SëMA

**BOLETÍN NÚMERO 44**

**Septiembre 2008**

# sumario

Editorial . . . . . 5

Artículos . . . . . 7

*Introducción al estudio de la ecuación de Euler de algunos funcionales del Cálculo de Variaciones*, por D. Arcoya, L. Boccardo 7

*The numerical solution of discontinuous IVPs by Runge–Kutta codes: A Review*, por M. Calvo, J.I. Montijano, L.Rández . . . . . 33

*Analysis of a cell system with finite divisions*, por B. Perthame, T.M. Touaoula . . . . . 55

*Sobre una interpolación no lineal: Aplicación al procesado de señales*, por S. Amat . . . . . 81

Educación Matemática . . . . . 101

*Un viaje matemático por el Espacio Europeo de Educación Superior*, por M.V. Cuevas, A. Nevot . . . . . 101

Resúmenes de tesis doctorales . . . . . 115

Anuncios . . . . . 117

**Boletín de la Sociedad Española de Matemática Aplicada** SëMA



**Página web de** SëMA
`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: Solución de la ecuación de ondas con el software *Elmer*.

Estimados socios,

A través de esta editorial queremos haceros partícipes de las últimas novedades de la Sociedad que han acaecido en el marco de la XIII Escuela Jacques-Louis Lions Hispano-Francesa sobre Simulación Numérica en Física e Ingeniería. Como sabéis ha tenido lugar en Valladolid en este mes de Septiembre, y durante su celebración ha tenido lugar la asamblea anual de SëMA, con la elección de presidente y tres miembros del comité ejecutivo. Podéis encontrar más información en la página web de la Sociedad.

Queremos también aprovechar para felicitar a los compañeros de Valladolid por el éxito en la organización de la Escuela, que ha contado con cursos y conferencias de gran interés, además de un buen número de participantes.

En cuanto a los contenidos de este nuevo número del boletín incluimos cuatro interesantes artículos de D. Arcoya y L. Boccardo; M. Calvo, J.I. Montijano y L. Rández; B. Perthame y T.M. Touaoula y S. Amat, además de un artículo de Educación Matemática sobre el Espacio Europeo de Educación Superior firmado por M.V. Cuevas y A. Nevot. Como siempre esperamos que lo disfrutéis.

Recibid un cordial saludo,

Grupo Editor
boletin.sema@uclm.es

*Bol. Soc. Esp. Mat. Apl.*
*nº 44(2008), 7–31*

ARTÍCULOS

# INTRODUCCIÓN AL ESTUDIO DE LA ECUACIÓN DE EULER DE ALGUNOS FUNCIONALES DEL CÁLCULO DE VARIACIONES

DAVID ARCOYA* Y LUCIO BOCCARDO**

*Departamento de Análisis Matemático
Universidad de Granada

**Dipartimento di Matematica
Università di Roma 1

darcoya@ugr.es boccardo@mat.uniroma1.it

## 1 Introducción

Este trabajo se corresponde con las notas de un curso de introducción al estudio de algunas cuestiones relativas a determinados funcionales del Cálculo de Variaciones. Fue impartido en abril de 1995 dentro del Ciclo de Conferencias del Departamento de Análisis Matemático de la Universidad de Granada y en Itla 2004 - Italian-Latin American Conference on Applied and Industrial Mathematics, Trujillo (Perú) del 15 al 18 de diciembre del 2005. En concreto, está pensado como una introducción a los teoremas clásicos de semicontinuidad, regularidad y teorema de paso de montaña de E. De Giorgi, G. Stampacchia y A. Ambrosetti - P.H. Rabinowitz. Complementariamente a estos resultados, se presentan algunos desarrollos más recientes de los mismos obtenidos en [3, 9]. El núcleo del curso está constituido por el estudio de determinados funcionales del Cálculo de Variaciones que aunque aparecen de una forma natural no son diferenciables. En contraste con la costumbre de los cursos básicos del Análisis Matemático, esta falta de diferenciabilidad no se debe a la aparición de funciones similares al valor absoluto. En nuestro caso, las funciones que aparecen son *suaves*. De hecho, podríamos suponerlas de clase $C^\infty$ y, aún así, el funcional correspondiente no será diferenciable más que en un subespacio de direcciones. Para este tipo de funcionales consideraremos aspectos relativos a sus posibles mínimos y, más generalmente, puntos críticos, así como a la ecuación de Euler asociada.

Fecha de recepción: 20/09/2007. Aceptado (en forma revisada): 30/01/2008.

## 2 Minimización de funcionales

Comenzamos con el problema de minimización de funcionales $\Phi$ definidos en un espacio de Banach $X$. El teorema básico lo constituye el Teorema de Weierstrass:

**Teorema 1** *Sea $A$ un espacio métrico compacto y $\Phi : A \longrightarrow \mathbb{R}$ un funcional semicontinuo inferiormente. Entonces $\Phi(A)$ tiene mínimo.* □

**Notas 3** i) Observemos que las dos hipótesis del teorema, esto es, la compacidad de $A$ y la semicontinuidad inferior de $\Phi$ van en direcciones opuestas: *"cuanto más sencillo es de verificar una, más díficil se hace la otra"*. En efecto, observemos que si $\tau, \tau'$ denotan distintas topologías en $A$ con $\tau \subset \tau'$, entonces se tiene

$$\Phi :(\ A, \tau) \longrightarrow \mathbb{R} \text{ s.c.i. } \Longrightarrow \Phi :(\ A, \tau') \longrightarrow \mathbb{R} \text{ s.c.i.}$$

$$(A, \tau) \text{ compacto } \Longleftarrow (A, \tau') \text{ compacto}$$

ii) Supongamos que $X$ es reflexivo y $\Phi : X \longrightarrow \mathbb{R}$ es coercivo (es decir, $\lim_{\|x\|_\infty \to \infty} \Phi(x) = \infty$). Entonces la coercividad permite reducir el problema de minimización en todo $X$ a minimizar en una bola cerrada $\overline{B}(0, R)$ centrada en 0 y de radio $R$ suficientemente grande. Recordemos que las bolas cerradas son débilmente compactas, pero que las funciones $\Phi$ suelen ser continuas en $(X, \|\cdot\|_X)$.

Recordemos que se llama epigráfo de $\Phi$ al conjunto

$$\text{Epi } \Phi = \{(x, t) \in X \times \mathbb{R} \ / \ \Phi(x) \leq t\}$$

y que se tiene

**Lema 2** *Si $\tau$ es cualquier topología en $X$ y $\Phi : X \longrightarrow \mathbb{R}$ entonces*

*i) Epi $\Phi$ es $\tau$-cerrado si y solamente si $\Phi$ es $\tau$-semicontinuo inferiormente.*

*ii) Epi $\Phi$ es convexo si y solamente si $\Phi$ es convexo.* □

**Teorema 3** *Sean $X$ un espacio de Banach reflexivo y $\Phi : X \longrightarrow \mathbb{R}$ un funcional coercivo, convexo y continuo.*[1] *Entonces existe $x_0 \in X$ tal que*

$$\Phi(x_0) = \text{mín}\{\Phi(x) \ / \ x \in X\}.$$

*Demostración.* Observemos:

1. Puesto que $\Phi$ es coercivo, entonces $\text{mín}\{\Phi(x) \ / \ x \in X\} = \text{mín}\{\Phi(x) \ / \ x \in B(0, R)\}$, para $R$ suficientemente grande.

[1] Si no escribimos nada se entiende que los adjetivos son en la toplogía fuerte

2. Por el lema precedente, la convexidad y continuidad de $\Phi$ implican que Epi $\Phi$ es cerrado y convexo, y así, como consecuencia del Teorema de Hahn-Banach, Epi $\Phi$ es débilmente cerrado; y otra vez por el lema anterior, $\Phi$ es d.i.s.c. (débil inferiormente semicontinua).

Por los puntos 1. y 2. se concluye el teorema a partir del Teorema de Weierstrass. □

A continuación mostramos algunos ejemplos simples que ilustran el teorema:

**Ejemplo 1** Sean $X = W_0^{1,p}(\Omega)$, $p > 1$, $g \in L^{p'}(\Omega)$,

$$\Phi(v) = \int_\Omega f(x, Dv)\, dx - \int_\Omega g(x)v\, dx, \quad v \in X.$$

Las siguientes condiciones son suficientes (y casi necesarias) para que se verifiquen las hipótesis del Teorema:

i) $f : \Omega \times \mathbb{R}^N \longrightarrow \mathbb{R}$ es de Carathéodory.[2]

ii) Existe $\alpha > 0$ tal que

$$\alpha|\xi|^p \leq f(x, \xi) \leq \beta(1 + |\xi|^p), \quad \text{a.e. } x \in \Omega, \quad \forall \xi \in \mathbb{R}^N.$$

iii) $\xi \mapsto f(x, \xi)$ es convexa.

En efecto, por el teorema de Nemickii, las condiciones i)-ii) implican que el operador (no lineal) $v \mapsto f(x, Dv)$ es continuo y acotado de $X$ en $L^1(\Omega)$. Además, $\Phi$ es coercivo y convexo por ii) y iii), respectivamente.

**Notas 4** 1. Algunos casos particulares de este ejemplo son:

(a) $f(x, \xi) = f(\xi) = \frac{1}{p}|\xi|^p + \frac{1}{q}|\xi|^q$, $1 < q < p$, $\xi \in \mathbb{R}^N$.

(b) $f(x, \xi) = M(x)\xi \cdot \xi$, $(x \in \Omega, \quad \xi \in \mathbb{R}^N)$, con $M(x)$ una matriz elíptica, simétrica y acotada (es decir, las hipótesis del Teorema de Lax-Milgram).

2. En el caso particular $f(\xi) = \frac{1}{p}|\xi|^p$, aprovechando la diferenciabilidad de $f$, puede probarse la d.s.c.i. de $\Phi$ de una forma simple sin usar el teorema de Hahn-Banach. En efecto, para $f(\xi) = \frac{1}{p}|\xi|^p$ tenemos

$$f(\xi) \geq f(\xi_0) + \nabla f(\xi_0) \cdot (\xi - \xi_0)$$

[2] es decir, la función $x \mapsto f(x, \xi)$ es medible en $\Omega$ para todo $\xi \in \mathbb{R}^N$, y la función $\xi \mapsto f(x, \xi)$ es continua a.e. $x \in \Omega$ ($\Omega$ es un conjunto abierto y acotado de $\mathbb{R}^N$)

y así

$$\int_\Omega f(Dv_n)\,dx \geq \int_\Omega f(Dv)\,dx + \int_\Omega |Dv|^{p-2}Dv\cdot(Dv_n - Dv)\,dx.$$

Por tanto, si $v_n \rightharpoonup v$ in $X = W_0^{1,p}(\Omega)$ entonces al ser $|Dv|^{p-2}Dv \in L^{p'}(\Omega)$

$$\lim_{n\to\infty}\int_\Omega |Dv|^{p-2}Dv\cdot(Dv_n - Dv)\,dx = 0$$

y

$$\liminf_{n\to\infty}\int_\Omega f(Dv_n)\,dx \geq \int_\Omega f(Dv)\,dx.$$

La condición iii) anterior sobre la convexidad de $f(x,\xi)$ en la variable $\xi$ es la más extraña. En el teorema siguiente analizamos con detalle ésta y veremos su relación con la d.s.c.i. No lo haremos en el caso general $\Phi(v) = \int_\Omega f(x,v,Dv)\,dx$ (Teorema de De Giorgi [12]); sino que consideramos un ejemplo simple que reúne las principales ideas del caso general:

$$X = L^p(\Omega), \quad \Phi(v) = \int_\Omega f(v)\,dx, \quad v \in X \tag{1}$$

con $f : \mathbb{R} \longrightarrow \mathbb{R}$ una función continua tal que $|f(t)| \leq \beta|t|^p$ para algún $\beta > 0$.

**Teorema 4 (Condición necesaria)** *Si el funcional* $\Phi$ *dado por* (1) *es d.s.c.i. entonces* $f$ *es convexa.*

**Notas 5** i) Así, la convexidad de $f$ es condición necesaria y suficiente para la convexidad de $\Phi$.

ii) Un problema interesante sería minimizar cuando el espacio no es reflexivo, por ejemplo, si $p = 1$.

Para la prueba necesitamos el siguiente lema que es, más o menos, el lema de Riemann-Lebesgue:

**Lema 5** *Supongamos que* $p : \mathbb{R} \longrightarrow \mathbb{R}$ *es una función* $(b-a)-$*periódica y acotada. Sea* $u_n(t) := p(nt)$, $t \in \mathbb{R}$. *Entonces*

$$u_n(t) \rightharpoonup \overline{p} \text{ en } L^p(a,b)$$

*y*

$$u_n(t) \overset{*}{\rightharpoonup} \overline{p} \text{ en } L^\infty(a,b).$$

*donde* $\overline{p}$ *denota la media de* $p$, *es decir,* $\overline{p} = \int_a^b p(s)\,ds$.

*Demostración.* Mediante la consideración de una conveniente traslación de $p$ se observa que no hay pérdida de generalidad en suponer que $p \geq 0$, $a = 0$ y $b = 1$. Por la densidad de las funciones simples en $L^{p'}(0,1)$ y teniendo en cuenta que

para cualesquiera $A, B \in \mathbb{R}$ con $A < B$ se tiene que la función característica $\xi_{(A,B)}$ se expresa como $\xi_{(A,B)} = \xi_{(0,B)} - \xi_{(0,A)}$, bastará probar que

$$\int_0^1 u_n(t)\chi_{(0,A)}(t)\,dt \longrightarrow \overline{p}\int_0^1 \chi_{(0,A)}(t)\,dt \quad \forall A \in (0,1),$$

o sea

$$\int_0^A u_n(t)\,dt \longrightarrow \overline{p}A, \quad \forall A \in (0,1).$$

Ahora bien para cualquier $A \in (0,1)$,

$$\int_0^A u_n(t)\,dt = \int_0^A p(nt)\,dt = \frac{1}{n}\int_0^{nA} p(y)\,dy,$$

y como $p \geq 0$ llegamos a

$$\frac{A}{nA}\int_0^{[nA]} p(y)\,dy \leq \int_0^A u_n(t)\,dt \leq \frac{A}{nA}\int_0^{[nA]+1} p(y)\,dy.$$

Por la 1-periódicidad de $p$, esto significa

$$\frac{A}{nA}[nA]\overline{p} \leq \int_0^A u_n(t)\,dt \leq \frac{A}{nA}\left([nA]+1\right)\overline{p}.$$

Finalmente, tomando límites en estas desigualdades deducimos

$$\lim_{n\to\infty}\int_0^A u_n(t)\,dt = \overline{p}A.$$

□

**Nota 1** En la prueba del teorema, usaremos el lema anterior para la función $1-$periódica $p = p_\lambda$ $(\lambda \in (0,1))$ dada en $[0,1)$ por

$$p_\lambda(t) = \begin{cases} a, & \text{si } 0 \leq t \leq \lambda \\ b, & \text{si } \lambda < t < 1 \end{cases}. \tag{2}$$

En este caso, la tesis del lema afirma:

$$u_n(t) = p_\lambda(nt) \rightharpoonup \lambda a + (1-\lambda)b \ \text{ en } L^p(0,1)$$

*Demostración del Teorema 4.* Elegimos $p_\lambda$ la función dada por (2). Entonces por la d.s.c.i. de $\Phi$, si $u_n(t) = p_\lambda(nt)$,

$$\begin{aligned} \Phi(\lambda a + (1-\lambda)b) &\leq \liminf_{n\to\infty}\Phi(u_n) \\ &= \liminf_{n\to\infty}\int_0^1 f(u_n(t)\,dt \\ &= \liminf_{n\to\infty}\int_0^1 f(p_\lambda(nt))\,dt. \end{aligned}$$

Pero, teniendo en cuenta el lema anterior:

$$\lim_{n\to\infty}\int_0^1 f(p_\lambda(nt))\,dt = \lim_{n\to\infty}\int_0^1 f(p_\lambda(nt))1\,dt = \lambda f(a) + (1-\lambda)f(b)$$

y en consecuencia:

$$\Phi(\lambda a + (1-\lambda)b) = f(\lambda a + (1-\lambda)b) \le \lambda f(a) + (1-\lambda)f(b).$$

□

**Notas 6** 1. Con una idea parecida a esta, construyendo sucesiones oscilantes cuya derivada en una dirección sea como $p(nt)$ es posible probar una condición necesaria en el marco del Ejemplo 1 (Teorema de De Giorgi).

2. El lema anterior también resulta útil para encontrar condiciones necesarias para la continuidad en las topologías débiles del operador de Nemickii asociado a una función $f$:

**Proposición 6** *Supongamos que $f : \mathbb{R} \longrightarrow \mathbb{R}$ es una función continua verificando $|f(t)| \le \beta_1|t| + \beta_2$, $\forall t \in \mathbb{R}$ ($\beta > 0$). Sea $T : L^p(0,1) \longrightarrow L^p(0,1)$ el operador de Nemickii asociado a $f$, es decir, $T(v) = f \circ v$, $\forall v \in L^p(0,1)$. Entonces $T$ es continuo de $L^p(0,1)$ con la topología débil en $L^p(0,1)$ con la topología débil si y solamente si $f$ es una recta afín.*

*Demostración.* En efecto, si consideramos la función $p_\lambda$ dada por (2), tenemos por el lema anterior que

$$u_n(t) = p_\lambda(nt) \rightharpoonup \lambda a + (1-\lambda)b \ \text{ en } L^p(0,1)$$

y

$$T(u_n) = f \circ u_n \rightharpoonup \lambda f(a) + (1-\lambda)f(b) \ \text{ en } L^p(0,1).$$

Así, $T$ es continuo en las topologías débiles si y solamente si

$$\lambda f(a) + (1-\lambda)f(b) = T(\lambda a + (1-\lambda)b) = f(\lambda a + (1-\lambda)b),$$

es decir, si y solamente si $f$ es afín. □

**Ejemplo 2** Sean $X = W_0^{1,p}(\Omega)$, $g \in L^{p'}(\Omega)$ y $a : \Omega \times \mathbb{R} \longrightarrow \mathbb{R}$ una función de Carathéodory verificando $0 < \alpha \le a(x,s) \le \beta$, a.e. $x \in \Omega, \forall s \in \mathbb{R}$. Sea $\Phi : X \longrightarrow \mathbb{R}$ el funcional definido por

$$\Phi(v) = \frac{1}{p}\int_\Omega a(x,v)|Dv|^p\,dx - \int_\Omega g(x)v\,dx, \quad v \in X.$$

Observemos que $X$ es reflexivo y el funcional $\Phi$ está bien definido (puesto que $a(x,s) \leq \beta$) y es coercivo (porque $a(x,s) \geq \alpha > 0$). Además, por el teorema de Nemickii, $\Phi$ es continuo.

Así, la pregunta es como conseguir la débil semicontinuidad inferior de $\Phi$. Para ello se pueden seguir dos posibles caminos. Uno consistiría en aplicar el Teorema 4, para lo cual precisamos probar que $\Phi$ es convexo. No obstante, no proporciona información más que en el caso $a(x,s) = a(x)$, siendo, por tanto, un subcaso del Ejemplo 1. En efecto, puede probarse:

$$\Phi \text{ es convexo } \iff a(x,s) = a(x).$$

El segundo método a seguir es más artesanal y se basa en una aplicación conveniente del teorema del valor medio:

Sea $\{v_n\} \rightharpoonup v$ en $X$. Si $f(\xi) = \frac{1}{p}|\xi|^p$, $\xi \in \mathbb{R}^N$, entonces como consecuencia del teorema del valor medio tenemos la desigualdad

$$f(\xi) \geq f(\xi_0) + \nabla f(\xi_0) \cdot (\xi - \xi_0), \quad \forall \xi, \xi_0 \in \mathbb{R}^N,$$

de donde se deduce

$$\begin{aligned} \int_\Omega \frac{1}{p} a(x, v_n)|Dv_n|^p \, dx \quad \geq \quad & \int_\Omega \frac{1}{p} a(x, v_n)|Dv|^p \, dx \\ & + \int_\Omega \frac{1}{p} a(x, v_n)|Dv|^{p-2} Dv \cdot D(v_n - v) \, dx. \end{aligned}$$

Para ver que $\Phi$ es d.s.c.i. bastara probar el siguiente resultado:

**Lema 7** *Supongamos que $a : \Omega \times \mathbb{R} \longrightarrow \mathbb{R}$ es una función de Carathéodory verificando $0 < \alpha \leq a(x,s) \leq \beta$ a.e. $x \in \Omega, \forall s \in \mathbb{R}$. Si $\{v_n\} \rightharpoonup v$ en $W_0^{1,p}(\Omega)$ entonces*

$$a(x, v_n)|Dv|^p \xrightarrow{L^1} a(x, v)|Dv|^p$$

$$a(x, v_n)|Dv|^{p-2} Dv \xrightarrow{L^{p'}} a(x, v)|Dv|^{p-2} Dv.$$

*Demostración.* Puesto que $W_0^{1,p}(\Omega)$ está inmerso compactamente en $L^p(\Omega)$, tendremos que $v_n$ converge en $L^p(\Omega)$ a $v$, por lo que también tendremos convergencia en medida de $v_n$ a $v$ y de $a(x, v_n)$ a $a(x, v)$. Además,

$$|a(x, v_n)|Dv|^p| \leq \beta |Dv|^p.$$

El teorema de la convergencia (*en medida*) dominada de Lebesgue concluye la prueba del lema. □

Como consecuencia directa de las observaciones anteriores obtenemos el teorema siguiente:

**Teorema 8** *Supongamos que* $g \in L^{p'}(\Omega)$ *y* $a : \Omega \times \mathbb{R} \longrightarrow \mathbb{R}$ *es una función de Carathéodory verificando* $0 < \alpha \leq a(x,s) \leq \beta$ *a.e.* $x \in \Omega, \forall s \in \mathbb{R}$. *Sea* $\Phi : X = W_0^{1,p}(\Omega) \longrightarrow \mathbb{R}$ *el funcional definido por*

$$\Phi(v) = \frac{1}{p}\int_\Omega a(x,v)|Dv|^p\,dx - \int_\Omega g(x)v\,dx, \quad v \in X.$$

*Entonces existe* $u \in X$ *tal que* $\Phi(u) = \min\{\Phi(v) \ / \ v \in X\}$.

## 7 Ecuación de Euler

A continuación estudiaremos cual es la ecuación de Euler asociada al problema de minimización del funcional $\Phi$ del ejemplo anterior cuando $p = 2$. Para ello observemos que si $u \in W_0^{1,2}(\Omega)$ verifica $\Phi(u) = \min\{\Phi(v) \ / \ v \in X\}$, entonces

$$\Phi(u) \leq \Phi(u + tv), \quad \forall v \in W_0^{1,2}(\Omega), \quad \forall t \in \mathbb{R}.$$

Así, considerando la función real

$$\phi(t) = \Phi(u + tv) = \frac{1}{2}\int_\Omega a(x, u + tv)|D(u + tv)|^2\,dx - \int_\Omega g(x)(u + tv)\,dx,$$

tendremos $\phi(0) \leq \phi(t)$, por lo que, si $\phi$ es diferenciable en $t = 0$, necesariamente será $\phi'(0) = 0$.

Ahora bien, si suponemos que existe $\frac{\partial}{\partial s}a(x,s)$ y que está acotada, es decir,

$$\left|\frac{\partial}{\partial s}a(x,s)\right| \leq \gamma, \quad \text{a.e. } x \in \Omega, \ \forall s \in \mathbb{R}$$

e **imponemos que** $v \in W_0^{1,2}(\Omega) \cap L^\infty(\Omega)$, entonces $\phi$ es diferenciable en $t = 0$ con

$$\phi'(0) = \int_\Omega a(x,u)Du \cdot Dv\,dx + \frac{1}{2}\int_\Omega a_s(x,u)|Du|^2 v\,dx - \int_\Omega g(x)v\,dx = 0.$$

Por tanto, la función $u$ es una solución débil del problema

$$\left.\begin{array}{c} -\mathrm{div}(a(x,u)Du) + \frac{1}{2}a_s(x,u)|Du|^2 = g(x), \quad x \in \Omega \\ u \in W_0^{1,2}(\Omega). \end{array}\right\}$$

Este problema de contorno es un caso particular del más general que sigue:

$$\left.\begin{array}{c} Q(u) + a_0(x,u)u = f(x) + H(x,u,Du), \quad x \in \Omega \\ u \in W_0^{1,2}(\Omega) \cap L^\infty(\Omega), \end{array}\right\} \tag{3}$$

con

- $Q(v) = -\text{div}\,(a(x,v)Dv)$, siendo $a(x,s)$ una matriz de Carathéodory verificando para ciertos $\alpha, \beta > 0$,

$$a(x,s)\xi \cdot \xi \geq \alpha|\xi|^2, \tag{4}$$

y

$$|a(x,s)| \leq \beta, \quad a.e.\ x \in \Omega, \quad \forall s \in \mathbb{R}, \quad \forall \xi \in \mathbb{R}^N. \tag{5}$$

- $a_0(x,s)$ es una función de Carathéodory verificando

$$0 < \alpha_0 \leq a_0(x,s) \leq \beta_0, \quad a.e.\ x \in \Omega, \quad \forall s \in \mathbb{R}. \tag{6}$$

- El dato $f$ verifica

$$f \in L^\infty(\Omega). \tag{7}$$

- $H(x,s,\xi)$ es una función de Carathéodory verificando

$$|H(x,s,\xi)| \leq \gamma|\xi|^2, \quad a.e.\ x \in \Omega, \quad \forall s \in \mathbb{R}, \quad \forall \xi \in \mathbb{R}^N. \tag{8}$$

Si , $\langle \cdot, \cdot \rangle$ denota la dualidad, por solución de (3) entendemos una función $u \in W_0^{1,2}(\Omega) \cap L^\infty(\Omega)$ que satisface

$$\langle Q(u), v\rangle + \int_\Omega a_0(x,u)uv\,dx = \int_\Omega fv\,dx + \int_\Omega H(x,u,Du)v\,dx,$$

para cualquier $v \in W_0^{1,2}(\Omega) \cap L^\infty(\Omega)$.

**Nota 2** En realidad, en lugar de trabajar para $f \in L^\infty(\Omega)$, podríamos hacerlo con términos más generales como son $f \in L^q(\Omega)$, $q > \frac{N}{2}$.

El caso $Q = -\Delta$, $a_0 \equiv 0$, $H(x,s,\xi) = f(s) + |\xi|^2$, esto es, el problema

$$\left.\begin{array}{c} -\Delta u = f(u) + |Du|^2, \quad x \in \Omega \\ u \in W_0^{1,2}(\Omega) \cap L^\infty(\Omega), \end{array}\right\} \tag{9}$$

fue estudiado por Kazdan y Kramer [13]. Entre otros probaron el siguiente resultado:

**Lema 9** *Si $\lambda_1$ denota el primer valor propio de $-\Delta$ con condiciones nulas de Dirichlet en el borde de $\Omega$ y la función $f$ verifica $f(u) > \lambda_1 u$, para todo $u \in \mathbb{R}$, entonces el problema* (9) *no tiene solución.*

*Demostración.* Razonamos por contradicción suponiendo que $u \in W_0^{1,2}(\Omega) \cap L^\infty(\Omega)$ es una solución de (9). Observemos que $u \geq 0$ puesto que $f \geq 0$. Si consideramos $v = e^u - 1 \in W_0^{1,2}(\Omega) \cap L^\infty(\Omega)$ podemos ver fácilmente que $v$ es solución del problema

$$\left.\begin{array}{c} -\Delta v = -f(v+1), \quad x \in \Omega \\ v \in W_0^{1,2}(\Omega) \cap L^\infty(\Omega). \end{array}\right\}$$

Tomando en este problema como función test una función propia $\varphi_1 > 0$ asociada a $\lambda_1$ llegamos a

$$\lambda_1 \int_\Omega v\varphi_1 \, dx = \int_\Omega f(v+1)\varphi_1 \, dx > \lambda_1 \int_\Omega (v+1)\varphi_1 \, dx,$$

de donde, se deduce $0 > \lambda_1 \int_\Omega \varphi_1 \, dx$, que evidentemente es contradictorio. □

**Teorema 10 (Boccardo, Murat y Puel, [9])** *Bajo las hipótesis* $(4)-(8)$ *anteriores, el problema* (3) *tiene al menos una solución.*

*Demostración.* Para estudiar (3) consideraremos los problemas aproximados

$$\left.\begin{array}{c} Q(u_n) + a_0(x,u_n)u_n = f(x) + H_n(x,u_n,Du_n), \quad x \in \Omega \\ u_n \in W_0^{1,2}(\Omega), \end{array}\right\} \tag{10}$$

siendo la función $H_n$ la dada por

$$H_n(x,s,\xi) = \frac{h(x,s,\xi)}{1+\frac{|\xi|^2}{n}}, \quad x \in \Omega, s \in \mathbb{R}, \xi \in \mathbb{R}^N.$$

Como $H_n$ es acotada!, no será díficil probar la existencia de solución $u_n$ de (10). Unas estimas a priori apropiadas nos van a permitir pasar al límite y resolver (3). Señalemos de antemano que la estima

$$|H_n(x,s,\xi)| \le \frac{\gamma|\xi|^2}{1+\frac{|\xi|^2}{n}} \le n\gamma$$

no parece muy *inteligente* pués depende de $n$.
Concretando ya, nuestro proceso de demostración será el siguiente:

Paso 1. Existencia de solución $u_n$ de (10).
Paso 2. $L^\infty$-acotación de la solución $u_n$.
Paso 3. $L^\infty$-acotación de la sucesión $\{u_n\}$.
Paso 4. $W_0^{1,2}$-acotación de la sucesión $\{u_n\}$.
Paso 5. Compacidad relativa de la sucesión $\{u_n\}$ en $W_0^{1,2}(\Omega)$ .
Paso 6. Existencia de solución de (3).

Pasos 1. y 2. Los Pasos 1. y 2. son consecuencia directa del siguiente lema:

**Lema 11** *Si* $B(x,s,\xi)$ *es una función de Carathéodory acotada, i.e. verificando*

$$|B(x,s,\xi)| \le \delta, \quad x \in \Omega, \ s \in \mathbb{R}, \ \xi \in \mathbb{R}^N$$

*para algún $\delta > 0$. Entonces existe $z \in W_0^{1,2}(\Omega)$ tal que*

$$Q(z) + a_0(x,z)z = f(x) + B(x,z,Dz), \quad x \in \Omega.$$

*Además $z \in L^\infty(\Omega)$.*

*Demostración.* Observemos que por el Teorema de Lax-Milgram, dada $v \in W_0^{1,2}(\Omega)$ existe una única solución w del problema

$$\left.\begin{array}{c} -\text{div } (a(x,v)Dw) + a_0(x,v)w = f(x) + B(x,v,Dv), \quad x \in \Omega \\ w \in W_0^{1,2}(\Omega), \end{array}\right\}$$

Además, por ser $B$ acotada, existe $R > 0$, independiente de $v \in W_0^{1,2}(\Omega)$, tal que $\|w\|_{W_0^{1,2}} \leq R$, $\forall v \in W_0^{1,2}(\Omega)$.

Así, la aplicación $S : W_0^{1,2}(\Omega) \longrightarrow W_0^{1,2}(\Omega)$ definida mediante $S(v) = w$, $(v \in W_0^{1,2}(\Omega))$ es continua.

La prueba de la primera parte del lema se concluye, via el Teorema del punto fijo de Schauder, si probamos que $S$ es compacta. Ahora bien, observemos que si $\rho > 0$ es cualquier positivo entonces el conjunto

$$\Big\{ f(x) + B(x,v,Dv) \ / \ \|v\|_{W_0^{1,2}} \leq \rho \Big\}$$

es $L^\infty-$acotado y, por tanto, compacto en $W^{-1,2}(\Omega)$. Consecuentemente, el conjunto

$$\Big\{ w = S(v) \ / \ \|v\|_{W_0^{1,2}} \leq \rho \Big\}$$

es compacto en $W_0^{1,2}(\Omega)$.

Para demostrar que $z \in L^\infty(\Omega)$ consideremos como función test $(z-k)^+$ $(k > 0)$. Entonces, si denotamos $\widetilde{f}(x) = f(x) + B(x,z,Dz) \in L^\infty(\Omega)$ tenemos

$$\int_\Omega a(x,z)Dz \cdot D(z-k)^+ \, dx + \int_\Omega a_0(x,z)z(z-k)^+ \, dx = \int_\Omega \widetilde{f}(z-k)^+ \, dx,$$

es decir,

$$\begin{aligned} 0 &\leq \int_\Omega a(x,z)|D(z-k)^+|^2 \, dx + \int_\Omega a_0(x,z)|(z-k)^+|^2 \, dx \\ &= \int_\Omega \Big[\widetilde{f} - a_0(x,z)k\Big] (z-k)^+ \, dx \\ &\leq \int_\Omega \Big[\widetilde{f} - \alpha_0 k\Big] (z-k)^+ \, dx. \end{aligned}$$

Notemos que si escogemos como $k$ una cota superior de $\frac{\widetilde{f}}{\alpha_0}$ entonces la positividad del ultimo miembro de estas desigualdades implica $(z-k)^+=0$, o sea, $z\leq k$. Así $z\leq \frac{\|\widetilde{f}\|_\infty}{\alpha_0}$. De una forma análoga se prueba también que $z\geq -\frac{\|\widetilde{f}\|_\infty}{\alpha_0}$ y por tanto:

$$\|z\|_\infty \leq \frac{\|\widetilde{f}\|_\infty}{\alpha_0} = K(\alpha_0, \|f\|_\infty, \|B\|_\infty).$$

Paso 3. En esta etapa y las siguientes vamos a considerar para $\lambda\in\mathbb{R}$ la función real $\varphi(t)=te^{\lambda t^2}$ $(t\in\mathbb{R})$. Escogeremos $\lambda>0$ de tal forma que

$$\alpha\varphi'(t)-\gamma|\varphi(t)|\geq\frac{\alpha}{2},\quad \forall t\in\mathbb{R}.$$

Por el Paso 2 podemos tomar $\varphi((u_n-k)^+)$ como función test en (10) y deducimos:

$$\int_\Omega a(x,u_n)D((u_n-k)^+)\cdot D((u_n-k)^+)\varphi'((u_n-k)^+)\,dx+$$

$$+\int_\Omega a_0(x,u_n)u_n\varphi((u_n-k)^+)\,dx=$$

$$=\int_\Omega f(x)\varphi((u_n-k)^+)\,dx+\int_\Omega H(x,u_n,Du_n)\varphi((u_n-k)^+)\,dx,$$

de donde,

$$\int_\Omega a(x,u_n)D((u_n-k)^+)\cdot D((u_n-k)^+)\varphi'((u_n-k)^+)\,dx+$$

$$+\int_\Omega a_0(x,u_n)(u_n-k)^+\varphi((u_n-k)^+)\,dx=$$

$$=\int_\Omega [f(x)-a_0(x,u_n)k]\,\varphi((u_n-k)^+)\,dx+\int_\Omega H(x,u_n,Du_n)\varphi((u_n-k)^+)\,dx\leq$$

$$\leq\int_\Omega [f(x)-a_0(x,u_n)k]\,\varphi((u_n-k)^+)\,dx+\gamma\int_\Omega |D((u_n-k)^+)|^2|\varphi((u_n-k)^+)|\,dx$$

en virtud de la desigualdades $|H_n(x,s,\xi)|\leq|H(x,s,\xi)|\leq\gamma|\xi|^2$, $(x\in\Omega,\ s\in\mathbb{R}$ y $\xi\in\mathbb{R}^N)$.

En consecuencia, por la elección de $\lambda$,

$$\frac{\alpha}{2}\int_\Omega |D((u_n-k)^+)|^2\,dx\leq$$

$$\leq\int_\Omega |D((u_n-k)^+)|^2\left\{\alpha\varphi'((u_n-k)^+)-\gamma|\varphi((u_n-k)^+)|\right\}\,dx\leq$$

$$\leq \int_\Omega [f(x) - \alpha_0 k]\, \varphi((u_n - k)^+)\, dx.$$

Un argumento similar al hecho en la demostración de la segunda parte del lema anterior nos permite deducir $u_n \leq \frac{\|f\|_\infty}{\alpha_0}$. Análogamente, se prueba $u_n \geq -\frac{\|f\|_\infty}{\alpha_0}$ y así,

$$\|u_n\|_\infty \leq \frac{\|f\|_\infty}{\alpha_0} \quad \forall n \in \mathbb{N}$$

quedando, por tanto, probado el Paso 3.

Paso 4. El paso anterior nos permite considerar $\varphi(u_n)$ como función test. De esta forma,

$$\int_\Omega a(x,u_n) Du_n \cdot Du_n \varphi'(u_n)\, dx + \int_\Omega a_0(x,u_n) u_n \varphi(u_n)\, dx =$$

$$= \int_\Omega f(x)\varphi(u_n)\, dx + \int_\Omega H_n(x,u_n,Du_n)\varphi(u_n)\, dx \leq$$

$$\leq C_3 + \gamma \int_\Omega |Du_n|^2 |\varphi(u_n)|\, dx.$$

La elípticidad de $Q$ y la elección de $\lambda$ implican:

$$\frac{\alpha}{2} \int_\Omega |Du_n|^2\, dx \leq \int_\Omega |Du_n|^2 \left\{\alpha\varphi'(u_n) - \gamma|\varphi(u_n)|\right\}\, dx \leq C_3$$

o sea, $\|u_n\|_{W_0^{1,2}} \leq \frac{2C_3}{\alpha}$ y queda probado el paso 4.

Paso 5. Por el Paso 4. podemos, pasando a una subsucesión, suponer las siguientes convergencias:

$$\{u_n\} \overset{*L^\infty}{\searrow} u,$$

$$\{u_n\} \overset{W_0^{1,2}}{\searrow} u,$$

$$\{u_n\} \overset{L^2}{\longrightarrow} u,$$

$$\{u_n(x)\} \longrightarrow u(x), \text{ a.e. } x \in \Omega.$$

Sin embargo, esto no es suficiente para concluir. En el caso semilineal $(a(x,s) = a(x))$ se consigue mejorar la convergencia considerando $u_n - u$ como función test. En nuestro caso, tomamos $\varphi(u_n - u)$ como función test y deducimos

$$\int_\Omega a(x,u_n) Du_n \cdot D(u_n - u)\varphi'(u_n - u)\, dx + \int_\Omega a_0(x,u_n) u_n \varphi(u_n - u)\, dx =$$

$$= \int_\Omega f(x)\varphi(u_n - u)\,dx + \int_\Omega H_n(x, u_n, Du_n)\varphi(u_n - u)\,dx,$$

de donde, por las convergencias anteriores,

$$\int_\Omega a(x, u_n)D(u_n - u)\cdot D(u_n - u)\varphi'(u_n - u)\,dx =$$

$$= -\int_\Omega a_0(x, u_n)u_n\varphi(u_n - u)\,dx + \int_\Omega f(x)\varphi(u_n - u)\,dx +$$

$$+ \int_\Omega H_n(x, u_n, Du_n)\varphi(u_n - u)\,dx - \int_\Omega a(x, u_n)Du\cdot D(u_n - u)\varphi'(u_n - u)\,dx =$$

$$= \varepsilon_n + \int_\Omega H_n(x, u_n, Du_n)\varphi(u_n - u)\,dx$$

siendo $\{\varepsilon_n\}$ una sucesión real convergente a cero. Así, por la elípticidad de $Q$, llegamos a la desigualdad

$$\alpha\int_\Omega |D(u_n - u)|^2\varphi'(u_n - u)\,dx \le \varepsilon_n + \gamma\int_\Omega |Du_n|^2|\varphi(u_n - u)|\,dx.$$

Ahora, teniendo en cuenta que

$$|Du_n|^2 \le \left(|Du_n - Du| + |Du|\right)^2 \le 2|Du_n - Du|^2 + 2|Du|^2,$$

obtenemos

$$\alpha\int_\Omega |D(u_n - u)|^2\varphi'(u_n - u)\,dx \le$$

$$\le \varepsilon_n + 2\gamma\int_\Omega |D(u_n - u)|^2|\varphi(u_n - u)|\,dx + 2\gamma\int_\Omega |Du|^2|\varphi(u_n - u)|\,dx.$$

Como, por el teorema de la convergencia dominada de Lebesgue, el último sumando del segundo miembro de esta desigualdad es convergente a cero, concluimos

$$\int_\Omega |D(u_n - u)|^2\left[\alpha\varphi'(u_n - u) - 2\gamma|\varphi(u_n - u)|\right]\,dx \le \varepsilon_n' \longrightarrow 0$$

y escogiendo $\lambda > 0$ tal que $\alpha\varphi'(t) - 2\gamma|\varphi(t)| \ge \frac{\alpha}{2}\ \forall t \in \mathbb{R}$ tendremos que

$$\int_\Omega |D(u_n - u)|^2\,dx \le \varepsilon_n' \longrightarrow 0$$

y, por tanto, $\{u_n\} \longrightarrow u$ en $W_0^{1,2}(\Omega)$, que prueba el Paso 5.

Paso 6. Observemos que $u_n$ verifican

$$\int_\Omega a(x, u_n)Du_n\cdot Dv\,dx + \int_\Omega a_0(x, u_n)u_n v\,dx =$$

$$= \int_\Omega f(x)v\,dx + \int_\Omega H_n(x,u_n,Du_n)v\,dx, \quad \forall v \in W_0^{1,2}(\Omega)\cap L^\infty(\Omega).$$

El paso anterior nos permite pasar al límite cuando $n$ tiende a $\infty$ en los tres primeros sumandos de esta desigualdad. Respecto del cuarto, observemos que se tiene la *dominación*

$$|H_n(x,u_n(x),Du_n(x))| \le \gamma |Du_n(x)|^2, \text{ a.e. } x\in\Omega, \quad \forall n\in\mathbb{N},$$

y así por el teorema de la convergencia dominada (por una sucesión convergente en $L^1(\Omega)$[3]) deducimos la convergencia del cuarto sumando por lo que

$$\int_\Omega a(x,u)Du\cdot Dv\,dx + \int_\Omega a_0(x,u)uv\,dx = \int_\Omega f(x)v\,dx + \int_\Omega H(x,u,Du)v\,dx$$

para cualesquiera $v \in W_0^{1,2}(\Omega)\cap L^\infty(\Omega)$. Esto concluye el Paso 6. y en consecuencia la demostración. □

**Nota 3** La demostración de la convergencia del cuarto sumando en el Paso 6. podría haberse hecho de otras dos formas distintas: o bien usando el teorema de Vitali, o bien usando una observación que puede encontrarse en el libro de Brezis [10] sobre la posibilidad de extraer una subsucesión dominada por una función de $L^2(\Omega)$ de cualquier sucesión convergente de funciones en $L^2(\Omega)$.

A continuación estudiamos diversos aspectos relativos a funcionales del tipo

$$\Phi(v) = \frac{1}{2}\int_\Omega a(x,v)|Dv|^2\,dx - \frac{1}{\theta+1}\int_\Omega \rho(x)|v|^{\theta+1}\,dx, \quad v\in W_0^{1,2}(\Omega),$$

con $a:\Omega\times\mathbb{R}\longrightarrow\mathbb{R}$ una función de Carathéodory, $0<\theta<1$, $\rho:\Omega\longrightarrow\mathbb{R}$ verificando

$$0<\alpha\le a(x,s)\le\beta, \quad \text{a.e. } x\in\Omega, \quad \forall s\in\mathbb{R} \tag{11}$$

$$0<\rho_0\le\rho(x)\le R_0, \quad \text{a.e. } x\in\Omega.$$

**Teorema 12** *Bajo las hipótesis anteriores, existe $u\in W_0^{1,2}-\{0\}$ tal que*

$$\Phi(u) = \min\{\Phi(v) \ / \ v\in W_0^{1,2}\}.$$

*Además, si la función $a(x,s)$ tiene derivada $a_s(x,s)$ respecto de la variable $s$ con*

$$|a_s(x,s)|\le\gamma, \quad \textit{a.e. } x\in\Omega, \quad \forall s\in\mathbb{R} \tag{12}$$

*entonces $u$ satisface*

$$\int_\Omega a(x,u)Du\cdot Dv\,dx + \frac{1}{2}\int_\Omega a_s(x,u)|Du|^2v\,dx = \int_\Omega \rho(x)|u|^{\theta-1}uv\,dx. \tag{13}$$

[3]Se puede demostrar esta versión con la misma idea clásica de demostración del Teorema de la convergencia dominada a partir del Lema de Fatou.

*Demostración.* Por el Ejemplo 1 sabemos que el funcional

$$v \mapsto \frac{1}{2}\int_\Omega a(x,v)|Dv|^2\,dx$$

es d.i.s.c. El teorema de Rellich garantiza la continuidad del funcional $v \mapsto \frac{1}{\theta+1}\int_\Omega \rho(x)|v|^{\theta+1}\,dx$, y por tanto, $\Phi$ es d.i.s.c.

De otra parte observemos que, por la desigualdad de Holder, para cualquier $v \in W_0^{1,2}(\Omega)$

$$\frac{1}{2}\int_\Omega a(x,v)|Dv|^2\,dx - \frac{1}{\theta+1}\int_\Omega \rho(x)|v|^{\theta+1}\,dx$$

$$\geq \frac{\alpha}{2}\int_\Omega |Dv|^2\,dx - \frac{R_0}{\theta+1}\int_\Omega |v|^{\theta+1}\,dx \geq$$

$$\geq \frac{\alpha}{2}\int_\Omega |Dv|^2\,dx - \frac{R_0}{\theta+1}\left(\int_\Omega |v|^{2^*}\,dx\right)^{\frac{\theta+1}{2^*}} (\text{ meas } \Omega)^{1-\frac{\theta+1}{2^*}}$$

$$\geq \frac{\alpha}{2}\int_\Omega |Dv|^2\,dx - C_1\left(\int_\Omega |v|^{2^*}\,dx\right)^{\frac{\theta+1}{2^*}},$$

y así $\Phi$ es coercivo.

Por el Teorema 1 deducimos que existe $u \in W_0^{1,2}(\Omega)$ tal que

$$\Phi(u) = \text{mín}\{\Phi(v) \ / \ v \in W_0^{1,2}(\Omega)\}.$$

Para probar que $u \not\equiv 0$ basta observar que para $t > 0$ suficientemente pequeño y $\varphi_1 > 0$ denotando una función propia asociada al primer valor propio del operador de Laplace con condición nula de Dirichlet en la frontera, se tiene

$$\Phi(u) \leq \Phi(t\varphi_1) \leq \frac{\beta}{2}t^2\lambda_1\int_\Omega \varphi_1^2\,dx - \frac{t^{\theta+1}}{\theta+1}\int_\Omega \rho(x)|\varphi_1|^{\theta+1}\,dx < 0 = \Phi(0).$$

Para probar la segunda parte del Teorema, observemos que si definimos para $v \in W_0^{1,2}(\Omega) \cap L^\infty(\Omega)$ fija, la función real $\Psi(t) = \Phi(u+tv)$ , $t \geq 0$, entonces $\Psi$ tiene un mínimo global en $t = 0$, por lo que $\Psi'(0) = 0$, es decir, $u$ verifica (13). □

La regularidad de la solución puede ser estudiada. De hecho, como consecuencia de la técnica de Stampacchia [14] se puede probar el siguiente resultado.

**Teorema 13** *Si además de las hipótesis anteriores suponemos que*

$$a_s(x,s)s \geq 0, \ \ a.e.\ x \in \Omega, \quad s \in \mathbb{R}, \tag{14}$$

*entonces* $u \in L^\infty(\Omega)$. □

También resulta interesante estudiar el caso de una función $\rho$ que cambia de signo. A este respecto tenemos el resultado siguiente para el caso $a(x,s) \equiv 1$:

**Teorema 14** *Supongamos que* $0 < \theta < 1$ *y* $\rho \in L^\infty(\Omega)$ *es una función para la que existe* $\omega \subset\subset \Omega$ *tal que* $\rho(x) \geq \rho_0 > 0$, *a.e.* $x \in \omega$. *Entonces existe una solución* $u \not\equiv 0$ *del problema*

$$\left.\begin{array}{c} -\Delta u = \rho(x) u^\theta, x \in \Omega \\ u \in W_0^{1,2}(\Omega), u \geq 0 \end{array}\right\}$$

*Demostración.* Usaremos la técnica de sub-supersoluciones. Para construir la supersolución consideramos la única solución $\psi$ del problema lineal

$$\left.\begin{array}{c} -\Delta \psi = \|\rho^+\|_\infty, x \in \Omega \\ \psi \in W_0^{1,2}(\Omega), \end{array}\right\}$$

Por el principio del Máximo, $\psi \geq 0$ y $\|\psi\|_\infty \leq c_0 \|\rho^+\|_\infty$.

Afirmamos que si $T > 0$ es suficientemente grande, entonces $\overline{u} = T\psi$ es una supersolución. En efecto, para $T$ grande se tiene

$$T\|\rho^+\|_\infty \geq T^\theta \|\rho\|_\infty^{1+\theta} \rho_0^\theta \geq \rho^+(x)(T\psi)^\theta \geq \rho(x)(T\psi)^\theta$$

y así

$$-\Delta \overline{u} = T\|\rho^+\|_\infty \geq \rho(x)(T\psi)^\theta.$$

Para conseguir la subsolución, tomamos una función propia positiva $\varphi_{1,\omega} : \omega \longrightarrow \mathbb{R}$ asociada al primer valor propio $\lambda_1^\omega$ del operador de Laplace en $\omega$ con condiciones nulas de Dirichlet en el borde. Sea $\overline{\varphi}_{1,\omega} : \Omega \longrightarrow \mathbb{R}$ la extensión cero fuera de $\omega$ de $\varphi_{1,\omega}$. Se tiene:

**Lema 15 (G. Stampacchia [14]** (véase también [7]))

$$-\Delta \overline{\varphi}_{1,\omega} \leq \lambda_1^\omega \overline{\varphi}_{1,\omega}, \ x \in \Omega.$$

□

Usando este lema es fácil comprobar que si $t > 0$ entonces $\underline{u} \equiv t\overline{\varphi}_{1,\omega}$ es una subsolución con $\underline{u} \leq \overline{u}$. La conclusión del teorema es ya estándar. □

## 8 Funcionales con términos supercuadráticos

En el Teorema 12 estudiamos para $0 < \theta < 1$ el problema de minimización del funcional

$$\Phi(v) = \frac{1}{2}\int_\Omega a(x,v)|Dv|^2\,dx - \frac{1}{\theta+1}\int_\Omega \rho(x)|v|^{\theta+1}\,dx, \quad v \in W_0^{1,2}(\Omega).$$

En esta sección, en lugar del término subcuadrático $\rho(x)|v|^{\theta+1}$, $(\theta + 1 < 2)$, pondremos un término supercuadrático. En el siguiente teorema, aprovecharemos la homogeneidad para deducir la existencia de solución de un problema con una no linealidad superlineal. Concretamente estudiamos para $q \in (1, 2^*)$, $q \neq 2$,

$$\left.\begin{array}{c} -\Delta u = |u|^{q-2}u, \quad x \in \Omega \\ u = 0, \quad x \in \partial\Omega \end{array}\right\} \tag{15}$$

**Teorema 16** *Sea $q \in (1, 2^*)$, $q \neq 2$. Entonces* (15) *tiene al menos una solución no nula.*

**Demostración**. Consideremos el funcional $\Phi$ definido en $W_0^{1,2}(\Omega)$ mediante

$$\Phi(v) = \frac{1}{2}\int_\Omega |Dv|^2\, dx$$

Observemos que, trivialmente, $\Phi$ es d.i.s.c. Planteamos el problema de minimización condicionada

$$\text{mín}\ \left\{\ \Phi(v)\ /\ v \in W_0^{1,2}(\Omega),\ \int_\Omega |v|^q\, dx = 1\ \right\}$$

Por el teorema de Rellich se sigue la existencia de una solución $\overline{u}$ de este problema. Además por el teorema de los multiplicadores de Lagrange, $\overline{u}$ verifica

$$-\Delta\overline{u} = \lambda|\overline{u}|^{q-2}\overline{u},$$

para algún $\lambda \neq 0$. Finalmente, por la homogeneidad, $u = \lambda^{\frac{-1}{q-2}}\overline{u}$ es la solución buscada. □

**Notas 9** i) En el caso $q = 2^*$, se tiene el famoso resultado de Pohozaev sobre la no existencia de solución positiva de (15). Posteriormente, Brézis y Nirenberg [11] demostraron la existencia de solución positiva del problema

$$\left.\begin{array}{c} -\Delta u = \lambda u + |u|^{2^*-1}, \quad x \in \Omega \\ u = 0, \quad x \in \partial\Omega \end{array}\right\}$$

para ciertos $\lambda > 0$.

ii) En el trabajo de L. Boccardo, M. Escobedo e I. Peral [8] se estudia la existencia de solución del problema

$$\left.\begin{array}{c} -\Delta u = \lambda u^\theta + |u|^{2^*-1}, \quad x \in \Omega \\ u \geq 0,\ u \in W_0^{1,2}(\Omega) \end{array}\right\} \tag{16}$$

con $0 < \theta < 1$. Concretamente, se tiene:

**Teorema 17** *Si $0 < \theta < 1$, entonces existe $\lambda^* > 0$ tal que el problema* (16) *tiene al menos una solución no cero para cualquier $\lambda \in (0, \lambda^*)$.*

**Demostración**. Usaremos la técnica de sub y supersoluciones. Empezaremos discutiendo la existencia de una supersolución: Sea $\Psi$ la única solución del problema

$$\left.\begin{array}{c} -\Delta\Psi = \lambda, \quad x \in \Omega \\ \Psi \in W_0^{1,2}(\Omega) \end{array}\right\}$$

Es bien sabido que $0 < \Psi < c_0\lambda$ para alguna constante $c_0$. Observemos que existe $\lambda^* > 0$ tal que si $\lambda \in (0, \lambda^*)$ entonces es posible elegir $T > 0$ tal que

$$c_0^\theta T^{\theta-1}\lambda^\theta + c_0^{2^*} T^{2^*-1}\lambda^{2^*} \leq 1.$$

Entonces $\overline{u} \equiv T\Psi$ es una supersolución ya que

$$-\Delta\overline{u} = T\Psi \geq \lambda(T\Psi)^\theta + (T\Psi)^{2^*}.$$

De otra parte es fácil comprobar que si $t > 0$ es suficientemente pequeño, entonces $\underline{u} \equiv t\varphi_1$ es una subsolución con $0 < \underline{u} < \overline{u}$. □

**Notas 10** i) Usando la misma técnica de demostración, se puede sustituir el término $u^{2^*-1}$ por un término del tipo $u^s$ con $s > 2^* - 1$. Obsérvese que en este caso $v^s$ no tiene sentido para todo $v \in W_0^{1,2}(\Omega)$, pero, puesto que, en nuestro caso, tenemos $t\varphi_1 \leq u \leq T\Psi$, deducimos $u \in L^\infty(\Omega)$ y $u^s$ tiene sentido.

ii) El resultado del teorema es valido también para operadores homogéneos mas generales, por ejemplo, para el $p$-laplaciano.

iii) A. Ambrosetti, H. Brézis y G. Cerami [1] estudiaron posteriormente la multiplicidad de solución para este problema con $p = 2$.

## 11 Puntos críticos via el Teorema de Paso de Montaña: El problema de la compacidad

En esta última sección volvemos nuestra atención al estudio de la existencia de puntos críticos de funcionales

$$\Phi(v) = \frac{1}{2}\int_\Omega a(x,v)|Dv|^2\,dx - \int_\Omega F(v)\,dx, \quad v \in W_0^{1,2}(\Omega), \tag{17}$$

donde la función $a(x,s)$ cumple (11), (12) y (14) y la función $F(s) = \int_0^s f(t)\,dt$ con $f$ continua, superlineal con crecimiento subcrítico y no necesariamente homogénea:

$$\lim_{u\to+\infty} \frac{f(u)}{u} = +\infty \tag{18}$$

y,

$$\text{si } N \geq 2 \text{ entonces } \limsup_{u\to+\infty} \frac{|f(u)|}{|u|^{2^*-1}} < +\infty \tag{19}$$

donde $2^* = 2N/(N-2)$ si $N \geq 3$ y $2^* = +\infty$ si $N = 2$. Ya que estaremos interesados en puntos críticos $u \geq 0$, impondremos que $f \equiv 0$ en $(-\infty, 0]$. Empezamos recordando el caso semilineal, esto es, cuando $a(x,s) \equiv 1$. En este caso, los puntos críticos del funcional se corresponden con las soluciones del problema

$$\left.\begin{array}{l} -\Delta u = f(u), \quad x \in \Omega \\ u \geq 0, \ u \in W_0^{1,2}(\Omega) \end{array}\right\}$$

y para probar la existencia de al menos una solución no trivial se puede seguir el método de Ambrosetti y Rabinowitz [2]. Este se basa en la aplicación del Teorema de Paso de Montaña al funcional

$$\Phi(v) = \frac{1}{2}\int_\Omega |Dv|^2\,dx - \int_\Omega F(v)\,dx, \quad v \in X = W_0^{1,2}(\Omega).$$

Las hipótesis de dicho teorema son una mezcla de condiciones geométricas y de compacidad del funcional. Respecto de las primeras se impone que el funcional presente un mínimo local estricto en $v = 0$ que no sea un mínimo absoluto, esto es:

- (Mínimo local estricto) Existen $r > 0$ y $\delta > 0$ tales que

$$\Phi(v) \geq 0 = \Phi(0), \quad \forall \|v\| \leq r$$

  y

$$\Phi(v) \geq \delta, \quad \forall \|v\| = r$$

- (No mínimo global) Existe $w \in X$ (con $\|e\| > r$) tal que $\Phi(w) < 0$.

La comprobación de estas propiedades no es difícil. Concretamente, la verificación de la primera requiere que, además de la hipótesis de subcriticalidad (19), se tenga que

$$\lim_{u \to 0^+} \frac{f(u)}{u} = 0. \tag{20}$$

Ciertamente, bajo estas hipótesis es posible encontrar constantes positivas $C$ y $\varepsilon \in (0, \lambda_1)$ ($\lambda_1$ el primer valor propio positivo del operador de Laplace con condiciones nulas en el borde) de forma que

$$F(s) \leq \frac{\varepsilon}{2}s^2 + Cs^{2^*}, \quad \forall s \geq 0.$$

Por tanto,

$$\Phi(v) \geq \int_\Omega |Dv|^2 - \int_\Omega \frac{\varepsilon}{2}v^2 - C\int_\Omega v^{2^*}, \quad \forall v \in W_0^{1,2}(\Omega).$$

La inmersión continua de $W_0^{1,2}(\Omega) \hookrightarrow L^{2^*}(\Omega)$ y la caracterización variacional de $\lambda_1$ implican entonces que $v = 0$ es un mínimo local estricto de $\Phi$.

Respecto de la segunda hipótesis, basta observar que la condición de superlinealidad (18) nos da que

$$\lim_{t\to+\infty} \Phi(tv) = \lim_{t\to+\infty} t^2 \left[ \int_\Omega |Dv|^2 - \int_\Omega \frac{F(tv)}{t^2} \right] = -\infty$$

para todo $v \in W_0^{1,2}(\Omega)$.

En el Teorema de Paso de Montaña también juega un papel muy importante la llamada condición de compacidad de Palais-Smale:

$$\left.\begin{array}{c} |\Phi(u_n)| \le \text{Const.} \\ \\ \Phi'(u_n) \overset{W^{-1,2}}{\longrightarrow} 0 \end{array}\right\} \Longrightarrow \exists u_{n_k} \overset{W_0^{1,2}}{\longrightarrow} u.$$

Para la verificación de esta condición aparece de una forma natural la hipótesis

$$\exists m > 2 \text{ tal que } mF(s) \le sf(s), \quad \forall s \ge 0. \tag{21}$$

En efecto, recordemos brevemente como se puede probar la condición de compacidad en este caso. Puesto que $|\Phi(u_n)| \le$ Const. tenemos

$$\frac{1}{2}\int_\Omega |Du_n|^2\,dx - \int_\Omega F(u_n)\,dx \le \text{Const.}$$

De otra parte, de la convergencia de $y_n \equiv \Phi'(u_n)$ a 0, tomando $v = u_n$ como función test y dividiendo por $m$, deducimos que

$$\frac{1}{m}\int_\Omega |Du_n|^2\,dx - \frac{1}{m}\int_\Omega u_n f(u_n)\,dx = \frac{1}{m}\langle y_n, u_n\rangle \ge -\varepsilon_n \|u_n\|,$$

con $\varepsilon_n \longrightarrow 0$. Restando estas expresiones llegamos a

$$\begin{aligned} \left(\frac{1}{2} - \frac{1}{m}\right)\int_\Omega |Du_n|^2\,dx &\le -\int_\Omega \left[F(u_n) - \frac{1}{m}\int_\Omega u_n f(u_n)\right] dx \\ &\quad +\text{Const.} + \varepsilon_n\|u_n\| \\ &\le \text{Const.} + \varepsilon_n\|u_n\|, \end{aligned}$$

de donde, usando que $m > 2$ concluimos que $\|u_n\|$ está acotada y argumentos standard prueban que existe una subsucesión $u_{n_k}$ convergente, y, por tanto, la condición de Palais-Smale.

Una pregunta natural a cuestionarse es ¿qué pasa si la parte principal del funcional es $\frac{1}{2}\int_\Omega a(x,v)|Dv|^2\,dx$ con $a(x,s)$ verificando (11) y (12)?. En este caso, en general (concretamente, si $N > 1$), el funcional no es de clase $C^1$ y así no se puede aplicar la teoría clásica de puntos críticos de Ambrosetti y Rabinowitz [2]. No obstante, se puede extender dicha teoría y probar [3] un teorema

abstracto de puntos críticos para este tipo de funcionales no diferenciables (véase también [4, 5, 6]). En lugar de presentarlo con todo detalle, preferimos reducirnos en estas notas a comentar algunas ideas que clarifican la situación.

Comenzamos discutiendo el caso $N = 1$ (el más simple). En este caso, como $W_0^{1,2}(\Omega) \cap L^\infty(\Omega) = W_0^{1,2}(\Omega)$, tenemos que el funcional sí es de clase $C^1$. Fácilmente se comprueba que $u = 0$ es un mínimo local estricto[4] que no es un mínimo global del funcional (pues no está acotado inferiormente); es decir, se verifican las hipótesis geométricas del Teorema de Paso de Montaña clásico. Así, podremos aplicar este teorema si somos capaces de probar la condición de Palais-Smale. Intentando *copiar* la demostración de ésta para el caso semilineal, llegamos a las dos expresiones

$$\frac{1}{2}\int_\Omega a(x,u_n)|u_n'|^2\,dx - \int_\Omega F(u_n)\,dx \le \text{Const.}$$

$$\frac{1}{m}\int_\Omega a(x,u_n)|u_n'|^2\,dx + \frac{1}{2m}\int_\Omega a_s'(x,u_n)u_n|u_n'|^2\,dx =$$

$$= \frac{1}{m}\int_\Omega u_n f(u_n)\,dx + \frac{1}{m}\langle y_n, u_n\rangle \ge -\varepsilon_n\|u_n\|,$$

con $\varepsilon_n \longrightarrow 0$. Restando estas expresiones obtenemos

$$\int_\Omega |u_n'|^2\left[\left(\frac{1}{2}-\frac{1}{m}\right)a(x,u_n) - \frac{1}{2m}a_s'(x,u_n)u_n\right]dx \le$$

$$\begin{aligned} &\le \quad -\int_\Omega \left[F(u_n) - \frac{1}{m}\int_\Omega u_n f(u_n)\right]dx + \text{Const.} + \varepsilon_n\|u_n\| \\ &\le \quad \text{Const.} + \varepsilon_n\|u_n\|. \end{aligned}$$

Así, una hipótesis razonable para conseguir la acotación de $u_n$ en $W_0^{1,2}(\Omega)$ sería imponer:

$$\left(\frac{1}{2}-\frac{1}{m}\right)a(x,s) - \frac{1}{2m}a_s'(x,s)s \ge \delta > 0, \quad \text{a.e. } x \in \Omega, \quad \forall s \in \mathbb{R}. \tag{22}$$

Suponiendo esta hipótesis, podemos obtener que la sucesión $u_n$ está acotada y, por tanto, pasando a una subsucesión, si es necesario, suponer que es débilmente convergente en $W_0^{1,2}(\Omega)$ a algún $u \in W_0^{1,2}(\Omega)$.

Ahora, puesto que $\Phi'(u_n) = y_n \longrightarrow 0$, tenemos

$$-\text{div}\,(a(x,u_n)u_n') + \frac{1}{2}a'(x,u_n)|u_n'|^2 = f(u_n) + y_n,$$

con $f(u_n)$ e $y_n$ compactas en $W^{-1,2}(\Omega)$. Respecto del término $\frac{1}{2}a'(x,u_n)|Du_n|^2$ observemos que está acotado en $L^1(\Omega)$ y como $L^1(\Omega)$ está inmerso compactamente en $W^{-1,2}(\Omega)$ (ya que $W_0^{1,2}(\Omega)$ lo está en $L^\infty(\Omega)$), deducimos

[4] El único cambio se debe a la condición (11) que obliga a escoger $\varepsilon \in (0, \alpha\lambda_1)$.

que este término también es compacto en $W^{-1,2}(\Omega)$. Luego $-\mathrm{div}\,(a(x,u_n)Du_n)$ es compacto en $W^{-1,2}(\Omega)$ y, por tanto, existe una subsucesión $u_{n_k}$ convergente a algún punto crítico.

En consecuencia hemos probado el siguiente teorema.

**Teorema 18** *Supongamos $N = 1$ y que la función $a(x,s)$ verifica las condiciones* (11), (12), (14) *y* (22) *y la función $f(s)$ verifica la condición de superlinealidad* (18) *y* (20). *Entonces el funcional $\Phi$ dado por* (17) *posee al menos un punto crítico $0 \le u \in W_0^{1,2}(\Omega) - \{0\}$.* □

**Nota 4** Respecto de las hipótesis que hemos impuesto al coeficiente $a(x,s)$ destaquemos que estas pueden contrastarse en un ejemplo *super-sencillo* como es el caso en el que $a(x,s) = b(s)^2$, para todo $s \in \mathbb{R}$. En este caso, el funcional es

$$\Phi(v) = \frac{1}{2}\int_\Omega b(v)^2|v'|^2\,dx - \int_\Omega F(v)\,dx, \quad v \in W_0^{1,2}(\Omega).$$

Si consideramos una primitiva $B$ de $b$ (es decir, tal que B'=b) y hacemos, para todo $v \in W_0^{1,2}(\Omega)$, el cambio $\widehat{v} = B(v)$ veremos que el estudio de los puntos críticos de $\Phi$ queda reducido al estudio de éstos para el funcional

$$\widehat{\Phi}(\widehat{v}) = \frac{1}{2}\int_\Omega |\widehat{v}'|^2\,dx - \int_\Omega F(B^{-1}(\widehat{v}))\,dx, \quad \widehat{v} \in W_0^{1,2}(\Omega),$$

el cual es del tipo (semilineal) estudiado por Ambrosetti y Rabinowitz [2] y que hemos discutido al principio de esta sección. Pues bien, nuestras condiciones impuestas al coeficiente $a(x,s) = b(s)^2$ implican las impuestas por estos autores en su trabajo para estudiar el caso superlineal.

El teorema anterior puede extenderse al caso general $N \ge 1$ (véase [3]), siendo ahora la prueba mucho más difícil. En efecto, en contraste con el caso unidimensional, forzosamente deberemos trabajar con los dos espacios de Banach $W_0^{1,2}(\Omega)$ y $W_0^{1,2}(\Omega) \cap L^\infty(\Omega)$. El primer paso será un teorema abstracto de puntos críticos para funcionales que son diferenciables únicamente a lo largo de direcciones de un subespacio vectorial. En este teorema, la geometría que se le requiere al funcional es análoga al caso clásico, pero la condición de compacidad necesitada es profundamente distinta. A saber, imponemos que se verifique

(*C*) *Cualquier sucesión $\{u_n\}$ en $W_0^{1,2}(\Omega) \cap L^\infty(\Omega)$ satisfaciendo las siguientes afirmaciones para ciertas $\{K_n\} \subset \mathbb{R}^+$ y $\{\varepsilon_n\} \longrightarrow 0$*

*i)* $\{\Phi(u_n)\}$ *es acotada,*

*ii)* $\|u_n\|_\infty + \|u_n\| \le 2K_n \quad \forall n \in \mathbb{N}$,

*iii)* $\langle \Phi'(u_n), v\rangle \leq \varepsilon \left[ \dfrac{\|v\|_\infty}{K_n} + \|v\| \right], \quad \forall v \in W_0^{1,2}(\Omega) \cap L^\infty(\Omega)$

*tiene una subsucesión convergente en* $W_0^{1,2}(\Omega)$.

El lector interesado podrá encontrar en la referencia citada [3] la verificación de esta condición $(C)$. Por nuestra parte, nos contentamos en estas notas con mostrar cómo, aunque aparentemente la prueba de esta condición de compacidad debería ser parecida a la del Teorema 10, no es posible, en realidad, seguir los pasos de este teorema. En efecto, si siguiéramos las ideas de la demostración de éste, tendríamos que probar para una cierta sucesión $\{u_n\}$ (de soluciones aproximadas por la condición *iii)*) que se verifican los siguientes pasos:

1. $\|u_n\|_\infty \leq$ Const.
2. $\|u_n\| \leq$ Const.
3. Compacidad.

Sin embargo, la verificación del punto 1. no es trivial ni siquiera en el caso semilineal y para la condición clásica de Palais-Smale. En efecto, el hecho de la convergencia $\Phi'(u_n) = y_n \overset{W^{-1,2}}{\longrightarrow} 0$ significa

$$-\Delta u_n = y_n \in W^{-1,2}(\Omega)$$

de donde no podemos deducir que $u_n$ estén en $L^\infty(\Omega)$. (Recordemos que la condición suficiente que suele usarse para concluir $u_n \in L^\infty(\Omega)$ es que $y_n \in W^{-1,r}(\Omega)$ con $r \geq N$).

## Referencias


[1] Ambrosetti, A., Brezis, H. y Cerami, G., *Combined effects of concave and convex nonlinearities in some elliptic problems.* J. Funct. Anal. **122** (1994), 519–543.

[2] Ambrosetti, A. y Rabinowitz, P. H., *Dual variational methods in critical point theory and applications.* J. Funct. Anal. **14** (1973), 349–381.

[3] Arcoya, D. y Boccardo, L., *Critical points for multiple integrals of Calculus of Variations.* Arch. Rat. Mech. Anal. **134** no. 3 (1996), 249–274.

[4] Arcoya, D. y Boccardo, L., *Some remarks on critical point theory,* NoDEA 6 (1999), 79–100.

[5] Arcoya, D. y Boccardo, L., *An introduction to critical points for integral functionals.* En *Nonlinear Partial Differential Equations and their applications: Collège de France Seminar, Volume XIII*, Stud. Math. Appl., 31, North-Holland, Amsterdam, 2002, pag. 1–12.

[6] Arcoya, D., Boccardo, L. y Orsina, L.,*Existence of critical points for some noncoercive functionals*. Annals. of Inst. Henri Poincaré (C) Analyse non linéaire Volume 18 - Numéro 4, (2001) 437–457.

[7] Boccardo, L. *The role of truncates in nonlinear Dirichlet problems in* $L^1$. *Nonlinear partial differential equations* (Fès, 1994), 42–53, Pitman Res. Notes Math. Ser., **343**, Longman, Harlow, 1996.

[8] Boccardo, L., Escobedo, M. y Peral, I., *A Dirichlet problem involving critical exponent.* To appear in Nonlinear Anal. TMA., **24** (1995), 1639–1648.

[9] Boccardo, L., Murat, F. y Puel, J.P., *Existence de solutions faibles pour des équations quasilinéaires à croissance quadratique.* Research Notes in Mathematics 84, Pitman, 1983, 19–73.

[10] Brezis, H., *Análisis Funcional*, Alianza Universidad Textos, 1984.

[11] Brezis, H. y Nirenberg, L., *Positive solutions of nonlinear elliptic equations involving critical Sobolev exponents.* Comm. Pure Appl. Math. **36** (1983), 437–477.

[12] Dacorogna, B., *Direct Methods in the Calculus of Variations.* Springer-Verlag, 1989.

[13] Kazdan, J.L. y Kramer, R.J., *Second-order quasilinear elliptic equations.* Comm. Pure Appl. Math. **31** (1978), 619–645.

[14] Stampacchia, G., *Equations elliptiques du second ordre à coefficients discontinus.* Les Presses de L'Université du Montréal, 1966.

*Bol. Soc. Esp. Mat. Apl.*
*nº 44(2008), 33–53*

# THE NUMERICAL SOLUTION OF DISCONTINUOUS IVPS BY RUNGE–KUTTA CODES: A REVIEW

M. CALVO, J.I. MONTIJANO AND L. RÁNDEZ

IUMA. Departamento Matemática Aplicada
Pza. San Francisco s/n. Universidad de Zaragoza.
50009-Zaragoza, Spain.

calvo@unizar.es monti@unizar.es randez@unizar.es

**Abstract**

In this paper several techniques for handling discontinuous IVPs when they are solved by means of adaptive Runge–Kutta codes are examined. The aim of the techniques is to provide, in an easy way, a tool to get a precise location of the detected discontinuities and a crossing of them retaining the accuracy of the numerical solution. Two remarkable features of these techniques is that they do not require an a priori knowledge about the location of the discontinuities and also that add a valuable capability to the existing software with a small computational cost. Some numerical experiments are presented to illustrate the reliability and efficiency of the proposed algorithms.

**Key words:** *Discontinuous Initial Value Problems, Adaptive Runge–Kutta methods, Detection, location and crossing of discontinuities*
**AMS subject classifications:** *65L06 65L05*

## 1 Introduction

Many standard adaptive Runge–Kutta (RK) codes for the numerical solution of non stiff IVPs of systems of differential equations [1, 13, 10] have been designed so that, for a given tolerance (TOL) given by the user, they provide an approximate solution of the IVP at some given points with maximum accuracy and efficiency. To achieve these goals it is not enough to select the best possible formula (one with the highest order, smallest local error and large stability region) and to apply it repeatedly with a fixed step size along the integration interval to get the solution at the required points. As it has been widely recognized, the behavior of the solution may vary strongly along the integration interval and consequently it is reasonable to take small step sizes when the

This work was supported by project MTM2004-06466-C02-01
Fecha de recepción: 09/11/2007. Aceptado (en forma revisada): 26/11/2007.

solution has a quick variation and the numerical integrator introduces large errors in the steps and larger step sizes when the solution varies slowly. A standard approach to control the step has been to include in the integrator a reliable estimator of the local error that allows the code to choose dynamically the step size according to the given tolerance. Hence, for a local error larger than TOL, the size of the step is reduced to make it about the same size of the TOL and if the local error is smaller than TOL the step size is increased. In the error estimation technique by embedded pairs, two RK formulas that share the same set of evaluations of the vector field of the differential equation and possess different orders e.g. $p$ and $p-1$ are constructed. Hence, since the local errors behave as $\mathcal{O}(h^{p+1})$ and $\mathcal{O}(h^p)$ respectively, the difference between the two solutions is an asymptotically correct estimate of the local error of the lower order solution. A pair of RK formulas of this type is called an embedded RK$(p, p-1)$ pair and when the code advances with the higher order solution it is said that employs local extrapolation. In this paper we will assume that the RK codes use local extrapolation.

The application of an explicit RK formula only provides approximations to the solution at a discrete set of grid points but there are important applications that require approximations between steps. Thus, a graphical display of solutions has become standard in may computer environments like MATLAB and in this context a continuous solution with suitable accuracy is required. Other application that requires continuous solutions is the so called events location in which we must find the time in which the solution $y = y(t)$ of a given IVP satisfies an algebraic equation $g(t, y) = 0$, for a given event function $g$. Here we must compute $(t^*, y^*)$ so that $y^* = y(t^*)$ and $g(t^*, y(t^*)) = 0$. In this kind of problems we need a continuous approximation $\widetilde{y}(t) \simeq y(t)$ and then to determine the time $t = t^*$ in which the event equation $g(t, \widetilde{y}(t))$ has a change of sign. In view of these applications, modern RK codes must include the capability to have a continuous approximation with the same accuracy as the original discrete formula.

The above properties of actual RK codes are well supported theoretically for IVPs whose solution is sufficiently smooth in the integration interval and further the vector field of the differential system is also sufficiently smooth in some neighborhood of the solution. However, when the vector field is not smooth enough the above mentioned theory that supports the good behavior of the codes does not hold and the accuracy and efficiency may suffer serious drawbacks. Observe that, when the differential equation is sufficiently smooth, the local error estimate (EST) of an adaptive code behaves as a certain positive power $\mathcal{O}(h^p)$ of the step size $h$ whereas in the presence of a discontinuity of order $q < p$, EST behaves as $\mathcal{O}(h^q)$. Hence, the presence of a low order discontinuity implies that EST is usually larger than the prescribed TOL and therefore to proceed the integration the code reduces the current step size. In general, when the numerical solution attains the discontinuity, the step size will be reduced one or more times until the step size is sufficiently small. In this process it may happen that after successive reductions of the step size, it is too small so that the integration is stopped (due to an excessive number of steps) and

then the code signals the user the detected problem or else the code crosses the discontinuity and the integrator recovers their normal step sizes. Nevertheless, in the last case it is not infrequent that significative errors be introduced in the crossing of the singularity and they affect adversely the global errors in the rest of the integration.

The aim of this paper is to give a survey about some approaches that have been proposed to overcome the shortcomings that can be found when an adaptive RK code is applied to the solution of discontinuous IVPs. In general all approaches have four stages: First of all, since in many practical IVPs the discontinuities are not known in advance by the user, the process starts with the design of algorithms that are able to suspect the presence of discontinuities by using the available information on the numerical solution computed by the RK code. Note that for problems where singularities are known in advance such a guessing process can be skipped and different techniques have been proposed e.g. in [5, 7, 11, 12]. In the second stage, after the detection process has been activated, a pre–location of the discontinuity within the current step with low computational cost is carried out. This stage allows us to confirm the existence of a discontinuity and at the same time to give a rough location of it. In the third stage the discontinuity is located with the desired accuracy and finally in the last stage the numerical solution crosses it maintaining the accuracy of the global errors along the integration.

Among the techniques for solving discontinuous IVPs, we may mention the paper of Enright, Jackson, Nørsett and Thompson [8]. Here by using the defects of the local interpolants of a RK pair of formulas the authors detect and locate a singularity and the restart the integration from it. This technique is applied to a RK(3,4) pair developed by Nørsett and also to a RK(5,6) developed by Verner and used in the classical code DVERK (see [10], p. 253). More recently an alternative technique based in the use of linear forms associated to the RK pair has been proposed by the authors of the present paper in [2, 3]. This technique has been justified by means of asymptotic reasonings and when applied to the DOPRI5(4) ([10], p. 253) it has given very good results for some test problems with low computational cost.

The paper is organized as follows: In section 2 some notations and basic assumptions that will be used along the rest of the paper are collected. In section 3 we explain and justify our criterion to activate the discontinuity detection process. In section 4 such a criterion is complemented with pre–location algorithms. In section 5 some algorithms to locate a discontinuity with the desired accuracy are examined and in section 6 the effects of crossing the discontinuity are analyzed. Finally, in section 7 the results of several numerical examples are presented to show that the proposed techniques are efficient and improve the reliability of adaptive standard RK methods in the solution of discontinuous IVPs.

## 2 Background

We consider IVPs for systems of first order differential equations that will be written in the form

$$y' = f(t, y), \quad y(t_0) = y_0 \in \mathbb{R}^d, \quad t \in [t_0, t_f], \tag{1}$$

where the function $f$ that defines the vector field or some derivatives of it may have bounded discontinuities.

In the simplest cases $f(t, y)$ can be written in closed form. For example, in the scalar equation $y' = f(y) \equiv \sum_{j=1}^{3} |y - j|$ ([8], Problem 2) $\partial_y f(y)$ has a bounded discontinuity at $y = 1, 2, 3$ and, depending on the initial conditions, the corresponding solution may have up to three discontinuities in the first derivative. In many mechanical applications discontinuous problems have the form $y'' = F(t, y) + \widetilde{F}(t, y)$. where $F$ comes from a smooth field of forces and $\widetilde{F}$ is a forcing term that acts only in special cases and is discontinuous. Clearly this $d$–dim second order system can be written as a $2d$–dim first order system in the form (1).

For a precise mathematical formulation of the (local) crossing of a discontinuity at some $(x_d, y_d)$, we will assume that in a neighborhood of this point there exist a smooth hypersurface $g(t, y) = 0$, (switching surface) so that $f(t, y)$ can be written in the form:

$$f(t, y) = \begin{cases} f_-(t, y) & \text{for} \quad g(t, y) \le 0, \\ f_+(t, y) & \text{for} \quad g(t, y) > 0, \end{cases}$$

with sufficiently smooth functions $f_-$ and $f_+$ satisfying a local Lipschitz condition with respect to $y$ in a tubular domain around the solution of (1).

We will say that a discontinuity of $f$ has order $q$ ($\ge 1$) if $f$ has a finite jump in at least one of the partial derivatives of order $(q-1)$ in the switching surface and it has continuous derivatives of all orders $< (q-1)$. Further $(t_d, y_d) \in \mathbb{R} \times \mathbb{R}^d$ is a switching point of the solution $y = y(t)$ of (1) if $y_d = y(t_d)$ and $g(t_d, y_d) = 0$. We will assume that the solution of (1) satisfies the so called transversality condition (see e.g. [11], eq. (1.6)); this implies that there is a unique solution $y(t)$ of (1) with a finite number of switching points and at each switching point $y^{(q)}(t)$ has a finite jump.

There are some types of discontinuous IVPs that do not fit into the above form (1) but can be solved with the proposed techniques to be given below with a careful definition of $f$. As a first example we mention a simple model to describe the heating of a furnace that is turned on and off under the control of a thermostat ([8], Problem 3). Here a status switch ($sw$) that has the value 1 when the furnace is on and 0 otherwise and will be reset whenever the temperature reaches a critical value is introduced. For example, an equation of this type could be

$$y' = \begin{cases} y & \text{if} \quad sw = 1, \\ -y/2 & \text{if} \quad sw = 0, \end{cases} \qquad t \ge 0,$$

with $y(0) = 1$, $sw = 1$ at $t = 0$ and $sw$ is reset to 0 when $y \geq 2$ and reset to 1 when $y \leq 1$. Now the solution is a (periodic) function that oscillates between $y = 1$ and $y = 2$ and the numerical difficulties arise when crossing the critical values $y = 1, 2$ in which the derivative is discontinuous.

Another discontinuous IVP considered by Shampine, Gladwell and Thompson in [14] in connection with the MATLAB facility for event location is the motion of an elastic ball bouncing down in an inclined plane. Considering the problem in a vertical plane $(x, y)$ and assuming that the equation of the inclined plane is $x + y = 1$, we have a second order system in $x = x(t), y = y(t)$ in which the events are the successive contacts of the ball with the plane. Now at each event we must restart the differential equation from this point with the reflected velocity and therefore this variable is discontinuous at each event.

For the numerical solution of (1) we consider $s$ stages adaptive Runge–Kutta codes based on an embedded pair of explicit formulas of orders $p$ and $(p-1)$ defined by the coefficients $A = (a_{ij}) \in \mathbb{R}^{s\times s}$, with $a_{ij} = 0$, for $j \geq i$, the weights $b = (b_i) \in \mathbb{R}^s, \bar{b} = (\bar{b}_i) \in \mathbb{R}^s$ and the nodes $c_j = \sum_{l=1}^{j-1} a_{jl}$ [10]. In a successful step $(t_n, y_n) \to (t_{n+1} = t_n + h_n, y_{n+1})$ with step size $h_n$ the code advances the numerical solution with the higher order solution defined by

$$y_{n+1} = y_n + h_n \sum_{i=1}^{s} b_i \, f_i, \tag{2}$$

and computes an estimate $\mathrm{EST}_n = E(t_n, y_n, h_n)$ of the local discretization error as the difference of the solutions of orders $p$ and $(p-1)$ by means of

$$\mathrm{EST}_n = h_n \sum_{i=1}^{s} (b_i - \bar{b}_i) \, f_i, \tag{3}$$

with

$$f_i = f(t_{n,i}, Y_{n,i}) \equiv f\left( t_n + c_i h_n, y_n + h_n \sum_{j=1}^{i-1} a_{ij} \, f_j \right), \quad i = 1, \ldots, s. \tag{4}$$

Note that since the exact solution of $y' = f(t, y)$ starting from $(t_n, y_n)$ satisfies $y(t_{n+1}) = y_n + \int_{t_n}^{t_{n+1}} f(t, y(t))\mathrm{d}t$, the RK solution (2) may be viewed as an approximation to the integral over the vector field by a discrete average over the points $(t_{n,i}, Y_{n,i})$.

In all codes the acceptance of the step $t_n \to t_{n+1}$ depends on the fact that a norm of $\mathrm{EST}_n$ remains below the prescribed TOL. Here it will be assumed that the step size control is based on an estimate of type (3) obtained from an embedded pair, however there are alternative techniques (Enright *et al.*, [8]) in which this control is based on an estimate of the defect $\delta(t) = p'(t) - f(t, p(t))$ of some interpolant of the numerical solution in the step. Further in the case of an accepted step such an $\mathrm{EST}_n$ is also used to predict the new step $h_{n+1}$ from $(t_{n+1}, y_{n+1})$. Here in the case of (possibly) discontinuous IVPs we will restrict the step size increase by a factor of two, i.e. $(h_{n+1}/h_n) \leq 2$.

In addition, some conditions on the nodes $c_i$ and coefficients $(b_i - \overline{b}_i)$ of (3) must be imposed. First of all, if some $c_i > 1$, the corresponding $t_{n,i} \notin [t_n, t_n + h_n]$, and we could have discontinuous problems for which the local solution at $(t_n, y_n)$, $y(t; t_n, y_n)$ is on the same side of the switching surface for all $t \in [t_n, t_n + h_n]$ while the discrete solution employs function evaluations on both sides of this surface. As a consequence of this fact we will assume $0 \le c_i \le 1, i = 1, \ldots, s$ with

$$c_1 = 0, \quad c_s = 1, \qquad \text{and} \quad b_1 - \overline{b}_1 \neq 0,\ b_s - \overline{b}_s \neq 0. \tag{5}$$

The last two conditions imply that in the local error estimation (3) we are taking into account both ends of the interval of the step. Furthermore the last function evaluation of (4) could be on one side of the switching surface while $(t_{n+1}, y_{n+1})$ could be on the other side. Thus it is convenient that $Y_{n,s} = y_{n+1}$, i.e. the Runge–Kutta pair satisfies the so called FSAL condition. In conclusion we will consider Runge–Kutta pairs that satisfy (5) and the FSAL condition.

## 3 Detecting a discontinuity

Our first task is to establish a test that detects whether or not the numerical solution has crossed a discontinuity in a typical step $t_n \to t_{n+1}$. Since this detection will be tested frequently it should have two natural requirements: low computational cost and to fit well into the code, i.e. it should be based on quantities that the code computes at every step. Observe that this information about the vector field concerns a finite set of points and this implies that no criterion will be able to detect all kinds of discontinuities.

Several criteria have been proposed in the literature (Gear and Østerby [9], Enright *et al.* [8], Calvo, Montijano and Rández [3]). Although all of them have in common the rejection of a step as a consequence of an unusually large estimate of the local error, they are all different and it can be seen that neither of them implies the others.

Next, we will state the criterion employed by the authors in [3] to start the pre–detection process of a discontinuity. Suppose that from $(t_n, y_n)$ with a predicted step size $h_n^{(a1)}$, the step $t_n \to t_n + h_n^{(a1)}$ is a failed step because $\|\mathrm{EST}_n\| > \mathrm{TOL}$ and the step size control predicts a new step size $h_n^{(a2)} \le h_n^{(a1)}/2$. Then, in agreement with Gear and Østerby, we will consider that this condition is sufficient to suspect a possible discontinuity in the failed step $(t_n, t_n + h_n^{(a1)}]$. On the other hand in the case that $h_n^{(a2)} > h_n^{(a1)}/2$ and there is a discontinuity in $(t_n, t_n + h_n^{(a1)}]$, if the step is accepted the new $t_{n+1} = t_n + h_n^{(a2)}$ will be closer to this discontinuity. But practical experience shows that this process could lead to a sequence of accepted–rejected steps with a geometrically decreasing size and, in our opinion, this situation should be avoided (see e.g. Hairer, Nørsett, Wanner, p. 197, Fig. 6.3). Thus, if two rejections are encountered in the last three steps we propose to activate also a pre–location process. In conclusion we activate the discontinuity pre–location process when the code detects either of the following situations:

- The new step size $h_n^{(a2)}$ predicted by the code after a rejected step of size $h_n^{(a1)}$ implies a reduction factor $\leq 0.5$, i.e. $h_n^{(a2)} \leq h_n^{(a1)}/2$.
- In the last three steps the code had at least two rejections.

It may be argued that our test is too sensitive in the sense that for some smooth problems with a strongly varying function $f$ it may detect discontinuities which mathematically do not exist. In fact, such a case may appear, but the more refined study of the suspected discontinuities that will be presented below will help to disregard this possibility. On the contrary we have found that our test is able to detect some discontinuities that otherwise would be undetectable.

## 4 Pre–locating the discontinuity

Suppose that the above detection process signals the existence of a discontinuity in the step $t_n \to t_n + h_n^{(a1)}$, then our next aim is to confirm or not the suspected discontinuity in this step and to prelocate it in a subinterval of the step by using the available information on the numerical solution in the previous successful step $t_{n-1} \to t_n = t_{n-1} + h_{n-1}$ as well as in the rejected step $t_n \to t_n + h_n^{(a1)}$.

We start comparing the size of the previous accepted step $h_{n-1}$ with the failed step $h_n^{(a1)}$ $(< 2h_{n-1})$. Since an increase of size of the step may be dangerous in a numerical integration that attempts to search discontinuities, if $h_n^{(a1)} > h_{n-1}$ then we try a new step from $t_n$ with the size $h_n^{(a1)}/2$. Now if $t_n \to t_n + (h_n^{(a1)}/2)$ turns out to be a successful step the suspected discontinuity would be located in second half of the failed step i.e. in the interval $(t_n + h_n^{(a1)}/2, t_n + h_n^{(a1)}]$ and otherwise it would be contained in $(t_n, t_n + h_n^{(a1)}/2]$. In any case we may assume that after a successful step $t_{n-1} \to t_n = t_{n-1} + h_{n-1}$ (that for simplicity will be denoted by $t_0 \to t_0 + h$ ), we have a failed step $t_n \to t_n + h_n^{(a1)}$ $(t_1 \to t_1 + rh)$ with $h_n^{(a1)} < h_{n-1}$ $(r < 1)$.

Next, as our pre–location process depends on the particular RK pair under consideration we will consider the particular case of the DOPRI54 pair, in which have at our disposal the following evaluations of the vector field:

$$f_i = f(t_{0,i}, Y_{0,i}), \quad i = 1, \ldots, 7, \quad \text{of the last successful step } t_0 \to t_0 + h$$
$$f(t_{1,i}, Y_{1,i}), \quad i = 1, \ldots, 7, \qquad \text{of the current failed step } t_1 \to t_1 + rh$$

with $t_1 = t_0 + h, r < 1$ and $t_{0,i} = t_0 + c_i h, t_{1,i} = t_1 + c_i rh$. Observe that in this particular pair of Dormand and Prince the nodes have the values

$$c_1 = 0,\ c_2 = 0.2,\ c_3 = 0.3,\ c_4 = 0.8,\ c_5 = 8/9,\ c_6 = c_7 = 1,$$

and taking into account that $t_{1i} = t_1 + c_i rh$, the available information of the vector field in the critical pre–detection interval $(t_1, t_1 + rh]$ concentrates on both ends of the interval. In view of this fact, to improve the reliability of our search process we have included between 0.3 and 0.8 three extra nodes so that, in a discontinuity step, we have the extended 10–dim vector of nodes

$$\widehat{c} = (\widehat{c}_i) = (0, 0.2, 0.3, 0.38, 0.48, 0.62, 0.8, 8/9, 1, 1)$$

with the corresponding approximations

$$f_{7+i} = f\left(t_1 + \widehat{c}_i rh, \widehat{Y}_{1,i}\right), \quad i = 1, \ldots, 10,$$

and the coefficients $a_{ij}$ that define $\widehat{Y}_{1,i}$ in the stage equations (4) are chosen as in the original DOPRI54 pair taking into account the same simplifying assumptions.

### 4.1 Pairs of embedded forms

To detect a possible discontinuity in the first subinterval $[t_1, t_{1,2} = t_1 + \widehat{c}_2 rh]$ of the failed step we propose to construct two linear forms $\phi_{12}$ and $\psi_{12}$ on the vector field evaluations

$$\phi_{12}(h) = h \sum_{i=1}^{7} C_{2,i}\, f_i, \qquad \psi_{12}(h) = h \sum_{i=1}^{9} D_{2,i}\, f_i,$$

where $C_{2,i}$ are constant coefficients and $D_{2,i}$ depend on $r$ so that $\phi_{12}$ is based on function evaluations of the back successful step (presumably on the left of the discontinuity) and $\psi_{12}$ includes also $f_9$ i.e. the function evaluation at the the point $(t_{1,2}, \widehat{Y}_{1,2})$ with $D_{2,9} \neq 0$. The coefficients of these forms have been determined so that both have the same highest order $\kappa$

$$\phi_{12}(h) = \mathcal{O}(h^\kappa) = \psi_{12}(h),$$

and are asymptotically equivalent, i.e. with the same leading terms of their Taylor series expansion in $h$ if no discontinuity is found in $[t_1, t_{1,2}]$.

For the DOPRI54 pair, taking into account the simplifying assumptions employed in their derivation, it is found that, by using the Butcher series expansions of the two forms, $\kappa = 3$. Thus, for $r = 1$ we have (see [4] and [3] for a detailed derivation of them for arbitrary step sizes)

$$\begin{aligned}
\phi_{12}(h) &= (h/9)\left[f_4 - 9f_5 + 8f_7\right], \\
\psi_{12}(h) &= h\left[-\frac{247}{440}f_1 - f_2 + \frac{1280}{371}f_3 - \frac{9631}{396}f_4\right. \\
&+ \left.\frac{823049}{23320}f_5 - \frac{75821}{2485}f_6 + \frac{585154}{35145}f_7 + f_9\right].
\end{aligned}$$

In this way if the two points $(t_1, y_1)$ and $(t_{1,2}, Y_{1,2})$ are on the same side of the switching surface, the quotient $q_{12}(h) = \|\psi_{12}(h)\|/\|\phi_{12}(h)\| \to 1$ as $h \to 0^+$. Nevertheless if the two points are on different sides of the switching surface and there is a discontinuity of order $q$ we have $\phi_{12}(h) = \mathcal{O}(h^3)$ whereas $\psi_{12}(h) = \mathcal{O}(h^q)$ and therefore for $q < 3$ the quotient $q_{12}(h) \to +\infty$ when $h \to 0^+$. This means that a large value of $q_{12}$ signals that $(t_{1,2}, Y_{1,2})$ is on the other side of the switching surface.

In practical computation (depending on the tolerance) $h$ may not be too small. Thus, in order to give a simple test that allows us, from the computed

value of $q_{12}(h)$, to detect the presence of a low order discontinuity, we have considered the condition

$$q_{12}(h) \leq K. \tag{6}$$

The value of $K = 10$ has been chosen from practical experience with several test problems.

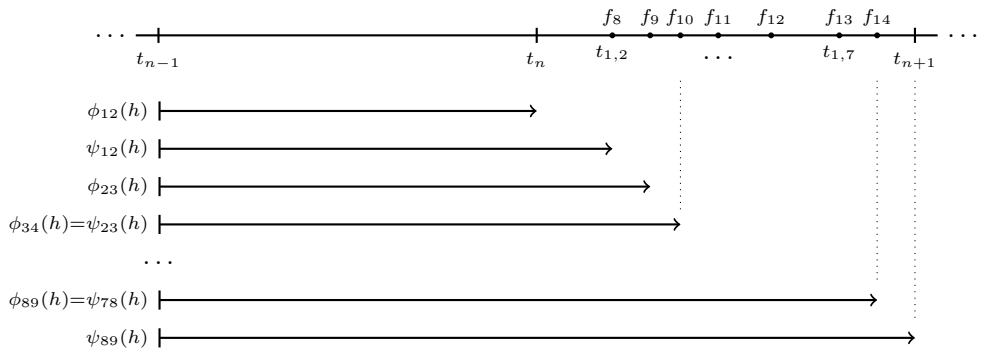
Suppose that the above test does not show the existence of a discontinuity between $t_1$ and $t_{1,2}$, then we will continue with the detection process in the next interval $(t_{1,2} = t_1 + \widehat{c}_2 rh, t_{1,3} = t_1 + \widehat{c}_3 rh]$. To this end we follow the same approach as above, we take two linear forms $\phi_{23}$ and $\psi_{23}$ given by

$$\phi_{23}(h) = h \sum_{i=1}^{7} C_{3,i}\ f_i, \qquad \psi_{23}(h) = h \sum_{i=1}^{10} D_{3,i}\ f_i,$$

with $D_{3,10} \neq 0$, so that both have the same maximum order and are asymptotically equivalent. It is found that this order is four and in particular, for $r = 1$ we get

$$\begin{aligned}
\phi_{23}(h) &= h\left[-\frac{53}{243}f_1 + \frac{128}{243}f_2 - \frac{106}{81}f_4 + f_5\right], \\
\psi_{23}(h) &= h\left[-(76831/311040)f_1 + (591763/901530)f_3 - (184723/51840)f_4\right. \\
&+ \left.(323719/67840)f_5 - (13035/3976)f_6 + (554/355)f_7 + (1/10)f_{10}\right].
\end{aligned}$$

This process is continued until the discontinuity is detected in some subinterval $(t_{1,i}, t_{1,i+1}]$ or else no discontinuity is detected in the whole interval $(t_1, t_1 + rh]$, taking $\phi_{i\ i+1}(h) = \psi_{i-1\ i}(h)$. In the last case since no discontinuity is pre–detected in the suspected interval the integration restarts again from $t_1$ in the usual way.




### 4.2 Using the defect

Now, we try again to determine the subinterval $(t_1 + rh\hat{c}_{i-1}, t_1 + rh\hat{c}_i]$ in which the discontinuity is located, but using in this case the defect of the solution.

The idea is to construct for each subinterval a continuous method

$$p_i(t) = y_0 + h \sum_{j=1}^{7+i} b_{i,j}(\theta)\ f_j, \quad i = 2, \ldots, 9,$$

where $t = t_0 + h\theta$, with the highest order possible. Imposing the order conditions and using the free parameters to minimize the coefficient of the leading error term (see e.g. [2] to get more information about the minimization process) it is seen that in all cases the methods have order four except for the subinterval $(t_1, t_1 + rc_2h]$ in which only order three can be attained.

For each $i$, we choose a point $s_i^* = t_0 + \theta_i^* h \in [t_0, t_1]$ and compute the defect

$$\delta_i(s_i^*) = p_i'(s_i^*) - f(t_0 + h\theta_i^*, p_i(s_i^*)).$$

If the discontinuity is located before the subinterval $(t_1 + rh\hat{c}_{i-1}, t_1 + rh\hat{c}_i]$, $\delta_i(s_i^*) = \mathcal{O}(h^p)$, being $p$ the order of the continuous method, but otherwise it can be proved that $\delta_i(s_i^*) = \mathcal{O}(h^q)$, where $q$ is the order of the discontinuity. As for the case of pairs of embedded forms, if

$$\frac{\|\delta_i(s_i^*)\|}{\|\delta_0(s_i^*)\|} > K, \quad \text{and} \quad \frac{\|\delta_j(s_j^*)\|}{\|\delta_0(s_j^*)\|} \leq K \text{ for } j = 1, \ldots, i-1,$$

we will assume that the discontinuity belongs to $[t_1 + rh\hat{c}_{i-1}, t_1 + rh\hat{c}_i]$.

## 5 Locating the discontinuity

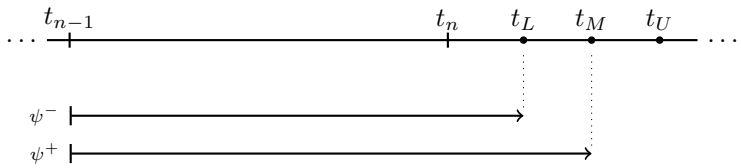
So far the pre–location process either disregards the existence of the suspected discontinuity in the failed step $t_n \to t_n + h_n$ (with $h_n = rh_{n-1}, r < 1$) or else confirms their existence in some (critical) subinterval $(t_{n,k}, t_{n,k+1}]$ with $t_{n,k} = t_n + \widehat{c}_k h_n$ defined by the extended vector of nodes $\widehat{c} = (\widehat{c}_k)$. Our next issue is to locate such a discontinuity in a sufficiently small critical interval depending on our accuracy requirements. Before to do that, note that due to the fact that the pre–detection is based on asymptotic reasonings with a discrete set of points, we must admit some inaccuracies in the definition of the critical subinterval. To improve the reliability of our critical interval we will start our location process from an extended interval $[t_{n,k} - \varepsilon, t_{n,k+1} + \varepsilon]$ with some $\varepsilon > 0$. In our numerical experiments we have found that $\varepsilon = 0.05h_n$ is enough to have a reliable starting critical interval.

If $k \geq 3$, we advance a step from $t_n$ to $t_n + r\hat{c}_k h_n$. In this way we can always ensure that we have given a successful step from $t_{n-1}^*$ to $t_n^* = t_{n-1}^* + h$ (with $t_{n-1}^* = t_n$ and $h = r\hat{c}_k h_n$ or else $t_{n-1}^* = t_{n-1}$ and $h = h_n$) and the discontinuity is pre–located into an interval $[t_n^*, t_n^* + ch]$, with $c = (\hat{c}_{k+1} - \hat{c}_k)rh_n/h \leq 0.35$.

A natural approach to reduce the length of the critical interval $[t_n^*, t_n^* + ch]$, used in [9], is based on the bisection technique. Now, restarting the integration from $(t_n^*, y_n^*)$ with step size $ch/2$ and taking into account the local error estimate in this step we may decide whether the discontinuity is located either on

$(t_n^*, t_n^* + ch/2]$ or else on $(t_n^* + ch/2, t_n^* + ch]$. This process can be repeated until the length of the interval is sufficiently small but the main drawback is that their computational cost can be very high because each bisection has the cost of an integration step and further, since the integration step is smaller in each bisection, the computation may be strongly contaminated by rounding off errors. A similar bisection technique in connection with the defects of suitable polynomial interpolant of the numerical solution has been used by Enright *et al.* in [8].

Next, we will describe two more elaborated alternative techniques that are cheaper than the bisection technique and have been proved very effective in the numerical experiments performed by the authors in many discontinuous IVPs. In the first technique, referred to as pair of linear forms and proposed by the authors in [3], is based again in the comparison of two asymptotically equivalent linear forms $\psi^-$ and $\psi^+$ in the evaluations $f_i$ of the vector field. The first one $\psi^-$ is based on the available information in the previous accepted step whereas the second one $\psi^+ = \psi^+(t)$ takes into account this information and also an extra point $(t, u(t))$ with $t \in (t_n^*, t_n^* + ch)$ that may vary along the interval of discontinuity and $u(t) \simeq y(t; t_{n-1}, y_{n-1})$ near the local solution starting from $(t_{n-1}, y_{n-1})$. This approximation may be obtained from a continuous extension associated to the RK pair and in general it has an extra cost. Then the behavior of the quotient $||\psi^+(t)||/||\psi^-||$ with $t$ will allow us to determine if the discontinuity $t_d$ satisfies $t_d > t$ or else $t_d \leq t$. The second technique given below is based on the construction of a continuous method $p(t)$ that allows us to evaluate the defect $\delta(t) = p'(t) - f(t, p(t))$ at different points.




### 5.1 Pairs of linear forms

To illustrate the construction of the linear forms $\psi^-$ and $\psi^+$ we will consider again the DOPRI54 pair that uses 7 stages (including the FSAL). For the continuous extension of this pair we have used the one proposed by the authors in [2]. This interpolant requires two additional function evaluations in each step and provides for each successful step $t_{n-1} \to t_{n-1} + h_{n-1}$ a polynomial $p(t)$ with $p(t_{n-1}) = y_{n-1}, p(t_n) = y_n$ that has the same order of accuracy as the extrapolated discrete solution between steps. In addition, for extrapolation purposes, $p(t)$ it is also a reliable approximation outside the interval provided that $t \in [t_{n-1}, t_{n-1} + 1.35h_{n-1}]$.

Assume that the critical subinterval is the first one $[t_{n,1}, t_{n,2}] = [t_n, t_n + c_1 hr]$. The coefficients of the form $\psi^-$, based on the available information of the vector

field in the previous step $t_{n-1} \to t_n^* = t_{n-1} + h$, have been chosen so that

$$\psi^- = \sum_{i=1}^{7} \beta_i^- f(t_{n-1,i}, Y_{n-1,i}) = C^- h^5 + \mathcal{O}(h^6), \quad (C^- \neq 0),$$

with the normalization $\beta_7 = 1$. This leads to the values

$$\beta^- = \left( -\frac{71}{1440}, 0, \frac{568}{3339}, -\frac{71}{48}, \frac{17253}{8480}, -\frac{176}{105}, 1 \right)^T.$$

For the form $\psi^+$ we take

$$\psi^+ = \sum_{i=1}^{9} \beta_i^+ f(t_{n-1} + c_i h, Y_{n-1,i}) + \beta_{10}^+ f(t, p(t)) = C^+ h^5 + \mathcal{O}(h^6).$$

that takes into account, apart of the information of the previous step, information in the point $(t, p(t))$ of the critical interval, where $p(t)$ is the above mentioned fifth order interpolant. The coefficients $\beta_i^+$ are selected so that $\psi^+$ has order five i.e $\psi^+ = C^+ h^5 + \mathcal{O}(h^6)$ with the additional requirement that $C^+ = C^-$. It can be seen that with the available parameters all these requirements may be satisfied with $\beta^+ = (\beta_1^+, \ldots, \beta_{10}^+)^T$ given by

$$\begin{aligned}
\beta_1^+ &= -\tfrac{2043009701250\theta^4 - 4988244082500\theta^3 + 4363161405550\theta^2 - 1667789853100\theta + 258805757340}{2208960(104000\theta^2 - 166880\theta + 71091)},\\
\beta_2^+ &= 0,\\
\beta_3^+ &= \tfrac{8}{2561013}\,\tfrac{2810707410000\theta^4 - 6037191121500\theta^3 + 4123999290500\theta^2 - 931758140600\theta + 38714025870}{104000\theta^2 - 166880\theta + 71091},\\
\beta_4^+ &= \tfrac{3425937112500\theta^4 - 5860560195000\theta^3 + 1727616777500\theta^2 + 1391857537000\theta - 77428051740}{73632(104000\theta^2 - 166880\theta + 71091)},\\
\beta_5^+ &= -\tfrac{243}{13008320}\,\tfrac{1257206737500\theta^4 - 2167088175000\theta^3 - 6002466550\theta^2 + 1654757335000\theta - 77428051740}{104000\theta^2 - 166880\theta + 71091},\\
\beta_6^+ &= \tfrac{11}{80535}\,\tfrac{6996456250\theta^4 - 12068117250\theta^3 - 6781987600\theta^2 + 1957028220\theta - 872428752}{104000\theta^2 - 166880\theta + 71091},\\
\beta_7^+ &= -\tfrac{(\theta-1)(9948995000\theta^3 - 553147700\theta^2 - 63064089\theta + 54526797)}{767(104000\theta^2 - 166880\theta + 71091)},\\
\beta_8^+ &= -\tfrac{330150625}{6136}\,\tfrac{\theta(5\theta-2)(\theta-1)^2}{104000\theta^2 - 166880\theta + 71091},\\
\beta_9^+ &= -\tfrac{342125}{12}\,\tfrac{\theta(\theta-1)(25\theta^2 - 23\theta + 4)}{104000\theta^2 - 166880\theta + 71091},\\
\beta_{10}^+ &= \tfrac{8211}{104000\theta^2 - 166880\theta + 71091},
\end{aligned}$$

with $\theta = (t - t_{n-1})/h$.

An analysis of the above forms [4] shows that, on the left of the discontinuity $\nu(t) := \|\psi^+(t)\|/\|\psi^-\| = 1 + \mathcal{O}(h)$ whereas on the right $\nu(t) = \mathcal{O}(h^{-j})$ with $j \geq 1$. In the left side of Figure 1 we have displayed for problem 1 (see section 7), integrated with error tolerance $10^{-5}$, the function $\nu(t_{n-1} + \theta h)$ for $\theta \in [1, 1.35]$. The step size from $t_{n-1}$ to $t_n$ is the one provided by the detection process in such a way that the discontinuity point $t_d = 1$ has been pre–located in the

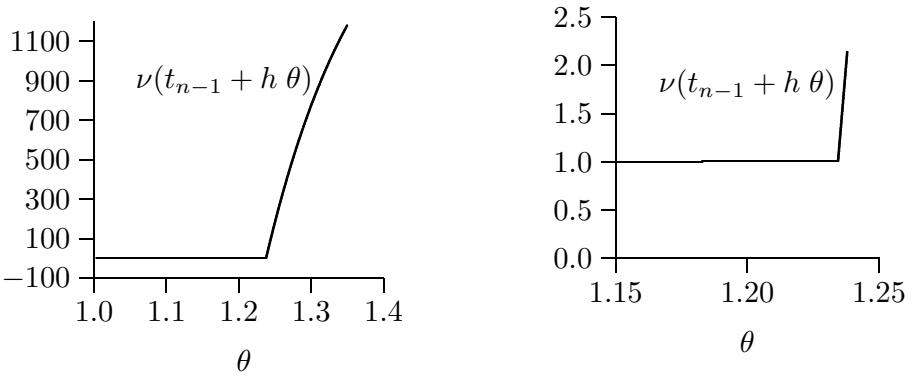



Figure 1: Function $\nu(t)$ crossing a discontinuity for problem 1

interval $(t_n, t_n + 1.35h]$. On the right side of the figure we have displayed the behavior of the same function $\nu(t_{n-1} + \theta h)$ for $\theta \in [1.15, 1.25]$. As it can be seen, for $t < t_d$, the function $\nu(t)$ remains very close to 1. However, when we cross the discontinuity, $\nu(t)$ increases very quickly. This behavior in crossing the discontinuities has been observed in all tested problems with first order discontinuities.

We must remark that for discontinuities of order greater than or equal to 2, $\nu(t)$ is a continuous function. For values of $t \geq t_d$ close to the discontinuity, $\nu(t)$ can be near to 1 and the discontinuity point can not be located with high precision with this criteria. However, the function $\nu(t)$ is almost constant before crossing the discontinuity, whereas after doing it, the function varies quickly. This fact can be used in the location process and thus we can consider that a point $t$ is to the right of the discontinuity if $(\nu(t) - \nu(t_L))/(t - t_L) > M$ being $t_L$ the last point that has been considered in the bisection process as being located to the left of the discontinuity and therefore the available point closest to the discontinuity with $t_L < t_d$. In the numerical results to be presented below we have taken $M = 10$.

### 5.2 Using the defect

Enright *et al.* [8] use a continuous extension $p(t)$ of the method in the back successful step $[t_{n-1}, t_n]$ and compare the defect $\delta(t) = p'(t) - f(t, p(t))$ of this extension at $t$ with the defect $\delta(t^*)$ at some suitably chosen fixed point $t^* \in (t_{n-1}, t_n)$ to locate the discontinuity. However, there are serious objections to this technique when the time $t$ is far from $t_n$. To remedy this inconvenient in the DOPRI pair, we have proposed a modification that consist in adding an explicit stage to the Runge–Kutta formula, at the point $t = t_{n-1} + (1 + \eta)h$

defined by

$$f_8(\eta) = f(t_{n-1} + (1+\eta)h, y_{n-1} + h\sum_{i=1}^{7} a_{8i}(\eta) g_i), \tag{7}$$

so that we have an extended DOPRI method with the Butcher's coefficients

$$\begin{array}{c|c} \hat{c} & \hat{A} \\ \hline & \hat{b}^T(\eta) \end{array} = \begin{array}{c|c} c & A \\ 1+\eta & a_8^T(\eta) \\ \hline & \hat{b}^T(\eta) \end{array}$$

that provides a new $y_{n+\eta} \simeq y(t_{n+1} + (1+\eta)h; t_{n-1}, y_{n-1})$ given by

$$y_{n+\eta} = y_{n-1} + h\sum_{i=1}^{8} b_i(\eta) f_i. \tag{8}$$

Note that the calculation of $y_{n+\eta}$ requires only one extra evaluation of $f_8(\eta)$.

Although we do not enter into details about the explicit computation of the coefficients $a_{8j}$ and $b_j(\eta)$ of (7) and (8), it must be noticed that in the above formulas there are 15 parameters $a_{8j}(\eta)$, $j = 1, \ldots, 7$ and $\hat{b}_j(\eta)$, $j = 1, \ldots, 8$ that must be chosen so that $y_{n+\eta}$ is fifth order solution at $t = t_{n-1} + (1+\eta)h$. Then, taking into account the order conditions and the simplifying assumptions used in the derivation of the DOPRI pair it is found that there are still two free parameters. In our study they have been used so that minimize the coefficients of the leading term of the local error of the solution (8).

Observe that for a fixed $h$ and variable $\eta$, $v(t) = y_{n+\eta}$ (with $t = t_{n-1} + (1+\eta)h$) may be viewed as a continuous approximation to the solution to the local solution in $(t_{n-1}, y_{n-1})$ at each $t$ that, in contrast with the standard interpolant on the step $t_{n-1} \to t_{n-1} + h$, includes the additional information $f_8$ on the vector field in the last point $t = t_{n-1} + (1+\eta)h$ and therefore (8) is more reliable than the standard interpolant on intervals that extend on the right of $t_{n-1} + h$. The coefficient of the leading term of the local error of the solution $y_{n+\eta}$ has a moderate size whenever $\eta \lesssim 0.35$.

By using this continuous extension whose computation only requires one function evaluation for each value of $\eta$ we may approximate the defect of $v(t)$, $\delta(t) = v'(t) - f(t, v(t))$ by the finite difference expression

$$\widetilde{\delta}(t) = \frac{v(t) - v(t - h\varepsilon)}{h\varepsilon} - f(t, v(t)) = \frac{y_{n+\eta}(\eta) - y_{n+\eta-\varepsilon}(\eta - \varepsilon)}{h\varepsilon} - f(t, y_{n+\eta}).$$

It can be seen that, for smooth $f$, $\widetilde{\delta}(t) = \mathcal{O}(h^5)$ and otherwise $\delta(t)$ is not continuous in a neighborhood of a discontinuity of order $\leq 2$.

Finally, for the precise location of the discontinuity, we follow, like with pair of linear forms, a bisection process. Assuming that the discontinuity $t_d$ is contained in a critical interval $[t_L, t_U]$, for the mid point $t_m = (t_L + t_U)/2$ we compute the quotient $\nu(t_m) = \|\widetilde{\delta}(t_m)\|/\|\widetilde{\delta}(t_L)\|$ and if $\nu(t_m) < K$ we conclude

that the discontinuity is contained in $[t_m, t_U]$. This process is continued until the length of the interval is sufficiently small.

For discontinuities of order greater than 2, $\nu(t)$ is a continuous function and for values of $t \geq t_d$ close to the discontinuity, $\nu(t)$ near to 1, the discontinuity point can not be located with high precision with the criteria $\nu(t) \leq K$. As with pairs of embedded forms, we can consider that a point $t$ is to the right of the discontinuity if $(\nu(t) - \nu(t_L))/(t - t_L) > M$. In our experiments we have taken $M = 10$.

## 6 Crossing the discontinuity

As remarked above, in the iterative location process a discontinuity in the solution of the IVP at some $t_d$ may be located in a sufficiently small critical interval $[t_L, t_H]$ depending on the required tolerance TOL. In particular, in our numerical experiments with the DOPRI54 pair, we have found that the requirement $t_H - t_L \leq \text{TOL}/8$ is enough for practical purposes.

Our next goal is to cross the critical interval retaining the prescribed accuracy of the numerical solution. To do that, recall that it has been proved by Enright et al. in [8] that an asymptotically correct bound of the error $\varepsilon_c$ in crossing the discontinuity is given by $\varepsilon_c = (t_H - t_L)\|\delta(t_H)\|$ where $\delta(t)$ is the defect of an interpolant with the same order of accuracy as the numerical solution. To estimate the last term observe that when the order of the discontinuity is $q \geq 2$ the defect of an interpolant $p(t)$ of the computed solution in some previous interval $[t_{n-1}, t_{n-1}+h]$ given by $\delta(t) = p'(t) - f(t, p(t))$ is a continuous function. Hence, provided that the considered interpolant has order five and also that $t_H - t_{n-1} \leq 1.35h$ to ensure that the reliability of the interpolant, $\|\delta(t_H)\|$ will not be large and $\varepsilon_c$ has the size of the TOL. Now the integration may be restarted from $(t_H, p(t_H))$.

In the case of a first order discontinuity at $(t_d, y(t_d))$, $f(t, y(t))$ has a bounded discontinuity when $y(t)$ crosses the switching surface $g(t, y) = 0$ and consequently the defect has also a discontinuity. A similar reasoning can be given along an interpolant $p(t)$ of the numerical solution and it can be seen that now $\varepsilon_c$ may be estimated by the jump of $f(t, p(t))$ in the crossing of the switching surface $\Delta f_d = f_+(t_d, z_0(t_d)) - f_-(t_d, z_0(t_d))$. Once again, for the DOPRI54 pair, by assuming that $t_H - t_{n-1} \leq 1.35h$, we may estimate the jump of the interpolant in the critical interval $[t_L, t_H]$ by $\Delta f_d = f(t_H, p(t_H)) - f(t_L, p(t_L))$ and reduce, if necessary, the size of the critical interval before to restart the integration.

## 7 Numerical experiments

In order to test the techniques proposed for the pre–location and location of the discontinuities, we have integrated a number of initial value problems for which either the derivative function or else some of its derivatives have a finite jump. Here we present the results obtained with the following four problems

whose behaviour can be a representative test for the above techniques.

**Problem 1.** (see [10, pp. 237])

$$\begin{cases} y' = -\mathrm{sgn}(t)|1-|t||y^2, \\ y(-2) = 2/3, \qquad\qquad t \in [-2,2]. \end{cases}$$

This problem has three discontinuity points. Two second order discontinuities at $t = \pm 1$ and a first order discontinuity at $t = 0$.

**Problem 2.** (see [8])

$$\begin{cases} y' = \sum_{i=1}^{3} |y-i|, \\ y(0) = 1, \qquad\qquad t \in [0,1]. \end{cases}$$

The solution $y(t)$ is a monotonically increasing function with second order discontinuities when it crosses the straight lines $y = i$, $i = 2, 3$. This happens for $t_i = i \log(3/2) = 0.405465\, i$ hence there are two second order discontinuities at $t_1, t_2$.

**Problem 3.** (see [13])

$$\begin{cases} x'' = 0, \\ y'' = -g, \ g = 9.81, \\ x(0) = 0, x'(0) = 0, \\ y(0) = 2, y'(0) = 0, \quad t \in [0,1]. \end{cases}$$

It is a linear system that appears when considering a bouncing ball down a ramp. If the ball hit the ramp $x + y - 1 = 0$ at time $t = t^*$, it rebounds and its subsequent motion is described by another IVP for the same ODEs with new initial conditions

$$(x(t^*), x'(t^*), y(t^*), y'(t^*)) \to (x(t^*), -ky'(t^*), y(t^*), kx'(t^*)),$$

where $k \in (0,1)$ is called the coefficient of restitution.

In order to solve this problem with our codes, we have redefined the problem by taking

$$y'' = \begin{cases} -g, \text{ if } x + y - 1 \geq 0, \\ -Mg, \text{ if } x + y - 1 < 0, \end{cases}$$

with a large value of $M$, so that we introduce a first order discontinuity when $g(x,y) = x + y - 1$ changes the sign and the codes can locate the points $t^*$. In our experiments we have taken $k = 0.35$ and $M = 5000$.

We have integrated the above problems with variable stepsize codes based on the DOPRI54 formula for tolerances (with the same absolute and relative error tolerance) ranging from $10^{-5}$ to $10^{-8}$. This pair has been implemented in several modes:

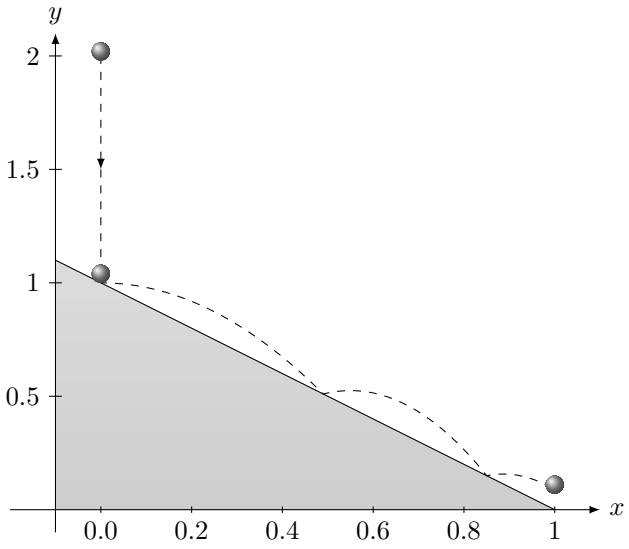



Figure 2: Problem 3. Bouncing ball down a ramp

**Std** With no discontinuity control technique.

**Pair–Pair** Prelocating and locating the discontinuities by means of pairs of forms.

**Def–Pair** Prelocating the discontinuities by using the defect and locating them with pairs.

**Def–Def** Prelocating and locating the discontinuities by using the defect.

All numerical experiments have been carried out with initial step–size $h_0 = 0.1$ and for the test to stop the bisection method we have chosen $t_H - t_L \leq \mathrm{TOL}/8$.

In Figures 3 and 4 we display the efficiency plots (number of evaluations of the derivative function (nfcn) against the error of the numerical solution at the end point of the integration interval (GE)) for Problems 1 and 2, comparing the performance of the above four implementations. It is clear that the use of the proposed techniques improve the efficiency of the code. The technique based on pairs of forms is in general more efficient since it requires fewer additional evaluations of the derivative function.

Also, in order to measure the reliability of the proposed techniques we also give in Tables 1 and 2 the error of each method in the localization of the last discontinuity point for the two first problems.

In the numerical integration of problem 1 with TOL=$10^{-6}$, the discontinuity at $t = 1$ was not detected (Table 2) with the Def–Pair code because the

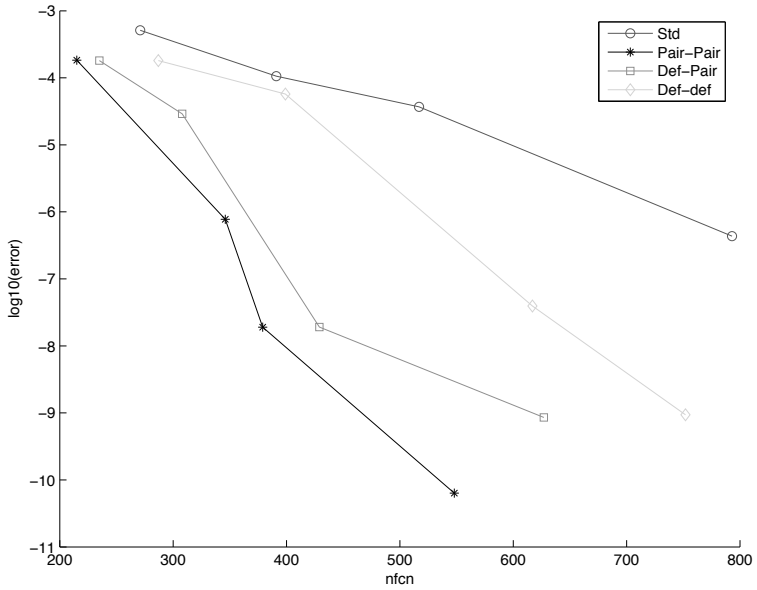



Figure 3: Efficiency plot for problem 1

Table 1: Problem 1. Error in the discontinuity location

| TOL | Pair–Pair | Def–Pair | Def–Def |
|---|---|---|---|
| $10^{-5}$ | — | — | — |
| $10^{-6}$ | 7.9e−08 | — | 7.8e−8 |
| $10^{-7}$ | 1.4e−09 | 7.8e−10 | 2.1e−09 |
| $10^{-8}$ | 4.8e−10 | 1.9e−10 | 8.5e−10 |

DOPRI54 advances a successful step from $t_n = 0.87352$ to $t_{n+1} = 0.99636$ and $t_{n+1}$ is very close to the last discontinuity point. Note that if a grid point $t_n$ hits a discontinuity point of order $\geq 2$ (function $f$ is continuous), the numerical approximation and the error estimation do not suffer a lack of order and it is not possible to detect the discontinuity with our techniques.

Regarding Problem 3, for the shake of comparison, we have also integrated it with the function `ode45` of the `matlab` package activating the detection of events. For tolerance TOL=$10^{-6}$, the computational cost, in terms of number of evaluations of the derivative function $f$, of our code (Pair–Pair) and the `ode45` are very similar (130 and 136 respectively). In both cases the discontinuities are located and the numerical solution is obtained with the required precision (see Figure 2).

**Conclusions**

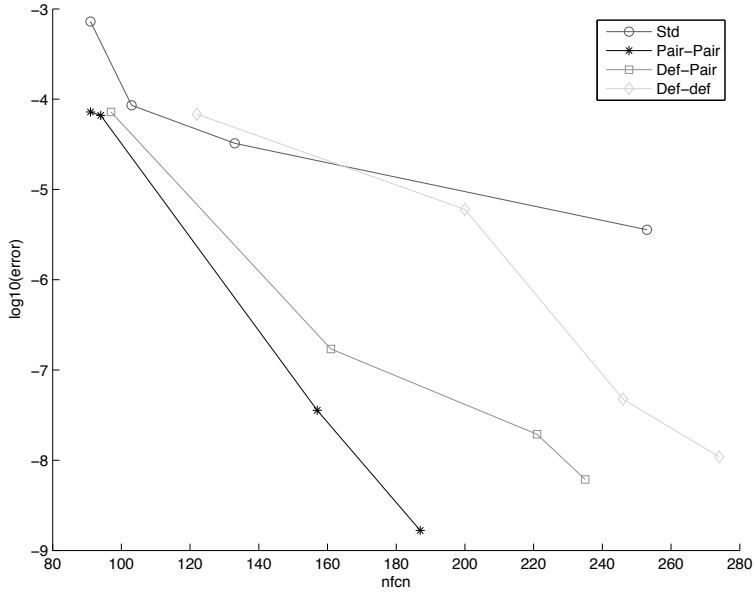



Figure 4: Efficiency plot for problem 2

Table 2: Problem 2. Error in the discontinuity location

| TOL | Pair–Pair | Def–Pair | Def–Def |
|---|---|---|---|
| $10^{-5}$ | 1.42e−07 | 1.37e−05 | 1.53e−05 |
| $10^{-6}$ | 1.23e−08 | 5.35e−09 | 1.19e−05 |
| $10^{-7}$ | 1.71e−09 | 1.67e−09 | 7.89e−09 |
| $10^{-8}$ | 7.16e−11 | 2.01e−10 | 3.92e−08 |

Several techniques for detection, accurate location and reliable crossing of discontinuities for the numerical solution of IVPs by means of adaptive RK codes have been presented and compared. These techniques have been justified by using asymptotic reasonings. They have been applied to the well known DOPRI54 embedded pair of formulas of Dormand and Prince. Numerical experiments show that the proposed algorithms are very reliable and efficient whenever the stepsize control can detect the presence of the discontinuity and the detection–location process is activated. Further in all cases in which the detection–location process was activated the discontinuity was located with the required accuracy and with low additional computational cost.

**References**

[1] R. W. Brankin, I. Gladwell and L.F. Shampine *RKSUITE: A suite of explicit Runge–Kutta codes* Contiributions in Numerical Mathematics (ed. R. Agarwal) World Scientific Series in Applicable Analysis 2, World Scientific, Singapore, p. 85-98, (1993).

[2] M. Calvo, J.I. Montijano and L. Rández, *A fifth–order interpolant for the Dormand and Prince Runge–Kutta method*, J. Comp. Appl. Math. 29 (1990), p. 91–100.

[3] M. Calvo, J.I. Montijano and L. Rández, *On the solution of discontinuous IVPs by adaptive Runge–Kutta codes*, Numerical Algorithms 33 (2003), p. 163–182.

[4] M. Calvo, J.I. Montijano and L. Rández, *On the numerical solution of IVPs with discontinuities by adaptive Runge–Kutta codes*, Technical report, Department of Matemática Aplicada (2002), http://pcmap.unizar.es/numerico/reports.php.

[5] M.B. Carver, *Efficient integration over discontinuities in ordinary differential equation simulations*, Math. Comput. Simul., 20, 3, (1978), pp. 190-196.

[6] J. Dormand and P. Prince, *A family of embedded Runge–Kutta formulae*, J. Comp. Appl. Math. 6 (1980), 19–26.

[7] D. Ellison, *Efficient Automatic integration of Ordinary Differential with Discontinuities*, Math. and Comput. in Simul., 23,1,(1981), pp. 12-20.

[8] W.H. Enright, K.R. Jackson, S.P. Nørsett and P.G. Thomsen, *Effective Solution of Discontinuous IVPs Using a Runge–Kutta Formula Pair with Interpolants*, Appl. Math. and Comp., 27, (1988), pp. 313-335.

[9] C.W. Gear, O. Østerby, *Solving Ordinary Differential Equations with Discontinuities*, ACM Trans. Math. Soft., 10, 1, (1984), pp. 23–44.

[10] E. Hairer, S.P. Nørsett and G. Wanner, *Solving Ordinary Differential Equations I, Nonstiff Problems, second revised edition* (Springer–Verlag, Berlin, 1992).

[11] R. Mannshardt, *One–step Methods of Any Order for Ordinary Differential Equations with Discontinuous Right–Hand Sides*, Numer. Math., 31, (1978), pp. 131-152.

[12] P.G. O'Regan, *Step size adjustement at discontinuities for fourth order Runge–Kutta methods*, Comput. J., 13, 4, (1970), pp. 401-404.

[13] L. F. Shampine and M. W. Reichelt, *The MATLAB ODE suite* , *SIAM Journal on Scientific Computing*, 18 , 1, (1997) p. 1-22.

[14] L. F. Shampine, I. Gladwell and S. Thompson, *Solving ODEs with MATLAB*, Cambridge, 2003.

*Bol. Soc. Esp. Mat. Apl.*
*n° 44(2008), 55–79*

# ANALYSIS OF A CELL SYSTEM WITH FINITE DIVISIONS

B. PERTHAME* AND T.M. TOUAOULA†

* UPMC, Univ. Paris 06, UMR 7598
Laboratoire Jacques-Louis Lions F-75005, Paris, France
and Institut Universitaire de France

† Département de Mathématiques, Université de Tlemcen,
BP 119, 13000, Algérie

perthame@ann.jussieu.fr touaoula_tarik@yahoo.fr

**Abstract**

In this paper we consider a model of cell division which describes the continuous growth of cells and their division. The originality of the model is to assume only a finite number of possible divisions. We are concerned with three questions.

First, we prove both the existence of a stable steady dynamics (first positive eigenvector) and, second, we show that the long time asymptotics of any solution can be described by this stable steady dynamics. Our third result concerns the limit of a high number of possible divisions; in the particular case where the division rates are independent of the number of earlier divisions, we show that, in this limit, we recover the classical cell division model for equal mitosis.

**Key words:** *cell division equations, growth processes, general relative entropy, asymptotic analysis*

**AMS subject classifications:** *35B40 35F10 92D25*

## 1 Introduction and main results

This paper is concerned with the system of equations

$$\begin{cases} \dfrac{\partial n_1}{\partial t}(t,x) + \dfrac{\partial n_1}{\partial x}(t,x) + B_1(x)n_1(t,x) = 2\displaystyle\sum_{i=1}^{I-1} B_i(2x)n_i(t,2x), \quad t>0, \quad x\geq 0, \\ \dfrac{\partial n_i}{\partial t}(t,x) + \dfrac{\partial n_i}{\partial x}(t,x) + B_i(x)n_i(t,x) = 2B_{i-1}(2x)n_{i-1}(t,2x), \quad 2\leq i\leq I, \\ n_i(t,0) = 0, \qquad 0\leq i\leq I, \\ n_i(0,x) = n_i^0(x) \in L^1(\mathbb{R}^+), \qquad 0\leq i\leq I. \end{cases} \tag{1}$$

Fecha de recepción: 04/04/2008. Aceptado (en forma revisada): 17/05/2008.

It arises in particular as a model for size-structured populations, $n_i(t,x)$ denotes the $i^{th}$ generation of cells of size $x$ at time $t$. We choose to let $x \in \mathbb{R}^+$ because on the one hand the dynamics will set by itself the 'average size' and the variance around it, with an exponential decay for large sizes of the population density, and on the other hand the variability in living systems can hardly lead to a meaningful maximum size. In the above model we take into account that cells can divide at most $I$ times, with $I$ large. The $i$-th generation cells grow at a constant rate but also divide into two cells of equal size at different rates $B_i(x)$ when mitosis occurs, see [12, 2, 18] and the references therein. One of them, the daughter cell, resumes a cycle at generation 1 while the mother cell turns to generation $i+1$. Our motivation comes from several areas of biology, cell cultures, tumor growth, where such models are very usual, see for instance,[3, 4, 6, 19, 20, 22]. This type of model is reminiscent from fragmentation equations in physics [1, 5, 7, 8, 10, 9, 11]. According to well accepted biological theories, our model contains a particular feature which is to assume that cells are programmed for a finite and fixed a priori number of divisions.

An interesting feature of this problem is the asymptotic behavior of $n_i(t,x)$ that gives the invasive capacity of the population and thus a fitness measure of populations. Our first purpose is to study existence of the first eigenpair $(\lambda, (N_i(x))_{1\le i\le I})$ given by

$$\begin{cases} \dfrac{\partial N_1(x)}{\partial x} + (B_1(x)+\lambda)N_1(x) = 2\displaystyle\sum_{i=1}^{I-1} B_i(2x)N_i(2x), \quad x \ge 0, \\ \dfrac{\partial N_i(x)}{\partial x} + (B_i(x)+\lambda)N_i(x) = 2B_{i-1}(2x)N_{i-1}(2x), \quad 2 \le i \le I, \\ N_i(0) = 0, \quad \text{and} \quad \displaystyle\sum_{i=1}^{I}\int_0^\infty N_i(x)dx = 1, \quad N_i(x) > 0, \quad \forall x > 0, \end{cases} \tag{2}$$

with $\lambda$ the first eigenvalue (sometimes called the Malthus parameter in biology). Notice that this can be understood as a systems of $(I-1)$ equations for $1 \le i \le I-1$, and the equation of $N_I$ is in fact decoupled. Notice this is not a system of delay differential equations because the knowledge of larger sizes is needed to determine the density at size $x$. For this reason the analysis uses tools closer to Laplace type equations than ODEs. Our second purpose is to make rigorous that this density plays the role of the so called stable steady dynamics, that is after a time renormalization to keep the density bounded from above and below, all the solutions to (1) converge to a multiple of the solution

to (2). To be more precise, we need to introduce the adjoint operator

$$\begin{cases} \dfrac{\partial \phi_i(x)}{\partial x} - (B_i(x)+\lambda)\phi_i(x) = -B_i(x)\Big(\phi_1(\frac{x}{2}) + \phi_{i+1}\left(\frac{x}{2}\right)\Big), \qquad 1 \le i \le I-1, \\ \dfrac{\partial \phi_I(x)}{\partial x} - (B_I(x)+\lambda)\phi_I(x) = 0, \\ \displaystyle\sum_{i=1}^{I} \int_0^\infty N_i(x)\phi_i(x)dx = 1, \quad \phi_i > 0, \quad 1 \le i \le I-1. \end{cases} \tag{3}$$

It plays a fundamental role in the analysis of the dynamics of (1), because its solution allows us to define a conservation law for the equation for $n_i(t,x)$ :

$$\sum_{i=1}^{I} \int_0^\infty n_i(t,x)e^{-\lambda t}\phi_i(x)dx = \sum_{i=1}^{I} \int_0^\infty n_i^0(x)\phi_i(x)dx.$$

Now we state the main theorems of the paper.

**Theorem 1.1** *Assume that for* $1 \le i \le I$, $B_i \in C(\mathbb{R}^+)$, *and* $0 < b_m = \min_{i,x} B_i(x)$, $B_M = \max_{i,x} B_i(x) < \infty$. *There is a unique solution* $\big((N_i), \lambda, (\phi_i)\big)_{1\le i\le I}$ *to equations (2), (3), with* $N_i, \phi_i \in C^1(\mathbb{R}^+)$, *and we have*

$$b_m - \frac{2B_M}{I} \le \lambda \le B_M. \tag{4}$$

*Moreover,* $\forall \mu \in [0,\lambda)$, *and* $C$ *a universal positive constant, we have the bounds*

$$\sum_{i=1}^{I} \int_0^\infty N_i(x)e^{\mu x}dx \le \frac{\lambda}{\lambda-\mu}, \quad \sum_{i=1}^{I} N_i(x)e^{\mu x} \le \lambda + \frac{B_M\lambda}{\lambda-\mu}, \quad \text{and} \quad \frac{\partial}{\partial x}\sum_{i=1}^{I} N_i(x)e^{\mu x} \in L^1(\mathbb{R}^+) \tag{5}$$

$$0 < \phi_i(x) \le C(1+x), \quad 1 \le i \le I-1, \quad x \ge 0. \tag{6}$$

The long time behavior follows and we have the

**Theorem 1.2** *Assume that, for* $1 \le i \le I$, *the initial data satisfies* $n_0^i \in L^1(\mathbb{R}^+, \phi_i(x)dx)$, *then the solution to (1) tends to a steady state. More precisely, with* $\rho = \displaystyle\sum_{i=1}^{I-1} \int_0^\infty n_i^0(x)\phi_i(x)dx$, *we have*

$$\lim_{t\to\infty} \sum_{i=1}^{I-1} \int_0^\infty |n_i(t,x)e^{-\lambda t} - \rho N_i(x)|\phi_i(x)dx = 0,$$

$$\lim_{t\to\infty} \int_0^A |n_I(t,x)e^{-\lambda t} - \rho N_I(x)|dx = 0, \quad \forall A > 0.$$

Our third purpose is devoted to the limit of $I \to \infty$ with equal division rates

$$B_i(x) = B(x), \quad \forall \ 1 \le i \le I, \qquad 0 < b_m \le B(x) \le B_M < \infty. \tag{7}$$

We prove uniform bounds independently of $I$ and show that the solution $\left(\lambda^I, \sum_{i=1}^{I} N_i^I, \frac{1}{I-1}\sum_{i=1}^{I-1}\phi_i^I\right)$ converges to $(\bar{\lambda}, N, \phi)$ solution to the problem

$$\begin{cases} \dfrac{\partial N(x)}{\partial x} + (B(x)+\bar{\lambda})N(x) = 4B(2x)N(2x), \quad x \ge 0, \\ N(0) = 0, \quad \int_0^\infty N(x)dx = 1 \text{ and } N(x) > 0 \text{ for } x > 0, \end{cases} \tag{8}$$

with the associated adjoint problem

$$\begin{cases} \dfrac{\partial \phi(x)}{\partial x} - (B(x)+\bar{\lambda})\phi(x) = -2B(x)\phi\left(\frac{x}{2}\right), \quad x \ge 0, \\ \phi(x) > 0 \ \text{ for } \ x \ge 0, \ \text{ and } \ \int_0^\infty N(x)\phi(x)dx = 1. \end{cases} \tag{9}$$

The same is true for the evolution equation, and $\sum_{i=1}^{I} n_i^I(t,x)$ converges to $n(t,x)$ solution to the problem

$$\begin{cases} \dfrac{\partial n}{\partial t} + \dfrac{\partial n}{\partial x} + B(x)n = 4B(2x)n(t,2x), \quad x \ge 0, \\ n(t,0) = 0, \\ n(0,x) = n_0(x). \end{cases} \tag{10}$$

These problems are the standard equal mitosis equations and were treated in [17]. Noticing that the bounds stated in Theorem 1.1 are uniform, we can derive the following result

**Theorem 1.3** *Assume (7), then the Malthus parameter $\lambda^I$ satisfies*

$$b_m - \frac{2B_M}{I} \le \lambda^I \le \bar{\lambda}.$$

*As $I \to \infty$, we have $\lambda^I \to \bar{\lambda}$,*

$$\sum_{i=1}^{I} N_i^I \to N, \ \textit{(in } L^1(\mathbb{R}^+)\textit{)}, \quad \textit{and} \ \ \frac{1}{I-1}\sum_{i=1}^{I-1}\phi_i^I \to \phi \ \textit{(in } C^0_{\text{loc}}(\mathbb{R}^+)\textit{)}.$$

Finally, we state the following theorem, which gives the relationship between the two evolution problems namely, (1) and (10). We have

**Theorem 1.4** *Assume that (7) holds, and the initial data $(n_i^0)_{1\le i\le I}$ is a decreasing sequence in $i$, such that $\sum_{i=1}^{I} n_i^0$ converges to $n_0$ as $I$ tends to infinity, then*

$$\sum_{i=1}^{I} n_i^I(t,x) \to n(t,x) \ \ \textit{as} \ \ I \to \infty \tag{11}$$

*strongly in* $L^1\big(0,T;L^1_{loc}(\mathbb{R}^+)\big)$, $\forall T>0$.

The paper is organized as follows. In Section 2, we study the eigenproblem (2)–(3) and we prove the Theorem 1.1. Section 3 is devoted to the proof of the long time behavior for the evolution problem in Theorem 1.2; here we mainly use the General Relative Entropy (GRE). The goal of Section 4 is to prove the asymptotic results as $I\to\infty$ stated in Theorems 1.3 and 1.4.

## 2 Mathematical analysis of the eigenproblem

The purpose of this section is to prove the existence of a first eigenvalue and a positive eigenvectors for the problem (2), (3) and to prove Theorem 1.1. We follow the proof in [17], using a solution $\big((N_i^L),\lambda^L,(\phi_i^L)\big)_{1\le i\le I}$ of the eigenvalue on a bounded interval $[0,L]$ and then passing to the limit.

### 2.1 Bounded domain

The problem on a bounded interval [0,L] is to find the first eigenvalue $\lambda^L$ and $(N_i^L,\phi_i^L)_{1\le i\le I}$ such that

$$\begin{cases} \dfrac{\partial N_1^L(x)}{\partial x}+(B_1(x)+\lambda^L)N_1^L(x)=2\displaystyle\sum_{i=1}^{I-1}B_i(2x)N_i^L(2x)1_{\{2x\le L\}}, \quad 0\le x\le L,\\ \dfrac{\partial N_i^L(x)}{\partial x}+(B_i(x)+\lambda^L)N_i^L(x)=2B_{i-1}(2x)N_{i-1}^L(2x)1_{\{2x\le L\}}, \quad 2\le i\le I,\\ N_i^L(0)=0, \displaystyle\sum_{i=1}^{I}\int_0^L N_i^L(x)dx=1, \quad \text{and} \quad N_i^L(x)>0, \quad \forall x>0, \end{cases} \tag{12}$$

and

$$\begin{cases} -\dfrac{\partial\phi_i^L(x)}{\partial x}+(B_i(x)+\lambda^L)\phi_i^L(x)=B_i(x)\Big(\phi_1^L\left(\frac{x}{2}\right)+\phi_{i+1}^L\left(\frac{x}{2}\right)\Big), \quad 1\le i\le I-1,\\ -\dfrac{\partial\phi_I^L(x)}{\partial x}+(B_I(x)+\lambda^L)\phi_I^L(x)=0,\\ \phi_i^L(L)=0, \quad \text{and} \quad \displaystyle\sum_{i=1}^{I}\int_0^L N_i^L(x)\phi_i^L(x)dx=1, \quad \phi_i^L(x)>0. \end{cases} \tag{13}$$

Notice that $\phi_I^L=0$, and similarly to system (2), this is in fact a system of $(I-1)$ equations. Arguing as in [13] or [18] p. 176, by the Krein-Rutman Theorem, we obtain directly the existence of these eigenelements. We do not detail this point and concentrate on the limit $L\to\infty$.

### 2.2 The limit $L\to\infty$

We begin with the existence of a limit as $L\to\infty$ of the eigenfunctions and eigenvalues constructed above. We first establish the some uniform bounds,

then pass to the limit and prove uniqueness.

**The first step. Uniform bounds for** $(N_i^L)$.

**Lemma 1** *The solution to (12) satisfies*

$$b_m - \frac{2B_M}{I} - \frac{1}{L} \leq \lambda^L \leq B_M.$$

*Moreover,* $\forall \mu \in [0, \lambda^L)$, *we have*

$$\sum_{i=1}^{I} \int_0^\infty N_i^L(x) e^{\mu x} dx \leq \frac{\lambda^L}{\lambda^L - \mu}, \quad \sum_{i=1}^{I} N_i^L(x) e^{\mu x} \leq \lambda^L + \frac{1}{L} + \frac{B_M \lambda^L}{\lambda^L - \mu},$$

$$\text{and} \quad \frac{\partial}{\partial x} \sum_{i=1}^{I} N_i^L(x) e^{\mu x} \in L^1(\mathbb{R}^+).$$

**Proof** First of all summing the equations of problem (12),

$$\sum_{i=1}^{I} \frac{\partial N_i^L}{\partial x} + \sum_{i=1}^{I} (B_i(x) + \lambda^L) N_i^L = 4 \sum_{i=1}^{I-1} B_i(2x) N_i^L(2x) 1_{\{2x \leq L\}}. \tag{14}$$

**Estimate on** $\lambda^L$**:** After integrating (14) in $x$ we have

$$\sum_{i=1}^{I} N_i^L(L) + \lambda^L \sum_{i=1}^{I} \int_0^L N_i^L(x) dx = \sum_{i=1}^{I-1} \int_0^L B_i(x) N_i^L(x) dx - \int_0^L B_I(x) N_I^L(x) dx, \tag{15}$$

so,

$$\lambda^L \leq \sum_{i=1}^{I-1} \int_0^L B_i(x) N_i^L(x) dx \leq B_M. \tag{16}$$

Now, multiplying the equation (14) by $x$ and integrating,

$$\begin{aligned} \sum_{i=1}^{I} L N_i^L(L) - \sum_{i=1}^{I} \int_0^L N_i^L(x) dx + \lambda^L \sum_{i=1}^{I} \int_0^L x N_i^L(x) dx \\ = - \int_0^L x B_I(x) N_I^L(x) dx \leq 0. \end{aligned} \tag{17}$$

Multiplying (15) by $L$ and substracting it to (17), we obtain

$$\lambda^L \sum_{i=1}^{I} \int_0^L x N_i^L(x) dx + L \sum_{i=1}^{I-1} \int_0^L B_i(x) N_i^L(x) dx - L \int_0^L B_I(x) N_I^L(x) dx \leq L\lambda^L + 1,$$

so,

$$\lambda^L \geq \sum_{i=1}^{I-1} \int_0^L B_i(x) N_i^L(x) dx - \int_0^L B_I(x) N_I^L(x) dx - \frac{1}{L}. \tag{18}$$

Moreover for $2 \leq i \leq I$, and after integration the equation (12) we obtain

$$\int_0^L B_{i-1}(x)N_{i-1}^L(x)dx \geq \int_0^L B_i(x)N_i^L(x)dx, \tag{19}$$

thus we have

$$I\int_0^L B_I(x)N_I^L(x)dx \leq \sum_{i=1}^{I}\int_0^L B_i(x)N_i^L(x)dx \leq B_M,$$

therefore

$$\int_0^L B_I(x)N_I^L(x)dx \quad \leq \quad \frac{B_M}{I}. \tag{20}$$

Combining (18) with (20) we conclude that

$$\lambda^L \geq b_m - \frac{2B_M}{I} - \frac{1}{L}.$$

**Estimate on $N_i^L$:** We multiply the equation (12) by $e^{\mu x}$ and integrate on $[0, L]$,

$$\begin{aligned}
&\sum_{i=1}^{I} N_i^L(L)e^{\mu L} + (\lambda^L - \mu)\sum_{i=1}^{I}\int_0^L e^{\mu x}N_i^L(x)dx + \sum_{i=1}^{I}\int_0^L B_i(x)N_i^L(x)e^{\mu x}dx \\
= \quad & 4\sum_{i=1}^{I-1}\int_0^L B_i(2x)N_i^L(2x)e^{\mu x}1_{\{2x\leq L\}}dx, \\
\leq \quad & 2\sum_{i=1}^{I-1}\int_0^L B_i(2x)N_i^L(2x)(1+e^{2\mu x})1_{\{2x\leq L\}}dx, \\
\leq \quad & \sum_{i=1}^{I-1}\int_0^L B_i(x)N_i^L(x)(1+e^{\mu x})dx,
\end{aligned}$$

therefore

$$\begin{aligned}
&\sum_{i=1}^{I} N_i^L(L)e^{\mu L} + (\lambda^L - \mu)\sum_{i=1}^{I}\int_0^L e^{\mu x}N_i^L(x)dx \\
\leq \quad & \sum_{i=1}^{I-1}\int_0^L B_i(x)N_i^L(x)dx - \int_0^L B_I(x)N_I^L(x)e^{\mu x}dx.
\end{aligned}$$

Now using (15), we have

$$\sum_{i=1}^{I} N_i^L(L)e^{\mu L} + (\lambda^L - \mu)\sum_{i=1}^{I}\int_0^L e^{\mu x}N_i^L(x)dx \leq \sum_{i=1}^{I} N_i^L(L) + \lambda^L.$$

As a consequence we find

$$\sum_{i=1}^{I}\int_0^L e^{\mu x}N_i^L(x)dx \le \lambda^L/(\lambda^L-\mu).$$

Arguing as in the above estimate but after integration between 0 and $x$, (see also (17) which implies that $\sum_{i=1}^{I} N_i^L(L) \le \frac{1}{L}$ ) we obtain

$$\sum_{i=1}^{I} e^{\mu x}N_i^L(x) \le (\lambda^L+\frac{1}{L}) + B_M\lambda^L/(\lambda^L-\mu),$$

and similarly after multiplying (12) by $e^{\mu x}$ and integration,

$$\sum_{i=1}^{I}\int_0^L \left|\frac{\partial}{\partial x}(e^{\mu x}N_i^L(x))\right| dx \le |\lambda^L-\mu|\sum_{i=1}^{I}\int_0^L e^{\mu x}N_i^L(x)dx$$
$$+ \quad \sum_{i=1}^{I}\int_0^L B_i(x)N_i^L(x)e^{\mu x}dx + 4\sum_{i=1}^{I-1}\int_0^L B_i(2x)N_i^L(2x)e^{\mu x}1_{\{2x\le L\}}dx,$$

therefore $\sum_{i=1}^{I}\frac{\partial}{\partial x}(e^{\mu x}N_i^L(x)) \in L^1(\mathbb{R}^+)$ for all $\mu \in [0,\lambda^L)$. □

**The second step. Positivity of $N_i$.** Now, as $L\to\infty$, one can extract a subsequence (still labelled by $L$ to simplify notations) such that $\lambda^L\to\lambda$ and $N_i^L$ converges strongly to $N_i \ge 0$ (thanks to the estimates gathered above), eigen pairs for (2) in $C^1(\mathbb{R}^+)$ and $1\le i\le I$. The exponential control shows that $\sum_{i=1}^{I}\int_0^\infty N_i(x)dx = 1$. It remains to prove the positivity of $(N_i)_{i=1}^I$. Let us denote $a = inf\{x \ \ s.t., N_1(x) > 0\}$. Then using the method of characteristics, we have

$$\begin{aligned} N_1(x) &= 2e^{-J_1(x)}\sum_{i=1}^{I-1}\int_0^x B_i(2y)N_i(2y)e^{J_1(y)}dy, \\ &\ge 2e^{-J_1(x)}\int_0^x B_1(2y)N_1(2y)e^{J_1(y)}dy, \end{aligned}$$

with

$$J_i(x) = \int_0^x (B_i(y)+\lambda)dy.$$

Thus, for $x > \frac{a}{2}$, we deduce that $N_1(x) > 0$ since there is an open subset ($N_1$ is continuous) where $N_1(2y) > 0$ in this integral. Then $a = 0$ and $N_1(x) > 0$ for $x > 0$. In addition, for $2\le i\le I$, we have

$$N_i(x) = 2e^{-J_i(x)}\int_0^x B_{i-1}(2y)N_{i-1}(2y)e^{J_i(y)}dy,$$

therefore $N_i(x) > 0$ for $x > 0$ for all $1 \leq i \leq I$.

**The third step. The limit of $\phi_i^L$.** Still follows from the Krein-Rutman theorem. We begin by establishing the following lemma.

**Lemma 2** *The solution to (13) satisfies*

$$\sum_{i=1}^{I-1} N_i^L(x)\phi_i^L(x) \leq \frac{2}{x}. \tag{21}$$

*Moreover, for $1 \leq i \leq I-1$ and with a constants $C$ independent of $L$, we have*

$$0 < \phi_i^L(x) \leq C(1+x). \tag{22}$$

**Proof** We prove the first inequality, recalling that $\phi_I^L \equiv 0$. Firstly using (12) and (13), we observe that the function $\sum_{i=1}^{I-1} N_i^L(x)\phi_i^L(x)$ satisfies

$$\begin{cases} \sum_{i=1}^{I-1} \dfrac{\partial(N_i^L\phi_i^L)(x)}{\partial x} = 2\sum_{i=1}^{I-1} B_i(2x)N_i^L(2x)\phi_1^L(x) - \sum_{i=1}^{I-1} B_i(x)N_i^L(x)\phi_1^L\left(\dfrac{x}{2}\right) \\ +2\sum_{i=1}^{I-2} B_i(2x)N_i^L(2x)\phi_{i+1}^L(x) - \sum_{i=1}^{I-2} B_i(x)N_i^L(x)\phi_{i+1}^L\left(\dfrac{x}{2}\right), \\ \sum_{i=1}^{I} N_i^L(0)\phi_i^L(0) = \sum_{i=1}^{I} N_i^L(L)\phi_i^L(L) = 0. \end{cases}$$

Arguing as in [17], we find after integration of the above equation and a change of variable in the right hand side that

$$\sum_{i=1}^{I-1} N_i^L(x)\phi_i^L(x) =$$

$$\int_x^{min(2x,L)} \left(\phi_1^L\left(\frac{z}{2}\right)\sum_{i=1}^{I-1} B_i(z)N_i^L(z) + \sum_{i=1}^{I-1} B_i(z)N_i^L(z)\phi_{i+1}^L\left(\frac{z}{2}\right)\right) dz, \tag{23}$$

$$\int_0^L \int_x^{min(2x,L)} \left(\phi_1^L\left(\frac{z}{2}\right)\sum_{i=1}^{I-1} B_i(z)N_i^L(z) + \sum_{i=1}^{I-1} B_i(z)N_i^L(z)\phi_{i+1}^L\left(\frac{z}{2}\right)\right) dzdx = 1,$$

so,

$$\int_0^L z\left(\phi_1^L\left(\frac{z}{2}\right)\sum_{i=1}^{I-1} B_i(z)N_i^L(z) + \sum_{i=1}^{I-1} B_i(z)N_i^L(z)\phi_{i+1}^L\left(\frac{z}{2}\right)\right) dz = 1.$$

We can use this bounds in the right side of (23) again, and we obtain the first inequality of the lemma. Secondly we prove an intermediate upper bound, namely

$$\sup_{0\leq x\leq A} \sum_{i=1}^{I-1} \phi_i^L(x) \leq C(A). \tag{24}$$

For this inequality we use (3) and first derive, for $y < z$ and $1 \leq i \leq I-1$,

$$\phi_i^L(\frac{z}{2})e^{-J_i(\frac{z}{2})} \leq \phi_i^L(\frac{y}{2})e^{-J_i(\frac{y}{2})} \leq \phi_i^L(\frac{y}{2}). \tag{25}$$

Summing the equations of the dual problem,

$$\sum_{i=1}^{I-1} \frac{\partial \phi_i^L(x)}{\partial x} - \sum_{i=1}^{I-1} \left(B_i(x) + \lambda^L\right) \phi_i^L(x) = \\ - \left(\sum_{i=1}^{I-1} B_i(x)\right) \phi_1^L\left(\frac{x}{2}\right) - \sum_{i=1}^{I-2} B_i(x)\phi_{i+1}^L\left(\frac{x}{2}\right), \tag{26}$$

after integration over $y$ and $x^L$, with $y \leq x_L$,

$$\begin{aligned} & \sum_{i=1}^{I-1} \phi_i^L(x^L) - \sum_{i=1}^{I-1} \phi_i^L(y) - \sum_{i=1}^{I-1} \int_y^{x^L} (B_i(x) + \lambda^L)\phi_i^L(x)dx \\ = \; & -\int_y^{x^L} (\sum_{i=1}^{I-1} B_i(x))\phi_1^L\left(\frac{x}{2}\right)dx - \sum_{i=1}^{I-2} \int_y^{x^L} B_i(x)\phi_{i+1}^L\left(\frac{x}{2}\right)dx. \end{aligned}$$

We now insert this bound into the inequality

$$\sum_{i=1}^{I-1} \phi_i^L(y) \leq \sum_{i=1}^{I-1} \phi_i^L(x^L) + (I-1)B_M\phi_1^L\left(\frac{y}{2}\right)e^{J_1(\frac{x_L}{2})}x_L + \\ B_M \sum_{i=1}^{I-2} \phi_{i+1}^L\left(\frac{y}{2}\right)e^{J_{i+1}(\frac{x_L}{2})}x_L.$$

Moreover for all $1 \leq i \leq I$ we have,

$$e^{J_i(x_L/2)} \leq e^{B_M x_L},$$

and so,

$$\sum_{i=1}^{I-1} \phi_i^L(y) \leq \sum_{i=1}^{I-1} \phi_i^L(x^L) + (I-1)B_M e^{2B_M x_L} x_L \sum_{i=1}^{I-1} \phi_i^L\left(\frac{y}{2}\right),$$

choosing $x_L$ such that

$$(I-1)B_M e^{B_M x_L} x_L = \frac{1}{2}.$$

Note that $x_L$ is uniformly bounded ($c \leq x_L \leq C$). We thus obtain,

$$\sup_{0<y<x_L} \sum_{i=1}^{I-1} \phi_i^L(y) \leq 2 \sum_{i=1}^{I-1} \phi_i^L(x^L).$$

It remains to bound $\sum_{i=1}^{I-1}\phi_i^L(x^L)$. Indeed by using the fact that $\phi_i^L(x)e^{-J_i(x)}$ is decreasing and $e^{J_i(x)}$ is bounded we have

$$\phi_i^L(x_L) \le \phi_i^L(y)e^{2B_M C},$$

thus, for some $a > 0$,

$$\begin{aligned}\phi_i^L(x_L)\int_0^a N_i^L(y)dy &\le e^{2B_M C}\int_0^a \phi_i^L(y)N_i^L(y)dy,\\ &\le e^{2B_M C}\sum_{i=1}^{I-1}\int_0^a \phi_i^L(y)N_i^L(y)dy,\\ &\le e^{2B_M C},\end{aligned}$$

since $\int_0^a N_i^L(y)dy$ is uniformly positive, then the inequality (24) follows for $x \in (0, A)$ for all $A > 0$. Moreover we find a supersolution (independent of $L$) for the $\phi_i^L$ equation following an argument of [17]. For $x = L - y$ and denoting by $\bar\phi_i^L(y) = \phi_i^L(x)$. We have

$$\begin{cases}\dfrac{\partial\bar\phi_i^L(y)}{\partial y} + (\lambda + \bar B_i(y))\bar\phi_i^L(y) = \bar B_i(y)\Big(\bar\phi_1^L\left(\frac{L+y}{2}\right) + \bar\phi_{i+1}^L\left(\frac{L+y}{2}\right)\Big),\\ \phi_i^L(0) = 0,\end{cases}$$

for each $1 \le i \le I-1$. We notice that $\bar v_i(y) := \bar v(y) = C(L-y)$ for all $1 \le i \le I$, is a supersolution of the equation on $\bar\phi_i^L(y)$. Indeed for $L - y \ge \dfrac{1}{\lambda}$, we have

$$\begin{cases}\dfrac{\partial\bar v(y)}{\partial y} + (\lambda + \bar B_i(y))\bar v(y) - 2\bar B_i(y)\bar v\left(\frac{L+y}{2}\right) = C\big(-1 + \lambda(L-y)\big) \ge 0,\\ \bar v(0) > 0.\end{cases}$$

Henceforth, by the maximum principle, for all $1 \le i \le I-1$ we have $\phi_i(x) \le C(1+x)$. This concludes the proof of the upper bound of the Lemma 2. Thus we conclude that $\phi_i^L$ converges locally uniformly to a solution $\phi_i$ of (3) (this is local strong convergence) satisfying the same bounds as in the Lemma 2. Thanks to the uniform decay of $N_i^L$ at infinity faster than any polynomial, we deduce that

$$\sum_{i=1}^{I-1}\int_0^\infty \phi_i(x)N_i(x)dx = 1. \tag{27}$$

On the other hand, and from (27), there exists $1 \le i_0 \le I-1$ such that $\phi_{i_0}$ does not vanish and from (25), we deduce that $\phi_{i_0}$ does not vanish on some interval $[0, x_0]$, with $x_0 > 0$. In addition according to (3), and using the method of characteristics, we see that for all $1 \le i \le i_0$, $\phi_i$ does not vanish on $[0, x_0]$. Finally since $\phi_1 > 0$ and from the structure of the dual problem (3), we show that $\phi_i(x) > 0$ for all $x \ge 0$. This conclude the proof of the Lemma 2. □

Now it is easy to prove the estimates (4)-(6), so it remains to see the uniqueness of the limit.

**The fourth step. The Uniqueness of the limit.** By the same idea as in [13], given another eigenpair $(\bar{\lambda}, (\bar{N}_i)_{1\leq i\leq I})$ of this problem, and using the adjoint equation (3) for $N$ we have

$$\sum_{i=1}^{I-1}\frac{\partial(\bar{N}_i\phi_i)(x)}{\partial x}+(\bar{\lambda}-\lambda)\sum_{i=1}^{I-1}\bar{N}_i\phi_i(x)=2\sum_{i=1}^{I-1}B_i(2x)\bar{N}_i(2x)\phi_1(x)-$$
$$\sum_{i=1}^{I-1}B_i(x)\bar{N}_i(x)\phi_1\left(\frac{x}{2}\right)+2\sum_{i=1}^{I-2}B_i(2x)\bar{N}_i(2x)\phi_{i+1}(x)-\sum_{i=1}^{I-2}B_i(x)\bar{N}_i(x)\phi_{i+1}\left(\frac{x}{2}\right),$$

after integration, we find

$$(\bar{\lambda}-\lambda)\sum_{i=1}^{I-1}\int_0^\infty\bar{N}_i(x)\phi_i(x)dx=0,$$

which implies that $\bar{\lambda}=\lambda$. Next we prove that $N_i=\bar{N}_i$. By using the general relative entropy method, see [14, 15, 18], we conlclude that, for all $k>0$,

$$\sum_{i=1}^{I-1}\int_0^\infty B_i(2x)N_i(2x)\phi_1(x)$$
$$\cdot\left[\operatorname{sgn}\left(\frac{\bar{N}_1(x)}{N_1(x)}-k\right)\left(\frac{\bar{N}_i(2x)}{N_i(2x)}-k\right)-\left|\frac{\bar{N}_i(2x)}{N_i(2x)}-k\right|\right]dx=0,$$

and

$$\sum_{i=2}^{I-1}\int_0^\infty B_{i-1}(2x)N_{i-1}(2x)\phi_i(x)$$
$$\cdot\left[\operatorname{sgn}\left(\frac{\bar{N}_i(x)}{N_i(x)}-k\right)\left(\frac{\bar{N}_{i-1}(2x)}{N_{i-1}(2x)}-k\right)\left|\frac{\bar{N}_{i-1}(2x)}{N_{i-1}(2x)}-k\right|\right]dx=0,$$

thus, for $2\leq i\leq I-1$, we obtain $\operatorname{sgn}\left(\frac{\bar{N}_1(x)}{N_1(x)}-k\right)=\operatorname{sgn}\left(\frac{\bar{N}_i(2x)}{N_i(2x)}-k\right)$ and $\operatorname{sgn}\left(\frac{\bar{N}_i(x)}{N_i(x)}-k\right)=\operatorname{sgn}\left(\frac{\bar{N}_{i-1}(2x)}{N_{i-1}(2x)}-k\right)$, for all $k>0$. Therefore $\frac{\bar{N}_1(x)}{N_1(x)}=\frac{\bar{N}_i(2x)}{N_i(2x)}$ and $\frac{\bar{N}_i(x)}{N_i(x)}=\frac{\bar{N}_{i-1}(2x)}{N_{i-1}(2x)}$. Now combining these two relations, we obtain

$$\frac{\bar{N}_i(x)}{N_i(x)}=\frac{\bar{N}_i(2x)}{N_i(2x)}.$$

On the other hand, we deduce from this equality that,

$$\begin{aligned}\frac{\partial}{\partial x}\frac{\bar{N}_i(x)}{N_i(x)} &= 2B_{i-1}(2x)\frac{\bar{N}_{i-1}(2x)}{N_i(x)}\left(\frac{\bar{N}_{i-1}(2x)}{N_{i-1}(2x)}-\frac{\bar{N}_i(x)}{N_i(x)}\right),\\ &= 0.\end{aligned}$$

That is $\dfrac{\bar{N}_i(x)}{N_i(x)} = k$ for all $1 \le i \le I$. Because of both solutions are probability measures, then the uniqueness is proved. The same argument proves that $\phi_i = \bar{\phi}_i$.

## 3 Trend to a stable size distribution

In order to prove Theorem 1.2, we need to use a family of entropy inequalities to equation (1) which generalizes the usual conservation law (1). Firstly we start by the following theorem.

**Theorem 3.1** *Assume that $n_i^0 \in L^1(\mathbb{R}^+, \phi_i(x)dx)$ such that the initial data $|n_i^0(x)| \le CN_i(x)$ for $1 \le i \le I$, then there is a unique solution $n_i \in C\big(\mathbb{R}^+; L^1(\mathbb{R}^+, \phi_i(x)dx)\big)$ to (1) which satisfies*

$$|n_i(t,x)e^{-\lambda t}| \le CN_i(x) \quad \text{for all} \quad t > 0. \tag{28}$$

$$\sum_{i=1}^{I} \int_0^\infty \phi_i(x) n_i(t,x) e^{-\lambda t} dx = \sum_{i=1}^{I} \int_0^\infty \phi_i(x) n_i^0(x) dx, \tag{29}$$

$$\sum_{i=1}^{I} \int_0^\infty \phi_i(x) |n_i(t,x)| e^{-\lambda t} dx \le \sum_{i=1}^{I} \int_0^\infty \phi_i(x) |n_i^0(x)| dx, \tag{30}$$

In order to prove this theorem we need to state the generalized relative entropy introduced in [14, 15]. Indeed a straightforward computation leads to the following result,

**Lemma 3** *(General Relative Entropy) Assume that there exist eigenelements $\big((N_i), \lambda, (\phi_i)\big)_{1 \le i \le I}$ to (2)-(3), then for all convex and Lipschitz functions $H : \mathbb{R} \to \mathbb{R}$ with $H(0) = 0$, we have,*

$$\frac{d}{dt} \sum_{i=1}^{I} \int_0^\infty N_i(x)\phi_i(x) H\left(\frac{n_i(t,x)e^{-\lambda t}}{N_i(x)}\right) = -D_H(t) \le 0, \quad \forall t > 0,$$

*where the entropy dissipation $D_H(t) \ge 0$ is given by*

$$\begin{aligned} D_H(t) = &-2 \sum_{i=1}^{I-1} \int_0^\infty B_i(2x) N_i(2x) \phi_1(x) \\ &\cdot \left[ H'\left(\frac{n_1(t,x)}{N_1(x)}\right) \left(\frac{n_i(t,2x)}{N_i(2x)} - \frac{n_1(t,x)}{N_1(x)}\right) + H\left(\frac{n_1(t,x)}{N_1(x)}\right) - H\left(\frac{n_i(t,2x)}{N_i(2x)}\right) \right] \\ &- 2 \sum_{i=2}^{I-1} \int_0^\infty B_{i-1}(2x) N_{i-1}(2x) \phi_i(x) \left[ H'\left(\frac{n_i(t,x)}{N_i(x)}\right) \left(\frac{n_{i-1}(t,2x)}{N_{i-1}(2x)} - \frac{n_i(t,x)}{N_i(x)}\right) \right. \\ &\left. + H\left(\frac{n_i(t,x)}{N_i(x)}\right) - H\left(\frac{n_{i-1}(t,2x)}{N_{i-1}(2x)}\right) \right] \end{aligned}$$

We do not prove these two results which follow exactly the general theory developed in ([18], p 61 or [15]). We use the Generalized Relative Entropy method also to show that the solution of (1) converges to a steady state, see also ( [17, 14, 15, 16, 18]). We will need also some regularity results that we state now,

**Theorem 3.2** *With the same assumptions as in Theorem 3.1 and with the initial data satisfying*

$$|n_i^0(x)| \leq CN_i(x), \qquad \frac{\partial n_i^0(x)}{\partial x} \in L^1(\mathbb{R}^+, \phi_i(x)dx), \quad 1 \leq i \leq I, \tag{31}$$

*the solution to (1) satisfies, setting* $\tilde{n} = ne^{-\lambda t}$,

$$\sum_{i=1}^{I} \int_0^\infty \left| \frac{\partial \tilde{n}_i(t,x)}{\partial x} \right| \phi_i(x)dx + \sum_{i=1}^{I} \int_0^\infty \left| \frac{\partial \tilde{n}_i(t,x)}{\partial t} \right| \phi_i(x)dx \leq C, \quad \forall t \geq 0. \tag{32}$$

**Proof First step. Time derivative.** We follow the same ideas as in [18]. Indeed differentiating the equation (1) in time and setting $q_i(t,x) = \dfrac{\partial \tilde{n}_i}{\partial t}$ we show that $q_i(t,x)$ satisfies the same equation. From the contraction principle in Theorem 3.1, we conclude

$$\sum_{i=1}^{I} \int_0^\infty |q_i(t,x)|\phi_i(x)dx \leq \sum_{i=1}^{I} \int_0^\infty |q_i(0,x)|\phi_i(x)dx, \tag{33}$$

with

$$q_1(0,x) = -\frac{\partial n_1^0(x)}{\partial x} - B_1(x)n_1^0(x) + 2\sum_{i=1}^{I-1} B_i(2x)n_i^0(2x), \tag{34}$$

and for $2 \leq i \leq I$,

$$q_i(0,x) = -\frac{\partial n_i^0(x)}{\partial x} - B_i(x)n_i^0(x) + 2B_{i-1}(2x)n_{i-1}^0(2x). \tag{35}$$

Now we estimate each term. Since $|n_i^0(x)| \leq CN_i(x)$ for all $1 \leq i \leq I$, we have

$$|q_1(0,x)| \leq \left| \frac{\partial n_i^0(x)}{\partial x} \right| + CB_M|N_1(x)| + 2C\sum_{i=1}^{I-1} B_i(2x)N_i(2x), \tag{36}$$

replace $2\sum_{i=1}^{I-1} B_i(2x)N_i(2x)$ by the other terms of the equation on $N_1$, we arrive at

$$|q_1(0,x)| \leq \left| \frac{\partial n_1^0(x)}{\partial x} \right| + CB_M|N_1(x)| + C\Big( \left| \frac{\partial N_1(x)}{\partial x} \right| + B_M N_1(x) \Big). \tag{37}$$

After multiplying by $\phi_1$ and integrating we obtain

$$\int_0^\infty |q_1(0,x)|\phi_1(x)dx \leq \int_0^\infty \left(\left|\frac{\partial n_1^0(x)}{\partial x}\right| + C\left|\frac{\partial N_1(x)}{\partial x}\right|\right)\phi_1(x)dx + 2C(\lambda + B_M). \tag{38}$$

Similarly we have

$$\int_0^\infty |q_i(0,x)|\phi_i(x)dx \leq \int_0^\infty \left(\left|\frac{\partial n_i^0(x)}{\partial x}\right| + C\left|\frac{\partial N_i(x)}{\partial x}\right|\right)\phi_i(x)dx + 2C(\lambda + B_M), \tag{39}$$

for all $2 \leq i \leq I$. Thus summing these last equations for $1 \leq i \leq I-1$ we obtain

$$\sum_{i=1}^{I-1}\int_0^\infty |q_i(0,x)|\phi_i(x)dx \leq$$
$$\sum_{i=1}^{I-1}\int_0^\infty \left(\left|\frac{\partial n_i^0(x)}{\partial x}\right| + C\sum_{i=1}^{I-1}\left|\frac{\partial N_i(x)}{\partial x}\right|\right)\phi_i(x)dx + 2C(\lambda + B_M) \leq C,$$

since $\sum_{i=1}^{I}\int_0^\infty e^{\mu x}\left|\frac{\partial N_i(x)}{\partial x}\right|dx < \infty$ and $\phi_i(x) \leq C(1+x)$ for all $1 \leq i \leq I$ then $\sum_{i=1}^{I-1}\int_0^\infty \left|\frac{\partial N_i(x)}{\partial x}\right|\phi_i(x)dx < \infty$. Therefore the estimate on time derivatives is proved.

**Second step. Space derivative.** We have

$$\sum_{i=1}^{I}\frac{\partial \tilde{n}_i(t,x)}{\partial x} = -\frac{\partial \tilde{n}_i(t,x)}{\partial t} - \sum_{i=1}^{I}(B_i(x)+\lambda)\tilde{n}_i(t,x) + 4\sum_{i=1}^{I}B_i(2x)\tilde{n}_i(t,2x). \tag{40}$$

The control of $\frac{\partial \tilde{n}_i(t,x)}{\partial t}$ in the first step and $|\tilde{n}_i(t,x)| \leq C|N_i(x)|$, gives us a control similar to that on the time derivative, namely

$$\sum_{i=1}^{I}\int_0^\infty \left|\frac{\partial \tilde{n}_i(t,x)}{\partial x}\right|\phi_i(x)dx \leq C(n^0).$$

□

We are now ready to prove the Theorem 1.2.

**Proof of Theorem 1.2.** With the same ideas of [18], we can prove this theorem. Indeed we set $h_i(t,x) = \tilde{n}_i(t,x) - \rho N_i(x)$, we first notice that $h_i(t,x)$ being a solution to a cell division equation (1). The contraction principle shows that for some $l \geq 0$,

$$\sum_{i=1}^{I-1}\int_0^\infty |h_i(t,x)|\phi_i(x)dx \to l, \quad \text{as} \quad t \to \infty.$$

And it remains to show that $l = 0$. As we mentioned in Theorem 3.2, we have $|h_i| \leq CN_i$, for $1 \leq i \leq I-1$, $\sum_{i=1}^{I-1} \int_0^\infty \left|\frac{\partial h_i(t,x)}{\partial t}\right| \phi_i(x)dx \leq C(n^0)$, and $\sum_{i=1}^{I-1} \int_0^\infty \left|\frac{\partial h_i(t,x)}{\partial x}\right| \phi_i(x)dx \leq C(n^0)$. We then introduce the sequence of functions $h_i^n(t,.) = h_i(t+t_n,.)$, when $t_n \to \infty$ as $n \to \infty$. After extracting a subsequence still denoted $h_i^n$, we have $h_i^n \to g_i$ strongly in $L^1([0,T] \times \mathbb{R}^+; \phi_i(x)dx)$, $1 \leq i \leq I-1$. And that $g_i$ for $1 \leq i \leq I-1$ is solution to the cell division equations (1) and $g_i(t,x) \leq CN_i(x)$ for all $1 \leq i \leq I-1$. We can now work on the entropy dissipation of $h_i(t,x)$. From the Generalized Relative Entropy inequality, we have using the square entropy $H(u) = u^2$

$$\sum_{i=1}^{I-1} \int_0^T \int_0^\infty B_i(2x)N_i(2x)\phi_1(x) \left(\frac{h_i(t,2x)}{N_i(2x)} - \frac{h_1(t,x)}{N_1(x)}\right)^2 dxdt$$
$$+ \sum_{i=2}^{I-1} \int_0^T \int_0^\infty B_{i-1}(2x)N_{i-1}(2x)\phi_i(x) \left(\frac{h_{i-1}(t,2x)}{N_{i-1}(2x)} - \frac{h_i(t,x)}{N_i(x)}\right)^2 dxdt \leq C.$$

Therefore as $n \to \infty$

$$\sum_{i=1}^{I-1} \int_0^T \int_0^\infty B_i(2x)N_i(2x)\phi_1(x) \left(\frac{h_i^n(t,2x)}{N_i(2x)} - \frac{h_1^n(t,x)}{N_1(x)}\right)^2 dxdt \to 0,$$

and

$$\sum_{i=2}^{I-1} \int_0^T \int_0^\infty B_{i-1}(2x)N_{i-1}(2x)\phi_i(x) \left(\frac{h_{i-1}^n(t,2x)}{N_{i-1}(2x)} - \frac{h_i^n(t,x)}{N_i(x)}\right)^2 dxdt \to 0.$$

By the strong limit of $h_n$ we arrive at

$$\sum_{i=1}^{I-1} \int_0^T \int_0^\infty B_i(2x)N_i(2x)\phi_1(x) \left(\frac{g_i(t,2x)}{N_i(2x)} - \frac{g_1(t,x)}{N_1(x)}\right)^2 dxdt = 0,$$

and

$$\sum_{i=2}^{I-1} \int_0^T \int_0^\infty B_{i-1}(2x)N_{i-1}(2x)\phi_i(x) \left(\frac{g_{i-1}(t,2x)}{N_{i-1}(2x)} - \frac{g_i(t,x)}{N_i(x)}\right)^2 dxdt = 0.$$

In other words $\frac{g_i(t,2x)}{N_i(2x)} = \frac{g_1(t,x)}{N_1(x)}$, for $1 \leq i \leq I-1$, and $\frac{g_i(t,x)}{N_i(x)} = \frac{g_{i-1}(t,2x)}{N_{i-1}(2x)}$, for $2 \leq i \leq I-1$. On the other hand, by a simple calculus we have

$$\frac{\partial}{\partial t}\left(\frac{g_1}{N_1}\right) + \frac{\partial}{\partial x}\left(\frac{g_1}{N_1}\right) = 2\sum_{i=1}^{I-1} B_i(2x)\frac{N_i(2x)}{N_1(x)}\left(\frac{g_i(t,2x)}{N_i(2x)} - \frac{g_1(t,x)}{N_1(x)}\right) = 0,$$

and for $2 \leq i \leq I-1$,

$$\frac{\partial}{\partial t}\left(\frac{g_i}{N_i}\right) + \frac{\partial}{\partial x}\left(\frac{g_i}{N_i}\right) = 2B_{i-1}(2x)\frac{N_{i-1}(2x)}{N_i(x)}\left(\frac{g_{i-1}(t,2x)}{N_{i-1}(2x)} - \frac{g_i(t,x)}{N_i(x)}\right) = 0.$$

Thanks to Lemma 4-5 in [18] we have $\frac{g_i}{N_i}$ is constant for all $1 \leq i \leq I-1$ and the mass condition $\sum_{i=1}^{I-1} \int_0^\infty g_i(t,x)\phi_i(x)dx = 0$ allows us to conclude that $g_i = 0$, $1 \leq i \leq I-1$ and thus $l = 0$. It remains to show that $h_I(t,.) \to 0$ in $L^1_{loc}(\mathbb{R}^+)$ as $t \to \infty$. To do that we use the equation of $h_I$, namely

$$\begin{cases} \dfrac{\partial h_I}{\partial t}(t,x) + \dfrac{\partial h_I}{\partial x}(t,x) + B_I(x)h_I(t,x) = 2B_{I-1}(2x)h_{I-1}(t,2x), \\ h_I(t,0) = 0, \quad \text{and} \quad h_I(0,x) = h_I^0(x) \in L^1(\mathbb{R}^+). \end{cases} \tag{41}$$

After multiplying the equation (41) by $\operatorname{sgn}(h_I)$ and integrating over $(0,A)$ for all $A > 0$ we have

$$\frac{d}{dt}\int_0^A |h_I(t,x)|dx + b_m \int_0^A |h_I(t,x)|dx \leq 2B_M \int_0^A |h_{I-1}(t,2x)|dx,$$

by taking $q_I(t) := \int_0^A |h_I(t,x)|dx$ we obtain

$$\frac{d}{dt}(q_I(t)e^{b_m t}) \leq 2B_M e^{b_m t} \int_0^A |h_{I-1}(t,2x)|dx, \tag{42}$$

thus, after integrating (42) over $(0,t_n)$ with $t_n \to \infty$ as $n \to \infty$ we have

$$q_I(t_n) \leq e^{-b_m t_n} q_I(0) + 2B_M \int_0^{t_n} e^{-b_m(t_n - s)} \int_0^A |h_{I-1}(s,2x)|dxds,$$

and so $|q_I(t_n)| \leq C_1$ because $|h_{I-1}(s,x)| \leq CN_{I-1}(x)$. Again after integrating (42) over $(t_n, t+t_n)$ we have

$$q_I^n(t) := q_I(t+t_n) \leq e^{-b_m(t+t_n)} q_I(t_n) + 2B_M \int_{t_n}^{t+t_n} e^{-b_m(t+t_n-s)} \int_0^A |h_{I-1}(s,2x)|dxds,$$

by taking $F^n(s) = \int_0^A |h_{I-1}(s+t_n,2x)|dx := \int_0^A |h_{I-1}^n(s,2x)|dx$ and after a change of the variable we have

$$q_I^n(t) \leq e^{-b_m(t+t_n)} q_I(t_n) + 2B_M \int_0^t e^{-b_m(t-s)} F^n(s)ds,$$

since $F^n \to 0$ (because $\int_0^A |h_{I-1}^n(s,2x)|\phi_{I-1}(2x)dx \to 0$ as $n \to \infty$, and $\phi_{I-1} > 0$), then $|F^n| \leq C$ and $\int_0^t e^{-b_m(t-s)} F^n(s)ds < \infty$. Now we apply the Lebesgue theorem to conclude that $q_I^n \to 0$ as $n \to \infty$, and the claim is proved. □

## 4 Limit $I \to \infty$ with equal division rates

In this section we prove Theorems 1.3 and 1.4 for the case when division rates are independent of the number of earlier divisions as explained in the introduction. Throughout this section, we set $V^I(x) = \sum_{i=1}^{I} N_i^I(x)$ and $W^I(x) = \frac{1}{I-1}\sum_{i=1}^{I-1} \phi_i^I(x)$, where $N_i^I(x), \phi_i^I(x)$ are solutions to problem (2)-(3). We prove, using a priori estimates, that the limit of $\left(\lambda^I, V^I, W^I\right)$ as $I$ tends to $\infty$, called $(\theta, V, W)$, is solution to (8)-(9). From the result of [17], we know that problem (8)-(9) has a unique solution $(\bar{\lambda}, N, \phi)$. Hence we conclude that $(\theta, V, W) = (\bar{\lambda}, N, \phi)$.

The proof is divided in several steps : (i) To begin with, we recall the uniform estimates on $(N_i^I)_{1\le i\le I}$, (see Theorem 1.1) (ii) We refine the estimates on $(N_i^I)_{1\le i\le I}$ and $(\phi_i^I)_{1\le i\le I}$ independently of $I$. (iii) We conclude the proof passing to the limit. We begin with estimates,

**Lemma 4** *The solution $(N_i^I)_{1\le i\le I}$ satisfies*

$$N_i^I(x) \le N_j^I(x), \quad \forall i \ge j. \tag{43}$$

*Moreover there exists a constant $1 < \alpha < 2$ such that*

$$N_i^I(x) \ge \frac{N_1^I(x)}{2^{i-1}} \ge \frac{N_i^I(x)}{\alpha^{i-1}}, \quad \forall \ \ 1 \le i \le I, \tag{44}$$

$$\frac{1}{2^{I-1}} \sum_{i=1}^{I-1} 2^{i-1} N_i^I(x) \to 0 \ \ as \ \ I \to \infty. \tag{45}$$

**Proof** First of all we set $L(N_i^I) = \dfrac{\partial N_i^I}{\partial x} + (B(x) + \lambda)N_i^I$. Remarking that $L(N_1^I) \ge L(N_2^I)$, so by the maximum principle we obtain $N_1^I \ge N_2^I$. By induction, this proves the inequalities (43) by the comparison principle. Concerning the second result, notice that from system (2), $L(N_1^I) = L(\sum_{i=2}^{I} N_i^I)$, hence

$$N_1^I(x) = \sum_{i=2}^{I} N_i^I(x). \tag{46}$$

Moreover $L(2N_2^I) = 4B(2x)N_1^I(2x) \ge L(N_1^I)$, thus $2N_2^I \ge N_1^I$, also we have $L(2^2N_3^I) = 8B(2x)N_2^I(2x) \ge 4B(2x)N_1^I(2x) \ge L(N_1^I)$, which implies $2^2N_3^I \ge$

$N_1^I$, thus by induction we deduce the first inequality of (44). Furthermore by (46) we have $N_I^I \le \dfrac{N_1^I}{I-2}$, and

$$\begin{aligned} L(N_1^I) &= 2B(2x)N_1^I(2x) + 2B(2x)\sum_{i=2}^{I-1} N_i^I(x) \\ &= 4B(2x)N_1^I(2x) - 2B(2x)N_I^I(2x), \end{aligned}$$

henceforth $L(N_1^I) \ge B(2x)N_1^I(2x)\left(4 - \dfrac{2}{I-2}\right)$. Now choose a constant $1 < \alpha < 2$, such that $\alpha\left(4 - \dfrac{2}{I-2}\right) > 4$, for $I$ large enough, so

$$L(\alpha N_1^I) \ge B(2x)N_1^I(2x)\alpha\left(4 - \frac{2}{I-2}\right) > 4B(2x)N_1^I(2x) = L(2N_2^I),$$

thus $\alpha N_1^I \ge 2N_2^I$, then similarly we prove (45). Now we have

$$\frac{1}{2^{I-1}}\sum_{i=1}^{I-1} 2^{i-1}N_i^I(x) \le \frac{N_1^I(x)}{2^{I-1}}\sum_{i=1}^{I-1}\alpha^{i-1} \le \frac{N_1^I(x)}{\alpha-1}\frac{\alpha^{I-1}-1}{2^{I-1}},$$

since $1 < \alpha < 2$ we conclude (45). □

Now we turn to estimate the solution $(\phi_i^I)_{1\le i\le I-1}$ to the adjoint equation.

**Lemma 5** *The solution $(\phi_i^I)_{1\le i\le I-1}$ satisfies*

$$\phi_i^I(x) \le \phi_j^I(x), \quad \forall\ i \ge j. \tag{47}$$

$$\sum_{i=1}^{I-1}\phi_i^I(x) \ge (I-2)\phi_j^I(x), \quad 1 \le j \le I-1. \tag{48}$$

$$\frac{\phi_i^I}{2}(x) \le \phi_j^I(x), \quad \forall\ 1 \le i \le j \le I-1. \tag{49}$$

$$\phi_{I-j}^I(x) \ge \frac{2^j-1}{2^j}\phi_1^I(x), \quad \forall\ 1 \le j \le I-1. \tag{50}$$

**Proof** We set $S(\phi_i^I) := \dfrac{\partial\phi_i^I}{\partial x} - (B(x)+\lambda^I)\phi_i^I$, then

$$S(\phi_{I-1}^I) = -B(x)\phi_1^I\left(\frac{x}{2}\right) \ge S(\phi_{I-2}^I),$$

so by the maximum principle we obtain $\phi_{I-1}^I(x) \le \phi_{I-2}^I(x)$. By induction we prove the first result, namely (47). Secondly, on one hand we have

$$S\Big((I-2)\phi_{I-1}^I\Big) = -B(x)(I-2)\phi_1^I\left(\frac{x}{2}\right) \le -B(x)\sum_{i=1}^{I-1}\phi_i^I\left(\frac{x}{2}\right),$$

and on the other hand

$$S\left(\sum_{i=1}^{I-1}\phi_i^I(x)\right) = -B(x)(I-2)\phi_1^I\left(\frac{x}{2}\right) - B(x)\sum_{i=1}^{I-1}\phi_i^I\left(\frac{x}{2}\right) \leq S\Big((I-2)\phi_{I-1}^I\Big),$$

once again by the maximum principle we have

$$(I-2)\phi_{I-1}^I(x) \leq \sum_{i=1}^{I-1}\phi_i^I(x).$$

Henceforth by the structure of equation (3), we conclude (48). Concerning (49) we have

$$S(2\phi_{I-1}^I) = -2B(x)\phi_1^I\left(\frac{x}{2}\right) \leq S(\phi_i^I), \quad \forall \ 1 \leq i \leq I-1,$$

thus $2\phi_{I-1}^I(x) \geq \phi_i^I(x)$, for $1 \leq i \leq I-1$. Furthermore

$$S(2\phi_{I-2}^I) = -2B(x)\Big(\phi_1^I\left(\frac{x}{2}\right) + \phi_{I-1}^I\left(\frac{x}{2}\right)\Big) \leq S(\phi_i^I), \quad \forall \ 1 \leq i \leq I-1,$$

then by induction we obtain (49). We finish this proof by establishing (50). Indeed

$$S(\phi_{I-1}^I) = -B(x)\phi_1^I\left(\frac{x}{2}\right),$$

$$\begin{aligned} S(\phi_{I-2}^I) &= -B(x)\left(\phi_1^I\left(\frac{x}{2}\right) + \phi_{I-1}^I\left(\frac{x}{2}\right)\right) \\ &\leq -B(x)\left(\phi_1^I\left(\frac{x}{2}\right) + \frac{1}{2}\phi_1^I\left(\frac{x}{2}\right)\right) \\ &= -\frac{3}{2}B(x)\phi_1^I\left(\frac{x}{2}\right). \end{aligned}$$

In addition $S\left(\frac{3}{4}\phi_1^I\right) \geq -\frac{3}{2}B(x)\phi_1^I\left(\frac{x}{2}\right) \geq S(\phi_{I-2}^I)$, so, applying the maximum principle we find $\phi_{I-2}^I(x) \geq \dfrac{2^2-1}{2^2}$, therefore, once more by induction we conclude (50). □

Now we state the uniform estimate of $(\phi_i)_{1\leq i\leq I-1}$.

**Lemma 6** *The solution $(\phi_i^I)_{1\leq i\leq I-1}$ satisfies, with a constant $C$ is independent of $I$,*

$$\frac{1}{I-1}\sum_{i=1}^{I-1}\phi_i^I(x) \leq C(1+x). \tag{51}$$

**Proof** We follow the proof of Lemma 2. First step. Using a solution $(\phi_i^L)_{1\le i\le I}$ of (13) on a bounded interval $(0, L)$. Summing and integrating the equation (13) over $(y, x_L)$ we obtain, setting $W_L^I(x) = \dfrac{1}{I-1}\displaystyle\sum_{i=1}^{I-1}\phi_i^L(x)$,

$$\begin{aligned} W_L^I(y) &= W_L^I(x_L) + 2\int_y^{x_L} B(x)W_L^I\left(\frac{x}{2}\right)dx - \int_y^{x_L}(B(x)+\lambda_L^I)W_L^I(x)dx \\ &- \int_y^{x_L} B(x)\left(W_L^I(x) - (I-2)\phi_1\left(\frac{x}{2}\right)\right)dx \\ &\le W_L^I(x_L) + 2\int_y^{x_L} B(x)W_L^I\left(\frac{x}{2}\right)dx, \end{aligned}$$

now arguing as in the proof of Lemma 2, we deduce that $\sup\limits_{0\le y\le A} W_L^I(y) \le W_L^I(x_L)$. Since $\displaystyle\sum_{i=1}^{I-1}\phi_i^I(x_L)N_i^I(x_L) \le \frac{2}{x_L}$, then $\phi_{I-1}^I(x_L)\displaystyle\sum_{i=1}^{I-1}N_i^I(x_L) \le \frac{2}{x_L}$, and thus $\phi_{I-1}^I(x_L)\big(V_L^I(x_L) - N_I^I(x_L)\big) \le \dfrac{2}{x_L}$, therefore by (49) we have $\phi_1^I(x_L)\big(V_L^I(x_L) - N_I^I(x_L)\big) \le \dfrac{4}{x_L}$, but $\phi_1^I(x) \ge W_L^I(x)$, then $W_L^I(x_L) \le \dfrac{4}{c\big(V_L^I(x_L) - N_I^I(x_L)\big)} \le C_1$, where $C_1$ is independent of $I$, and $c \le x_L \le C$. For $x$ large, we have

$$S(W_L^I) = -B(x)W_L^I\left(\frac{x}{2}\right) - \frac{I-2}{I-1}B(x)\phi_1^I\left(\frac{x}{2}\right),$$

thus, using (48) we find $S(W_L^I) \ge -2B(x)W_L^I\left(\frac{x}{2}\right)$. Consequently still as Lemma 2 or [17] we obtain the result (51). □

Now we are able to prove Theorems 1.3.

**Proof of Theorem 1.3** Summing equations (2) we have

$$\frac{\partial V^I(x)}{\partial x} + (B(x)+\lambda^I)V^I(x) = 4B(2x)V^I(2x) - 4B(2x)N_I^I(2x),$$

the function $\phi V^I$ satisfies

$$\frac{\partial(\phi V^I)}{\partial x} \le 4B(2x)\phi(x)V^I(2x) - 2B(2x)\phi(\frac{x}{2})V^I(x) + (\bar\lambda - \lambda^I)\phi(x)V^I(x),$$

after integration over $(0,\infty)$ we deduce

$$(\bar\lambda - \lambda^I)\int_0^\infty \phi(x)V^I(x)dx \ge 0,$$

then $\lambda^I \le \bar\lambda$. From Theorem 1.1, it follows that, $\{V^I\}$ converges weakly to $V$ in $W^{1,1}(\mathbb{R}^+)$, $\lambda^I \to \theta$ a.e and that $\int_0^\infty V(x)dx = 1$, so $V \ne 0$. We

claim that $(\theta, V)$ satisfy the equation (8). To prove the claim we have just to show that $N_I^I \to 0$ strongly in $L^1(\mathbb{R}^+)$. Notice that for fixed $I$ we have $V^I(x) = \sum_{i=1}^{I} N_i^I(x) \geq I N_I^I(x)$, where the last affirmation follows easily by using (43). Since $\{V^I\}$ is bounded in $W^{1,1}(\mathbb{R}^+)$, it follows that $N_I^I \to 0$ strongly in $L^1(\mathbb{R}^+)$ and the claim follows. Therefore the proof is achieved. The uniqueness of the solution to (8) implies that $\bar{\lambda} = \theta$ and $N(x) = V(x)$. We turn out to $W^I$. From Lemma 6 we can see that $W^I$ converges strongly (local convergence) to a function $W$. Since $W^I$ satisfies the equation

$$S(W^I) = -B(x)W^I\left(\frac{x}{2}\right) - \frac{I-2}{I-1}B(x)\phi_1^I\left(\frac{x}{2}\right),$$

and by (47)-(48),

$$\frac{I-2}{I-1}W^I(x) \leq \frac{I-2}{I-1}\phi_1^I(x) \quad \leq \quad W^I(x),$$

then the limit function $W$ satisfies

$$S(W) = -2B(x)W\left(\frac{x}{2}\right).$$

In addition, on the one hand we have

$$1 = \int_0^\infty \sum_{i=1}^{I-1} \phi_i^I(x) N_i^I(x)dx \leq \int_0^\infty \phi_1^I(x) \sum_{i=1}^{I-1} N_i^I(x)dx \leq \frac{I-1}{I-2}\int_0^\infty W^I(x)V^I(x)dx,$$

thus

$$\int_0^\infty W(x)V(x)dx \geq 1.$$

On the other hand, by using (50), we have

$$\begin{aligned}\sum_{i=1}^{I-1} N_i^I(x)\phi_i^I(x) &\geq \phi_1^I(x)\sum_{i=1}^{I-1} N_i^I(x)\frac{2^{I-i}-1}{2^{I-i}},\\ &\geq W^I(x)\Big(V^I(x) - N_I^I(x)\Big) - \frac{1}{2^{I-1}}\sum_{i=1}^{I-1} 2^{i-1}N_i^I(x),\end{aligned}$$

then by (45), and passing to the limit we obtain

$$\int_0^\infty W(x)V(x)dx \leq 1,$$

henceforth, the uniqueness of the solution to (9) implies that $W = \phi$. The theorem is proved. □

We turn out to the solution of the hyperbolic problem, namely the problem (1).

**Proof of Theorem 1.4** First of all $v^I(t,x) := \sum_{i=1}^{I} n_i^I(t,x)$ being a solution to

$$\frac{\partial v^I}{\partial t} + \frac{\partial v^I}{\partial x} + B(x)v^I = 4B(2x)v^I(t,2x) - 4B(2x)n_I^I(t,2x),$$

with $(n_i^I)_{1\le i\le I}$ is solution to (1). Because of global boundary variation regularity on $v^I$ proved in Theorem 3.2, we have $v^I \to m$ strongly in $L^1\big(0,T;L^1_{loc}(\mathbb{R}^+)\big)$. Also we can see that $(n_i^I)_{1\le i\le I}$ is a decreasing sequence in $i$ provided that the initial data $(n_i^0)_{1\le i\le I}$ is decreasing in $i$. Hence $n_I^I(x) \le \frac{1}{I}v^I(x)$. Therefore the limit $m$ is solution to (1). According to the uniqueness of this solution, the result is proved. □

## 5 Perspectives

We have considered a model of cell division for the continuous growth of cells and their division in two new-born cells of equal sizes, one is the mother cell and passes to the next generation, the other, the daughter cell begins at generation 0. We have included the standard phenomena that only a finite number of divisions is possible.

We have shown that recently developed methods based on generalized entropy can be used to give a natural background for existence of a stable steady dynamics (first positive eigenvector), for the long time asymptotics of solutions. Our main new result concerns the limit of a high number of possible divisions. In the particular case where the division rates are independent of the number of earlier divisions, we have shown that we recover the classical cell division model for equal mitosis. This involves new estimates and convergence in some average sense.

Several questions are still to study in this problem. One can mention for instance general, non-necessarily symmetric divisions [13, 18] and more general growth rates; these are important for many applications as parasites or plasmids contents [21]. In the direction of improving the very new point of this paper, one can wonder how to find the limit of high generation numbers when cell division rates depend upon the generation since it is natural to suppose that successive inner degradations decrease the birth rate.

## References

[1] L. Arlotti, J. Banasiak *Strictly substochastic semigroups with application to conservative and shattering solutions to fragmentation equations with mass loss.* J. Math. Anal. Appl. 293, (2004), no. 2, 693-720.

[2] B. Basse, B. C. Baguley, E. S. Marshall, W. R. Joseph, B. van Brunt, G. Wake, D. J. N. Wall, *A mathematical model for analysis of the cell cycle in cell lines derived from human tumors.* J. Math. Biol. Vol. 47 (4) (2003), 295-312.

[3] N. Bellomo, M. Maini, Preface : *The cellular scale.* Math Mod. Meth. Appl. Sci. 15, 2005.

[4] O. Diekmann, H. J. A. M. Heijmans, H. R. Thieme, *On the stability of the cell size distribution.* J. Math. Biol. Vol. 19, 2, (1984), 227-248.

[5] J. F. Collet, T. Goudon, S. Hariz, F. Poupaud, A. Vasseur, *Some recent results on the kinetic theory of phase transitions.* 103–120, IMA Vol. Math. Appl., 135, Springer, New York, 2004.

[6] O. Diekmann, H. J. A. M. Heijmans, H. R. Thieme, *On the stability of the cell size distribution. II. Time periodic developmental rates.* Hyperbolic partial differential equations, III. Comput. Math. Appl. Part A. Vol. 12, 4-5, (1986), 491-512.

[7] M. Escobedo, P. Laurencot, S. Mishler, B. Perthame, *Gelation and Mass Conservation in Coagulation-Fragmentation Models.* J. Differential Equations. 195, (2003), no. 1, 143-174.

[8] M. Escobedo, S. Mischler, M. Rodriguez Ricard, *On self-similarity and stationary problem for fragmentation and coagulation models.* Ann. Inst. H. Poincaré Anal. Non Linéaire 22 (2005), no. 1, 99-125.

[9] M. L. Greer, P. van den Driessche, Lin Wang, G. F. Webb, *Effects of General Incidence and Polymer Joining on Nucleated Polymerization in a Model of Prion Proliferation.* SIAM J. Appl. Math. 68, 154 (2007) pp. 154-170.

[10] P. Laurençot, C. Walker, *Steady states for a coagulation-fragmentation equation with volume scattering.* SIAM J. Math. Anal. 37 (2005), no. 2, 531-548 (electronic).

[11] G. Menon, R. L. Pego, *Dynamical scaling in Smoluchowski's coagulation equations: uniform convergence.* SIAM Rev. 48, (2006), no. 4, 745-768

[12] J. A. J. Metz, O. Diekmann, *The dynamics of phisiologically structured populations.* LN in biomathematics 68, Springel-Verlag (1986)

[13] P. Michel, *Existence of a solution to the cell division eigenproblem.* Math. Mod. Meth. Appl. Sci. Vol. 16, suppl. issue 1, (2006), 1125-1153.

[14] P. Michel, S. Mishler, B. Perthame, *General entropy equations for structured population models and scattering,* C. R. Math. Acad. Sci. Paris, Vol. 338, (2004), 9, 697-702.

[15] P. Michel, S. Mishler, B. Perthame, *General relative entropy inequality : an illustration on growth models,* J. Math. Pures. Appl. Vol. 84, (2005), 9, 1235-1260.

[16] S. Mishler, B. Perthame, L. Ryzhik, *Stability in a nonlinear population maturation model,* Math. Model, Methods, Appl. Sci., Vol. 12, 1751-1772.

[17] B. Perthame, L. Ryzhik, *Exponential decay for the fragmentation or cell division equation.* J. Diff. Eq. Vol. 210, (2005), 155-177.

[18] B. Perthame, *Transport Equations in Biology*, Birkhauster, Berlin, 2007.

[19] L. Preziosi, *Modeling Cancer Growth.* CRC-Press-Chapman-Hall, Boca Raton, 2003.

[20] M. Rotenberg, *Transport theory for growing cell populations.* J. Theor. Biol. Vol. 103, (1983), 181-199.

[21] C. Shene, B. Andrews, J. A. Asenjo, *Study of recombinant micro-organism populations characterized by their plasmid content per cell using a segregated model.* Bioprocess Biosyst. Eng. 25 (2003) 333-340.

[22] E. J. Stewart, R. Madden, G. Paul, F. Taddei, *Aging and death in an organism that reproduces by morphologically symmetric division* Plos Biology, Vol. 3, issue 2, (2005), 295-300.

*Bol. Soc. Esp. Mat. Apl.*
*nº 44(2008), 81–99*

# SOBRE UNA INTERPOLACIÓN NO LINEAL: APLICACIÓN AL PROCESADO DE SEÑALES

SERGIO AMAT PLATA

Departamento de Matemática Aplicada y Estadística.
Universidad Politécnica de Cartagena (Spain)

sergio.amat@upct.es

*Este artículo está dedicado a mi madre: Carmen Plata Jiménez*

**Resumen**

En el presente trabajo se presenta un esquema interpolatorio no lineal. Se trata de una reconstrucción a trozos, de cuarto orden en regiones de suavidad, centrada (con soporte óptimo) y sin presencia del fenómeno de Gibbs cerca de las discontinuidades. Se estudia la estabilidad del esquema de multirresolución no lineal asociado. Dado el carácter no lineal no se puede hacer uso de las técnicas habituales de la teoría de los wavelets. Finalmente, se presentarán varias aplicaciones dentro del campo del procesado de señales.

**Palabras clave:** *multirresolución, esquemas de subdivisión no lineales, interpolación, media armónica, procesado de señales.*
**Clasificación por materias AMS:** *41A05 41A10 65D05 65D17*

## 1 Introducción

Un esquema de multirresolucón conecta de forma biyectiva una sucesión discreta $f^L$, donde $L$ representa el nivel de resolución más fino, con una sucesión de la forma

$$\{f^0, d^1, \ldots, d^L\},$$

donde $f^0$ representa una versión de la señal inicial en el nivel de resolución más grosero y cada una de las sucesiones $d^k$ representan los detalles necesarios para recuperar $f^k$ a partir de $f^{k-1}$. En el caso de algoritmos lineales se trata simplemente de un cambio de base.

El objetivo es tener la misma información pero escrita de otra forma que permita distinguir entre las partes más y menos importantes de la señal. Más concretamente, por construcción, los detalles serán pequeños en regiones donde

Fecha de recepción: 10/03/2008. Aceptado (en forma revisada): 17/05/2008.

se aproximen bien las sucesiones $f^k$ a partir del nivel inferior $f^{k-1}$. Esta propiedad es crucial en las aplicaciones.

En los últimos años, varias propuestas para mejorar las multirresoluciones lineales clásicas de tipo wavelet han dado lugar a multirresoluciones no lineales. El carácter no lineal es usado para intentar obtener mejores aproximaciones que permitan tener el mayor número de detalles pequeños.

En estos contextos, pocos resultados de convergencia y estabilidad han sido probados [16], [21], [17], [23], [38], [40].

En [9], en el contexto de compresión de imágenes, se presenta una nueva multirresolución no lineal. Usando un esquema tipo producto tensorial, esta multirresolución esta basada en una reconstrucción no lineal llamada PPH (Piecewise Polynomial Harmonic). Se analizó en términos de convergencia y estabilidad del esquema de subdivisión asociado, siguiendo las ideas presentadas en [21]. Los experimentos numéricos para la compresión de imágenes demostraron que puede ser considerado como una buena alternativa a los algoritmos lineales.

La estabilidad de los esquemas de multirresolución es una propiedad imprescindible, ya que, en las aplicaciones modificaremos la multirresolución antes de recuperar la sucesión original. Un algoritmo no estable no permitiría tener un control del error que se va a cometer. Por otro lado, destacar que en caso no lineal la estabilidad no es consecuencia ni de la convergencia ni de la estabilidad del esquema de subdivisión asociado.

En [11], se establece la estabilidad de la multirresolución PPH. Otros estudios similares pueden consultarse en [28], [40] y [23].

Este tipo de reconstrucciones no lineales tienen diversas aplicaciones como: reconstrucciones para métodos capturadores de choques en la aproximación de leyes de conservación [3], [4], [25], [37], procesamiento de señales [1], [5], [14], [12], compresión de imágenes y videos [9], [6], [7], [13], [18] eliminación de ruido en imágenes [8], [39], esquemas de subdivision [9], [20], [29], [30], [31].

El resto del artículo está organizado como sigue: En la sección 2 presentamos la multirresolución PPH como una perturbación de la multirresolución lineal interpolatoria. En la sección 3 se establece una propiedad de contracción entre las diferencias de segundo orden de los distintos niveles de resolución, lo que permite deducir resultados tanto de convergencia para el esquema de subdivision como de estabilidad para la multirresolución PPH. Finalmente, en la sección 4 se presenta y analiza varias aplicaciones de la reconstrucción PPH para el procesado de señales.

## 2 Multirresolución interpolatoria

Se considera un conjunto encajado de mallados en $\mathbb{R}$:

$$X^k = \{x_j^k\}_{j\in\mathbb{Z}}, \quad x_j^k = jh_k, \quad h_k = 2^{-k},$$

y la discretización puntual

$$\mathcal{D}_k : \begin{cases} C_B(\mathbb{R}) & \to V^k \\ f & \mapsto f^k = (f_j^k)_{j\in\mathbb{Z}} = (f(x_j^k))_{j\in\mathbb{Z}}, \end{cases} \tag{1}$$

donde $V^k$ es el espacio de sucesiones reales para la resolución dada por $X^k$ y $C_B(\mathbb{R})$ el conjunto de funciones continuas y acotadas sobre $\mathbb{R}$. Un operador reconstrucción $\mathcal{R}_k$ asociado a esta discretización es cualquier inversa por la derecha de $\mathcal{D}_k$, lo que significa que para toda $f^k \in V^k$, $\mathcal{R}_k f^k \in C_B(\mathbb{R})$ y

$$(\mathcal{R}_k f^k)(x_j^k) = f_j^k = f(x_j^k). \tag{2}$$

Las sucesiones de operadores $\{\mathcal{D}_k\}$ y $\{\mathcal{R}_k\}$ definen un esquema de multirresolución. El operador de predicción, es decir, $\mathcal{D}_{k+1}\mathcal{R}_k : V^k \to V^{k+1}$, define un esquema de subdivisión. La relación (2) implica que el esquema de subdivisión es interpolatorio. Si $\mathcal{R}_k$ es una reconstrucción no lineal, el esquema de subdivisión correspondiente es también no lineal, (ver [15] para más detalles).

A partir de estas reconstrucciones, usando funciones primitivas, se pueden obtener reconstrucciones y algoritmos de multirresolución para otra tipo de discretizaciones [15]. Por ejemplo, para medias en celda

$$\mathcal{D}_k : \begin{cases} L^1(\mathbb{R}) & \to V^k \\ f & \mapsto f^k = (f_j^k)_{j\in\mathbb{Z}} = (\int_{x_{j-1}^k}^{x_j^k} f(x)dx)_{j\in\mathbb{Z}}. \end{cases} \tag{3}$$

### 2.1 Reconstrucciones lineales: Interpolación de Lagrange

Las técnicas interpolatorias de Lagrange son independientes de los datos y se usan para definir operadores de reconstrucción interpolatorios *lineales* $(\mathcal{R}_k^{\mathcal{L}} f^k)(x)$ que son polinomios a trozos definidos en cada subintervalo $[x_j^k, x_{j+1}^k]$ como la única interpolación polinómica asociada al conjunto de datos

$$\{f_{j+m}^k, m \in \mathcal{S}\}$$

con $\mathcal{S} = \mathcal{S}(r,s) = \{-s, -s+1, \ldots, -s+r\}$.

Las interpolaciones lineales de Lagrange pierden su orden de aproximación con la presencia de singularidades ([15]). De hecho, si $f$ tiene una discontinuidad de salto en $[x_{j-1}^k, x_j^k]$, es fácil ver que cualquier diferencia dividida[1] basada en un conjunto de $s+1$ puntos conteniendo a $\{x_{j-1}^k, x_j^k\}$ verifica que

$$f[x_l^k, \ldots, x_{l+s}^k] = O([f])/h_k^s,$$

donde $[f] = |f_j^k - f_{j-1}^k|$. Por lo tanto, cada vez que se cruza la singularidad el error es de la forma

$$f(x) = (\mathcal{R}_k^{\mathcal{L}} f^k)(x) + O([f]),$$

lo que significa que el orden de la predicción es cero.

[1] Las diferencias divididas se definen como $f[x_{i-1}^k, x_i^k] := \frac{f(x_i^k) - f(x_{i-1}^k)}{x_i^k - x_{i-1}^k}$ y $f[x_{i-m}^k, \ldots, x_{i-m+n}^k] := \frac{f[x_{i-m+1}^k, \ldots, x_{i-m+n}] - f[x_{i-m}^k, \ldots, x_{i-m+n-1}^k]}{x_{i-m+n}^k - x_{i-m}^k}$

### 2.2 Reconstrucciones no lineales: Interpolación PPH

En esta sección introduciremos una reconstrucción no lineal de cuarto orden en regiones de suavidad. Se basa en una interpolación polinómica a trozos siguiendo las ideas introducidas en [3] para la reconstrucción de flujos en leyes de conservación. Se le llamará PPH (Piecewise Polynomial Harmonic). Para más detalles ver [9].

Esta técnica no lineal de interpolación nos permite construir un operador de reconstrucción con ciertas propiedades interesantes. En primer lugar, cada trozo de polinomio es construido con un conjunto de cuatro datos centrados $\{x^k_{j-1}, x^k_j, x^k_{j+1}, x^k_{j+2}\}$. En segundo lugar, tiene el mismo orden que la interpolación lineal en regiones de suavidad. El orden se reduce cerca de las singularidades pero se mantiene suficientemente para no generar el fenómeno de Gibbs.

Cuando una función tiene una discontinuidad de salto en un punto, las aproximaciones lineales tienen un comportamiento especial alrededor de dicho punto. Este comportamiento se llama fenómeno de Gibbs. Este fenómeno consiste en que cerca del punto se generan unas oscilaciones que no se hacen pequeñas al aumentar el orden de la reconstrucción. Este fenómeno fue observado por el físico A. Michelson, quien en 1898 construyó una máquina para sumar series de Fourier. Alrededor de las discontinuidades de las funciones siempre aparecían saltos, que no se hacían pequeños por mucho que se aumentara el número de sumandos de la serie. El fenómeno fue explicado en 1899 por J.W. Gibbs.

Considerando las diferencias divididas

$$e^k_{j-\frac{1}{2}} = f[x^k_{j-1}, x^k_j], \qquad e^k_{j+\frac{1}{2}} = f[x^k_j, x^k_{j+1}], \qquad e^k_{j+\frac{3}{2}} = f[x^k_{j+1}, x^k_{j+2}],$$

$$D^k_j = f[x^k_{j-1}, x^k_j, x^k_{j+1}], \qquad D^k_{j+1} = f[x^k_j, x^k_{j+1}, x^k_{j+2}],$$

la interpolación PPH asociada a conjunto centrado $f^k_{j-1}, f^k_j, f^k_{j+1}, f^k_{j+2}$ tiene la siguiente forma

$$\tilde{P}_j(x^k_{j+\frac{1}{2}}) = \frac{f^k_j + f^k_{j+1}}{2} - \frac{1}{4}\tilde{D}^k h^2. \tag{4}$$

donde

$$\tilde{D}^k = \begin{cases} \frac{2D^k_j D^k_{j+1}}{D^k_j + D^k_{j+1}} & \text{si } D^k_j D^k_{j+1} > 0, \\ 0 & \text{en otro caso.} \end{cases} \tag{5}$$

Es interesante compararla con la interpolación de Lagrange $P_j(x)$

$$P_j(x^k_{j+\frac{1}{2}}) = \frac{f^k_j + f^k_{j+1}}{2} - \frac{1}{4}\frac{D^k_j + D^k_{j+1}}{2}h^2. \tag{6}$$

Aplicando que

$$|2\frac{D^k_j D^k_{j+1}}{D^k_j + D^k_{j+1}}| \leq 2\,\text{mín}(|D^k_j|, |D^k_{j+1}|) = O(1), \tag{7}$$

se obtiene que $\tilde{D}^k = O(1)$, en lugar de $O(\frac{1}{h})$, como en el caso lineal cuando una discontinuidad existe en $[x^k_{j-1}, x^k_j]$ o en $[x^k_{j+1}, x^k_{j+2}]$.

Cuando los valores son próximos las dos medias son parecidas, pero la media armónica siempre es dos veces más pequeña que el mínimo de los valores. Esta propiedad es la clave para la no aparición del fenómeno de Gibbs en la reconstrucción PPH.

Otras dos formulaciones de la reconstrucción PPH pueden encontrase en [9].

Las principales propiedades que verifica la reconstrucción PPH son las siguientes:

1) Por construcción los datos utilizados en la reconstrucción son siempre centrados.

2) Si $f$ es un polinomio de grado menor o igual a 2,

$$D^k_j = D^k_{j+1} = \frac{D^k_j + D^k_{j+1}}{2} = \tilde{D}^k,$$

y por lo tanto la reconstrucción PPH reproduce polinomios de grado 2.

3) Si $f \in C^4$ y $D^k_j D^k_{j+1} > 0$, usando desarrollos de Taylor se obtiene que

$$\begin{aligned} 2\frac{D^k_j D^k_{j+1}}{D^k_j + D^k_{j+1}} &= \frac{f''(x^k_{j+\frac{1}{2}})}{2} + O(h^2), \\ \frac{D^k_j + D^k_{j+1}}{2} &= \frac{f''(x^k_{j+\frac{1}{2}})}{2} + O(h^2). \end{aligned}$$

Por lo tanto, en regiones de suavidad, la diferencia entre las dos medias es de orden $O(h^2)$, y la reconstrucción alcanza su orden óptimo 4.

4) Cuando $D^k_j D^k_{j+1} \leq 0$, $\tilde{P}(x^k_{j+1/2}) = \frac{f^k_{j+1} + f^k_j}{2}$. En este caso el orden se reduce a dos incluso en regiones de suavidad. Mediante traslaciones se puede evitar esta reducción de orden.

5) Si existe una discontinuidad en $[x^k_{j+1}, x^k_{j+2}]$ y $D^k_j D^k_{j+1} > 0$, por la propiedad (7) el fenómeno de Gibbs típico en las reconstrucciones lineales no aparece. Además, el orden de aproximación permanece $O(h^2)$ en este caso.

Para más detalles ver [9], [44].

## 3 Estabilidad de la multirresolución PPH

Introduciendo las diferencias $D^k f_j = f^k_{j+1} - 2f^k_j + f^k_{j-1}$, la reconstrucción PPH (es otra forma de verla, como decíamos en la sección 2) $\tilde{P}_j$ cuando $|D^k f_j| \leq |D^k f_{j+1}|$, es el polinomio de grado 3 definido mediante

$$\begin{cases} \tilde{P}_j(x^k_l) & = f^k_l, \quad \text{para } j-1 \leq l \leq j+1, \\ \tilde{P}_j(x^k_{j+2}) & = \tilde{f}^k_{j+2}, \end{cases} \tag{8}$$

con

$$\tilde{f}^k_{j+2} = f^k_{j+1} + f^k_j - f^k_{j-1} + 2H(D^k f_j, D^k f_{j+1}),$$

donde $H$ está definida como:

$$(x,y) \in I\!R^2 \mapsto H(x,y) := \frac{xy}{x+y}(\text{sign}(xy)+1),$$

donde $\text{sign}(x) = 1$ si $x \geq 0$ y $\text{sign}(x) = -1$ si $x < 0$.

Para la prueba de la estabilidad se necesitan unos lema técnicos de la función $H$ que acontinuación enunciaremos (ver [11] para más detalles).

**Lema 1** *Dados $(x,y),(x',y') \in I\!R^2$, la función $H$ satisface las siguientes propiedades:*

*1)* $H(x,y) = H(y,x)$

*2)* $H(x,y) = 0$ si $xy \leq 0$

*3)* $H(-x,-y) = -H(x,y)$

*4)* $|H(x,y)| \leq \text{máx}(|x|,|y|)$

*5)* $|H(x,y)| \leq 2\,\text{mín}(|x|,|y|)$

*6)* $|H(x,y) - H(x',y')| \leq 2\,\text{máx}\{|x-x'|,|y-y'|\}$.

**Lema 2** *La función $Z$ definida sobre $I\!R^3$ mediante $Z(x,y,z) = \frac{x}{2} - \frac{1}{8}(H(x,y) + H(x,z))$ satisface las siguientes propiedades:*

*1)* $|Z(x,y,z)| \leq \frac{|x|}{2}$

*2)* $sign(Z(x,y,z)) = sign(x)$.

*3)* $|Z(x,y,z) - Z(x^{'},y^{'},z^{'})| \leq \frac{1}{2}|x-x^{'}| + \frac{1}{2}\,\text{máx}\{|y-y^{'}|,|z-z^{'}|\}$

Ahora nos centramos en el esquema de subdivisión no lineal $S_{PPH}$ asociado a la predicción PPH, que viene dado por

$$f^{k-1} \mapsto S_{PPH}(f^{k-1}) = \mathcal{D}_k \mathcal{R}_{k-1} f^{k-1},$$

con

$$\begin{cases} (\mathcal{D}_k \mathcal{R}_{k-1} f^{k-1})_{2j+1} = \tilde{P}_j(x^k_{j+\frac{1}{2}}), \\ (\mathcal{D}_k \mathcal{R}_{k-1} f^{k-1})_{2j} = f^{k-1}_j. \end{cases} \tag{9}$$

Se tiene la siguiente contracción en dos pasos:

**Proposición 1** *Si (omitiendo $k$ por sencillez) $\hat{f} = S_{PPH}(f)$, $\hat{g} = S_{PPH}(g)$, $\bar{f} = S_{PPH}(\hat{f})$ y $\bar{g} = S_{PPH}(\hat{g})$ entonces*

$$\begin{array}{ll} 1) & ||D\hat{f}||_{l_\infty(\mathbb{Z})} \leq \frac{1}{2}||Df||_{l_\infty(\mathbb{Z})}, \\ 2) & |D(\hat{f}_j - \hat{g}_j)| \leq \frac{1}{2}||D(f-g)||_{l_\infty(\mathbb{Z})}, \quad \textit{para } j = 2n+1, \\ & |D(\hat{f}_j - \hat{g}_j)| \leq ||D(f-g)||_{l_\infty(\mathbb{Z})}, \quad \textit{para } j = 2n, \\ y & \\ 3) & ||D(\bar{f} - \bar{g})||_{l_\infty(\mathbb{Z})} \leq \frac{3}{4}||D(f-g)||_{l_\infty(\mathbb{Z})}. \end{array} \tag{10}$$

**Nota 1** *El ejemplo siguiente,* $(D^k f_n, D^k f_{n+1}, D^k f_{n-1}) = (M+1, 0, 0)$ *y* $(D^k g_n, D^k g_{n+1}, D^k g_{n-1}) = (M, 1, 1)$ *con* $M \to +\infty$ *muestra que una contracción en un sólo paso en el sentido de la propiedad 2)* ($j = 2n+1$) *de la proposición 1 no es cierta en general para* $j = 2n$.

Con estas condiciones podemos demostrar directamente la convergencia del esquema de subdivisión [2] $S_{PPH}$, aplicando el Teorema 3.3 de [23]. Para nuestro esquema se obtiene regularidad Hölder igual a uno.

Ahora nos centramos en la multirresolución PPH.

Primero damos el siguiente resultado donde ya aparecen los detalles $d(f), d(g)$ y $d(\dot{f}), d(\dot{g})$ del esquema de multirresolución:

**Proposición 2** *Si (omitiendo k por sencillez)* $\dot{f} = S_{PPH}(f) + d(f)$, $\dot{g} = S_{PPH}(g) + d(g)$, $\ddot{f} = S_{PPH}(\dot{f}) + d(\dot{f})$ *y* $\ddot{g} = S_{PPH}(\dot{g}) + d(\dot{g})$ *entonces*

$$
\begin{array}{ll}
1) & ||D\dot{f}||_{l_\infty(\mathbb{Z})} \le \frac{1}{2}||Df||_{l_\infty(\mathbb{Z})} + ||Dd(f)||_{l_\infty(\mathbb{Z})}, \\
2) & |D(\dot{f}_j - \dot{g}_j)| \le \frac{1}{2}||D(f-g)||_{l_\infty(\mathbb{Z})} + ||Dd(f) - Dd(g)||_{l_\infty(\mathbb{Z})}, \text{ para } j = 2n+1, \\
 & |D(\dot{f}_j - \dot{g}_j)| \le ||D(f-g)||_{l_\infty(\mathbb{Z})} + ||Dd(f) - Dd(g)||_{l_\infty(\mathbb{Z})}, \text{ para } j = 2n, \\
y & \\
3) & ||D(\ddot{f} - \ddot{g})||_{l_\infty(\mathbb{Z})} \le \frac{3}{4}||D(f-g)||_{l_\infty(\mathbb{Z})} \\
 & \qquad + ||Dd(f) - Dd(g)||_{l_\infty(\mathbb{Z})} + ||Dd(\dot{f}) - Dd(\dot{g})||_{l_\infty(\mathbb{Z})}.
\end{array}
\tag{11}
$$

**Estabilidad de la multirresolución PPH** $\{f^0, d^0, \ldots, d^{L-1}\} \mapsto f^L$**:**

**Teorema 1** *Para cualquier par de elementos* $f^L, \widetilde{f}^L \in l_\infty(\mathbb{Z})$ *y sus descomposiciones PPH:* $\{f^0, d^0, \ldots, d^{L-1}\}$ *y* $\{\widetilde{f}^0, \widetilde{d}^0, \ldots, \widetilde{d}^{L-1}\}$, *se tiene que*

$$
||f^L - \widetilde{f}^L||_{l_\infty(\mathbb{Z})} \le 9\left(||f^0 - \widetilde{f}^0||_{l_\infty(\mathbb{Z})} + \sum_{k=0}^{L-1} ||d^k - \widetilde{d}^k||_{l_\infty(\mathbb{Z})}\right). \tag{12}
$$

**Estabilidad de la descomposición PPH** $f^L \mapsto \{f^0, d^0, \ldots, d^{L-1}\}$**:**

**Teorema 2** *Dados* $\{f^0, d^0, \ldots, d^{L-1}\}$ *y* $\{\widetilde{f}^0, \widetilde{d}^0, \ldots, \widetilde{d}^{L-1}\}$ *dos descomposiciones PPH, correspondientes a* $f^L, \widetilde{f}^L \in l_\infty(\mathbb{Z})$, *se tiene que*

$$
\begin{array}{rcl}
||f^0 - \widetilde{f}^0||_{l_\infty(\mathbb{Z})} & \le & ||f^L - \widetilde{f}^L||_{l_\infty(\mathbb{Z})}, \\
||d^k - \widetilde{d}^k||_{l_\infty(\mathbb{Z})} & \le & 3\ ||f^L - \widetilde{f}^L||_{l_\infty(\mathbb{Z})},\ \forall 0 \le k \le L-1.
\end{array}
$$

Las pruebas de todos estos resultados pueden encontrarse en [11].

[2] Un esquema de subdivisión $S$ se dice convergente con suavidad Hölder $s$ si, para toda sucesión $f^0 \in l_\infty(\mathbb{Z})$, la sucesión de funciones lineales a trozos $\phi^k$ interpolando los puntos $f_j^k$ en $x_j^k$ converge a una función $\phi$ de regularidad Hölder $s$.

## 4 Aplicaciones en el procesado de señales

En esta sección introducimos varias aplicaciones de la reconstrucción PPH.

### 4.1 Compresión de señales localmente oscilatorias con discontinuidades

Para este tipo de aplicaciones ni las técnicas locales (basadas en descomposiciones multiescala) ni las globales (basadas en Analísis de Fourier) obtienen buenas tasas de compresión. En esta sección presentamos un esquema combinado, mezclando la reconstrucción PPH con la transformada discreta de Fourier.

Nuestro esquema está basado en el trabajo [26], donde se utilizan esquemas interpolatorios tipo ENO (essential non oscillatory). El inconveniente de este tipo de interpolaciones es que son menos locales que el PPH y su falta de estabilidad [21].

Siguiendo [26], definimos el siguiente *esquema combinado*:

**1)** Determinar y marcar los límites de cada intervalo oscilatorio $I_m$.

**2)** Computar la multirresolución PPH de $f^L$.

**3)** Comprimir esta representación del siguiente modo. Dado $\epsilon$ el parámetro de truncación, entonces

$$\tilde{d}_j^k = \mathbf{tr}(d_j^k, \epsilon) = \begin{cases} 0 & |d_j^k| \leq \epsilon \\ 0 & x_j^k \in I_m \\ d_j^k & \text{en otro caso.} \end{cases}$$

donde $d_j^k$ representa los diferentes detalles.

**4)** Computar el error entre la reconstrucción a partir del nivel más bajo de resolución 0 y la señal original,

$$E_j = f_j^L - R_0(x_j^L).$$

**5)** Computar localmente la TDF (Transformada Discreta de Fourier) de la sucesión $\{E_j\}$ para cada $j$ tal que $x_j^L \in I_m$.

**6)** Comprimir los coeficientes $A_j^m$ y $B_j^m$ de la TDF,

$$\tilde{C}_j^m = \mathbf{tr}(C_j^m, \epsilon^F) = \begin{cases} 0 & |C_j^m| \leq \epsilon^F \\ C_j^m & \text{en otro caso.} \end{cases}$$

Al final del proceso obtenemos una representación de $f^L$:

$$\{f^0, (\tilde{d}^1, \tilde{d}^2, \ldots, \tilde{d}^L), \tilde{A}_j^m, \tilde{B}_j^m\}.$$

Consideramos la siguiente señal localmente oscilatoria y con discontinuidades:

$$f(x) = \begin{cases} sin(0{,}2x) & x \in [-2\pi, 0) \cap [2\pi, 4\pi), \\ sin(0{,}2x) + 0{,}2sin(10x) & x \in [0, 2\pi), \\ sin(0{,}2x) + 1 & \text{en otro caso.} \end{cases} \tag{13}$$

Consideramos LIN4-la transformada lineal wavelet interpolatoria de 4 puntos, TDF-la transformada discreta de Fourier, PPH-la multirresolución PPH y finalmente PPH-TDF-el esquema combinado.

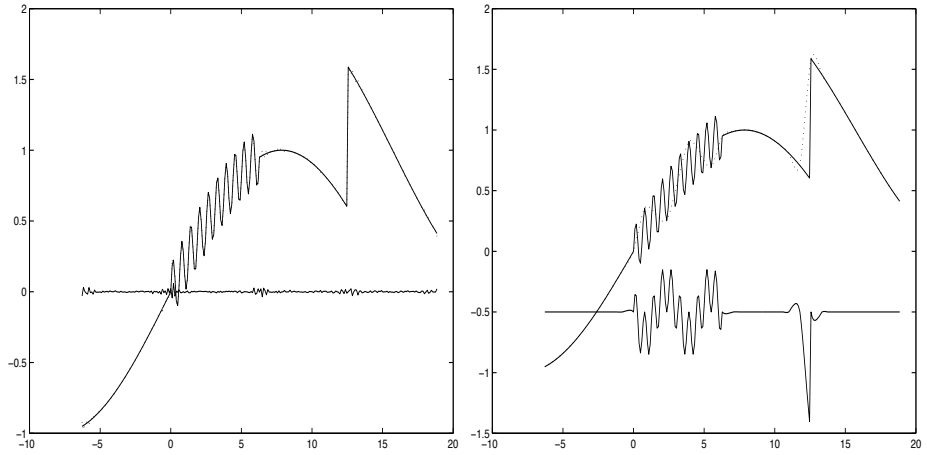


Figura 1: izquierda TDF, derecha LIN4, original, reconstrucción y error

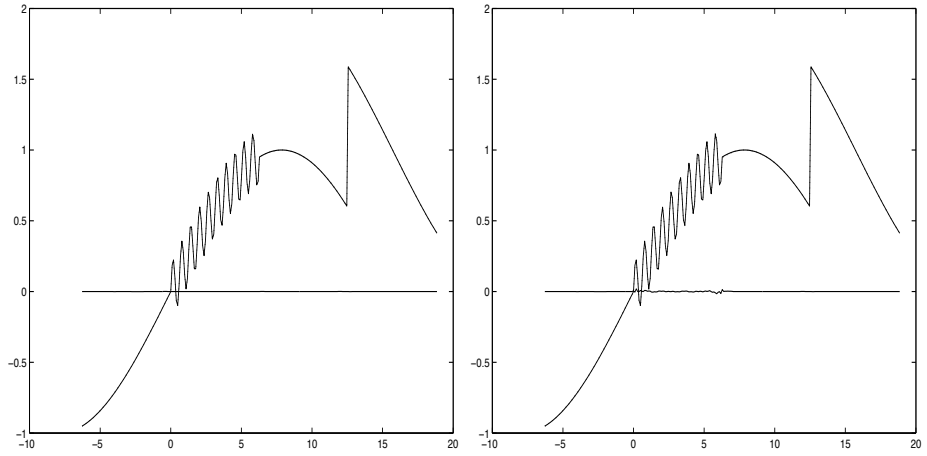


Figura 2: izquierda PPH, derecha PPH-TDF, original, reconstrucción y error

En la tabla 1 y en las figuras 1-2 podemos observar el buen comportamiento del esquema combinado. El error más pequeño es producido por el PPH pero necesita 20 coeficientes más, correspondientes a la zona oscilatoria. También, se observa el fenómeno de Gibbs en la interpolación lineal.

Para más detalles y experimentos ver [5].

| $\epsilon = 0{,}001$, $L = 4$ | LIN4 | TDF | PPH | PPH-TDF |
|---|---|---|---|---|
| nnc | 94 | 185 | 88 | 68 |
| $r_c$ | $3{,}67\ e-01$ | $7{,}23\ e-01$ | $3{,}44\ e-01$ | $2{,}66\ e-01$ |
| $l_1$ | $14{,}39\ e+00$ | $1{,}55\ e+00$ | $2{,}82\ e-02$ | $2{,}52\ e-01$ |
| $l_\infty$ | $9{,}02\ e-01$ | $4{,}40\ e-02$ | $6{,}34\ e-04$ | $1{,}55\ e-02$ |
| $l_2^2$ | $2{,}15\ e+00$ | $1{,}56\ e-01$ | $3{,}10\ e-03$ | $4{,}09\ e-02$ |

Cuadro 1: Número de coeficientes no ceros, ratio de compresión, errores en las normas $l_\infty$, $l_1$ y $l_2^2$, función $f(x)$, $(J_L + 1) = 257$ puntos, $\epsilon^F = 0{,}003$.

### 4.2 Aplicación al procesado de imágemes

Esta sección está dedica a la aplicación de la multirresolución PPH en la compresión y eliminación de ruido en imágenes.

Para analizar las propiedades de compresión de un algoritmo de multirresolución bi-dimensional $M$

$$A^k \leftrightarrow \left( \begin{array}{c|c} A^{k-1} & \Delta_2^k \\ \hline \Delta_3^k & \Delta_1^k \end{array} \right) \tag{14}$$

se introduce $\epsilon$, un parámetro de truncación, y un operador de truncamiento (*hard-threshold*) $\mathbf{tr}^\epsilon$ definido como

$$\mathbf{tr}^\epsilon(A^0, \Delta) = (A^0, \hat{\Delta}),$$

con

$$(\hat{\Delta}_l^k)_{ij} = \left\{ \begin{array}{ll} 0 & |(\Delta_l^k)_{ij}| \leq \epsilon, \\ (\Delta_l^k)_{ij} & \text{en otro caso.} \end{array} \right.$$

Usaremos el mismo truncamiento $\epsilon$ para todos los niveles $k$ de multirresolución. Otro truncamiento, más adaptado a la eliminación de ruido, es el *soft-threshold* [27]

$$\hat{\Delta}_i^k = \eta_\epsilon \left(\Delta_i^k\right) = sign\left(\Delta_i^k\right) * \text{máx}\left(abs(\Delta_i^k) - \epsilon, 0\right).$$

Después de la truncación, la transforma inversa de multirresolución $M^{-1}$ es aplicada para obtener

$$\hat{A}^L = M^{-1}\mathbf{tr}^\epsilon(MA^L),$$

la cual compararemos usando las siguientes normas:

$$\begin{aligned} ||A^L - \hat{A}^L||_{l_p} &= \left( \frac{1}{(J_L+1)^2} \sum_i |A_i^L - \hat{A}_i^L|^p \right)^{1/p}, \quad p = 1, 2, \\ ||A^L - \hat{A}^L||_{l_\infty} &= \max_i(|A_i^L - \hat{A}_i^L|). \end{aligned}$$

En este sentido, la estabilidad de la multirresolución es una condición imprescindible para recuperar una buena reconstrucción de la imagen.

El ratio de compresión lo definimos como en [33], [34] mediante

$$r_c = \frac{nnc}{(J_L+1)\times(J_L+1)-(J_0+1)\times(J_0+1)},$$

donde $nnc$ denota el número de detalles no cero.

Con el fin de obtener mejor eliminación de ruido hemos utilizado el *soft-threshold* variando el parámetro de truncación en cada subbanda y en cada escala de la siguiente forma:

$$\begin{aligned}
\epsilon_1^k &= 2 * \frac{\sigma\sqrt{2\ln(M_1^k)}}{(k+1)*(k+1)} \\
\epsilon_2^k &= \frac{\sigma\sqrt{2\ln(M_2^k)}}{(k+1)*(k+1)} \\
\epsilon_3^k &= \frac{\sigma\sqrt{2\ln(M_3^k)}}{(k+1)*(k+1)}
\end{aligned}$$

donde $M_1^k$, $M_2^k$, $M_3^k$ son respectivamente los tamaños de las matrices $\Delta_1^k$, $\Delta_2^k$ y $\Delta_3^k$, $k = 1, 2, \ldots, L$, y $\sigma^2$ denota la varianza del ruido (Gaussiano) de la imágen.

### 4.2.1 Multirresolución mediante producto tensor

Analizaremos los resultados obtenidos para el procesamiento de la imagen del cámara, ver figura 3.


Figura 3: Imagen del cámara

Consideramos $J_L = 256$ (el tamaño de la imagen es $257 \times 257$) donde $L = 4$ (el nivel más fino de resolución) y $\epsilon = 10$ (parámetro de truncación).

Construiremos las multirresoluciones provenientes de productos tensoriales, tanto de la interpolación lineal de 4 puntos como de nuestra reconstrucción PPH. Un zoom de las imágenes reconstruidas sobre una zona donde la presencia de los ejes (corresponde a la discontinuidad de la imagen) es importante se muestra en la figura 4. La multirresolución PPH es la que obtiene mejores resultados (ver también la tabla 2).


Figura 4: Arriba-izquierda Original, arriba-derecha LIN4, abajo PPH, $L = 4$, $\epsilon = 10$

Más ejemplos se pueden consultar en [9].

### 4.2.2 Multirresolución mediante el formato Quincunx

Con el fin de adaptarse mejor a la presencia de los ejes de las imágenes, se consideran formatos más sofisticados que el proveniente de productos

| $\epsilon = 10$ | LIN4 | PPH |
|---|---|---|
| nnc | 12580 | 12100 |
| $r_c$ | $1{,}91e-01$ | $1{,}84e-01$ |
| $l_1$ | 3,82 | 3,25 |
| $l_\infty$ | 31,30 | 29,93 |
| $l_2$ | 5,23 | 4,56 |

Cuadro 2: Imagen del cámara ($J_L = 256$): Número de coeficientes no cero, ratio de compresión, error en la reconstrucción $l_\infty$, $l_1$ y $l_2$.

tensoriales. En [7], introducimos la multirresolución de Harten en uno de estos formatos no separables, en concreto en el Quincunx.

Consideramos el cuadrado unidad $[0,1] \times [0,1]$.

La transformación $T(x,y) = (x+y, x-y)$ define la submalla de la pirámide quincunx. Notar que $T^2 = 2Id$, que es la submalla del producto tensorial. Así, tomaremos '$L$' el nivel de resolución más fino, un número par.

Sean $X^L = \{x_i^L, y_j^L\}_{i,j=0}^{J_L}$, $x_i^L = ih_L$, $y_j^L = jh_L$, $h_L = 2^{-L}h_0$, $J_L = 2^L J_0$, $J_0$ entero, $h_0 = \frac{1}{J_0}$, $L$ par.

Como $T^2 = 2Id$, se obtiene, para $i,j = 0,\ldots,\frac{J_L}{2}$, $x_{2i}^L = x_i^{L-2}$ y $y_{2j}^L = y_j^{L-2}$.

Las conexiones entre $L$ y $L-1$ o $L-1$ y $L-2$ son más complicadas. Para el primer paso, se tiene para $j = 0,\ldots,J_L$

$$\begin{aligned}(x_{2i}^L, y_j^L) &= (x_i^{L-1}, y_j^{L-1}), \quad i = 0,\ldots\frac{J_L}{2}\ j\ par,\\ (x_{2i-1}^L, y_j^L) &= (x_i^{L-1}, y_j^{L-1}), \quad i = 1,\ldots\frac{J_L}{2}\ j\ impar,\end{aligned}$$

y para el segundo

$$(x_i^{L-1}, y_{2j}^{L-1}) = (x_i^{L-2}, y_j^{L-2}), \quad i,j = 0,\ldots\frac{J_L}{2}.$$

Los siguientes pasos se hacen de forma similar.

Consideramos la discretización

$$\mathcal{D}_k : \mathcal{C}([0,1]\times[0,1]) \longrightarrow V^k \qquad \bar{f}_{i,j}^k = (\mathcal{D}_k f)_{i,j} = f(x_i^k, y_j^k), \tag{15}$$

donde si $k$ es par $0 \le i,j \le J_k$, con $J_k := \frac{J_L}{2^{\frac{L-k}{2}}}$ y si $k$ es impar $0 \le j \le 2J_k$ y

$$\begin{aligned}0 \le i \le J_k,\ j\ par,\\ 1 \le i \le J_k,\ j\ impar,\end{aligned}$$

con $J_k := \frac{J_L}{2^{\frac{L-k+1}{2}}}$.

En este caso, $\dim V^k = (J_k+1)\times(J_k+1)$ o $\dim V^k = ((J_k+1)\times(J_k+1)) + (J_k \times J_k)$ respectivamente. Los operadores de decimación son para $k$ par

$$\begin{aligned}\bar{f}_{i,j}^{k-1} &= (D_k^{k-1}\bar{f}^k)_{i,j} = \bar{f}_{2i,j}^k, \quad j\ par,\\ \bar{f}_{i,j}^{k-1} &= (D_k^{k-1}\bar{f}^k)_{i,j} = \bar{f}_{2i-1,j}^k, \quad j\ impar,\end{aligned}$$

y para $k$ impar

$$\bar{f}^{k-1}_{i,j} \quad = \quad (D^{k-1}_k \bar{f}^k)_{i,j} = \bar{f}^k_{i,2j}.$$

En particular, se obtiene para $k$ par

$$\mathcal{N}(D^{k-1}_k) = \{v^k \in V^k \ : \ v^k_{2i,j} = 0, \ j \ par, \ v^k_{2i-1,j} = 0, \ j \ impar\},$$

y para $k$ impar

$$\mathcal{N}(D^{k-1}_k) = \{v^k \in V^k : v^k_{i,2j} = 0\}.$$

Así, si se denota por $e^k$ los errores en la predicción, sólo se necesita guardar $e^k_{2i-1,j}$ $j$ par, $e^k_{2i,j}$ $j$ impar y y $e^k_{i,2j-1}$ respectivamente.

Una reconstrucción para esta discretización es un operador $\mathcal{R}_k$ tal que

$$\mathcal{R}_k : V^k \longrightarrow \mathcal{C}([0,1] \times [0,1]); \qquad \mathcal{D}_k \mathcal{R}_k \bar{f}^k = \bar{f}^k, \tag{16}$$

lo que significa que

$$(\mathcal{R}_k \bar{f}^k)(x^k_i, y^k_j) = \bar{f}^k_{i,j} = f(x^k_i, y^k_j). \tag{17}$$

Por lo tanto, $\mathcal{R}_k$ será una función continua que interpola $\bar{f}^k$ en el mallado $X^k$. Finalmente, se definen los operadores de predicción como

$$P^k_{k-1} := \mathcal{D}_k \mathcal{R}_{k-1}. \tag{18}$$

En [7] se propone una reconstrucción quincunx-PPH, usando la interpolación PPH en una dimensión. Para cada detalle se elige entre las dos direcciones principales, la dirección con diferencias divididas asociadas más pequeñas en valor absoluto (ver figura 5). Entonces se aplica la reconstrucción PPH en 1-D en la dirección seleccionada.

Figura 5: *Los círculos son usados para predecir los cuadrados. Para k par, derecha: reconstrucción desde el nivel $k-2$ a $k-1$, izquierda: reconstrucción desde el nivel $k-1$ a $k$.*

En la implementación usamos algoritmos de error control para prevenir inestabilidades causadas por la elección de la dirección a interpolar. Estos algoritmos permiten tener un control total del error. La idea es primero computar el nivel más grosero $f^0$ y procesarlo obteniendo $\hat{f}^0$, a partir de él aproximar $f^1$, calcular los detalles asociados $d^1$ y procesarlos. Con $\hat{f}^0$ y $\hat{d}^1$ obtener una aproximación $\hat{f}^1$ y usarla para aproximar $f^2$ calculando los detalles $d^2$. Una vez procesados dichos detalles encontrar una aproximación $\hat{f}^2$ y usarla

para aproximar $f^3$ y calcular los detalles $d^3$. El proceso continúa hasta llegar a computar los detalles asociados al último nivel de resolución, obteniendo la representación

$$\{\hat{f}^0, \hat{d}^1, \ldots, \hat{d}^L\}.$$

En [7] se puede consultar todos los detalles.

Realizamos un estudio comparativo utilizando como indicador de calidad el $PSNR$ (Peak Signal Noise Ratio) [41]. Para una imagen 8 bit $(0 - 255)$, el $PSNR$ se define como

$$PSNR = 20 \log_{10} \left(\frac{255}{||f^L - \hat{f}^L||_{l_2}}\right)$$

En la tabla 3, se considera el $PSNR$ contra el número de coeficientes no cero. Se puede observar que fijado un nivel de calidad en la reconstrucción la compresión obtenida mediante el quincunx-PPH con error-control es considerablemente mayor a la obtenida mediante una multirresolución lineal separable del mismo orden.

El objetivo de los operadores no lineales tipo quincunx es mejorar la aproximación en regiones cercanas a las singularidades (ejes en la imagen).

| PSNR | LIN4 | PPH-Quincunx (E-C) |
|---|---|---|
| 30 | 9481 | 3559 |
| 35 | 14147 | 7635 |
| 40 | 21518 | 12922 |
| 45 | 32430 | 19015 |

Cuadro 3: Imagen del cámara, Número de detalles no cero, $L = 4$, Quincunx

### 4.2.3 Eliminación de ruido en imágenes

En esta sección, presentaremos una comparación entre multirresoluciones lineales y no lineales para la eliminación de ruido en imágenes. Se usará el truncamiento *soft-threshold* propuesto por Donoho.

Denotaremos por $r_{scheme}$ el $PSNR$ entre la imagen que da el esquema y la imagen original sin ruido. Calcularemos $R_{PPH/LIN4} = \frac{r_{PPH}}{r_{LIN4}}$.

Consideramos diferentes niveles de ruido entre 10 y 50.

La mejor adaptación de la multirresolución PPH a la presencia de los ejes hace que también obtenga mejores resultados en la eliminación de ruido (ver figura 6).

En [8] se pueden encontrar más experimentos y detalles.

## 5 Conclusiones e investigación actual

El objetivo de este artículo ha sido estudiar un operador de reconstrucción no lineal adaptado a la presencia de discontinuidades. Dentro de un algoritmo

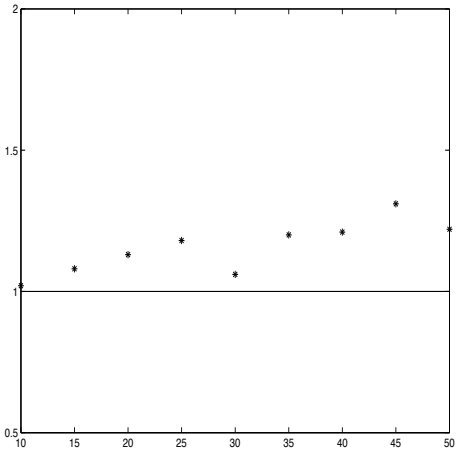



Figura 6: Imagen del cámara, $R_{PPH/LIN4}$ contra nivel de ruido

de multirresolución se ha estudiado teóricamente y se han presentado varios ejemplos comparándolo de forma favorable con esquemas lineales del mismo orden.

Nuestro grupo continúa trabajando sobre ideas relacionadas con la reconstrucción PPH:

- Obtención de algoritmos PPH para discretizaciones en valores en media (en lugar del caso puntual) que son más adaptadas para el procesado de imágenes.
- Definición y análisis de reconstrucciones no lineales de alto orden.
- Generalización del análisis de estabilidad para otro tipo de esquemas no lineales.
- Definición y análisis de esquemas tipo Hermite-PPH.

**Agradecimientos**

Quisiera agradecer a todos los coautores, que han trabajado conmigo en la reconstrucción PPH: I.Ali, S.Busquier, V.F.Candela, H.Cherif, K.Dadourian, R.Donat, D.El Kebir, J.Liandrat, J.Molina, J.Ruiz y J.C.Trillo.

Agradecer las sugerencias propuestas por el *referee* y el editor.

## Referencias

[1] Ali I., Amat S. and Trillo J.C., (2006). Point values Hermite multiresolution for non-smooth noisy signals. *Computing*, **77** 3, 223–236.

[2] Amat S., Aràndiga F., Cohen A. and Donat R., (2002). Tensor product multiresolution analysis with error control for compact image representation. *Signal Processing*, **82**(4), 587-608.

[3] Amat S., Busquier S. and Candela V.F., (2003). A polynomial approach to Piecewise Hyperbolic Method, *Int.J. Computational Fluid Dynamics* **17**(3), 205-217.

[4] Amat S., Busquier S. and Candela V.F.,, (2003). Local Total Variation Bounded methods for hyperbolic conservation laws, *J. of Comp. Methods in Sciences and Engineering*, **3**(3), 193-200.

[5] Amat S., Busquier S., El Kebir D. and Molina J., (2002). Compression of locally oscillatory signals with discontinuities, *International Mathematical J.*, **2**(12), 1141-1156.

[6] Amat S., Busquier S. and Trillo J.C., (2005). Stable Interpolatory Multiresolution in 3D , *Applied Numerical Analysis and Computational Mathematics*, **2**(2), 177-188.

[7] Amat S., Busquier S. and Trillo J.C., (2005). Non-linear Harten's Multiresolution on the Quincunx Pyramid , *J. of Comp. and App. Math.*, **189** (1-2), 555-567.

[8] Amat S., Cherif H. and Trillo J.C., (2005). Denoising using Linear and Nonlinear Multiresolutions, *Engineering of Computations*, **22** 7, 877-891.

[9] Amat S., Donat R., Liandrat J. and Trillo J.C., (2006). Analysis of a new nonlinear subdivision scheme. Applications in image processing. *Foundations of Computational Mathematics*, **6** 2, 193–225.

[10] Amat S., Donat R., Liandrat J. and Trillo J.C., (2007). A fully adaptive PPH multiresolution scheme for image processing. *Mathematical and Computer Modelling*, **46** 1-2, 2-11.

[11] Amat S. and Liandrat J., (2005). On the stability of the PPH nonlinear multiresolution, *Appl. Comp. Harm. Anal.*, **18**(2), 198-206.

[12] Aràndiga, F., Baeza, A. and Donat, R., (2004). Discrete multiresolution based on Hermite interpolation: computing derivatives. Recent advances in computational and mathematical methods for science and engineering. *Commun. Nonlinear Sci. Numer. Simul.*, **9** 2, 263–273.

[13] Aràndiga, F. and Belda, Ana M., (2004). Weighted ENO interpolation and applications. Recent advances in computational and mathematical methods for science and engineering. *Commun. Nonlinear Sci. Numer. Simul.*, **9** 2, 187–195.

[14] Aràndiga, F., Cohen, A., Donat, R. and Dyn, N., (2005). Interpolation and approximation of piecewise smooth functions. (English summary) *SIAM J. Numer. Anal.*, **43** 1, 41–57 (electronic).

[15] Aràndiga F. and Donat R., (2000). Nonlinear Multi-scale Decomposition: The Approach of A.Harten, *Numerical Algorithms*, **23**, 175-216.

[16] Binev, P., Dahmen, W., DeVore, R. and Dyn, N., (2004). Adaptive approximation of curves. Approximation theory: a volume dedicated to Borislav Bojanov, 43–57, Prof. M. Drinov Acad. Publ. House, Sofia.

[17] Cohen, A. Theoretical, applied and computational aspects of nonlinear approximation. (English summary) Multiscale problems and methods in numerical simulations, 1–29, Lecture Notes in Math., 1825, Springer, Berlin, 2003.

[18] Cohen, A. Numerical analysis of wavelet methods. Studies in Mathematics and its Applications, 32. North-Holland Publishing Co., Amsterdam, 2003.

[19] Cohen, A., Dahmen, W., Daubechies, I. and DeVore, R., (2001). Tree approximation and optimal encoding. *Appl. Comput. Harmon. Anal.*, **11** 2, 192–226.

[20] Cohen, A. and Dyn, N., (1996). Nonstationary subdivision schemes and multiresolution analysis. *SIAM J. Math. Anal.*, **27** 6, 1745–1769.

[21] Cohen A., Dyn N. and Matei B., (2003). Quasi linear subdivision schemes with applications to ENO interpolation. *Applied and Computational Harmonic Analysis*, **15**, 89-116.

[22] Dahmen, W. Multiscale and wavelet methods for operator equations. Multiscale problems and methods in numerical simulations, 31–96, Lecture Notes in Math., 1825, Springer, Berlin, 2003.

[23] Daubechies I., Runborg O. and Sweldens W., (2004). Normal multiresolution approximation of curves, *Const. Approx.*, **20** 3, 399–463.

[24] Deslauriers G. and Dubuc S., (1989). Symmetric iterative interpolation processes, *Const. Approx.*, **5**, 49-68.

[25] Donat, R. and Marquina, A., (1996). Capturing shock reflections: an improved flux formula. *J. Comput. Phys.*, **5** 1, 42–58.

[26] Donat R. and Harten A., (1993). Data compression Algorithms for Locally Oscillatory Data, UCLA CAM Report 93-26.

[27] Donoho D., (1995). Denoising by soft thresholding, *IEEE Trans. on Inform. Theory*, **41**(3), 613-627.

[28] Donoho D., Yu T.P.-Y, (2000). Nonlinear pyramid transforms based on median interpolation. *SIAM J. Math. Anal.*, **31**(5), 1030-1061.

[29] Dyn, N., Kuijt, F., Levin, D. and van Damme, R., (1999). Convexity preservation of the four-point interpolatory subdivision scheme. *Comput. Aided Geom. Design*, **16** 8, 789–792.

[30] Dyn, N. and Levin, D., (2002) Subdivision schemes in geometric modelling. *Acta Numer.*, **11**, 73–144.

[31] Dyn, N., Levin, D. and Luzzatto, A., (2003) Refining oscillatory signals by non-stationary subdivision schemes. Modern developments in multivariate approximation, 125–142, Internat. Ser. Numer. Math., 145, Birkhäuser, Basel.

[32] Floater M. S. and Michelli C.A., (1998). Nonlinear stationary subdivision, *Approximation theory: in memory of A.K. Varna, edt: Govil N.K, Mohapatra N., Nashed Z., Sharma A., Szabados J.*, 209-224.

[33] Harten A., (1993). Discrete multiresolution analysis and generalized wavelets, *J. Appl. Numer. Math.* **12**,153-192.

[34] Harten A., (1996). Multi resolution representation of data II, *SIAM J. Numer. Anal.*, **33**(3), 1205-1256.

[35] Harten A., Osher S.J., Engquist B. and Chakravarthy S. (1987). Some results on uniformly high-order accurate essentially non-oscillatory schemes. *Appl. Numer. Math.*, **2**, 347-377.

[36] Kuijt F. and van Damme R., (1998). Convexity preserving interpolatory subdivision schemes. *Const. Approx.*, **14**, 609-630.

[37] Marquina, A., (1994). Local piecewise hyperbolic reconstruction of numerical fluxes for nonlinear scalar conservation laws. *SIAM J. Sci. Comput.*, **15** 4, 892–915.

[38] Matei, B., (2005). Smoothness characterization and stability in nonlinear multiscale framework: theoretical results. *Asymptot. Anal.*, **41** 3-4, 277–309.

[39] Matei, B., (2004). Denoising using nonlinear multiscale representations. *C. R. Math. Acad. Sci. Paris*, **338** 8, 647–652.

[40] Oswald P., (2004). Smoothness of Nonlinear Median-Interpolation Subdivision, *Adv. Comput. Math.*, **20**(4), 401-423.

[41] Rabbani M. and Jones P.W., (1991). Digital Image Compression Techniques. Tutorial Text, Society of Photo-Optical Instrumentation Engineers (SPIE), TT07.

[42] Shu C.W. and Osher S.J. (1987). Efficient implementation of essential non-Oscillatory shock capturing schemes, *J.Comput.Phys.*, 77, 231-303.

[43] Shu C.W. and Osher S.J. (1989). Efficient implementation of essential non-Oscillatory shock capturing schemes II, *J.Comput.Phys.*, 83, 32-78.

[44] Trillo, J.C. Multirresolución no lineal y Aplicaciones, PhD in the University of Valencia, Spain, 2006.

*Bol. Soc. Esp. Mat. Apl.*
*nº 44(2008), 101–114*



# UN VIAJE MATEMÁTICO POR EL ESPACIO EUROPEO DE EDUCACIÓN SUPERIOR

MARíA VICTORIA CUEVAS Y ANTONIO NEVOT

Departamento de Matemática Aplicada
Escuela Universitaria de Arquitectura Técnica
Universidad Politécnica de Madrid

mariavictoria.cuevas@upm.es antonio.nevot@upm.es

**Resumen**

La adaptación y preparación, tanto por parte de los profesores como de los estudiantes, al Espacio Europeo de Educación Superior, supone un reto y al mismo tiempo una investigación permanente. Así, utilizando el símil de un viaje en tren, este artículo pretende mostrar algunas pinceladas de una experiencia piloto desarrollada con un grupo numeroso de estudiantes en la asignatura de primer curso, Fundamentos Matemáticos, de una Escuela Técnica y con la participación de dos profesores. Combinar de una forma regular y eficiente a lo largo del curso diversas actuaciones que fomenten el trabajo autónomo, el trabajo en equipo, las búsquedas en la Red, lecturas históricas y sus personajes o exposiciones orales, entre otros, para lograr la adquisición de determinadas competencias generales y específicas, son las aportaciones que aquí se reflejan.

## 1 Los profesores, ¿sabemos y queremos viajar al EEES?

> "Los grandes profesores aparecen, pasan por la vida de los estudiantes, y sólo unos pocos de ellos quizás consigan alguna influencia en el vasto arte de la enseñanza. En la mayoría de los casos, su ingenio perece con ellos".
>
> Kein Bain [1]

Es verdad que las condiciones en las que trabajamos los profesores no siempre hacen fácil la ilusión, la profesionalidad, la creatividad. Pero aún podemos encontrar a muchos profesores, maestros anónimos, que siguen buscando que entrar en el aula sea un placer y motivo de deseo tanto para ellos como para sus alumnos (L.Bazarra, O.Casanova y J.García [2]).

La docencia es una creación científica y artística y, por tanto, muy personal. Lo que a un profesor le funciona en su clase, en su asignatura, con una personalidad determinada, a otro puede que no le funcione.

Si nos detenemos en el profesorado universitario y los deseos de embarcarse en este proyecto que es el Espacio Europeo de Educación Superior, la situación

es tremendamente compleja y variopinta. La mayoría del profesorado se siente seguro de sus conocimientos, del dominio de la asignatura y de cómo desarrolla sus clases. Además, en muchos casos, las relaciones con los estudiantes son gratificantes y les posibilita mantenerse en esa juventud permanente. Sin embargo, es imprescindible que el profesorado se convenza de que hay que modificar sustancialmente la labor docente y para ello hay que prepararse y formarse. No es menos cierto, también, que el estudiante se tiene que preparar y concienciar de que el trabajo diario, la asistencia y participación en las clases o el trabajo en equipo van a ser ingredientes entre otros muchos que van a formar parte de su formación en las universidades.

Con el fin de conocer la realidad de nuestras aulas universitarias, a pesar de cometer errores, evitar generalizaciones y no mostrar las singularidades tan importantes que existen, ha parecido conveniente mostrar aquí algunas de las conclusiones del Seminario organizado por la Comisión Académica, constituida por la Secretaría de Estado de Universidades e Investigación y el Consejo de Coordinación Universitaria, y encargada de realizar un diagnóstico sobre la situación de las metodologías docentes universitarias y proponer medidas para su renovación, y celebrado en la Universidad Politécnica de Madrid (2005). Entre las conclusiones dadas a conocer se ponen de manifiesto diversas causas que dificultan la renovación y que, entre otras, son las siguientes:

- El bajo reconocimiento de la labor docente frente a la investigadora.
- La concentración de los esfuerzos de los docentes en la transmisión de contenidos.
- La escasa preparación pedagógica de los docentes, derivada de una ausencia de formación inicial y permanente.
- La resistencia del profesorado al cambio metodológico.
- La falta de información y concienciación del profesorado del cambio de cultura pedagógica que comporta el EEES.
- La falta de tradición de trabajo cooperativo en la docencia.
- La carencia de modelos universalmente aceptados para evaluar competencias genéricas.
- El tamaño de los grupos, excesivos en algunas titulaciones.
- La dificultad de implicar a los estudiantes en su propio proceso formativo.
- La falta de adecuación de los procesos administrativos a un modelo diversificado que incrementa considerablemente las tareas de planificación y gestión académicas.
- La inadecuada infraestructura de muchos centros cuyas aulas están pensadas para clases magistrales y grupos numerosos.

## 2 De la enseñanza al aprendizaje: los nuevos papeles del profesor y el alumno

Señala I. Pozo [9] que "no es sólo que lo que ayer debía ser aprendido, hoy ya no lo sea, que lo que ayer era culturalmente relevante, hoy lo sea menos. Sino que lo que ha de aprenderse evoluciona a tal velocidad que la forma de aprender y enseñar también debería evolucionar". Nos encontramos, pues, en lo que ha venido en denominarse la "nueva cultura del aprendizaje".

En cuanto a la docencia, los principios básicos de adaptación al EEES (J. González y R.Wagenaar [7]), se pueden citar los siguientes:

- La docencia está centrada en el alumno, preparándolo, sobre todo, para el aprendizaje autónomo.
- El papel del profesor cambia, de estar centrado en la mera transmisión de contenidos, pasa a ser el gesto del proceso de aprendizaje de los alumnos.
- La formación está orientada a la consecución de competencias.
- Una nueva definición del papel formativo de la Universidad, pasando de ser una formación durante un tiempo limitado a ser referencia en la formación a lo largo de toda la vida.
- Los materiales didácticos se transforman en recursos actualizados que incorporan las TIC.

El profesor, por tanto, asume así el papel de entrenador de un equipo al que debe efectuar un seguimiento permanente, cuidando especialmente la comunicación con los alumnos. Aunque pareciera lo contrario, el papel del profesor en estas metodologías activas es crucial. En la tabla 1 se muestran los nuevos papeles del profesor y el alumno según A.Benito y A.Cruz [3].

| | |
|---|---|
| Clase magistral | Clase magistral<br>Metodologías activas<br>Seguimiento académico |
| Exámenes | Evaluaciones alternativas |
| Asistencia a clase<br>Estudio | Asistencia y participación en clase<br>Trabajo guiado<br>Trabajo en equipo<br>Trabajo autónomo<br>Estudio |

Tabla 1: Nuevos papeles del profesor y del alumno

Evidentemente esta transformación supone un reto para el estudiante, puesto que pasar de ser sujeto pasivo y sólo recibir a una situación de ser activo y participar. De tal manera que, lleva aparejado un cambio de la concepción estudiantil y por supuesto personal.

## 3 Un viaje piloto de adaptación al EEES

En nuestra opinión no se puede entender la labor de un profesor sin una buena práctica docente, es por ello que en este artículo se pretende dar a conocer, utilizando el símil de un viaje en tren, una experiencia piloto de adaptación al Espacio Europeo de Educación Superior desarrollada durante los cursos académicos 2005-2006 y 2006-2007 en sendos grupos de 80 estudiantes de la asignatura de Fundamentos Matemáticos de primer curso de los estudios conducentes a la titulación de Arquitecto Técnico en la Universidad Politécnica de Madrid. Se trata de una asignatura troncal con una carga docente actual de 15 créditos, lo que equivale realizando la conversión a 10 ECTS.

Obviamente se trata de una experiencia reciente y, por tanto, sujeta a múltiples interpretaciones, modificaciones y sugerencias de los propios participantes y del resto de compañeros. Se podrá coincidir en algunos planteamientos y por supuesto disentir en otros. Pero de lo que no cabe la menor duda es que se ha realizado con honestidad, entusiasmo y mucho trabajo, sobre todo, con el objetivo de mejorar, por un lado, el aprendizaje y en general la formación de los estudiantes, y por otro, nuestra labor docente.

## 4 Preparación del viaje

Se podría considerar con cierta imaginación que la experiencia realizada con los alumnos ha sido un viaje en tren, donde el punto de partida era la LOU y el destino el EEES. Los alumnos son los pasajeros y los profesores asumimos diferentes papeles como conductores, guías, animadores, revisores, etc.

Si se está preparando un viaje de nueve meses de duración parece razonable conocer algunos datos de los viajeros, que en este caso, obviamente, son los estudiantes. Pues bien, las características más destacables de los dos grupos de estudiantes que realizaron esta experiencia son las siguientes:

- Nota media de ingreso en torno a 6.
- Un alto porcentaje ha elegido estos estudios en primera opción.
- Uno de cada cinco no viven habitualmente en Madrid, en su mayoría proceden de Comunidades Autónomas limítrofes- Castilla León y Castilla La Mancha- y de Andalucía.
- Aproximadamente un 60 % procede de centros públicos y el resto de centros concertados o privados.
- Aproximadamente el 40 % son mujeres y el 60 % hombres.
- En cuanto a los conocimientos previos de Matemáticas de Bachillerato existe una gran dispersión de resultados. Diríamos que es necesaria una escala de 0 a 100 para poder reflejar mejor la situación de partida. Una parte de estas diferencias puede considerarse achacable a los contenidos reales de los centros de estudio de procedencia y/o autonómicos.

El tren utilizado para realizar este viaje está dotado de distintos vagones como, por ejemplo, el de equipaje, el de trabajo guiado con profesor, el de trabajo en equipo con seguimiento académico o el de trabajo autónomo activo con seguimiento académico.

## 5 Viajeros al tren

Para viajar se necesita, obviamente, un billete y, en este caso el billete venía dado en forma de "contrato de aprendizaje" entre los estudiantes y los profesores de la asignatura de tal forma que permitiera que el desarrollo del viaje se desarrollase con ciertas obligaciones por ambas partes. Para ello, los profesores entregamos unas guías de trabajo autónomo y en equipo en las que figuraban las actividades correspondientes y las fechas de entrega. El estudiante, por otra parte, se comprometía a la entrega de los trabajos y a la realización de las pruebas escritas en los plazos y fechas establecidos. Además, asumían conjuntamente el compromiso de colaboración para que el trabajo en equipo cumpla los objetivos propuestos.

Se estableció un plan para los primeros días de viaje. Así, comenzamos por explicarles en qué consistía esta experiencia piloto y cómo la íbamos a llevar a cabo. Posteriormente, hicimos una presentación del Espacio Europeo de Educación Superior, sus fases, el estado actual y las previsiones.

Se fijaron como objetivos, siguiendo la metodología *Tuning* [7], determinadas competencias específicas de la asignatura, además de determinadas competencias generales. Las competencias específicas fueron:

- Utilizar y contrastar diversas estrategias para la resolución de cuestiones y ejercicios.
- Adquirir la habilidad de utilizar la terminología adecuada.
- Transcribir problemas reales al lenguaje matemático.
- Desarrollar la capacidad para identificar los mecanismos básicos característicos de cada problema.
- Compartir y comprobar diversas fuentes de información efectuando un análisis crítico.

Mientras que como competencias generales se eligieron las siguientes:

- Capacidad de análisis y síntesis.
- Capacidad de trabajar y aprender en equipo y de forma autónoma.
- Capacidad de organizar y planificar.
- Capacidad de adaptarse a nuevas situaciones.
- Habilidades de expresión oral y escrita.

## 6 Plan de viaje

El programa de la asignatura y el calendario académico nos obligaron a la realización de una planificación muy estricta del desarrollo de esta aventura para poder llevar a buen fin nuestro propósito. No había cabida a la improvisación en ningún momento y mucho menos al azar.

Por ello, pareció conveniente elaborar un protocolo de actuación de todas aquellas tareas que implicase en mayor o menor medida a los profesores y a los estudiantes. Así, elaboramos la "Guía Docente de presentación de la asignatura" que el segundo día lectivo entregamos a cada uno de los estudiantes y en la que intentamos plasmar nuestro trabajo conjunto y, al mismo tiempo, dejar constancia escrita de cada una de las cuestiones más significativas (M.V. Cuevas y A. Nevot [5]).

Además, elaboramos dos tipos de guías, una de trabajo autónomo y otra de trabajo en equipo, para cada uno de los seis bloques en que se dividió el curso. Cada bloque estaba compuesto por uno o dos temas, dependiendo de la naturaleza y duración de los mismos. En estas guías se les proporcionaba una serie de actividades, tanto de forma autónoma como en equipo, así como el formato de entrega, fechas de entrega de los trabajos y de la realización de las pruebas.

Como disponíamos de cinco horas de docencia oficiales para la asignatura, se estableció un plan semanal y, que de forma regular se mantuvo durante todos el curso, con el fin de que permitiera que profesores y alumnos compartiéramos los distintos vagones del tren al EEES (véase tabla 2).

| | |
|---|---|
| Lunes (2 horas) | *Trabajo guiado con profesor.* |
| Martes (1 hora) | *Trabajo guiado con profesor.* |
| Martes (1 hora) | *-Trabajo en equipo con seguimiento académico.*<br>*-Prueba en equipo.*<br>*-Exposición oral.*<br>*-Entrega de trabajos de equipo.* |
| Miércoles (1 hora) | *-Trabajo autónomo activo con seguimiento académico.*<br>*-Prueba individual.*<br>*-Entrega de trabajos individuales.*<br>*-Investigación en el aula de informática.* |

Tabla 2: Plan de trabajo semanal

Desde el primer momento decidimos que el diseño semanal siempre se pusiera en común y que, además, las aportaciones individuales las asumiéramos en su totalidad como propias. Por otra parte, también consideramos fundamental reunirnos para analizar cómo nos sentíamos después de haber guiado o planteado alguna actividad, sobre todo en la hora de trabajo de equipo, qué cosas habían ido bien y cuáles no tan bien, para intentar tomar medidas y modificarlas, siempre lógicamente dentro de nuestro contrato de aprendizaje. Intentamos, además, por todos los medios posibles que las posibles respuestas a preguntas

e interrogantes de los alumnos sobre las actividades propuestas estuvieran consensuadas. Era importantísimo no sembrar ambigüedades.

Por otra parte, la estructura y organización de las diversas sesiones debían estar perfectamente planificadas, señalando las actividades que deben realizar, el tiempo aproximado que llevarán y otros elementos relevantes. Las instrucciones tenían que ser claras, sin ambigüedades y que las entendiera todo el grupo.

## 7 Equipaje

No cabe la menor duda de que uno de los vagones más importantes de este viaje, debido a la delicadeza de su carga, ha sido el vagón del equipaje. Todos los participantes en esta experiencia, profesores y alumnos, partíamos con un lastre (equipaje) que no siempre resulta sencillo dejar a un lado. Por una parte, los profesores llevábamos años impartiendo la asignatura de manera más o menos tradicional, si bien habíamos ido incorporando algunas novedades dentro del estrecho margen que ofrecía el departamento, ya que eran nueve grupos en los que se impartía la asignatura con unos contenidos, objetivos y exámenes comunes. Debíamos, pues, prescindir de una parte de nuestro equipaje anterior para poder incorporar equipaje nuevo.

Sin embargo, no consideramos haber vivido esta experiencia como una ruptura con el trabajo realizado en cursos anteriores, es más, quizá sin ese bagaje de muchos años dedicados a la enseñanza no hubiéramos podido realizarla. Pero sí se puede considerar como un proceso lleno de incorporaciones y de nuevas posibilidades de abordar otras formas de trabajo, tanto entre los propios profesores como entre éstos y los estudiantes.

La procedencia de los pasajeros alumnos era relativamente diversa. Algunos alumnos eran repetidores o, mejor dicho, veteranos. Pero la mayoría eran alumnos que cursaban la asignatura por primera vez, unos pocos procedentes de otras escuelas o facultades, la mayoría de bachillerato y una minoría de módulos profesionales.

## 8 El vagón de trabajo guiado con profesor

En el vagón del trabajo guiado con profesor utilizamos diversos formatos combinados. En algunos casos, a la vieja usanza se utilizaban las clases magistrales pero evitando que los alumnos se dedicasen a copiar apuntes, puesto que ya disponían de todo el material necesario que facilitábamos por Internet a través de la plataforma Aulaweb. En otras ocasiones, se les proponía la resolución de alguna cuestión o ejercicio, que después de unos minutos de elaboración, exponía algún alumno o el profesor en la pizarra. En cualquier caso, en este vagón intentamos despertar el lado crítico de los viajeros y, sobre todo, fomentar el aspecto participativo.

## 9 El vagón de trabajo autónomo

Además de las actividades que debía hacer cada estudiante fuera del aula de forma autónoma, disponía de una hora semanal en el aula, que denominamos "Trabajo Autónomo con seguimiento académico", siempre con la presencia de los dos profesores, los alumnos maduraban los conceptos trabajados en el vagón del trabajo guiado con profesor, realizaban parte del trabajo autónomo que debían entregar al final de cada escala y, preguntaban las posibles dudas que se les planteaban, tanto del tema objeto del trabajo como de temas básicos para el desarrollo de la asignatura. La asistencia de estudiantes se ha mantenido durante todo el curso en torno a un 65 %.

A diferencia de las tutorías, en las que es el alumno el que se desplaza al despacho del profesor, en el vagón de trabajo autónomo, es el profesor el que se desplaza al lugar del trabajo del alumno.

En las "Guías de Trabajo Autónomo" que se facilitaba a cada alumno al comenzar cada uno de los seis bloques (compuesto de uno o dos temas) en los que dividimos el curso, figuraban las actividades, el formato y la fecha de entrega. Las actividades que figuraban en la guía consistían fundamentalmente en la realización y entrega de algunos ejercicios de las prácticas seleccionados por su variedad y en muchos casos de aplicabilidad inmediata, evitando en todo caso la repetición. Además, también se proponían trabajos voluntarios de diversa índole.

## 10 El vagón de trabajo en equipo

Este vagón ha sido uno de los tratados con más delicadeza, pues de su buena utilización dependían muchos resultados. Por otra parte, los profesores éramos conscientes de la importancia y de la dificultad que iba a entrañar la coordinación del vagón.

El objetivo esencial de la enseñanza en pequeño grupo es facilitar la comunicación al animar a los estudiantes a que hablen, reflexionen y se comuniquen con mucha mayor facilidad que en un grupo grande (K. Exley y R. Dennick [6]).

Sin lugar a dudas uno de los temas más complejos de la experiencia ha sido el trabajo en equipo. Comenzamos por el método de formación rompiendo estándares establecidos al agruparlos por proximidad en el aula (vagón), pensando, creo que acertadamente como después comprobamos, que hay cierto grado de afinidad en la mayoría de los estudiantes por el sitio que ocupan, por lo menos con el que se sienta a su lado. Cada equipo de trabajo estaba formado por cuatro estudiantes y un total de veinte equipos. Posteriormente, después de una reunión de media hora de duración donde de alguna forma se conocieron algo más, cada equipo de común acuerdo nombró un coordinador y un secretario, indicándoles en líneas generales cuáles iban a ser sus funciones. En principio pensamos que tuvieran esos puestos durante el primer semestre, para luego cambiar y que todos asumieran alguna responsabilidad más precisa, pero las cosas después no fueron así.

Las "Guías de Trabajo de Equipo" también se entregaron al comienzo de cada uno de los seis bloques. En ellas, figuraban como regla general las actividades que debía realizar cada equipo y que estaba estructurado, en la mayoría de los casos, en varios apartados: resolución de problemas con un grado de dificultad mayor al del trabajo autónomo, investigación de procesos de la vida cotidiana o aplicación a otras materias del tema objeto de estudio, historia de las unidades, curiosidades, búsqueda de enlaces de interés en Internet relacionados con los diversos temas y, en algunas ocasiones, exposiciones orales. A modo de ejemplo, en el tema 1 de Cónicas figuraban las siguientes:

A. Resolución de ejercicios de las prácticas de Fundamentos Matemáticos (Trabajando las cónicas con parámetros).

B. Lectura: Resumen comentado ($\approx$ 2 páginas) del capítulo 2 del libro Lugares geométricos. Cónicas de Río, J. (documento impreso).

C. Investigación: Búsqueda de algunos objetos o fenómenos de la vida cotidiana en la ciencia, la técnica y el arte, incluyendo detalle e indicando expresamente mediante gráficos, fotografías, vídeos,... las características cónicas encontradas.

D. Búsqueda de información abierta en Internet sobre historia de las cónicas (máximo cinco enlaces).

Las actividades de búsqueda abierta de información en Internet, en libros o la lectura de determinados textos, estaban fundamentalmente relacionadas con la historia de las matemáticas y algunos de sus personajes más importantes en el desarrollo de algunos de los temas de los contenidos del curso de esta asignatura. Asimismo, también intentamos que ligaran las matemáticas a otras disciplinas, sobre todo desde la perspectiva de la arquitectura, pero no exclusivamente. En algunos casos les facilitamos textos concretos de referencia como punto de partida.

Quizá la sesión semanal de trabajo de equipo en el aula haya sido la que más altibajos y cambios de rumbo nos haya supuesto sobre todo el primer año de la experiencia. Cada uno de los profesores nos encargamos de coordinar la mitad de los equipos. En cada sesión nos reuníamos con cada uno de los grupos, aclarándoles y guiándoles en las actividades que debían hacer, respondiendo a sus preguntas, observando cómo llevaban los trabajos, escuchándoles en sus sugerencias. En suma, que la cercanía del profesor fuera la nota dominante en esos minutos.

Durante el tiempo que duró el viaje, la asistencia a este vagón era obligatoria para poder garantizar, en primer lugar, que se conociesen y, en segundo lugar, que todos los componentes del equipo participaran en las diversas actividades que debían realizar. Por ello, se pasaba lista y, en cierto modo, se premiaba la fidelidad a las clases en la evaluación.

## 11 El diario de viaje

Señala V. Klenoswski [8] que "el portafolio es el diario de viaje del aprendizaje" y contendrá las pruebas escritas y sus soluciones, los problemas o trabajos voluntarios, los resúmenes de cada tema, las aportaciones al grupo, las consultas en tutorías individuales o grupales, el proceso de búsqueda en Internet o en capítulos de libros y sus consecuencias, las preparaciones de la exposición oral, el tiempo de dedicación, etc.

Digamos que el portafolio ha sido una herramienta fundamental, tanto para el profesor como para el alumno. Al profesor le ha permitido hacer un seguimiento permanente del trabajo realizado por el alumno y, por tanto, evaluarle de forma continua. Y, al alumno, además de conocer de forma casi inmediata cómo se evalúa, dónde y por qué se encontraban los errores, el poder corregir el rumbo, si no es el acertado, y mejorar, si procediera, conforme avanza el curso.

## 12 Escalas

Dependiendo de cómo sea la evaluación que planteamos a los estudiantes, conseguimos unos resultados de aprendizaje y no otros. Así, la evaluación determina el qué y cómo se aprende (M. A. Zabalza [10]).

Durante primer viaje se realizaron seis paradas o escalas, aproximadamente una cada mes y medio, en las que se efectuó una evaluación de lo aprendido durante el trayecto, no sólo entre cada parada sino entre el comienzo del viaje y esa escala. Durante el segundo viaje se redujeron a cuatro escalas.

Difícilmente puede llevarse a cabo una innovación educativa sin modificar o al menos adaptar un modelo de evaluación que valore los resultados del aprendizaje logrado por el estudiante. Por ello, con el fin de implicar al estudiante en su propio proceso de aprendizaje, en la evaluación nos habíamos propuesto incorporar todas las actividades realizadas con unos pesos específicos. Por otra parte, queríamos insistir en que las pruebas fuesen acumulativas en el sentido de que en cada bloque se incorporara los temas anteriores para así lograr una visión de conjunto y, además, lograr averiguar si los fallos de los bloques anteriores se habían solucionado.

La mayoría de las pruebas autónomas fueron tipo test. Los 80 estudiantes y la lectora óptica facilitaban tremendamente la labor. Normalmente al día siguiente de haber realizado la prueba les facilitábamos las soluciones junto con sus respuestas en Internet a través de la plataforma Aulaweb, de tal modo que podían conocer sus fallos de forma inmediata (M.C. Cuevas y A. Nevot [5]).

En dos ocasiones, una en cada cuatrimestre, hicimos sendas pruebas de ensayo en las que figuraban diversos ejercicios y problemas. Además, también planteamos dos pruebas escritas de equipo a lo largo del curso en las que el secretario se encargaba de la redacción final de las soluciones.

En otra ocasión, tuvieron que elaborar de forma original y en equipo un enunciado de un problema y, posteriormente, obtener su solución.

De forma más concreta los criterios y métodos de evaluación que se han tenido en cuenta son los siguientes:

- Pruebas escritas (50 % de la calificación global) se realizaron al finalizar cada uno de los temas bien mediante pruebas objetivas de respuesta múltiple o mediante resolución de cuestiones, ejercicios o problemas, en la mayoría de los casos de forma individual y en alguna otra en grupo.

- Trabajo autónomo (25 % de la calificación global) valorándose aspectos como la entrega de problemas, trabajos voluntarios y participación activa en las clases.

- Trabajo en equipo (25 % de la calificación global), teniendo en cuenta planteamiento y resolución de problemas, exposición oral, búsqueda de información dirigida o abierta en Internet o en determinados libros, así como la fidelidad a la asistencia a las sesiones de equipo.

Las presentaciones orales se realizaron en tres de los seis bloques. Sin lugar a dudas, la dinámica de estas sesiones fue compleja y difícil de encajar y las que más tiempo de consulta por parte de cada uno de los equipos supuso. Sin embargo, también fue una actividad muy gratificante porque tanto el enfoque dado a cada tema que abordaron los equipos, las presentaciones en sí, la gran creatividad desarrollada en algunos equipos y la adaptación a los contenidos nos dejó un huella profunda. Bien es verdad, y no podía ser de otra manera, que en algunos casos las presentaciones dejaban bastante que desear. Pero, no lo es menos, que a lo largo del curso se produjo una superación en muchos de los equipos.

## 13 Cafetería de profesores

Uno de los objetivos que nos habíamos marcado al comenzar esta experiencia piloto era la necesidad de efectuar un seguimiento permanente de la misma, con el fin de ir adaptando todas aquellas cuestiones que no funcionasen como estaba previsto e incorporar aquellas otras que pudieran mejorarla.

Además de las adaptaciones que íbamos haciendo en las reuniones semanales y en ocasiones diarias los dos profesores, consideramos conveniente conocer cómo vivían la experiencia los estudiantes.

El seguimiento individual y de grupo que realizábamos nos permitía tener cierta información sobre algunos aspectos, sobre todo, por parte de aquellos estudiantes más abiertos y espontáneos. Pero no era suficiente. Por ello, después de los tres primeros meses de funcionamiento de los equipos, pasamos un cuestionario acerca de cómo se había desarrollado la elaboración de las actividades. Queríamos averiguar, entre otras cosas, si en algún equipo había estudiantes que no participaban en la elaboración de las prácticas, si alguno estaba sobrecargado de trabajo, cómo se habían repartido el trabajo en la realidad, cuándo, cómo y dónde se habían reunido para elaborarlas. Además, como sabíamos que veinte estudiantes no tenía acceso a Internet desde su casa

(todos los estudiante tienen acceso desde la Escuela), queríamos también conocer si había supuesto un obstáculo para ellos el realizar aquellas búsquedas de información previstas en alguna actividad.

Diez estudiantes, tres de ellos secretarios o coordinadores, manifestaron, o bien tener sobrecarga de trabajo, o bien que alguno del grupo no hacía su trabajo y, por tanto, repercutía en el resto.

Éramos conscientes de que inicialmente, por una solidaridad mal entendida, los alumnos no iban a delatar a los compañeros que no colaboraran haciendo las actividades propuestas. En algunos casos nos equivocamos, y en otros fuimos averiguando y analizando individualmente la calificación que obtenía cada estudiante de forma individual y grupal. En aquellos casos que detectamos unas diferencias significativas hablamos con todos los miembros del grupo y en cierta medida fue corrigiéndose, bien porque modificaron la dinámica de funcionamiento, o bien porque los estudiantes más implicados por su no participación (eran muy pocos) abandonaron la asignatura. No obstante, a mitad de curso y debido a bajas, en algún caso y a incompatibilidades manifiestas entre los miembros de un grupo, de común acuerdo con los estudiantes afectados hicimos diversos reajustes de grupos integrándolos en otros. En el curso 2005-2006 surgió problemas en dos equipos y en el curso 2006-2007 en tres equipos.

Por otra parte, al final de curso mantuvimos dos reuniones o más bien debates informales: una, con todos los coordinadores de equipo, y, otra, con alumnos repetidores. En ellas, se trataba de hacer un balance del curso y conocer en otro ambiente las opiniones de los responsables de grupo sobre la dinámica de funcionamiento de los mismos. Verdaderamente aportaron sugerencias y propuestas de mejora que sin duda, después de analizar su viabilidad, incorporaremos el próximo curso. Por otra aparte, en cuanto a la reunión mantenida con los alumnos repetidores se trataba de conocer las ventajas e inconvenientes que habían encontrado a lo largo del curso con esta nueva metodología en comparación con el curso anterior, puesto que necesitábamos tener también una referencia que los alumnos nuevos no nos podían facilitar.

## 14 Cafetería de alumnos

De diversas reuniones, bien individualmente o bien en grupo con los estudiantes, se pueden destacar las siguientes opiniones:

- Es más asequible aprobar, si bien obtener nota alta es más complicado al intervenir tantas pruebas y actividades.
- El tiempo dedicado a preparar la asignatura en su conjunto ha sido muy superior al dedicado a otras asignaturas del curso.
- Se tiene constancia de haber aprendido mucho más y, además, otras cosas interesantes que permiten ver las matemáticas aplicadas.

- La metodología utilizada obliga a estar estudiando a diario, pero a cambio ves los resultados de forma inmediata y te permite corregir las equivocaciones porque sabes cuáles son.
- Esta forma de trabajo te permite conocer mucho más a tus compañeros de clase.
- Con esta forma de trabajar logras una gran cercanía y confianza con los profesores.
- El trabajo en equipo me ha permitido hacer verdaderos amigos.

## 15 Algunas conclusiones

Después de realizar durante dos cursos esta experiencia en dos grupos numerosos de alumnos, podemos extraer las siguientes conclusiones:

- Los alumnos y los profesores aprenden a trabajar en equipo. Por un lado los alumnos han aprendido que el trabajo del equipo repercute en todo el grupo ya sea de forma positiva o de forma negativa. Los profesores, por otra parte, estando acostumbrados a un trabajo individual, también hemos aprendido a trabajar en equipo.
- Los alumnos han sido conscientes de su propio proceso de aprendizaje. El seguimiento más personal de cada uno de ellos les ha permitido conocer durante todo el proceso las calificaciones en la asignatura y en todos sus componentes.
- La investigaciones, tanto históricas como de ampliación de determinados temas, ha aumentado el interés y la curiosidad de los alumnos por la asignatura.
- Los profesores hemos aprendido muchísimo, por una parte al pensar y buscar actividades que debían realizar los alumnos y, por otra, de los propios trabajos y actividades de los estudiantes.
- Con esta forma de enseñanza el trabajo del profesor se ha cuadriplicado, ya que no sólo debe dar las clases, sino que debe tutelar trabajos, corregir trabajos, tutelar equipos, preparar exámenes, corregir exámenes, evaluar,... Y todo lo anterior por cada unidad temática.

## Referencias

[1] K. Bain. *Lo que hacen los mejores profesores universitarios*. PUV, Valencia, 2005.

[2] L. Bazarra, O. Casanova y J. García. *Ser profesor y dirigir profesores en tiempos de cambio*. Narcea, Madrid, 2004.

[3] A. Benito, A. Cruz. *Nuevas claves para la Docencia Unviersitaria*. Narcea, Madrid, 2005.

[4] M. V. Cuevas y A. Nevot. *Aulaweb como herramienta en la enseñanza de las Matemáitcas*. I Jornadas de Enseñanza y Aprendizaje de la EUATM. UPM, 2005.

[5] M. V. Cuevas y A. Nevot. *Algunos instrumentos para la Enseñanza y el Aprendizaje de las Matemáticas en el EEES*. III Jornadas Internacionales de Innovación Universitaria. UEM, 2006.

[6] K. Exley y R. Dennick. *Enseñanza en Pequeños Grupos en Educación Superior*. Narcea, Madrid, 2007.

[7] J. González, R. Wagenaar (Coordinadores). *Tuning Educational Structures in Europe*. Universidad de Deusto, Bilbao, 2003.

[8] V. Klenowski. *Desarrollos de Portafolios*. Narcea, Madrid, 2005.

[9] I. Pozo. *Aprendices y Maestros*. Alianza Editorial, Madrid, 1999.

[10] M.A. Zabalza. *Competencias docentes del profesorado universitario. Calidad y desarrollo profesional*. Narcea, Madrid, 2003.

| | |
|---|---|
| **Título:** | Sistemas Singulares. Invariantes y Formas Canónicas. |
| **Doctorando:** | Adolfo Díaz Cordero. |
| **Director/es:** | María Isabel García Planas. |
| **Defensa:** | 22 de Septiembre de 2008, UPC. |
| **Calificación:** | Sobresaliente Cum laude. |

**Resumen:**

Los objectos tratados en la tesis son sistemas dinámicos lineales singulares invariantes en el tiempo $E\dot{x} = Ax + Bu, y = Cx$, que representamos mediante cuaternas de matrices $(E, A, B, C) \in M_n(\mathbb{C}) \times M_n(\mathbb{C}) \times M_{n\times m}(\mathbb{C}) \times M_{p\times n}(\mathbb{C})$.

El estudio se centra en la relación de equivalencia, que llamamos semejanza por realimentación proporcional y derivada y inyección externa proporcional y derivada, que es la que admite cambios de base en los espacios de estados, de entradas y de salidas, realimentación de estados tanto proporcional como derivada, inyección externa también proporcional y derivada y además, premultiplicación de la ecuación de estados por una matriz invertible.

Esta relación se puede considerar como la generalización natural de la semejanza para matrices cuadradas y la semejanza por bloques de la cual se obtiene la forma reducida de Kronecker. En todos los estudios realizados hasta ahora, la realimentación derivada así como la inyección externa derivada, no estaban incluidos en la relación, pero si se tiene en cuenta que los conceptos de controlabilidad y observabilidad de un sistema tan importantes en teoría de control, llevan implícito la condición necesaria de que las matrices $E$ y $A$ del sistema son invertibles o se pueden transformar en matrices inertibles mediante realmentaciones proporcionales y/o derivadas e inyecciones externas porporcionales y/o derivadas, nos inducen a introducir estas acciones en la relación de equivalencia. Observemos que en el caso de sistemas estándar las realimentaciones proporcinales y las inyecciones externas han estat incluidas en la relación de equivalencia por muchos autores desde hace tiempo.

Para esta relación de equivalencia, el encontrar una forma reducida canónica es un problema abierto, de la cual en esta tesis se da solución para el caso de sistemas regularizables, es decir para aquellos que o bien son regulares o bien mediante realimentación tanto proporcional como derivada e inyección externa también tanto proporcional como derivada, el sistema se transforma en uno regular. Recordemos que los sistemas regulares son aquellos para los cuales se garantiza la existencia de solución única para cualquier condición inicial consistente.

Para esta relación de equivalencia sobre el conjunto abierto y denso, de los sistemas regularizables, también se halla un conjunto completo de invariantes

que permite decidir, dada una cuaterna cualquiera, a que clase de equivalencia pertenece.

Se aborda también el cálculo de deformaciones versales por la relación d'equivalencia siguiendo las técnicas geométricas introducidas por V. I. Arnold en el caso particular de la variedad diferenciable de las matrices cuadradas en las que actúa el grupo lineal. Una aplicación de la descripción de deformaciones miniversales explícitas es el estudio de perturbaciones locales y la obtención de la dimensión de las distintas órbitas. Se realiza también el análisis de la estabilidad estructural caracterizando las cuaternas estructuralmente estables.

| | |
|---|---|
| **Tipo de evento:** | Congreso |
| **Nombre:** | International Conference on Modeling of Engineering & Technological Problems (ICMETP) and 9th National Conference of Indian Society of Industrial and Applied Mathematics |
| **Lugar:** | BMAS Engineering College, Sharda Group of Institutions, National Highway #2, Keetham, Agra, 282007, UP, India |
| **Fecha:** | January 14–16, 2009 |
| **Organiza:** | |
| **Información:** | |
| **E-mail:** | ajit_46@yahoo.com; kapurami@gmail.com; siddiqi.abulhasan@gmail.com |
| **WWW:** | http://www.siam-india.org/ |

| | |
|---|---|
| **Tipo de evento:** | Workshop |
| **Nombre:** | FEM and BEM for time–dependent wave problems |
| **Lugar:** | Max–Planck–Institute for Mathematics in the Sciences, Leipzig, ALEMANIA |
| **Fecha:** | January 22–24, 2009 |
| **Organiza:** | Wolfgang Hackbusch (MPI Leipzig), Lehel Banjai (University of Zurich/MPI Leipzig), Stefan Sauter (University of Zurich) |
| **Información:** | |
| **E-mail:** | lehelb@math.uzh.ch |
| **WWW:** | http://www.ipam.ucla.edu/programs/le2009/ |

| | |
|---|---|
| **Tipo de evento:** | Cursos y Workshop |
| **Nombre:** | Numerical Approaches to Quantum Many-Body Systems |
| **Lugar:** | Institute for Pure and Applied Mathematics (IPAM), Los Angeles, CA, USA |
| **Fecha:** | January 22–24, 2009: Lectures and Tutorials; January 26-30, 2009: Workshop |
| **Organiza:** | Ulrich Schollwöck (RWTH Aachen), Simon Trebst (Microsoft Research, Station Q), Guifre Vidal (University of Queensland) |
| **Información:** | |
| **E-mail:** | qs2009@ipam.ucla.edu |
| **WWW:** | http://www.ipam.ucla.edu/programs/qs2009/ |

| | |
|---|---|
| **Tipo de evento:** | Coloquios y Escuelas Temáticas |
| **Nombre:** | Session résidentielle au CIRM: Calcul Scientifique et Equations aux Dérivées Partielles |
| **Lugar:** | Marseille, FRANCE |
| **Fecha:** | February 2–6, 2009 |
| **Organiza:** | F. Boyer y F. Hubert |
| **Información:** | |
| **E-mail:** | |
| **WWW:** | http://www.latp.univ-mrs.fr/cirm09/ |

| | |
|---|---|
| **Tipo de evento:** | Workshop |
| **Nombre:** | Laplacian Eigenvalues and Eigenfunctions: Theory, Computation, Application |
| **Lugar:** | Institute for Pure and Applied Mathematics (IPAM), Los Angeles, CA, USA |
| **Fecha:** | February 9–13, 2009 |
| **Organiza:** | |
| **Información:** | |
| **E-mail:** | le2009@ipam.ucla.edu |
| **WWW:** | http://www.ipam.ucla.edu/programs/le2009/ |

| | |
|---|---|
| **Tipo de evento:** | Workshop |
| **Nombre:** | Third School and Workshop on Mathematical Methods in Quantum Mechanics |
| **Lugar:** | Bressanone, ITALIA |
| **Fecha:** | February 16–21, 2009 |
| **Organiza:** | |
| **Información:** | |
| **E-mail:** | info.mmqm@unimore.it |
| **WWW:** | http://www.mmqm.unimore.it/ |

| | |
|---|---|
| **Tipo de evento:** | Workshop |
| **Nombre:** | Many–body systems far from equilibrium: Fluctuations, slow dynamics and long–range interactions |
| **Lugar:** | Max–Planck–Institut for the Physics of Complex Systems, Dresden, ALEMANIA |
| **Fecha:** | February 16–27, 2009 |
| **Organiza:** | |
| **Información:** | |
| **E-mail:** | mbsffe09@mpipks-dresden.mpg.de |
| **WWW:** | http://www.mpipks-dresden.mpg.de/ mbsffe09/ |

| | |
|---|---|
| **Tipo de evento:** | Congreso |
| **Nombre:** | SIAM Conference on Computational Science and Engineering (CSE09) |
| **Lugar:** | Miami, FL, USA |
| **Fecha:** | March 2–6, 2009 |
| **Organiza:** | Society for Industrial and Applied Mathematics |
| **Información:** | |
| **E-mail:** | |
| **WWW:** | http://www.siam.org/meetings/cse09/ |

| | |
|---|---|
| **Tipo de evento:** | Cursos y Workshops |
| **Nombre:** | Quantum and Kinetic Transport: Analysis, Computations, and New Applications |
| **Lugar:** | Institute for Pure and Applied Mathematics (IPAM), Los Angeles, CA, USA |
| **Fecha:** | March 9 – June 12, 2009 |
| **Organiza:** | |
| **Información:** | |
| **E-mail:** | kt2009@ipam.ucla.edu |
| **WWW:** | http://www.ipam.ucla.edu/programs/kt2009/ |

| | |
|---|---|
| **Tipo de evento:** | Congreso |
| **Nombre:** | 17th International Conference on Computing in High Energy and Nuclear Physics |
| **Lugar:** | Praga, REPÚBLICA CHECA |
| **Fecha:** | March 21–27, 2009 |
| **Organiza:** | |
| **Información:** | |
| **E-mail:** | chep2009@particle.cz |
| **WWW:** | http://www.particle.cz/conferences/chep2009/ |

| | |
|---|---|
| **Tipo de evento:** | Congreso |
| **Nombre:** | 33rd SIAM Southeastern-Atlantic Section Conference |
| **Lugar:** | University of South Carolina, Columbia, South Carolina, USA |
| **Fecha:** | April 4-5, 2009 |
| **Organiza:** | |
| **Información:** | |
| **E-mail:** | |
| **WWW:** | http://www.math.sc.edu/ siamseas/ |

| | |
|---|---|
| **Tipo de evento:** | Congreso |
| **Nombre:** | MAFELAP 2009: The Mathematics of Finite Elements and Applications 2009 |
| **Lugar:** | Brunel University, London, UK |
| **Fecha:** | 9–12 June, 2009 |
| **Organiza:** | |
| **Información:** | |
| **E-mail:** | Carolyn.Sellers@brunel.ac.uk |
| **WWW:** | http://people.brunel.ac.uk/~icsrsss/bicom/mafelap2009/ |

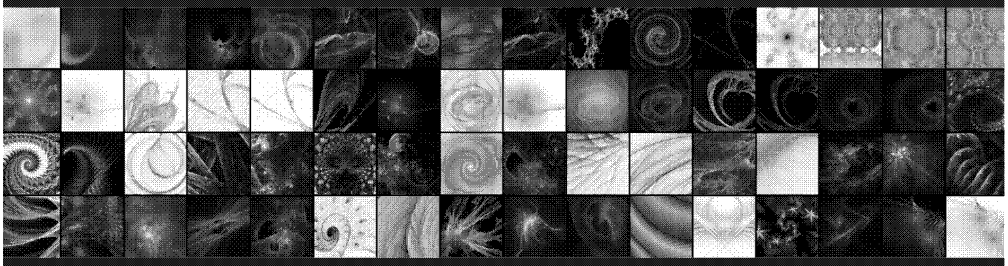
| | |
|---|---|
| **Tipo de evento:** | Congreso |
| **Nombre:** | CEDYA 2009: XXI Congreso de Ecuaciones Diferenciales y Aplicaciones / XI Congreso de Matemática Aplicada |
| **Lugar:** | Universidad de Castilla-La Mancha, Ciudad Real |
| **Fecha:** | 21–25 Septiembre, 2009 |
| **Organiza:** | SëMA, UCLM |
| **Información:** | |
| **E-mail:** | Congreso.CEDYA09.secretaria@uclm.es |
| **WWW:** | http://matematicas.uclm.es/cedya09/ |

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
SēMA
Castilla-La Mancha

**Bravo Trinidad, José Luis**
Profesor Colaborador. *Líneas de investigación:* Problema 16 de Hilbert, ecuación de Abel, problema del centro-foco – Univ. de Extremadura – Fac. de Ciencias – Dpto. de Matemáticas – Avda. de Elvas, s/n. 06071 Badajoz.
*Tlf.:* 924.289.570. *Fax:* 924.272.911.
*e-mail:* trinidad@unex.es.
http://kolmogorov.unex.es/~trinidad

**Calvo Garrido, María del Carmen**
Estudiante. *Líneas de investigación:* Modelos matemáticos en finanzas – Univ. de La Coruña – Fac. de Informática – Dpto. de Matemáticas – Campus de Elviña, s/n. 15071 La Coruña.
*Tlf.:* 981.167.000, Ext. 1301. *Fax:* 981.167.160.
*e-mail:* mcalvog@udc.es.

**Granero Belinchón, Rafael**
Estudiante.

*e-mail:* rafa_g1988@hotmail.com.

**López Pérez, Sergio**
Director. Mathlan Matematika S. A. – Alda. Mazarredo, 47, 6-5. 48009 Bilbao.
*Tlf.:* 944.242.203.
*e-mail:* sergio.lopez@mathlan.es.
http://www.mathlan.es

**Univ. de Bío-Bío**

*Dep. de Ciencias Básicas*
Campus Fernando May. Avda. Andrés Bello, s/n. Casilla 447. Chillán (Chile).
*Tlf.:* +56-42-253050. *Fax:* +56-42-253046.
*e-mail:* ftoledo@roble.fdo-may.ubiobio.cl.

**Univ. de Huelva**

*Dep. de Matemáticas*
Avda. de las Fuerzas Armadas, s/n. 21071 Huelva. *Fax:* 959.219.909.
*e-mail:*

# Direcciones útiles

## Consejo Ejecutivo de SëMA

**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 SëMA

**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 SēMA

**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 SēMA

**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.

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 S̈eMA
   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 SëMA

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 SëMA (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 SëMA (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 SëMA

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