator (resp. the minimum number $k_r = 2M_r^* - 1$ of stages for the kernel) required by a method of order r (resp. effective order r)

In this way one has to solve a system of M_r or M_r^* nonlinear polynomial equations with the same number of unknowns β_i. The number of real solutions typically increase a good deal with r. In general, these equations have to be solved numerically and getting *all* solutions is a very challenging task, even for moderate values of r. Once a number of solutions for the parameters β_i have been obtained, there remains to select that solution one expects will give the best performance when applied on practical problems, typically by minimizing some objective function. What is the most appropriate objective function in this case? A frequently used criterion is to choose the solution which minimizes $C = \sum_{i=1}^{s} |\beta_i|$.

If one takes additional stages in (62), for instance $s = s_r + 2$, then one has an extra free parameter (notice the scheme is symmetric and two stages are required to introduce one parameter). By choosing β_1 as this free parameter, then it is clear that 1-parameter families of solutions are obtained. For instance, taking $\beta_1 = 0$ one has the previous solutions and by continuation it is possible to get several of these 1-parameter families of solutions, but this procedure does not guarantee to find all solutions.

Finally, one has to select that solution minimizing the value of C. Of course, additional stages can be introduced and the process is similar but technically much more involved. This objective function allows one to find very efficient methods involving additional stages, although the efficiency of methods with the same order but different number of stages cannot be compared from the value obtained for C.

When we stop including additional stages in the composition (62)? Two criteria are possible: (i) when one has enough stages available to achieve a higher order; (ii) when the performance of the actual methods constructed with additional stages do not show a significant improvement in numerical experiments.

For instance, the simple 4th-order scheme (29) can be improved just by the 5-stage generalized composition [78]

$$\mathcal{S}_{\alpha h}^{[2]} \circ \mathcal{S}_{\alpha h}^{[2]} \circ \mathcal{S}_{\beta h}^{[2]} \circ \mathcal{S}_{\alpha h}^{[2]} \circ \mathcal{S}_{\alpha h}^{[2]}, \tag{82}$$

Order			
4	6	8	10
3-[34, 30, 88, 78] **5**-[78, 54]	**7**-[88] **9**-[54, 45] **11-13**-[76]	**15**-[88, 81, 54, 45] **17**-[54, 45] **19-21**-[76] **24**-[20]	**31**-[81, 45, 41, 76] **33**-[45, 40, 83, 76] **35**-[40, 76]
P:3-17-[55]	**P:5-15**-[7]	**P:9-19**-[7]	**P:15-25**-[7]

Table 4: Symmetric compositions of symmetric methods published in the literature. We indicate the number of stages (in boldface) and the pertinent reference. Processed methods are preceded by **P**.

with $\alpha = 1/(4 - 4^{1/3})$, $\beta = 1 - 4\alpha$, as numerical experiments clearly indicate. This is a particular case of (62) where the value of C reaches a minimum and if we add two new stages with a new parameter then a 6th-order method can be obtained.

In Table 4 we collect some of the most relevant methods from the literature with different orders and number of stages. At each order, r, we label the methods by the number of stages **s** and the reference where this method can be found. We also include methods obtained by using the processing technique, which are referred as **P**:**s**, where s is the number of stages for the kernel. We write $\mathbf{s_1}$-$\mathbf{s_2}$ for indicating that methods from $\mathbf{s_1}$ up to $\mathbf{s_2}$ stages are analyzed in that particular reference.

Splitting into two parts. Composition of method and its adjoint. Next we review methods of the form (36) (for ODEs that can be split into two parts) and (35). It is important to emphasize that, although the order conditions for both classes of methods are equivalent, the optimization procedures carried out to identify the most efficient schemes may differ. In consequence, a particular method optimized for equations separable into two parts is not necessarily the best choice for a composition (35), although their performances are closely related.

Considering, as before, symmetric compositions, i.e., $a_{s+1-i} = a_i$, $b_{s+2-i} = b_i$ in (36) and $\alpha_{2s+1-i} = \alpha_i$ in (35), it is easy to verify, from Table 3, that the minimum number of stages required to get a method of order r is $s_r = N_r$ and of effective order r it is $k_r = N_r^*$.

Note that schemes of order six or higher require more stages than compositions (62), and only fourth-order methods seem promising. Nevertheless, one should recall that by including additional stages more efficient methods could be obtained. For instance, sixth-order methods require at least 9 stages (unless they are considered as composition of symmetric-symmetric methods in which case the 9 equations can be solved with only 7 unknowns) and the coefficients a_i, b_i or α_i have to solve a system of eight nonlinear equations (in addition to consistency conditions). These equations have a very large number of solutions and it might be the case that one of them could correspond to a

Order			
3	4	6	8
3-[72]	**3**-[34, 30, 88, 78] **4-5**-[54] **6**-[16]	**9**-[33] **10**-[16]	**27**-?
	P:3,4-[11] **P:2-7**-[7]	**P:5**-[11] **P:5-10**-[7]	**P:14**-?

Table 5: Symmetric composition schemes of the form (36) (appropriate when the ODE is split in two parts) and (35) (composition of a method and its adjoint). At order eight, we have not found methods. They would require at least 27 stages or at least 14 stages for processed schemes. The notation is the same as in Table 4.

method with very small error terms.

One optimization criterion frequently used when dealing with composition (36) is to work with the homogeneous subspace $\mathcal{L}_{r+1} = \langle F_{r+1,1}, \ldots, F_{r+1,n_{r+1}} \rangle$ (where by $F_{r+1,i}$ we denote the elements of the basis of the Lie algebra generated by $F^{[a]}$, $F^{[b]}$ at order $r+1$) and the leading error term, which can be expressed as $\sum_{i=1}^{n_{r+1}} c_i F_{r+1,i}$. In this setting, one selects the solution minimizing $E_{r+1} = \left(\sum_{i=1}^{n_{r+1}} |c_i|^2 \right)^{1/2}$. This optimization criterion allows one to compare the performance of methods with different number of stages by introducing the effective error, $\mathcal{E}_f = s E_{r+1}^{1/r}$, which normalizes with respect to the number of stages.

For the composition (35), it is not so clear how to assign a weight to each element of the associated Lie algebra since their contribution on the error can differ significantly. One accepted choice consists in minimizing the objective function $C = \sum_{i=1}^{2s} |\alpha_i|$.

Methods up to order six built by applying this procedure can be found in the literature. They show for most problems a better efficiency than compositions (62) at the same order when applied to the same class of problems. We collect some of the most relevant schemes in Table 5. As before, we also include processed methods.

We have not found methods of order eight. In fact, it is an open problem to determine if such a large system of polynomial equations admits solutions leading to more efficient methods than those collected in Table 4.

Runge–Kutta–Nyström methods. As we have seen in section 3.7, methods of this class may be considered as particular examples of composition (36). Nevertheless, their wide range of applicability to relevant physical problems has originated an exhaustive search of efficient schemes. Moreover, since in this case the associated vector fields $F^{[a]}$ and $F^{[b]}$ have different qualitative properties, methods with different features may be found in the literature. Thus, one may find non-symmetric methods of the form

$$\begin{array}{rcl} AB \equiv \psi_h & = & \varphi^{[a]}_{a_s h} \circ \varphi^{[b]}_{b_s h} \circ \cdots \circ \varphi^{[a]}_{a_1 h} \circ \varphi^{[b]}_{b_1 h} \\ BA \equiv \psi_h & = & \varphi^{[b]}_{b_s h} \circ \varphi^{[a]}_{a_s h} \circ \cdots \circ \varphi^{[b]}_{b_1 h} \circ \varphi^{[a]}_{a_1 h} \end{array} \qquad (83)$$

where AB and BA are conjugate to each other, leading to the same performance. However, to take profit of the FSAL (First Same As Last) property, we can consider the following non equivalent compositions

$$ABA \equiv \psi_h = \varphi^{[a]}_{a_{s+1} h} \circ \varphi^{[b]}_{b_s h} \circ \varphi^{[a]}_{a_s h} \circ \cdots \circ \varphi^{[b]}_{b_1 h} \circ \varphi^{[a]}_{a_1 h} \qquad (84)$$

and

$$BAB \equiv \psi_h = \varphi^{[b]}_{b_{s+1} h} \circ \varphi^{[a]}_{a_s h} \circ \varphi^{[b]}_{b_s h} \circ \cdots \circ \varphi^{[a]}_{a_1 h} \circ \varphi^{[b]}_{b_1 h}. \qquad (85)$$

The symmetric case ($a_{s+2-i} = a_i, b_{s+1-i} = b_i$ for the composition ABA and $b_{s+2-i} = b_i, a_{s+1-i} = a_i$ for the composition BAB) has also been proposed in this setting to get more efficient schemes. In this case, the minimum number of stages is, from Table 3, $s_r = L_r - 1$ (resp. $k_r = L^*_r - 1$), to get a method of order r (resp. effective order r). For non-symmetric compositions this minimum number can be obtained from Table 2.

Highly efficient methods up to order six have been published. In Table 6 we collect the most representative within this class. We add **S** or **N** to distinguish symmetric from non-symmetric schemes and the subindex AB, ABA and BAB to denote compositions (83), (84) and (85), respectively. Processed methods have also been included.

To achieve order eight, the coefficients a_i, b_i in a non-processed scheme have to solve (in addition to consistency) a system of 16 nonlinear equations. A large number of solutions could exist, although, as far as we know, only one attempt to solve these equations has been reported [64] (the performance of such method was not clearly superior symmetric-symmetric methods).

In [16] the authors have carried out a detailed analysis of the order conditions for symmetric compositions ABA and BAB. In this work 4th-order methods from 3 to 6 stages, and also 6th-order methods from 7 to 14 stages are analysed. The integrators selected perform extraordinarily well indeed. For instance, on the Hénon–Heiles Hamiltonian (19) the 4th-order 6-stage method is more accurate (at constant work) than leapfrog in a wide range of step sizes, whereas its global error is about 0.00175 times that of the classical 4th-order Runge–Kutta method. In consequence, its computational cost for a given error is about 0.31. This has to be compared with the composition (29) based on leapfrog, which have truncation errors about 10 times larger than the classical Runge–Kutta scheme.

On the other hand, as we have seen in subsection 4.1, the particular structure of problem (68) allows one to use modified potentials in compositions (83)-(85). This is appealing when the evaluation of such modified potentials is not particularly costly. In such circumstances one may replace in (83)-(85) flows associated to $hb_iF^{[b]}$ with the corresponding to $hC_h(b_i, c_i, d_i, e_{i1}, e_{i2})$, as given in

Order			
4	5	6	8
3S-[34, 30, 88, 78] **4N**$_{AB}$-[56] **4N**$_{BAB}$-[19] **4-5S**$_{ABA}$-[54] **5S**$_{BAB}$-[15] **6S**$_{ABA,BAB}$-[16]	**5N**$_{ABA}$-[65] **6N**$_{AB}$-[56] **6N**$_{AB}$-[29]	**7S**$_{ABA}$-[33, 65] **7S**$_{BAB}$-[33] **8-15S**$_{ABA,BAB}$-[16] **7,11S**$_{BAB}$-[16]	**17S**$_{ABA}$-[64]
P:2N$_{AB}$-[11]		**P:4-6S**$_{ABA,BAB}$-[13] **P:7S**$_{BAB}$-[14]	**P:9S**$_{ABA}$-[13] **P:11S**$_{BAB}$-[14]

Table 6: RKN splitting integrators. Since the role of the flows $\varphi_t^{[a]}$ and $\varphi_t^{[b]}$ is not interchangeable here, we distinguish symmetric **S** and non-symmetric **N** compositions with a subindex AB, ABA, BAB for the compositions (83), (84) and (85). As usual, processed methods are preceded by **P**.

Order			
3	4	6	8
2N$_{AB}$-[72]	**2S**$_{ABA,BAB}$-[47, 26] **4S**$_{ABA,BAB}$-[80] **3,4S**$_{ABA,BAB}$-[28, 67]	**4,5S**$_{ABA,BAB}$-[67]	**11S**$_{ABA,BAB}$-[67]
	P:1S$_{BAB}$-[82, 71, 87, 11] **P:2S**$_{BAB}$-[51]	**P:3S**$_{ABA,BAB}$-[11]	**P:4S**$_{ABA}$-[13] **P:5S**$_{BAB}$-[13, 14]

Table 7: RKN splitting methods with modified potentials. Schemes are coded as in Table 6.

(77). The coefficients c_i, d_i, etc. can be used to solve some order conditions, so that methods with a reduced number of stages can be obtained. We emphasize that these schemes are of interest when the extra cost due to the modified potentials is moderate, as is the case in many problems arising in classical and quantum mechanics. In Table 7 we collect some relevant methods we have found in the literature, both processed and non-processed.

Methods for near-integrable systems. As we have seen in section 3.6, splitting methods designed for equation (66) have typically two relevant parameters: h (the step size) and ε (the size of the perturbation). In consequence, the dominant error in a given scheme depends on their relative size, and this depends usually on the particular problem considered (and sometimes even on the initial conditions). For this reason, a number of methods at different orders in both parameters h and ε are found in the literature. We collect some of them in Table 8. Here the notation is a bit clumsy: a method of order (n,4), say, means that the exact and the modified vector fields, i.e., hF and $\log(\Psi(h))$ in (67), differ in terms $\mathcal{O}(\varepsilon h^{n+1} + \varepsilon^2 h^5 + \cdots)$, whereas in a method (7,6,4) this difference is $\mathcal{O}(\varepsilon h^8 + \varepsilon^2 h^7 + \varepsilon^3 h^5 + \cdots)$. In both cases, the order of consistency

Order		
(n,2)	(n,4)	(n,5)
1(2,2)**S**-[86] **n**(2n,2)$\mathbf{S}_{ABA,BAB}$-[53, 48]	**4**(6,4)$\mathbf{S}_{ABA,BAB}$-[53] **5**(8,4)$\mathbf{S}_{ABA,BAB}$-[53]	
P:1(32,2)-[87]	**P:3**(7,6,4)$\mathbf{S}_{ABA}$-[12]	
	P:2(6,4)$\mathbf{S}_{AB}$-[12]	**P:3**(7,6,5)$\mathbf{S}_{AB}$-[12]
	P:1(6,4)$\mathbf{S}_{ABA}$-[12] **P:n**(n,4)$\mathbf{S}_{ABA}$-[48]	**P:2**(7,6,5)$\mathbf{S}_{AB}$-[12]

Table 8: Splitting methods for near-integrable systems. For processed methods we also include methods applicable when $[[[F^{[b]}, F^{[a]}], F^{[b]}], F^{[b]}] = 0$ (second row) and schemes with modified flows (last two rows of processed methods).

in the limit $h \to 0$ is four, but the last method incorporates more terms in the asymptotic expansion of the error.

In [53] both families (ABA and BAB) of symmetric $(2s, 2)$ schemes for $s \leq 5$ with positive coefficients are proposed which are about three times more accurate (at constant work) than leapfrog, whereas in [48] a systematic study of $(2s, 2)$ methods is carried out, obtaining new schemes up to $s = 10$ with positive coefficients.

In some near-integrable problems, the identity $[[[F^{[b]}, F^{[a]}], F^{[b]}], F^{[b]}] = 0$ still holds, where $F^{[i]}$ is the vector field associated to $f^{[i]}$, $i = a, b$, in (66). This takes place, in particular, in Hamiltonian problems $H = H_0 + \varepsilon H_1$ where H_0 is quadratic in the kinetic energy and εH_1 depends only on the coordinates (e.g. examples (19)-(22) can be split in this way, where H_0 is the harmonic oscillator or the Kepler problem and H_1 depends only on the coordinates). In consequence, the previous techniques used to obtain RKN methods still apply here, as well as the inclusion of flows of modified potentials in the composition.

In Table 8 we separate, as usual, non-processed from processed schemes (preceded by **P**). In the later case we also include methods applicable when $[[[F^{[b]}, F^{[a]}], F^{[b]}], F^{[b]}] = 0$ (second row) and schemes with modified potentials (last two rows of processed methods).

6 Preserving properties and backward error analysis

Much insight into the long-time behavior of splitting methods (including preservation of invariants and structures in the phase space) can be gained by applying backward error analysis techniques. We will summarize here some of the main issues involved and refer the reader to [41] for a detailed treatment of the theory.

When we analyzed in the Introduction the symplectic Euler scheme as applied to the simple harmonic oscillator, we associated its good qualitative properties with the fact that the numerical solution can be interpreted as the exact solution of a perturbed Hamiltonian system. This remarkable feature constitutes a simple illustration of the insight provided by backward error analysis (BEA) in this setting. More generally, suppose that we apply the splitting method (36) to solve equation (11). Then the corresponding numerical solution at time $t = h$ is given by

$$x(h) = K(h)x_0 \equiv \mathrm{e}^{b_{s+1}hB}\,\mathrm{e}^{a_s hA}\,\mathrm{e}^{b_s hB}\cdots\mathrm{e}^{b_2 hB}\,\mathrm{e}^{a_1 hA}\,\mathrm{e}^{b_1 hB}x_0,$$

where the so-called stability matrix $K(h)$ is given explicitly by

$$K(h) = \begin{pmatrix} 1 & 0 \\ -b_{s+1}h & 1 \end{pmatrix}\begin{pmatrix} 1 & a_s h \\ 0 & 1 \end{pmatrix}\cdots\begin{pmatrix} 1 & a_1 h \\ 0 & 1 \end{pmatrix}\begin{pmatrix} 1 & 0 \\ -b_1 h & 1 \end{pmatrix}.$$

In this way, one gets

$$K(h) = \begin{pmatrix} K_1(h) & K_2(h) \\ K_3(h) & K_4(h) \end{pmatrix}$$

where $K_1(h)$, $K_4(h)$ (respectively, $K_2(h)$, $K_3(h)$) are even (resp. odd) functions and $\det K(h) = 1$. As a matter of fact, any splitting method is uniquely determined by its stability matrix, so that the analysis can be carried out only with $K(h)$ [9]. If in addition the splitting method is symmetric then $K(h)^{-1} = K(-h)$ and we can write

$$K(h) = \begin{pmatrix} p(h) & K_2(h) \\ K_3(h) & p(h) \end{pmatrix}$$

where $p(h) = \frac{1}{2}\mathrm{tr}(K(h)) = \frac{1}{2}(K_1(h) + K_4(h))$. It can be shown that the matrix $K(h)$ is stable for a given $h \in \mathbb{R}$, i.e., $K(h)^n$ is bounded for all the iterations n, if and only if there exist real functions $\phi(h), \gamma(h)$ such that $p(h) = \cos(\phi(h))$ and $K_2(h) = -\gamma(h)^2 K_3(h)$. In that case

$$K(h) = \begin{pmatrix} \cos(\phi(h)) & \gamma(h)\sin(\phi(h)) \\ -\frac{\sin(\phi(h))}{\gamma(h)} & \cos(\phi(h)) \end{pmatrix} = \exp\begin{pmatrix} 0 & \gamma(h)\phi(h) \\ -\frac{\phi(h)}{\gamma(h)} & 0 \end{pmatrix}$$

where, by consistency, $\phi'(0) = 1$ and $\gamma(0) = 1$, whereas symmetry imposes $\phi(-h) = -\phi(h)$ and $\gamma(-h) = \gamma(h)$.

This result implies, in particular, that the numerical solution (q_n, p_n) at time $t_n = nh$ obtained by applying the splitting method to the linear system (11) verifies

$$\begin{pmatrix} q_n \\ p_n \end{pmatrix} = \begin{pmatrix} \cos(t_n\tilde{\omega}) & \gamma(h)\sin(t_n\tilde{\omega}) \\ -\gamma(h)^{-1}\sin(t_n\tilde{\omega}) & \cos(t_n\tilde{\omega}) \end{pmatrix}\begin{pmatrix} q_0 \\ p_0 \end{pmatrix}$$

for values of h such that $K(h)$ is stable. Here $\tilde{\omega} = \phi(h)/h$. Equivalently,

$$q_n = \tilde{q}(t_n), \qquad p_n = (\phi(h)\gamma(h)/h)^{-1}\frac{d}{dt}\tilde{q}(t_n),$$

where $\tilde{q}(t)$ is the exact solution of

$$\frac{d^2}{dt^2}\tilde{q} + \tilde{\omega}^2\tilde{q} = 0$$

with initial condition $\tilde{q}(0) = q_0$, $\tilde{q}'(0) = (\phi(h)\gamma(h)/h)p_0$. In other words, the numerical solution provided by the splitting method is the exact solution of a harmonic oscillator with frequency $\tilde{\omega} \approx 1$, i.e., of a system of equations satisfying the same geometric properties as the original system. The existence of such a backward error interpretation has direct implications for the qualitative behavior of the numerical solution, as well as for its global error.

The main idea can be extended to an arbitrary non-linear ODE (1). Recall from Subsection 2.2 that each integrator ψ_h has associated a series $\Psi(h) = I + h\Psi_1 + h^2\Psi_2 + \cdots$ of differential operators acting on smooth functions $g \in C^\infty(\mathbb{R}^D, \mathbb{R})$, and its formal logarithm $\log(\Psi(h))$ is a series of vector fields (viewed as first order differential operators) $\log(\Psi(h)) = hF_1 + h^2F_2 + \cdots$. For $g \in C^\infty(\mathbb{R}^D, \mathbb{R})$, the result of acting each F_k on g is of the form $F_k[g] = g'(x)f_k(x)$, for a certain smooth map $f_k : \mathbb{R}^D \longrightarrow \mathbb{R}^D$. Now, consider the *modified differential equation* (defined as a formal series in powers of h)

$$\tilde{x}' = f_h(\tilde{x}) \equiv f(\tilde{x}) + hf_2(\tilde{x}) + h^2f_3(\tilde{x}) + \cdots \tag{86}$$

associated to the integrator ψ_h. Then one has that $x_n = \tilde{x}(t_n)$, with $t_n = nh$, which allows studying the long-time behaviour of the numerical integrator by analysing the solutions of the system (86) viewed as a small perturbation of the original system (1). This allows one to get important qualitative information about the numerical solution. In particular,

- for symmetric methods, the modified differential equation only contains even powers of h;
- for volume-preserving methods applied to a divergence-free dynamical system, the modified equation is also divergence-free;
- for symplectic methods applied to a Hamiltonian system, the modified differential equation is (locally) Hamiltonian.

In the particular case of a symplectic integration method, this means that there exist smooth functions $H_j : \mathbb{R}^{2d} \longrightarrow \mathbb{R}$ for $j = 2, 3, \ldots$, such that $f_j(x) = J\nabla H_j(x)$, where J is the canonical symplectic matrix. In consequence, there exists a modified Hamiltonian of the form

$$\tilde{H}(q,p) = H(q,p) + hH_2(q,p) + h^2H_3(q,p) + h^3H_4(q,p) + \cdots \tag{87}$$

such that the modified differential equation is given by

$$q' = \nabla_p\tilde{H}(q,p), \qquad p' = -\nabla_q\tilde{H}(q,p).$$

Of course, if the method has order r, say, then $H_i = 0$ for $i \leq r$ in (87). In other words, the modified Hamiltonian has the form $\tilde{H} = H + h^rH_{r+1} + \cdots$. In particular, for the Störmer-Verlet method (10) applied to the Hamiltonian $H(q,p) = T(p) + V(q)$, one has

$$\tilde{H} = H + h^2\left(-\frac{1}{24}V_{qq}(T_p, T_p) + \frac{1}{12}T_{pp}(V_q, V_q)\right) + \cdots$$

Apart from the linear case analyzed before, the series in (86) does not converge in general. To make this formalism rigorous, one has to give bounds on the coefficient functions $f_j(x)$ of the modified equation so as to determine an optimal truncation index and finally one has to estimate the difference between the numerical solution x_n and the exact solution $\tilde{x}(h)$ of the modified equation.

These estimates constitute in fact the basis for rigorous statements about the long term behavior of the numerical solution. For instance, this theory allows one to proof rigorously that a symplectic numerical method of order r with constant step size h applied to a Hamiltonian system H verifies that $H(x_n) = H(x_0) + \mathcal{O}(h^r)$ for exponentially long time intervals [41].

On the other hand, since the modified differential equation of a numerical scheme depends explicitly on the step size used, one has a different modified equation each time the step size h is changed. This fact seems to be the reason of the poor long time behavior observed in practice when a symplectic scheme is implemented directly with a standard variable step-size strategy.

7 Special methods for special problems

7.1 Splitting methods for linear systems

Suppose one is interested in solving numerically the differential equations arising from the generalized harmonic oscillator with Hamiltonian function (18). Although RKN methods with modified potentials can be always used for this purpose, we will see in the sequel that the particular structure of this system allows one to design specially tailored schemes which are orders of magnitude more efficient than other integrators frequently used in the literature.

At this point, the reader could reasonably ask about the convenience of designing new numerical methods for the harmonic oscillator (18). It turns out, however, that efficient splitting methods for this system can be of great interest for the numerical treatment of partial differential equations appearing in quantum mechanics, optics and electrodynamics previously discretized in space.

Suppose, in particular, that we have to solve numerically the time dependent Schrödinger equation (26) with initial wave function $\psi(x,0) = \psi_0(x)$. We can write (26) as

$$i\frac{\partial}{\partial t}\psi = (T(P) + V(X))\,\psi, \tag{88}$$

where $T(P) = \dfrac{1}{2m}P^2$, and the operators X, P are defined by their actions on $\psi(x,t)$ as

$$X\psi(x,t) = x\psi(x,t), \qquad P\psi(x,t) = -i\nabla\psi(x,t).$$

For simplicity, let us consider the one-dimensional problem and suppose that it is defined in a given interval $x \in [x_0, x_N]$ ($\psi(x_0,t) = \psi(x_N,t) = 0$ or it has periodic boundary conditions). A common procedure consists in taking first a discrete spatial representation of the wave function $\psi(x,t)$: the interval is split in N parts of length $\Delta x = (x_N - x_0)/N$ and the vector $\mathbf{u} = (u_0, \ldots, u_{N-1})^T \in \mathbb{C}^N$

is formed, with $u_n = \psi(x_n, t)$ and $x_n = x_0 + n\Delta x$, $n = 0, 1, \dots, N-1$. The partial differential equation (88) is then replaced by the N-dimensional linear ODE

$$i\frac{d}{dt}\mathbf{u}(t) = \mathbf{H}\,\mathbf{u}(t), \qquad \mathbf{u}(0) = \mathbf{u}_0 \in \mathbb{C}^N, \tag{89}$$

where $\mathbf{H} \in \mathbb{R}^{N\times N}$ represents the (in general Hermitian) matrix associated with the Hamiltonian [32]. The formal solution of equation (89) is given by $\mathbf{u}(t) = \mathrm{e}^{-it\mathbf{H}}\mathbf{u}_0$, but to exponentiate this $N \times N$ complex and full matrix can be prohibitively expensive for large values of N, so in practice other methods are preferred.

In general $\mathbf{H} = \mathbf{T} + \mathbf{V}$, where $\mathbf{V}$ is a diagonal matrix associated with the potential energy V and $\mathbf{T}$ is a full matrix related to the kinetic energy T. Their action on the wave function vector is obtained as follows. The potential operator being local in this representation, one has $(\mathbf{Vu})_n = V(x_n)u_n$ and thus the product $\mathbf{Vu}$ requires to compute N complex multiplications. Since periodic boundary conditions are assumed, for the kinetic energy one has $\mathbf{T}\,\mathbf{u} = \mathcal{F}^{-1}\mathbf{D}_T\mathcal{F}\mathbf{u}$, where $\mathcal{F}$ and $\mathcal{F}^{-1}$ correspond to the forward and backward discrete Fourier transform, and $\mathbf{D}_T$ is local in the momentum representation (i.e., it is a diagonal matrix). The transformation $\mathcal{F}$ from the discrete coordinate representation to the discrete momentum representation (and back) is done via the fast Fourier transform (FFT) algorithm, requiring $\mathcal{O}(N \log N)$ operations. It is therefore possible to use the methods of subsection 2.1 with this splitting.

There are other ways, however, of using splitting techniques in this context. To this end, notice that $\mathrm{e}^{-it\mathbf{H}}$ is not only unitary, but also symplectic with canonical coordinates $\mathbf{q} = \mathrm{Re}(\mathbf{u})$ and momenta $\mathbf{p} = \mathrm{Im}(\mathbf{u})$. Thus, equation (89) is equivalent to [37, 38]

$$\frac{d}{dt}\mathbf{q} = \mathbf{H}\,\mathbf{p}, \qquad\qquad \frac{d}{dt}\mathbf{p} = -\mathbf{H}\,\mathbf{q}, \tag{90}$$

where $\mathbf{H}\,\mathbf{q}$ and $\mathbf{H}\,\mathbf{p}$ require both a real-complex FFT and its inverse. In addition, system (90) can be seen as the classical evolution equations corresponding to the Hamiltonian function (18) with $M = N = \mathbf{H}$. Thus, efficient schemes for solving numerically the generalized harmonic oscillator can be applied directly to this problem. Also the Maxwell equations (28) in an isotropic, lossless and source free medium, when they are previously discretized in space have a similar structure [70]. In consequence, numerical methods of this class are well adapted for their numerical treatment.

Clearly, one may write

$$\frac{d}{dt}\left\{\begin{array}{c}\mathbf{q}\\ \mathbf{p}\end{array}\right\} = \left(\begin{array}{cc}\mathbf{0} & \mathbf{H}\\ -\mathbf{H} & \mathbf{0}\end{array}\right)\left\{\begin{array}{c}\mathbf{q}\\ \mathbf{p}\end{array}\right\} = (\mathbf{A}+\mathbf{B})\left\{\begin{array}{c}\mathbf{q}\\ \mathbf{p}\end{array}\right\}, \tag{91}$$

with the $2N \times 2N$ matrices $\mathbf{A}$ and $\mathbf{B}$ given by

$$\mathbf{A} \equiv \left(\begin{array}{cc}\mathbf{0} & \mathbf{H}\\ \mathbf{0} & \mathbf{0}\end{array}\right), \qquad\qquad \mathbf{B} \equiv \left(\begin{array}{cc}\mathbf{0} & \mathbf{0}\\ -\mathbf{H} & \mathbf{0}\end{array}\right).$$

The evolution operator corresponding to (91) is

$$\mathbf{O}(t) = \begin{pmatrix} \cos(t\mathbf{H}) & \sin(t\mathbf{H}) \\ -\sin(t\mathbf{H}) & \cos(t\mathbf{H}) \end{pmatrix}, \tag{92}$$

which is an orthogonal and symplectic $2N \times 2N$ matrix. As before, its evaluation is computationally very expensive and thus some approximation is required. The usual procedure is to split the whole time interval into M steps of length $h = t/M$, so that $\mathbf{O}(t) = [\mathbf{O}(h)]^M$, and then approximate $\mathbf{O}(h)$ acting on the initial condition at each step.

In this respect, observe that

$$\mathrm{e}^{\mathbf{A}} = \begin{pmatrix} \mathbf{I} & \mathbf{H} \\ \mathbf{0} & \mathbf{I} \end{pmatrix}, \qquad \mathrm{e}^{\mathbf{B}} = \begin{pmatrix} \mathbf{I} & \mathbf{0} \\ -\mathbf{H} & \mathbf{I} \end{pmatrix}$$

and the cost of evaluating the action of $\mathrm{e}^{\mathbf{A}}$ and $\mathrm{e}^{\mathbf{B}}$ on $\mathbf{z} = (\mathbf{q}, \mathbf{p})^T$ is essentially the cost of computing the products $\mathbf{H}\,\mathbf{p}$ and $\mathbf{H}\,\mathbf{q}$, respectively. It makes sense, then, to use splitting methods of the form (36), which in this context read

$$\mathbf{O}_n(h) = \mathrm{e}^{hb_{s+1}\mathbf{B}}\,\mathrm{e}^{ha_s\mathbf{A}}\,\cdots\,\mathrm{e}^{hb_2\mathbf{B}}\,\mathrm{e}^{ha_1\mathbf{A}}\,\mathrm{e}^{hb_1\mathbf{B}}. \tag{93}$$

Several methods with different orders have been constructed along these lines indeed [37, 50, 89]. Of particular relevance are the schemes presented in [37], since only $s = r$ exponentials $\mathrm{e}^{ha_i\mathbf{A}}$ and $\mathrm{e}^{hb_i\mathbf{B}}$ are used to achieve order r for $r = 4, 6, 8, 10$ and 12. By contrast, in a general composition (93) the minimum number s of exponentials $\mathrm{e}^{ha_i\mathbf{A}}$ and $\mathrm{e}^{hb_i\mathbf{B}}$ (or stages) required to attain order $r = 8, 10$ is $k = 15, 31$, respectively [41, 45].

Furthermore, one can use processing to reduce even more the number of exponentials. A different approach can also be followed, however: to take a number of stages larger than strictly necessary to solve all the order conditions to improve the efficiency and stability of the resulting schemes. The idea is to use the extra cost to reduce the size of the error terms, enlarge the stability interval and achieve therefore a higher efficiency but without raising the order.

Kernels with up to 19, 32 and 38 stages have been proposed, and for each kernel the corresponding coefficients a_i, b_i have been determined according to two different criteria. The first set of solutions is taken so as to provide methods of order $r = 10$, 16 and 20. The second set of coefficients bring highly accurate *second* order methods with an enlarged domain of stability. A more detailed treatment can be found in [8, 9].

7.2 Splitting methods for non-autonomous systems

So far we have considered the problem of designing splitting methods for the numerical integration of autonomous differential equations (1). As we have shown, there are a large number of schemes of different orders in the literature, and some of them are particularly efficient when the system possesses some additional structure, e.g., for the second-order differential equation $y'' = g(y)$ and the generalized harmonic oscillator (18). In this section we will review two different strategies to apply the splitting schemes when there is an explicit time dependency in the original problem.

To fix ideas, let us assume that our system is non-autonomous and can be split as

$$x' = f(x,t) = f^{[a]}(x,t) + f^{[b]}(x,t), \qquad x(0) = x_0. \tag{94}$$

The first, most obvious procedure consists in taking t as a new coordinate, so that (94) is transformed into an equivalent autonomous equation to which standard splitting algorithms can be applied. More specifically, equation (94) is equivalent to the enlarged system

$$\frac{d}{dt}\left\{\begin{array}{c} x \\ x_{t1} \\ x_{t2} \end{array}\right\} = \underbrace{\left\{\begin{array}{c} f^{[a]}(x,x_{t1}) \\ 0 \\ 1 \end{array}\right\}}_{\hat{f}^{[1]}} + \underbrace{\left\{\begin{array}{c} f^{[b]}(x,x_{t2}) \\ 1 \\ 0 \end{array}\right\}}_{\hat{f}^{[2]}} \tag{95}$$

with $x_{t1}, x_{t2} \in \mathbb{R}$. Note that if the systems

$$y' = \hat{f}^{[A]}(y), \qquad y' = \hat{f}^{[B]}(y)$$

with $y = (x, x_{t1}, x_{t2})$ are solvable, then a splitting method similar to (36) can be used, since x_{t1} is constant when integrating the first equation and x_{t2} is constant when solving the second one. This, in fact, can be considered as a generalization of the procedure proposed in [74] for time-dependent and separable Hamiltonian systems, and is of interest if the time-dependent part in $f^{[a]}$ and $f^{[b]}$ is cheap to compute. Otherwise the overall algorithm may be computationally costly, since these functions have to be evaluated s times (the number of stages in (36)) per time step.

Another disadvantage of this simple procedure is the following. Suppose that, when the time is frozen, the function f in (94) has a special structure which allows to apply highly efficient splitting schemes. If now t is a variable, with (95) this time dependency is eliminated but the structure of the equation might be modified so that one is bound to resort to more general and less efficient integrators. This issue has been analyzed in detail in [5].

A second procedure which avoids the difficulties exhibited by the previous example consists in approximating the exact solution of (94) or equivalently the flow φ_h by the composition

$$\psi_{s,h}^{[r]} = \varphi_h^{[\hat{B}_{s+1}]} \circ \varphi_h^{[\hat{A}_s]} \circ \varphi_h^{[\hat{B}_s]} \circ \cdots \circ \varphi_h^{[\hat{B}_2]} \circ \varphi_h^{[\hat{A}_1]} \circ \varphi_h^{[\hat{B}_1]}, \tag{96}$$

where the maps $\varphi_h^{[\hat{A}_i]}$, $\varphi_h^{[\hat{B}_i]}$ are the exact 1-flows corresponding to the time-independent differential equations

$$x' = \hat{A}_i(x), \qquad x' = \hat{B}_i(x), \qquad i = 1, 2, \ldots \tag{97}$$

respectively, with

$$\hat{A}_i(x) \equiv h \sum_{j=1}^{k} \rho_{ij} f^{[a]}(x, \tau_j), \qquad \hat{B}_i(x) \equiv h \sum_{j=1}^{k} \sigma_{ij} f^{[b]}(x, \tau_j). \tag{98}$$

Here $\tau_j = t_0 + c_j h$ and the (real) constants c_j, ρ_{ij}, σ_{ij} are chosen such that $\varphi_h = \psi_{s,h}^{[r]} + \mathcal{O}(h^{r+1})$. Furthermore, the new schemes, when applied to (94) with the time frozen, reproduce the standard splitting (36). This is accomplished by ensuring that $\sum_j \rho_{ij} = a_i$ and $\sum_j \sigma_{ij} = b_i$. The c_j coefficients, on the other hand, are typically chosen as the nodes of a symmetric quadrature rule of order at least r. In particular, if a Gauss–Legendre quadrature rule is adopted, with k evaluations of $f^{[a]}(x, \tau_j)$ and $f^{[b]}(x, \tau_j)$ a method of order $r = 2k$ can be built (taking s sufficiently large).

Once the quadrature nodes τ_j and the number of stages s are fixed, there still remains to obtain the coefficients ρ_{ij}, σ_{ij} such that $\psi_{s,h}^{[r]}$ has the desired order. This is done by requiring that the composition (96) match the solution of (94) as given by the Magnus expansion [10]. The task is made easier by noticing that the order conditions to be satisfied by ρ_{ij} and σ_{ij} are identical both for linear and nonlinear vector fields. Thus, the problem for the linear case is solved first and then one generalizes the treatment to arbitrary nonlinear separable problems.

The integrators of order four and six constructed along these lines in [5] are generally more efficient than standard splitting methods applied to the enlarged system (95).

8 Numerical examples with selected methods

This section intends to illustrate the relative performance between different splitting methods, and occasionally we compare with other standard methods. We consider first a relatively simple problem where most of the methods previously mentioned can be used, showing their good features. We show the interest of the high order methods when accurate results are desired and the improvement which can be achieved when choosing a method from the most appropriate family of methods for each problem. Next, we consider a problem which, due to its very particular structure, allows to build tailored methods whose performance is much superior to other splitting methods.

8.1 The perturbed Kepler problem

As a first example, we take the perturbed Kepler problem with Hamiltonian (20)

$$H = \frac{1}{2}(p_1^2 + p_2^2) - \frac{1}{r} - \frac{\varepsilon}{2r^3}\left(1 - \alpha\frac{3q_1^2}{r^2}\right), \tag{99}$$

where $r = \sqrt{q_1^2 + q_2^2}$ and the additional parameter α has been introduced for convenience. This Hamiltonian describes in first approximation the dynamics of a satellite moving into the gravitational field produced by a slightly oblate spheric planet. The motion takes place in a plane containing the symmetry axis of the planet when $\alpha = 1$, whereas $\alpha = 0$ corresponds to a plane perpendicular to that axis [60].

This simple (but non trivial) example constitutes in fact an excellent test bench for most of the methods of this paper. Notice that the system is separable into kinetic and potential parts, and we can use, for instance, the symmetric second order method (10) which allows us to get higher order methods by composition, as given in (62). On the other hand, since the system is separable into two solvable parts, then we can also use methods from Table 5, which should show better performances than methods of the same order considered from the previous family of methods. In addition, the kinetic energy is quadratic in momenta, so that RKN methods from Table 6 can be used, and one expect a further improvement. Finally, observe that one may split the system as

$$H = H_0 + \varepsilon H_I, \tag{100}$$

where H_0 corresponds to the Kepler problem, which is exactly solvable. The Keplerian part of the Hamiltonian can be solved in action-angle coordinates, where two changes of variables are needed. Alternatively, if desired, H_0 can be integrated in cartesian coordinates using the f and g Gauss functions, but then a nonlinear equation must be solved with an iterative scheme [31]. In any case, if $\varepsilon \ll 1$, methods from Table 8 can be used which should be superior to all previous methods in the limit $\varepsilon \to 0$.

We must also mention that the performance of all methods previously mentioned can be further improved by using the processing technique, and even additional improvements can be achieved if modified potentials are considered.

We take $\varepsilon = 0.001$, which approximately corresponds to a satellite moving under the influence of the Earth [46] and initial conditions $q_1 = 1 - e$, $q_2 = 0$, $p_1 = 0$, $p_1 = \sqrt{(1+e)/(1-e)}$, with $e = 0.2$ (this would be the eccentricity for the unperturbed Kepler problem). In general, no closed orbits are present and a precession is observed. Notice that for the Hamiltonian (99) the strength of the perturbation depends obviously of the value of ε, but also on the initial conditions. We take $\alpha = 1$ and determine numerically the trajectory for up to the final time $t_f = 500 \cdot 2\pi$ (the exact solution is accurately approximated using a high order method with a very small time-step, and this computation was repeated with different time steps and methods to assure the accuracy is reached up to round off).

To compare the performance of different methods it is usual to consider efficiency curves. We measure the average error in energy computed at times $t_k = k \cdot 2\pi$ for $k = 401, 402, \ldots, 500$ and this is repeated several times for each method and using different time steps (changing the computational cost for the numerical integration).

In the first numerical test, we compare the relative performance between different symmetric-symmetric methods collected in Table 4. We choose as the basic method the symmetric second order composition (10) to build higher order methods with the composition (62). As mentioned, in general, the performance of the methods of the same order increase with the number of stages for the methods in Table 4. This is illustrated in Figure. 4 where we show the performance of two 4th-order methods with three stages (given by

(29)) and five stages (given by (82)). The results show that for this problem the five stage method is more accurate for all computational costs considered. A similar feature is observed for the other methods at higher orders (with the exception of the 24-stage 8th-order method which was obtained in a different way and the 21-stage methods shows a better performance). We choose the best method from Table 4 at each order (including the well known three-stages 4th-order method as a reference) where we denote by $\mathrm{SS}_s r$ the corresponding method of order r using a s-stage composition:

- $\mathrm{SS}_1 2$: The 2nd-order method (10) which has the highest possible stability among splitting methods.
- $\mathrm{SS}_3 4$ The well known 3-stage 4th-order method (29).
- $\mathrm{SS}_5 4$ The 5-stage 4th-order method (82) [78].
- $\mathrm{SS}_{13} 6$, $\mathrm{SS}_{21} 8$, $\mathrm{SS}_{35} 10$: The composition from Table 4 and whose coefficients are given in [76].

The results are shown in Fig. 4, where we clearly observe that the high order methods have better performance when high accuracy is desired.

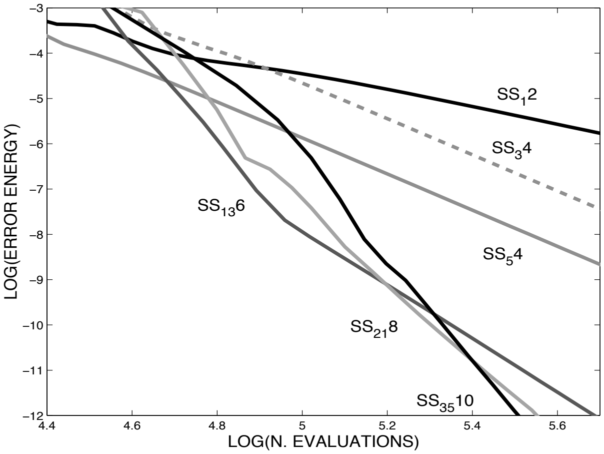


Figure 4: Average error in energy versus number of force evaluations in a double logarithmic scale for the numerical integration of the Hamiltonian system (99). It is shown performance of the most efficient non-processed symmetric-symmetric methods from Table 4.

The following numerical experiment intends to illustrate the interest of the methods designed for problems with some particular structure. For simplicity, in this numerical test we only consider fourth-order methods from different families of methods which can be used on this problem, in order to observe the benefit of tuned methods for problems with particular structures. The following methods are considered in addition to $\mathrm{SS}_3 4$ and $\mathrm{SS}_5 4$:

- $\mathrm{S}_6 4$: The symmetric 6-stage 4th-order method for separable problems [16] from Table 5.
- $\mathrm{RKN}_6 4$: The symmetric $\mathbf{6S}_{BAB}$ 4th-order method for Nyström problems [16] from Table 6.
- NI(8,4): The 5-stage fourth-order method $\mathbf{5}(8,4)\mathbf{S}_{BAB}$ given in [53] from Table 8.
- $\mathrm{RK}_4 4$: The standard 4-stage 4th-order non-symplectic Runge-Kutta methods, used as a reference method.

Figure 5 shows in double logarithmic scale the results obtained. In the left panel we show the average error in energy versus the number of force evaluations and in the right panel we repeated the same experiment, but we measured the average error in position (computed at the same instants). For the method NI(8,4) this counting of the computational cost is not an appropriate measure. Its computational cost strongly depends on each particular problem since the evolution of H_0 has to be computed exactly (or very accurately). For simplicity, in our experiments, we have considered that one stage of NI(8,4) is twice as expensive as one evaluation of the force. We have also included as a reference the curve obtained in Fig. 4 by $\mathrm{SS}_{35}10$.

Observe that in the first case the results will be largely independent of t_f because the average error in energy does not increase secularly for symplectic integrators. For comparison, we have also included the results obtained by the standard 4-stage fourth-order Runge-Kutta method whose error in energy grows linearly and the error in positions quadratically.

Even more accurate results could be obtained as follows. As mentioned, for this particular problem, modified potentials could be used and this can be done at a very low computational cost. Then, methods from Tables 7 and 8 can be used. For instance, for the split (100) we can apply methods which incorporate modified perturbations $\exp(\varepsilon C_h(b,c))$ into the algorithm. Then the following map has to be evaluated:

$$\begin{aligned} \mathrm{e}^{\varepsilon C_h(b,c)} p_1 &= p_1 + h\varepsilon \left(b\frac{A}{r^7} - h^2 \varepsilon c \frac{C}{(r^7)^2} \right) q_1 \\ \mathrm{e}^{\varepsilon C_h(b,c)} p_2 &= p_2 + h\varepsilon \left(b\frac{B}{r^7} - h^2 \varepsilon c \frac{D}{(r^7)^2} \right) q_2, \end{aligned} \tag{101}$$

where $A = (3/2)(\alpha(3q_1^2 - 2q_2^2) - r^2)$, $B = (3/2)(\alpha 5q_1^2 - r^2)$, $C = 9(2r^4 + 3\alpha r^2(q_2^2 - 4q_1^2) + \alpha^2(18q_1^4 + q_1^2 q_2^2 - 2q_2^4))$ and $D = 9(2r^4 - 15\alpha r^2 q_1^2 + 5\alpha^2 q_1^2(5q_1^2 + 2q_2^2))$. Notice that the increment in the computational cost with respect to the evaluation of $\mathrm{e}^{h\varepsilon bF^{[b]}}$ (which corresponds to $c = 0$) is only due to a few very simple additional operations. For this particular example, the evaluation of the modified perturbation $\mathrm{e}^{\varepsilon C_h(b,c)}$ is about a $10 - 20\%$ more expensive than $\mathrm{e}^{hbF^{[b]}}$.

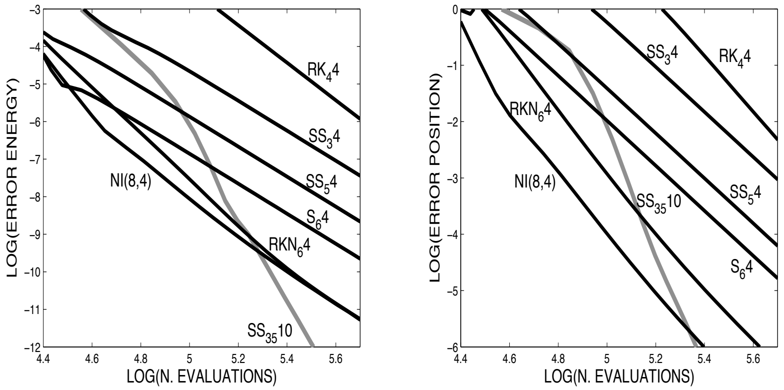


Figure 5: Average error in energy (left panel) and position (right panel) versus number of force evaluations in a double logarithmic scale for the numerical integration of the Hamiltonian system (99). The performance of different 4th-order methods from Tables 4-8 is shown. As a reference, we also shown the results obtained by the standard 4-stage fourth-order Runge-Kutta method.

As a result, more elaborated and efficient methods are obtained by considering the schemes $(n, 4)$ and $(n, 5)$ from Table 8 with processing (which only requires a few more code lines to program) for RKN problems and which incorporate modified potentials (see [12]).

8.2 The Schrödinger equation

As a second example we consider the one-dimensional time-dependent Schrödinger equation (26) with the Morse potential $V(x) = D\,(1 - e^{-\alpha x})^2$. We fix the parameters to the following values in atomic units (a.u.): $\mu = 1745$ a.u., $D = 0.2251$ a.u. and $\alpha = 1.1741$ a.u., which are frequently used for modelling the HF molecule. As initial conditions we take the Gaussian wave function $\psi(x,t) = \rho \exp\big(-\beta(x-\bar{x})^2\big)$, with $\beta = \sqrt{k\mu}/2$, $k = 2D\alpha^2$, $\bar{x} = -0.1$ and ρ is a normalizing constant. Assuming that the system is defined in the interval $x \in [-0.8, 4.32]$, we split it into $d = 128$ parts of length $\Delta x = 0.04$, take periodic boundary conditions and integrate along the interval $t \in [0, 20 \cdot 2\pi/w_0]$ with $w_0 = \alpha\sqrt{2D/\mu}$ (see [8] for more details on the implementation of the splitting methods to this particular problem).

Figure 6 shows the error in the Euclidean norm of the vector solution at the end of the integration versus the number of FFT calls in double logarithmic scale. The integrations are done starting from a sufficiently small time step and repeating the computation by slightly increasing the time step until an overflow occurs, which we identify with the stability limit. We present the results for the following methods (in addition to the previous ones SS$_1$2, SS$_3$4 and SS$_{21}$8):

- $\mathrm{RKN}_{11}6$: The 11-stage 6th-order method $\mathbf{11S}_{BAB}$-[16] from Table 6.
- $\mathrm{GM}_{12}12$: The 12-stage 12th-order method from [37] tailored for linear problems with this particular structure.
- $\mathrm{P}_{38}2$: The 38-stage second order processed method with coefficients given in [8] tailored for linear problems with this particular structure (only the computational cost required to evaluate the kernel has been taken into consideration).
- $\mathrm{T}_r r,\ r = 8, 12$: r-stage rth-order Taylor methods obtained by truncating the exponential up to order r.

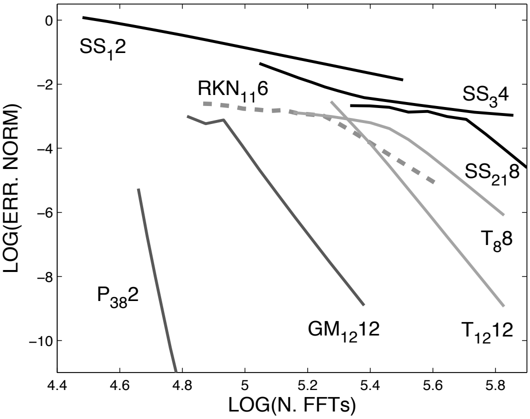


Figure 6: Error of the vector solution for the Schrödinger equation versus the number of FFT calls in a log-log scale for the symmetric-symmetric composition methods: $\mathrm{SS}_k\mathrm{n}$, for methods of order n using k-stage compositions; $\mathrm{RKN}_{11}6$ is an 11-stage 6th-order methods from [16] for Nyström problems; the 12-stage 12th-order method, $\mathrm{GM}_{12}12$, from [37]; and $\mathrm{P}_{38}2$, a 38-stage second order processed method.

From the figure we observe that, for this numerical experiment, standard Taylor methods outperform to general symmetric-symmetric splitting methods and are also more accurate than the RKN method. This is because this problem has a very particular structure and these splitting methods are not optimized for them. However, the schemes $\mathrm{GM}_{12}12$ and $\mathrm{P}_{38}2$ are built for linear problems with this structure and their superiority is clearly manifest. It is important to remember that Taylor methods are non-geometric integrators. The numerical experiments are carried out for a relatively short time, and the relative performances of the Taylor methods deteriorates with respect to splitting methods for longer integrations.

9 Conclusions and outlook

Splitting methods are a flexible and powerful way to solve numerically the initial value problem defined by (1) when f can be decomposed into two or more parts and each of them is simpler to integrate than the original problem. This is especially true when the exact flow possesses some structural features which seems natural to reproduce at the discrete level, as happens, for instance, in Hamiltonian, Poisson, volume-preserving or time-reversible dynamical systems. They are explicit, usually simple to apply in practice and constitute an important class of geometric numerical integrators. Closely connected with splitting schemes are composition methods. In this case, the idea is to construct numerical integrators of arbitrarily high order by composing one or more basic schemes of low order with appropriately chosen coefficients. The resulting method inherits the relevant properties that the basic integrator shares with the exact solution, provided these properties are preserved by composition.

In this paper we have reviewed some of the main features of splitting and composition methods in the numerical integration of ordinary differential equations. We have presented a novel approach to get the order conditions of this class of schemes based on Lyndon words and we have seen how these order conditions particularize when coping with special classes of dynamical systems (near-integrable systems and second-order differential equations of the form $y'' = g(y)$). It turns out that the number of equations to be solved increases dramatically with the order considered, as so does the complexity of the problem of finding efficient high order methods. One way to circumvent (up to a certain point) this difficulty consists in applying the processing technique, since then it is possible to design algorithms with fewer evaluations per time step. In this sense, one could say that the use of processing is perhaps the most economical path to achieve high order.

Since splitting methods are widely applied in many areas of science, it is not surprising that a great number of different schemes are available in the mathematical, physical and chemical literature. We have collected here some of the most representative integrators, classified according to the particular structure of the differential equations, the number of stages and the order of consistency, citing in each case the actual reference where the method has been first proposed.

The good qualitative behavior exhibited by splitting methods (including preservation of invariants and structures in phase space), as well as their favorable error propagation in long-time integrations can be accounted for by applying the theory of backward error analysis. Loosely speaking, the observed performance is related with the fact that the numerical solution provided by the splitting method is the exact solution of a differential equation with the same geometric properties as the original system. This interpretation constitutes in addition the basis for rigorous estimates on the numerical solution.

In contrast with standard integration methods (Runge–Kutta, multistep), whose efficiency is essentially independent of the particular differential equation considered, splitting schemes can be designed to incorporate in their formulation some of the most relevant properties of the original system. This feature has to be taken into account when comparing the efficiency of splitting methods with respect to other general purpose integrators. In this sense, Figures 4 and 5 are quite illustrative. For this particular problem, specially adapted 4th-order splitting schemes are up to 6 orders of magnitude more accurate with the same computational cost than the well known Runge–Kutta method. They even outperform other standard higher order composition integrators for a wide range of values of the step size h.

As an additional evidence of the extraordinary flexibility of splitting methods, we have considered the problem of designing specially tailored schemes for the numerical integration of the generalized harmonic oscillator (18). It turns out that several partial differential equations appearing in quantum mechanics, optics and electrodynamics give rise, once discretized in space, to this system with different matrices M and N. The particular structure of this dynamical system can be exploited to build an optimized processed second order method involving a large number of stages that nevertheless is far more efficient than other integrators.

There are other issues in connection with splitting and composition methods that we have *not* tackled here, however, and that are also important in this context. Among them we can mention the following.

- As was remarked in the introduction, no general rule is provided here to split any given function f in the differential equation (1). It turns out that, for f within a certain class of ODEs, this can be done systematically, whereas for other functions one has to proceed on a case by case basis. Sometimes, several splittings are possible, and the different schemes built from them lead to the preservation of distinctive geometric properties. It makes sense, then, to classify the ODEs and their corresponding integration methods into different categories. This aspect has been analyzed in [57]. Moreover, in many physical problems there are several geometric properties that are conserved simultaneously along the evolution and it is not clear at all how to design methods preserving all of them. In that case, which one is the most relevant from a numerical point of view?

- In this paper we have only considered the initial value problem defined by eq. (1) and integration methods with constant step size h. Backward error analysis provides an argument why this has to be the case in geometric numerical integration: the modified equation corresponding to the numerical method depends explicitly on h, so that if h is changed so does the modified equation and no preservation of geometric properties is guaranteed. There are problems, however, where the use of an adaptive step size is mandatory, for instance in configurations of the N-body problem allowing close encounters. In this case one may apply splitting methods with variable step size by using some specifically designed transformations involving the time variable, in such a way that in the new variables the resulting time step is constant (see, e.g., [3]).

- As we have shown in section 3.5, the presence of negative coefficients in splitting methods of order higher than two is unavoidable. This is not a problem when the flow of the differential equation evolves in a group (such as in the Hamiltonian case), but may be unacceptable when the ODE originates from a partial differential equation that is ill-posed for negative time progression. Several alternatives have been proposed in the literature, mainly by considering, when possible, modified potentials [4], as noted in section 4.1. One should observe, however, that the analysis done in section 3.5 does not preclude the existence of *complex* coefficients with positive real parts. As a matter of fact, splitting methods with complex coefficients have been developed and tested for problems in which the Hamiltonian is split into kinetic and potential energy terms [24], for the time-dependent Schrödinger equation [2], for generic parabolic equations [23] and also in the more abstract setting of evolution PDEs in analytic semigroups [42].

- An important characteristic of any numerical integration method is *stability*. Roughly speaking, the numerical solution provided by a stable numerical integrator does not tend to infinity when the exact solution is bounded. Although important, this feature has received considerably less attention in the specific case of splitting methods. To test the (linear) stability of the method (36), instead of the linear equation $y' = ay$ as in the usual stability analysis for ODE integrators, one considers the harmonic oscillator $y'' + \lambda^2 y = 0$, $\lambda > 0$, as a model problem with a splitting of the form (11). The idea is to find the time steps for which all numerical solutions remain bounded. The integrator (36) typically will be unstable for $|h\lambda| > x_*$, where the parameter x_* determines the stability threshold of the numerical scheme. In particular, for the leapfrog method one has $x_* = 2$. Although the stability threshold imposes restrictions on the step size, in the process of building high order schemes, linear stability is not usually taken into account, ending sometimes with methods possessing such a small relative stability threshold that they are useless in practice. In this way, constructing high order splitting methods with relatively large linear stability intervals and highly accurate is of great interest. This has been achieved in reference [9] for linear systems, but remains an open problem in general.

- In section 5 we have mentioned an optimization criterion to choose the free parameters in splitting and composition methods, which consist in minimizing the Euclidean norm of the coefficients that constitute the leading error term of the method. It is clear, however, that minimizing the leading error term does not guarantee that the method thus obtained is the most efficient: it might occur that the influence of the subsequent error terms is the decisive factor in the performance of the scheme. In this sense, it would be extremely interesting to have estimates on all the error terms in the asymptotic expansion of the modified equation and get the coefficients of the method that minimize these estimates.

- The numerical analysis of second-order differential equations with oscillatory solutions has aroused much interest during the past few years. The typical test problem in this setting is the equation $q'' + \Omega^2 q = f(q)$, where Ω is a symmetric and positive definite matrix. Here the aim is to design new methods which improve in accuracy and stability the standard Störmer–Verlet integrator. We refer the reader to [41] and references therein for a comprehensive study of this problem.

- Although only ODEs have been considered here, splitting methods have been also applied with success to stochastic differential equations (SDEs). Here the aim is, as in the deterministic case, to design integration methods which automatically incorporate conservation properties the SDE possesses [61].

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). SB also aknowledges the support of the UPV through the project 20070307. We are especially grateful to Prof. Arieh Iserles for his kind invitation to write and submit this paper.

References

[1] V. I. Arnold. *Mathematical Methods of Classical Mechanics.* Springer-Verlag, GTM 60, Second edition, 1989.

[2] A.D. Bandrauk, E. Dehghanian, and H. Lu. Complex integration steps in decomposition of quantum exponential evolution operators. *Chem. Phys. Lett.*, 419:346–350, 2006.

[3] S. Blanes and C.J. Budd. Adaptive geometric integrators for hamiltonian problems with approximate scale invariance. *SIAM J. Sci. Comput.*, 26:1089–1113, 2005.

[4] S. Blanes and F. Casas. On the necessity of negative coefficients for operator splitting schemes of order higher than two. *Appl. Numer. Math.*, 54:23–37, 2005.

[5] S. Blanes and F. Casas. Splitting methods for non-autonomous separable dynamical systems. *J. Phys. A: Math. Gen.*, 39:5405–5423, 2006.

[6] S. Blanes, F. Casas, and A. Murua. On the numerical integration of ordinary differential equations by processed methods. *SIAM J. Numer. Anal.*, 42:531–552, 2004.

[7] S. Blanes, F. Casas, and A. Murua. Composition methods for differential equations with processing. *SIAM J. Sci. Comput.*, 27:1817–1843, 2006.

[8] S. Blanes, F. Casas, and A. Murua. Symplectic splitting operator methods tailored for the time-dependent Schrödinger equation. *J. Chem. Phys.*, 124:234105, 2006.

[9] S. Blanes, F. Casas, and A. Murua. On the linear stability of splitting methods. *Found. Comp. Math.*, 8:357–393, 2008.

[10] S. Blanes, F. Casas, J.A. Oteo, and J. Ros. The Magnus expansion and some of its applications. *Phys. Rep.*, 2008. In press.

[11] S. Blanes, F. Casas, and J. Ros. Symplectic integrators with processing: a general study. *SIAM J. Sci. Comput.*, 21:711–727, 1999.

[12] S. Blanes, F. Casas, and J. Ros. Processing symplectic methods for near-integrable Hamiltonian systems. *Celest. Mech. and Dyn. Astro.*, 77:17–35, 2000.

[13] S. Blanes, F. Casas, and J. Ros. High-order Runge–Kutta–Nyström geometric methods with processing. *Appl. Numer. Math.*, 39:245–259, 2001.

[14] S. Blanes, F. Casas, and J. Ros. New families of symplectic runge-kutta-nyström integration methods. In *Numerical Analysis and its Applications, LNCS 1988*, pages 102–109. Springer, 2001.

[15] S. Blanes and P.C. Moan. Splitting methods for non-autonomous Hamiltonian equations. *J. Comp. Phys.*, 170:205–230, 2001.

[16] S. Blanes and P.C. Moan. Practical symplectic partitioned Runge–Kutta and Runge–Kutta–Nyström methods. *J. Comput. Appl. Math.*, 142:313–330, 2002.

[17] C.J. Budd and A. Iserles. Geometric integration: numerical solution of differential equations on manifolds. *Phil. Trans. Royal Soc. A*, 357:945–956, 1999.

[18] J. Butcher. The effective order of Runge–Kutta methods. In *Conference on the Numerical Solution of Differential Equations*, Lecture Notes in Math. 109, pages 133–139, Berlin, 1969. Springer.

[19] M.P. Calvo and J.M. Sanz-Serna. The development of variable-step symplectic integrators, with applications to the two-body problem. *SIAM J. Sci. Comput.*, 14:936–952, 1993.

[20] M.P. Calvo and J.M. Sanz-Serna. High-order symplectic Runge–Kutta–Nyström methods. *SIAM J. Sci. Comput.*, 14:1237–1252, 1993.

[21] J. Candy and W. Rozmus. A symplectic integration algorithm for separable Hamiltonian functions. *J. Comp. Phys.*, 92:230–256, 1991.

[22] F. Casas and A. Murua. An efficient algorithm for computing the Baker–Campbell–Hausdorff series and some of its applications. Technical report, Universitat Jaume I, 2008.

[23] F. Castella, P. Chartier, S. Decombes, and G. Vilmart. Splitting methods with complex times for parabolic equations. Technical report, Université de Rennes, September 2008.

[24] J.E. Chambers. Symplectic integrators with complex time steps. *Astron. J.*, 126:1119–1126, 2003.

[25] P. Chartier and A. Murua. An algebraic theory of order. Technical report, 2008. Submitted.

[26] S.A. Chin. Symplectic integrators from composite operator factorizations. *Phys. Lett. A*, 226:344–348, 1997.

[27] S.A. Chin. Structure of positive decomposition of exponential operators. *Phys. Rev. E*, 71:016703, 2005.

[28] S.A. Chin and C.R. Chen. Fourth order gradient symplectic integrator methods for solving the time-dependent Schrödinger equation. *J. Chem. Phys.*, 114:7338–7341, 2001.

[29] L.Y. Chou and P.W. Sharp. Order 5 symplectic explicit Runge–Kutta–Nyström methods. *J. Appl. Math. Decision Sci.*, 4:143–150, 2000.

[30] M. Creutz and A. Gocksch. Higher-order hybrid Monte Carlo algorithms. *Phys. Rev. Lett.*, 63:9–12, 1989.

[31] J.M.A. Danby. *Fundamentals of Celestial Mechanics.* Willmann-Bell, 1988.

[32] M.D. Feit, Jr. J.A. Fleck, and A. Steiger. Solution of the Schrödinger equation by a spectral method. *J. Comp. Phys.*, 47:412–, 1982.

[33] E. Forest. Sixth-order Lie group integrators. *J. of Comp. Phys.*, 99:209–213, 1992.

[34] E. Forest and R.D. Ruth. Fourth-order symplectic integration. *Physica D*, 43:105–117, 1990.

[35] D. Goldman and T.J. Kaper. Nth-order operator splitting schemes and nonreversible systems. *SIAM J. Numer. Anal.*, 33:349–367, 1996.

[36] H. Goldstein. *Classical Mechanics.* Addison Wesley, Second edition, 1980.

[37] S. Gray and D.E. Manolopoulos. Symplectic integrators tailored to the time-dependent Schrödinger equation. *J. Chem. Phys.*, 104:7099–7112, 1996.

[38] S. Gray and J.M. Verosky. Classical Hamiltonian structures in wave packet dynamics. *J. Chem. Phys.*, 100:5011–5022, 1994.

[39] J. Guckenheimer and P. Holmes, editors. *Nonlinear Oscillations, Dynamical Systems, and Bifurcations of Vector Fields.* Springer, 1983.

[40] E. Hairer, Ch. Lubich, and G. Wanner. *Geometric Numerical Integration. Structure-Preserving Algorithms for Ordinary Differential Equations.* Springer-Verlag, 2002.

[41] E. Hairer, Ch. Lubich, and G. Wanner. *Geometric Numerical Integration. Structure-Preserving Algorithms for Ordinary Differential Equations.* Springer-Verlag, Second edition, 2006.

[42] E. Hansen and A. Ostermann. High order splitting methods for analytic semigroups exist. Technical report, Institut für Mathematik, Universität Innsbruck, 2008.

[43] M. Hénon and C. Heiles. The applicability of the third integral of motion: some numerical experiments. *Astron. J.*, 69:73–79, 1964.

[44] A. Iserles, H. Z. Munthe-Kaas, S. P. Nørsett, and A. Zanna. Lie-group methods. *Acta Numerica*, 9:215–365, 2000.

[45] W. Kahan and R.C. Li. Composition constants for raising the order of unconventional schemes for ordinary differential equations. *Math. Comp.*, 66:1089–1099, 1997.

[46] U. Kirchgraber. An ODE-solver based on the method of averaging. *Numer. Math.*, 53:621–652, 1988.

[47] P.-V. Koseleff. *Formal Calculus for Lie Methods in Hamiltonian Mechanics.* PhD thesis, Lawrence Berkeley Laboratory, 1994.

[48] J. Laskar and P. Robutel. High order symplectic integrators for perturbed Hamiltonian systems. *Celest. Mech. and Dyn. Astro.*, 80:39–62, 2001.

[49] B. Leimkuhler and S. Reich. *Simulating Hamiltonian Dynamics.* Cambridge University Press, 2004.

[50] X. Liu, P. Ding, J. Hong, and L. Wang. Optimization of symplectic schemes for time-dependent Schrödinger equations. *Comput. Math. Appl.*, 50:637–, 2005.

[51] M.A. López-Marcos, J.M. Sanz-Serna, and R.D. Skeel. Explicit symplectic integrators using Hessian-vector products. *SIAM J. Sci. Comput.*, 18:223–238, 1997.

[52] E. Lorenz. Deterministic nonperiodic flows. *J. Atmos. Sci.*, 20:130–141, 1963.

[53] R. I. McLachlan. Composition methods in the presence of small parameters. *BIT*, 35:258–268, 1995.

[54] R. I. McLachlan. On the numerical integration of ordinary differential equations by symmetric composition methods. *SIAM J. Numer. Anal.*, 16:151–168, 1995.

[55] R. I. McLachlan. Families of high-order composition methods. *Numer. Alg.*, 31:233–246, 2002.

[56] R. I. McLachlan and P. Atela. The accuracy of symplectic integrators. *Nonlinearity*, 5:541–562, 1992.

[57] R.I. McLachlan and R. Quispel. Splitting methods. *Acta Numerica*, 11:341–434, 2002.

[58] R.I. McLachlan and R. Quispel. Geometric integrators for ODEs. *J. Phys. A: Math. Gen.*, 39:5251–5285, 2006.

[59] R.I. McLachlan and B. Ryland. The algebraic entropy of classical mechanics. *J. Math. Phys.*, 44:3071–3087, 2003.

[60] L. Meirovich. *Methods of Analytical Dynamics.* McGraw-Hill, 1988.

[61] T. Misawa. A Lie algebraic approach to numerical integration of stochastic differential equations. *SIAM J. Sci. Comput.*, 23:866–890, 2001.

[62] A. Murua. The Hopf algebra of rooted trees, free Lie algebras, and Lie series. *Found. Comp. Math.*, 6:387–426, 2006.

[63] A. Murua and J. M. Sanz-Serna. Order conditions for numerical integrators obtained by composing simpler integrators. *Phil. Trans. Royal Soc. A*, 357:1079–1100, 1999.

[64] D.I. Okunbor and E.J. Lu. Eight-order explicit symplectic Runge–Kutta–Nyström integrators. Technical Report CSC 94-21, Dept. of Computer Science, University of Missouri-Rolla, 1994.

[65] D.I. Okunbor and R.D. Skeel. Canonical Runge–Kutta–Nyström methods of orders five and six. *J. Comput. Appl. Math*, 51:375–382, 1994.

[66] P. J. Olver. *Applications of Lie Groups to Differential Equations.* GTM 107. Springer-Verlag, Second edition, 1993.

[67] I.P. Omelyan, I.M. Mryglod, and R. Folk. On the construction of high-order force gradient algorithms for integration of motion in classical and quantum systems. *Phys. Rev. E*, 66:026701, 2002.

[68] L.P. Pitaevskii and S. Stringari. *Bose–Einstein Condensation.* Clarendon Press, 2003.

[69] C. Reutenauer. *Free Lie algebras*, volume 7 of *London Math. Soc. monographs (new series).* Oxford University Press, 1993.

[70] R. Rieben, D. White, and G. Rodrigue. High-order symplectic integration methods for finite element solutions to time dependent Maxwell equations. *IEEE Trans. Antennas Propagat.*, 52:2190–2195, 2004.

[71] G. Rowlands. A numerical algorithm for Hamiltonian systems. *J. of Comp. Phys.*, 97:235–239, 1991.

[72] R. Ruth. A canonical integration technique. *IEEE Trans. Nucl. Sci.*, 30:26–69, 1983.

[73] J. M. Sanz-Serna and M. P. Calvo. *Numerical Hamiltonian Problems.* AMMC 7. Chapman & Hall, 1994.

[74] J.M. Sanz-Serna and A. Portillo. Classical numerical integrators for wave-packet dynamics. *J. Chem. Phys.*, 104:2349–2355, 1996.

[75] Q. Sheng. Solving linear partial differential equations by exponential splitting. *IMA J. Numer. Anal.*, 9:199–212, 1989.

[76] M. Sofroniou and G. Spaletta. Derivation of symmetric composition constants for symmetric integrators. *Optimization Methods Software*, 20:597–613, 2005.

[77] G. Strang. On the construction and comparison of difference schemes. *SIAM J. Numer. Anal.*, 5:506–517, 1968.

[78] M. Suzuki. Fractal decomposition of exponential operators with applications to many-body theories and Monte Carlo simulations. *Phys. Lett. A*, 146:319–323, 1990.

[79] M. Suzuki. General theory of fractal path integrals with applications to many-body theories and statistical physics. *J. Math. Phys.*, 32:400–407, 1991.

[80] M. Suzuki. Hybrid exponential product formulas for unbounded operators with possible applications to Monte Carlo simulations. *Phys. Lett. A*, 201:425–428, 1995.

[81] M. Suzuki and K. Umeno. Higher-order decomposition theory of exponential operators and its applications to QMC and nonlinear dynamics. In *Computer Simulation Studies in Condensed-Matter Physics VI*, Springer Proceedings in Physics 76, pages 74–86, Berlin, 1993. Springer.

[82] M. Takahashi and M. Imada. Montecarlo calculation of quantum system. II. Higher order correction. *J. Phys. Soc. Japan*, 53:3765–3769, 1984.

[83] Ch. Tsitouras. A tenth order symplectic Runge–Kutta–Nyström method. *Celest. Mech. and Dyn. Astron.*, 74:223–230, 1999.

[84] V. S. Varadarajan. *Lie Groups, Lie Algebras, and Their Representations.* GTM 102. Springer-Verlag, 1984.

[85] L. Verlet. Computer “experiments” on classical fluids. I. Thermodynamical properties of Lennard–Jones molecules. *Phys. Rev.*, 159:98–103, 1967.

[86] J. Wisdom and M. Holman. Symplectic maps for the N-body problem. *Astron. J.*, 102:1528–1538, 1991.

[87] J. Wisdom, M. Holman, and J. Touma. Symplectic correctors. In *Integration Algorithms and Classical Mechanics*, pages 217–244, Providence, RI, 1996. American Mathematical Society.

[88] H. Yoshida. Construction of higher order symplectic integrators. *Phys. Lett. A*, 150:262–268, 1990.

[89] W. Zhu, X. Zhao, and Y. Tang. Numerical methods with high order of accuracy applied in the quantum system. *J. Chem. Phys.*, 104:2275–, 1996.

LA ECUACIÓN DE RICCATI

JOSÉ MARíA AMIGÓ

Centro de Investigación Operativa
Universidad Miguel Hernández de Elche

jm.amigo@umh.es

Resumen

En este artículo repasamos algunas propiedades y aspectos interesantes de la ecuación de Riccati.

Palabras clave: *ecuación de Riccati, propiedades generales, métodos de resolución.*
Clasificación por materias AMS: *3401*

1 Introducción

Al igual que ocurre con otras ecuaciones diferenciales ordinarias con 'pedigrí', la ecuación de Riccati (o Ricatti, que en esto parece haber distintas opiniones) tiene la maravillosa costumbre de aparecer en campos muy diversos —tan diversos, en el caso que nos ocupa, como la física aplicada y teórica, la geometría diferencial y proyectiva, el cálculo de variaciones, la teoría de control y programación dinámica, filtros lineales, matemática financiera, ecología, ... La búsqueda de *Ricatti equation* en Google produce aproximadamente 184,000 resultados, mientras que *Riccati equation* 'sólo' produce unos 167,000. Parte de esta universalidad, tanto en matemática pura como aplicada, se debe a sus propiedades proyectivas y a su relación con la ecuación lineal de segundo orden (en particular, con la ecuación de Bessel). No es, por tanto, de extrañar que fuera estudiada por algunos de los matemáticos más eminentes de los siglos pasados, como Euler, Cayley, Liouville, Schlafli y Glaisher, entre otros.

La ecuación de Riccati puede ser considerada como la ecuación de primer orden no-lineal más sencilla, puesto que se obtiene de la ecuación lineal completa de primer orden sin más que añadir un término cuadrático en la variable dependiente. Sin embargo, esta ecuación sólo se sabe resolver cuando se conoce una solución particular o en casos especiales. Es más, Liouville probó la imposibilidad, en general, de resolver la ecuación de Riccati mediante cuadraturas. Más allá de este aspecto puramente calculístico, la ecuación de Riccati tiene muchas propiedades y aspectos interesantes que vale la pena

Fecha de recepción: 20/04/2008. Aceptado (en forma revisada): 30/04/2008.

recordar. En este artículo intentaremos, precisamente, hacer una excursión por una selección de ellos. Aparte de los temas propios de la teoría de las ecuaciones diferenciales (algunos quizá ya olvidados), ante nuestros ojos desfilarán temas de la mecánica clásica (caída de graves), de la geometría diferencial (ecuaciones de Frenet-Serret), de funciones especiales (funciones de Bessel) y hasta de la teoría de números (fracciones continuas). Pero todos ellos compartiendo la belleza de la buena matemática, sin distinción de de pura o aplicada.

Este artículo está estructurado de la siguiente manera. Después de una breve reseña histórica, veremos en la Sección 3 cinco muestras, tomadas de otras tantas áreas de la matemática o de la física, en las que hace acto de presencia la ecuación de Riccati —en alguna de ellas, diría yo, de forma inesperada. En la Sección 4 repasaremos tres propiedades básicas de la ecuación de Riccati, antes de pasar a estudiar en las tres últimas secciones, el problema de la integración de la ecuación general (Sección 5) y de su forma original (Sección 6 y 7). A lo largo de todas estas secciones utilizaremos la "ecuación del paracaidista" como ejemplo recurrente para ilustrar las distintas técnicas de resolución. De hecho, la resolveremos nada menos que con seis técnicas distintas —eso sí, ¡la solución siempre es la misma!

2 Referencias históricas

Si eliminamos la constante K del haz de curvas

$$y = \frac{a(x) + Kb(x)}{c(x) + Kd(x)}, \tag{1}$$

donde

1. $a(x), ..., d(x)$ son funciones diferenciables en cierto intervalo abierto $I \subset \mathbb{R}$ o dominio $\Omega \subset \mathbb{C}$,
2. $\Delta(x) := b(x)c(x) - a(x)d(x) \not\equiv 0$,

llegamos a una ecuación diferencial no lineal (la ecuación diferencial de primer orden del haz (1)) de la forma

$$y' + q(x)y + r(x)y^2 = p(x), \tag{2}$$

donde

$$q = \frac{a'd - ad' + bc' - b'c}{\Delta}, \quad r = \frac{cd' - cd'}{\Delta}, \quad p = \frac{a'b - ab'}{\Delta}.$$

Observemos que la ecuación (2) incluye a la ecuación lineal ($r(x) \equiv 0$) y a la ecuación de Bernoulli de exponente 2 ($p(x) \equiv 0$).

La ecuación (2) se llama *ecuación de Riccati* en honor del conde italiano Jacopo Francesco Riccati (Venecia 1676 - Treviso 1754), un sabio italiano que se dedicó principalmente a la hidráulica y que, en 1724, publicó[1] una ecuación

[1] Animadversationes in aequationes differentiales secundi gradus, *Actorum Eruditorum quae Lipsiae publicantur. Supplementa* **8** (1724), 66-73.

equivalente a

$$y' + Ay^2 = Bx^n \quad (A, B \in \mathbb{R} \text{ o } \mathbb{C}, \quad n \in \mathbb{R}), \tag{3}$$

que es un caso especial de la (2). Nosotros nos referiremos a (3) como *la ecuación original de Riccati*, cuando queramos particularizar. Riccati sólo estudió algunos casos concretos de esta ecuación, sin hallar solución alguna. Añadamos que dos de los tres hijos del conde Riccati, Vincenzo y Giordano, se dedicaron también a las matemáticas y a la física: el primero (un jesuita) estudió las funciones hiperbólicas y el segundo midió el módulo de Young de algunos metales —¡adelantándose a Young en 25 años!—, amén de dedicarse a la arquitectura, hidráulica y música.

Pero tuvo que ser otra familia de matemáticos, los Bernoulli, la que resolviera la ecuación (3). En efecto, con anterioridad a Riccati, en 1694, Johannes Bernoulli había intentado ya resolver sin éxito la ecuación

$$y' + y^2 = x^2.$$

Tuvieron que pasar 9 años para que su hermano Jakob obtuviera una solución en serie de potencias,

$$y = \frac{1}{3}x^3 + \frac{1}{3^2 \cdot 7}x^7 + \frac{2}{3^3 \cdot 7 \cdot 11}x^{11} + \frac{13}{3^4 \cdot 5 \cdot 7^2 \cdot 11}x^{15} + ...$$

Finalmente, en 1725, Daniel Bernoulli publicó la solución de la ecuación de Riccati original para aquellos valores de n para los que existe solución elemental (véase Corolario 3, Sección 6).

Nota 1 *El cambio de función*

$$u(x) = y(x) \exp\left(\int_{x_0}^{x} q(\xi) d\xi\right)$$

en la ecuación de Riccati permite eliminar el término lineal en la función incógnita, transformándola en la ecuación

$$u' + \widetilde{r}(x)u^2 = \widetilde{p}(x)$$

con

$$\widetilde{r}(x) := r(x) \exp\left(-\int_{x_0}^{x} q(\xi) d\xi\right), \quad \widetilde{p}(x) := p(x) \exp\left(\int_{x_0}^{x} q(\xi) d\xi\right).$$

Luego, siempre que sea conveniente, podemos suponer sin pérdida de generalidad que la ecuación de Riccati está en forma simplificada, *es decir, que* $q(x) \equiv 0$ *en* (2).

3 Ejemplos

3.1 Aproximación de ecuaciones no lineales

Supongamos que el campo vectorial de la ecuación diferencial normal

$$y' = f(x,y) \tag{4}$$

se puede desarrollar en potencias de y uniformemente en $x \in \mathbb{R}$, es decir,

$$f(x,y) = f(x,0) + f_y(x,0)y + \frac{1}{2}f_{yy}(x,0)y^2 + ... \quad (x \in I \subset \mathbb{R},\ |y| < M).$$

Si aproximamos $f(x,y)$ por los dos primeros sumandos de este desarrollo en una banda suficientemente estrecha del eje X, el haz integral de la ecuacion lineal

$$y' = f(x,0) + f_y(x,0)y$$

representará, en primera aproximación, el haz integral de la ecuación (4) en la banda $I \times (-M,M)$.

Si, con objeto de mejorar la aproximación en $I \times (-M,M)$, tomamos un término más del desarrollo de $f(x,y)$ en potencias de y, nos vemos conducidos a una ecuación de Riccati:

$$y' = f(x,0) + f_y(x,0)y + \frac{1}{2}f_{yy}(x,0)y^2.$$

3.2 Mecánica

Consideremos la caída vertical de un cuerpo de masa m bajo la fuerza de la gravedad en un medio que ofrece una resistencia proporcional al cuadrado de la velocidad del cuerpo. Si $s(t)$ es la distancia recorrida por el grave en el tiempo t (siendo $t_0 = 0$ el tiempo inicial) y $v = ds/dt \geq 0$ la velocidad de caída, entonces su movimiento viene dado por la ley de Newton,

$$\frac{dv}{dt} = g - \frac{\kappa}{m}v^2 \quad \text{o} \quad \frac{dv}{dt} = g\left(1 - \frac{\kappa}{gm}v^2\right), \tag{5}$$

donde $g = 9{,}81...\ \mathrm{ms}^{-2}$ es la aceleración de la gravedad y $\kappa > 0$ es una constante llamada *coeficiente de fricción* del cuerpo (que depende de la densidad del medio y de la geometría del cuerpo). La ecuación (5), que denominaremos en lo sucesivo *la ecuación del paracaidista* por describir la caída de graves en la atmósfera a la velocidad que alcanza un paracaidísta, es una ecuación original de Riccati que pasamos a resolver.

En primer lugar, la ecuación del paracaidista puede ser adimensionalizada mediante los cambios de variable

$$v \leftarrow \frac{v}{v_\infty},\ \ t \leftarrow \frac{g}{v_\infty}t = \gamma t, \tag{6}$$

donde

$$v_\infty := \sqrt{\frac{gm}{\kappa}} \ \text{ y } \ \gamma := \frac{g}{v_\infty} = \sqrt{\frac{g\kappa}{m}} \tag{7}$$

tienen dimensiones de velocidad y frecuencia, repectivamente. En la nuevas variables adimensionales, que por economía notacional seguiremos llamando v y t, la ecuación queda:

$$\frac{dv}{dt} = 1 - v^2. \tag{8}$$

1. *Solución singular*: Escribiendo el miembro derecho de (8) como

$$\left(1 - v^2\right) = (1 + v)\,(1 - v)\,,$$

llegamos a la conclusión de que (8) admite la solución constante

$$v(t) \equiv 1.$$

2. *Solución general*: Si $v \not\equiv 1$, entonces podemos resolver (8) separando variables e integrando la ecuación resultante con la condición inicial $v(0) = v_0 \geq 0$,

$$\int_0^t dt = \frac{1}{g} \int_{v_0}^{v} \frac{dv}{1 - v^2},$$

con el resultado

$$t = \begin{cases} \operatorname{arg\,tanh} v - C & \text{si } v < 1, \\ \\ \operatorname{arg\,coth} v - C & \text{si } v > 1, \end{cases} \tag{9}$$

donde C es la constante de integración, distinta en cada caso. Observemos que la solución tiene una discontinuidad (no está definida) en $v = 1$, por lo que el movimiento de descenso de nuestro imaginario paracaidista será acelerado si $v_0 < 1$ (tal y como ocurre antes de abrir el paracaídas), correspondiendo a la solución con el $\operatorname{arg\,tanh} v$, y desacelerado si $v_0 > 1$ (tal y como ocurre después de abrir el paracaídas), correspondiendo a la solución con el $\operatorname{arg\,coth} v$. La solución singular viene a llenar el hueco que no cubre (9): si $v_0 = 1$, el movimiento es uniforme.

Despejando v de (9), deducimos

$$v(t) = \frac{e^{t+C} \mp e^{-(t+C)}}{e^{t+C} \pm e^{-(t+C)}} = \frac{e^t - Ke^{-t}}{e^t + Ke^{-t}} \tag{10}$$

con $K = \pm e^{-2C} \gtrless 0$, según sea $v_0 \lessgtr 1$. La solución singular $v \equiv 1$ puede ser incluida en el haz general permitiendo que $K = 0$. Puesto que

$$\lim_{t \to \infty} v(t) = 1, \quad \text{con} \quad \begin{cases} v(t) \nearrow 1 & \text{si } v_0 < 1, \\ v(t) \searrow 1 & \text{si } v_0 > 1, \end{cases}$$

vemos que la solución singular, $v(t) \equiv v_\infty$ en dimensiones físicas, corresponde a la *velocidad límite de caída*.

La constante K se puede fijar fácilmente mediante la velocidad inicial v_0, resultando

$$K = \frac{1 - v_0}{1 + v_0}. \tag{11}$$

En particular, si $v_0 = 0$, entonces $K = 1$. Sustituyendo (11) en la solución (10) deducimos

$$v(t) = \frac{v_0 + \tanh t}{1 + v_0 \tanh t}.$$

Finalmente, integrando $ds/dt = v$ con la condición inicial $s(0) = 0$, obtenemos

$$s(t) = \int_0^t v(\tau)d\tau = \ln\left(\cosh t + v_0 \operatorname{senh} t\right).$$

En los casos prácticos ($v_0 = 0$ y $t \gg 1$), esta fórmula se aproxima por

$$s(t) = t - \ln 2$$

con gran precisión. Recuperando las variables dimensionadas,

$$t \leftarrow \frac{t}{\gamma}, \quad s \leftarrow \frac{v_\infty}{\gamma} s,$$

obtenemos la aproximación

$$s(t) = v_\infty t - \frac{v_\infty^2 \ln 2}{g}.$$

3.3 Sistemas lineales homogéneos de segundo orden

Consideremos el sistema lineal homogéneo de orden 2 real (o sistema dinámico lineal autónomo en el plano real)

$$\left.\begin{aligned} \frac{dx}{dt} &= a(t)x + b(t)y \\ \frac{dy}{dt} &= c(t)x + d(t)y \end{aligned}\right\} \qquad (12)$$

Los métodos tradicionales de análisis de este tipo de sistemas se basan en cambios sencillos de coordenadas (por ejemplo, coordenadas polares) o en el estudio del comportamiento cualitativo de las soluciones en función de los autovalores de la matriz de coeficientes. Aquí consideraremos la transformación proyectiva de coordenadas

$$\left.\begin{array}{l} u = xy \\ v = y/x \end{array}\right\} \Leftrightarrow \left\{\begin{array}{l} x^2 = u/v \\ y^2 = uv \end{array}\right.$$

la cual transforma el sistema (12) en

$$\left.\begin{aligned} \frac{1}{u}\frac{du}{dt} &= bv + (a + d) + cv^{-1} \\ \frac{dv}{dt} &= -bv^2 + (d - a)v + c \end{aligned}\right\}$$

que nos dice que tanto la derivada logarítmica de u como la derivada de v dependen sólo de v. Observemos que la segunda ecuación es de Riccati. Basta, pues, resolver esta ecuación de Riccati para obtener la solución del sistema lineal homogéneo de segundo orden.

Este resultado conduce directamente a la siguente proposición:

Corolario 1 *La ecuación de Riccati*

$$v' + (a(t) - d(t))\, v + b(t)v^2 = c(t)$$

tiene solución en el intervalo $I \subset \mathbb{R}$ *si y sólo si el sistema lineal homogéneo* (12) *tiene una solución* $(x(t), y(t))$ *tal que* $x(t) \neq 0$ *y* $v(t) = y(t)/x(t)$ *para* $t \in I$.

3.4 Las ecuaciones de Frenet-Serret

Dada una curva diferenciable (o 'alabeada') $\mathbf{r} : [\alpha, \beta] \to \mathbb{R}^3$ parametrizada por la longitud s, recordemos que el *vector tangente* $\mathbf{t}(s) = \dot{\mathbf{r}}(s)$ (donde el punto denota derivación respecto de s), el *vector normal principal* $\mathbf{n}(s) = \dot{\mathbf{t}}(s)/\left|\dot{\mathbf{t}}(s)\right|$ y el *vector binormal* $\mathbf{b}(s) = \mathbf{t}(s) \times \mathbf{n}(s)$ forman en todo punto $\mathbf{r}(s)$ del rango $C = \mathbf{r}([\alpha, \beta])$ un conjunto ortonormal de vectores, llamado *triedro de Frenet* de la curva (nosotros no distinguiremos entre curva y rango). El triedro de Frenet verifica las *ecuaciones de Frenet-Serret*:

$$\begin{cases} \dot{\mathbf{t}} = \kappa \mathbf{n} \\ \dot{\mathbf{n}} = -\kappa \mathbf{t} + \tau \mathbf{b} \\ \dot{\mathbf{b}} = -\tau \mathbf{n} \end{cases}$$

donde $\kappa(s)$ y $\tau(s)$ son la *curvatura* y *torsión*, respectivamente, de C en el punto $\mathbf{r}(s)$. Escritas en componentes, las ecuaciones de Frenet-Serret forman un sistema lineal inhomogéneo de nueve ecuaciones

$$\dot{t}_j = \kappa n_j, \quad \dot{n}_j = -\kappa t_j + \tau b_j, \quad \dot{b}_j = -\tau n_j \quad (j = 1, 2, 3). \tag{13}$$

Sin embargo, sólo 6 de las 9 funciones incógnitas son funcionalmente independientes, ya que se verifican las 3 ligaduras siguientes:

$$t_j^2 + n_j^2 + b_j^2 = 1 \quad (j = 1, 2, 3). \tag{14}$$

En efecto: de las ecuaciones (13) se deduce, para cada $j \in \{1, 2, 3\}$,

$$\begin{aligned} \frac{1}{2}\frac{d}{ds}\left(t_j^2 + n_j^2 + b_j^2\right) &= t_j\dot{t}_j + \ n_j\dot{n}_j + \ b_j\dot{b}_j \\ &= t_j(\kappa n_j) + \ n_j(-\kappa t_j + \tau b_j) + \ b_j(-\tau n_j) \\ &= 0 \end{aligned}$$

de manera que $t_j^2 + n_j^2 + b_j^2 = const.$ Que esta constante vale 1 es consecuencia de la ortonormalidad del triedro $\{\mathbf{t}, \mathbf{n}, \mathbf{b}\}$.

El teorema fundamental de existencia y unicidad de las curvas alabeadas en $\mathbb{R}^3$ (que es consecuencia del teorema de Picard) afirma que, dadas dos funciones continuas $\kappa(s)$ y $\tau(s)$ en $[\alpha,\beta] \subset \mathbb{R}$, existe (salvo congruencias) una y sólo una curva C cuya curvatura es $\kappa(s)$, su torsión es $\tau(s)$ y s es el parámetro longitud a lo largo de ella. Las ecuaciones $\kappa = \kappa(s)$, $\tau = \tau(s)$ se llaman *ecuaciones intrínsecas* de la curva C.

Fijemos $j \in \{1,2,3\}$ en (13) y procedamos a integrar las 3 ecuaciones resultantes. Puesto que cada conjunto de soluciones $t_j(s)$, $n_j(s)$, $b_j(s)$ tiene que satisfacer la restricción (14), introducimos con Darboux dos funciones complejas $\sigma_j(s)$ y $\omega_j(s)$ definidas por

$$\begin{aligned} \sigma_j &= \frac{t_j + in_j}{1-b_j} = \frac{1+b_j}{t_j - in_j}, \\ -\frac{1}{\omega_j} &= \overline{\sigma_j} = \frac{t_j - in_j}{1-b_j} = \frac{1+b_j}{t_j + in_j}. \end{aligned}$$

La transformación inversa es la siguiente:

$$t_j = \frac{1-\sigma_j\omega_j}{\sigma_j - \omega_j}, \quad n_j = i\frac{1+\sigma_j\omega_j}{\sigma_j - \omega_j}, \quad b_j = i\frac{\sigma_j + \omega_j}{\sigma_j - \omega_j}. \tag{15}$$

Si estas expresiones se sustituyen en (13), resulta que las funciones σ_j y ω_j son soluciones de la ecuación de Riccati

$$\frac{dy}{ds} + i\kappa y - \frac{i}{2}\tau y^2 = -\frac{i}{2}\tau. \tag{16}$$

Recíprocamente, una solución particular $\sigma_j(s)$ de la ecuación (16) define, a través de las sustituciones (15) con $\omega_j = -1/\overline{\sigma_j}$, una solución $(t_j(s), n_j(s), b_j(s))$ del sistema (13), que satisface la restricción (14). La integración de las ecuaciones de Frenet-Serret se reduce, pues, a la integración de la ecuación (16).

3.5 La forma canónica de Forsyth-Laguerre

Consideremos la ecuación lineal homogénea de orden n

$$L[x,y] \equiv y^{(n)} + \binom{n}{1} p_1(x) y^{(n-1)} + \binom{n}{2} p_2(x) y^{(n-2)} + ... + p_n(x) = 0. \tag{17}$$

Se puede probar que el cambio de variables $x \leftarrow \xi$, $y \leftarrow \eta$ más general que transforma (17) en otra ecuación del mismo tipo y orden es de la forma

$$x = f(\xi), \quad y = \lambda(\xi)\eta, \tag{18}$$

donde $f(\xi)$ y $\lambda(\xi)$ son funciones “arbitrarias” de ξ. Esta libertad puede utilizarse para eliminar los términos correspondientes a las derivadas $n-1$ y $n-2$, obteniendo de este modo la denominada *forma canónica de Forsyth-Laguerre*

$$y^{(n)} + \binom{n}{3} p_3(x) y^{(n-3)} + ... + p_n(x) = 0$$

de una ecuación homogénea lineal de orden n. A continuación estudiaremos separadamente los cambios (18).

1. El cambio de variable dependiente $y = \lambda(x)\eta$ transforma $L[x, y] = 0$ en la ecuación

$$L[x, \eta] \equiv \eta^{(n)} + \binom{n}{1}\pi_1(x)\eta^{(n-1)} + \binom{n}{2}\pi_2(x)\eta^{(n-2)} + ... + \pi_n(x) = 0, \qquad (19)$$

donde

$$\begin{aligned}
\pi_1(x) &= \frac{1}{\lambda}\left[\lambda' + p_1\lambda\right], \\
\pi_2(x) &= \frac{1}{\lambda}\left[\lambda^{''} + 2p_1\lambda' + p_2\lambda\right], \ldots \\
\pi_n(x) &= \frac{1}{\lambda}\left[\lambda^{(n)} + \binom{n}{1}p_1\lambda^{(n-1)} + \binom{n}{2}p_2\lambda^{(n-2)} + ... + p_n\lambda\right].
\end{aligned}$$

Luego si sustituimos

$$\lambda(x) = e^{-\int p_1(x)dx},$$

en la ecuación (19), resulta $\pi_1(x) \equiv 0$ y la ecuación queda en la *forma semicanónica*

$$\eta^{(n)} + \binom{n}{2}P_2(x)\eta^{(n-2)} + \binom{n}{3}P_3(x)\eta^{(n-3)} + ... + P_n(x) = 0, \qquad (20)$$

con

$$P_2 = p_2 - p_1^2 - p_1'$$

y, en general,

$$P_k(x) = e^{-\int p_1(x)dx}\sum_{j=0}^{k}\binom{k}{j}\frac{d^{k-j}}{dx^{k-j}}\left(e^{-\int p_1(x)dx}\right) \qquad (k = 2, 3, ..., n).$$

2. El cambio de variable independiente $\xi = \xi(x)$ transforma las derivadas $y^{(m)}(x)$ en $L[x, y] = 0$ de la siguiente manera:

$$\frac{dy^m}{dx^m} = \sum_{k=1}^{m}\frac{A_{mk}}{k!}\frac{d^k y}{d\xi^k} \qquad (m = 1, 2, ..., n), \qquad (21)$$

donde los coeficientes A_{mk} son polinomios en las derivadas de $\xi(x)$. Los tres más sencillos son:

$$\begin{aligned}
\frac{A_{mm}}{m!} &= \left(\xi'\right)^m, \\
\frac{A_{m,m-1}}{(m-1)!} &= \binom{m}{2}\xi''\left(\xi'\right)^{m-2}, \\
\frac{A_{m,m-2}}{(m-2)!} &= \binom{m}{3}\xi^{'''}\left(\xi'\right)^{m-3} + 3\binom{m}{4}\left(\xi^{''}\right)^2\left(\xi'\right)^{m-4}.
\end{aligned}$$

La sustitución de las expresiones (21) en la ecuación $L[x,y]=0$ la convierte en

$$L[\xi,y]\equiv y^{(n)}+\binom{n}{1}\bar{p}_1(\xi)y^{(n-1)}+\binom{n}{2}\bar{p}_2(\xi)y^{(n-2)}+...+\bar{p}_n(\xi)=0,$$

donde

$$\binom{n}{r}(\xi')^n\,\bar{p}_r=\sum_{i=0}^{r}\binom{n}{i}\frac{A_{n-i,n-r}}{(n-r)!}\,p_i\qquad(r=0,1,...,n),$$

con $\bar{p}_0=p_0=1$. En particular,

$$\begin{aligned}\bar{p}_1(x)&=\frac{1}{\xi'}\left[p_1+\frac{n-1}{2}\zeta\right]\quad\text{con }\zeta=\frac{\xi''}{\xi'},\\ \bar{p}_2(x)&=\frac{1}{(\xi')^2}\left[p_2+(n-2)p_1\zeta+\frac{1}{12}(3n^2-11n+10)\zeta^2+\frac{n-2}{3}\zeta'\right].\end{aligned}$$

Vemos, pues, que la forma semicanónica (es decir, $p_1(x)\equiv 0$) no es invariante bajo cambios generales de la variable independiente $\xi=\xi(x)$.

3. Consideremos nuevamente la ecuación $L[x,y]=0$. Según **1.** y **2.**, si hacemos sucesivamente los cambios de variables $y=\lambda(x)\bar{y}$ y $\bar{x}=\xi(x)$, entonces

$$\begin{aligned}p_1(x)&\mapsto\pi_1(x)=\frac{1}{\lambda}\left[\lambda'+p_1\lambda\right]=\frac{\lambda'}{\lambda}+p_1\\ &\mapsto\bar{p}_1(\xi)=\frac{1}{\xi'}\left[\pi_1+\frac{n-1}{2}\zeta\right]=\frac{1}{\xi'}\left[\frac{\lambda'}{\lambda}+p_1+\frac{n-1}{2}\zeta\right],\end{aligned}\tag{22}$$

(donde $\zeta=\xi''/\xi'$) y

$$\begin{aligned}p_2(x)&\mapsto\pi_2(x)=\frac{1}{\lambda}\left[\lambda''+2p_1\lambda'+p_2\lambda\right]=\frac{\lambda''}{\lambda}+2p_1\frac{\lambda'}{\lambda}+p_2\\ &\mapsto\bar{p}_2(\xi)=\frac{1}{(\xi')^2}\left[\pi_2+(n-2)\pi_1\zeta+\frac{1}{12}(3n^2-11n+10)\zeta^2+\frac{n-2}{3}\zeta'\right].\end{aligned}\tag{23}$$

Si suponemos, sin pérdida de generalidad, que la ecuación de partida estaba en forma semicanónica (es decir, $p_1(x)\equiv 0$) y queremos que, después de los cambios efectuados, siga estándolo, se ha de verificar (véase (22))

$$\frac{\lambda'}{\lambda}+\frac{n-1}{2}\frac{\xi''}{\xi'}=0\quad\Rightarrow\quad\lambda(x)=C\,(\xi')^{-(n-1)/2}\quad(C\in\mathbb{R}).$$

Sustituyendo ahora

$$\pi_2=\frac{\lambda''}{\lambda}+p_2=-\frac{n-1}{2}\left(\zeta'-\frac{n-1}{2}\zeta^2\right),\quad\pi_1=\frac{\lambda'}{\lambda}=-\frac{n-1}{2}\zeta$$

en (23) y exigiendo $\bar{p}_2\equiv 0$, obtenemos la ecuación de Riccati simplificada

$$\zeta'-\frac{1}{2}\zeta^2=\frac{6}{n+1}p_2\tag{24}$$

para $\zeta = \xi''/\xi' = d\ln\xi'/dx$.

En conclusión, la transformación

$$\bar{x} = \int_{x_0}^{x} dt \left(\exp \int_{t_0}^{t} ds\zeta(s) \right), \quad \bar{y} = C \left(\exp \int_{x_0}^{x} dt\zeta(t) \right) y$$

donde ζ es una solución de la ecuación de Riccati (24), reduce la ecuación lineal homogénea de orden n en forma semicanónica, a la forma canónica de Forsyth-Laguerre .

4 Propiedades de la ecuación de Riccati

4.1 Relación con la ecuación lineal de segundo orden

Si en la ecuación de Riccati simplificada

$$y' + r(x)y^2 = p(x),$$

hacemos el cambio de función

$$y = \frac{1}{ru}\frac{du}{dx},$$

la ecuación resultante es lineal homogénea de segundo orden:

$$\frac{d}{dx}\left(\frac{1}{r(x)}\frac{du}{dx}\right) = p(x)u(x).$$

Recíprocamente, si en la ecuación general lineal homogénea de segundo orden,

$$\frac{d}{dx}\left(a(x)\frac{du}{dx}\right) = c(x)u(x), \tag{25}$$

hacemos el cambio de función

$$\frac{du}{dx} = (ry)u,$$

resulta la siguiente ecuacion de Riccati simplificada

$$\frac{dy}{dx} + \frac{1}{a(x)}y^2 = c(x).$$

Luego a toda ecuación de Riccati le corresponde una ecuación lineal homogénea de segundo orden y, recíprocamente, a toda ecuación lineal homogénea de segundo orden le corresponde una ecuación de Riccati. El paso de una a otra se realiza mediante el cambio de función

$$y = \frac{u'}{ru}.$$

Como acabamos de ver, esta relación tiene una forma especialmente simétrica si la ecuación de Riccati está en forma simplificada y la ecuación de segundo orden se escribe en forma autoadjunta.

Corolario 2 *Si* u_1, u_2 *son soluciones linealmente independientes de la ecuación lineal homogénea de segundo orden*

$$\frac{d}{dx}\left(\frac{1}{r(x)}\frac{du}{dx}\right) = p(x)u$$

y C_1, C_2 *son constantes arbitrarias* (*no ambas cero*), *entonces la función*

$$y(x) = \frac{1}{r(x)}\frac{C_1 u_1'(x) + C_2 u_2'(x)}{C_1 u_1(x) + C_2 u_2(x)}$$

es solución de la ecuación de Riccati

$$\frac{dy}{dx} + r(x)y^2 = p(x).$$

El resolver un problema resolviendo, en su lugar, otro problema 'asociado' más accesible, es una técnica muy generalizada en matemáticas.

Ejemplo 1 (La ecuación del paracaidista II) *La ecuación lineal de segundo orden asociada a la ecuación del paracaidista*

$$\frac{dv}{dt} + v^2 = 1$$

es

$$\frac{d^2u}{dt^2} - u = 0,$$

con

$$v = \frac{u'}{u}.$$

Luego la solución general es

$$u(t) = C_1 e^t + C_2 e^{-t} \quad (C_1, C_2 \in \mathbb{R}),$$

de donde, aplicando el Corolario 2,

$$v(t) = \frac{C_1 e^t - C_2 e^{-t}}{C_1 e^t + C_2 e^{-t}} = \frac{e^t - Ke^{-t}}{e^t + Ke^{-t}},$$

con $K = C_2/C_1$. □

4.2 El teorema de la razón doble

Por la *razón doble* de los números x_1, x_2, x_3 y x_4 entendemos el cociente

$$\frac{(x_1 - x_3)(x_2 - x_4)}{(x_1 - x_4)(x_2 - x_3)}.$$

Sabemos (véase (1)) que la solución general de la ecuación de Riccati es de la forma

$$y(x) = \frac{g_1(x) + Kg_2(x)}{g_3(x) + Kg_4(x)}, \tag{26}$$

donde $g_i(x), 1 \leq i \leq 4$, son funciones diferenciables en cierto intervalo abierto y K es una constante arbitraria. Consideremos ahora cuatro soluciones particulares linealmente independientes y_1, y_2, y_3 e y_4, obtenidas de (26) dando a K los valores K_1, K_2, K_3 y K_4, respectivamente. Calculando se obtiene entonces

$$y_i - y_j = \frac{(K_i - K_j)(g_2 g_3 - g_1 g_4)}{(g_3 - K_i g_4)(g_3 - K_j g_4)} \qquad (1 \leq i, j \leq 4)$$

de donde

$$\frac{(y_1 - y_3)(y_2 - y_4)}{(y_1 - y_4)(y_2 - y_3)} = \frac{(K_1 - K_3)(K_2 - K_4)}{(K_1 - K_4)(K_2 - K_3)}.$$

Luego: *la razón doble de cualesquiera cuatro soluciones linealmente independientes de la ecuación de Riccati es constante.* O en el lenguaje de la geometría proyectiva: *las series que resultan de cortar el haz integral por dos rectas paralelas al eje Y son proyectivas.*

4.3 Singularidades de la ecuación de Riccati

Si comparamos la solución general de la ecuación de Riccati

$$y' + y^2 = 0$$

con la de la ecuación de Abel

$$y' + \frac{1}{2}y^3 = 0,$$

a saber,

$$y(x) = \frac{1}{x + C} \qquad \text{e} \qquad y(x) = \frac{1}{\sqrt{x + C}}$$

respectivamente, observamos que la primera tiene un *polo móvil* (es decir, que la posición de la singularidad depende de la constante de integración), mientras que la segunda tiene un *punto de ramificación móvil.* Recordemos que, en teoría de ecuaciones diferenciales ordinarias, se llaman *puntos críticos* de una solución a los puntos de ramificación y a las singularidades esenciales (o polos infinitos). Precisamente, el criterio de propuesto por Painlevé y Gambier para clasificar las ecuaciones no lineales de segundo orden fue el carácter de los puntos singulares de sus soluciones. Sus investigaciones concluyeron con el descubrimiento de 50 tipos canónicos de ecuaciones cuyas soluciones tenían puntos críticos fijos, es decir, cuyas soluciones sólo podían tener polos como singularidades móviles.

Teorema 1 *Sean $P(x, y)$, $Q(x, y)$ funciones analíticas en $x, y \in \mathbb{C}$ y polinomios en y. Si la solución general de la ecuación*

$$\frac{dy}{dx} = \frac{P(x, y)}{Q(x, y)} \tag{27}$$

no tiene puntos de ramificación móviles, entonces la ecuación es necesariamente de Riccati.

Obsérvese, por otra parte, que las hipótesis sobre el campo vectorial de la ecuación (27) son poco restrictivas en la práctica y que tales ecuaciones surgen de manera natural al determinar los retratos de fase de sistemas dinámicos autónomos en el plano.

Prueba. Si (x_0, y_0) es un punto del plano tal que

$$Q(x_0, y_0) = 0, \quad P(x_0, y_0) \neq 0 \tag{28}$$

entonces, (x_0, y_0) es una singularidad de la ecuación (27). Por otra parte, este mismo punto es un punto regular de la ecuación diferencial

$$\frac{dx}{dy} = \frac{Q(x,y)}{P(x,y)} \tag{29}$$

en la que ahora x es la variable dependiente. Por el teorema de Cauchy, la ecuación (29) tiene localmente una única solución $x = x(y)$ tal que $x(y_0) = x_0$ y que, además, es analítica en el punto y_0, con lo que podemos representar $x(y)$ mediante una serie de potencias,

$$x(y) = x_0 + \sum_{n=1}^{\infty} c_n (y - y_0)^n \quad (c_n \in \mathbb{C}),$$

en un entorno de y_0. Además, $c_1 = 0$ ya que $x'(y_0) = Q(x_0, y_0)/P(x_0, y_0) = 0$, de manera que

$$x - x_0 = c_k (y - y_0)^k \ + c_{k+1} (y - y_0)^{k+1} + ...,$$

donde $c_k \neq 0$, $k \geq 2$. Invirtiendo ahora esta serie de potencias, podemos expresar $y - y_0$ como una serie de potencias de $(x - x_0)^{1/k}$,

$$y - y_0 = d_1 (x - x_0)^{1/k} + d_2 (x - x_0)^{2/k} + ...$$

en un entorno de x_0, con lo que x_0 es un punto de ramificación. Además, como x_0 es arbitrario, sujeto únicamente a la existencia de un y_0 tal que la condición (28) se verifique, concluimos que x_0 es un punto de ramificación móvil.

De las hipótesis del teorema y del argumento anterior deducimos que, si la solución de la ecuación (27) ha de estar libre de puntos de ramificación móviles, la única posibilidad es que $Q(x,y)$ dependa sólo de x (de otro modo, siempre podemos encontrar x_0 e y_0 tales que verifiquen (28)). Pero esto significa que la ecuación (27) ha de tener la forma

$$\frac{dy}{dx} = P_0(x) + P_1(x) y + P_2(x) y^2 + ... + P_m(x) y^m.$$

Falta probar que $m \leq 2$. Haciendo el cambio $y = 1/z$, obtenemos

$$\frac{dz}{dx} = -z^2 \left(P_0(x) + \frac{P_1(x)}{z} + \frac{P_2(x)}{z^2} + ... + \frac{P_m(x)}{z^m} \right).$$

Ahora bien, si $m > 2$ estaríamos en la misma situación que antes (el campo de velocidades sería un cociente de funciones analíticas en x y z y, en particular, el denominador sería un polinomio en z) y, como acabamos de ver, sus soluciones tendrían puntos de ramificación móviles y, por ende, también las tendría $y = 1/z$. La única manera de evitar esta posibilidad es que $m \leq 2$. □

5 Integración de la ecuación de Riccati

Como dijimos en la Introducción, Liouville probó que, en general, la ecuación de Riccati no se puede resolver mediante cuadraturas.

A. Sea, pues, y_1 una solución particular de la ecuación de Riccati

$$y' + Qy + Ry^2 = P.$$

Entonces, haciendo el cambio de función

$$y = y_1 + \frac{1}{u},$$

la ecuación de Riccati se transforma en la ecuación lineal

$$\frac{du}{dx} - (2Ry_1 + Q)u = R, \tag{30}$$

la cual se resuelve mediante dos cuadraturas: una para la solución general de la ecuación homogénea asociada y otra, para una solución particular de la ecuación completa.

B. Si conocemos dos soluciones particulares y_1 e y_2, la solución general de la ecuación de Riccati puede obtenerse mediante una sola cuadratura. En efecto, haciendo los cambios de función

$$y = y_1 + \frac{1}{u}, \quad y = y_2 + \frac{1}{v},$$

obtenemos, análogamente al caso anterior,

$$\frac{du}{dx} - (2Ry_1 + Q)u = R, \quad \frac{dv}{dx} - (2Ry_2 + Q)v = R,$$

respectivamente. A continuación, de

$$\left.\begin{array}{l} u'v - (2Ry_1 + Q)uv = Rv \\ uv' - (2Ry_2 + Q)uv = Ru \end{array}\right\} \Rightarrow \frac{u'v - uv'}{v^2} - 2R(y_1 - y_2)\frac{u}{v} + R\frac{u - v}{v^2} = 0$$

y

$$\left.\begin{array}{l} 1/u = y - y_1 \\ 1/v = y - y_2 \end{array}\right\} \Rightarrow \frac{u - v}{v^2} = (y_1 - y_2)\frac{u}{v},$$

deducimos la ecuación

$$\frac{d}{dx}\left(\frac{u}{v}\right) - R(y_1 - y_2)\frac{u}{v} = 0, \tag{31}$$

la cual es de variables separables y, por tanto, resoluble mediante una sola cuadratura.

C. Finalmente, si se conocen tres soluciones particulares y_1, y_2 e y_3, la solución general de la ecuación de Riccati puede obtenerse sin ninguna cuadratura mediante el teorema de la razón doble:

$$\frac{(y-y_2)(y_1-y_3)}{(y-y_3)(y_1-y_2)} = K.$$

Ejemplo 2 (La ecuación del paracaidista III) *Es evidente que la ecuación*

$$\frac{dv}{dt} + v^2 = 1$$

tiene la solución particular

$$v(t) = 1.$$

Por tanto, podemos hallar la solución general mediante el cambio de función

$$v(t) = 1 + \frac{1}{u(t)},$$

que transforma la ecuación del paracaidista en la ecuación lineal

$$\frac{du}{dt} - 2u = 1,$$

cuya solución general es

$$u(t) = Ce^{2t} - \frac{1}{2} \quad (C \in \mathbb{R}).$$

Invirtiendo el cambio $v(t) \mapsto u(t)$, obtenemos

$$v(t) = \frac{2Ce^{2t}+1}{2Ce^{2t}-1} = \frac{e^t - Ke^{-t}}{e^t + Ke^{-t}},$$

con $K := -1/(2C)$. □

6 Integración de la ecuación original de Riccati I

Hay varias técnicas para integrar la ecuación original de Riccati. En esta sección explicaremos el que acaso sea el método más elegante y que permite expresar las soluciones mediante funciones de Bessel, dejando para la próxima sección otro basado en fracciones continuas.

Recordemos que la *funciones de Bessel de primera especie*,

$$J_\nu(x) = \sum_{j=0}^{\infty} \frac{(-1)^j}{j!\Gamma(\nu+j+1)} \left(\frac{x}{2}\right)^{\nu+2j} \qquad (\nu \in \mathbb{R})$$

(la serie converge para todo $x \in \mathbb{R}$ o $\mathbb{C}$) y de *segunda especie*,

$$Y_\nu(x) = \frac{J_\nu(x)\cos\nu\pi - J_{-\nu}(x)}{\operatorname{sen}\nu\pi}$$

(donde, si $\nu \in \mathbb{Z}$, se toma el límite del miembro derecho), son soluciones linealmente independientes de la *ecuación de Bessel* de orden $\nu \in \mathbb{R}$,

$$L(u;\nu) \equiv x^2\frac{d^2u}{dx^2} + x\frac{du}{dx} + (x^2-\nu^2)u = 0.$$

Los cambios de variables independiente y dependiente,

$$t = \left(\frac{x}{2}\right)^2, \quad u(t) = t^{\nu/2}w(t),$$

respectivamente, transforman la ecuación de Bessel en la ecuación

$$M(w;\nu) \equiv t\frac{d^2w}{dt^2} + (1+\nu)\frac{dw}{dt} + w = 0.$$

Las soluciones correspondientes son, pues:

$$\begin{aligned} L(u;\nu) = 0: &\quad u(x) = C_1J_\nu(x) + C_2Y_\nu(x), \\ M(w;\nu) = 0: &\quad w(t) = t^{-\nu/2}\left(C_1J_\nu(2\sqrt{t}) + C_2Y_\nu(2\sqrt{t})\right). \end{aligned}$$

Si $\nu \notin \mathbb{Z}$, entonces J_ν y $J_{-\nu}$ son linealmente independientes, de manera que, en este caso, pueden utilizarse de forma alternativa estas dos funciones de primera especie como sistema fundamental de soluciones.

Por otro lado, los cambios de función $y \mapsto w$ y de variable independiente $x \mapsto s$,

$$y(x) = \frac{u(x)}{x}, \quad s = -\frac{AB}{(n+2)^2}x^{n+2}, \quad u(s) = \frac{(n+2)s}{A}\frac{w'(s)}{w(s)},$$

transforman la ecuación original de Riccati,

$$y' + Ay^2 = Bx^n \quad (n \neq -2)$$

en la ecuación de Bessel de orden $\nu = -1/(n+2)$,

$$M\left(w; -\frac{1}{n+2}\right) \equiv s\frac{d^2w}{ds^2} + \left(1 - \frac{1}{n+2}\right)\frac{dw}{ds} + w = 0,$$

de donde deducimos que

$$w(s) = s^{\frac{1}{2(n+2)}}\left(C_1J_{\frac{1}{n+2}}(2\sqrt{s}) + C_2Y_{\frac{1}{n+2}}(2\sqrt{s})\right). \tag{32}$$

Invirtiendo los cambios de variable realizados, obtenemos la solución general de la ecuación original de Riccati para $n \neq -2$.

Recordemos, finalmente, que J_ν y $J_{-\nu}$ (y con ellas, Y_ν) son funciones elementales cuando $\nu \in \mathbb{Z}+1/2$. Luego la solución $w(s)$ (y con ella $y(x)$) es una función elemental cuando

$$\frac{1}{n+2} = k + \frac{1}{2} \quad (k \in \mathbb{Z})$$

El caso $n = -2$, que está todavía por dilucidar, tiene también solución elemental. En efecto: el cambio de función

$$y(x) = \frac{u(x)}{x},$$

transforma la ecuación

$$y' + Ay^2 = Bx^{-2}$$

en la ecuación de variables separables

$$x\frac{du}{dx} - u + Au^2 = B,$$

cuya solución general es un haz de funciones elementales.

Corolario 3 *La ecuación original de Riccati*

$$y' + Ay^2 = Bx^n$$

es resoluble mediante funciones elementales si y sólo si

1. $n = -2$.

2. $n = -\dfrac{4k}{(2k+1)} = 0, -4, -\dfrac{4}{3}, -\dfrac{8}{3}, -\dfrac{8}{5}, -\dfrac{12}{5}, -\dfrac{12}{7}, -\dfrac{18}{7}, \ldots$

Nota 2 *De la relación de recurrencia para las funciones de Bessel de primera especie*

$$J_{\nu+1}(x) = \frac{2\nu}{x}J_\nu(x) - J_{\nu-1}(x),$$

se deduce que

$$\begin{aligned} \frac{J_{\nu-1}(x)}{J_\nu(x)} &= \frac{2\nu}{x} - \frac{1}{J_\nu(x)/J_{\nu+1}(x)} \\ &= \frac{2\nu}{x} - \cfrac{1}{\frac{2\nu+2}{x} - \cfrac{1}{\frac{2\nu+4}{x} - \ldots}} \end{aligned}$$

En particular, para $\nu = 1/2$,

$$J_{1/2}(x) = \sqrt{\frac{2}{\pi x}}\,\mathrm{sen}\, x \quad y \quad J_{-1/2}(x) = \sqrt{\frac{2}{\pi x}}\cos x,$$

de donde

$$\begin{aligned}\operatorname{cotg} x &= \frac{J_{-1/2}(x)}{J_{1/2}(x)} = \frac{1}{x} - \frac{1}{\frac{3}{x}-}\frac{1}{\frac{5}{x}-}\frac{1}{\frac{7}{x}-\ldots}\\ &= \frac{1}{x} - \frac{x}{3-}\frac{x^2}{5-}\frac{x^2}{7-\ldots}\end{aligned}$$

y

$$\begin{aligned}\operatorname{tg} x &= \frac{1}{\operatorname{cotg} x} = \frac{1}{\frac{1}{x}-}\frac{1}{\frac{3}{x}-}\frac{1}{\frac{5}{x}-\ldots}\\ &= \frac{x}{1-}\frac{x^2}{3-}\frac{x^2}{5-\ldots} \quad (x \neq \frac{\pi}{2} \pm n\pi) \qquad (33)\end{aligned}$$

Sustituyendo x por ix obtenemos las siguientes representaciones de $\coth x$ *y* $\tanh x$ *en fracciones continuas:*

$$\begin{aligned}\coth x &= i \operatorname{cotg} ix = \frac{1}{x} + \frac{x}{3+}\frac{x^2}{5+}\frac{x^2}{7+\ldots}\\ \tanh x &= \frac{1}{i} \operatorname{tg} ix = \frac{x}{1+}\frac{x^2}{3+}\frac{x^2}{5+\ldots}\end{aligned}$$

Estas representaciones, llamadas a veces de Lambert, serán utilizadas en la siguiente sección.

Recordemos de paso que las fracciones continuas de $\operatorname{tg} x$ y $\tanh x$ fueron utilizadas por J. Lambert en 1761 para demostrar, por primera vez, la irracionalidad de π y e, respectivamente.

Ejemplo 3 (La ecuación del paracaidista IV) *La ecuación adimensionalizada del paracaidista es del tipo*

$$y' + Ay^2 = Bx^n$$

con

$$A = 1, \quad B = 1, \quad n = 0,$$

de manera que, como ya sabemos, tiene solución elemental. Haciendo los cambios de variable dependiente e independiente explicados más arriba,

$$y(t) = \frac{u(t)}{t}, \quad s = -\frac{t^2}{4}, \quad u(s) = 2s\frac{w'(s)}{w(s)}, \qquad (34)$$

deducimos de (32)

$$\begin{aligned}w(s) &= s^{1/4}\left(C_1 J_{1/2}(2\sqrt{s}) + C_2 J_{-1/2}(2\sqrt{s})\right)\\ &= \sqrt{\frac{1}{\pi}}\left(C_1 \operatorname{sen}(2\sqrt{s}) + C_2 \cos(2\sqrt{s})\right),\end{aligned}$$

y de (34)

$$u(s) = \frac{2\sqrt{s}\,(C_1 \cos(2\sqrt{s}) - C_2 \operatorname{sen}(2\sqrt{s}))}{C_1 \operatorname{sen}(2\sqrt{s}) + C_2 \cos(2\sqrt{s})},$$

donde (véase (34)) $2\sqrt{s} = \pm it$. *Luego*

$$2\sqrt{s}\cos(2\sqrt{s}) = \pm it\cosh t, \quad 2\sqrt{s}\,\mathrm{sen}(2\sqrt{s}) = -t\,\mathrm{senh}\,t$$

y, finalmente,

$$y(t) = \frac{u(t)}{t} = \frac{\pm iC_1\cosh t + C_2\,\mathrm{senh}\,t}{\pm iC_1\,\mathrm{senh}\,t + C_2\cosh t} = \frac{e^t - Ke^{-t}}{e^t + Ke^{-t}},$$

donde $K = (C_2 \mp iC_1)/(C_2 \pm iC_1)$. □

7 Integración de la ecuación original de Riccati II

Otra técnica ingeniosa de probar el Corolario 3 y obtener, de paso, soluciones particulares de la ecuación original de Riccati consiste en desarrollar las soluciones en fracciones continuas. Veamos cómo se aplica esta técnica.

El cambio de función

$$y(x) = \frac{u(x)}{x}$$

en la ecuación original de Riccati

$$y' + Ay^2 = Bx^n \quad (A, B, n \in \mathbb{R})$$

la transforma en otra ecuación de Riccati de la forma

$$x\frac{du}{dx} - u + Au^2 = Bx^p \tag{35}$$

con $p = n + 2$. Si $p = 0$, esta ecuación es de variables separables. En el caso general $p \neq 0$, podemos tomar la ecuación (35) como punto de partida para una serie de cambios de función. La siguiente tabla muestra los tres primeros cambios junto con las ecuaciones transformadas correspondientes:

$u(x) = \dfrac{1}{A} + \dfrac{x^p}{u_1(x)}$	$x\dfrac{du_1}{dx} - (1+p)u_1 + Bu_1^2 = Ax^p$
$u_1(x) = \dfrac{1+p}{B} + \dfrac{x^p}{u_2(x)}$	$x\dfrac{du_2}{dx} - (1+2p)u_2 + Au_2^2 = Bx^p$
$u_2(x) = \dfrac{1+2p}{A} + \dfrac{x^p}{u_3(x)}$	$x\dfrac{du_3}{dx} - (1+3p)u_3 + Bu_3^2 = Ax^p$

Continuando con esta serie de transformaciones, después del k-ésimo cambio de función llegaríamos a la ecuación

$$x\frac{du_k}{dx} - (1+kp)u_k + A_k u_k^2 = B_k x^p \tag{36}$$

con $k \geq 1$ y

$$\begin{cases} A_k = A,\ B_k = B & \text{si } k \text{ es par,} \\ A_k = B,\ B_k = A & \text{si } k \text{ es impar.} \end{cases}$$

Componiendo las transformaciones realizadas, vemos que la fracción continua

$$\begin{aligned} xy(x) &= u(x) = \frac{1}{A} + \frac{x^p}{u_1(x)} = \frac{1}{A} + \frac{x^p}{\frac{1+p}{B} + \frac{x^p}{u_2(x)}} = \frac{1}{A} + \frac{x^p}{\frac{1+p}{B} + \frac{x^p}{\frac{1+2p}{A} + \frac{x^p}{u_3(x)} \cdots}} \\ &= \frac{1}{A} + \frac{x^p}{\frac{1+p}{B} +} \frac{x^p}{\frac{1+2p}{A} +} \frac{x^p}{\frac{1+3p}{B} + \ldots} \qquad (37) \end{aligned}$$

nos da una solución particular de la ecuación original de Riccati.

Por otro lado, si p es de la forma

$$p = -\frac{2}{2k-1} \qquad (k \in \mathbb{N}),$$

entonces

$$1 + kp = -\frac{1}{2k-1} = -\frac{1}{2}p,$$

con lo que la ecuación (36), que es la ecuación transformada de la (4) tras k cambios de función, quedaría en la forma

$$x\frac{du_k}{dx} - \frac{1}{2}pu_k + A_k u_k^2 = B_k x^p. \qquad (38)$$

El nuevo cambio $u_k(x) = x^{p/2}v(x)$ la transforma ahora en la ecuación

$$x^{1-p/2}\frac{dv}{dx} + A_k v^2 = B_k,$$

que es de variables separables. La solución, que se obtiene mediante una cuadratura, es una función elemental, en consonancia con lo que vimos en la sección anterior ya que

$$\begin{aligned} n &= p - 2 = -\frac{4k}{2k-1} \qquad (k = 1, 2, \ldots) \\ &= -4, -\frac{8}{3}, -\frac{12}{5}, \ldots \end{aligned}$$

¿Dónde están las soluciones elementales de la ecuación original de Riccati que, según el Corolario 3, aparecen cuando

$$n = 0, -\frac{4}{3}, -\frac{8}{5}, \ldots$$

se cumple? Consideremos de nuevo la ecuación de partida,

$$x\frac{du}{dx} - u + Au^2 = Bx^p,$$

que, recordemos, se obtiene de la ecuación original de Riccati tras el cambio $y(x) = u(x)/x$, y hagamos esta vez la sucesión de cambios de función en las ecuaciones resultantes, que a continuación se muestran:

$u(x) = \dfrac{x^p}{u_1(x)}$	$x\dfrac{du_1}{dx} - (p-1)u_1 + Bu_1^2 = Ax^p$
$u_1(x) = \dfrac{p-1}{B} + \dfrac{x^p}{u_2(x)}$	$x\dfrac{du_2}{dx} - (2p-1)u_2 + Au_2^2 = Bx^p$
$u_2(x) = \dfrac{2p-1}{A} + \dfrac{x^p}{u_3(x)}$	$x\dfrac{du_3}{dx} - (3p-1)u_3 + Bu_3^2 = Ax^p$

etc. Después de $j \geq 1$ transformaciones, llegaríamos a la ecuación

$$x\frac{du_j}{dx} - (jp-1)u_j + A_j u_j^2 = B_j x^p, \tag{39}$$

donde, igual que antes,

$$\begin{cases} A_j = A,\ B_j = B & \text{si } j \text{ es par,} \\ A_j = B,\ B_j = A & \text{si } j \text{ es impar.} \end{cases}$$

De la sucesión de cambios de función, deducimos ahora que la ecuación original de Riccati tiene una solución particular con representación en fracción continua

$$xy(x) = u(x) = \frac{x^p}{\frac{p-1}{B}+}\,\frac{x^p}{\frac{2p-1}{A}+}\,\frac{x^p}{\frac{3p-1}{B}+...} \tag{40}$$

Si, análogamente al caso previo, ocurre que $p = 2/(2j-1)$, de manera que

$$jp - 1 = \frac{1}{2j-1} = \frac{1}{2}p$$

en (39), entonces, tras j transformaciones, la ecuación de partida (39) quedaría en la forma

$$x\frac{du_j}{dx} - \frac{1}{2}pu_j + A_j u_j^2 = B_j x^p.$$

Esta ecuación, que es idéntica a la (38), se puede integrar con una cuadratura, tal y como se explicó más arriba. Así, pues, las ecuaciones originales de Riccati con

$$\begin{aligned} n &= p - 2 = -\frac{4(j-1)}{2j-1} = -\frac{4k}{2k+1} \quad (k = j-1 = 0, 1, 2, ...) \\ &= 0, -\frac{4}{3}, -\frac{8}{5}, ... \end{aligned}$$

tienen asimismo soluciones elementales.

Ejemplo 4 (La ecuación del paracaidista V) *La fórmula* (37)*, con* $A = B = 1$*,* $n = 0$*,* $p = n + 2 = 2$*, da como solución particular de la ecuación*

$$\frac{dy}{dt} + y^2 = 1$$

la función

$$y_1(t) = \frac{1}{t} + \frac{t}{3+}\frac{t^2}{5+}\frac{t^2}{7+...} = \coth t$$

y la fórmula (40), *la función*

$$y_2(t) = \frac{t}{1+}\frac{t^2}{3+}\frac{t^2}{5+...} = \tanh t$$

(*véase Nota* 2). *Como tenemos dos soluciones particulares, podemos encontrar la solución general* $y(t)$ *mediante una sola cuadratura. Recordemos cómo se hace* (*Sección* 5, *apartado* B)*:*

$$\left.\begin{array}{l} y - y_1 = 1/u \\ y - y_2 = 1/v \end{array}\right\} \Rightarrow \frac{y-y_2}{y-y_1} = \frac{1/v}{1/u} = \frac{u}{v} \tag{41}$$

donde la función u/v *satisface la ecuación* (31),

$$\begin{aligned} 0 &= \frac{d}{dt}\left(\frac{u}{v}\right) - R(y_1 - y_2)\frac{u}{v} = \frac{d}{dt}\left(\frac{u}{v}\right) - (\coth t - \tanh t)\frac{u}{v} \\ &= \frac{d}{dt}\left(\frac{u}{v}\right) - \frac{2}{\operatorname{senh} 2t}\frac{u}{v}. \end{aligned}$$

Separando variables,

$$\frac{u(t)}{v(t)} = C\exp\left(\int \frac{2dt}{\operatorname{senh} 2t}\right) = C\tanh t$$

y ustituyendo en (41), *obtenemos:*

$$\frac{y - \tanh t}{y - \coth t} = C\tanh t \;\Rightarrow\; y(t) = \frac{\tanh t - C}{1 - C\tanh t}$$

Luego, la solución buscada es

$$y(t) = \frac{\operatorname{senh} t - C\cosh t}{\cosh t - C\operatorname{senh} t} = \frac{(1-C)e^t - (1+C)e^{-t}}{(1-C)e^t + (1+C)e^{-t}} = \frac{e^t - Ke^{-t}}{e^t + Ke^{-t}}$$

con $K := (1+C)/(1-C)$. □

Ejercicio de repaso. (*La ecuación del paracaidista VI*) Sabiendo que la ecuación del paracaidista (5) tiene las tres soluciones particulares

$$v_1(t) = v_\infty,\; v_2(t) = v_\infty \tanh \gamma t,\;\; v_3(t) = v_\infty \coth \gamma t,$$

con los parámetros v_∞ y γ definidos en (7), hallar el haz integral sin realizar ninguna cuadratura.

Referencias

[1] L.T. Eisenhart. *A Treatise on the Differential Geometry of Curves and Surfaces.* Dover, New York 1960.

[2] P. Puig Adam. *Ecuaciones Diferenciales.* Biblioteca Matemática, Madrid 1974.

[3] W.T. Reid. *Riccati Differential Equations.* Academic Press, New York 1972.

[4] G.F. Simmons. *Ecuaciones Diferenciales* (2^a edición). MacGraw-Hill, Madrid 1993.

[5] E.J. Wilczynski. *Projective Differential Geometry of Curves and Ruled Surfaces.* Teubner, Leipzig 1906.

Título:	Análisis Numérico de algunos Modelos Diferenciales acoplados de la Mecánica de Fluidos.
Doctorando:	Juan Vicente Gutiérrez-Santacreu.
Director/es:	Francisco Manuel Guillén González.
Defensa:	9 de enero de 2008, Sevilla.
Calificación:	Sobresaliente cum laude por unanimidad.

Resumen:

La tesis está dedicada al análisis numérico de varios sistemas de ecuaciones en derivadas parciales de la Mecánica de Fluidos; en concreto, las ecuaciones de Navier-Stokes con difusión de masa y densidad variable, modelos de cristales líquidos nemáticos y sistemas de solidificación de aleaciones binarias que incorporan campos de fase. Las características comunes a estos modelos son: i) verificación de un principio del máximo de una de las variables de estado (la densidad de masa en el caso de los modelos de difusión de masa, el vector de orientación de las macromoléculas de cristales nemáticos y la concentración de los compuestos de la aleación para los modelos de campo de fases), ii) distinta regularidad de las incógnitas del sistema, iii) influencia de un pequeño parámetro (coeficiente de difusión de masa, penalización sobre la restricción unitaria del vector de orientación o sobre los dos valores que estabilizan las fases).

En el primer capítulo, para la versión bidimensional del modelo de difusión de masa simplificado, se propone un esquema numérico lineal y desacoplado en densidad y velocidad-presión basado en elementos finitos globalmente continuos para la densidad y un par de elementos finitos estables para la velocidad-presión. Se demuestra estabilidad incondicional y convergencia hacia soluciones débiles. Las herramientas esenciales son: un operador de truncamiento discreto para la densidad y una desigualdad discreta de tipo Gagliardo-Nirenberg que permiten eludir el uso de estimaciones puntuales para la densidad (que no está claro como conseguir al usar elementos finitos sólo continuos). Esta desigualdad se obtiene usando una variable auxiliar que representa el laplaciano discreto de la densidad.

En el capítulo segundo, se plantean los mismos objetivos para la versión tridimensional del problema de difusión de masa simplificado. La técnica empleada en el caso anterior es exclusivamente bidimensional, y por tanto, son necesarias nuevas ideas tanto para el diseño del esquema numérico como para el estudio de su estabilidad y convergencia. Ahora no se proyecta la densidad para que verifique cotas puntuales pero a cambio hay que considerar una velocidad de convección en la ecuación de la densidad en un espacio de

divergencia discreta nula ”más rico”. Se precisan ahora restricciones sobre los pasos de discretización (temporal y espacial) para conseguir una versión discreta del principio del máximo para la densidad. Esto permite obtener estimaciones en normas fuertes para la densidad que conducen a la convergencia del esquema. Además, se estudia el comportamiento asintótico de la solución aproximada cuando el parámetro de difusión de masa λ tiende a cero, probando que, bajo ciertas restricciones que relacionan los parámetros de discretización y el coeficiente de difusión de masa, el esquema converge hacia una solución débil del problema de Navier-Stokes con densidad variable.

En el tercer capítulo, se amplía el estudio numérico al modelo de difusión de masa completo (considerando ahora términos en λ^2 antes despreciados). Haciendo uso de las técnicas desarrolladas para el modelo simplificado y de un principio de inducción en las etapas de tiempo, bajo adecuadas restricciones sobre los parámetros se establecen resultados de convergencia y estabilidad condicional en el sentido del capítulo segundo.

En el capítulo cuarto, se diseña un esquema numérico para el modelo de difusión de masa simplificado, para el que se obtiene estimaciones de error (en tiempo y espacio), respecto de una solución regular. Es importante destacar, que al contrario que para el modelo de Navier-Stokes, el orden de convergencia obtenido no es optimal, a menos que se imponga regularidad a la solución que conlleve las conocidas condiciones de compatibilidad para la presión en el tiempo inicial.

En el quinto capítulo, abordamos la construcción de un esquema numérico para el modelo penalizado de tipo Ginzburg-Landau de cristales líquidos nemáticos. Se consigue un esquema estable, totalmente acoplado entre el vector orientación y el par velocidad-presión, y convergente para dominios tridimensionales, basado en la introducción de una variable auxiliar para aproximar el laplaciano del vector de orientación, que deberá aproximarse en un espacio de elementos finitos compatible con los espacios de aproximación de la velocidad y el vector de orientación. Además, realizamos experiencias numéricas relativas a este esquema.

Finalmente, en el capítulo sexto, para un modelo de campo de fases para un proceso de solidificación de una mezcla binaria con propiedades térmicas, se presentan dos esquemas que desacoplan el cálculo de las 3 incógnitas del problema: campo de fase, temperatura y concentración relativa (de un material respecto al otro). El primero de ellos realiza una discretización no lineal de la ecuación de campo de fase y lineal para la temperatura y concentración, que resulta incondicionalmente estable y convergente hacia soluciones débiles. El segundo esquema utiliza también una discretización lineal para la ecuación de campo de fase, pero la estabilidad (y convergencia) se consigue a costa de imponer una restricción sobre los parámetros de discretización. Para ambos esquemas es preciso utilizar un operador de truncamiento discreto para garantizar estimaciones puntuales para la concentración.

Título:	Diseño óptimo aerodinámico a través del método adjunto continuo.
Doctorando:	Francisco de Asís Palacios Gutiérrez.
Director/es:	Prof. Enrique Zuazua Iriondo.
Defensa:	30 de junio de 2008, Madrid.
Calificación:	Sobresaliente Cum Laude.

Resumen:

El diseño de forma aerodinámica es una de las disciplinas de la aeronáutica de mayor tradición y que en la actualidad, gracias al avance de la computación y los métodos numéricos, está adquiriendo una gran relevancia en el sector industrial y en los centros de investigación. Como es lógico, este interés ha venido precedido por la exactitud de cálculo que empieza a lograrse mediante el empleo de técnicas numéricas para simular el comportamiento de los fluidos mediante computación.

El objetivo del diseño de forma aerodinámica es minimizar un determinado funcional de interés aeronáutico, a través de controlar el sistema de EDPs que modeliza el comportamiento del fluido sobre la aeronave. En este caso, el control se ejerce sobre la forma de la superficie a diseño con el objeto de mejorar las prestaciones de la aeronave.

Existe una importante tradición en la aplicación de la teoría de control al diseño óptimo de forma en sistemas gobernados por EDPs, desde los trabajos iniciales de J.-L. Lions, pasando por los primeros desarrollos en el ámbito de la mecánica de los fluidos de O. Pironneau y terminando en A. Jameson quien en una serie de artículos, trató la aplicación de estas técnicas en sistemas gobernados por las ecuaciones de Euler y Navier-Stokes. Mención aparte merecen los trabajos de M. Giles y S. Ulbrich en el ámbito de la optimización con discontinuidades en las variables de flujo.

A lo largo de esta tesis se plantean y resuelven cuestiones relevantes en el ámbito de los problemas de diseño óptimo de forma en sistemas gobernados por las ecuación de Burgers y las ecuaciones de Euler y Navier-Stokes. Tras exponer una panorámica general de los métodos de diseño óptimo en aerodinámica, en esta tesis se aborda un análisis riguroso del control en problemas de diseño inverso gobernados por la ecuación de Burgers, en este ámbito se introduce un método original denominado "método de las direcciones de descenso alternantes". Más adelante, se aborda el problema del diseño óptimo de forma en sistemas gobernados por las ecuaciones de Euler y Navier Stokes mediante la formulación denominada "adjunta continua", en este contexto se realizan aportaciones originales de tipo metodológico y computacional entre las que destacan un análisis riguroso de la influencia de las relaciones de Rankine-Hugoniot en el estado adjunto y la consecución de un proceso de optimización aerodinámica mediante la técnica de "conjuntos de nivel". Finalmente se presentan resultados numéricos en configuraciones aeronáuticas de interés industrial. Por último, destacar que los resultados más relevantes de esta tesis han sido publicados en las revistas M3AS y AIAA Journal.

Título:	ANÁLISIS NUMÉRICO DE ESQUEMAS FRACCIONADOS EN TIEMPO PARA NAVIER-STOKES $3D$ Y ECUACIONES PRIMITIVAS.
Doctorando:	María Victoria Redondo Neble.
Director/es:	Francisco Manuel Guillén González.
Defensa:	26 de septiembre de 2008, Sevilla.
Calificación:	Sobresaliente cum laude por unanimidad.

Resumen:

La Tesis se centra en el análisis numérico de dos esquemas fraccionados en tiempo: un método de descomposición de la viscosidad (DV) y un método de segregación de presión (SP) basado en un esquema de proyección incremental (PI), aplicados a dos problemas: las Ecuaciones de Navier-Stokes incompresibles (NS) y las Ecuaciones Primitivas del Océano (EP).

Para el esquema DV aplicado a NS, en la tesis se mejoran las estimaciones de error obtenidas por J.Blasco y R.Codina. Se trata de un esquema con dos subetapas en tiempo, separando la no linealidad de la incompresibilidad del problema pero conservando el término de viscosidad y las condiciones de contorno para la velocidad en ambas subetapas (primero se calcula $\mathbf{u}^{m+1/2}$ como una primera aproximación de la solución $\mathbf{u}(t_{m+1})$ y después $(\mathbf{u}^{m+1}, p^{m+1})$ como aproximación de $(\mathbf{u}(t_{m+1}), p(t_{m+1}))$. La convergencia de las velocidades es obtenida por Blasco, Codina y Huerta, quienes obtuvieron además estimaciones de error de orden $O(k)$ en $l^2(\mathbf{H}^1) \cap l^\infty(\mathbf{L}^2)$ para la velocidad final $\mathbf{u}^{m+1}$ y orden $O(k^{1/2})$ en $l^2(L^2)$ para la presión. Por otra parte, se realizaron cálculos numéricos que conducían a orden óptimo $O(k)$ tanto en velocidad como en presión. Otros estudios fueron realizados con esquemas de DV obteniendo también, en el caso totalmente discreto, estimaciones suboptimales para la presión en el análisis numérico y observando resultados numéricos con orden óptimo. En la Tesis se rellena este salto demostrando analíticamente las estimaciones de orden optimal también en presión observadas en los experimentos numéricos. En concreto, para el esquema discreto en tiempo, se mejora el orden de error para la presión, de $O(\sqrt{k})$ a $O(k)$ y la norma del error en velocidad y presión, pasando de $l^\infty(\mathbf{L}^2)$ a $l^\infty(\mathbf{H}^1)$ en velocidad y de $l^2(L^2)$ a $l^\infty(L^2)$ en presión. Posteriormente, se usa este esquema semidiscreto en tiempo como un problema auxiliar, para obtener las estimaciones de error en el esquema totalmente discreto con elementos finitos, extendiendo el orden en velocidad y presión, de $O(k)$ a $O(k+h)$ y mejorando el orden del error para la velocidad en norma $l^2(\mathbf{L}^2)$, de $O(k+h)$ a $O(k+h^2)$. Juntando estos dos argumentos y bajo la restricción $h^2 \leq C\,k$, se obtiene orden óptimo de error para los errores totales.

Posteriormente, se aproxima el problema de NS 3D con un esquema de SP inspirado en un método de PI. Para el esquema en tiempo de proyección no incremental, Shen obtuvo estimaciones de orden $O(k^{1/2})$ en $l^2(\mathbf{H}^1) \cap l^\infty(\mathbf{L}^2)$

y de orden $O(k)$ en $l^2(\mathbf{L}^2)$ para la velocidad y de orden $O(k^{1/2})$ en $l^2(L^2)$ para la presión y, para el método de PI con elementos finitos y formulación mixta velocidad-presión, estas estimaciones son mejoradas por Guermond y Quartapelle a orden $O(k+h)$ en $l^\infty(\mathbf{H}^1)$ para la velocidad y a orden $O(k+h)$ en $l^\infty(L^2)$ para la presión, bajo la restricción de los parámetros $k^2 \leq \alpha\, h$. En la Tesis, se obtienen también estimaciones óptimas de error, para un esquema de SP (desacoplando la velocidad y presión) imponiendo una restricción diferente: $h \leq \alpha\, k$.

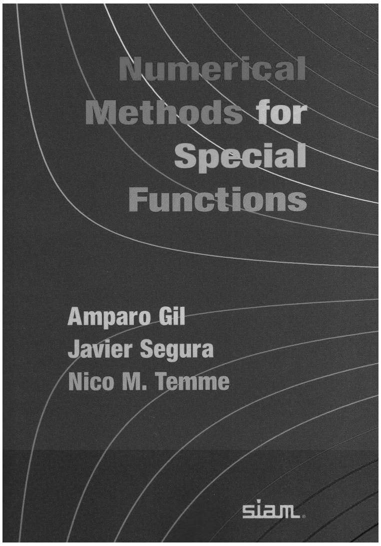
Respecto al problema de EP, se hace un análisis numérico de los esquemas DV y PI, apareciendo dos nuevas dificultades: la pérdida de regularidad de la convección vertical y la aproximación efectiva de la velocidad vertical. Así, nuevas técnicas son necesarias: uso de estimaciones anisótropas, suposición de soluciones más regulares y restricciones sobre los parámetros de discretización de distinto tipo.

En el caso del método DV, no es posible el razonamiento hecho para el problema de NS, de utilizar el problema semidiscreto en tiempo como problema auxiliar para la obtención de las estimaciones de los errores totales. Entonces, se compara la solución exacta directamente con un esquema totalmente discreto. Para una reformulación completamente diferencial del problema de EP, se propone un esquema DV con elementos finitos sobre una malla general no estructurada, obteniendo estabilidad y convergencia cuando $(k, h) \to 0$, hacia una solución débil y estimaciones de error respecto de una solución suficientemente regular. Concretamente para $l = 1, 2$ (donde l es el orden de aproximación de los elementos finitos), se obtiene error de orden $O(k + h^l)$ para la velocidad y $O(\sqrt{k} + h^l)$ para la derivada discreta de la velocidad en $l^2(\mathbf{L}^2)$, que conducen a orden $O(\sqrt{k} + h^l)$ para la presión. Seguidamente, sólo para $l = 2$, se obtienen estimaciones de orden $O(k + h^2)$ para la derivada discreta de la velocidad, que conducen a estimaciones de orden $O(k+h^2)$ para la presión. Finalmente, considerando una malla específica estructurada en vertical, se obtienen estimaciones de error de orden $O(k + h^{l+1})$ para la velocidad en $l^2(\mathbf{L}^2)$ y orden $O(k + h)$ para la presión en $l^2(L^2)$ en el caso $l = 1$.

Finalmente, para el esquema semidiscreto en tiempo de PI sobre una formulación íntegro-diferencial del problema de EP, se obtienen estimaciones de error óptimas $O(k)$ para la velocidad y la presión, que pueden servir como etapa intermedia para la obtención de estimaciones de los errores totales.

Numerical Methods for Special Functions.
Amparo Gil, Javier Segura y Nico M. Temme
SIAM
ISBN: 978-0-898716-34-4 (xvi + 415 páginas) – 2007

Por José Luis López



Esta obra es el resultado natural de los numerosos años que sus autores han dedicado a la investigación de computación de funciones especiales. Nico Temme, ya retirado, ha sido investigador destacado del MAS (Modelling, Analysis and Simulation department) del CWI (Centrum Wiskunde & Informatica) de Amsterdam, con más de cien publicaciones en revistas de relevancia de Matemática Aplicada, autor de varios libros sobre funciones especiales; un investigador de alto prestigio internacional. Javier Segura y Amparo Gil han colaborado con Nico Temme durante más de una década (al igual que el que suscribe) fundamentalmente en aspectos numéricos de funciones especiales. Profesores del Departamento de Matemáticas, Estadística y Computación del la Universidad de Cantabria, son también autores de multitud de publicaciones en revistas internacionales de prestigio sobre funciones especiales y expertos a nivel mundial en computación numérica de funciones especiales.

A pesar del elevado nivel del tema del libro, éste está escrito en un lenguaje que resulta sencillo de seguir y puede considerarse como el libro básico para todo aquel estudiante de doctorado o investigador en general que desee adentrarse en el mundo de la computación numérica de funciones especiales. El libro se divide en cuatro grandes bloques: métodos básicos, métodos avanzados, ejemplos y tópicos relacionados y un cuarto bloque dedicado al software. La primera parte contiene una selección de métodos considerados por los autores como los más importantes y populares: series convergentes y series asintóticas, desarrollos de Chebyshev, relaciones de recurrencia lineales y métodos de cuadratura, todos ellos ilustrados con abundantes ejemplos de funciones especiales. Además, esta primera parte es muy intuitiva y auto-contenida. La segunda parte es un compendio de métodos de computación de funciones especiales más avanzados, quizá menos conocidos pero no por ello menos eficientes que los anteriores. Probablemente el ejemplo más destacado sea el uso de desarrollos asintóticos uniformes, una técnica relativamente reciente y muy potente que proporciona aproximaciones válidas en regiones de los parámetros involucrados en el problema más amplias que los métodos no uniformes. En esta segunda parte encontramos métodos tales como fracciones continuas, cálculo de ceros de funciones especiales, los ya mencionados desarrollos asintóticos uniformes, aproximantes de Padé, mejor aproximación racional, el método de Frobenius y métodos de cuadratura más avanzados como son los métodos de Clenshaw-Curtis y Filon. La tercera parte del libro se centra en métodos menos generales aunque no por ello menos interesantes, ya que son métodos diseñados específicamente para ciertas funciones especiales. Incluye la inversión numérica y asintótica de una clase de funciones de distribución con aplicaciones en estadística, con especial énfasis en las funciones beta y gamma. También contiene otros tópicos tales como la fórmula de sumación de Euler (con algunas aplicaciones), el método de Carlson de computación de integrales elípticas y la inversión numérica de transformadas de Laplace. La última parte es la parte más aplicada. El libro se adentra aquí en la implementación práctica de algunos de estos métodos ofreciendo algorítmos de programación para el cómputo de algunas (las más relevantes) funciones especiales como son las funciones de Airy, de Legendre o las funciones cilíndricas parabólicas. Incluso ofrece el sitio web `http://funciones.unican.es` donde pueden encontrarse las rutinas de Fortran90 correspondientes a esos algorítmos. En suma, este libro viene a cubrir un importante hueco en la bibliografia básica de evaluación de funciones especiales y no debería faltar en la biblioteca de ningún investigador que trabaje directa o indirectamente con funciones especiales.

X PREMIO S$\vec{e}$MA AL JOVEN INVESTIGADOR

SOCIEDAD ESPAÑOLA DE MATEMÁTICA APLICADA

PREÁMBULO

La Sociedad Española de Matemática Aplicada (S$\vec{e}$MA), en cumplimiento de su objetivo de contribuir al desarrollo en nuestro país de las Matemáticas y sus aplicaciones y, más en concreto, de promover y estimular la investigación y procurar medios para efectuarla, consciente del notable desarrollo que las Matemáticas están experimentando y de la necesidad de promover el interés de las jóvenes generaciones por la tarea de la creación científica, convencida del papel positivo que el aprecio de la comunidad juega en la vida científica de los investigadores y siguiendo con una tradición honrosa y habitual tanto en las Artes como en las Ciencias, convoca el "Décimo Premio S$\vec{e}$MA al Joven Investigador", según las bases que se adjuntan.

BASES GENERALES

1. La Sociedad Española de Matemática aplicada (S$\vec{e}$MA) convoca el "Premio S$\vec{e}$MA al Joven Investigador", que se concederá anualmente.

2. Son posibles candidatos todos los investigadores españoles que, a la fecha del límite de presentación de candidaturas, no rebasen la edad de 33 años. También pueden serlo aquellos investigadores de otras nacionalidades que tengan un puesto de trabajo permanente en una Universidad o Centro de investigación español y cumplan la condición de edad. No pueden concurrir al Premio candidatos galardonados en convocatorias precedentes.

3. El Premio está destinado a promover la excelencia en el trabajo matemático original en todas las ramas de las Matemáticas que tienen una componente aplicada. Su objetivo es premiar la contribución personal del candidato. El limite de edad fijado pretende señalar candidatos que hayan tenido tiempo de desarrollar su creatividad matemática independiente tras la etapa formativa correspondiente a la Tesis Doctoral. El Premio tiene así por objetivo abrirles el camino de su periodo de madurez y reconocer al mismo tiempo sus capacidades demostradas.

4. Los méritos serán juzgados por un Comité Científico de cinco miembros, nombrado por el Consejo Ejecutivo de la Sociedad entre investigadores de probado prestigio. Este Comité tendrá su propio reglamento de funcionamiento. En todo caso, será presidido por el Presidente de la

Sociedad u otro miembro del Consejo Ejecutivo en quien delegue, no pudiendo ser miembros del Comité Científico más de dos miembros del Consejo Ejecutivo.

5. Los candidatos habrán de presentar, dentro del plazo que se cite, una Memoria exponiendo la trayectoria vital y los méritos que concurren, un curriculum normalizado, así como otros documentos que puedan ser pertinentes para acreditar sus contribuciones originales a las Matemáticas y sus aplicaciones. Las candidaturas pueden ser presentadas también por otros investigadores. El Comité se reserva el derecho de recabar la información complementaria necesaria del candidato o de quien le haya presentado

6. El galardonado con el Premio recibirá de la Sociedad un Diploma acreditativo y una cuantía que será establecida en cada convocatoria por la Sociedad.

7. La Sociedad requerirá al candidato galardonado un resumen de su trabajo de investigación escrito en estilo divulgativo, con una extensión a convenir entre las 6 y las 20 páginas para su publicación en el Boletín de la Sociedad. Este resumen puede formar parte de la Memoria mencionada en el punto 5.

8. El fallo del concurso es irrevocable. El Comité acompañará la concesión del Premio de una exposición de los méritos hallados en el candidato galardonado. Por lo demás, las deliberaciones y resoluciones del Comité serán regidas por su reglamento.

BASES PARTICULARES DE LA CONVOCATORIA DE 2009

9. La fecha límite de presentación de candidaturas es el 30 de abril de 2009. Podrán concursar por tanto las personas que hayan nacido después del 30 de abril de 1975.

10. La documentación presentada constará de la Memoria y el curriculum citados, así como copia de las cinco contribuciones más importantes del investigador a las Matemáticas y sus aplicaciones, todo ello por quintuplicado.

    Se recomienda a los candidatos que presenten su propia candidatura que la Memoria se adecúe o en su caso contenga el resumen del trabajo de investigación referido en el apartado 7.

11. La documentación debe ser dirigida a

    Prof. Carlos Vázquez Cendón
    Premio S$\vec{e}$MA Joven Investigador 2009
    Departamento de Matemáticas
    Facultad de Informática
    Universidad de La Coruña
    15071 - La Coruña

12. La cuantía actual del Premio es de 1800€. El Premio es indivisible. Además, el candidato galardonado quedará eximido del pago de las cuotas como socio de S$\vec{e}$MA correspondientes a los años 2010 y 2011. En caso de no ser miembro de S$\vec{e}$MA, pasaría automáticamente a serlo.

13. El Premio será fallado antes del 31 de agosto de 2009 y será entregado con ocasión de la Asamblea anual de la Sociedad, en el marco del XXI Congreso de Ecuaciones Diferenciales y Aplicaciones (CEDYA)-XI Congreso de Matemática Aplicada (CMA)), que tendrá lugar en Ciudad Real, entre el 21 y el 25 de septiembre de 2009.

La Coruña, a 1 de diciembre de 2008

PREMIO AL MEJOR ARTÍCULO DEL BOLETÍN Se⃗MA

PREÁMBULO

La Sociedad Española de Matemática Aplicada (Se⃗MA), consciente del notable desarrollo que las Matemáticas están experimentando, del incremento de su influencia sobre todos los aspectos de la vida en las sociedades desarrolladas y de la necesidad de estimular el interés del público por la cultura científica, en cumplimiento de sus objetivos, anuncia el "Premio al mejor artículo del Boletín Se⃗MA", según las bases que se adjuntan.

Es uno de los intereses principales de Se⃗MA promover la divulgación de las Matemáticas, su relevancia y su eficacia. Dada la enorme variedad de intereses aplicados de las Matemáticas, las Bases del concurso pretenden dar preferencia a los temas que tradicionalmente han estado ligados a Se⃗MA de una u otra manera. Muy en especial, deben ser mencionados el análisis teórico y numérico, el control y los aspectos computacionales de sistemas que permiten modelar fenómenos con origen en otras Ciencias.

BASES GENERALES

1. La Sociedad Española de Matemática Aplicada (Se⃗MA) concederá anualmente el "Premio al mejor artículo del Boletín Se⃗MA".

2. Serán posibles candidatos todos los artículos publicados en los volúmenes del Boletín Se⃗MA del año indicado en la convocatoria.

3. El artículo premiado será elegido por un Comité Científico de cinco miembros nombrados por el Comité Ejecutivo de la Sociedad. El Comité estará formado por el Editor Jefe del Boletín, el Presidente de la Sociedad u otro miembro del Comité Ejecutivo en quien delegue, y tres miembros del Comité Científico del Boletín propuestos por el Editor Jefe, quien actuará como presidente del Comité. Una vez constituido, el Comité tendrá sus propias normas de funcionamiento.

4. El Premio tendrá una dotación económica, a repartir entre los autores del artículo. La concesión del Premio llevará aparejada la entrega de un diploma acreditativo y las certificaciones correspondientes. Además, los autores quedarán eximidos del pago de las cuotas como socios de Se⃗MA durante un año. En caso de no ser miembros de Se⃗MA, pasarían automáticamente a serlo de manera gratuita durante ese período.

5. El fallo del Premio es irrevocable. El Comité acompañará la concesión del Premio de una exposición de los méritos hallados en el artículo galardonado.

6. El trabajo premiado deberá ser presentado por alguno de sus autores en alguna de las actividades organizadas por la Sociedad en el año en que se otorga, durante la cual se entregará el diploma.

BASES DEL PREMIO AL MEJOR ARTÍCULO DEL BOLETÍN S$\vec{e}$MA 2008

1. Serán posibles candidatos al “Premio al mejor artículo del Boletín S$\vec{e}$MA 2008” todos los artículos publicados en el Boletín S$\vec{e}$MA en el año 2008.

2. La dotación económica de este Premio es de 1.500€.

3. El fallo del Comité se hará público antes del 30 de abril de 2009.

4. El trabajo premiado deberá ser presentado por alguno de sus autores en el contexto del Congreso CEDYA-CMA 2009, que se celebrará en Ciudad Real, del 21 al 25 de septiembre de 2009. En dicho Congreso se entregará el correspondiente Diploma acreditativo.

BOLETIN S$\vec{e}$MA BEST PAPER AWARD

The Sociedad Española de Matematica Aplicada (S$\vec{e}$MA), aware of the remarkable development of Applied Mathematics and of its growing influence in the life of modern societies, and aware as well of the necessity to encourage the transfer of scientific knowledge within our community, is proud to announce the “Boletín S$\vec{e}$MA Best Paper Award”, for outstanding papers published in the Boletín S$\vec{e}$MA. The guidelines of this prize are established below.

One of the main interests of S$\vec{e}$MA is to promote the dissemination of Applied Mathematics, emphasizing its relevance and its potential. There is an increasingly large body of applied interests in Mathematics, and the guidelines below seek to enforce those subjects that traditionally have formed the core of the scientific activities in S$\vec{e}$MA, that is theoretical and numerical analysis, control theory and computational techniques in mathematical modeling in science.

GENERAL GUIDELINES

1. The Sociedad Española de Matemática Aplicada (S$\vec{e}$MA) will award the “Boletin S$\vec{e}$MA Best-Paper Award” on a yearly basis.

2. All papers published in the volumes of the Boletín S$\vec{e}$MA of the corresponding year will be eligible for the prize.

3. There will be an ad-hoc prize Committee, composed by 5 members: The editor in chief of the Boletín S$\vec{e}$MA, who will act as chair of the committee, the president of S$\vec{e}$MA (or a member of the executive committee of S$\vec{e}$MA acting in his name) and three members of the scientific committee of the Boletín S$\vec{e}$MA, proposed by the editor in chief and appointed by the executive committee of the Society. The prize committee will devise its own procedures to carry out the selection process.

4. The award includes a diploma containing the citation and a cash prize to be shared among the authors of the awarded paper. If the authors are S$\vec{e}$MA members, the S$\vec{e}$MA membership fee will be waived for a one year period . If any of the authors is not a member of S$\vec{e}$MA, he or she will be automatically granted membership in the Society, and the fee will be waived for one year.

5. The decision of the Committee is final. The committee will report on the merits found in the awarded paper.

6. One of the authors of the awarded paper should agree to deliver a presentation about the contents of the paper in one of the activities organized by S$\vec{e}$MA during the year in which the prize is awarded.

BOLETÍN S$\vec{e}$MA BEST PAPER AWARD 2008. GUIDELINES

1. All papers published in the Boletín S$\vec{e}$MA in 2008 are elegible for the “Boletin SeMA Best Paper Award”.

2. The award includes a cash prize of 1.500€.

3. The decision of the committee will be published before April 30th 2009.

4. The Best-Paper of the year will be the object of a lecture, delivered by one of its authors.

POSTDOCTORAL POSITIONS

Applications are invited for several postdoctoral positions at the Center of Mathematical Modeling CMM - Universidad de Chile. CMM is an associate research unit to CNRS-France.

Candidates in all areas of mathematics will be considered, with an emphasis in those currently cultivated at CMM: discrete mathematics, probability, dynamical systems, optimization and equilibria/mathematical economics, partial differential equations/nonlinear analysis, numerical analysis/ fluid dynamics/ mathematical and computational mechanics, mathematical biology. Candidates interested in participating in applied projects in fields such as Transportation, Telecommunications, Mininig and Energy are also welcome.

Candidates should exhibit evidence of outstanding research potential and are required to have received a Ph.D. degree in mathematics or a related field, not before 2003. They are expected to dedicate fully to research in an active research-oriented environment, with no teaching responsabilities.

Spanish language is not required. Positions are for one year, renewable for up to three years. Appointments do not have a fixed starting date but should preferably begin between June and September 2009.

Monthly salary is $ 1.200.000, about US$ 2.000. As a reference, the rent of a furnished apartment near CMM is about US$ 500 per month. Flight expenses (roundtrip) will be covered. Possibility to apply for national grants in case of a longer stay is also open.

Candidates should submit curriculum vitae, research statement and arrange for at least three recommendation letters sent no later than March 30th 2009 by email to María Inés Rivera (`mrivera@dim.uchile`) with additional hard copies to

POSTDOCTORAL SEARCH COMMITTEE
Center for Mathematical Modelling
Universidad de Chile
Casilla 170/3, Correo 3
Santiago, CHILE

SEMA MEMBERS: JOIN SIAM AT REDUCED RATES!

Expand your network! If you are a member of SEMA and live outside the United States, you can now become a reciprocal member of the Society for Industrial and Applied Mathematics (SIAM) for 30 % less than the standard dues. For 2009, the SIAM regular member dues of US $127 will be reduced to US $88.90 for SEMA members in good standing who reside outside the USA.

You can join online or download a reciprocal member application at
http://www.siam.org/membership/individual/reciprocal.php

Join SIAM for networking opportunities, visibility in the applied mathematics and computational science communities, and access to cutting-edge research. Your membership will become active upon receipt of your application through December 31, 2009.

Let SIAM be another source for news and information about applied mathematics and computational science. A SIAM membership includes subscriptions to SIAM News and SIAM Review, and entitles you to substantial discounts on SIAM books, journals, and conferences.

CONTACT INFO FOR AD:

Society for Industrial and Applied Mathematics
3600 Market Street, 6th Floor, Philadelphia, PA 19104-2688 USA
Phone: +1-215-382-9800 or 1-800-447-7426 (toll free in USA and Canada)
Fax: +1-215-386-7999
Email: membership@siam.org
Web: http://www.siam.org

Tipo de evento:	Congreso
Nombre:	APLIMAT 2009: The 8th International Conference on Applied Mathematics
Lugar:	Bratislava, República Eslovaca
Fecha:	February 3–6, 2009
Organiza:	Faculty of Mechanical Engineering, Slovak University of Technology in Bratislava
Información:	
E-mail:	aplimat@aplimat.com
WWW:	http://www.aplimat.com/

Tipo de evento:	Congreso
Nombre:	Conference on Mathematical Biology: Modeling and Differential Equations
Lugar:	Campus de la Universitat Autònoma de Barcelona, Bellaterra, Barcelona
Fecha:	February 09-13, 2009
Organiza:	
Información:	
E-mail:	cmodeling@crm.cat
WWW:	http://www.crm.cat/CMODELING/

Tipo de evento:	Congreso
Nombre:	iCBBE2009: The 3rd International Conference on Bioinformatics and Biomedical Engineering
Lugar:	Beijing, China
Fecha:	June 11-13, 2009
Organiza:	
Información:	
E-mail:	`submit@icbbe.org`
WWW:	`http://www.icbbe.org/`

Tipo de evento:	Congreso
Nombre:	83ème rencontre entre physiciens théoriciens et mathématiciens : Théorie des représentations en mathématique et en physique
Lugar:	Strasbourg (Francia)
Fecha:	June 11-13, 2009
Organiza:	Institut de Recherche Mathématique Avancée (University of Strasbourg and CNRS)
Información:	
E-mail:	`papadop@math.u-strasbg.fr; souaifi@math.u-strasbg.fr`
WWW:	`http://www-irma.u-strasbg.fr/article717.html`

Tipo de evento:	Congreso
Nombre:	International Conference: Dynamical Systems and Applications
Lugar:	Constantza, Romania
Fecha:	June 15-18, 2009
Organiza:	Faculty of Mathematics and Computer Science and Faculty of Civil Engineering (Ovidius University of Constantza)
Información:	
E-mail:	`Cristina Gherghina (cgherghina@gmail.com); Cristina Dana Toncu (cristinatoncu@canals.ro)`
WWW:	`http://www.univ-ovidius.ro/faculties/civil_eng/conferinta%20iunie%202009/Home.html`

Tipo de evento:	Congreso
Nombre:	WAVES 2009
Lugar:	Pau, France
Fecha:	June 15-19, 2009
Organiza:	Institut National de Recherche en Informatique et en Automatique (INRIA)
Información:	

E-mail:	`helene.barucq@inria.fr;`
	`julien.diaz@inria.fr`

WWW:	`https://waves-2009.bordeaux.inria.fr/`

Tipo de evento:	Congreso
Nombre:	HPCC-09: The 11th IEEE International Conference on High Performance Computing and Communications
Lugar:	Korea University, Seoul, Korea
Fecha:	June 25-27, 2009
Organiza:	
Información:	

E-mail:	`juan@udc.es (program);`
	`parkjonghyuk1@hotmail.com (general);`
	`ysjeong@wonkwang.ac.kr (workshops)`

WWW:	`http://www.sersc.org/HPCC2009/`

Tipo de evento:	Congreso
Nombre:	ICAEM 2009: The 2009 International Conference of Applied and Engineering Mathematics
Lugar:	London, UK
Fecha:	July 01-03, 2009
Organiza:	International Association of Engineers
Información:	
E-mail:	`WCE@iaeng.org`
WWW:	`http://www.iaeng.org/WCE2009/ICAEM2009.html`

Tipo de evento:	Congreso
Nombre:	AFE 2009: The 6th International Conference on Applied Financial Economics
Lugar:	Samos Island, Greece
Fecha:	July 02-04, 2009
Organiza:	Research and Training Institute of East Aegean, University of Piraeus, University of the Aegean (Department of Statistics and Actuarial - Financial Mathematics), Democritus University of Thrace
Información:	
E-mail:	
WWW:	`http://www.ineag.gr/AFE/`

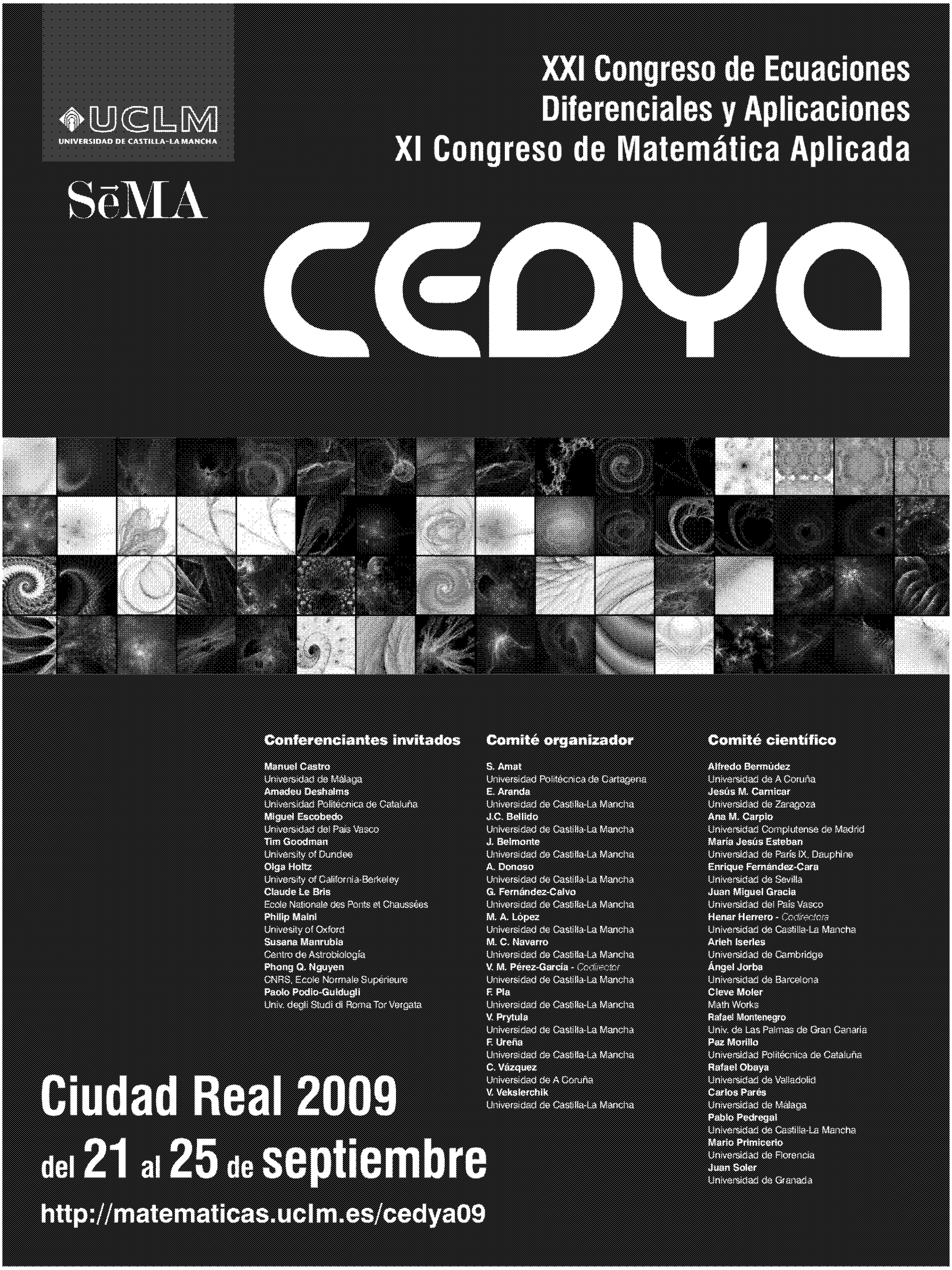
Tipo de evento:	Congreso
Nombre:	SIAM Conference on Control and its Applications
Lugar:	Sheraton Denver Hotel, Denver, Colorado, USA
Fecha:	July 06-08, 2009
Organiza:	Society for Industrial and Applied Mathematics (SIAM)
Información:	
E-mail:	
WWW:	`http://www.siam.org/meetings/ct09/`

Tipo de evento:	Congreso
Nombre:	ICMP09: XVI International Congress on Mathematical Physics
Lugar:	Praga, República Checa
Fecha:	August 03-08, 2009
Organiza:	
Información:	
E-mail:	`icmp09@cbttravel.cz`
WWW:	`http://www.icmp09.com/`

Tipo de evento:	Congreso
Nombre:	Workshop in nonlinear elliptic PDEs
Lugar:	Université Libre de Bruxelles, Brussels, Belgium
Fecha:	September 02-04, 2009
Organiza:	
Información:	
E-mail:	`wnpde09@ulb.ac.be`
WWW:	`http://wnpde09.ulb.ac.be/`

Tipo de evento:	Congreso
Nombre:	CEDYA 2009: XXI Congreso de Ecuaciones Diferenciales y Aplicaciones / XI Congreso de Matemática Aplicada
Lugar:	Universidad de Castilla–La Mancha, Ciudad Real.
Fecha:	21–25 septiembre, 2009
Organiza:	S$\vec{e}$MA, UCLM
Información:	
E-mail:	`Congreso.CEDYA09.secretaria@uclm.es`
WWW:	`http://matematicas.uclm.es/cedya09/`

Tipo de evento:	Congreso
Nombre:	2009 SIAM/ACM Joint Conference on Geometric and Physical Modeling
Lugar:	San Francisco, California, USA
Fecha:	October 05-09, 2009
Organiza:	Society for Industrial and Applied Mathematics (SIAM) y Association for Computer Machinery (ACM)
Información:	
E-mail:	`Congreso.CEDYA09.secretaria@uclm.es`
WWW:	`http://www.siam.org/meetings/gdspm09/`



UCLM
UNIVERSIDAD DE CASTILLA-LA MANCHA

SēMA

XXI Congreso de Ecuaciones Diferenciales y Aplicaciones
XI Congreso de Matemática Aplicada

CEDYA

Conferenciantes invitados

Manuel Castro
Universidad de Málaga
Amadeu Deshalms
Universidad Politécnica de Cataluña
Miguel Escobedo
Universidad del País Vasco
Tim Goodman
University of Dundee
Olga Holtz
University of California-Berkeley
Claude Le Bris
Ecole Nationale des Ponts et Chaussées
Philip Maini
Univesity of Oxford
Susana Manrubia
Centro de Astrobiología
Phong Q. Nguyen
CNRS, Ecole Normale Supérieure
Paolo Podio-Guidugli
Univ. degli Studi di Roma Tor Vergata

Comité organizador

S. Amat
Universidad Politécnica de Cartagena
E. Aranda
Universidad de Castilla-La Mancha
J.C. Bellido
Universidad de Castilla-La Mancha
J. Belmonte
Universidad de Castilla-La Mancha
A. Donoso
Universidad de Castilla-La Mancha
G. Fernández-Calvo
Universidad de Castilla-La Mancha
M. A. López
Universidad de Castilla-La Mancha
M. C. Navarro
Universidad de Castilla-La Mancha
V. M. Pérez-García - *Codirector*
Universidad de Castilla-La Mancha
F. Pla
Universidad de Castilla-La Mancha
V. Prytula
Universidad de Castilla-La Mancha
F. Ureña
Universidad de Castilla-La Mancha
C. Vázquez
Universidad de A Coruña
V. Vekslerchik
Universidad de Castilla-La Mancha

Comité científico

Alfredo Bermúdez
Universidad de A Coruña
Jesús M. Carnicar
Universidad de Zaragoza
Ana M. Carpio
Universidad Complutense de Madrid
María Jesús Esteban
Universidad de París IX, Dauphine
Enrique Fernández-Cara
Universidad de Sevilla
Juan Miguel Gracia
Universidad del País Vasco
Henar Herrero - *Codirectora*
Universidad de Castilla-La Mancha
Arieh Iserles
Universidad de Cambridge
Ángel Jorba
Universidad de Barcelona
Cleve Moler
Math Works
Rafael Montenegro
Univ. de Las Palmas de Gran Canaria
Paz Morillo
Universidad Politécnica de Cataluña
Rafael Obaya
Universidad de Valladolid
Carlos Parés
Universidad de Málaga
Pablo Pedregal
Universidad de Castilla-La Mancha
Mario Primicerio
Universidad de Florencia
Juan Soler
Universidad de Granada

Ciudad Real 2009
del 21 al 25 de septiembre

http://matematicas.uclm.es/cedya09

UCLM
UNIVERSIDAD DE CASTILLA-LA MANCHA

imaci
Instituto de Matemática Aplicada
a la Ciencia y la Ingeniería
Universidad de Castilla-La Mancha

SēMA

Castilla-La Mancha

Camas Jiménez, Inmaculada
Estudiante. *Líneas de investigación:* – UNIV. DE SEVILLA – Fac. de Matemáticas – Dpto. de Ecuaciones Diferenciales y Análisis Numérico – Tarfia, s/n. 41012 Sevilla.

e-mail: `inmcamjim@alum.us.es`.

Fernández García, Soledad
Estudiante. *Líneas de investigación:* Sistemas dinámicos – UNIV. DE SEVILLA – E. T. S. de Ingenieros – Dpto. de Matemática Aplicada II – Camino de los Descubrimientos, s/n. Isla Cartuja. 41092 Sevilla.
Tlf.: 954.486.169. *Fax:* 954.486.166.
e-mail: `soledadfdezgarcia@gmail.com`.
`http://www.ma2.us.es`

Santágueda Villanueva, María
Estudiante (Becario). *Líneas de investigación:* Multiresolución y análisis de imagen – UNIV. DE VALENCIA – Fac. de Matemáticas – Dpto. de Matemática Aplicada – Dr. Moliner, 50. 46100 Burjassot (Valencia).
Tlf.: 963.543.217.
e-mail: `sanvima@alumni.uv.es`.
`http://gata.uv.es`

Ureña Prieto, Francisco
Prof. Asociado. *Líneas de investigación:* Métodos numéricos – UNIV. DE CASTILLA-LA MANCHA – E. T. S. de Ingenieros Industriales – Dpto. de Matemáticas – Avda. Camilo José Cela, s/n. 13071 Ciudad Real.
Tlf.: 926.295.300. *Fax:* 926.295.361.
e-mail: `francisco.urena@uclm.es`.

Direcciones útiles

Consejo Ejecutivo de S$\vec{e}$MA

Presidente:

Carlos Vázquez Cendón. (carlosv@udc.es).
Dpto. de Matemáticas. Facultad de Informática. Univ. de A Coruña. Campus de Elviña, s/n. 15071 A Coruña. *Tel:* 981 16 7000-1335.

Vicepresidente:

Rosa María Donat Beneito. (Rosa.M.Donat@uv.es)
Dpto. de Matemática Aplicada. Fac. de Matemàtiques. Univ. de Valencia. Dr. Moliner, 50. 46100 Burjassot (Valencia) *Tel:* 963 544 727.

Secretario:

Carlos Castro Barbero. (ccastro@caminos.upm.es).
Dpto. de Matemática e Informática. E.T.S.I. Caminos, Canales y Puertos. Univ. Politécnica de Madrid. Av. Aranguren s/n. 28040 Madrid. *Tel:* 91 336 6664.

Vocales:

Sergio Amat Plata. (sergio.amat@upct.es)
Dpto. de Matemática Aplicada y Estadística. Univ. Politécnica de Cartagena. Paseo de Alfonso XIII, 52. 30203 Cartagena (Murcia). *Tel:* 968 325 694.

Rafael Bru García. (rbru@mat.upv.es)
Dpto. de Matemática Aplicada. E.T.S.I. Agrónomos. Univ. Politécnica de Valencia. Camí de Vera, s/n. 46022 Valencia. *Tel:* 963 879 669.

José Antonio Carrillo de la Plata. (carrillo@mat.uab.es)
Dpto. de Matemáticas. Univ. Autónoma de Barcelona. Edifici C. 08193 Bellaterra (Barcelona). *Tel:* 935 812 413.

Inmaculada Higueras Sanz. (higueras@unavarra.es).
Dpto de Matemática e Informática Univ. Pública de Navarra. Campus de Arrosadía, s/n. *Tel:* 948 169 526. 31006 Pamplona.

Carlos Parés Madroñal. (carlos_pares@uma.es).
Dpto. de Análisis Matemático. Fac. de Ciencias. Univ. de Málaga. Campus de Teatinos, s/n. 29080 Málaga. *Tel:* 952 132 017.

Pablo Pedregal Tercero. (Pablo.Pedregal@uclm.es).
Dpto. de Matemáticas. E.T.S.I. Industriales. Univ. de Castilla-La Mancha. Avda. de Camilo José Cela, s/n. 13071 Ciudad Real. *Tel:* 926 295 436

Luis Vega González. (luis.vega@ehu.es).
Dpto. de Matemáticas. Fac. de Ciencias. Univ. del País Vasco. Aptdo. 644. 48080 Bilbao (Vizcaya). *Tel:* 944 647 700.

Tesorero:

Íñigo Arregui Álvarez. (arregui@udc.es).
Dpto. de Matemáticas. Fac. de Informática. Univ. de A Coruña. Campus de Elviña, s/n. 15071 A Coruña. *Tel:* 981 16 7000-1327.

Comité Científico del Boletín de S$\vec{e}$MA

Enrique Fernández Cara. (cara@us.es).
Dpto. de Ecuaciones Diferenciales y An. Numérico. Fac. de Matemáticas. Univ. de Sevilla. Tarfia, s/n. 41012 Sevilla. *Tel:* 954 557 992.

Alfredo Bermúdez de Castro. (mabermud@usc.es).
Dpto. de Matemática Aplicada. Fac. de Matemáticas. Univ. de Santiago de Compostela. Campus Univ.. 15706 Santiago (A Coruña) *Tel:* 981 563 100.

Carlos Conca Rosende. (cconca@dim.uchile.cl).
Dpto. de Ingeniería Matemática. Univ. de Chile. Blanco Encalada 2120. Santiago (Chile) *Tel:* (+56) 0 978 4459.

Amadeus Delshams Valdés. (Amadeu.Delshams@upc.es).
Dpto. de Matemática Aplicada I. Univ. Politécnica de Cataluña. Diagonal 647. 08028 Barcelona. *Tel:* 934 016 052.

Martin J. Gander (Martin.Gander@math.unige.ch).
Section de Mathématiques. Université de Genève. 2-4 rue du Liévre, CP 64. CH-1211 Genève (Suiza). *Fax:* (+41) 22 379 11 76.

Vivette Girault (girault@ann.jussieu.fr). Laboratoire Jacques-Louis Lions. Université Paris VI. Boite Courrier 187, 4 Place Jussieu 75252 Paris Cedex 05 (Francia).

Arieh Iserles (A.Iserles@damtp.cam.ac.uk).
Department of Applied Mathematics and Theoretical Physics. University of Cambridge. Wilberforce Rd Cambridge (Reino Unido). *Tel:* (+44) 1223 337891.

José Manuel Mazón Ruiz. (Jose.M.Mazon@uv.es).
Dpto. de Análisis Matemático. Fac. de Matemáticas. Univ. de Valencia. Dr. Moliner, 50. 46100 Burjassot (Valencia) *Tel:* 963 664 721.

Pablo Pedregal Tercero. (Pablo.Pedregal@uclm.es).
Dpto. de Matemáticas. E.T.S.I. Industriales. Univ. de Castilla-La Mancha. Avda. Camilo José Cela s/n. 13071 Ciudad Real. *Tel:* 926 295 436 .

Ireneo Peral Alonso. (ireneo.peral@uam.es).
Dpto. de Matemáticas, C-XV. Fac. de Ciencias. Univ. Aut. de Madrid. Cantoblanco, Ctra. de Colmenar, km. 14. 28049 Madrid. *Tel:* 913 974 204.

Benoît Perthame. (benoit.perthame@ens.fr).
Laboratoire Jacques-Louis Lions. Université Paris VI. 175, rue du Chevaleret. 75013 Paris, (Francia). *Tel:* (+33) 1 44 32 20 36.

Olivier Pironneau (pironneau@ann.jussieu.fr).
Laboratoire Jacques-Louis Lions. Université Paris VI. 35 rue de Bellefond. 75009 Paris (Francia). *Tel:* (+33) 1 42 80 12 97.

Alfio Quarteroni. (alfio.quarteroni@epfl.ch).
Institute of Analysis and Scientific Computing. Ecole Polytechnique Fédérale de Lausanne. Piccard Station 8. CH-1015 Lausanne (Suiza) *Tel:* (+41) 21 69 35546.

Juan Luis Vázquez Suárez. (juanluis.vazquez@uam.es).
Dpto. de Matemáticas, C-XV. Fac. de Ciencias. Univ. Aut. de Madrid. Cantoblanco, Crta. de Colmenar, km. 14. 28049 Madrid. *Tel:* 913 974 935.

Luis Vega González. (mtpvegol@lg.ehu.es).
Dpto. de Matemáticas. Fac. de Ciencias. Univ. del País Vasco. Aptdo. 644. 48080 Bilbao (Vizcaya). *Tel:* 944 647 700.

Chi-Wang Shu. (shu@dam.brown.edu).
Division of Applied Mathematics Box F. 182 George Street Brown University Providence RI 02912 *Tel:* (401) 863-2549

Enrique Zuazua Iriondo. (enrique.zuazua@uam.es).
Dpto. de Matemáticas. Fac. de Ciencias. Univ. Aut. de Madrid. Cantoblanco, Ctra. de Colmenar, km. 14. 28049 Madrid. *Tel:* 913 974 368.

Grupo Editor del Boletín de S$\vec{e}$MA

Pablo Pedregal Tercero. (Pablo.Pedregal@uclm.es).
Dpto. de Matemáticas. E.T.S.I. Industriales. Univ. de Castilla-La Mancha. Avda. Camilo José Cela, s/n. 13071 Ciudad Real. *Tel:* 926 295 300 ext. 3809

Enrique Fernández Cara. (cara@us.es).
Dpto. de Ecuaciones Diferenciales y An. Numérico. Fac. de Matemáticas. Univ. de Sevilla. Tarfia, s/n. 41012 Sevilla. *Tel:* 954 557 992.

Ernesto Aranda Ortega. (Ernesto.Aranda@uclm.es).
Dpto. de Matemáticas. E.T.S.I. Industriales. Univ. de Castilla-La Mancha. Avda. Camilo José Cela, s/n. 13071 Ciudad Real. *Tel:* 926 295 300 ext. 3813

José Carlos Bellido Guerrero. (JoseCarlos.Bellido@uclm.es).
Dpto. de Matemáticas. E.T.S.I. Industriales. Univ. de Castilla-La Mancha. Avda. Camilo José Cela, s/n. 13071 Ciudad Real. *Tel:* 926 295 300 ext. 3859

Alberto Donoso Bellón. (Alberto.Donoso@uclm.es).
Dpto. de Matemáticas. E.T.S.I. Industriales. Univ. de Castilla-La Mancha. Avda. Camilo José Cela, s/n. 13071 Ciudad Real. *Tel:* 926 295 300 ext. 3859

Responsables de secciones del Boletín de S$\vec{e}$MA

Artículos:

Enrique Fernández Cara. (cara@us.es).
Dpto. de Ecuaciones Diferenciales y An. Numérico. Fac. de Matemáticas. Univ. de Sevilla. Tarfia, s/n. 41012 Sevilla. *Tel:* 954 557 992.

Matemáticas e Industria:

Mikel Lezaun Iturralde. (mepleitm@lg.ehu.es).
Dpto. de Matemática Aplicada, Estadística e I. O. Fac. de Ciencias. Univ. del País Vasco. Aptdo. 644. 48080 Bilbao (Vizcaya). *Tel:* 944 647 700.

Educación Matemática:

Roberto Rodríguez del Río. (rr_delrio@mat.ucm.es).
Dpto. de Matemática Aplicada. Fac. de Químicas. Univ. Compl. de Madrid. Ciudad Universitaria. 28040 Madrid. *Tel:* 913 944 102.

Resúmenes de libros:

Fco. Javier Sayas González. (jsayas@posta.unizar.es).
Dpto. de Matemática Aplicada. Centro Politécnico Superior . Universidad de Zaragoza. C/María de Luna, 3. 50015 Zaragoza. *Tel:* 976 762 148.

Noticias de S$\vec{e}$MA:

Carlos Castro Barbero. (`ccastro@caminos.upm.es`).
Dpto. de Matemática e Informática. E.T.S.I. Caminos, Canales y Puertos. Univ. Politécnica de Madrid. Av. Aranguren s/n. 28040 Madrid. *Tel:* 91 336 6664.

Anuncios:

Óscar López Pouso. (`oscarlp@usc.es`).
Dpto. de Matemática Aplicada. Fac. de Matemáticas. Univ. de Santiago de Compostela. Campus sur, s/n. 15782 Santiago de Compostela *Tel:* 981 563 100, ext. 13228.

Responsables de otras secciones de S$\vec{e}$MA

Gestión de Socios:

Íñigo Arregui Álvarez. (`arregui@udc.es`).
Dpto. de Matemáticas. Fac. de Informática. Univ. de A Coruña. Campus de Elviña, s/n. 15071 A Coruña. *Tel:* 981 16 7000-1327.

Página web: `www.sema.org.es/`**:**

Carlos Castro Barbero. (`ccastro@caminos.upm.es`).
Dpto. de Matemática e Informática. E.T.S.I. Caminos, Canales y Puertos. Univ. Politécnica de Madrid. Av. Aranguren s/n. 28040 Madrid. *Tel:* 91 336 6664.

INFORMACIÓN PARA LOS AUTORES

1. Los artículos publicados en este Boletín podrán ser escritos en español o inglés y deberán ser enviados por correo certificado a

   Prof. E. FERNÁNDEZ CARA
   Presidente del Comité Científico, Boletín S$\vec{e}$MA
   Dpto. E.D.A.N., Facultad de Matemáticas
   Aptdo. 1160, 41080 SEVILLA

   También podrán ser enviados por correo electrónico a la dirección

   `boletin.sema@uclm.es`

   En ambos casos, el/los autor/es deberán enviar por correo certificado una carta a la dirección precedente mencionando explícitamente que el artículo es sometido a publicación e indicando el nombre y dirección del autor corresponsal. En esta carta, podrán sugerirse nombres de miembros del Comité Científico que, a juicio de los autores, sean especialmente adecuados para juzgar el trabajo.

   La decisión final sobre aceptación del trabajo será precedida de un procedimiento de revisión anónima.

2. Las contribuciones serán preferiblemente de una longitud inferior a 24 páginas y se deberán ajustar al formato indicado en los ficheros a tal efecto disponibles en la página web de la Sociedad (`http://www.sema.org.es/`).

3. El contenido de los artículos publicados corresponderá a un área de trabajo preferiblemente conectada a los objetivos propios de la Matemática Aplicada. En los trabajos podrá incluirse información sobre resultados conocidos y/o previamente publicados. Se anima especialmente a los autores a presentar sus propios resultados (y en su caso los de otros investigadores) con estilo y objetivos divulgativos.

Ficha de Inscripción Individual

Sociedad Española de Matemática Aplicada S$\vec{e}$MA

Remitir a: Iñigo Arregui, Dpto de Matemáticas, Fac. de Informática,
Universidad de A Coruña. Campus de Elviña, s/n. 15071 A Coruña.
CIF: G-80581911

Datos Personales

- Apellidos: ....
- Nombre: ....
- Domicilio: ....
- C.P.: .... Población: ....
- Teléfono: .... DNI/CIF: ....
- Fecha de inscripción: ....

Datos Profesionales

- Departamento: ....
- Facultad o Escuela: ....
- Universidad o Institución: ....
- Domicilio: ....
- C.P.: .... Población: ....
- Teléfono: .... Fax: ....
- Correo electrónico: ....
- Página web: `http://` ....
- Categoría Profesional: ....
- Líneas de Investigación: ....

  ....

Dirección para la correspondencia: □ Profesional **□ Personal**

Cuota anual para el año 2008

□ Socio ordinario: 30€ □ Socio de reciprocidad con la RSME: 12€
□ Socio estudiante: 15€ □ Socio extranjero: 25€

Datos bancarios

. . . de . . . . de 200. .

Muy Sres. Míos:

Ruego a Uds. que los recibos que emitan a mi cargo en concepto de cuotas de inscripción y posteriores cuotas anuales de S$\vec{e}$MA (Sociedad Española de Matemática Aplicada) sean pasados al cobro en la cuenta cuyos datos figuran a continuación

Entidad (4 dígitos)	Oficina (4 dígitos)	D.C. (2 dígitos)	Número de cuenta (10 dígitos)

- Entidad bancaria: . . . .
- Domicilio: . . . .
- C.P.: . . . . Población: . . . .

Con esta fecha, doy instrucciones a dicha entidad bancaria para que obren en consecuencia.

Atentamente,

Fdo. . . . .

Para remitir a la entidad bancaria

. . . de . . . . de 200. .

Muy Sres. Míos:

Ruego a Uds. que los recibos que emitan a mi cargo en concepto de cuotas de inscripción y posteriores cuotas anuales de S$\vec{e}$MA (Sociedad Española de Matemática Aplicada) sean cargados a mi cuenta corriente/libreta . . . . en esa Agencia Urbana y transferidas a

SEMA: 0128 - 0380 - 03 - 0100034244
Bankinter
C/ Hernán Cortés, 63
39003 Santander

Atentamente,

Fdo. . . . .

Ficha de Inscripción Institucional

Sociedad Española de Matemática Aplicada S$\vec{e}$MA

Remitir a: Iñigo Arregui, Dpto de Matemáticas, Fac. de Informática,
Universidad de A Coruña. Campus de Elviña, s/n. 15071 A Coruña.
CIF: G-80581911

Datos de la Institución

- Departamento: ....
- Facultad o Escuela: ....
- Universidad o Institución: ....
- Domicilio: ....
- C.P.: .... Población: ....
- Teléfono: .... DNI/CIF: ....
- Correo electrónico: ....
- Página web: `http://` ....
- Fecha de inscripción: ....

Forma de pago

La cuota anual para el año 2008 como Socio Institucional es de 150€.
El pago se realiza mediante transferencia bancaria a

SEMA: 0128 - 0380 - 03 - 0100034244
Bankinter
C/ Hernán Cortés, 63
39003 Santander