SēMA

BOLETÍN NÚMERO 42

Marzo 2008

sumario

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: Teselaciones de Penrose

Estimados socios,

Os presentamos en esta ocasión un nuevo número de nuestro Boletín muy especial en todos los sentidos. Como habíamos anunciado previamente, se trata de dejar constancia del contenido y el nivel científico del XX Congreso de Ecuaciones Diferenciales y Aplicaciones / X Congreso de Matemática Aplicada, celebrado el pasado mes de septiembre en Sevilla. Consideramos que puede ser una buena ocasión para que muchos de nosotros nos actualicemos en algunos temas, unos más clásicos, otros emergentes. Confiamos que este número especial sea muy leído y acabe siendo de gran utilidad para todos.

Con esa intención han colaborado en su confección varios comités. Recibid un cordial saludo.

Grupo Editor
boletin.sema@uclm.es

Sesiones plenarias

Russel E. Caflish

Martin J. Gander

Jaume Llibre

Julio D. Rossi

Bol. Soc. Esp. Mat. Apl.
n° 42(2008), 9–19

HYBRID MONTE CARLO METHODS FOR FLUID AND PLASMA DYNAMICS

RUSSEL CAFLISCH

Mathematics Department, University of California at Los Angeles, Los Angeles, CA 90095 USA

Abstract

For small Knudsen number, simulation of rarefied gas dynamics by the Direct Simulation Monte Carlo (DSMC) method becomes computationally intractable because of the large collision rate. To overcome this problem we have developed a hybrid simulation method, combining DSMC and a fluid dynamic description into a single method. The molecular distribution function f is represented as a linear combination of a Maxwellian distribution M and a particle distribution g; i.e., $f = \beta M + (1 - \beta)g$. The density, velocity and temperature of M are governed by fluid-like equations, while the particle distribution g is simulated by DSMC. In addition there are interaction terms between M and g. The coefficient β is determined automatically, by a thermalization approximation. Numerical results will be presented to demonstrate the validity of this method, as well as the acceleration that it provides over DSMC. This method has been extended to simulation of Coulomb collisions in a plasma. For this extension, the underlying Monte Carlo method is Nanbu's method for Coulomb collisions.

1 Introduction

Since the early 1970's the dominant method for computation of rarefied gas dynamics (RGD) has been the Direct Simulation Monte Carlo (DSMC) method pioneered by Graeme Bird [1], which moves particles according to their velocities and performs collisions between randomly chosen particles. This method has been tremendously successful in a wide range of applications. There is an important flow regime, however, in which the DSMC method loses its effectiveness: flow for which the Knudsen number ε is small enough that the collision rate is large, but not small enough that the flow is well described by fluid mechanics. In this *near-continuum* regime, the appropriate length and time scales are nearly those for fluid mechanics, but the collisional length and time scales are quite small. Since accuracy of DSMC depends on resolution of the collisional length and time scales, it becomes slow and inaccurate in this regime.

This presentation is for a simulation method for RGD that is formulated to overcome this difficulty by combining DSMC with a computational fluid dynamics (CFD) solver. In this method the velocity distribution function f is written as a combination of a Maxwellian distribution M and a particle distribution g as, $f = \beta M + (1-\beta)g$, in which β is a parameter representing the degree of local thermalization. In this Interpolated Fluid/Monte Carlo (IFMC) method, evolution of the fluid component M and the particle component g is governed by CFD and DSMC, respectively. In addition there are interaction terms between the fluid and particle components. The parameter β is determined from the local velocity distribution and Knudsen number ε. As $\varepsilon \to 0$, the computation becomes purely CFD; while for moderate values of ε, it becomes purely DSMC. In the near continuum regime, this method provides considerable acceleration (or equivalently, reduced statistical error) over DSMC, while maintaining accuracy.

We have generalized this method to apply to Coulomb collisions in a plasma. Simulation of Coulomb collisions can be a computational bottleneck, since the collision times are often very disparate from the characteristic times of interest. This difficulty is compounded by the wide range of collision rates for many problems. For example, consider a velocity distribution in the form of a bump-on-tail; i.e., a near-equilibrium distribution at low velocity with an isolated spike far out on its tail (the "bump"). The rate of collisions between two particles of velocity $\mathbf{v}_1$ and $\mathbf{v}_2$ is proportional to u^{-3} for $u = |\mathbf{v}_1 - \mathbf{v}_2|$. The average rate of collisions between the particles in the central distribution $f \approx M$ is of size $T_M^{-3/2}$ in which T_M is the temperature of the Maxwellian distribution M. The bump may be concentrated at a velocity difference u_B from the center of M with $u_B >> T_M^{1/2}$, so that its rate of interaction with M is of size $u_B^{-3} << T_M^{-3/2}$. Direct simulation of the Coulomb collisions for a bump-on-tail distribution is dominated by collisions between M and itself, which preserve M but do not affect the evolution of f, and the important interactions of the bump with M will be rare events. This shows that direct simulation of this problem is highly inefficient.

We present a hybrid method for accelerating the simulation of Coulomb collisions. It represents the distribution function as a combination of a thermal component m (a Maxwellian distribution) and kinetic component k (numerically represented as a set of particles). Evolution of the thermal component m is performed using continuum methods based on conservation principles; while evolution of the kinetic component k is performed by Monte Carlo simulation of binary collisions. An interaction between m and k is performed by sampling a particle from m and selecting a particle from k, then treating the interaction as a particle collision. In addition, thermalization (particles moved from k to m) and dethermalization (particles moved from m to k) are performed with probabilities p_T and p_D respectively.

2 The Boltzmann Equation and the IFMC Method

For the simple case of a monatomic gas, the Boltzmann equation is

$$\partial_t f + \mathbf{v} \cdot \nabla f = \frac{1}{\varepsilon} Q(f, f) \tag{1}$$

in which $f = f(\mathbf{x}, \mathbf{v}, t)$ is the molecular density function for particles with position $\mathbf{x}$ and velocity $\mathbf{v}$ at time t. The intermolecular collision process is described by the bilinear operator Q. The Knudsen number ε is the ratio of the mean free time (i.e., the average time between collisions) to the macroscopic time scale of interest. For large ε, collisions are infrequent and the gas particles mostly perform free-streaming. For small ε, the collision rate is large, and the distribution function f is rapidly driven towards a Maxwellian equilibrium given by

$$M(\mathbf{v}) = \rho(2\pi T)^{-3/2} \exp\left(-|\mathbf{v} - \mathbf{u}|^2/2T\right) \tag{2}$$

in which ρ, $\mathbf{u}$ and T are the macroscopic density, velocity and temperature. In this continuum regime, the spatial fluctuations in ρ, $\mathbf{u}$ and T are described by the fluid equations.

The IFMC method uses a mixed representation of f as a combination of a Maxwellian M and a non-Maxwellian g; i.e.,

$$f = (1 - \beta)g + \beta M \tag{3}$$

in which β is a parameter that describes the degree of equilibration of f.

The acceleration of the IFMC method relies on several steps: an implicit time step, a thermalization approximation, a mixed particle/analytic representation, an advection step and a Monte Carlo collision step.

3 IFMC for Spatially Homogeneous Problems

Consider the spatially homogeneous Boltzmann equation

$$\partial_t f = \frac{1}{\varepsilon} Q(f, f) \tag{4}$$

which describes the collision step within a single cell for DSMC. In the Boltzmann equation (4), the collision operator $Q(f, f)$ can be written as a sum of positive and negative parts, i.e.

$$Q(f, f) = Q^+(f, f) - Q^-(f)f. \tag{5}$$

Let μ be a constant with

$$\mu > Q^-(f) \tag{6}$$

and rewrite Q as

$$Q(f, f) = P(f, f) - \mu f. \tag{7}$$

Now set

$$\begin{aligned} \tau &= (1 - e^{-\mu t/\varepsilon}) \\ F &= f e^{\mu t/\varepsilon}. \end{aligned} \tag{8}$$

Then (4) can be written as

$$\frac{\partial F}{\partial \tau} = \frac{1}{\mu} P(F, F). \tag{9}$$

This formulation involves "implicit" time evolution, in the sense that the small parameter ε has been removed from (9) and is only in the mapping from τ to t. It was first proposed by Gabetta et al. [6].

Next is the thermalization approximation. The solution of (9) can be written in a Wild expansion [10] as

$$F(\mathbf{v}, \tau) = \sum_{k=0}^{\infty} \tau^k f_k(\mathbf{v}) \tag{10}$$

where the functions f_k are given by the recurrence formula

$$f_{k+1}(\mathbf{v}) = \frac{1}{k+1} \sum_{h=0}^{k} \frac{1}{\mu} P(f_h, f_{k-h}), \quad k = 0, 1, \ldots, \qquad f_{k=0}(\mathbf{v}) = f(\mathbf{v}, t = 0). \tag{11}$$

The term f_k represents the contribution from particles that suffer exactly k collisions in the time period t. We now make an approximation that a particle having K or more collisions in a short period t becomes thermalized; i.e., it becomes part of the Maxwellian distribution M, which is chosen to have the same velocity and temperature as the original f. This approximation amounts to replacing f_k by M for $k \geq K$. Taking $K = 2$, replacing t by the time step Δt and reinserting f leads to the following approximation for f:

$$f(\Delta t) = A f(0) + B f_1 + C M \tag{12}$$

in which

$$\begin{aligned} A &= (1 - \tau) \\ B &= \tau(1 - \tau) \\ C &= \tau^2 \\ f_1 &= \frac{1}{\mu} P(f_0, f_0). \end{aligned} \tag{13}$$

This is the thermalization approximation that is applied for each small time step Δt.

Next we use the mixed representation (3) of f as a combination of a Maxwellian M and a non-Maxwellian g. The collisions in f_1 are now performed as follows: Collisions between M and M lead to M, so that they do not need to be performed. Collisions between g and g are performed as in DSMC. Collisions between M and g are performed by first sampling a particle from M, then using DSMC.

4 IFMC for Spatially Inhomogeneous Problems: Convection Step

Here we describe a convection method for the Maxwellian component of f that is based on motion of the particles in a Maxwellian distribution.

For a spatially inhomogeneous problem, the spatial domain is divided into cells of size dx, and the time into discrete time steps of size dt, as in the DSMC method. In each time step, a splitting method is used, so that each time step is described by a collisional step and a convective step. The spatially homogeneous method, described in Section 3, is used in each collisional step. At the beginning of a convection step, the density f consists of a Maxwellian and a collection of particles, as in (3). The particles advect for the time step dt by their velocity, as in DSMC; i.e., $x_k(t + dt) = x_k(t) + dt\ v_k$.

Convection of the Maxwellian component is more complicated, but can be performed by three different methods. In the first method, all of the particles in M are sampled and then moved by their velocity. In the second method, the Maxwellian distribution is decomposed into pieces that move from one spatial cell to another. This is equivalent to a Boltzmann scheme for solving the fluid equations. In the third method, a numerical method for the fluid equations is used directly.

5 Computational Results for RGD

In this section, we present results from simulations for shock waves and flow past a leading edge. For these applications, we present simulations using DSMC and IFMC. For the IFMC shock simulations, we use the modification of Bird's shock boundary conditions.

5.1 Shocks

We performed comparisons of shock wave simulations from the IFMC method and Bird's DSMC code for shock waves over a range of Mach numbers and Knudsen numbers. The results presented here are for a moderate strength shock with Mach number of 1.4 in Figure 1. The IFMC shock results presented here are somewhat sensitive to the choice of time step dt and it has been picked for optimal results.

5.2 Flow Past a Leading Edge

In these simulations the incoming flow at the top has density $\rho = 4.65\ 10^{-6}$, vertical velocity $v = -1412.5$ and temperature $T = 300.0$. The leading edge is a semi-infinite line starting at $y = 9$ (taking the bottom to be at $y = 0$) and extending downwards. On the edge the particles have thermal reflection at temperature $T = 500.0$. All of these parameters are in SI units. The boundary conditions at the bottom are vacuum type and those on the wall opposite the leading edge are the same as at the incoming boundary. Figure 2 shows good results from a comparison of DSMC and IFMC for this problem.

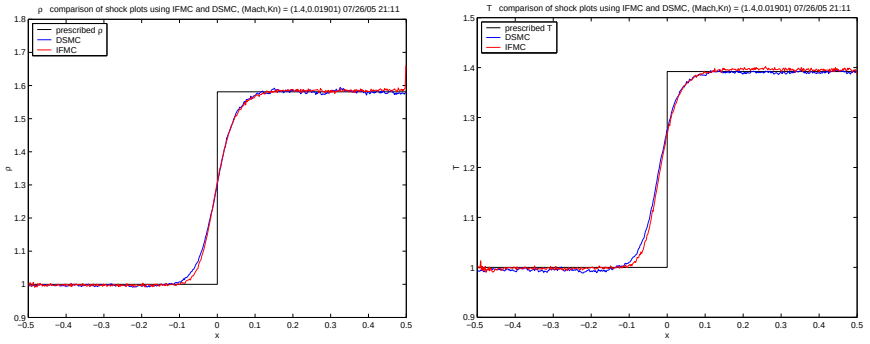


Figure 1: Comparison of simulation results for ρ (left) and T (right) from IFMC (red) and DSMC (blue) for a shock with Mach number 1.4 and Knudsen number 0.019.

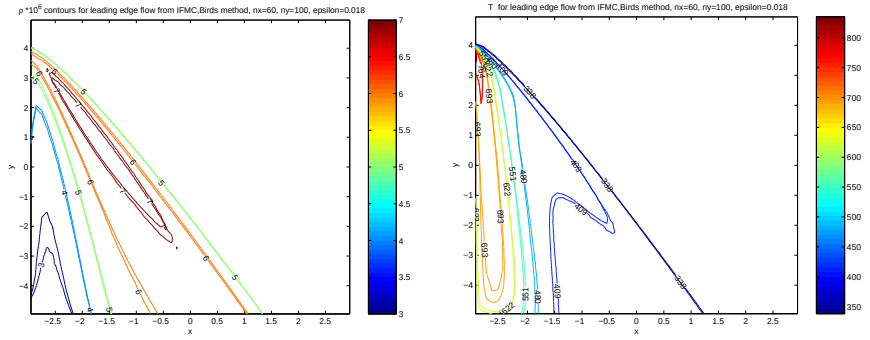


Figure 2: Comparison of simulation results for ρ (left) and T (right) from IFMC and DSMC for flow past a leading edge.

6 Monte Carlo Simulation of Coulomb Collisions

We first introduce the governing equation for the physical process, and describe the TA and Nanbu Monte Carlo binary collision models for a spatially homogeneous plasma. We consider collisions between N particles consisting of $N/2$ particles from each of two species α and β.

6.1 Governing equation

Coulomb collisions in a plasma can be treated as the simulation of many continuous small-angle binary collisions [7]. The time evolution of the particle distribution in a spatially homogeneous, non-equilibrium plasma is described by the Fokker-Planck equation:

$$\frac{\partial f_\alpha}{\partial t} = (\frac{\delta f_\alpha}{\delta t})_c \tag{14}$$

in which f_α is the distribution function of the α species and $(\frac{\delta f}{\delta t})_c$ is the collision operator defined as (MKS units)

$$(\frac{\delta f_\alpha}{\delta t})_c = -\sum_\beta \frac{\partial}{\partial v_j} \frac{e_\alpha^2 e_\beta^2 \log \Lambda}{8\pi\epsilon_0^2 m_\alpha} \int dv' [\frac{\delta_{jk}}{u} - \frac{u_j u_k}{u^3}][\frac{f_\alpha}{m_\beta} \frac{\partial f_\beta'}{\partial v_k'} - \frac{f_\beta'}{m_\alpha} \frac{\partial f_\alpha}{\partial v_k}]. \quad (15)$$

in which we use the notation $\mathbf{u} = \mathbf{v}_\alpha - \mathbf{v}_\beta$, $u = |\mathbf{u}|$ and $f_\beta' = f_\beta(\mathbf{v}')$. The equation for f_β is similar. Bobylev and Nanbu [2] derived a general formulation for a binary collision model that approximates the solution of (14) over a time step Δt. A detailed comparison of the methods of TA and Nanbu is presented in [9].

7 The Hybrid Method

The hybrid method is based on representation of the velocity distribution function f as a combination of a thermal component m and a kinetic component k; i.e.,

$$f(\mathbf{v}) = m(\mathbf{v}) + k(\mathbf{v}). \quad (16)$$

The thermal component is a Maxwellian distribution

$$m(\mathbf{v}) = n_m (2\pi T_m)^{-3/2} \exp(-|\mathbf{v} - \mathbf{u}_m|^2 / 2T_m). \quad (17)$$

Because of the (expected) slow interaction of the thermal component m with the kinetic component k, the average density, velocity and temperature n_m, $\mathbf{u}_m$ and T_m of m are not assumed to be those of the full distribution f. This explains the difference between the notation m and M, since M is assumed to density, velocity and temperature that are equal to those of f.

Collision are performed as in the hybrid method for RGD: The $m - m$ collisions do not change the distribution m, so they do not need to be performed. The $k - k$ collision are performed as in the method of TA or Nanbu. The $m - k$ collisions are performed by sampling a particle from m then using the method of TA or Nanbu.

After the collisions, thermalization and dethermalization are applied to all of the particles that collided in $k-k$ or $m-k$ collisions. Particles of velocity v that started in k are thermalized with probability $p_T(\mathbf{v})$. This is done by removing $\mathbf{v}$ from k and adding its number, momentum and energy to m. Particles of velocity $\mathbf{v}$ that started in M are dethermalized with probability $p_D(\mathbf{v})$. This is done by adding $\mathbf{v}$ to k and subtracting its number, momentum and energy from m.

A possible problem with this algorithm is that sampling velocities from m may remove too much energy from m. This can be avoided by conservative sampling. First sample all of the required velocities from m and then shift and scale these so that the average momentum and energy of the sampled particles is the same as the average momentum and energy of m.

A choice of thermalization and dethermalization proabilities p_T and p_D is described in [4]. It is based on a detailed balance condition, but is otherwise somewhat ad hoc.

8 Computational Results for Coulomb Collisions

8.1 Bump-on-Tail and Maxwellian Initial Data

As a test of the hybrid method, we performed a series of computations for initial data that is a bump-on-tail. As discussed in Section 1, this problem involves two widely separated time scales for Coulomb interactions, so that it is well suited for the hybrid method: a fast time scale for collisions between particles within the central Maxwellian and a slower time scale for those between particles from the central Maxwellian and the bump. We also performed computation for initial data that is Maxwellian, in order to test the consistency of the hybrid method.

The bump-on-tail initial distribution $f_0(\mathbf{v})$ is specified to be a combination of a Maxwellian $M_0(\mathbf{v})$ and a bump $g_0(\mathbf{v})$. The bump is specified to be approximately a δ-function containing 10% of the mass of the distribution and centered at $\mathbf{v} = (v_b, 0, 0)$ with $v_b = a\sqrt{T_e/m_e}$. The Maxwellian M_0 is centered and scaled so that the average velocity is 0 and the temperature is T_e. The examples presented here uses $a = 4$ and is referred to as BOT4.

The computation is performed in a dimensionless formulation in which the electron mass is $m_e = 1$, and the electron density n_e and temperature T_e were chosen to be $n_e = 0.1$ and $T_e = 0.05065776$. For a characteristic time for the collision process, we use

$$\begin{aligned} t_c &= u_{th}^3 \left(\frac{q_e^2}{\epsilon_0 m_e/2} \right)^{-2} \left(\frac{n_e \log \Lambda_e/2}{4\pi} \right)^{-2} \\ u_{th} &= \sqrt{6T_e/m_e} \end{aligned} \tag{18}$$

which has value $t_c = 5.348275$. Unless otherwise state, the number of particles is $N = 128,000$.

Note that in all the simulation examples reported here, the plasma is spatially homogeneous so that there are no electromagnetic fields and no convection.

8.2 Simulation for the Evolution of a Bump-on-Tail

Figure 3 shows a comparison of the solutions computed by the hybrid (blue dashed line) and Nanbu (red solid line) methods for bump-on-tail problem BOT4, at various times between the initial time and a final time $T = 7.2t_c$. The time step is $\Delta t = t_c/10$. The thermal component of the hybrid representation (16) (green dotted line) is also plotted. The figure shows very agreement between the hybrid and Nanbu curves, providing a measure of validation for the hybrid method. In Figure 3 the thermal component of the hybrid representation (16), which contains about 1/3 of the particles.

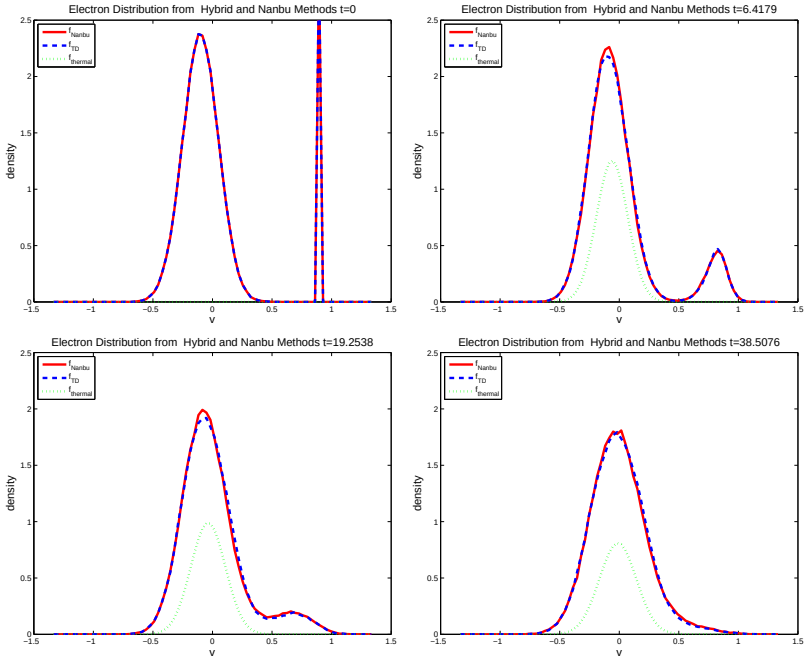


Figure 3: Comparison of the hybrid (blue dashed line) and Nanbu (red solid line) solutions at different times $t = 0$ (upper left), $t = 1.2t_c$ (upper right), $t = 3.6t_c$ (lower left) and $t = 7.2t_c$ (lower right). The computations use $\Delta t = t_c/10$ for the problem BOT4.

9 Conclusions

The IFMC method presented here combines DSMC and CFD in a hybrid simulation method for RGD. This method provides significant reduction in statistical error and computational time, as compared to DSMC, for flow in the near continuum regime.

In a similar way, the hybrid method for Coulomb collisions combines continuum and particle descriptions for the evolution of a velocity distribution function. The method includes particle interactions, but since the examples here are spatially homogeneous, the continuum description is just an equilibrium Maxwelllian distribution. Because of the variation of the interaction rate as a function of particle velocity, the division of f between particles and continuum must be performed as a function of velocity. In the hybrid method of this paper, the velocity dependence is effected through velocity dependence of the thermalization and dethermalization probabilities.

10 Acknowledgments

Work partly performed under the auspices of the U.S. Department of Energy by the University of California, Los Angeles, under grant DE-FG02-05ER25710. The work was supported by the Office of Advanced Scientific Computing Research, DOE Office of Science, under the Multiscale Initiative program.

References

[1] G.A. Bird. *Molecular Gas Dynamics*. Oxford University Press, London, 1976.

[2] A. V. Bobylev and K. Nanbu, Physical Review E, Vol 61, No. 4, p. 4576-4582 (2000).

[3] R. E. Caflisch and L. Pareschi. J. Compt. Phys. 154, 96 (1999).

[4] R.E. Caflisch, C. Wang, G. Dimarco, B. Cohen and A. Dimits, preprint (2007).

[5] K. Nanbu, Phys. Rev. E. 55 (1997).

[6] E. Gabetta, L. Pareschi, and G. Toscani. Relaxation schemes for nonlinear kinetic equations. *SIAM Journal on Numerical Analysis*, 34:2168–2194, 1997.

[7] L. Spitzer, Jr., Physics of Fully Ionized Gases, 2nd ed. (Interscience, New York, 1967).

[8] T. Takizuka and H. Abe, J. Comp. Phys. 25 (1977).

[9] C. Wang, T. Lin, R.E. Caflisch, B. Cohen and A. Dimits, "Particle Simulation of Coulomb Collisions: Comparing the methods of Takizuka & Abe and Nanbu" (2006) under review.

[10] E. Wild. On Boltzmann's equation in the kinetic theory of gases. *Proc. Camb. Phil. Soc.*, 47:602–609, 1951.

Bol. Soc. Esp. Mat. Apl.
nº 42(2008), 21–35

ANALYSIS OF THE PARAREAL ALGORITHM APPLIED TO HYPERBOLIC PROBLEMS USING CHARACTERISTICS

MARTIN J. GANDER

Section de Mathématiques, Université de Genève.

martin.gander@math.unige.ch

Abstract

The parareal algorithm is a time domain decomposition algorithm for the time parallel approximation of solutions of evolution problems. It can be interpreted as a multiple shooting method for initial value problems with a particular choice of the approximate Jacobian on a coarse grid. The method can give significant speedup for non-linear systems of ordinary differential equations and discretized diffusive partial differential equations, but has been reported to be less effective for hyperbolic problems. We prove in this paper a convergence result for the advection equation using the technique of characteristics. Our analysis also reveals limitations of the method when applied to the second order wave equation.

Key words: *Time parallel time integration methods, multiple shooting for initial value problems, parareal algorithm, hyperbolic problems, advection equation, second order wave equation.*

AMS subject classifications: *65R20, 45L05, 65L20*

1 Introduction

Time domain decomposition methods have a long history: already Nievergelt proposed in [19] a parallel algorithm based on a decomposition of the time direction for the solution of ordinary differential equations. While his idea targeted large scale parallelism, Miranker and Liniger proposed a little later in [18] a family of naturally parallel Runge Kutta methods for small scale time parallelism. Waveform relaxation methods, introduced by Lelarasmee, Ruehli and Sangiovanni-Vincentelli in [14] for the large scale simulation in VLSI design, are another fundamental way to introduce time parallelism into the solution of evolution problems. A more recent time parallel algorithm which we will study in this paper is the parareal algorithm, see Lions, Maday and Turinici [15]. For an up to date historical review and further references, see [11].

The parareal algorithm is a time domain decomposition method, based on multiple shooting with an approximate Jacobian on a coarse grid. A detailed derivation of the algorithm and relations to other algorithms can be found in [11]. This reference also contains sharp convergence estimates for linear

problems, including discretizations of the heat equation and the advection equation on unbounded domains, which show that while the method works well in the diffusive case, it is not effective in the advective case on unbounded domains. For nonlinear problems, the parareal algorithm has been analyzed in [9], and a sequence of numerical experiments showed that substantial speedup can be obtained for a problem from chemical reactions, the computation of satellite orbits, and for the Lorentz equations, which are a simplistic model for weather prediction, one of the key applications for time parallel algorithms, since computations need to be performed in real time, and thus any speedup is welcome, even if it is suboptimal, which is often the case for time parallel algorithms. More substantial numerical experiments can be found for fluid and structure problems in [6], for the Navier-Stokes equations in [8], and for reservoir simulation in [12]. Several variants of the method have been proposed, see for example [6, 13]. The algorithm has been further analyzed in [16, 17], and its stability is investigated in [3, 20].

For hyperbolic problems, the parareal algorithm can have performance problems, as it was pointed out in [6] and [11]. An interesting modification of the parareal algorithm was then proposed in [5] and [7], and further analyzed in the context of shooting methods in [10]. Numerical experiments in [11] however had shown that for advection equations on bounded domains, the parareal algorithm can be effective, and approximately linear convergence was observed, which could not be explained by the Fourier analysis used in [11]. We prove in this paper a convergence result for the advection equation on bounded domains. Our result is based on the technique of characteristics, which is the main novelty in the analysis of the parareal algorithm, and permits a generalization to non-constant coefficient and non-linear problems. We then show, using again the method of characteristics, the limitations of the parareal algorithm applied to the second order wave equation. We illustrate our results by numerical experiments.

2 Derivation of the Parareal Algorithm

The parareal algorithm is a time parallel algorithm for the solution of the general nonlinear system of ordinary differential equations

$$u'(t) = f(u(t)), \quad t \in (0,T), \quad u(0) = u^0, \tag{1}$$

where $f : \mathbb{R}^M \longrightarrow \mathbb{R}^M$ and $u : \mathbb{R} \longrightarrow \mathbb{R}^M$. To obtain a time parallel algorithm for (1), we follow the derivation in [9]: we decompose the time domain $\Omega = (0,T)$ into N time subdomains $\Omega_n = (T_n, T_{n+1})$, $n = 0, 1, \ldots N-1$, with $0 = T_0 < T_1 < \ldots < T_{N-1} < T_N = T$, and $\Delta T_n := T_{n+1} - T_n$, and consider on each time subdomain the evolution problem

$$u_n'(t) = f(u_n(t)),\ t \in (T_n, T_{n+1}),\ u_n(T_n) = U_n,\ n = 0, 1, \ldots, N-1, \tag{2}$$

where the initial values U_n need to be determined such that the solutions on the time subdomains Ω_n coincide with the restriction of the solution of (1) to

Ω_n, i.e. the U_n need to satisfy the system of equations

$$U_0 = u^0, \quad U_n = \varphi_{\Delta T_{n-1}}(U_{n-1}), \quad n = 1, \ldots, N-1, \tag{3}$$

where $\varphi_{\Delta T_n}(U)$ denotes the solution of (1) with initial condition U after time ΔT_n. This time decomposition method is nothing else than a multiple shooting method for (1), see [4]. Letting $U = (U_0^T, \ldots U_{N-1}^T)^T$, the system (3) can be written in the form

$$F(U) = \begin{pmatrix} U_0 - u^0 \\ U_1 - \varphi_{\Delta T_0}(U_0) \\ \vdots \\ U_{N-1} - \varphi_{\Delta T_{N-2}}(U_{N-2}) \end{pmatrix} = 0, \tag{4}$$

where $F : \mathbb{R}^{M \cdot N} \longrightarrow \mathbb{R}^{M \cdot N}$. System (4) defines the unknown initial values U_n for each time subdomain, and needs to be solved, in general, by an iterative method. For a direct method in the case where (1) is linear and the system (4) can be formed explicitly, see [1].

Applying Newtons method to (4) leads after a short calculation to

$$\begin{aligned} U_0^{k+1} &= u^0, \\ U_n^{k+1} &= \varphi_{\Delta T_{n-1}}(U_{n-1}^k) + \varphi'_{\Delta T_{n-1}}(U_{n-1}^k)(U_{n-1}^{k+1} - U_{n-1}^k), \end{aligned} \tag{5}$$

where $n = 1, \ldots, N-1$. In [4], it was shown that the method (5) converges quadratically, once the approximations are close enough to the solution. However in general, it is too expensive to compute the Jacobian terms in (5) exactly. An interesting recent approximation is the parareal algorithm, which uses two approximations with different accuracy: let $F(T_n, T_{n-1}, U_{n-1})$ be an accurate approximation to the solution $\varphi_{\Delta T_{n-1}}(U_{n-1})$ on time subdomain Ω_{n-1}, and let $G(T_n, T_{n-1}, U_{n-1})$ be a less accurate approximation, for example on a coarser grid, or a lower order method, or even an approximation using a simpler model than (1). Then, approximating the time subdomain solves in (5) by $\varphi_{\Delta T_{n-1}}(U_{n-1}^k) \approx F(T_n, T_{n-1}, U_{n-1}^k)$, and the Jacobian term by

$$\varphi'_{\Delta T_{n-1}}(U_{n-1}^k)(U_{n-1}^{k+1} - U_{n-1}^k) \approx G(T_n, T_{n-1}, U_{n-1}^{k+1}) - G(T_n, T_{n-1}, U_{n-1}^k),$$

we obtain as approximation to (5)

$$\begin{aligned} U_0^{k+1} &= u^0, \\ U_n^{k+1} &= F(T_n, T_{n-1}, U_{n-1}^k) + G(T_n, T_{n-1}, U_{n-1}^{k+1}) - G(T_n, T_{n-1}, U_{n-1}^k), \end{aligned} \tag{6}$$

which is the parareal algorithm, see [15] for a linear model problem, and [2] for the formulation (6). A natural initial guess is the coarse solution, i.e. $U_n^0 = G(T_n, T_{n-1}, U_{n-1}^0)$ for $n = 1, 2, \ldots, N$.

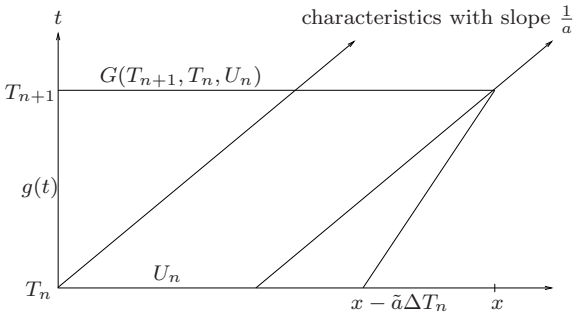


Figure 1: Propagation of the solution along the characteristics, and Assumption on the approximate solver G.

3 The Parareal Algorithm for the Advection Equation

We now study the convergence behavior of the parareal algorithm applied to the one dimensional linear advection equation on the domain $\Omega := (0, L)$, $L > 0$,

$$\begin{array}{rcll} u_t + a u_x & = & f & \text{in } \Omega \times (0,T), \\ u(x,0) & = & u^0(x) & \text{in } \Omega, \\ u(0,t) & = & g(t) & t \in (0,T), \end{array} \tag{7}$$

where we assume that $a > 0$, so that a boundary condition on the left needs to be imposed. We also assume for simplicity that a is constant; we will indicate at the end how the analysis can be generalized to the case of variable advection speed a.

Applying the parareal algorithm to (7), we obtain the same iteration (6), where now however the iterates are functions, $U_n^k : \Omega \longrightarrow \mathbb{R}$. Numerical experiments in [11] showed that the parareal algorithm converges approximately linearly for this problem, which could not be explained by the Fourier analysis in [11]. We assume in what follows that the fine approximation $F(T_n, T_{n-1}, U_{n-1})$ is the exact solution of (7) at time T_n with initial condition U_{n-1} at time T_{n-1}, and that $G(T_n, T_{n-1}, U_{n-1})$ is an approximate solution of (7) at time T_n with initial condition U_{n-1} at time T_{n-1}. In addition, G needs to satisfy the assumption

Assumption 1 *There exists a positive constant $\tilde{a}$, $0 < \tilde{a} \leq a$, such that $G(T_{n+1}, T_n, U_n)$ at x only depends on the boundary condition g and on $U_n(x - \tilde{a}\Delta T_n)$ for $x - \tilde{a}\Delta T_n \geq 0$.*

Remark 1 *Assumption 1 is natural for the advection equation, since the exact solution follows the characteristics, as illustrated in Figure 1, and for convergence, the CFL condition of the scheme requires that $\tilde{a} \leq a$.*

We need two lemmas to prove a convergence result of the parareal algorithm (6) applied to problem (7). The first Lemma holds in general for the parareal algorithm (6) applied to any problem.

Lemma 1 *If F is exact, then at iteration k of the parareal algorithm (6), U_n^k is the exact solution for $n \leq k$.*

Proof. The proof is by induction, in both k and n: for $k = 0$, we have $U_0^0 = u^0$ which is the initial condition and hence is exact. So assume that the result holds for k, i.e. U_n^k is exact for $n \leq k$. Then at iteration $k+1$, we still have for $n = 0$ that $U_0^{k+1} = u^0$, and we can now use induction on n: assuming that U_n^{k+1} is exact, algorithm (6) gives, since $U_n^{k+1} = U_n^k$,

$$\begin{aligned} U_{n+1}^{k+1} &= F(T_{n+1}, T_n, U_n^k) + G(T_{n+1}, T_n, U_n^{k+1}) - G(T_{n+1}, T_n, U_n^k) \\ &= F(T_{n+1}, T_n, U_n^k), \end{aligned}$$

which is exact by the assumption on F, and thus concludes the proof. □

The next lemma shows a similar property going out from the left boundary, if the parareal algorithm (6) is applied to the advection equation (7).

Lemma 2 *Let F be exact and G satisfy Assumption 1, when the parareal algorithm (6) is applied to the advection equation (7). If $U_n^k(x)$ at iteration k satisfies $U_n^k(x) = u(x, T_n)$, for $x \in [0, \alpha]$ for some $\alpha \geq 0$ and for all $n = 0, 1, \ldots, N$, then $U_n^{k+1}(x) = u(x, T_{n+1})$ for $x \in [0, \alpha + \tilde{a}\Delta T]$ and all n, where $\Delta T = \min_{n \in \{0,1,\ldots,N-1\}} \Delta T_n$.*

Proof. The proof is by induction on n. For $n = 0$, we have by definition in (6) that $U_0^{k+1} = u(x, 0)$ for all x and all k. So now we assume that for a given n, $U_n^{k+1}(x) = u(x, T_n)$ for $x \in [0, \alpha + \tilde{a}\Delta T]$. By assumption of the Lemma, we also have that $U_n^k(x) = u(x, T_n)$ for $x \in [0, \alpha]$, which implies that

$$U_n^{k+1}(x) = U_n^k(x) = u(x, T_n), \quad \text{for } x \in [0, \alpha].$$

Using now Assumption 1, the difference $G(T_{n+1}, T_n, U_n^{k+1}) - G(T_{n+1}, T_n, U_n^k)$ in algorithm (6) vanishes on the interval $[0, \alpha + \tilde{a}\Delta T_n]$, and thus the algorithm (6) gives on this interval $U_{n+1}^{k+1} = F(T_{n+1}, T_n, U_n^k)$, which implies that $U_{n+1}^{k+1} = u(x, T_{n+1})$ for $x \in [0, \alpha + \tilde{a}\Delta T_n]$, since F is the exact solution, $U_n^k(x) = u(x, T_n)$ for $x \in [0, \alpha]$ and $a \geq \tilde{a}$. Using now that $\Delta T = \min_{n \in \{0,1,\ldots,N-1\}} \Delta T_n$ completes the proof by induction. □

We are now ready to prove a convergence estimate for the parareal algorithm (6) applied to the advection equation (7). Figure 2 is illustrating the argument graphically, where we assumed that the coarse grid is equi-spaced, i.e. $T_n - T_{n-1} = \Delta T$ for all $n = 1, 2, \ldots, N$.

Theorem 1 *Let U_n^k be the approximations computed by the parareal algorithm (6) applied to the advection equation (7), where F is exact and G satisfies Assumption 1, and let $\Delta T = \min_{n \in \{0,1,\ldots,N-1\}} \Delta T_n$. Then we have the convergence estimate*

$$\sum_{n=0}^{N} ||u(\cdot, T_n) - U_n^k||_1 \leq C \max(L - k\tilde{a}\Delta T, 0) \times \max(N - k, 0), \tag{8}$$

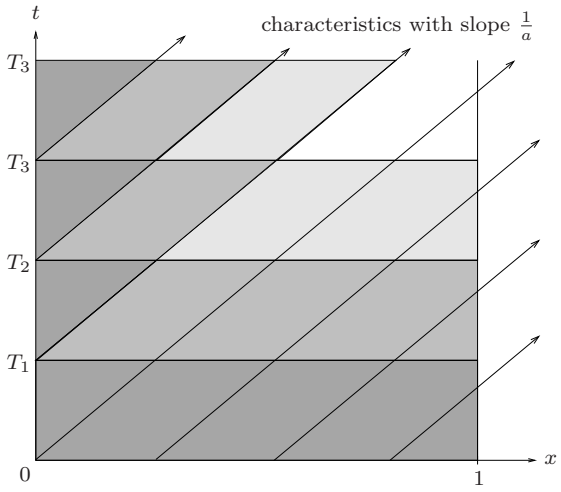


Figure 2: Convergence mechanism of the parareal algorithm applied to the advection equation.

where the constant C can be estimated by

$$C = \max_{n=1,2,\ldots,N} ||u(\cdot, T_n) - U_n^0||_\infty.$$

Proof. For $k = 0$, the estimate holds, since the sum of the L^1 norms is bounded by the maximum of the L^∞ norms multiplied by L and N. To obtain the decay estimate (8) for $k > 0$, Lemma 1 shows that $||u(\cdot, T_n) - U_n^k||_1 = 0$ for $n \leq k$, and using Lemma 2 inductively in k shows that $U_n^k(x) = u(x, T_n)$ for $x \in [0, k\tilde{a}\Delta T]$. Therefore, the difference $u(x, T_n) - U_n^k(x)$ is only non-zero for $n > k$ and $x > k\tilde{a}\Delta T$, which leads to the estimate (8). □

Remark 2 *The convergence result stated in Theorem 1 for the case of a constant advection term can be generalized to the case of variable advection, one simply needs to estimate the minimal distance over which the solution is transported. As long as this quantity remains positive, a similar convergence estimate holds.*

4 Numerical Experiments

We solve the advection equation (7) with $a = 1$ on the domain $\Omega = (0, L)$ with $L = 1$, and in time from zero up to $T = 2$, using for the initial condition $u^0(x) = e^{-100(x-\frac{L}{2})^2}$, for the boundary condition $g(t) = \sin 5t$, and for the source function $f(x, t) = 0$. We discretize the equation using the simple first order upwind scheme

$$\frac{u_{i+1}^j - u_i^j}{\Delta t} + a\frac{u_i^j - u_i^{j-1}}{\Delta x} = f_i^j,$$

with fine spatial and temporal discretization steps Δx and Δt to emulate F, and with coarser spatial and temporal discretization steps ΔX and ΔT to obtain the coarse approximation G. Note that we need to interpolate the solution from the coarse spatial grid to the fine spatial grid, and we use here linear interpolation.

In the first experiment, we chose for the coarse mesh $\Delta T = \frac{T}{12}$ and $\Delta X = \frac{1}{6}$, and for the fine mesh $\Delta t = \frac{T}{240}$ and $\Delta x = \frac{1}{120}$. We show in Figure 3 the initial guess and the first five iterations of the parareal algorithm. One can clearly see how the error is removed step by step, both from the initial line and also from the left boundary, as predicted by Theorem 8. We show in Figure 4 the convergence curve corresponding to this experiment, together with the theoretical estimate from Theorem 1. This shows that the rate estimate is quite sharp, the only overestimate is in this example the constant due to the use of the L^∞ and maximum norms in (8), but one could construct an example where this estimate is sharp as well.

We used in this first example a spatial and temporal discretization step which is precisely at the CFL condition, $a\frac{\Delta T}{\Delta X} = 1$, and thus Assumption 1 is verified with $\tilde{a} = a = 1$. In the next example, we change the coarse time step to $\Delta T = \frac{T}{13}$ and the fine time step to $\Delta t = \frac{T}{260}$, so that the discretization now stays below the CFL condition, $a\frac{\Delta T}{\Delta X} \approx 0.923$. In this case, Assumption 1 is not verified any more with a strictly positive $\tilde{a}$ for our discretization, since at each grid point, the scheme uses information from the same grid point one step earlier in time. We show in Figure 5 again the initial guess and the first five iterates. Even though our analysis does not apply any more, the behavior of the parareal algorithm is very similar to the case where Assumption 1 is verified: the error is still removed from the boundary as well, but now only approximately, as one can see in the error plots: on the left, at iteration two, and even more pronounced at iteration 3, the error is not identically zero any more in the corresponding spatial interval, it takes one more iteration to remove it there, as one can see in iteration 4. We show in Figure 6 the convergence curve for this case. Clearly the algorithm does not converge any more in the sixth step, but a more rapid convergence regime sets in, probably due to the slight diffusive nature of the discretized problem [11].

5 The Parareal Algorithm for the Wave Equation

It is tempting to try to generalize the convergence analysis using characteristics to the case of the second order wave equation,

$$\begin{array}{rcll} u_{tt} & = & c^2 u_{xx} & \text{in } \Omega \times (0,T), \\ u(x,0) & = & u^0(x) & \text{in } \Omega. \end{array} \tag{9}$$

The property of the advection equation (7) which led to the convergence result in Theorem 1 was the fact that the solution later in time is only affected by the boundary condition, and not the initial condition. This is however not the case

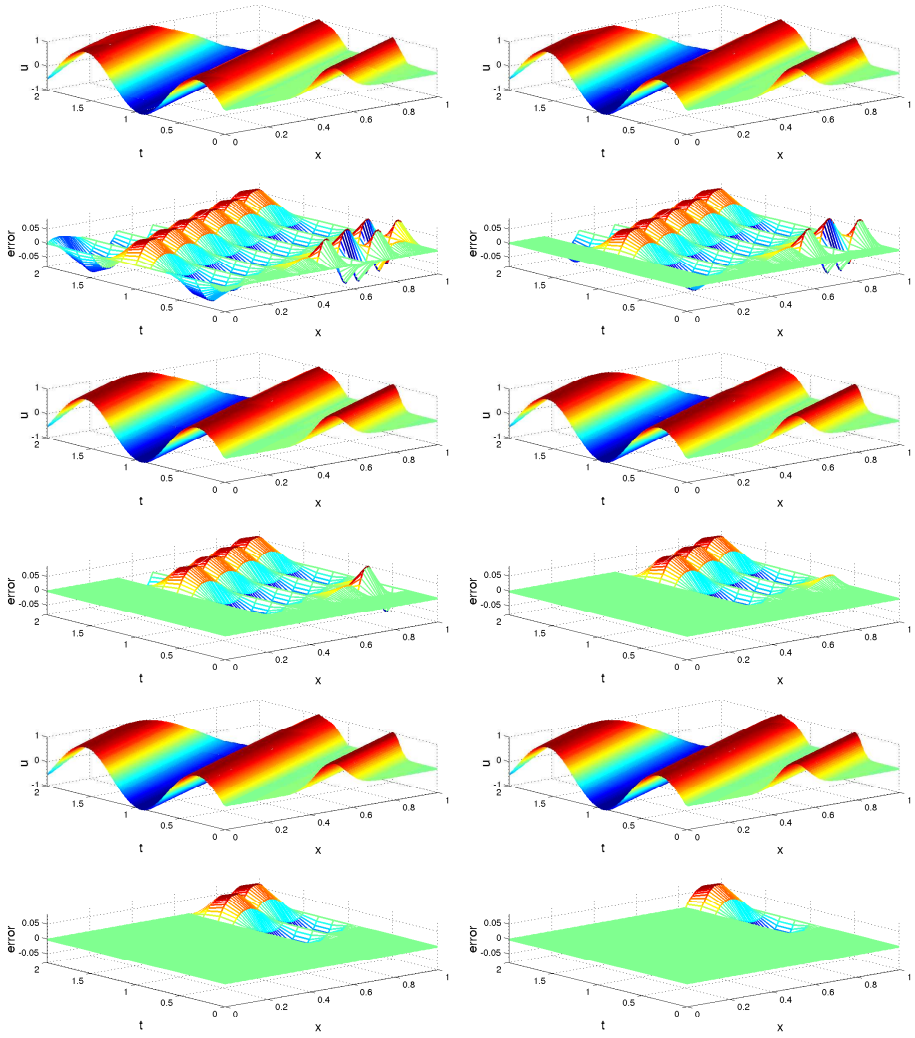


Figure 3: Initial guess and first five iterates of the parareal algorithm, on top the approximate solution, and underneath each time the error of the approximation, for the case where the discretization is precisely at the CFL condition.

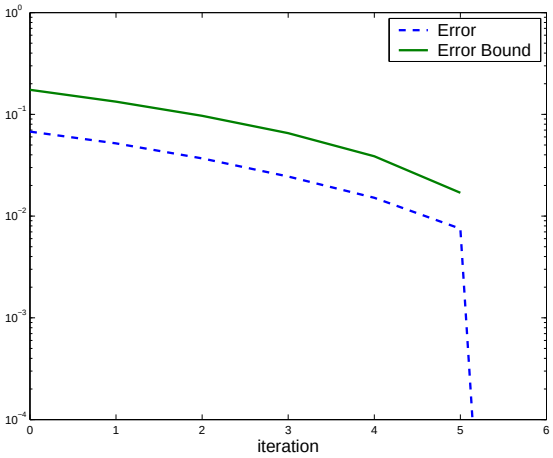


Figure 4: Convergence behavior of the parareal algorithm applied to the advection equation, together with the theoretical estimate.

for the wave equation (9) with Dirichlet boundary conditions

$$\begin{array}{rcll} u(0,t) & = & g_l(t) & t \in (0,T), \\ u(L,t) & = & g_r(t) & t \in (0,T), \end{array} \tag{10}$$

since with these conditions, the solution will be reflected on the boundary, and hence the initial condition can have an influence on the solution at an arbitrary later time.

The situation changes when one imposes transparent boundary conditions,

$$\begin{array}{rcll} cu_x(0,t) - u_t(0,t) & = & g_l(t) & t \in (0,T), \\ cu_x(L,t) + u_t(L,t) & = & g_r(t) & t \in (0,T). \end{array} \tag{11}$$

With these conditions, waves that arrive at the boundary are simply absorbed, and the only incoming information comes from the boundary functions $g_l(t)$ and $g_r(t)$. We show a numerical solution as an example in Figure 7, where we create an initial wave at $x = 1$, and also on each boundary a wave at $t = \frac{1}{2}$. A convergence result similar to the one for the advection equation could be shown for this case, as illustrated in Figure 8. There are now two propagation directions, so the algorithm can transport the correct boundary information both from the left and the right boundary, in addition to the initial line. As soon as in a region both the correct information from the left and from the right are available, the algorithm obtains the exact solution in that region. In order to prove such a result however, one will need a similar assumption on G as Assumption 1 for the advection equation, and unlike in the advection case, it is not natural for a discretization of the wave equation to solve the two propagation directions independently. For example the standard second order centered finite difference scheme for (9) is always using information from both

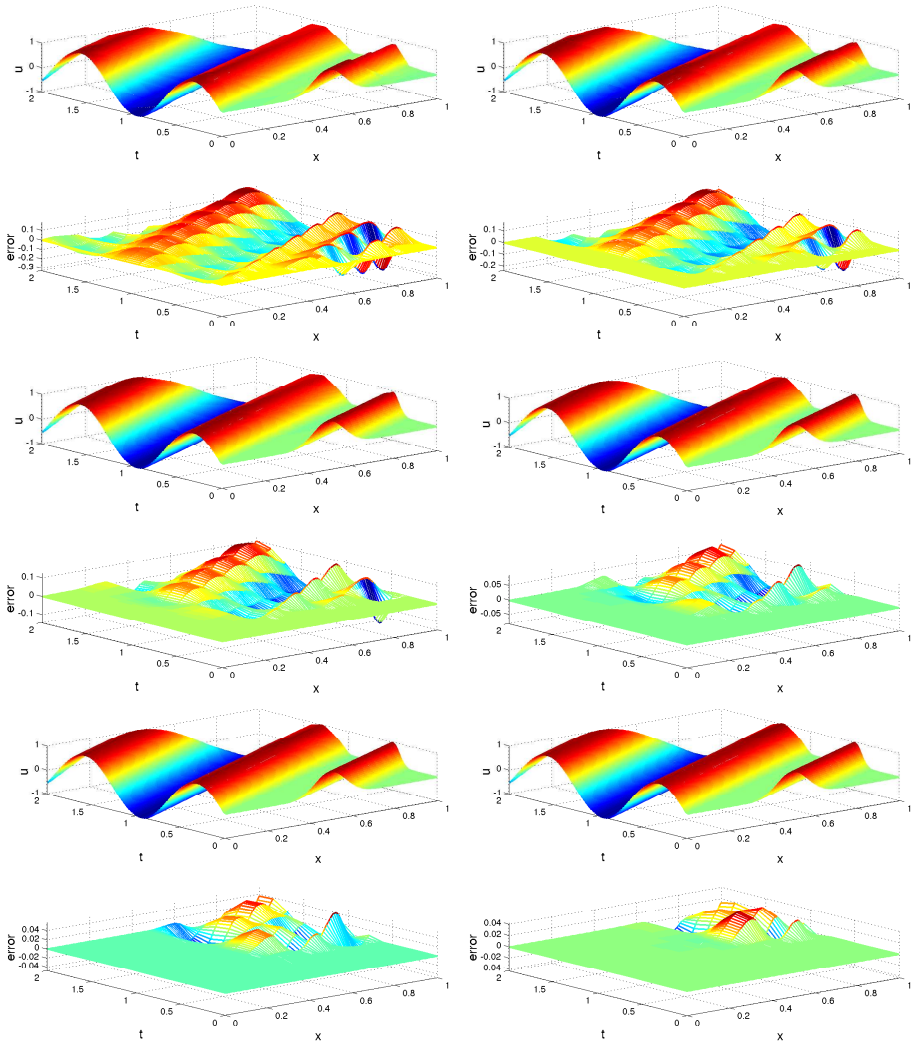


Figure 5: Initial guess and first five iterates of the parareal algorithm, on top the approximate solution, and underneath each time the error of the approximation, for the case where the discretization is below the CFL condition.

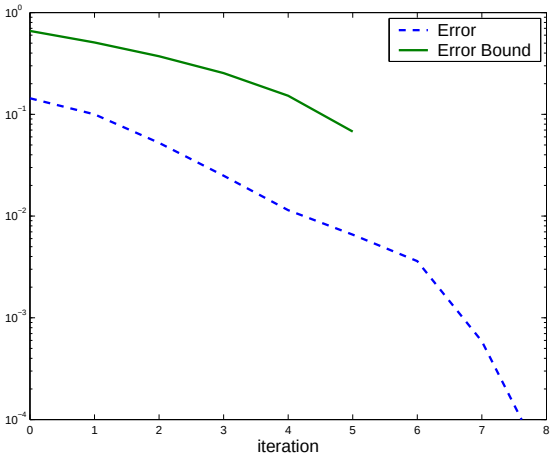


Figure 6: Convergence behavior of the parareal algorithm applied to the advection equation, together with the theoretical estimate, when Assumption 1 is violated.

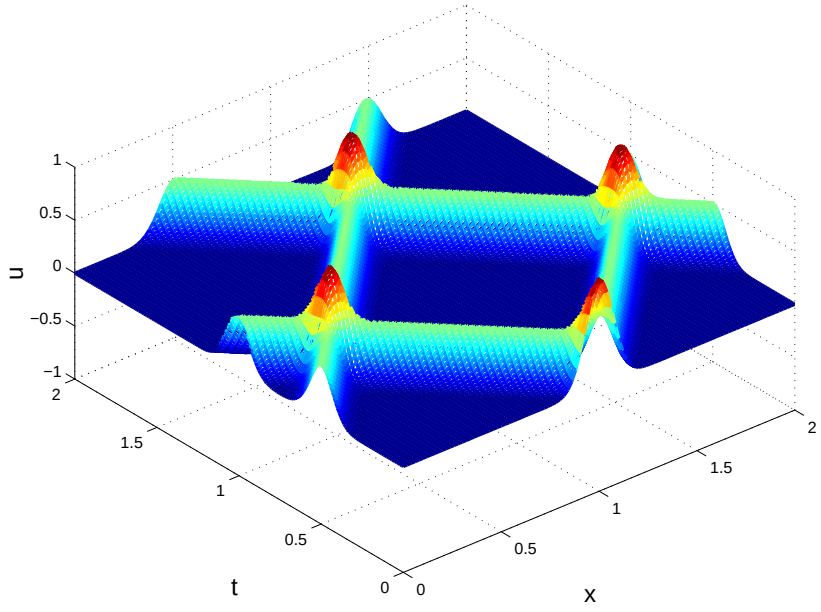


Figure 7: An example of a solution of the wave equation with transparent boundary conditions.

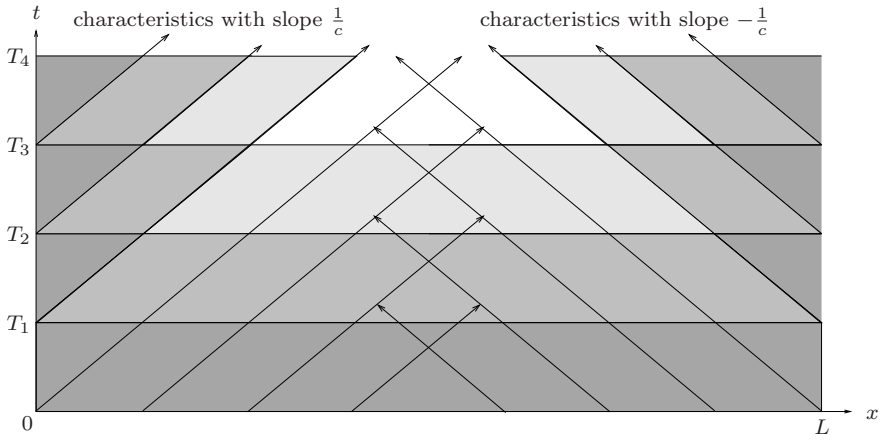


Figure 8: Idea how to generalize the convergence argument for the advection equation using characteristics to the wave equation.

directions, which prevents a convergence argument from going through. An illustrative example is shown in Figure 9. While the algorithm clearly proceeds to obtain the solution from the initial line in the time direction, as proved in Lemma 1, the solution at the two spatial boundaries is not obtained as in the case of the advection equation, only once the correct front from the initial condition reaches the point where the boundary condition is non-zero. Thus a convergence result like Theorem 1 does not hold when the parareal algorithm (6) is applied to an arbitrary discretization of the second order wave equation (9), not even with transparent boundary conditions (11).

6 Conclusions

Using characteristics, we have obtained a convergence result for the parareal algorithm applied to the advection equation with Dirichlet boundary condition. The algorithm computes in that case exact solution parts from the boundary inward, as it does usually from the initial line. This result can be generalized for variable coefficient advection problems, the only property needed in the proof is the transport of information along characteristics. Our analysis indicates however that the parareal algorithm is not the ideal tool to parallelize the solution of the advection equation: it would be much more efficient to solve such problems along characteristics, and then each characteristic can be solved independently, the problem becomes embarrassingly parallel.

In the case of the second order wave equation, the parareal algorithm can not obtain the exact solution from the boundary inward, even though the solution also has propagation directions, as in the case of the advection equation. For the algorithm to successfully do so, it would need a discretization which satisfies an assumption similar to the key assumption made for the advection equation, and

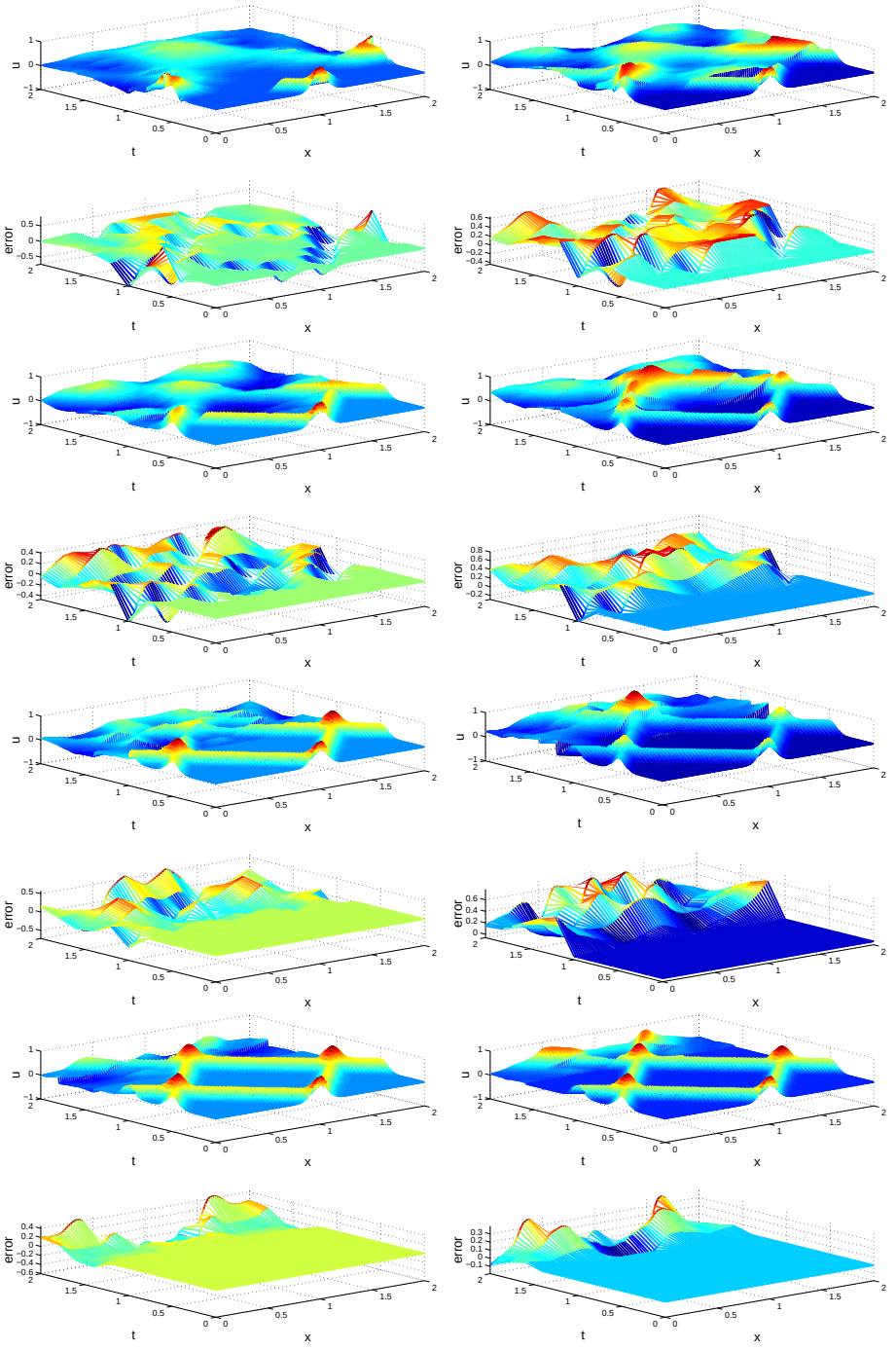


Figure 9: Initial guess and first seven iterates of the parareal algorithm, on top the approximate solution, and underneath each time the error of the approximation, for the case of the second order wave equation.

usual discretizations of the wave equation do not satisfy such an assumption. The algorithm is thus only converging from the initial line, which makes it not very useful for parallelizing the solution of the wave equation in time, since the number of iterations needed is then equal to the number of processors one can use in time.

If one needs to compute approximate solutions of the second order wave equation in a time parallel fashion, one therefore either needs to find discretizations of the wave equation which satisfy a propagation assumption, like in the case of the advection equation, or one needs to use a modified time parallel algorithm, a possibility being the modification of the parareal algorithm proposed in [5, 7, 10].

References

[1] P. Amodio and L. Brugnano. Parallel implementation of block boundary value methods for ODEs. *J. Comp. Appl. Math.*, 78:197–211, 1997.

[2] L. Baffico, S. Bernard, Y. Maday, G. Turinici, and G. Zérah. Parallel-in-time molecular-dynamics simulations. *Physical Review E*, 66:057706–1–4, 2002.

[3] G. Bal. On the convergence and the stability of the parareal algorithm to solve partial differential equations. In R. Kornhuber, R. H. W. Hoppe, J. Périaux, O. Pironneau, O. B. Widlund, and J. Xu, editors, *Proceedings of the 15th international domain decomposition conference*, pages 426–432. Springer LNCSE, 2003.

[4] P. Chartier and B. Philippe. A parallel shooting technique for solving dissipative ODEs. *Computing*, 51:209–236, 1993.

[5] J. Cortial and C. Farhat. A time-parallel implicit methodology for the near-real-time solution of systems of linear oscillators. In L. Biegler, O. Ghattaas, M. Heinkenschloss, D. Keyes, and B. van Bloemen Waanders, editors, *Real-Time PDE-Constrained Optimization*. SIAM, 2006.

[6] C. Farhat and M. Chandesris. Time-decomposed parallel time-integrators: theory and feasibility studies for fluid, structure, and fluid-structure applications. *Internat. J. Numer. Methods Engrg.*, 58(9):1397–1434, 2003.

[7] C. Farhat, J. Cortial, C. Dastillung, and H. Bavestrello. Time-parallel implicit integrators for the near-real-time prediction of linear structural dynamic responses. *Int. J. Numer. Meth. Engng*, 67(5):697–724, 2006.

[8] P. F. Fischer, F. Hecht, and Y. Maday. A parareal in time semi-implicit approximation of the Navier-Stokes equations. In R. Kornhuber, R. H. W. Hoppe, J. Périaux, O. Pironneau, O. B. Widlund, and J. Xu, editors, *Proceedings of the 15th international domain decomposition conference*, pages 433–440. Springer LNCSE, 2003.

[9] M. J. Gander and E. Hairer. Nonlinear convergence analysis for the parareal algorithm. In U. Langer, O. Widlund, and D. Keyes, editors, *Proceedings of the 17th international domain decomposition conference*. Springer LNCSE, 2007.

[10] M. J. Gander and M. Petcu. Analysis of a modified parareal algorithm for second-order ordinary differential equations. In *AIP Conference Proceedings of the International Conference of Numerical Analysis and Applied Mathematics ICNAAM*, volume 936, pages 233–236, 2007.

[11] M. J. Gander and S. Vandewalle. Analysis of the parareal time-parallel time-integration method. *SIAM J. Sci. Comp.*, 29(2):556–578, 2007.

[12] I. Garrido, M. S. Espedal, and G. E. Fladmark. A convergence algorithm for time parallelization applied to reservoir simulation. In R. Kornhuber, R. H. W. Hoppe, J. Périaux, O. Pironneau, O. B. Widlund, and J. Xu, editors, *Proceedings of the 15th international domain decomposition conference*, pages 469–476. Springer LNCSE, 2003.

[13] I. Garrido, B. Lee, G. E. Fladmark, and M. E. Espedal. Convergent iterative schemes for time parallelization. *Mathematics of Computation*, 75:1403–1428, 2006.

[14] E. Lelarasmee, A. E. Ruehli, and A. L. Sangiovanni-Vincentelli. The waveform relaxation method for time-domain analysis of large scale integrated circuits. *IEEE Trans. on CAD of IC and Syst.*, 1:131–145, 1982.

[15] J.-L. Lions, Y. Maday, and G. Turinici. A parareal in time discretization of pde's. *C.R. Acad. Sci. Paris, Serie I*, 332:661–668, 2001.

[16] Y. Maday and G. Turinici. A parareal in time procedure for the control of partial differential equations. *C.R.A.S. Sér. I Math*, 335:387–391, 2002.

[17] Y. Maday and G. Turinici. The parareal in time iterative solver: a further direction to parallel implementation. In R. Kornhuber, R. H. W. Hoppe, J. Périaux, O. Pironneau, O. B. Widlund, and J. Xu, editors, *Proceedings of the 15th international domain decomposition conference*, pages 441–448. Springer LNCSE, 2003.

[18] W. L. Miranker and W. Liniger. Parallel methods for the numerical integration of ordinary differential equations. *Math. Comp.*, 91:303–320, 1967.

[19] J. Nievergelt. Parallel methods for integrating ordinary differential equations. *Comm. ACM*, 7:731–733, 1964.

[20] G. A. Staff and E. M. Rønquist. Stability of the parareal algorithm. In R. Kornhuber, R. H. W. Hoppe, J. Périaux, O. Pironneau, O. B. Widlund, and J. Xu, editors, *Proceedings of the 15th international domain decomposition conference*, pages 449–456. Springer LNCSE, 2003.

Bol. Soc. Esp. Mat. Apl.
nº 42(2008), 37–48

ON THE LIMIT CYCLES OF THE LIÉNARD DIFFERENTIAL SYSTEMS

JAUME LLIBRE

Departament de Matemàtiques, Universitat Autònoma de Barcelona, 08193 Bellaterra, Barcelona, Catalonia, Spain

jllibre@mat.uab.cat

Abstract

One of the main interesting problems in the qualitative theory of planar differential equations is the problem of studying their limit cycles. For a particular subclass of planar differential systems, the Liénard systems, we shall present some old and new results on their limit cycles.

Key words: *Limit cycles, non–existence, uniqueness, Liénard system*
AMS subject classifications: *34C05, 34C07.*

1 Introduction

One of the most interesting and classical problems in the qualitative theory of planar differential equations is the study of their *limit cycles*, i.e. for a differential system of the form

$$\begin{aligned} \dot{x} &= P(x,y), \\ \dot{y} &= Q(x,y), \end{aligned} \tag{1}$$

where $P, Q : \mathbb{R}^2 \to \mathbb{R}$ are $\mathcal{C}^1$ functions, what are their isolated periodic orbits in the set of all periodic orbits?

One of the classes of planar differential systems more studied are those equivalent to the *generalized Liénard differential equation*

$$\ddot{x} + f(x)\dot{x} + g(x) = 0, \tag{2}$$

which were considered by many researchers, for instance see the references of this paper, or if the day that I was written this paper you looked in MathSciNet for the number of articles with the keywords *limit cycle* and *Liénard* you obtained 404.

Trabajo subvencionado por los proyectos MEC/FEDER número MTM2005-06098-C02-01 y CICYT número 2005SGR 00550.

A special subclass of Liénard differential equations are the *classical Liénard differential equations* (2) obtained when $g(x) = x$.

We consider the following assumptions on the functions $f, g : \mathbb{R} \to \mathbb{R}$:

(I) f and g are $\mathcal{C}^1$ functions defined in $\mathbb{R}$;

(II) $g(0) = 0$, $g'(0) > 0$ and $g(x)x > 0$ if $x \neq 0$.

The second order differential equation (2) is equivalent to the first order system:

$$\begin{aligned} \dot{x} &= y - F(x), \\ \dot{y} &= -g(x), \end{aligned} \tag{3}$$

where $F(x) = \displaystyle\int_0^x f(s)ds$. Hypothesis (I) implies that the theorem on the existence and uniqueness of the solutions of an ordinary differential equation holds for system (3). Assumption (II) guarantees that the origin is the only singular point of system (3), and also that the determinant of the linear part of system (3) at the origin is positive, so it is a node or a focus, see for instance [9]. Moreover if system (3) has periodic orbits, these turn clockwise around the origin.

Another way to write the second order differential equation (2) as a planar differential system of first order is

$$\begin{aligned} \dot{x} &= y, \\ \dot{y} &= -f(x)y - g(x). \end{aligned} \tag{4}$$

In this note we study the limit cycles of the Liénard differential systems (3), first when $F(x)$ is a polynomial and $g(x) = x$, see Section 2; and second when the functions $f(x)$ and $g(x)$ satisfy the assumptions (I) and (II), see Section 3.

2 Polynomial Liénard differential systems

Hilbert [15] in 1900 and in the second part of its 16–th problem proposed to find an estimation of the uniform upper bound for the number of limit cycles of all polynomial differential systems of a given degree, and also to study their distribution or configuration in the plane. Except for the Riemann hypothesis, the 16–th problem seems to be the most elusive of Hilbert's problems. It has been one of the main problems in the qualitative theory of planar differential equations in the XX century. The contributions of Écalle [11] and Ilyashenko [16] proving that any polynomial differential system has finitely many limit cycles have been the best results in this area. But until now it is not proved the existence of an uniform upper bound. This problem remains open even for the quadratic polynomial differential systems. However it is not difficult to see that any configuration of limit cycles is realizable for some polynomial differential system, see for details [20].

We have the finiteness of the number of limit cycles for every polynomial differential system of degree n, but we do not have uniform bounds for that number in the whole class of all polynomial differential systems of degree n.

Following to Smale [28] we consider a more easy and special class of polynomial differential systems, the *polynomial Liénard differential systems*:

$$\begin{aligned} \dot{x} &= y - F(x), \\ \dot{y} &= -x, \end{aligned} \tag{5}$$

where $F(x) = -a_1 x - \cdots - a_n x^n$. For these systems the existence of uniform bounds also remain unproved. But when the degree n of these systems is odd Ilyashenko and Panov in [17] obtained an uniform upper bound for the number of limit cycles in a subclass of systems such that F is monic and its coefficients satisfy some estimations. In short a first open problem for the polynomial Liénard differential systems is:

Open problem 1. *Show that there exists a positive integer $L = L(n)$ such that any polynomial Liénard differential systems* (5) *has at most L limit cycles if the degree of $F(x)$ is n.*

Lins, de Melo and Pugh [19] conjectured that $L(n) = [(n-1)/2]$, where $[a]$ denotes the integer part function of $a \in \mathbb{R}$. This conjecture was supported mainly by the following three facts.

(i) The Liénard systems of the form

$$\begin{aligned} \dot{x} &= y - \varepsilon F(x), \\ \dot{y} &= -x, \end{aligned}$$

with ε sufficiently small have at most k limit cycles bifurcating from the periodic orbits of the linear center $\dot{x} = y$, $\dot{y} = -x$, and there are examples with exactly k limit cycles. For the original proof see [19], or for a shorter proof see Section 4 where we reproduce the proof of [3].

(ii) It is known that systems (5) have a center at the origin if and only if $a_i = 0$ for all the odd i's, and consequently at most k small limit cycles can bifurcate from these centers through a Hopf bifurcation, when we perturb them inside the class of all Liénard differential systems of degree $n = 2k+1$ or $2k+2$, see Zuppa [35], and also Blows and Lloyd [1].

(iii) López and López–Ruiz [23] have studied the Liénard differential systems (5) in what they call the strongly nonlinear regime. In this regime they show that the conjecture is also true when n is odd.

(iv) Xiudong Chen and Yong Chen [5] show that the Lins, de Melo and Pugh conjecture holds when the polynomial function $F(x)$ is odd (i.e. $F(-x) = -F(x)$), so the previous result of López and López–Ruiz becomes a particular case of this last result.

In many other papers where some subclasses of Liénard differential systems (5) have been studied, the results provide support to the conjecture, see for instance [21]. The conjecture is true if $F(x)$ has degree 1 (see Corollary 2), 2 (see Corollary 4) and 3 (see [19]).

Recently Dumortier, Panazzolo and Roussarie [10] have proved that the Lins, de Melo and Pugh conjecture does not hold when the degree of $F(x)$ is 7, because in this case the maximum number of excepted limit cycles is $[(7-1)/2] = 3$ and they can construct an example with 4 limit cycles. The limit cycles that they found are relaxation oscillations which appear in slow–fast systems at the boundary of the polynomial Liénard differential systems (5). This paper shows that the slow–fast dynamics can provide new information about the limit cycles.

At this moment is unknown if the Lins, de Melo and Pugh conjectured is true when the degree of $F(x)$ is 4, 5 or 6. There are several papers showing that the conjecture is true for particular subfamilies of polynomials $F(x)$ of degree 4. So the easiest first open problem related with the Lins, de Melo and Pugh conjecture or with the 16–th Hilbert problem restricted to polynomial Liénard differential systems is the following.

Open problem 2. *The polynomial Liénard differential system*

$$\begin{aligned} \dot{x} &= y - a_1 x - a_2 x^2 - a_3 x^3 - a_4 x^4, \\ \dot{y} &= -x, \end{aligned}$$

with $a_4 \neq 0$ has at most one limit cycle.

3 $\mathcal{C}^1$ Liénard differential systems

We shall present seven different criteria for studying the limit cycles of the Liénard differential equation (2), or equivalently of the Liénard differential systems (3) or (4). Except the criteria given in Propositions 1 and 8, all the others are given in [13].

The first two criteria provide sufficient conditions for the non–existence of limit cycles. In fact Proposition 1 is well–known (see for instance [9] or [33]).

Set $G(x) = \displaystyle\int_0^x g(s)ds$. In what follows f, F, g and G will denote $f(x)$, $F(x)$, $g(x)$ and $G(x)$ respectively.

Proposition 1 (Bendixson Criterium) *Assume that the divergence function $\partial P/\partial x + \partial Q/\partial y = -f$ of the Liénard differential system (4) satisfies either > 0 or < 0 in a simply connected region R. Then this system has no periodic orbits contained in R.*

Corollary 2 *The Liénard differential system*

$$\dot{x} = y - a_1 x, \qquad \dot{y} = -x, \tag{6}$$

with $a_1 \neq 0$ has no limit cycles.

Proof. For system (6) we have that $f(x) = a_1 \neq 0$. So applying Proposition 1 to this system with $R = \mathbb{R}^2$, it follows that it has no limit cycles. □

Proposition 3 *Let k_0 be a fixed arbitrary real constant. Assume that the function $k_0 f - g$ satisfies either > 0 or < 0 in a simply connected region R. Then the Liénard differential system* (4) *has no periodic orbits contained in R.*

Corollary 4 *The Liénard differential system*

$$\dot{x} = y - a_1 x - a_2 x^2, \qquad \dot{y} = -x, \tag{7}$$

with $a_2 \neq 0$ has no limit cycles.

Proof. For system (7) we have that $f(x) = a_1 + 2a_2 x$. We distinguish two cases.
Case 1: $a_1 \neq 0$. Applying Proposition 3 to this system with $R = \mathbb{R}^2$ and $k_0 = 1/(2a_2)$, we get that $k_0 f - g = a_1/(2a_2) \neq 0$. So it follows that system (7) has no limit cycles.
Case 2: $a_1 = 0$. Then $H(x, y) = e^{-2a_2 y}(2a_2^2 x^2 - 2a_2 y - 1)$ is a first integral of system (7) defined in all $\mathbb{R}^2$. Hence this system cannot have limit cycles. □

An interesting case is when the Liénard differential equation (2) has a single limit cycle. There are many results providing sufficient conditions for the uniqueness of the limit cycles of Liénard differential systems, far of being exhaustive see [2, 4, 24, 26, 27, 29, 32] and the results in the books [33, 34]. Here we present three new criteria for the uniqueness of the limit cycles of Liénard differential systems (3) and (4), and also the Massera criterium (see Proposition 8).

Before we need to recall the following result, for a proof see for instance Theorem 1.23 of [9]. Suppose that system (1) has a periodic orbit $(x(t), y(t))$ of period T. Let

$$\sigma = \int_0^T \left(\frac{\partial P}{\partial x} + \frac{\partial Q}{\partial y} \right) (x(t), y(t))\, dt.$$

If $\sigma > 0$ (resp. $\sigma < 0$) then the periodic orbit $(x(t), y(t))$ is an unstable (resp. stable) limit cycle. A periodic orbit $(x(t), y(t))$ having $\sigma \neq 0$ is a *hyperbolic* limit cycle.

Proposition 5 *Let k_0 and k_1 be fixed arbitrary real constants. Assume that the function $-g(k_1 + F) + f(k_0 + 2G)$ satisfies either > 0 or < 0 in a simply connected region R. Then the Liénard differential system* (4) *has at most one periodic orbit which lies entirely in R, when it exists is a hyperbolic limit cycle.*

Proposition 6 *Let k_0, k_1 and k_2 be fixed arbitrary real constants with $k_2 \neq 0$. Assume that the function $k_2[-g(k_1 + k_2 x + F) + f(k_0 + 2G)]$ satisfies either > 0 or < 0 in a simply connected region R. Then the Liénard differential system* (4) *has at most one periodic orbit which lies entirely in R, when it exists is a hyperbolic limit cycle.*

Proposition 7 *Let k_0 and k_2 be fixed arbitrary real constants. Assume that the function $-g(k_2F + F^2 + 3G) + f(k_0 + 2k_2G + 4\int_0^x g(s)F(s)ds)$ satisfies either > 0 or < 0 in a simply connected region R. Then the Liénard differential system* (4) *has at most one periodic orbit which lies entirely in R, when it exists is a hyperbolic limit cycle.*

The following criterium was proved by Massera [24], and by Sansone [27] for a more particular subclass. In Section 5 we present a new proof due to Gabriele Villari and the author.

Proposition 8 (Massera's Theorem) *We consider the Liénard differential system* (4) *with $g(x) = x$, $f(0) < 0$ and $f'(x)x > 0$ if $x \neq 0$. Then system* (4) *has at most one limit cycle.*

There are in the literature few results providing sufficient conditions in order that a planar differential systems has at most n limit cycles with $n > 1$, see for instance [5, 6, 7, 12, 25]. The next criterium provide sufficient conditions in order that a Liénard differential system has at most 2 limit cycles.

Proposition 9 *Let k_i be arbitrary real constants for $k = 0, 1, 2, 3$. Assume that the function*

$$f\left(k_0 + 2k_2G + 4G^2 + 4k_3\int_0^x g(s)F(s)ds + 6\int_0^x g(s)F(s)^2ds\right) - g\left(k_1 + k_2F + 3k_3G + k_3F^2 + F^3 + 4FG + 5\int_0^x g(s)F(s)ds\right)$$

satisfies either > 0 or < 0 in a simply connected region R. Then the Liénard differential system (4) *has at most two periodic orbits contained in R, when they exist are hyperbolic limit cycles.*

4 Proof of a result of Lins, de Melo and Pugh

The goal of this section is to provide a shorter proof of the next result of Lins, de Melo and Pugh [19].

Proposition 10 *The Liénard differential system*

$$\begin{aligned} \dot{x} &= y - \varepsilon(a_1x + \cdots + a_nx^n), \\ \dot{y} &= -x, \end{aligned} \tag{8}$$

for ε sufficiently small has at most $[(n-1)/2]$ limit cycles bifurcating from the periodic orbits of the linear center $\dot{x} = y$, $\dot{y} = -x$, and there are examples with exactly $[(n-1)/2]$.

The proof of this proposition uses the first order averaging theory for studying the periodic orbits of a differential system. More precisely we consider the differential system

$$\dot{\mathbf{x}}(t) = \varepsilon F(t, \mathbf{x}(t)) + \varepsilon^2 R(t, \mathbf{x}(t), \varepsilon), \tag{9}$$

with $\mathbf{x} \in D \subset \mathbb{R}^m$, D a bounded domain and $t \geq 0$. Moreover, we assume that $F(t, \mathbf{x})$ and $R(t, \mathbf{x}, \varepsilon)$ are T–periodic in t.

The averaged system associated to system (9) is defined by

$$\dot{\mathbf{y}}(t) = \varepsilon f(\mathbf{y}(t)), \tag{10}$$

where

$$f(\mathbf{y}) = \frac{1}{T} \int_0^T F(s, \mathbf{y}) ds. \tag{11}$$

The next theorem says us under which conditions the singular points of the averaged system (10) provide T–periodic orbits of system (9). For a proof see Theorem 2.6.1 of [26], Theorems 11.5 and 11.6 of [31], and Theorem 4.1.1 of [14].

Theorem 11 *We consider system* (9) *and assume that the functions F, R, $D_{\mathbf{x}}F$, $D^2_{\mathbf{x}}F$ and $D_{\mathbf{x}}R$ are continuous and bounded by a constant M (independent of ε) in $[0, \infty) \times D$ with $-\varepsilon_0 < \varepsilon < \varepsilon_0$. Moreover, we suppose that F and R are T–periodic in t, with T independent of ε.*

(a) If $a \in D$ is a singular point of the averaged system (10) *such that $\det(D_{\mathbf{x}} f(a)) \neq 0$ then, for $|\varepsilon| > 0$ sufficiently small, there exists a T–periodic solution $\mathbf{x}_\varepsilon(t)$ of system* (9) *such that $\mathbf{x}_\varepsilon(t) \to a$ as $\varepsilon \to 0$.*

(b) If the singular point $\mathbf{y} = a$ of the averaged system (10) *is hyperbolic then, for $|\varepsilon| > 0$ sufficiently small, the corresponding periodic solution $\mathbf{x}_\varepsilon(t)$ of system* (9) *is unique, hyperbolic and of the same stability type as a.*

Proof Proof of Proposition 10. We write system (8) in polar coordinates (r, θ) where $x = r\cos\theta$, $y = r\sin\theta$, and we obtain

$$\dot{r} = -\varepsilon \sum_{k=1}^{n} a_k r^k \cos^{k+1}\theta,$$
$$\dot{\theta} = -1 + \varepsilon \sin\theta \sum_{k=1}^{n} a_k r^{k-1} \cos^k \theta,$$

or equivalently

$$\frac{dr}{d\theta} = -\varepsilon \sum_{k=1}^{n} a_k r^k \cos^{k+1}\theta + O(\varepsilon^2).$$

Then the averaged system associated to the previous system is

$$\frac{dy}{d\theta} = -\frac{\varepsilon}{2\pi} \sum_{k=1}^{n} a_k r^k \int_0^{2\pi} \cos^{k+1}\theta \, d\theta = -\frac{\varepsilon}{2\pi} \sum_{\substack{k=1 \\ k \text{ odd}}}^{n} a_k b_k r^k = p(r),$$

where $b_k = \displaystyle\int_0^{2\pi} \cos^{k+1}\theta \, d\theta \neq 0$ if k is odd. Then applying Theorem 11 the polynomial $p(r)$ has at most $[(n-1)/2]$ positive roots, and we can choose the

coefficients a_k with k odd in such a way that $p(r)$ has exactly $[(n-1)/2]$ simple positive roots. Hence the proposition follows. □

5 Proof of Massera's Theorem

We consider the family of straight lines $\mathcal{L}_k$ defined by $y = -k\,x$, when k runs in $\mathbb{R}$. We add to this family the straight line $\mathcal{L}_\infty$ defined by $x = 0$. So k runs in $\mathbb{R} \cup \{\infty\}$. Note that this family is formed by all the straight lines passing through the origin of coordinates.

Let γ be a periodic orbit of system (4). We say that γ has a *starshape with respect to the family of straight lines through the origin* if for every $k \in \mathbb{R}$ the straight line $\mathcal{L}_k$ intersects the closed curve γ exactly in two points.

Proposition 12 *Under the assumptions of Massera's Theorem and the additional assumption that any periodic orbit of a Liénard differential system* (4) *has a starshape with respect to the family of straight lines through the origin Massera's Theorem holds.*

Proof. Under the assumptions of Massera's Theorem the unique singular point of system (4) is an unstable node or focus at the origin.

We suppose that system (4) has two periodic orbits γ^* and γ. We shall arrive to a contradiction and the proposition will be proved. Without loss of generality we can assume that γ^* is contained inside the closed bounded region $R(\gamma)$ limited by γ.

We shall construct a new vector field on $R(\gamma)$ extending the vector field of system (4) restricted to the periodic orbit γ to the whole region $R(\gamma)$ producing a global center in this region.

First we introduce new coordinates in $R(\gamma) \setminus \{(0,0)\}$. Since γ has a starshape with respect to the straight lines through the origin for a point $(x_0, y_0) \in R(\gamma) \setminus \{(0,0)\}$ which is on the straight line $\mathcal{L}_{k_0}$ we can take the coordinates $(r, (x, y))$, where (x, y) is the point of $\gamma \cap \mathcal{L}_{k_0}$ which is in the same ray of $\mathcal{L}_{k_0} \setminus \{(0,0)\}$ than (x_0, y_0), and $r = x_0/x$ if $x_0 \neq 0$ or $r = y_0/y$ if $x_0 = 0$. Therefore, $(x_0, y_0) = (rx, ry)$ and $0 < r < 1$.

The vector field associated to the differential system (4) with $g(x) = x$ is

$$\mathcal{X}(x, y) = (y, -f(x)\,y - x),$$

for all $(x, y) \in \mathbb{R}^2$.

Now we define in $R(\gamma)$ the new vector field whose phase portrait will have a center at the origin. The new vector field in a point $(x_0, y_0) = (rx, ry) \in R(\gamma) \setminus \{(0,0)\}$ with $(x, y) \in \gamma$ and $r \in (0, 1]$ is defined as

$$\mathcal{Y}(x_0, y_0) = r(y, -f(x)y - x),$$

and $\mathcal{Y}(0, 0) = (0, 0)$.

If $(x(t), y(t))$ denotes the periodic solution γ of the vector field $\mathcal{X}$, it is immediate to check that the closed curve $(rx(t), ry(t))$ for every $r \in (0,1)$ constant is a periodic orbit of the vector field $\mathcal{Y}$. Therefore the origin is a global center for the vector field $\mathcal{Y}$ on $R(\gamma)$. We remark that all the periodic orbits of the vector field $\mathcal{Y}$ are homothetic to the periodic orbit γ.

We define the orthogonal vector field $\mathcal{Y}^\perp$ to the vector field $\mathcal{Y}$ in the points $(x_0, y_0) = (rx, ry) \in R(\gamma) \setminus \{(0,0)\}$ as

$$\mathcal{Y}^\perp(x_0, y_0) = r(f(x)y + x, y).$$

Let p be a point of the periodic orbit γ^* inside $R(\gamma)$ and let Γ be the periodic orbit of the vector field $\mathcal{Y}$ passing through p. Then for every point $(x_0, y_0) \in \Gamma$ we consider the inner product

$$\begin{aligned} \mathcal{Y}^\perp(x_0, y_0) \cdot \mathcal{X}(x_0, y_0) = & \quad r(f(x)y + x, y) \cdot (ry, -f(rx)ry - rx) \\ = & \quad r^2 y^2 (f(x) - f(rx)) \begin{cases} > 0 & \text{if } xy \neq 0, \\ = 0 & \text{if } xy = 0, \end{cases} \end{aligned}$$

because $0 < r < 1$, $f(0) < 0$ and $f'(x)x > 0$ if $x \neq 0$. Since the inner product between the orthogonal vector to the closed curve Γ at the point (x_0, y_0) and the vector field of system (4) at the same point never is negative, the closed curve Γ is transversal to the flow of system (4). Here a *transversal closed curve* Γ with respect to the flow of system (4) means that this flow either enters into the bounded region $R(\Gamma)$ limited by Γ through all the points of Γ, or exits out of $R(\Gamma)$ through all the points of Γ. Hence the orbit γ^* of system (4) through the point p cannot be periodic. This is a contradiction and consequently the proposition is proved. □

Proposition 13 *Assume that we have a Liénard differential system* (4) *satisfying the assumptions of Massera's Theorem. Then every periodic orbit of system* (4) *has a starshape with respect to the family of straight lines through the origin.*

Proof. We assume that system (4) has a periodic orbit γ without the starshape with respect to the family of straight lines through the origin. Then it exists a ray r with endpoint at the origin that intersects the closed curve γ in at least three points. So without taking into account the origin of the ray r the vector field $(\dot{x}, \dot{y})$ associated to system (4) has at least two contact points; i.e. there are at least two points in r where the vector field $(\dot{x}, \dot{y})$ is tangent to the ray r.

Let $y = -kx$ with either $x > 0$, or $x < 0$, the ray r. Then if $(x, y) \in r$ is a contact point with the vector field $(\dot{x}, \dot{y})$ it must satisfy

$$k\dot{x} + \dot{y}\Big|_{y=-kx} = -x(1 - kf(x) + k^2) = 0.$$

The solution $x = 0$ corresponds to the contact point at the origin. Since $f(0) < 0$ and $f'(x)x > 0$ if $x \neq 0$, it follows that at most there are two values of x

satisfying $f(x) = (1 + k^2)/k$, one positive and the other negative. Hence, since if there is a periodic orbit of system (4) without the starshape with respect to the family of straight lines through the origin it must exist a ray with at least two contact points, it follows that such a periodic orbit cannot exist. □

From Propositions 12 and 13 it follows Massera's Theorem.

References

[1] T.R. Blows and N.G. Lloyd, The number of small–amplitude limit cycles of Liénard equations, *Math. Proc. Cambridge Philos. Soc.* Vol. 95. pp. 359-366. 1984.

[2] T. Carletti and G. Villari, A note on existence and uniqueness of limit cycles for Liénard systems, *J. Math. Anal. Appl.* Vol. 307. pp 763-773. 2005.

[3] A. Cima, J. Llibre and M. A. Texeira, Limit cycles of some polynomial differential systems in dimension 2, 3 and 4, via averaging theory, to appear in *Applicable Analysis*.

[4] W. A. Coppel, Some Analytical Systems with at most one Limit Cycle, *Dynamics Reported*, Vol. 2, Edited by U. Kirchgraber & H.O.Walther, John Wiley Sons Ltd, pp. 61-88. 1989.

[5] Xiudong Chen and Yong Chen, On the conjecture of A. Lins, W. de Melo and C. C. Pugh, *Ann. Differential Equations* Vol. 22. pp 144-148. 2006.

[6] Xiudong Chen, J. Llibre and Zhifen Zhang, Sufficient conditions for the existence of at least n or exactly n limit cycles for the Liénard differential systems, to appear in *J. Differential Equations*.

[7] L.A. Cherkas, J.C. Artés and J. Llibre, Quadratic systems of limit cycles of normal size, *Bul. Acad. de Stiinte a Rep. Moldova, Matematica* Vol. 1 (41). pp 31-46. 2003.

[8] G.F.D. Duff and N. Levinson, On the non–uniqueness of periodic solutions for an asymmetric Liénard equation, *Quart. Appl. Math.* Vol. 10. pp 86-88. 1952.

[9] F. Dumortier, J. Llibre and J.C. Artés, *Qualitative theory of planar differential systems*, UniversiText, Springer–Verlag, New York, 2006.

[10] F. Dumortier, D. Panazzolo and R. Roussarie, More limit cycles than expected in Liénard systems, *Proc. Amer. Math. Soc.* Vol. 135. pp 1895-1904. 2007.

[11] J. Ecalle, *Introduction aux fonctions analysables et preuve constructive de la conjecture de Dulac*, Hermann, 1992.

[12] A. Gasull and H. Giacomini, A new criterion for controlling the number of limit cycles of some generalized Liénard equations, *J. Differential Equations* Vol. 185. pp 54-73. 2002.

[13] A. Gasull, H. Giacomini and J. Llibre, New criteria for the existence and non-existence of limit cycles in Liénard differential systems, preprint, 2007.

[14] J. Guckenheimer and P. Holmes, *Nonlinear oscillations, dynamical systems, and bifurcation of vector fields*, Springer, 1983.

[15] D. Hilbert, Mathematische Probleme, Lecture, Second Internat. Congr. Math. (Paris, 1900), Nachr. Ges. Wiss. G"ttingen Math. Phys. KL. pp. 253-297. 1900. English transl., Bull. Amer. Math. Soc. Vol. 8. pp 437-479. 1902. in Bull. Amer. Math. Soc. Vol. 37. pp 407-436. 2000.

[16] Yu. Ilyashenko, *Finiteness Theorems for Limit Cycles*, Translations of Math. Monographs Vol. 94. Amer. Math. Soc., 1991.

[17] Yu. Ilyashenko and A. Panov, Some upper estimates of the number of limit cycles of planar vector fields with applications to Liénard equations, *Moscow Math. J.* Vol. 1. pp 583-599. 2001.

[18] N. Levinson and O.K. Smith, A general equation for relation oscillations, *Duke Math. J.* Vol. 9. pp 382-403. 1942.

[19] A. Lins, W. de Melo and C.C. Pugh, On Liénard's Equation, Lecture Notes in Math. Vol. 597, Springer, Berlin. pp 335–357. 1977.

[20] J. Llibre and G. Rodríguez, Configurations of limit cycles and planar polynomial vector fields, *J. of Differential Equations* Vol. 198. pp 374-380. 2004.

[21] N.G. Lloyd, Liénard systems with several limit cycles, *Math. Proc. Cambridge Philos. Soc.* Vol. 102. pp 565-572. 1987.

[22] N.G. Lloyd and S. Lynch, Small-amplitude limit cycles of certain Liénard systems, *Proc. Roy. Soc. London Ser. A* Vol. 418. pp 199-208. 1988.

[23] J.L. López and R. López-Ruiz, The limit cycles of Liénard equations in the strongly nonlinear regime, *Chaos, Solitons and Fractals* Vol. 11. pp 747-756. 2000.

[24] J.L. Massera, Sur un théorème de G. Sansone sur l'équation di Liénard (French), *Boll. Un. Mat. Ital.* Vol. 9 (3). pp 367-369. 1954.

[25] G.S. Rychkov, The maximal number of limit cycles of the system $\dot{y} = -x$, $\dot{x} = y - \sum_{i=0}^{2} a_i x^{2i+1}$ is equal to two, (Russian) *Differencial Equations* Vol. 11. pp 390-391. 1975.

[26] M. Sabatini and G. Villari, Limit cycle uniqueness for a class of planar dynamical systems, *Appl. Math. Lett.* Vol. 19. pp 1180-1184. 2006.

[27] G. Sansone, Soluzioni periodiche dell'equazione di Liénard. Calcolo del periodo (Italian), *Univ. e Politecnico Torino. Rend. Sem. Mat.* Vol. 10. pp 155-171. 1951.

[28] S. Smale, Mathematical Problems for the Next Century, *Mathematical Intelligencer* Vol. 20. pp 7-15. 1998.

[29] U. Staude, Uniqueness of periodic solutions of the Liénard equation, in: *Recent Advances in Differential Equations*, Academic Press, pp. 421-429. 1981.

[30] J.A. Sanders and F. Verhulst, *Averaging Methods in Nonlinear Dynamical Systems*, Applied Mathematical Sciences Vol. 59. Springer, 1985.

[31] F. Verhulst, *Nonlinear Differential Equations and Dynamical Systems*, Universitext, Springer, 1991.

[32] Dongmei Xiao, Zhifen Zhang, On the uniqueness and nonexistence of limit cycles for predator-prey systems, *Nonlinearity* Vol. 16. pp 1185-1201. 2003.

[33] Ye Yenqian et al., *Theory of Limit Cycles*, Transl. Math. Monogr., Vol. 66, Amer. Math. Soc., Providence, RI, 1986.

[34] Zhang Zhifen, Ding Tongren, Huang Wenzao and Dong Zhenxi, *Qualitative Theory of Differential Equations*, Translations of Math. Monographs, Vol. 101, Amer. Math. Soc, Providence, 1992.

[35] C. Zuppa, Order of cyclicity of the singular point of Liénard's polynomial vector fields, *Bol. Soc. Brasil. Mat.* Vol. 12. pp 105–111. 1981.

Bol. Soc. Esp. Mat. Apl.
nº 42(2008), 49–65

APPROXIMATIONS OF LOCAL EVOLUTION PROBLEMS BY NONLOCAL ONES

JULIO D. ROSSI

Departamento de Matemática, FCEyN
UBA (1428) Buenos Aires, Argentina.
http://mate.dm.uba.ar/∼jrossi

jrossi@dm.uba.ar

Abstract

In this article we review recent results concerning limits of solutions to nonlocal equations when a rescaling parameter goes to zero. We recover some of the most frequently used diffusion models: the heat equation with Neumann or Direchlet boundary conditions, the $p-$Laplace equation with Neumann boundary conditions and a convection-diffusion equation.

Key words: *Non-local diffusion, Newmann boundary conditions.*
AMS subject classifications: *35B40 45M05 45G10.*

1 Introduction

Let $J : \mathbb{R}^N \to \mathbb{R}$ be a nonnegative, radial, continuous function with $\int_{\mathbb{R}^N} J(z)\,dz = 1$. Assume also that J is strictly positive in $B(0,d)$ and vanishes in $\mathbb{R}^N \setminus B(0,d)$. Nonlocal evolution equations of the form

$$u_t(x,t) = (J * u - u)(x,t) = \int_{\mathbb{R}^N} J(x-y)u(y,t)\,dy - u(x,t), \tag{1}$$

and variations of it, have been recently widely used to model diffusion processes. More precisely, as stated in [32], if $u(x,t)$ is thought of as a density at the point x at time t and $J(x-y)$ is thought of as the probability distribution of jumping from location y to location x, then $\int_{\mathbb{R}^N} J(y-x)u(y,t)\,dy = (J*u)(x,t)$ is the rate at which individuals are arriving at position x from all other places and $-u(x,t) = -\int_{\mathbb{R}^N} J(y-x)u(x,t)\,dy$ is the rate at which they are leaving location x to travel to all other sites. This consideration, in the absence of external or internal sources, leads immediately to the fact that the density u

Partially supported by SIMUMAT, Generalitat Valenciana under AINV2007/03, ANPCyT PICT 5009, UBA X066 and CONICET (Argentina).

satisfies equation (1). For recent references on nonlocal diffusion see, [7], [8], [9], [17], [19], [32], and references therein.

Here we will review some recent results concerning limits of nonlocal problems when a scaling parameter (that measures the radius of influence of the nonlocal term) goes to zero. We recover in these limits some well known diffusion problems, namely, the heat equation with Neumann or Dirichlet boundary conditions, the $p-$Laplace equation with Neumann boundary conditions and a convection-diffusion equation.

We will not present the proofs but we refer to [3], [4], [18], [20], [21], [22], [23], [35] for details. We provide here the main statements of the results with references and some discussion concerning the hypotheses involved.

The paper is organized as follows: in Section 2 we deal with linear diffusion with Neumann boundary conditions, in Section 3 with Dirichlet boundary conditions, in Section 4 we face a nonlocal diffusion model analogous to the $p-$Laplacian and finally in Section 5 we present a nonlocal convection-diffusion equation.

2 A linear Neumann Problem

The purpose of this section is to show that the solutions of the usual Neumann boundary value problem for the heat equation can be approximated by solutions of a sequence of nonlocal "Neumann" boundary value problems.

Given a bounded, connected and smooth domain Ω, one of the most common boundary conditions that has been imposed in the literature to the heat equation, $u_t = \Delta u$, is the *Neumann boundary condition*, $\partial u/\partial \eta(x,t) = g(x,t)$, $x \in \partial\Omega$, which leads to the following classical problem,

$$\begin{cases} u_t - \Delta u = 0 & \text{in} \quad \Omega \times (0,T), \\ \dfrac{\partial u}{\partial \eta} = g & \text{on} \quad \partial\Omega \times (0,T), \\ u(x,0) = u_0(x) & \text{in} \quad \Omega. \end{cases} \tag{2}$$

Now we propose a nonlocal "Neumann" boundary value problem, namely

$$u_t(x,t) = \int_\Omega J(x-y)\big(u(y,t) - u(x,t)\big)\,dy + \int_{\mathbb{R}^N \setminus \Omega} G(x, x-y) g(y,t)\,dy, \tag{3}$$

where $G(x,\xi)$ is smooth and compactly supported in ξ uniformly in x.

In this model the first integral takes into account the diffusion inside Ω. In fact, as we have explained, the integral $\int J(x-y)(u(y,t)-u(x,t))\,dy$ takes into account the individuals arriving or leaving position x from or to other places. Since we are integrating in Ω, we are imposing that diffusion takes place only in Ω. The last term takes into account the prescribed flux of individuals that enter or leave the domain.

The nonlocal Neumann model (3) and the Neumann problem for the heat equation (2) share many properties. For example, a comparison principle holds

for both equations when G is nonnegative and the asymptotic behavior of their solutions as $t \to \infty$ is similar, see [22].

Existence and uniqueness of solutions of (3) with general G is proved by a fixed point argument. Also, a comparison principle when $G \geq 0$ holds. See [23].

As we have mentioned, our main goal is to show that the Neumann problem for the heat equation (2), can be approximated by suitable nonlocal Neumann problems (3). To this end, we consider the rescaled kernels

$$J_\varepsilon(\xi) = C_1 \frac{1}{\varepsilon^N} J\left(\frac{\xi}{\varepsilon}\right), \qquad G_\varepsilon(x,\xi) = C_1 \frac{1}{\varepsilon^N} G\left(x, \frac{\xi}{\varepsilon}\right) \tag{4}$$

with

$$C_1^{-1} = \frac{1}{2} \int_{B(0,d)} J(z) z_N^2 \, dz,$$

which is a normalizing constant in order to obtain the Laplacian in the limit instead of a multiple of it. Then, we consider the solution $u^\varepsilon(x,t)$ to

$$\begin{cases} u_t^\varepsilon(x,t) &= \dfrac{1}{\varepsilon^2} \displaystyle\int_\Omega J_\varepsilon(x-y)(u^\varepsilon(y,t) - u^\varepsilon(x,t))\, dy \\ & \qquad + \dfrac{1}{\varepsilon} \displaystyle\int_{\mathbb{R}^N \setminus \Omega} G_\varepsilon(x, x-y) g(y,t)\, dy, \\ u^\varepsilon(x,0) &= u_0(x). \end{cases} \tag{5}$$

We have that

$$u^\varepsilon \to u$$

in different topologies according to two different choices of the kernel G.

Let us give an heuristic idea in one space dimension, with $\Omega = (0,1)$, of why the scaling involved in (10) is the correct one. We assume that

$$\int_1^\infty G(1, 1-y)\, dy = -\int_{-\infty}^0 G(0,-y)\, dy = \int_0^1 J(y)\, y\, dy$$

and, as stated above, $G(x,\cdot)$ has compact support independent of x. In this case (5) reads

$$\begin{aligned} u_t(x,t) = \; & \frac{1}{\varepsilon^2} \int_0^1 J_\varepsilon\,(x-y)\,\big(u(y,t) - u(x,t)\big) dy \\ & + \frac{1}{\varepsilon} \int_{-\infty}^0 G_\varepsilon\,(x, x-y)\, g(y,t)\, dy \\ & + \frac{1}{\varepsilon} \int_1^{+\infty} G_\varepsilon\,(x, x-y)\, g(y,t)\, dy := \mathcal{A}_\varepsilon u(x,t). \end{aligned}$$

If $x \in (0,1)$ a Taylor expansion gives that for any fixed smooth u and ε small enough, the right hand side $\mathcal{A}_\varepsilon u$ in (5) becomes

$$\mathcal{A}_\varepsilon u(x) = \frac{1}{\varepsilon^2} \int_0^1 J_\varepsilon\,(x-y)\,(u(y) - u(x))\, dy \approx u_{xx}(x)$$

and if $x = 0$ and ε small,

$$\begin{aligned}\mathcal{A}_\varepsilon u(0) &= \frac{1}{\varepsilon^2}\int_0^1 J_\varepsilon(-y)\,(u(y)-u(0))\,dy + \frac{1}{\varepsilon}\int_{-\infty}^0 G_\varepsilon(0,-y)\,g(y)\,dy \\ &\approx \frac{C_2}{\varepsilon}(u_x(0)-g(0)).\end{aligned}$$

Analogously, $\mathcal{A}_\varepsilon u(1) \approx (C_2/\varepsilon)(-u_x(1)+g(1))$.

However, the proofs of the results are much more involved than simple Taylor expansions due to the fact that for each $\varepsilon > 0$ there are points $x \in \Omega$ for which the ball in which integration takes place, $B(x, d\varepsilon)$, is not contained in Ω. Moreover, when working in several space dimensions, one has to take into account the geometry of the domain.

Now we arrive to the precise statements of the results. First, we deal with homogeneous boundary conditions, this is, $g \equiv 0$.

Theorem 2.1 *Assume $g \equiv 0$. Let Ω be a bounded $C^{2+\alpha}$ domain for some $0 < \alpha < 1$. Let $u \in C^{2+\alpha,1+\alpha/2}(\overline{\Omega} \times [0,T])$ be the solution to* (2) *and let u^ε be the solution to* (5) *with J_ε as above. Then,*

$$\lim_{\varepsilon\to 0} \sup_{t\in[0,T]} \|u^\varepsilon(\cdot,t) - u(\cdot,t)\|_{L^\infty(\Omega)} = 0.$$

Note that this result holds for every G since $g \equiv 0$, and that the assumed regularity in u is guaranteed if $u_0 \in C^{2+\alpha}(\overline{\Omega})$ and $\partial u_0/\partial\eta = 0$, see [34].

The proof of Theorem 2.1 follows by constructing adequate super and subsolutions and then using comparison arguments to get bounds for the difference $u^\varepsilon - u$, see [23] for detalis.

Now we will make explicit the functions G we will deal with in the case $g \neq 0$. To define the first one let us introduce some notation. As before, let Ω be a bounded $C^{2+\alpha}$ domain. For $x \in \Omega_\varepsilon := \{x \in \Omega \mid \operatorname{dist}(x, \partial\Omega) < d\varepsilon\}$ and ε small enough we write $x = \bar{x} - s\,d\,\eta(\bar{x})$ where $\bar{x}$ is the orthogonal projection of x on $\partial\Omega$, $0 < s < \varepsilon$ and $\eta(\bar{x})$ is the unit exterior normal to Ω at $\bar{x}$. Under these assumptions we define

$$G_1(x,\xi) = -J(\xi)\,\eta(\bar{x})\cdot\xi \quad \text{for } x \in \Omega_\varepsilon. \tag{6}$$

Notice that the last integral in (5) only involves points $x \in \Omega_\varepsilon$ since when $y \notin \Omega$, $x - y \in supp\, J_\varepsilon$ implies that $x \in \Omega_\varepsilon$. Hence the above definition makes sense for ε small.

For this choice of the kernel, $G = G_1$, we have the following result.

Theorem 2.2 *Let Ω be a bounded $C^{2+\alpha}$ domain, $g \in C^{1+\alpha,(1+\alpha)/2}(\overline{(\mathbb{R}^N \setminus \Omega)} \times [0,T])$, $u \in C^{2+\alpha,1+\alpha/2}(\overline{\Omega} \times [0,T])$ the solution to* (2)*, for some $0 < \alpha < 1$. Let J as before and $G(x,\xi) = G_1(x,\xi)$, where G_1 is defined by* (6)*. Let u^ε be the solution to* (5)*. Then,*

$$\lim_{\varepsilon\to 0} \sup_{t\in[0,T]} \|u^\varepsilon(\cdot,t) - u(\cdot,t)\|_{L^1(\Omega)} \to 0.$$

Observe that G_1 may fail to be nonnegative and hence a comparison principle may not hold. However, in this case our proof of convergence to the solution of the heat equation does not rely on comparison arguments for (3). If we want a nonnegative kernel G, in order to have a comparison principle, we can modify $(G_1)_\varepsilon$ by taking

$$(\tilde{G}_1)_\varepsilon(x,\xi) = (G_1)_\varepsilon(x,\xi) + \kappa\varepsilon J_\varepsilon(\xi) = \frac{1}{\varepsilon} J_\varepsilon(\xi)\left(-\eta(\bar{x})\cdot\xi + \kappa\varepsilon^2\right)$$

instead.

Note that for $x \in \overline{\Omega}$ and $y \in \mathbb{R}^N \setminus \Omega$, $(\tilde{G}_1)_\varepsilon(x, x-y) = \frac{1}{\varepsilon} J_\varepsilon(x-y)\left(-\eta(\bar{x})\cdot(x-y) + \kappa\varepsilon^2\right)$ is nonnegative for ε small if we choose the constant κ as a bound for the curvature of $\partial\Omega$, since $|x-y| \leq d\,\varepsilon$. We observe that Theorem 2.2 remains valid with $(G_1)_\varepsilon$ replaced by $(\tilde{G}_1)_\varepsilon$.

Finally, the other "Neumann" kernel we propose is

$$G(x,\xi) = G_2(x,\xi) = C_2 J(\xi),$$

where C_2 is such that

$$\int_0^d \int_{\{z_N > s\}} J(z)\big(C_2 - z_N\big)\,dz\,ds = 0. \tag{7}$$

This choice of G is natural since we are considering a flux with a jumping probability that is a scalar multiple of the same jumping probability that moves things in the interior of the domain, J.

Several properties of solutions to (3) have been recently investigated in [22] in the case $G = G_2$ for different choices of g.

For the case of G_2 we can still prove convergence but in a weaker sense.

Theorem 2.3 *Let Ω be a bounded $C^{2+\alpha}$ domain, $g \in C^{1+\alpha,(1+\alpha)/2}(\overline{(\mathbb{R}^N \setminus \Omega)} \times [0,T])$, $u \in C^{2+\alpha,1+\alpha/2}(\overline{\Omega} \times [0,T])$ the solution to (2), for some $0 < \alpha < 1$. Let J as before and $G(x,\xi) = G_2(x,\xi) = C_2 J(\xi)$, where C_2 is defined by (7). Let u^ε be the solution to (5). Then, for each $t \in [0,T]$*

$$u_\varepsilon(x,t) \rightharpoonup u(x,t) \quad *-\text{weakly in } L^\infty(\Omega)$$

as $\varepsilon \to 0$.

3 A linear Dirichlet Problem

Following [21], now we propose the following nonlocal "Dirichlet" boundary value problem: Given $g(x,t)$ defined for $x \in \mathbb{R}^N \setminus \Omega$ and $u_0(x)$ defined for $x \in \Omega$, find $u(x,t)$ such that

$$\begin{cases} u_t(x,t) = \displaystyle\int_{\mathbb{R}^N} J(x-y)(u(y,t) - u(x,t))dy, & x \in \Omega,\ t > 0, \\ u(x,t) = g(x,t), & x \notin \Omega,\ t > 0, \\ u(x,0) = u_0(x), & x \in \Omega. \end{cases} \tag{8}$$

In this model we prescribe the values of u outside Ω which is the analogous of prescribing the so called Dirichlet boundary conditions for the classical heat equation. However, the boundary data is not understood in the usual sense, see [18]. As explained before in this model the right hand side models the diffusion, the integral $\int J(x-y)(u(y,t)-u(x,t))\,dy$ takes into account the individuals arriving or leaving position $x \in \Omega$ from or to other places while we are prescribing the values of u outside the domain Ω by imposing $u = g$ for $x \notin \Omega$. When $g = 0$ we get that any individuals that leave Ω die, this is the case when Ω is surrounded by a hostile environment.

Existence and uniqueness of solutions of (8) is proved by a fixed point argument and also a comparison principle holds for this problem.

Let us consider the classical Dirichlet problem for the heat equation,

$$\begin{cases} v_t(x,t) - \Delta v(x,t) = 0, & x \in \Omega,\ t > 0, \\ v(x,t) = g(x,t), & x \in \partial\Omega,\ t > 0, \\ v(x,0) = u_0(x), & x \in \Omega. \end{cases} \tag{9}$$

The nonlocal Dirichlet model (8) and the classical Dirichlet problem (9) share many properties, among them the asymptotic behavior of their solutions as $t \to \infty$ is similar as was proved in [18].

The main goal of this section is to show that the Dirichlet problem for the heat equation (9) can be approximated by suitable nonlocal problems of the form of (8). More precisely, for a given J and a given $\varepsilon > 0$ we consider the rescaled kernel

$$J_\varepsilon(\xi) = C_1 \frac{1}{\varepsilon^N} J\left(\frac{\xi}{\varepsilon}\right), \quad \text{with} \quad C_1^{-1} = \frac{1}{2}\int_{B(0,d)} J(z) z_N^2\,dz. \tag{10}$$

Here C_1 is a normalizing constant in order to obtain the Laplacian in the limit instead of a multiple of it. Let $u^\varepsilon(x,t)$ be the solution of

$$\begin{cases} u_t^\varepsilon(x,t) = \displaystyle\int_\Omega \frac{J_\varepsilon(x-y)}{\varepsilon^2}(u^\varepsilon(y,t) - u^\varepsilon(x,t))dy, & x \in \Omega, t > 0, \\ u(x,t) = g(x,t), & x \notin \Omega, t > 0, \\ u(x,0) = u_0(x), & x \in \Omega. \end{cases} \tag{11}$$

Our main result now reads as follows.

Theorem 3.1 *Let Ω be a bounded $C^{2+\alpha}$ domain for some $0 < \alpha < 1$. Let $v \in C^{2+\alpha,1+\alpha/2}(\overline{\Omega} \times [0,T])$ be the solution to (9) and let u^ε be the solution to (11) with J_ε as above. Then, there exists $C = C(T)$ such that*

$$\sup_{t\in[0,T]} \|v - u^\varepsilon\|_{L^\infty(\Omega)} \le C\varepsilon^\alpha \to 0, \qquad \text{as } \varepsilon \to 0. \tag{12}$$

Note that the assumed regularity of v is a consequence of regularity assumptions on the boundary data g, the domain Ω and the initial condition u_0, see [34].

4 A non-local $p-$Laplacian with Neumann boundary conditions

Our main goal in this section is to study the following nonlocal nonlinear diffusion problem, which we call the *nonlocal p-Laplacian problem* (with homogeneous Neumann boundary conditions),

$$P_p^J(u_0) \quad \begin{cases} u_t(t,x) = \displaystyle\int_\Omega J(x-y)|u(t,y)-u(t,x)|^{p-2}(u(t,y)-u(t,x))\,dy, \\ u(x,0) = u_0(x). \end{cases}$$

Here $1 \le p < +\infty$ and $\Omega \subset \mathbb{R}^N$ is a bounded domain.

When dealing with local evolution equations, two models of nonlinear diffusion has been extensively studied in the literature, the porous medium equation, $u_t = \Delta u^m$, and the $p-$Laplacian evolution, $u_t = \operatorname{div}(|\nabla u|^{p-2}\nabla u)$. In the first case (for the porous medium equation) a nonlocal analogous equation was studied in [3] (see also [20]). Now we want to study the nonlocal equation P_p^J, that is, the nonlocal analogous to the $p-$Laplacian evolution.

As in the previous section, since we are integrating only in Ω, problem $P_p^J(u_0)$ has to be seen as a problem with homogeneous Neumann boundary condition. For the linear case, $p = 2$, see the previous section and [23], [22].

We will see in this section that solutions to problem $P_p^J(u_0)$ converge to the solution of the classical $p-$Laplacian if $p > 1$ and to the total variation flow when $p = 1$ with Neumann boundary conditions when the convolution kernel J is also rescaled in a suitable way.

First, let us state the precise definition of a solution. Solutions to $P_p^J(u_0)$ will be understood in the following sense.

Definition 4.1 *Let $1 < p < +\infty$. A solution of $P_p^J(u_0)$ in $[0,T]$ is a function $u \in C([0,T];L^1(\Omega)) \cap W^{1,1}(]0,T[;L^1(\Omega))$ which satisfies $u(0,x) = u_0(x)$ a.e. $x \in \Omega$ and*

$$u_t(t,x) = \int_\Omega J(x-y)|u(y,t)-u(x,t)|^{p-2}(u(y,t)-u(x,t))\,dy \quad \text{a.e in }]0,T[\times\Omega.$$

Let us note that, with this definition of solution, the evolution problem $P_p^J(u_0)$ is the gradient flow associated to the functional

$$J_p(u) = \frac{1}{2p}\int_\Omega\int_\Omega J(x-y)|u(y)-u(x)|^p\,dy\,dx,$$

which is the nonlocal analogous to the energy functional associated to the $p-$Laplacian

$$F_p(u) = \frac{1}{p}\int_\Omega |\nabla u(y)|^p\,dy.$$

Our first result shows existence and uniqueness of a global solution for this problem. Moreover, a contraction principle holds.

Theorem 4.1 *Assume $p > 1$ and let $u_0 \in L^p(\Omega)$. Then, there exists a unique solution to $P_p^J(u_0)$ in the sense of Definition 4.1.*

Moreover, if $u_{i0} \in L^1(\Omega)$, $i = 1, 2$, and u_i is a solution in $[0, T]$ of $P_p^J(u_{i0})$. Then

$$\int_\Omega (u_1(t) - u_2(t))^+ \leq \int_\Omega (u_{10} - u_{20})^+ \quad \text{for every } t \in]0, T[.$$

If $u_{i0} \in L^p(\Omega)$, $i = 1, 2$, then

$$\|u_1(t) - u_2(t)\|_{L^p(\Omega)} \leq \|u_{10} - u_{20}\|_{L^p(\Omega)} \quad \text{for every } t \in]0, T[.$$

Let us now deal with existence and uniqueness for the extreme case $p = 1$. We have that the formal evolution problem

$$u_t(t, x) = \int_\Omega J(x - y) \frac{u(t, y) - u(t, x)}{|u(t, y) - u(t, x)|}\, dy,$$

is the gradient flow associated to the functional

$$J_1(u) = \frac{1}{2} \int_\Omega \int_\Omega J(x - y)|u(y) - u(x)|\, dy\, dx,$$

which is the nonlocal analogous to the energy functional associated to the total variation

$$F_1(u) = \int_\Omega |\nabla u(y)|\, dy.$$

For $p = 1$ we give the following definition of what we understand as a solution.

Definition 4.2 *A solution of $P_1^J(z_0)$ in $[0, T]$ is a function*

$$u \in C([0, T]; L^1(\Omega)) \cap W^{1,1}(]0, T[; L^1(\Omega))$$

which satisfies $u(0, x) = u_0(x)$ a.e. $x \in \Omega$ and

$$u_t(t, x) = \int_\Omega J(x - y) g(t, x, y)\, dy \quad \text{a.e in }]0, T[\times\Omega,$$

for some $g \in L^\infty(0, T; L^\infty(\Omega \times \Omega))$ with $\|g\|_\infty \leq 1$ such that $g(t, x, y) = -g(t, y, x)$ and

$$J(x - y)g(t, x, y) \in J(x - y)sign(u(t, y) - u(t, x)).$$

To get existence and uniqueness of these kind of solutions, the idea is to take the limit as $p \searrow 1$ of solutions to P_p^J with $p > 1$.

Theorem 4.2 *Assume $p = 1$ and let $u_0 \in L^1(\Omega)$. Then, there exists a unique solution to $P_1^J(u_0)$ in the sense of Definition 4.2.*

Moreover, for $i = 1, 2$, let $u_{i0} \in L^1(\Omega)$ and u_i be a solution in $[0, T]$ of $P_1^J(u_{i0})$. Then

$$\int_\Omega (u_1(t) - u_2(t))^+ \leq \int_\Omega (u_{10} - u_{20})^+ \quad \text{for almost every } t \in]0, T[.$$

Our next step is to rescale the kernel J appropriately and take the limit as the scaling parameter goes to zero. To be more precise, for every $p \geq 1$, we consider the local $p-$Laplace evolution equation with homogeneous Neumann boundary conditions

$$N_p(u_0) \quad \begin{cases} u_t = \Delta_p u & \text{in } (0,T)\times\Omega, \\ |\nabla u|^{p-2}\nabla u \cdot \eta = 0 & \text{on } (0,T)\times\partial\Omega, \\ u(x,0) = u_0(x) & \text{in } \Omega, \end{cases}$$

where η is the unit outward normal on $\partial\Omega$, $\Delta_p u = \operatorname{div}(|\nabla u|^{p-2}\nabla u)$ is the p-laplacian of u. We obtain that the solutions of this local problem, $N_p(u_0)$, can be approximated by solutions of a sequence of nonlocal *p-Laplacian* problems of the form P_p^J.

Problem $N_1(u_0)$, that is, the Neumann problem for the Total Variation Flow, was studied in [1] (see also [2]), motivated by problems in image processing. This PDE appears when one uses the steepest descent method to minimize the Total Variation, a method introduced by L. Rudin, S. Osher and E. Fatemi [38] in the context of image denoising and reconstruction. Then, solving $N_1(u_0)$ amounts to regularize or, in other words, to filter the initial datum u_0. This filtering process has less destructive effect on the edges than filtering with a Gaussian, i.e., than solving the heat equation with initial condition u_0. In this context the given *image* u_0 is a function defined on a bounded, smooth or piecewise smooth open subset Ω of $\mathbb{R}^N$, typically, Ω will be a rectangle in $\mathbb{R}^2$.

S. Kindermann, S. Osher and P. W. Jones in [36] have studied deblurring and denoising of images by nonlocal functionals, motivated by the use of neighborhood filters [16]. Such filters have originally been proposed by Yaroslavsky, [43], [44], and further generalized by C. Tomasi and R. Manduchi, [42], as bilateral filter. The main aim of [36] is to relate the neighborhood filter to an energy minimization. Now in this case the Euler-Lagrange equations are not partial differential equations but include integrals. The functional considered in [36] takes the general form

$$J_g(u) = \int_{\Omega\times\Omega} g\left(\frac{|u(x)-u(y)|^2}{h^2}\right) w(|x-y|)\,dxdy, \tag{13}$$

with $w \in L^\infty(\Omega)$, $g \in C^1(\mathbb{R}^+)$ and $h > 0$ is a parameter. The Fréchet derivative of J_g as a functional from $L^2(\Omega)$ into $\mathbb{R}$ is given by

$$J_g'(u)(x) = \frac{4}{h^2}\int_\Omega g'\left(\frac{|u(x)-u(y)|^2}{h^2}\right)(u(x)-u(y))w(|x-y|)\,dy.$$

Note that the nonlocal functional J_p is of the form (13) with $g(t) = \frac{1}{2p}|t|^{\frac{p}{2}}$, $w = J$ and $h = 1$. Then, problem $P_p^J(u_0)$ appears when one uses the steepest descent method to minimize this particular nonlocal functional.

For given $p \geq 1$ and J we consider the rescaled kernels

$$J_{p,\varepsilon}(x) := \frac{C_{J,p}}{\varepsilon^{p+N}} J\left(\frac{x}{\varepsilon}\right),$$

where

$$C_{J,p}^{-1} := \frac{1}{2}\int_{\mathbb{R}^N} J(z)|z_N|^p\,dz$$

is a normalizing constant in order to obtain the p-Laplacian in the limit instead a multiple of it.

Associated with these rescaled kernels we have solutions u_ε to the equation in P_p^J with J replaced by $J_{p,\varepsilon}$ and the same initial condition u_0 (we shall call this problem $P_p^{J_{p,\varepsilon}}$). Our next result states that these functions u_ε converge strongly in $L^p(\Omega)$ to the solution to the local $p-$Laplacian $N_p(u_0)$.

Theorem 4.3 *Let Ω be a smooth bounded domain in $\mathbb{R}^N$ and $p \geq 1$. Assume $J(x) \geq J(y)$ if $|x| \leq |y|$. Let $T > 0$, $u_0 \in L^p(\Omega)$ and u_ε the unique solution of $P_p^{J_{p,\varepsilon}}(u_0)$. Then, if u is the unique solution of $N_p(u_0)$,*

$$\lim_{\varepsilon\to 0}\sup_{t\in[0,T]}\|u_\varepsilon(t,.)-u(t,.)\|_{L^p(\Omega)} = 0.$$

Note that the above result states that P_p^J is a nonlocal analogous to the $p-$Laplacian.

In order to study the asymptotic behaviour as $t \to \infty$ of the solutions of the nonlocal problems, we first prove a Poincaré's type inequality. This inequality allows to show that the solutions of the nonlocal problems converge to the mean value of the initial condition.

Theorem 4.4 *Let $p \geq 1$ and $u_0 \in L^\infty(\Omega)$. Let u be the solution to $P_p^J(u_0)$, then*

$$\|u(t)-\overline{u_0}\|_{L^p(\Omega)} \leq \left(\frac{||u_0||^2_{L^2(\Omega)}}{t}\right)^{1/p} \to 0, \qquad \text{as } t\to\infty,$$

where $\overline{u_0}$ is the mean value of the initial condition, $\overline{u_0} = \frac{1}{|\Omega|}\int_\Omega u_0(x)\,dx$.

5 A Non-local convection diffusion equation

In this section we analyze a nonlocal equation that takes into account convective and diffusive effects. We deal with the nonlocal evolution equation

$$\begin{cases} u_t(t,x) = (J*u-u)(t,x) + (G*(f(u))-f(u))(t,x), & t>0,\ x\in\mathbb{R}^N,\\ u(0,x) = u_0(x), & x\in\mathbb{R}^N. \end{cases} \tag{14}$$

Let us state first our basic assumptions. The functions J and G are nonnegatives and verify $\int_{\mathbb{R}^N} J(x)dx = \int_{\mathbb{R}^N} G(x)dx = 1$. Moreover, we consider J smooth, $J \in \mathcal{S}(\mathbb{R}^N)$, the space of rapidly decreasing functions, and radially symmetric and G smooth, $G \in \mathcal{S}(\mathbb{R}^N)$, but not necessarily symmetric. To obtain a diffusion operator similar to the Laplacian we impose in addition that J verifies

$$\frac{1}{2}\partial^2_{\xi_i\xi_i}\widehat{J}(0) = \frac{1}{2}\int_{supp(J)} J(z)z_i^2\,dz = 1.$$

This implies that

$$\widehat{J}(\xi) - 1 + \xi^2 \sim |\xi|^3, \qquad \text{for } \xi \text{ close to } 0.$$

Here $\widehat{J}$ is the Fourier transform of J and the notation $A \sim B$ means that there exist constants C_1 and C_2 such that $C_1 A \leq B \leq C_2 A$. We can consider more general kernels J with expansions in Fourier variables of the form $\widehat{J}(\xi) - 1 + A\,\xi^2 \sim |\xi|^3$. Since the results (and the proofs) are almost the same, we do not include the details for this more general case, but we comment on how the results are modified by the appearance of A.

The nonlinearity f will de assumed nondecreasing with $f(0) = 0$ and locally Lipschitz continuous (a typical example that we will consider below is $f(u) = |u|^{q-1}u$ with $q > 1$).

Equations like $w_t = J * w - w$ and variations of it, have been recently widely used to model diffusion processes, for example, in biology, dislocations dynamics, etc. See, for example, [9], [17], [19], [20], [32], [33].

In our case, see the equation in (14), we have a diffusion operator $J * u - u$ and a nonlinear convective part given by $G * (f(u)) - f(u)$. Concerning this last term, if G is not symmetric then individuals have greater probability of jumping in one direction than in others, provoking a convective effect.

First, we prove existence, uniqueness and well-possedness of a solution with an initial condition $u(0, x) = u_0(x) \in L^1(\mathbb{R}^N) \cap L^\infty(\mathbb{R}^N)$.

Theorem 5.1 *For any $u_0 \in L^1(\mathbb{R}^N) \cap L^\infty(\mathbb{R}^N)$ there exists a unique global solution*

$$u \in C([0,\infty); L^1(\mathbb{R}^N)) \cap L^\infty([0,\infty); \mathbb{R}^N).$$

If u and v are solutions of (14) corresponding to initial data $u_0, v_0 \in L^1(\mathbb{R}^N) \cap L^\infty(\mathbb{R}^N)$ respectively, then the following contraction property

$$\|u(t) - v(t)\|_{L^1(\mathbb{R}^N)} \leq \|u_0 - v_0\|_{L^1(\mathbb{R}^N)}$$

holds for any $t \geq 0$. In addition,

$$\|u(t)\|_{L^\infty(\mathbb{R}^N)} \leq \|u_0\|_{L^\infty(\mathbb{R}^N)}.$$

We have to emphasize that a lack of regularizing effect occurs. This has been already observed in [18] for the linear problem $w_t = J * w - w$. In [27], the authors prove that the solutions to the local convection-diffusion problem, $u_t = \Delta u + b \cdot \nabla f(u)$, satisfy an estimate of the form $\|u(t)\|_{L^\infty(\mathbb{R}^N)} \leq C(\|u_0\|_{L^1(\mathbb{R}^N)})\, t^{-d/2}$ for any initial data $u_0 \in L^1(\mathbb{R}^N) \cap L^\infty(\mathbb{R}^N)$. In our nonlocal model, we cannot prove such type of inequality independently of the $L^\infty(\mathbb{R}^N)$-norm of the initial data. Moreover, in the one-dimensional case with a suitable bound on the nonlinearity that appears in the convective part, f, we can prove that such an inequality does not hold in general, see Section 3. In addition, the $L^1(\mathbb{R}^N) - L^\infty(\mathbb{R}^N)$ regularizing effect is not available for the linear equation, $w_t = J * w - w$, see [18].

Concerning (14) we can obtain a solution to a standard convection-diffusion equation

$$v_t(t,x) = \Delta v(t,x) + b \cdot \nabla f(v)(t,x), \qquad t > 0,\, x \in \mathbb{R}^N, \tag{15}$$

as the limit of solutions to (14) when a scaling parameter goes to zero. In fact, let us consider

$$J_\varepsilon(s) = \frac{1}{\varepsilon^N} J\left(\frac{s}{\varepsilon}\right), \qquad G_\varepsilon(s) = \frac{1}{\varepsilon^N} G\left(\frac{s}{\varepsilon}\right),$$

and the solution $u_\varepsilon(t,x)$ to our convection-diffusion problem rescaled adequately,

$$\begin{cases} (u_\varepsilon)_t(t,x) &= \dfrac{1}{\varepsilon^2}\displaystyle\int_{\mathbb{R}^N} J_\varepsilon(x-y)(u_\varepsilon(t,y) - u_\varepsilon(t,x))\,dy \\ &\quad + \dfrac{1}{\varepsilon}\displaystyle\int_{\mathbb{R}^N} G_\varepsilon(x-y)(f(u_\varepsilon(t,y)) - f(u_\varepsilon(t,x)))\,dy, \\ u_\varepsilon(x,0) &= u_0(x). \end{cases} \tag{16}$$

Remark that the scaling is different for the diffusive part of the equation $J * u - u$ and for the convective part $G * f(u) - f(u)$. The same different scaling properties can be observed for the local terms Δu and $b \cdot \nabla f(u)$.

Theorem 5.2 *With the above notations, for any $T > 0$, we have*

$$\lim_{\varepsilon \to 0} \sup_{t \in [0,T]} \|u_\varepsilon - v\|_{L^2(\mathbb{R}^N)} = 0,$$

where $v(t,x)$ is the unique solution to the local convection-diffusion problem (15) *with initial condition $v(x,0) = u_0(x) \in L^1(\mathbb{R}^N) \cap L^\infty(\mathbb{R}^N)$ and $b = (b_1, ..., b_d)$ given by*

$$b_j = \int_{\mathbb{R}^N} x_j\, G(x)\,dx, \qquad j = 1, ..., d.$$

This result justifies the use of the name "nonlocal convection-diffusion problem" when we refer to (14).

From our hypotheses on J and G it follows that they verify $|\widehat{G}(\xi) - 1 - ib\cdot\xi| \leq C|\xi|^2$ and $|\widehat{J}(\xi) - 1 + \xi^2| \leq C|\xi|^3$ for every $\xi \in \mathbb{R}^N$. These bounds are exactly what we are using in the proof of this convergence result.

Remark that when G is symmetric then $b = 0$ and we obtain the heat equation in the limit. Of course the most interesting case is when $b \neq 0$ (this happens when G is not symmetric). Also we note that the conclusion of the theorem holds for other $L^p(\mathbb{R}^N)$-norms besides $L^2(\mathbb{R}^N)$, however the proof is more involved.

We can consider kernels J such that

$$A = \frac{1}{2}\int_{supp(J)} J(z) z_i^2\,dz \neq 1.$$

This gives the expansion $\widehat{J}(\xi) - 1 + A\xi^2 \sim |\xi|^3$, for ξ close to 0. In this case we will arrive to a convection-diffusion equation with a multiple of the Laplacian as the diffusion operator, $v_t = A\Delta v + b \cdot \nabla f(v)$.

Next, we want to study the asymptotic behaviour as $t \to \infty$ of solutions to (14). To this end we first analyze the decay of solutions taking into account only the diffusive part (the linear part) of the equation. These solutions have a similar decay rate as the one that holds for the heat equation, see [18] where the Fourier transform play a key role. Using similar techniques we can prove the following result that deals with this asymptotic decay rate.

Theorem 5.3 *Let $p \in [1, \infty]$. For any $u_0 \in L^1(\mathbb{R}^N) \cap L^\infty(\mathbb{R}^N)$ the solution $w(t, x)$ of the linear problem*

$$\begin{cases} w_t(t,x) = (J * w - w)(t,x), & t > 0,\ x \in \mathbb{R}^N, \\ u(0,x) = u_0(x), & x \in \mathbb{R}^N, \end{cases} \tag{17}$$

satisfies the decay estimate

$$\|w(t)\|_{L^p(\mathbb{R}^N)} \le C(\|u_0\|_{L^1(\mathbb{R}^N)}, \|u_0\|_{L^\infty(\mathbb{R}^N)})\, \langle t \rangle^{-\frac{N}{2}(1-\frac{1}{p})}.$$

We use the notation $A \le \langle t \rangle^{-\alpha}$ to denote $A \le (1+t)^{-\alpha}$.

Now we are ready to face the study of the asymptotic behaviour of the complete problem (14). To this end we have to impose some grow condition on f. Therefore, in the sequel we restrict ourselves to nonlinearities f that are pure powers

$$f(u) = |u|^{q-1}u \tag{18}$$

with $q > 1$.

The analysis is more involved than the one performed for the linear part and we cannot use here the Fourier transform directly (of course, by the presence of the nonlinear term). Our strategy is to write a variation of constants formula for the solution and then prove estimates that say that the nonlinear part decay faster than the linear one. For the local convection diffusion equation this analysis was performed by Escobedo and Zuazua in [27]. However, in the previously mentioned reference energy estimates were used together with Sobolev inequalities to obtain decay bounds. These Sobolev inequalities are not available for the nonlocal model, since the linear part does not have any regularizing effect. Therefore, we have to avoid the use of energy estimates and tackle the problem using a variant of the Fourier splitting method proposed by Schonbek to deal with local problems, see [39], [40] and [41].

We state our result concerning the asymptotic behaviour (decay rate) of the complete nonlocal model as follows:

Theorem 5.4 *Let f satisfies (18) with $q > 1$ and $u_0 \in L^1(\mathbb{R}^N) \cap L^\infty(\mathbb{R}^N)$. Then, for every $p \in [1, \infty)$ the solution u of equation (14) verifies*

$$\|u(t)\|_{L^p(\mathbb{R}^N)} \le C(\|u_0\|_{L^1(\mathbb{R}^N)}, \|u_0\|_{L^\infty(\mathbb{R}^N)})\, \langle t \rangle^{-\frac{N}{2}(1-\frac{1}{p})}. \tag{19}$$

Finally, we look at the first order term in the asymptotic expansion of the solution. For $q > (d+1)/d$, we find that this leading order term is the same as the one that appears in the linear local heat equation. This is due to the fact that the nonlinear convection is of higher order and that the radially symmetric diffusion leads to gaussian kernels in the asymptotic regime, see [18].

Theorem 5.5 *Let f satisfies* (18) *with $q > (d+1)/d$ and let the initial condition u_0 belongs to $L^1(\mathbb{R}^N, 1+|x|) \cap L^\infty(\mathbb{R}^N)$. For any $p \in [2, \infty)$ the following holds*

$$t^{-\frac{N}{2}(1-\frac{1}{p})} \|u(t) - MH(t)\|_{L^p(\mathbb{R}^N)} \leq C(J, G, p, d)\, \alpha_q(t),$$

where $M = \int_{\mathbb{R}^N} u_0(x)\, dx$, $H(t)$ is the Gaussian,

$$H(t) = \frac{e^{-\frac{x^2}{4t}}}{(2\pi t)^{\frac{N}{2}}},$$

and

$$\alpha_q(t) = \begin{cases} \langle t \rangle^{-\frac{1}{2}} & \text{if} \quad q \geq (N+2)/N, \\ \langle t \rangle^{\frac{1-N(q-1)}{2}} & \text{if} \quad (N+1)/N < q < (N+2)/N. \end{cases}$$

Remark that we prove a weak nonlinear behaviour, in fact the decay rate and the first order term in the expansion are the same that appear in the linear model $w_t = J * w - w$, see [18].

As before, recall that our hypotheses on J imply that $\widehat{J}(\xi) - (1 - |\xi|^2) \sim B|\xi|^3$, for ξ close to 0. This is the key property of J used in the proof of Theorem 5.5. We note that when we have an expansion of the form $\widehat{J}(\xi) - (1 - A|\xi|^2) \sim B|\xi|^3$, for $\xi \sim 0$, we get as first order term a Gaussian profile of the form $H_A(t) = H(At)$.

Also note that $q = (d+1)/d$ is a critical exponent for the local convection-diffusion problem, $v_t = \Delta v + b \cdot \nabla(v^q)$, see [27]. When q is supercritical, $q > (N+1)/N$, for the local equation it also holds an asymptotic simplification to the heat semigroup as $t \to \infty$.

The first order term in the asymptotic behaviour for critical or subcritical exponents $1 < q \leq (N+1)/N$ is open. One of the main difficulties that one has to face here is the absence of a self-similar profile due to the inhomogeneous behaviour of the convolution kernels.

6 Acknowledgements

Part of this work was performed during a visit of the author to Univ. de Valencia and to Univ. Autonoma of Madrid. He wants to thank for the warm hospitality and the stimulating working atmosphere found in those places. Also, the author wants to thank to F. Andreu, J. Mazon, J. Toledo, L. Ignat, C. Cortazar and M. Elgueta for the enthusiasm and friendship working together.

References

[1] F. Andreu, C. Ballester, V. Caselles and J. M. Mazon, Minimizing Total Variation Flow, *Diff. and Int. Eq.*, **14**, (2001), 321–360.

[2] F. Andreu, V. Caselles, and J.M. Mazon, Parabolic Quasilinear Equations Minimizing Linear Growth Functionals, Progress in Mathematics, vol. 223, 2004. Birkhauser.

[3] F. Andreu, J. M. Mazón, J. D. Rossi and J. Toledo. The Neumann problem for nonlocal nonlinear diffusion equations. *Preprint*.

[4] F. Andreu, J. M. Mazón, J. D. Rossi and J. Toledo. A nonlocal $p-$Laplacian evolution equation with Neumann boundary conditions. *Preprint*.

[5] G. Aronsson, L. C. Evans and Y. Wu. Fast/slow diffusion and growing sandpiles. *J. Differential Equations*, 131 (1996), 304–335.

[6] H. Attouch. Familles d'opérateurs maximaux monotones et mesurabilité. *Ann. Mat. Pura Appl.*, **120** (1979), 35–111.

[7] P. Bates and A. Chmaj. An integrodifferential model for phase transitions: stationary solutions in higher dimensions. *J. Statistical Phys.*, **95**, (1999), 1119–1139.

[8] P. Bates and A. Chmaj. A discrete convolution model for phase transitions. *Arch. Rat. Mech. Anal.*, **150**, (1999), 281–305.

[9] P. Bates, P. Fife, X. Ren and X. Wang. Travelling waves in a convolution model for phase transitions. *Arch. Rat. Mech. Anal.*, **138**, (1997), 105–136.

[10] Ph. Benilan and M.G. Crandall. Completely Accretive Operators, in Semigroups Theory and Evolution Equations, Ph. Clement et al. editors, Marcel Dekker, 1991, pp. 41–76.

[11] Ph. Bénilan, M. G. Crandall and A. Pazy. Evolution Equations Governed by Accretive Operators. Book to appear.

[12] Ph. Bénilan, L.C. Evans and R.F. Gariepy. On some singular limits of homogeneous semigroups. *J. Evol. Equ.*, **3**, (2003), 203–214.

[13] H. Brezis. Équations et inéquations non linéaires dans les espaces vectoriels en dualité. *Ann. Inst. Fourier*, **18**, (1968), 115–175.

[14] H. Brezis. Opérateur Maximaux Monotones et Semi-groupes de Contractions dans les Espaces de Hilbert. North-Holland, 1973.

[15] H. Brezis and A. Pazy. Convergence and approximation of semigroups of nonlinear operators in Banach spaces. *J. Functional Analysis*, **9** (1972), 63–74.

[16] A. Buades, B. Coll and J. M. Morel. Neighborhood filters and PDE's. *Numer. Math.* **150** (2006), 1–34.

[17] C. Carrillo and P. Fife. Spatial effects in discrete generation population models. *J. Math. Biol.*, **50(2)**, (2005), 161–188.

[18] E. Chasseigne, M. Chaves and J. D. Rossi. Asymptotic behaviour for nonlocal diffusion equations. *J. Math. Pures Appl.*, **86**, (2006), 271–291.

[19] X. Chen. Existence, uniqueness and asymptotic stability of travelling waves in nonlocal evolution equations. *Adv. Differential Equations*, **2**, (1997), 125–160.

[20] C. Cortazar, M. Elgueta and J. D. Rossi. A non-local diffusion equation whose solutions develop a free boundary. *Annales Henri Poincaré*, **6(2)**, (2005), 269–281.

[21] C. Cortazar, M. Elgueta and J. D. Rossi. Nonlocal diffusion problems that approximate the heat equation with Dirichlet boundary conditions. To appear in *Israel J. Math.*

[22] C. Cortazar, M. Elgueta, J. D. Rossi and N. Wolanski. Boundary fluxes for non-local diffusion. *J. Differential Equations*, **234**, (2007), 360–390.

[23] C. Cortazar, M. Elgueta, J. D. Rossi and N. Wolanski. How to approximate the heat equation with Neumann boundary conditions by nonlocal diffusion problems. To appear in *Arch. Rat. Mech. Anal.*

[24] M. G. Crandall. An introduction to evolution governed by accretive operators. In *Dynamical systems (Proc. Internat. Sympos., Brown Univ., Providence, R.I., 1974), Vol. I*, pages 131–165. Academic Press, New York, 1976.

[25] M. G. Crandall. Nonlinear Semigroups and Evolution Governed by Accretive Operators. In *Proc. of Sympos. in Pure Mathematics, Part I, Vol. 45 (F. Browder ed.)*. A.M.S., Providence 1986, pages 305–338.

[26] HI. Ekeland and R. Temam. Convex Analysis and Variational Problems. North-Holland, 1972.

[27] M. Escobedo and E. Zuazua, Large time behavior for convection-diffusion equations in $\mathbb{R}^N$, *J. Funct. Anal.*, 100(1), (1991), 119–161.

[28] L. C. Evans. Partial differential equations and Monge-Kantorovich mass transfer. Current developments in mathematics, 1997 (Cambridge, MA), 65–126, Int. Press, Boston, MA, 1999.

[29] L. C. Evans, M. Feldman and R. F. Gariepy. Fast/slow diffusion and collapsing sandpiles. *J. Differential Equations*, **137** (1997), 166–209.

[30] L. C. Evans and Fr. Rezakhanlou. A stochastic model for growing sandpiles and its continuum limit. *Comm. Math. Phys.*, **197** (1998), 325–345.

[31] M. Feldman. Growth of a sandpile around an obstacle. Monge Ampère equation: applications to geometry and optimization (Deerfield Beach, FL, 1997), 55–78, Contemp. Math., 226, Amer. Math. Soc., Providence, RI, 1999.

[32] P. Fife. Some nonclassical trends in parabolic and parabolic-like evolutions. Trends in nonlinear analysis, 153–191, Springer, Berlin, 2003.

[33] P. Fife and X. Wang. A convolution model for interfacial motion: the generation and propagation of internal layers in higher space dimensions. *Adv. Differential Equations*, **3(1)**, (1998), 85–110.

[34] A. Friedman. "Partial Differential Equations of Parabolic Type". Prentice-Hall, Englewood Cliffs, NJ, 1964.

[35] L. I. Ignat and J.D. Rossi. A nonlocal convection-diffusion equation. *J. Functional Analysis*, **251(2)** (2007), 399–437.

[36] S. Kindermann, S. Osher and P. W. Jones. Deblurring and denoising of images by nonlocal functionals. *Multiscale Model. Simul.*, **4**, (2005), 1091–1115.

[37] U. Mosco. Convergence of convex sets and solutions of variational inequalities. *Advances. Math.*, **3** (1969), 510–585.

[38] L. Rudin, S. Osher and E. Fatemi, Nonlinear Total Variation based Noise Removal Algorithms, *Physica D.*, **60**, (1992), 259–268.

[39] M. Schonbek, Decay of solutions to parabolic conservation laws, *Comm. Partial Differential Equations*, 5(5), 449–473, (1980).

[40] M. Schonbek, Uniform decay rates for parabolic conservation laws, *Nonlinear Anal.*, 10(9), 943–956, (1986).

[41] M. Schonbek, The Fourier splitting method, Advances in geometric analysis and continuum mechanics (Stanford, CA, 1993), Int. Press, Cambridge, MA, 1995, 269–274.

[42] C. Tomasi and R. Manduchi. Bilateral filtering for gray and color images, in Proceedings of the Sixth International Conference on Computer Vision, Bombay, India, 1998, 839–846.

[43] L. P. Yaroslavsky. Digital Picture Processing. An Introduction, Springer, Berlin, 1985.

[44] L. P. Yaroslavsky and M. Eden. Fundamentals of Digital Optics, Birkhäuser, Boston, 1996.

Sesiones monográficas

DINAMYCS AND BIFUCATION IN PIECEWISE SMOOTH SYSTEMS

V. Carmona

MATHEMATICS FOR HEALTH SCIENCES

M. Marín

APPROXIMATION THEORY AND SPECIAL FUNCTIONS WITH APPLICATIONS

J.L. López

J.L. Varona

RECENTS ADVANCES IN THE MATHEMATICAL AND NUMERICAL ANALYSIS OF OCEANOGRAPHY

L. Cea

M. Gómez Mármol

J.M. González Vida

GOAL ORIENTED ADAPTATIVE METHODS FOR THE NUMERICAL SOLUTION OF PDEs

J. Carpio

CONVOLUTIONS CODES

J.A. Domínguez

C. Perea

R.Pinto

Bol. Soc. Esp. Mat. Apl.
nº 42(2008), 69–78

BIFURCACIÓN SILLA–NODO DE CONOS INVARIANTES EN SISTEMAS LINEALES A TROZOS VIA BIFURCACIÓN FOCO-CENTRO-CICLO LíMITE

V. CARMONA, E. FREIRE, E. PONCE, J. ROS Y F. TORRES

Departamento de Matemática Aplicada II
Universidad de Sevilla

{vcarmona,efrem,eponcem,javieros,ftorres}@us.es

Resumen

En este trabajo se considera la existencia de conos invariantes en sistemasdinámicos continuos tridimensionales lineales a trozos, dada la relevancia que estas variedades invariantes tienen en la determinación de la estabilidad del origen en tales sistemas. Se recogen varios resultados de existencia de conos invariantes y se analiza una bifurcación silla-nodo de estas variedades invariantes. La relación biunívoca existente entre los conos invariantes y las órbitas periódicas de ciertos sistemas planos discontinuos (en particular, las que se generan en una bifurcación foco-centro-ciclo límite) constituye la herramienta fundamental en el estudio.

Palabras clave: *Sistemas dinámicos lineales a trozos, variedades invariantes, bifurcación silla-nodo.*
Clasificación por materias AMS: *34C23, 37G15*

1 Introducción

Los sistemas lineales a trozos se utilizan en diferentes disciplinas científicas para modelar una amplia gama de procesos y dispositivos. Entre estos sistemas, tienen especial relevancia los sistemas continuos que poseen dos zonas de linealidad, con el origen como único punto de equilibrio del sistema y situado en la frontera que separa dichas zonas. Una primera tarea en el estudio de estos sistemas es la determinación de la estabilidad y tipo topológico del origen. La estabilidad del origen puede garantizarse, como es bien sabido, mediante el uso de funciones de Liapunov. Sin embargo, la búsqueda de funciones de Liapunov en estos sistemas no es una tarea sencilla (véanse [8] y [9]) y, por otro lado, la existencia de una función de Liapunov no es una condición necesaria de estabilidad. Por lo tanto, resulta apropiado considerar otras técnicas para determinar la estabilidad del equilibrio.

En el caso bidimensional con dos zonas de linealidad, la estabilidad del origen está perfectamente caracterizada (véase, por ejemplo, [6]), mientras que si el sistema no es plano, el problema no es en absoluto trivial (ver [3] y [5]).

Todo sistema dinámico continuo tridimensional lineal a trozos con dos zonas y con un equilibrio en el origen localizado en la frontera de separación, puede escribirse en la forma

$$\dot{\mathbf{x}} = F(\mathbf{x}) = \begin{cases} A^+\mathbf{x} & \text{if} \quad x \geqslant 0, \\ A^-\mathbf{x} & \text{if} \quad x < 0, \end{cases} \tag{1}$$

donde $\mathbf{x} = (x, y, z)^T \in \mathbb{R}^3$ y las matrices A^+ y A^- de orden tres comparten, por continuidad, sus dos últimas columnas; esto es, $A^+ - A^- = (A^+ - A^-)\,\mathbf{e_1}\mathbf{e_1}^T$, siendo $\mathbf{e_1} = (1, 0, 0)^T$ el primer vector de la base canónica $\mathbb{R}^3$.

El campo vectorial F que define al sistema lineal a trozos (1) es homogéneo; es decir, $F(\mu\mathbf{x}) = \mu F(\mathbf{x})$ para todo $\mathbf{x} \in \mathbb{R}^3$ y $\mu \geqslant 0$. Por consiguiente, el flujo del sistema (1) transforma semirrectas contenidas en el plano de separación $x = 0$ que pasan por el origen en semirrectas del mismo tipo. Si una de estas semirrectas es invariante para el flujo, entonces el sistema (1) posee un cono invariante que calificaremos de bizonal porque interseca a los dos semiespacios de linealidad $\{x > 0\}$ y $\{x < 0\}$. El sistema (1) puede tener también conos invariantes que se encuentren en uno sólo de los semiespacios de linealidad, en cuyo caso se denominarán unizonales.

La estabilidad del origen para el sistema (1) está estrechamente relacionada con la presencia o ausencia de conos invariantes en dicho sistema. Así, tal y como se deduce de la Proposición 10 de [3] y el Teorema 2 de [5], si ambas matrices A^+ y A^- poseen autovalores complejos y el sistema (1) carece de conos invariantes bizonales, entonces el origen del sistema es globalmente asintóticamente estable si y sólo si los autovalores reales de A^+ y A^- son estrictamente negativos. Este resultado generaliza el teorema enunciado por Busenberg y Van Den Driessche en [1] para sistemas homogéneos de clase $\mathcal{C}^2$.

Por otra parte, la presencia de un cono invariante bizonal en el sistema complica el estudio del tipo topológico del origen. En efecto, tal y como se recoge en el Teorema 1 de [5], cuando el sistema posee un cono invariante, el punto de equilibrio puede ser inestable aunque ambas matrices A^+ y A^- tengan su espectro en el semiplano real negativo. Además, existen otras situaciones en las que el cono aparece foliado por un continuo no acotado de órbitas periódicas.

De los resultados expuestos se puede deducir la importancia del análisis de la existencia de conos invariantes en los sistemas lineales a trozos. En [3] se hace un estudio detallado de la existencia de conos invariantes, donde se demuestra que a lo sumo pueden aparecer dos conos invariantes aislados y se conjetura la existencia de una bifurcación silla–nodo de los mismos.

En este trabajo mostramos que los conos invariantes en el sistema tridimensional se corresponden biunívocamente con las órbitas periódicas de ciertos sistemas planos cuadráticos a trozos con dos zonas. Más aún, un adecuado cambio de variable transforma estos sistemas cuadráticos a trozos en sistemas lineales a trozos con dos zonas, pero ya no homogéneos y discontinuos. Esta relación entre conos invariantes y órbitas periódicas nos permitirá, entre otros resultados, probar la existencia de la bifurcación silla–nodo conjeturada en [3] y obtener la expresión analítica que deben satisfacer los parámetros del

sistema en esa bifurcación.

El resto del artículo se organiza de la siguiente forma. En la siguiente sección mostramos que los conos invariantes del sistema tridimensional, bajo condiciones genéricas, se corresponden con las órbitas periódicas de ciertos sistemas planos discontinuos lineales a trozos. En la tercera sección mostramos que dichos sistemas planos experimentan una bifurcación foco-centro-ciclo límite, lo que utilizaremos para obtener resultados de existencia de conos invariantes. En la última sección se analiza la degeneración de la bifurcación foco-centro-ciclo límite, demostrando que ésta proporciona una bifurcación silla-nodo de conos invariantes, cuya expresión puede darse de forma analítica. Por último, y como consecuencia del estudio realizado, se presentan nuevos resultados sobre la estabilidad del origen en los sistemas tridimensionales.

2 Conos Invariantes y Órbitas Periódicas en Sistemas Planos

La relación biunívoca existente entre los conos invariantes del sistema tridimensional (1) y las órbitas periódicas de determinados sistemas planos será considerada en esta sección. Sólo estudiaremos sistemas tridimensionales que no pueden ser desacoplados, pues en caso contrario, el problema a resolver sería de menor dimensión. Siguiendo muy de cerca los conceptos y resultados enunciados en [2], consideraremos sistemas tridimensionales observables, es decir, aquellos que pueden reducirse a la denominada forma canónica de Liénard,

$$\dot{\mathbf{x}} = \begin{cases} M^+\mathbf{x} & \text{si} \quad x \geqslant 0, \\ M^-\mathbf{x} & \text{si} \quad x < 0, \end{cases} \qquad \text{con} \quad M^\pm = \begin{pmatrix} t^\pm & -1 & 0 \\ m^\pm & 0 & -1 \\ d^\pm & 0 & 0 \end{pmatrix}. \tag{2}$$

Aquí, los parámetros $t^\pm$, $m^\pm$ y $d^\pm$ son los coeficientes de los polinomios característicos de las matrices $M^\pm$, a saber,

$$p_{M^\pm}(\lambda) = \det(M^\pm - \lambda I) = -\lambda^3 + t^\pm\lambda^2 - m^\pm\lambda + d^\pm.$$

Si λ^- es un autovalor real de la matriz M^-, entonces es directo observar que el plano $\Pi_F^- \equiv (\lambda^-)^2\, x - \lambda^- y + z = 0$ es una variedad invariante para el sistema lineal $\dot{\mathbf{x}} = M^-\mathbf{x}$. Análogamente, si λ^+ es un autovalor real de la matriz M^+, entonces el plano $\Pi_F^+ \equiv (\lambda^+)^2\, x - \lambda^+ y + z = 0$ es invariante para el sistema lineal $\dot{\mathbf{x}} = M^+\mathbf{x}$. La invariancia de estos planos para los sistemas lineales anteriores limita las zonas donde se localizan los conos invariantes. Obsérvese que si $\lambda^- \neq \lambda^+$, entonces tanto Π_F^+ como Π_F^- no son invariantes para el sistema completo (2) y si $\lambda^- = \lambda^+$, entonces los planos Π_F^+ y Π_F^- coinciden y conforman un cono invariante para el sistema (2). Además, utilizando el Lema 21 y la Proposición 22 de [3], podemos asegurar que los conos invariantes no planos del sistema (2), si existen, se encuentran simultáneamente por encima de ambos planos Π_F^+ y Π_F^- o simultáneamente por debajo de ellos. Entendemos que un objeto geométrico está por encima de otro cuando las terceras componentes de todos los puntos del primer objeto son mayores que las correspondientes del segundo.

En consecuencia, debemos buscar conos invariantes no planos por encima del plano Π_F^+ o por debajo de él. Si los buscamos por encima de dicho plano, los conos invariantes se corresponden, tal y como probamos en la siguiente proposición, con las órbitas periódicas de un sistema continuo cuadrático a trozos.

Proposición 1 *Si λ^+ es un autovalor real de la matriz M^+, entonces los conos invariantes del sistema* (2) *que están por encima del plano Π_F^+ se corresponden biunívocamente con las órbitas periódicas del sistema plano continuo cuadrático a trozos*

$$\begin{cases} \dot{u}_1 &= (t^- - \lambda^+)\, u_1 - u_2 - p_{M^-}(\lambda^+) u_1^2, \\ \dot{u}_2 &= \left[m^- + (\lambda^+)^2\right] u_1 - 2\lambda^+ u_2 - p_{M^-}(\lambda^+) u_1 u_2 - 1, \end{cases} \qquad u_1 \leqslant 0. \quad (3a)$$

$$\begin{cases} \dot{u}_1 &= (t^+ - \lambda^+)\, u_1 - u_2, \\ \dot{u}_2 &= \left[m^+ + (\lambda^+)^2\right] u_1 - 2\lambda^+ u_2 - 1, \end{cases} \qquad u_1 > 0. \quad (3b)$$

Demostración. Sólo es necesario realizar el cambio de variables

$$u_1 = \frac{x}{(\lambda^+)^2 x - \lambda^+ y + z},\ u_2 = \frac{y}{(\lambda^+)^2 x - \lambda^+ y + z},\ Z = \left(\lambda^+\right)^2 x - \lambda^+ y + z,$$

válido cuando $(\lambda^+)^2 x - \lambda^+ y + z > 0$. □

Si $\lambda^+ = \lambda^-$, entonces el sistema (3) es lineal en cada semiplano, mientras que si $\lambda^+ \neq \lambda^-$, entonces el sistema es lineal en el semiplano $u_1 \geqslant 0$ y cuadrático en el semiplano $u_1 \leqslant 0$. Seguidamente, probamos que el sistema cuadrático puede ser transformado en un sistema lineal.

Proposición 2 *Si $\lambda^+ \neq \lambda^-$, entonces el sistema cuadrático* (3a) *es equivalente al sistema lineal*

$$\begin{cases} \dot{u}_1 = (t^- - \lambda^-) u_1 - u_2, \\ \dot{u}_2 = \left[m^- + (\lambda^-)^2\right] u_1 - 2\lambda^- u_2 - 1, \end{cases} \quad (4)$$

en cada uno de los semiplanos abiertos que determina la recta de ecuación

$$1 - \left(\lambda^+ - \lambda^-\right)\left[\left(\lambda^+ + \lambda^-\right) u_1 - u_2\right] = 0. \quad (5)$$

Demostración. Es suficiente realizar el cambio de variables

$$\begin{aligned} U_1 &= \frac{u_1}{1 - (\lambda^+ - \lambda^-)\left[(\lambda^+ - \lambda^-)\, u_1 - u_2\right]} \\ U_2 &= \frac{u_2}{1 - (\lambda^+ - \lambda^-)\left[(\lambda^+ - \lambda^-)\, u_1 - u_2\right]} \end{aligned} \quad (6)$$

válido cuando no se satisface (5), y renombrar las variables U_1 y U_2. □

Debemos señalar que las órbitas periódicas del sistema (3), si existen, no pueden tener puntos comunes con la recta (5), ya que deben corresponder a

conos invariantes por encima de los planos Π_F^+ y Π_F^-, y entonces, las órbitas periódicas del sistema (3) deben estar localizadas en la región

$$\mathcal{R} = \left\{(u_1, u_2) \in \mathbb{R}^2 : 1 - \left(\lambda^+ - \lambda^-\right)\left[\left(\lambda^+ + \lambda^-\right) u_1 - u_2\right] > 0\right\}. \tag{7}$$

Por otra parte, obsérvese que el cambio dado en (6) deja invariante la recta de separación $u_1 = 0$ y el único punto fijo sobre ella es el origen. Nótese además, que el cambio (6) se reduce a la identidad cuando $\lambda^+ = \lambda^-$.

También debemos indicar que el sistema que se obtiene como unión del sistema lineal (4) actuando en la zona $u_1 < 0$ con el sistema lineal (3b) es discontinuo y es necesario introducir desplazamientos en la recta $u_1 = 0$ para recuperar las órbitas del sistema (3). Cuando una órbita alcance la recta de separación $u_1 = 0$ con $\dot{u}_1 < 0$, el punto de intersección debe sufrir en la recta de separación el desplazamiento

$$\delta(u_2) = \frac{u_2}{1 + (\lambda^+ - \lambda^-)\, u_2}, \tag{8}$$

antes de que el flujo del sistema de la zona $u_1 < 0$ actúe. Análogamente, cuando una órbita alcance la recta de separación $u_1 = 0$ con $\dot{u}_1 > 0$, el punto de intersección debe sufrir en la recta de separación el desplazamiento inverso

$$\delta^{-1}(u_2) = \frac{u_2}{1 - (\lambda^+ - \lambda^-)\, u_2}, \tag{9}$$

antes de que el flujo del sistema de la zona $u_1 > 0$ actúe. Puesto que debemos trabajar en la región $\mathcal{R}$ definida en (7), las funciones δ y δ^{-1} actúan en puntos de sus respectivos dominios de definición.

Así, la búsqueda de conos invariantes en el sistema (2) que están por encima del plano Π_F^+ se traslada a la búsqueda de órbitas periódicas del sistema lineal a trozos con impactos δ y δ^{-1},

$$\begin{cases} \dot{u}_1 = (t^- - \lambda^-)u_1 - u_2 \\ \dot{u}_2 = \left[m^- + (\lambda^-)^2\right] u_1 - 2\lambda^- u_2 - 1 \end{cases} \quad u_1 < 0,$$
$$\begin{cases} \dot{u}_1 = (t^+ - \lambda^+)\, u_1 - u_2, \\ \dot{u}_2 = \left[m^+ + (\lambda^+)^2\right] u_1 - 2\lambda^+ u_2 - 1, \end{cases} \quad u_1 > 0. \tag{10}$$

Ahora, teniendo en cuenta que el cambio $U_2 = -2\lambda^- u_1 + u_2$ transforma el sistema (4) en la forma de Liénard

$$\begin{cases} \dot{u}_1 = (t^- - 3\lambda^-)u_1 - U_2, \\ \dot{U}_2 = \left[m^- - 2\lambda^- t^- + 3\,(\lambda^-)^2\right] u_1 - 1, \end{cases}$$

es inmediato enunciar el siguiente resultado, ya que las funciones de impacto no se ven alteradas. La derivada del polinomio característico $p_{M^\pm}$ respecto de λ se denotará por $p'_{M^\pm}$.

Proposición 3 *Si λ^+ es un autovalor real de la matriz M^+, entonces los conos invariantes del sistema* (2) *que están por encima del plano Π_F^+ se corresponden biunívocamente con las órbitas periódicas del sistema plano discontinuo lineal a trozos con impactos δ y δ^{-1} definidos en* (8) *y* (9)

$$\begin{cases} \dot{u}_1 = (t^- - 3\lambda^-)u_1 - u_2 \\ \dot{u}_2 = -p'_{M^-}(\lambda^-)u_1 - 1 \end{cases} \quad u_1 < 0, \qquad \begin{cases} \dot{u}_1 = (t^+ - 3\lambda^+)\,u_1 - u_2, \\ \dot{u}_2 = -p'_{M^+}(\lambda^+)u_1 - 1, \end{cases} \quad u_1 > 0. \tag{11}$$

Evidentemente, si $p'_{M^+}(\lambda^+) \geqslant 0$ y $p'_{M^-}(\lambda^-) \leqslant 0$, entonces el sistema continuo cuadrático a trozos (3) no posee puntos de equilibrio y, por tanto, tampoco órbitas periódicas. Es decir, cuando $p'_{M^+}(\lambda^+) \geqslant 0$ y $p'_{M^-}(\lambda^-) \leqslant 0$, el sistema (2) no puede tener conos invariantes por encima de los planos Π_F^+ y Π_F^-. Un comentario análogo puede hacerse cuando se buscan conos invariantes por debajo de dichos planos. En particular, si todos los autovalores de las matrices M^+ y M^- son reales, entonces se podría probar que el sistema (2) no puede tener conos invariantes.

3 La Bifurcación Foco-Centro-Ciclo Límite para el Sistema con Impactos

En esta sección describiremos el fenómeno de bifurcación foco-centro-ciclo límite (véase [4]) y [7]) que tiene lugar en el sistema discontinuo plano lineal a trozos con impactos (11). Supondremos que en una de las zonas de linealidad es posible la existencia de un centro, lo que obliga a la existencia de un par de autovalores complejos en una de las zonas. También supondremos, aunque no es necesario, que la matriz de la otra zona tiene autovalores complejos. Por tanto, asumimos que las matrices M^+ y M^- del sistema tridimensional (2) poseen autovalores λ^+, $\alpha^+ \pm \mathrm{i}\beta^+$ y λ^-, $\alpha^- \pm \mathrm{i}\beta^-$, con $\beta^\pm > 0$. En consecuencia, el sistema (11) adopta (renombrando las variables u_1 y u_2 como x e y, respectivamente) la forma

$$\begin{cases} \dot{x} = 2(\alpha^- - \lambda^-)x - y \\ \dot{y} = \left[(\alpha^- - \lambda^-)^2 + (\beta^-)^2\right] x - 1 \end{cases} \quad x < 0, \qquad \begin{cases} \dot{x} = 2(\alpha^+ - \lambda^+)x - y \\ \dot{y} = \left[(\alpha^+ - \lambda^+)^2 + (\beta^+)^2\right] x - 1 \end{cases} \quad x > 0. \tag{12}$$

Los autovalores de las matrices que rigen al sistema (12) son $(\alpha^\pm - \lambda^\pm) \pm \mathrm{i}\beta^\pm$. Así, es fácil ver que el sistema (12) tiene un único punto de equilibrio, que se encuentra en la zona derecha, y que es de tipo centro si y sólo $\alpha^+ = \lambda^+$. En estas condiciones podemos enunciar el siguiente resultado, que por razones de brevedad acompañamos sólo con un esquema de la prueba. Como paso previo definimos los coeficientes

$$\eta = 3\lambda^+ - \lambda^- - 2\alpha^- \quad \text{y} \quad \tilde{\eta} = 3\lambda^- - \lambda^+ - 2\alpha^+. \tag{13}$$

Teorema 4 *Supongamos que las matrices M^+ y M^- del sistema continuo lineal a trozos tridimensional* (2) *poseen autovalores λ^+, $\alpha^+ \pm \mathrm{i}\beta^+$ y λ^-, $\alpha^- \pm \mathrm{i}\beta^-$, con $\beta^\pm > 0$, y sea η el valor definido en* (13)*. Entonces, el sistema plano lineal a trozos discontinuo* (12) *con impactos δ y δ^{-1} definidos en* (8) *y* (9) *posee un ciclo límite cuando $\alpha^+ - \lambda^+$ es suficientemente pequeño y $(\alpha^+ - \lambda^+) \cdot \eta > 0$.*

Este ciclo límite emerge de la órbita periódica de la configuración de centro, existente para $\alpha^+ = \lambda^+$, que es tangente a la recta de separación $x = 0$ en el origen. Además, el ciclo límite es único en un entorno de la órbita periódica.

Demostración. Supongamos que sistema (12) posee una órbita periódica de dos zonas, entonces existen dos puntos $(0, y_0)$ y $(0, y_1)$, con $y_1 > 0$ e $y_0 < 0$, y dos valores positivos τ^+ y τ^- tales que

$$\begin{cases} \phi^+\left((0, y_0), \tau^+\right) = (0, y_1), \\ \phi^-\left((0, \delta(y_1)), \tau^-\right) = (0, \delta(y_0)), \end{cases} \tag{14}$$

donde ϕ^+ y ϕ^- son los flujos de los sistemas lineales en la zona $x > 0$ y $x < 0$ que definen al sistema (12) y δ la función desplazamiento definida en (8).

Las expresiones dadas en (14) se denominan ecuaciones de cierre y caracterizan las órbitas periódicas bizonales del sistema (12) cuando se verifican las condiciones

$$\mathbf{e}_1^T \cdot \phi^+((0, y_0), t) > 0 \; \forall t \in \left(0, \tau^+\right), \; \mathbf{e}_1^T \cdot \phi^-(0, \delta(y_1)), t) < 0 \; \forall t \in \left(0, \tau^-\right), \tag{15}$$

donde $\mathbf{e}_1^T = (1, 0)$.

Supongamos fijos todos los parámetros que intervienen en el sistema a excepción de α^+. De esta forma, las ecuaciones de cierre (14) pueden ser entendidas como un sistema de cuatro ecuaciones con cinco incógnitas: $y_0 < 0, y_1 > 0, \tau^- > 0, \tau^+ > 0$ y α^+. Es inmediato comprobar que $\bar{q} = (\bar{y}_0, \bar{y}_1, \bar{\tau}^-, \bar{\tau}^+, \bar{\alpha}^+) = (0, 0, 0, 2\pi/\beta^+, \lambda^+)$ es solución de las ecuaciones de cierre y se corresponde con la órbita del centro lineal de la zona derecha tangente a la recta de separación en el origen. Dicho punto es singular y no es posible aplicar el teorema de la función implícita. Afortunadamente, el desarrollo en serie de la tercera ecuación de (14) es proporcional a τ^- y no es difícil conseguir, tras la eliminación del factor común τ^-, unas ecuaciones equivalentes para $\tau^- \neq 0$ y no singulares en el punto. La aplicación del teorema de la función implícita sobre estas últimas ecuaciones nos permite concluir, teniendo en cuenta que ϕ^+ y ϕ^- son flujos de sistemas lineales, que existe solución de las ecuaciones de cierre con $\tau^- \neq 0$ en un entorno del punto $\bar{q}$ y los siguientes desarrollos en las incógnitas son válidos para τ^- suficientemente pequeño:

$$\begin{aligned} y_0 &= -\tfrac{\tau^-}{2} + \tfrac{\eta}{12}\left(\tau^-\right)^2 + O\left(\tau^-\right)^3, & y_1 &= \tfrac{\tau^-}{2} + \tfrac{\eta}{12}\left(\tau^-\right)^2 + O\left(\tau^-\right)^3, \\ \tau^+ &= \tfrac{2\pi}{\beta^+} - \tau^- + O\left(\tau^-\right)^3, & \alpha^+ &= \lambda^+ + \tfrac{\left(\beta^+\right)^3 \eta}{24\pi}\left(\tau^-\right)^3 + O\left(\tau^-\right)^5, \end{aligned} \tag{16}$$

siendo η el coeficiente definido en (13).

Finalmente, teniendo en consideración el signo de τ^-, podemos afirmar que las soluciones (16) de las ecuaciones de cierre (14) se corresponden con un ciclo límite del sistema (12) si $\alpha^+ - \lambda^+$ es suficientemente pequeño y $(\alpha^+ - \lambda^+)\cdot\eta > 0$, pues es directo probar que las soluciones (16) satisfacen las condiciones (15) siempre que $\tau^- > 0$ sea suficientemente pequeño. □

El teorema anterior nos conduce al siguiente resultado de forma inmediata.

Teorema 5 *Bajo la hipótesis del Teorema 4, el sistema* (2) *posee un cono invariante bizonal por encima de los planos* Π_F^+ *y* Π_F^- *si* $\alpha^+ - \lambda^+$ *es suficientemente pequeño y* $(\alpha^+ - \lambda^+)\cdot\eta > 0$.

Notemos que se puede dar un resultado dual para los conos invariantes que se localizan por debajo de los planos Π_F^+ y Π_F^-.

Teorema 6 *Bajo la hipótesis del Teorema 4, el sistema* (2) *posee un cono invariante bizonal por debajo de los planos* Π_F^+ *y* Π_F^- *si* $\alpha^- - \lambda^-$ *es suficientemente pequeño y* $(\alpha^- - \lambda^-)\cdot\tilde{\eta} > 0$, *donde* $\tilde{\eta}$ *está definida en* (13).

4 Degeneración de la Bifurcación Foco-Centro-Ciclo Límite

En este apartado se analiza la situación de degeneración de la bifurcación foco-centro-ciclo límite que se produce cuando uno de los coeficientes $\eta = 3\lambda^+ - \lambda^- - 2\alpha^-$ ó $\tilde{\eta} = 3\lambda^- - \lambda^+ - 2\alpha^+$ es nulo. Si esto ocurre, entonces aparece una bifurcación silla-nodo de conos invariantes, como mostramos a continuación, y la conjetura apuntada en [3] queda así demostrada. Debemos señalar, como se deduce del Teorema 2 de [3], que $(\alpha^+ - \lambda^+)\cdot(\alpha^- - \lambda^-) < 0$ si el sistema (2) tiene más de un cono invariante bizonal. La prueba del siguiente teorema se realizará, por razones de brevedad, de forma esquemática.

Teorema 7 *Supongamos que* $(\alpha^+ - \lambda^+)$ *y* η *son suficientemente pequeños,* $(\alpha^+ - \lambda^+)\cdot(\alpha^- - \lambda^-) < 0$ *y* $(\alpha^+ - \lambda^+)\cdot\eta > 0$. *Entonces, existe una función* $\alpha_{SN} = \alpha_{SN}(\lambda^+, \lambda^-, \alpha^-, \beta^+, \beta^-)$ *definida localmente por*

$$\alpha_{SN} = \lambda^+ + \frac{729\sqrt{2}}{10\pi}\,(\beta^+)^3 \left[\frac{\lambda^+ - (\lambda^- + 2\alpha^-)/3}{(\lambda^- - \alpha^-)\left[(\lambda^- - \alpha^-)^2 + 9(\beta^-)^2\right]}\right]^{3/2} \cdot \frac{3\lambda^+ - \lambda^- - 2\alpha^-}{3} + \cdots$$

de manera que se satisfacen las siguientes propiedades:

1. *Si* $(\alpha^+ - \alpha_{SN})\cdot(\alpha^+ - \lambda^+) < 0$, *entonces el sistema* (2) *posee dos conos invariantes bizonales.*
2. *Si* $(\alpha^+ - \alpha_{SN})\cdot(\alpha^+ - \lambda^+) > 0$, *entonces el sistema* (2) *no posee conos invariantes y el origen del sistema* (2) *es globalmente asintóticamente estable si y sólo si* $\lambda^+ < 0$ *y* $\lambda^- < 0$.
3. *Si* $\alpha^+ = \alpha_{SN}(\lambda^+, \lambda^-, \alpha^-, \beta^+, \beta^-)$, *entonces el sistema* (2) *posee un único cono invariante bizonal y es semiestable, es decir, la correspondientes órbita periódica del sistema cuadrático* (3) *es semiestable.*

Demostración. El desarrollo de α^+ dado en (16) hasta orden cinco es

$$\alpha^+ = \lambda^+ + \frac{\left(\beta^+\right)^3\eta}{24\pi}\left(\tau^-\right)^3 + \frac{\left(\beta^+\right)^3\left(\lambda^- - \alpha^-\right)\left[\left(\lambda^- - \alpha^-\right)^2 + 9\left(\beta^-\right)^2\right] + O(\eta)}{2160\pi}\left(\tau^-\right)^5 + \cdots \tag{17}$$

El número de soluciones $\tau^- > 0$ de la ecuación (17) cuando τ^- es suficientemente pequeño se corresponde con el número de ciclos límite del sistema plano (12), y por tanto con el número conos invariantes bizonales del sistema tridimensional (2) que están por encima del plano Π_F^+.

Cuando η es suficientemente pequeño el número de soluciones positivas de la ecuación (17), para τ^- suficientemente pequeño, es el mismo que el de las soluciones positivas de la ecuación $h\left(\tau^-\right) = 0$, siendo

$$h\left(\tau^-\right) = \lambda^+ - \alpha^+ + \frac{\left(\beta^+\right)^3\eta}{24\pi}\left(\tau^-\right)^3 + \frac{\left(\beta^+\right)^3\left(\lambda^- - \alpha^-\right)\left[\left(\lambda^- - \alpha^-\right)^2 + 9\left(\beta^-\right)^2\right]}{2160\pi}\left(\tau^-\right)^5.$$

Las hipótesis aseguran que el número de soluciones positivas de la ecuación $h\left(\tau^-\right) = 0$ se discrimina a partir del signo del valor de la función h en su extremo relativo $\tau_*^- > 0$. Puesto que este valor viene dado por

$$h\left(\tau_*^-\right) = \alpha^+ - \lambda^+ + \frac{729\sqrt{2}}{10\pi}\left(\beta^+\right)^3\left[\frac{\eta}{3\left(\lambda^- - \alpha^-\right)\left[\left(\lambda^- - \alpha^-\right)^2 + 9\left(\beta^-\right)^2\right]}\right]^{3/2} \cdot \frac{\eta}{3}$$

la conclusión del Teorema es inmediata. □

El Teorema 7 tiene la siguiente versión dual.

Teorema 8 *Supongamos que* $(\alpha^- - \lambda^-)$ *y* $\tilde{\eta}$ *son suficientemente pequeños,* $(\alpha^+ - \lambda^+) \cdot (\alpha^- - \lambda^-) < 0$ *y* $(\alpha^- - \lambda^-) \cdot \tilde{\eta} > 0$. *Entonces para la función*

$$\bar{\alpha}_{SN} = \lambda^- + \frac{729\sqrt{2}}{10\pi}\left(\beta^-\right)^3\left[\frac{\lambda^- - (\lambda^+ + 2\alpha^+)/3}{(\lambda^+ - \alpha^+)\left[(\lambda^+ - \alpha^+)^2 + 9(\beta^+)^2\right]}\right]^{3/2} \cdot \frac{3\lambda^- - \lambda^+ - 2\alpha^+}{3} + \cdots$$

se satisfacen las siguientes propiedades:

1. *Si* $(\alpha^- - \bar{\alpha}_{SN}) \cdot (\alpha^- - \lambda^-) < 0$, *entonces el sistema* (2) *posee dos conos invariantes bizonales.*
2. *Si* $(\alpha^- - \bar{\alpha}_{SN}) \cdot (\alpha^- - \lambda^-) > 0$, *entonces el sistema* (2) *no posee conos invariantes y el origen del sistema* (2) *es globalmente asintóticamente estable si y sólo si* $\lambda^+ < 0$ *y* $\lambda^- < 0$.
3. *Si* $\alpha^- = \bar{\alpha}_{SN}$, *el sistema* (2) *posee un único cono invariante bizonal y es semiestable.*

En los dos últimos resultados, además de probar la existencia de la bifurcación silla–nodo conjeturada en [3], se avanza en la caracterización de la estabilidad del origen, pero el problema está aun lejos de ser resuelto completamente. La propiedad de homogeneidad hace que en estos sistemas se confunda la dinámica local y la global, de manera que bien puede ocurrir que sólo con técnicas de caracter global sea posible resolver definitivamente el problema de la estabilidad del origen.

Agradecimientos

Los autores agradecen la financiación de los proyectos MTM2004-04066 y MTM2006-00847 del Ministerio de Educación y Ciencia, así como del proyecto EXC/2005/FQM-872 de la Junta de Andalucía.

Referencias

[1] S. Busenberg y P. Van Den Driessche, *A Method for Proving the Non-existence of Limit Cycles*, Journal of Mathematical Analysis and Applications **172** 463–469 (1993).

[2] V. Carmona, E. Freire, E. Ponce & F. Torres, *On Simplifying and Classifying Piecewise Linear Systems*, IEEE Trans. Circuits Systems I Fund. Theory Appl. **49**, 609-620 (2002).

[3] V. Carmona, E. Freire, E. Ponce & F. Torres, *Bifurcation of Invariant Cones in Piecewise Linear Homogeneous Systems*, Internat. J. Bifur. Chaos Appl. Sci Engrg., **15**, 2469–2484 (2005).

[4] V. Carmona, E. Freire, E. Ponce, J. Ros & F. Torres, *Limit Cycle Bifurcation in 3D Continuous Piecewise Linear Systems With Two Zones: Application to Chua's Circuit*, Internat. J. Bifur. Chaos Appl. Sci Engrg., **15**, 3153-3164 (2005).

[5] V. Carmona, E. Freire, E. Ponce & F. Torres, *The continuous matching of two stable linear systems can be unstable,* Discrete and Continuous Dynamical Systems, **16**, 689-703 (2006).

[6] E. Freire, E. Ponce, F. Rodrigo & F. Torres, *Bifurcation Sets of Continuos Piecewise Linear Systems with Two Zones*, Internat. J. Bifur. Chaos Appl. Sci Engrg., **8**, 2073–2097 (1998).

[7] E. Freire, E. Ponce & J. Ros, *Limit cycle bifurcation from a center in symmetric piecewise linear systems*, Internat. J. Bifur. Chaos Appl. Sci Engrg., **9**, 895–907 (1999).

[8] M. Johansson & A. Rantzer, *Computation of Piecewise Quadratic Lyapunov Functions for Hybrid Systems.* IEEE Trans. Automat. Control **43**, 555–559. (1998)

[9] R. N. Shorten & K.S. Narendra, *On Common Quadratic Lyapunov Functions for Pairs of Stable LTI Systems whose System Matrices are in Companion Form.* IEEE Trans. Automat. Control **48**, 618–621 (2003).

Bol. Soc. Esp. Mat. Apl.
nº 42(2008), 79–90

SIMULACIÓN NUMÉRICA DE DIVERSOS PROBLEMAS RELATIVOS AL CRECIMIENTO DE TUMORES SÓLIDOS

MERCEDES MARÍN

Departamento de Informática y Análisis Numérico
Universidad de Córdoba

merche@uco.es

Palabras clave: *Crecimiento tumoral, simulación numérica*
Clasificación por materias AMS: *46N60 65J15 35R35*

1 Introducción

En los últimos años, el desarrollo de modelos matemáticos para simular el crecimiento de tumores malignos ha crecido con rapidez. De hecho, del millón y medio de artículos que se han escrito en el área de investigación sobre el cáncer, aproximadamente el 5 % tratan de su modelación matemática. Incluso ha surgido un nuevo término que añadir a las experimentaciones *in vitro* (en laboratorio) e *in vivo* (en seres vivos), denominándose experimentos *in silico* a aquellos que se realizan mediante simulaciones numéricas en ordenador.

En este trabajo se pretende dar una idea de algunos aspectos del cancer que pueden ser tratados desde el punto de vista matemático. Para una discusión detallada sobre el tema pueden consultarse los libros [1], [18] y [34], los especiales dedicados a la modelación de tumores [10], [15] y [16], los artículos recopilatorios [6], [11], [31], [35], [38], [40] y las referencias que aparecen en todos ellos.

El desarrollo de un tumor sólido comienza con la alteración de una célula como resultado de la mutación de ciertos genes. De esta forma, la célula pierde su ritmo natural de proliferación (mitosis) y muerte (apoptósis) dando lugar a un crecimiento inadecuado que origina el tumor.

La modelación matemática del crecimiento de tumores requiere de diferentes aproximaciones dependiendo del fenómeno que se desee simular ya que éstos tienen lugar a escalas diferentes: subcelular (cambios genéticos, alteración del ciclo celular...), celular (proliferación de las células tumorales, competición con el sistema inmune...) y macroscópica a nivel de tejidos (migración de células, convección y difusión de nutrientes y factores químicos,...).

Trabajo subvencionado por los proyectos EXC/2005/FQM-520 de la Junta de Andalucía y MTM2006-07932 del MEC.

Los modelos matemáticos relacionados con fenómenos celulares en los que interviene una única célula se escriben, generalmente, en términos de ecuaciones diferenciales ordinarias mientras que cuando se trata de fenómenos colectivos se usan ecuaciones cinéticas integro-diferenciales. Por otra parte, los fenómenos a escala macroscópica vienen descritos, en general, en términos de ecuaciones en derivadas parciales no lineales en los que pueden aparecer problemas de fronteras móviles (ver [9]).

Lo que sucede en cada escala está relacionado en gran medida con lo que sucede en las otras. Por ello, hoy día uno de los principales problemas abiertos es entender y modelar la relación entre la descripción microscópica y macroscópica.

A su vez, los modelos se pueden clasificar en dos categorías: modelos discretos y modelos continuos. En los modelos discretos las células se siguen de forma individual y se actualizan de acuerdo a un conjunto específico de reglas biológicas. Aquí se incluyen, por ejemplo, los modelos de autómatas celulares que aparecen en [2], [5] y [30]. Estas aproximaciones se utilizan, por ejemplo, en el estudio de la carcinogenesis que es el proceso por el cual las células normales se transforman en cancerosas, de la inestabilidad genética o para estudiar las interacciones de células individuales con cada una de las de su entorno. Estos métodos pueden ser difíciles de estudiar desde el punto de vista analítico y su coste computacional crece rápidamente con el número de células modeladas. Para problemas de cierto tamaño, teniendo en cuenta que en un tumor esférico de 1 mm hay del orden de 500000 células, la aplicación de estos métodos no es factible. Para sistemas de gran escala donde la población de células cancerígenas el del orden de millones, los modelos continuos proporcionan una buena alternativa. Los modelos que presentamos en este trabajo pertenecen a esta última categoría.

2 Modelos macroscópicos continuos

Las células tumorales forman masas que están hechas de diversos constituyentes (células tumorales, células inmunes, matriz extracelular...) cuyo crecimiento, además de los nutrientes, depende de varios factores que lo promueven e inhiben. Para realizar la modelación hay que distinguir dos tipos de componentes:

1. Los diferentes tipos de células, la matriz extracelular y el líquido extracelular que invade el tejido. Estos constituyentes ocupan espacio, y no pueden ser traspasados.

2. Los nutrientes, las macromoléculas y los factores químicos disueltos en el líquido, producidos y absorbidos por las células. Éstos se difunden por los tejidos y su dimensión relativa es despreciable.

Para describir las poblaciones de células, la matriz extracelular y el líquido extracelular se consideran sus razones volumétricas $\phi_j, j = 1, \ldots, P$ que se definen como el volumen ocupado por la población $j-$ésima sobre el volumen total. Así, se considera el continuo no en su estado real (a nivel celular en un punto espacial sólo está presente un constituyente al mismo tiempo) sino como

una mezcla: en cada punto de la mezcla hay una fracción de ϕ_j del $j-$ésimo constituyente (ver [36]). Para ϕ_j se puede escribir un sistema de ecuaciones de balance de masas

$$\rho_j \left[\frac{\partial \phi_j}{\partial t} + \nabla \cdot (\phi_j \mathbf{v}_j) \right] = \Gamma_j, \quad j = 1, \ldots, P, \tag{1}$$

aquí, $\mathbf{v}_j$ es la velocidad de la población $j-$ésima y ρ_j su densidad. Γ_j es un término que integra la producción y la desaparición del constituyente.

Por otra parte, la evolución de los factores químicos y nutrientes se describen por sus concentraciones u_i, $i = 1, \ldots, M$. Para éstas se puede escribir un sistema de ecuaciones de reacción-difusión

$$\frac{\partial u_i}{\partial t} + \nabla \cdot (u_i \mathbf{v}_e) = \nabla \cdot (Q_i \nabla u_i) + G_i, \quad i = 1, \ldots, M, \tag{2}$$

donde Q_i es el coeficiente de difusión del $i-$ésimo factor químico, $\mathbf{v}_e$ es la velocidad del líquido extracelular y G_i engloba la producción y desaparición de u_i.

Probablemente, el punto más delicado de los modelos (1)–(2) es definir cómo se mueven las células. Esto puede obtenerse basándose en argumentos biológicos o escribiendo las ecuaciones de balance de momentos o la ecuación de balance de fuerzas.

Con (1)–(2) y sus ecuaciones de cierre pueden describirse muchos de los modelos de proliferación tumoral que aparecen en la literatura. Los diferentes constituyentes a tener en cuenta dependen de la fase de crecimiento en la que se encuentre el tumor. Este crecimiento se puede dividir en tres etapas: crecimiento avascular, angiogénesis y crecimiento vascular(ver [3]).

La modelación del cáncer ha avanzado por dos caminos distintos. Mientras que algunos grupos apuntan a producir un modelo que describa las tres etapas del crecimiento del tumor, incluyendo los complicados detalles de enlace entre ellas, otros estudian etapas individuales con gran profundidad y consideran variaciones en el entorno del tumor.

3 Etapa avascular

En la etapa avascular el tumor es lo bastante pequeño como para recibir nutrientes y eliminar desechos por difusión. Sin embargo, la difusión no es suficiente para soportar su crecimiento continuado. Esto se debe a que la velocidad con que el tumor consume nutrientes es proporcional a su volumen mientras que el aporte de nutrientes es proporcional a su superficie. El tumor avascular puede así permanecer dormido durante un periodo indefinido en el cual el crecimiento se detiene. Cuando la concentración de oxígeno disminuye por debajo de un valor crítico, las células centrales del tumor mueren, formando un núcleo necrótico. Este núcleo está rodeado por una región exterior de células tumorales proliferantes y por una región intermedia de células inactivas (quiescentes).

Esta fase puede observarse y estudiarse en el laboratorio mediante cultivos *in vitro* de células cancerosas. Si el tumor crece esféricamente adquiere un tamaño limitado y rara vez sobrepasan los 1 ó 2 mm, por ello suelen ser lesiones asintomáticas y clínicamente no detectables. Sin embargo, si el tumor adquiere una forma irregular, comienzan a llegar nutrientes adicionales al interior del tumor debido al incremento de la razón entre el área superficial y el volumen y entonces continúa creciendo.

Para modelar esta etapa existen básicamente dos clases de modelos: modelos mixtos (más realistas) en los cuales se considera que todas las células están presentes de forma continua en todo el tumor, y modelos segregados (menos complicados) en los cuales las diferentes poblaciones de células están separadas por interfases (fronteras libres). Estos últimos suelen ser modelos unidimensionales, ya que se considera el tumor como un esferoide con simetría radial.

Un modelo más general es el considerado en [8] que simula el crecimiento de un tumor sólido entre la fase avascular y vascular sin tener en cuenta la angiogénesis intermedia.

El tumor es tratado como un medio poroso. No se considera la elasticidad del tejido y las fuerzas de adhesión célula-célula se modelan mediante una tensión superficial en el tejido tumoral exterior. La velocidad de la mitosis depende de la concentración de nutrientes dentro del tumor. Se considera que el tumor no tiene zona necrótica y que no existen inhibidores químicos en los tejidos externos.

Se supone, que en un instante t, el tumor ocupa una región $\omega(t)$ de frontera $\gamma(t)$. Las incógnitas son la concentración de nutrientes dentro del tumor $\sigma = \sigma(\mathbf{x}, t)$, y $p = p(x, t)$ la presión a la que están sometidas sus células. Una vez adimensionalizadas, las ecuaciones del modelo, para cada $t \in (0, T)$, son las siguientes:

$$\begin{array}{rcll} \alpha\sigma - \Delta\sigma & = & 0 & \text{en } \omega(t), \\ -\Delta p & = & \mu(\sigma - \tilde{\sigma}) & \text{en } \omega(t), \\ \sigma & = & 1 & \text{sobre } \gamma(t), \\ p & = & a\kappa & \text{sobre } \gamma(t), \\ \frac{\partial p}{\partial n} & = & -V_n & \text{sobre } \gamma(t). \end{array} \tag{3}$$

Aquí, $V_n(x, t)$ representa la componente normal de la velocidad $\vec{V}(x, t)$ con la que se mueve $\gamma(t)$ y κ es su curvatura. Se supone que μ, a, y α son constantes positivas y que $\tilde{\sigma}$ es una función dada. Para obtener la última condición, se ha supuesto que se verifica la Ley de Darcy ($\vec{V} = -\nabla p$).

Al tratarse de un problema de frontera libre presenta dificultades tanto teóricas como numéricas. Cuando se considera simetría radial se convierte, en esencia en un modelo 1D. Un estudio teórico de (3) y otros modelos relacionados con él puede verse en [20] y en las referencias que allí aparecen.

En [19] se resuelve numéricamente (3) utilizando un método de integrales frontera. En [27], [28] y [24] se utilizan diferencias finitas y level set para su resolución mientras que en [12] se utilizan elementos finitos junto con dominios ficticios y level set.

En cierto tipo de tumores (como el de mama) parece que es más adecuado

considerar el tumor no como un medio poroso sino como un fluido y utilizar una ecuación de Stokes, en lugar de la Ley de Darcy, para relacionar la velocidad y la presión (ver [21]).

4 Angiogénesis

La siguiente etapa de crecimiento del tumor es la angiogénesis que consiste en la formación de vasos sanguíneos nuevos a partir de los vasos preexistentes. Es un fenómeno normal durante el desarrollo embrionario, el crecimiento del organismo y en la cicatrización de las heridas. Sin embargo también es un proceso fundamental en la transformación maligna del crecimiento tumoral.

Como ya hemos comentado, cuando el tumor alcanza un cierto tamaño comienza a ser insuficiente el aporte de nutrientes por simple difusión. Ante la falta de oxígeno, las células que sufren hipoxia liberan unos factores químicos denominados TAF (*Tumor Angiogenic Factors*) que se difunden a través de los tejidos circundantes. Las células endoteliales que forman los vasos sanguíneos próximos al tumor, al ser alcanzadas y estimuladas por los TAF, liberan enzimas que degradan su membrana basal. De esta manera, proliferan y migran formando nuevos vasos sanguíneos que surten al tumor (neovascularización) e incrementan su progresión. Sin embargo, esta red de vasos tiene una estructura anormal que crece y se remodela de forma continua.

La angiogénesis es un paso necesario y requerido para la transición de un grupo inofensivo pequeño de células, a un tumor de gran tamaño. También es imprescindible para la diseminación de un cáncer. Las células cancerosas pueden desprenderse de un tumor sólido determinado, entrar en un vaso sanguíneo, y trasladarse a un lugar distante, donde pueden implantarse y comenzar el crecimiento de un tumor secundario o metástasis.

Los modelos que corresponden a la fase angiogénica describen el movimiento de las células endoteliales hacia el tumor en respuesta a los gradientes de TAF mediante un proceso de quimiotaxis (ver [25]).

En los últimos años han aparecido diversos modelos matemáticos que centran su atención en diferentes aspectos del proceso angiogénico ([4], [7], [13], [26], [32], [33], [39],[42], [43]).

Veamos, a continuación, dos de los modelos más representativos.

Modelo sin proliferación celular

En el modelo propuesto en [4], se consideran tres variables, la densidad de células endoteliales $E = E(x,t)$, la concentración de TAF $T = T(x,t)$ y la matriz extracelular a través de la concentración de fibronectina $F = F(x,t)$. El modelo supone que los TAF son segregados por las células tumorales y se difunden (de forma instantánea) activando a las células endoteliales de los vasos próximos al tumor. Las células endoteliales producen fibronectina y, a su vez, ésta se degrada por contacto con ellas. Por otra parte, se supone que el movimiento de las células endoteliales viene influenciado por una difusión aleatoria (análoga a la difusión molecular), por quimiotaxis ante los gradientes de TAF y por haptotaxis en

respuesta a los gradientes de fibronectina. El sistema de ecuaciones, una vez adimensionalizado, es

$$\begin{aligned}
\frac{\partial E}{\partial t} &= \overbrace{d_E \Delta E}^{\text{difusión aleatoria}} - \overbrace{\nabla \cdot \left(\frac{\chi}{1+\alpha T} E \nabla T \right)}^{\text{quimiotaxis}} - \overbrace{\nabla \cdot (\rho E \nabla F)}^{\text{haptotaxis}}, \\
\frac{\partial F}{\partial t} &= \overbrace{\beta E}^{\text{producción}} - \overbrace{\gamma E F}^{\text{consumo}}, \\
\frac{\partial T}{\partial t} &= - \overbrace{\eta E T}^{\text{consumo}},
\end{aligned}$$

con $d_E, \chi, \alpha, \rho, \beta, \gamma$ y η constantes positivas. El dominio considerado es $\Omega = (0,1)^2$ y representa un trozo del tejido de la córnea. El vaso sanguíneo se supone que está en un lateral de Ω y el tumor en el opuesto. Se consideran condiciones de flujo nulo sobre la frontera de Ω para las células endoteliales. En [4] se realizan diversas simulaciones numéricas comparando los resultados con los que aparecen en [23].

Modelo con proliferación celular

En el modelo anterior se supone que las células endoteliales no proliferan. Esto se justifica porque, en situaciones normales, su vida media es del orden de meses mientras que la escala de los procesos que intervienen en la angiogénesis es del orden de días. Sin embargo, el mecanismo de proliferación durante la angiogénesis y su tratamiento puede que tenga una mecánica diferente. En [14] se introduce otro modelo en el que se tiene en cuenta la proliferación y muerte de las células endoteliales. Se supone que la mitosis está regulada por un crecimiento logístico y que la pérdida de células es un proceso lineal. También se supone que las células no comienzan a proliferar hasta que no se alcanza un cierto nivel de TAF. Las ecuaciones son

$$\begin{aligned}
\frac{\partial E}{\partial t} &= d_E \Delta E - \nabla \cdot \left(\frac{\chi}{1+\alpha T} E \nabla T \right) - \nabla \cdot (\rho E \nabla F) \\
&\quad + \beta_r (1-E) E G(T) - \beta_d E, \\
\frac{\partial F}{\partial t} &= \beta E - \gamma E F, \\
\frac{\partial T}{\partial t} &= -\eta E T,
\end{aligned} \tag{4}$$

con

$$G(T) = \begin{cases} 0 & \text{si} \quad T \leq T^*, \\ T - T^* & \text{si} \quad T^* < T. \end{cases}$$

De nuevo $d_E, \chi, \alpha, \rho, \beta_r, \beta_d, \beta, \gamma$ y η son constantes positivas. Para las células endoteliales se consideran condiciones de contorno de flujo nulo.

A pesar de que estos modelos continuos son capaces de capturar algunos hechos de la angiogénesis, tales como una densidad media de ramificaciones y

la velocidad de crecimiento de la red, no pueden dar más detalles acerca de su estructura y su morfología.

Para salvar estas deficiencias, surgen modelos probabilísticos discretos que permiten el seguimiento del movimiento individual de las células endoteliales, que son las que conforman la red capilar. De esta manera, incorporando reglas que contemplan la bifurcación y la unión de capilares (anastomosis), se generan redes más realistas ([4]).

Existen también modelos que estudian el flujo de sangre dentro de la red capilar inducida por el tumor [32], [39]. El objetivo es investigar la eficiencia de los tratamientos de quimioterapia y cómo éstos pasan del flujo sanguíneo al tumor.

5 Etapa vascular

En la siguiente etapa del tumor, el crecimiento vascular, el tumor se abastece de nutrientes a través de la vasculatura creada y comienza a crecer rápidamente. Pueden ocurrir mutaciones adicionales que incrementan la movilidad celular y la producción de enzimas que degradan la matriz extracelular. Esto puede dar lugar a la invasión, en la cual, o bien individualmente, o bien colectivamente, las células cancerosas se salen y/o se separan del tumor y migran a través de los tejidos que lo rodean, o a la metástasis, en la cual las células tumorales entran en el flujo sanguíneo o en el sistema linfático y viajan a otras localizaciones.

En [17] se desarrolla un modelo de invasión en el que se consideran tres variables; la densidad de células tumorales $n = n(x,t)$, la densidad de matriz extracelular $f = f(x,t)$ y la concentración de enzimas que degradan la matriz $m = m(x,t)$. Se supone que la migración de las células tumorales está influenciada por su difusión aleatoria y por el gradiente de concentración de matriz extracelular (haptotaxis). También se supone que las células tumorales producen enzimas con un proceso lineal y que éstas se difunden por el tejido adyacente al tumor con una pérdida también lineal. Por otra parte, la velocidad con la que las células tumorales degradan la matriz extracelular depende de la probabilidad de encuentro entre ambas.

Así, el sistema de ecuaciones es el siguiente

$$\begin{aligned}
\frac{\partial n}{\partial t} &= d_n \Delta n - \gamma \nabla \cdot (n \nabla f)\,, \\
\frac{\partial f}{\partial t} &= -\eta m\, f, \\
\frac{\partial m}{\partial t} &= d_m \Delta m + \alpha n - \beta m,
\end{aligned}$$

con $d_n, \gamma, \eta, d_m \alpha$ y β constantes positivas. En [17], se consideran condiciones de contorno de flujo nulo tanto para las células tumorales como para las enzimas. El problema se resuelve en $\Omega = (0,1)$ y $\Omega = (0,1)^2$ considerando un amplio rango

de valores para el parámetro γ a fin de observar la influencia de la haptotaxis en el proceso de invasión.

En [44], se presenta un modelo más completo que simula desde el crecimiento en la etapa avascular, pasando por la transición de la fase avascular a la vascular hasta las últimas etapas de crecimiento e invasión de los tejidos sanos.

6 Comentarios finales

Los modelos matemáticos pueden utilizarse para comparar diferentes protocolos de tratamiento y pueden reducir significativamente el tiempo y el coste necesario para desarrollar y probar nuevos medicamentos. De esta manera se pueden determinar los protocolos óptimos antes de utilizar los medicamentos en pruebas clínicas.

Entre los tratamientos que hoy día se están aplicando, unos están orientados a eliminar directamente las células tumorales mientras que otros actúan en la fase angiogénica, bien destruyendo la nueva vasculatura que se genera, o bien anulando la acción de los TAF.

Así, por ejemplo, un modelo que incorpora la acción de un inhibidor, $I = I(x,t)$, de la proliferación de células endoteliales al modelo de angiogénesis (4), es el presentado en [41]:

$$\begin{aligned}
\frac{\partial E}{\partial t} &= D\Delta E - \nabla\cdot\left(\frac{\chi}{1+\alpha T}E\nabla T\right) - \nabla\cdot(\rho E\nabla F)\,,\\
&\quad +\beta_r(1-E)EG(T)\left(1-\tfrac{\varepsilon_{max}I_0 I}{(IC)_{50}+I_0 I}\right) - \beta_d E,\\
\frac{\partial F}{\partial t} &= \beta E - \gamma E\,F,\\
\frac{\partial T}{\partial t} &= -\eta E\,T,\\
\frac{\partial I}{\partial t} &= -\gamma_c I + \gamma_u U_{I,ex},
\end{aligned}$$

donde $\gamma_c, \gamma_u, \varepsilon_{max}, I_0$ y $(IC)_{50}$ son constantes que dependen del inhibidor y $U_{I,ex}$ representa la velocidad con que éste se administra.

Otro tipo de problemas matemáticos muy interesantes relacionados con el estudio del crecimiento tumoral son aquellos que pueden plantearse como un problema inverso y que se basan, entre otras, en las propiedades eléctricas, elásticas o térmicas que tienen los tejidos tumorales (ver, por ejemplo, [22], [37] y [29]).

Referencias

[1] Adam, J.A., Bellomo, N., Eds. *A Survey of Models on tumor Immune Systems Dynamics*, (Birkhäuser, Boston),(1996).

[2] Alarcon, T., Byrne, H.M., Maini, P.K. A cellular automaton model for tumour growth in inhomogeneous environment, *J. Theor. Biol.*, 225 , 257–274, (2003).

[3] Alarcon, T., Byrne, H.M., Maini, P.K. A multiple scale model for tumor growth. *Multiscale Model Simul.*, 3(2), 440–475, (2005).

[4] Anderson A.R.A., Chaplain M.A.J., Continuous and discrete mathematical models of tumor-induced angiogenesis, *Bull. Math. Biol.*, 60, 857-899 (1998).

[5] Anderson, A.R.A., A Hybrid Mathematical Model of Solid Tumour Invasion: The Importance of Cell Adhesion, *IMA Math. App. Med. Biol.*, 22, 163–186, (2005).

[6] Araujo R.P., McElwain D.L.S., A history of the study of tumor growth: the contribution of mathematical modeling, *Bull. Math. Biol.*, 66, 1039-1091 (2004).

[7] Balding, D., McElwain, D.L.S. A mathematical model of tumor-induced capillary growth.*J. Theor. Biol.*, 114, 53–73 (1985).

[8] Bazaliy, B.V. , Friedman, A. , A Free Boundary Problem for a Elliptic-Parabolic System: Application to a Model of Tumor Growth. *Comm. Partial Differential Equations*, 28 ,3-4, 517–560, (2003).

[9] Bellomo, N., De Angelis E., Preziosi L., Multiscale Modeling and Mathematical Problems Related to Tumor Evolution and Medical Therapy *Journal of Theoretical Medicine*, 5(2), 111–136 (2003).

[10] Bellomo, N.,De Angelis E., Eds. . Special Issue on Modeling and simulation of tumor development, treatment, and control, *Math. Comp. Modeling* 37, (2003).

[11] Byrne, H.M, Alarcon, T., Owen, M.R., Webb, S.D., Maini, P.K. Modelling aspects of cancer dynamics: a review. *Philosophical Transactions of the Royal Society A*, 364(1843), 1563–1578 (2006).

[12] Calzada, M.C., Camacho, G. Fernández-Cara, E., Marín, M., Resolución numérica de un modelo de frontera libre para el crecimiento tumoral. Actas del XX CEDYA, X Congreso de Mat. Aplicada, Sevilla (2007).

[13] Chaplain, M.A.J., Stuart, A.M. A model mechanism for the chemotactic response of endothelial cells to tumour angiogenesis factor. *IMA J. Math. Appl. Med Biol.* 10, 149–168 (1993).

[14] Chaplain, M.A.J. Avascular growth, angiogenesis and vascular growth in solid tumours: the mathematical modelling of the stages of tumour development. *Math. Comput. Model.* 23, 47–87, (1996).

[15] Chaplain, M.A.J. Ed., Special Issue *Math. Mod. Methods Appl. Sci.* 9.

[16] Chaplain, M.A.J. Ed., Special Issue on Mathematical Modeling and Simulations of Aspects of Cancer Growth, *J. Theor. Medicine* 4, (2002).

[17] Chaplain, M.A.J., Anderson, A.R.A. Mathematical modelling of Tissue invasion. En *Cancer Modelling and Simulation*, CRCPress/ Chapman Hall, pp.269-297, (2003).

[18] Chaplain, M.A.J., *Mathematical Modelling of Tumour Growth*, Springer (2006).

[19] Cristini V., Lowengrub J., Nie Q. Nonlinear simulation of tumor growth, *J. Math. Biol* 46, 191–224 (2003).

[20] Cui, S. Well-posedness of a multidimensional free boundary problem modelling the growth of nonnecrotic tumors. *Journal of Functional Analysis*, 245 (1), 1-18, (2007).

[21] Friedman, A., Hu, B. Bifurcation from stability to instability for a free boundary problem modeling tumor growth by Stokes equation. *J. Math. Anal. Appl.*, 327, 643-664, (2007).

[22] Gatenby, R.A., Maini, P.K., Gawlinski, E.T. Analysis of tumor as an inverse problem provides a novel theoretical framework for understanding tumor biology and therapy. *Appl. Math. Lett.*, 15, 339-345, (2002).

[23] Gimbrone,M.A., Cotran, R.S., Leapman S.B., Folkman, J. Tumor growth and neovascularization: An experimental model using the rabbit cornea. *J. Natn. Cancer Inst.* 52, 413–427 (1974).

[24] Hogea, C.S., Murray, B.T., Sethian, J.A., Implementation of the level set method for continuum mechanics based tumor growth models, *FDMP* 1(2), 109–130 (2005).

[25] Horstmann, D. From 1970 until present: the Keller-Segel model in chemotaxis and its consequences. Max Planck Institute for Mathematics in the Sciences, (2003).

[26] Levine, Sleeman, B.D., Nilsen-Hamilton, M.,Mathematical modeling of the onset of capillary formation initiating angiogenesis. *J. Math. Biol.*42, 195–238 (2001).

[27] Macklin P., Lowengrub J., Evolving interfaces via gradients of geometry dependent interior Poisson problems: application to tumor growth, *J. Comput. Phys.*, 203, 191–220 (2005).

[28] Macklin P., Lowengrub J., An improved geometry-aware curvature discretization for level-set methods: Application to tumor growth, *J. Comput. Phys.*, 215, 392–401 (2006).

[29] Majchrzak, E. , Paruch, M. Identification of Dimensions and Position of Tumor Region on the Basis of Skin Surface Temperature Using the Gradient Method Coupled with the Multiple Reciprocity BEM. *ICCES*, 1(1),7–13, (2007).

[30] Mallett, D.G., de Pillis, L.G., A Cellular Automata Model of Tumor-immune System Interactions. *J. Theor. Biol.* 239(3), 334–350, (2006).

[31] Mantzaris, N., Webb, S., Othmer, H.G. Mathematical modelling of tumour-induced angiogenesis, *J. Math. Biol.* 49, 111–187 (2004).

[32] McDougall, S.R., Anderson, A.R.A., Chaplain, M.A.J. Mathematical modelling of flow through vascular networks: Implications for tumour induced angiogenesis and chemotherapy strategies. *Bulletin of Mathematical Biology*, 64, 673–702, (2002).

[33] Orme, M.E., Chaplain, M.A.J. A mathematical-model of the first steps of tumour-related angiogenesis – capillary sprout formation and secondary branching. *IMA Jl. Math. Appl. Med. Biol.* 13, 73–98 (1996).

[34] Preziosi, L., Ed. *Cancer Modelling and Simulation*, CRCPress/ Chapman Hall, (2003).

[35] Quaranta, V., Weaver, A.M., Cummings P.T., Anderson A.R.A., Mathematical Modeling of Cancer: The future of prognosis and treatment, *Clinica Chimica Acta*, 357, 173–179, (2005).

[36] Rajagopal, K.R., Tao, L. *Mechanics of Mixtures*. World Scientific, Singapore,(1995).

[37] Samani A., Plewes D., An inverse problem solution for measuring the elastic modulus of intact ex vivo breast tissue tumours. *Phys. Med. Biol.* 52, 1247–1260, (2007).

[38] Sanga, S., Sinek, J.P., Frieboes, H.B., Fruehauf, J.P., Cristini, V. Mathematical modeling of cancer progression and response to chemotherapy, *Expert. Rev. Anticancer Ther.*, 6, 1361–1376, (2006).

[39] Stéphanou A., McDougall S.R., Anderson A.R.A., Chaplain, M.A.J., Mathematical Modelling of Flow in 2D and 3D Vascular Networks: Applications to Anti-Angiogenic and Chemotherapeutic Drug Strategies. *Mathematical and Computer Modelling* 41, 1137-1156, 2005.

[40] Swanson, K.R.,Bridge, C., Murray, J.D., Alvord, E.C. , Virtual and real brain tumors: Using mathematical modeling to quantify glioma growth and invasion, *J. Neuro. Sci.*, 216,1–10, (2003).

[41] Tee, D., DiStefano, J. Simulation of tumor-induced angiogenesis and its response to anti-angiogenic drug treatment: mode of drug delivery and clearance rate dependencies. *J. Cancer Res. Clin. Oncol.*, 130, 15–24, (2004).

[42] Valenciano J., Chaplain M.A.J., Computing highly accurate solutions of a tumour angiogenesis model *Math. Models Methods Appl. Sci.*, 13, 747-766, (2003).

[43] Valenciano J., Chaplain M.A.J., An explicit subparametric spectral element method of lines applied to a tumour angiogenesis system of partial differential equations *Math. Models Methods Appl. Sci.*, 14, 165–187, (2004).

[44] Zheng X., Wise S.M., V. Cristini. Nonlinear simulation of tumor necrosis, neo-vascularization and tissue invasion via an adaptive finite-element/level-set method. *Bulletin of Mathematical Biology* 67, 211-259. 2005.

Bol. Soc. Esp. Mat. Apl.
nº 42(2008), 91–101

ASYMPTOTIC METHODS FOR CONVOLUTION INTEGRALS UNIFIED AND DEMYSTIFIED

JOSÉ L. LÓPEZ

Departamento de Ingenieria Matemática e Informática
Universidad Pública de Navarra.

jl.lopez@unavarra.es

Abstract

This paper is a commented resume of [8]. We present a new method for deriving asymptotic expansions of $\int_0^\infty f(t)h(xt)dt$ for small x. We only require for $f(t)$ and $h(t)$ to have asymptotic expansions at $t = \infty$ and $t = 0$ respectively. Remarkably, it is a very general technique that unifies a certain set of asymptotic methods. Watson's Lemma and other classical methods, Mellin transform techniques, McClure and Wong's distributional approach and the method of analytic continuation turn out to be simple corollaries of this method. In addition, the most amazing thing about it is that its mathematics are absolutely elemental and do not involve complicated analytical tools as the aforesaid methods do: it consists of simple "sums and substractions". Many known and unknown asymptotic expansions of important integral transforms are trivially derived from the approach presented here.

Key words: *Asymptotic expansions of integrals, Mellin convolution integrals, Mellin transforms*

AMS subject classifications: *41A60 30B40 46F10*

1 Introducción

Despite the effort of many authors in asymptotics ([1], [5], [12] and [13] among others), it seems impossible to design a unique asymptotic method valid for any kind of integral containing an asymptotic parameter x: $\int_\Gamma f(x,t)dt$. Nevertheless, we show here that a simple method is possible for shedding some light on the "unification" of asymptotic methods of integrals of the form

$$I(x) \equiv \int_0^\infty f(t)h(xt)dt, \qquad x > 0. \tag{1}$$

The "Dirección General de Ciencia y Tecnología" (REF. MTM2007-63772) is acknowledged by its financial support.

Without loss of generality we can think of x as a small parameter. Many integral transforms can be put in the form (1): Laplace, Fourier, Stieltjes, Hankel, Poisson, Glasser, Lambert,... [14].

If we want to approximate (1) for small x, we may think that only the behaviour of $h(t)$ near the origin is relevant. Then, we require for $h(t)$ an expansion at $t = 0$:

$$h(t) = \sum_{k=0}^{n-1} b_k t^{k+\beta} + h_n(t), \tag{2}$$

replace this expansion for $h(xt)$ in (1) and interchange summation and integration. We obtain, formally, an asymptotic expansion for small x:

$$I(x) = \sum_{k=0}^{n-1} \left[b_k \int_0^\infty f(t) t^{k+\beta} dt \right] x^{k+\beta} + \int_0^\infty f(t) h_n(xt) dt. \tag{3}$$

In fact, classical methods such as Watson's Lemma, Laplace's method, saddle point techniques... are based on this idea. On the other hand, we may write (1) in a different form:

$$I(x) \equiv x^{-1} \int_0^\infty f\left(\frac{t}{x}\right) h(t) dt. \tag{4}$$

Written in this form, it seems plausible that only the behaviour of $f(t)$ at infinity is relevant to approximate $I(x)$ when $x \to 0$. Then, we require for $f(t)$ an expansion at $t = \infty$:

$$f(t) = \sum_{k=0}^{n-1} \frac{a_k}{t^{k+\alpha}} + f_n(t), \qquad t \to \infty. \tag{5}$$

Substituting this expansion in (4) and interchanging summation and integration we obtain the formal expansion:

$$I(x) = \sum_{k=0}^{n-1} \left[a_k \int_0^\infty t^{-k-\alpha} h(t) dt \right] x^{k+\alpha-1} + \int_0^\infty f_n(t) h(xt) dt. \tag{6}$$

If the negative moments of $h(t)$ exist, we can think of this formula as generating a new family of classical methods.

From (3) and (6) we see that classical methods require the existence of either all the positive moments of $f(t)$ or all the negative moments of $h(t)$. But if they do not exist, the coefficients of the expansion (3) or (6) are not defined and the classical expansion makes no sense.

Mclure and Wong (in the following M&W) solved this problem for certain families of functions $f(t)$ and $h(t)$ by using the theory of distributions and analytic continuation techniques [9], [10], [[12], Chaps. 5,6]. Different and more general proofs using only analytic continuation (in the following AC) have been proposed in [11], [6] and [7]. A different solution to this problem was proposed

by Handlesman and Lew by using the method of Mellin Transforms (in the following MT) [2], [3], [4], [[12], Chap. 3]. All of M&W, AC or MT are more difficult techniques than classical methods.

The main idea to be presented in this work is that the appearance of divergences in the asymptotic expansion is an artificial problem. An unnecessary problem is created when expanding h or f in (2) or (5) up to n terms, replacing this expansion in (1) and interchanging sum and integral. The idea is as simple as this: *expand h and f both simultaneously and replace these expansions in (1) in such a way that you do not create any divergence.* Moreover, this idea generates an extraordinarily simple method which contains, as straightforward corollaries, classical methods I and II, M&W, AC and MT techniques.

In the next section we give some definitions and technical results and MT, AC and M&W's theories are briefly resumed. Section 3 presents the main result of the paper: a unified and simple method to obtain asymptotic expansions of $I(x)$ for small x. Section 4 re-derives some classical results, MT, AC and M&W's theories as corollaries of the fundamental theorem of section 3.

2 Preliminaries

2.1 Definitions and technical results

Definition 1. *We denote by $\mathcal{F}$ the set of functions $f \in L^1_{\mathrm{Loc}}(0,\infty)$ verifying: (i) f has an asymptotic expansion at infinity:*

$$f(t) = \sum_{k=0}^{n-1} \frac{a_k}{t^{\alpha_k}} + f_n(t), \qquad n = 1, 2, 3, ..., \tag{7}$$

where, for $k = 0, 1, 2, ...$, $\{a_k\}$ and $\{\alpha_k\}$ are sequences of complex and real numbers respectively with α_k strictly increasing and $f_n(t) = \mathcal{O}(t^{-\alpha_n})$ as $t \to \infty$. (ii) $f(t) = \mathcal{O}(t^{-a})$ as $t \to 0^+$ with $a \in R$.

Definition 2. *We denote by $\mathcal{H}$ the set of functions $h \in L^1_{\mathrm{Loc}}(0,\infty)$ verifying: (i) h has an asymptotic expansion at $t = 0^+$:*

$$h(t) = \sum_{k=0}^{n-1} b_k t^{\beta_k} + h_n(t), \qquad n = 1, 2, 3, ..., \tag{8}$$

where, for $k = 0, 1, 2, ...$, $\{b_k\}$ and $\{\beta_k\}$ are sequences of complex and real numbers respectively with β_k strictly increasing and $h_n(t) = \mathcal{O}(t^{\beta_n})$ as $t \to 0^+$. (ii) $h(t) = \mathcal{O}(t^{-b})$ when $t \to \infty$ with $b \in R$.

Definition 3. *Let $g \in L^1_{\mathrm{Loc}}(0,\infty)$. We denote by $M[g;z]$ the Mellin transform of g, $\int_0^\infty t^{z-1} g(t) dt$ (when this integral exists), or its analytic continuation as a function of z.*

Remark 1. In the foregoing discussion we require for the parameters a, b, α_0 and β_0 to satisfy, without loss of generality, the following relations [8]:
CONDITION I: $a - \beta_0 < 1 < b + \alpha_0$.
CONDITION II. $-\beta_0 < b$ and $a < \alpha_0$.

The Mellin transform $M[f;z]$ of every function $f \in \mathcal{F}$ exists and defines a meromorphic function of z in the half plane $\Re z > a$. More precisely, for any $n \in N$,

$$M[f;z] = \begin{cases} \int_0^\infty t^{z-1} f(t)dt, & a < \Re z < \alpha_0 \\ \int_0^1 t^{z-1} f(t)dt - \sum_{k=0}^{n-1} \frac{a_k}{z-\alpha_k} + \int_1^\infty t^{z-1} f_n(t)dt, & a < \Re z < \alpha_n \\ \int_0^\infty t^{z-1} f_n(t)dt, & \alpha_{n-1} < \Re z < \alpha_n. \end{cases} \tag{9}$$

Observe that $M[f;z]$ has simple poles at the points $z = \alpha_k$, $k = 0, 1, 2, ...$ with residues $-a_k$.

The Mellin transform $M[h;z]$ of every function $h \in \mathcal{H}$ exists and defines a meromorphic function of z in the half plane $\Re z < b$. More precisely, for any $m \in N$,

$$M[h;z] = \begin{cases} \int_0^\infty t^{z-1} h(t)dt, & -\beta_0 < \Re z < b \\ \int_0^1 t^{z-1} h_m(t)dt + \sum_{k=0}^{m-1} \frac{b_k}{z+\beta_k} + \int_1^\infty t^{z-1} h(t)dt, & -\beta_m < \Re z < b \\ \int_0^\infty t^{z-1} h_m(t)dt, & -\beta_m < \Re z < -\beta_{m-1}. \end{cases} \tag{10}$$

Observe that $M[h;z]$ has simple poles at the points $z = -\beta_k$, $k = 0, 1, 2, ...$ with residues b_k.

2.2 Mellin transform techniques

Roughly speaking, the MT technique proceeds as follows. Let $h \in \mathcal{H}$ and $f \in \mathcal{F}$ and let c be any real number verifying $-\beta_0 < c < b$ and $1 - \alpha_0 < c < 1 - a$. If $M[f; 1 - c - i\cdot] \in L_1(-\infty, \infty)$ or $M[h; c + i\cdot] \in L_1(-\infty, \infty)$, then $I(x)$ may be written in the form [[12], Chap 3]:

$$I(x) = \frac{1}{2\pi i} \int_{c-i\infty}^{c+i\infty} x^{-z} M[f; 1 - z] M[h; z] dz. \tag{11}$$

A displacement of the integration contour to the straight line $\Re z = d < c$ and the use of the Cauchy residue theorem gives

$$\begin{aligned} I(x) = \sum_{d<\Re z<c} \mathrm{Res}\{x^{-z} M[f; 1 - z] M[h; z]; z = 1 - \alpha_k, -\beta_k\} \\ + \tfrac{1}{2\pi i} \int_{d-i\infty}^{d+i\infty} x^{-z} M[f; 1 - z] M[h; z] dz. \end{aligned}$$

Then, from formulas (9) and (10), when $\alpha_k - \beta_j \neq 1 \ \forall \ k, j \in N \bigcup \{0\}$ and for appropriate n and $m \in N$ [[12], Chap. 3],

$$\begin{aligned} \int_0^\infty h(xt) f(t) dt = \sum_{k=0}^{n-1} a_k M[h; 1 - \alpha_k] x^{\alpha_k - 1} + \sum_{j=0}^{m-1} b_j M[f; \beta_j + 1] x^{\beta_j} \\ + \tfrac{1}{2\pi i} \int_{d-i\infty}^{d+i\infty} x^{-z} M[f; 1 - z] M[h; z] dz. \end{aligned} \tag{12}$$

If $\alpha_k - \beta_j = 1$ for some $k, j \in N \bigcup \{0\}$, the pole $z = 1 - \alpha_k$ of $M[f; 1 - z]$ and the pole $z = -\beta_j$ of $M[h; z]$ coalesce and then, the integrand

$x^{-z}M[f;1-z]M[h;z]$ in (11) has a double pole. In this case the first line in the right hand side of (12) must be replaced by

$$\lim_{z\to 0}\left\{x^{\beta_j}\left[a_k x^{-z}M[h;1+z-\alpha_k]+b_j M[f;z+\beta_j+1]\right]\right\}.$$

Formally, the sum (12) yields an asymptotic expansion of $I(x)$ for small x. The difficulty of this method lies in the technical results required to write $I(x)$ in the form (11) and on the proof of the asymptotic character of (12).

2.3 McClure and Wong's distributional theory

Roughly speaking, M&W's theory proceeds as follows. Consider the tempered distributions $\mathbf{f}$, $\mathbf{t}_+^{-\mathbf{k}-\alpha}$ and $\mathbf{f_n}$ associated to the corresponding functions $f(t)$, $t^{-k-\alpha}$ and $f_n(t)$ in formula (7) for the particular case $\alpha_k = k+\alpha$, $k=0,1,2,....$ Consider first the case $0<\alpha<1$. Those distributions act over functions $h\in\mathcal{S}[0,\infty)$ (the Schwarz class of $\mathcal{C}^{(\infty)}[0,\infty)$ rapidly decreasing functions) in the following way [[12], Chap. 6]:

$$\begin{array}{ll} <\mathbf{f},h>=\int_0^\infty f(t)h(t)dt, & <\mathbf{f_n},h>=(-1)^n\int_0^\infty f_{n,n}(t)h^{(n)}(t)dt, \\ & <\mathbf{t}_+^{-\mathbf{k}-\alpha},h>=\frac{1}{(\alpha)_k}\int_0^\infty t^{-\alpha}h^{(k)}(t)dt \end{array} \tag{13}$$

for $k=0,1,2,...$, where

$$f_{n,n}(t)\equiv\frac{(-1)^n}{(n-1)!}\int_t^\infty (u-t)^{n-1}f_n(u)du. \tag{14}$$

From [[12], Chap 6, Lemma 1] we have that these distributions are related by the equality:

$$\mathbf{f}=\sum_{k=0}^{n-1}a_k\mathbf{t}_+^{-\mathbf{k}-\alpha}+\sum_{k=0}^{n-1}\frac{(-1)^k}{k!}M[f;k+1]\delta^{(\mathbf{k})}+\mathbf{f_n}, \tag{15}$$

where $\delta^{(k)}$ is the $k-$th derivative of the delta distribution at the origin: $<\delta^{(k)},h>=(-1)^k h^{(k)}(0)$. Applying (15) to specific kernels $h(xt)\in\mathcal{S}[0,\infty)$ and using (13) we can derive asymptotic expansions of certain integral transforms $I(x)$. For example, if $h(t)=e^{-t}$, we derive the asymptotic expansion of the Laplace transform near the origin for functions $f(t)\in\mathcal{F}$ [[12], Chap. 6, Theorem 13]:

$$\begin{array}{r} \int_0^\infty e^{-xt}f(t)dt=\sum_{k=0}^{n-1}a_k\Gamma(1-k-\alpha)x^{k+\alpha-1}+\sum_{k=0}^{n-1}(-1)^k\frac{M[f;k+1]}{k!}x^{k-1} \\ +x^n\int_0^\infty e^{-xt}f_{n,n}(t)dt. \end{array} \tag{16}$$

The case $\alpha=1$ is more complicated. In this case, the second line of (13) is replaced by

$$<\mathbf{t}_+^{-\mathbf{k}-\mathbf{1}},h>=-\frac{1}{k!}\int_0^\infty h^{(k+1)}(t)\log t\,dt, \qquad k=0,1,2,... \tag{17}$$

Formula (15) must be also replaced by [[12], Chap 6, Lemma 2]:

$$\mathbf{f} = \sum_{k=0}^{n-1} a_k \mathbf{t}_+^{\mathbf{-k-1}} + \sum_{k=0}^{n-1} \frac{(-1)^k}{k!} c_k \delta^{(\mathbf{k})} + \mathbf{f_n}, \tag{18}$$

with

$$c_k \equiv \lim_{z \to k} \left[M[f; z+1] + \frac{a_k}{z-k} \right] + a_k(\gamma + \psi(k+1)),$$

The complexity of M&W's method lies in the derivation of formulas (15) and (18) and their *a posteriori* implementation to specific kernels $h(t)$. Moreover, the calculation of a general error bound for the remainder is still a challenge.

2.4 The AC technique

As well as M&W's method, the method of analytic continuation considered in [6] requires $f \in \mathcal{F}$ for the particular case $\alpha_k = k + \alpha$, $k = 0, 1, 2, ...$ But for h it only requires $h \in \mathcal{C}^{(\infty)}[0, \infty)$ and not the more stringent condition $h \in \mathcal{S}[0, \infty)$ required in M&W's method. This method uses analytic continuation techniques instead of distributions. Nevertheless, it gives rise to a particular case of formula (12) with $\alpha_k = k + \alpha$, $k = 0, 1, 2, ...$, $M[h; 1 - \alpha_k]$ replaced by $\frac{1}{(\alpha)_k} \int_0^\infty t^{-\alpha} h^{(k)}(t) dt$, b_j replaced by $h^{(j)}(0)/j!$ and a "M&W form" for the remainder [[6], Theorems 1 and 2]:

$$\begin{array}{cc} \int_0^\infty h(xt) f(t) dt = \sum_{k=0}^{n-1} \frac{a_k}{(\alpha)_k} x^{k+\alpha-1} \int_0^\infty t^{-\alpha} h^{(k)}(t) dt + \sum_{k=0}^{n-1} \frac{M[f; k+1]}{k!} h^{(k)}(0) x^k \\ + (-1)^n x^{n-1} \int_0^\infty f_{n,n}\left(\frac{t}{x}\right) h^{(n)}(t) dt & \text{if } \ 0 < \alpha < 1 \end{array} \tag{19}$$

and

$$\begin{array}{cc} \int_0^\infty h(xt) f(t) dt = -\sum_{k=0}^{n-1} \frac{a_k}{k!} x^k \int_0^\infty h^{(k+1)}(t) \log t dt \\ + \sum_{k=0}^{n-1} \left[a_k(\psi(k+1) + \gamma - \log x) + \lim_{z \to k+1} \left(M[f; z] + \frac{a_k}{z-k-1} \right) \right] \frac{h^{(k)}(0)}{k!} x^k \\ + (-1)^n x^{n-1} \int_0^\infty f_{n,n}\left(\frac{t}{x}\right) h^{(n)}(t) dt & \text{if } \ \alpha = 1. \end{array} \tag{20}$$

This method is generalized in [11] and [7] to the case $h \in \mathcal{H}$. A formula similar to (12) is obtained in [11] and [7] for the particular case $\alpha_k = k + \alpha$ and $\beta_k = k + \beta$, $k = 0, 1, 2, ...$ Also, a different form for the remainder is obtained there.

3 The "sum up and subtract" method

We have seen in the previous section that the M&W method is a particular case of the method of analytic continuation considered in [6], which is a particular case of the method introduced in [11] and revisited in [7]. The asymptotic formula given in [11] and [7] is just formula (12) for the particular case $\alpha_k = k + \alpha$ and $\beta_k = k + \beta$, $k = 0, 1, 2, ...$ and a different form for the remainder. Then in principle, we may think that the AC method is a

particular case of the MT method. But the MT method requires the additional hypothesis $M[f;1-c-i\cdot] \in L_1(-\infty,\infty)$ or $M[h;c+i\cdot] \in L_1(-\infty,\infty)$ which is not required in the AC method. We present here a trivial proof of (12) without the restrictions $\alpha_k = k+\alpha$ or $\beta_j = j+\beta$ and which does not require $M[f;1-c-i\cdot] \in L_1(-\infty,\infty)$ either $M[h;c+i\cdot] \in L_1(-\infty,\infty)$. Moreover, it gives a simpler expression for the remainder from which a universal error bound is derived. Then, we derive a method which results in a generalization of the M&W, AC and MT methods.

We define $\alpha_{-1} \equiv a$ and $\beta_{-1} \equiv -b$ and observe that $\alpha_{-1} < \alpha_0$ and $\beta_{-1} < \alpha_0$.

Theorem 1. *Let $f \in \mathcal{F}$ and $h \in \mathcal{H}$. Then, for any n, $m \in N$ such that $\alpha_{n-1} - \beta_m < 1 < \alpha_n - \beta_{m-1}$,*

$$\int_0^\infty h(xt)f(t)dt = \sum_{k=0}^{n-1} a_k M[h;1-\alpha_k]x^{\alpha_k-1} + \sum_{j=0}^{m-1} b_j M[f;\beta_j+1]x^{\beta_j} + \int_0^\infty f_n(t)h_m(xt)dt. \tag{21}$$

If $\alpha_k - \beta_j = 1$ for some couple (k,j) then, in this formula, the sum of terms

$$a_k M[h;1-\alpha_k]x^{\alpha_k-1} + b_j M[f;\beta_j+1]x^{\beta_j}$$

must be replaced by

$$\begin{gathered}\lim_{z\to 0}\left\{x^{\beta_j}\left[a_k x^{-z} M[h;1+z-\alpha_k] + b_j M[f;z+\beta_j+1]\right]\right\} = \\ x^{\beta_j}\left\{\lim_{z\to 0}\left[a_k M[h;1+z-\alpha_k] + b_j M[f;z+\beta_j+1]\right] - a_k b_j \log x\right\}.\end{gathered} \tag{22}$$

Proof. Define $f_0(t) = f(t)$ and $h_0(t) = h(t)$. For any $k \in N\bigcup\{0\}$ it is always possible to find a $j \in N\bigcup\{0\}$ such that $\alpha_{k-1} - \beta_j < 1 < \alpha_k - \beta_{j-1}$. For the given (k,j) we have that the following integral exists:

$$\int_0^\infty f_k(t)h_j(xt)dt. \tag{23}$$

We start at $k = j = 0$ (the inequalities $\alpha_{-1} - \beta_0 < 1 < \alpha_0 - \beta_{-1}$ hold) and switch on the following algorithm which increases (k,j) step by step from $(0,0)$ up to (n,m):

(a) For a given (k,j) verifying $\alpha_{k-1} - \beta_j < 1 < \alpha_k - \beta_{j-1}$ do the following: If $\alpha_k - \beta_j < 1$ go to (b). If $\alpha_k - \beta_j > 1$ go to (c). If $\alpha_k - \beta_j = 1$ go to (d).

(b) Use $f_k(t) = a_k t^{-\alpha_k} + f_{k+1}(t)$ in (23) and formula (10)(c):

$$\int_0^\infty f_k(t)h_j(xt)dt = a_k x^{\alpha_k-1} M[h;1-\alpha_k] + \int_0^\infty f_{k+1}(t)h_j(xt)dt.$$

Go to (a) with k replaced by $k+1$.

(c) Use $h_j(xt) = b_j(xt)^{\beta_j} + h_{j+1}(xt)$ in (23) and and formula (9)(c):

$$\int_0^\infty f_k(t)h_j(xt)dt = b_j x^{\beta_j} M[f;\beta_j+1] + \int_0^\infty h_{j+1}(xt)f_k(t)dt.$$

Go to (a) with j replaced by $j+1$.

(d) Use first $f_k(t) = a_k t^{-\alpha_k} + f_{k+1}(t)$ and then $h_j(xt) = b_j (xt)^{\beta_j} + h_{j+1}(xt)$ in (23):

$$\begin{array}{c} \int_0^\infty f_k(t) h_j(xt) dt = \int_0^\infty \left[a_k t^{-\alpha_k} h_j(xt) + b_j x^{\beta_j} t^{\beta_j} f_{k+1}(t) \right] dt \\ + \int_0^\infty h_{j+1}(xt) f_{k+1}(t) dt. \end{array} \tag{24}$$

In [8] it is shown that

$$\begin{array}{c} \int_0^\infty f_k(t) h_j(xt) dt = x^{\beta_j} \left\{ \lim_{z\to 0} \left[a_k M[h; 1+z-\alpha_k] + b_j M[f; z+\beta_j+1] \right] \right. \\ \left. - a_k b_j \log x \right\} + \int_0^\infty h_{j+1}(xt) f_{k+1}(t) dt. \end{array}$$

Go to (a) with k replaced by $k+1$ and j replaced by $j+1$.

This algorithm generates formulas (21)-(22). A proof of the following theorem can be found in [8].

Theorem 2. *Within the hypothesis of Theorem 1, the expansion (21) is an asymptotic expansion for small x:*

$$\int_0^\infty f_n(t) h_m(xt) dt = \mathcal{O}(x^{\beta_m} + x^{\alpha_n - 1}) \qquad \text{when } x \to 0 \text{ and } \alpha_n \neq \beta_m + 1 \tag{25}$$

and

$$\int_0^\infty f_n(t) h_m(xt) dt = \mathcal{O}(x^{\beta_m} \log x) \qquad \text{when } x \to 0 \text{ and } \alpha_n = \beta_m + 1. \tag{26}$$

4 Demystification

Many classical techniques, the MT techniques, the M&W's theory and the AC method are easy corollaries of Theorems 1 and 2.

Corollary 1 (Classical methods I). If $t^n f(t) \in L^1(0,\infty)\ \forall\ n \geq 0$, then, in Definition 1, $a_k = 0\ \forall\ k$,

$$M[f; 1+\beta_k] = \int_0^\infty f(t) t^{\beta_k} dt, \qquad \text{and} \qquad f_n(t) = f(t).$$

Then, from Theorem 1,

$$\int_0^\infty h(xt) f(t) dt = \sum_{k=0}^{m-1} \left[b_k \int_0^\infty f(t) t^{\beta_k} dt \right] x^{\beta_k} + \int_0^\infty f(t) h_m(xt) dt.$$

An important example is Watson's Lemma: set $f(t) = e^{-t}$ and $x = \tilde{x}^{-1}$ in the above formula with $\tilde{x} \to \infty$:

$$\tilde{x} \int_0^\infty h(t) e^{-\tilde{x}t} dt = \int_0^\infty h(xt) e^{-t} dt = \sum_{k=0}^{m-1} \frac{b_k \Gamma(\beta_k + 1)}{\tilde{x}^{\beta_k}} + \tilde{x} \int_0^\infty e^{-\tilde{x}t} h_m(t) dt.$$

Corollary 2 (Classical methods II). If $t^{-n}h(t) \in L^1(0,\infty)\ \forall\ n \geq 0$, then, in Definition 2, $b_k = 0\ \forall\ k$,

$$M[h; 1-\alpha_k] = \int_0^\infty h(t)t^{-\alpha_k}dt \qquad \text{and} \qquad h_m(t) = h(t).$$

Then, from Theorem 1,

$$\int_0^\infty h(xt)f(t)dt = \sum_{k=0}^{n-1}\left[a_k\int_0^\infty h(t)t^{-\alpha_k}dt\right]x^{\alpha_k-1} + \int_0^\infty f_n(t)h(xt)dt.$$

Corollary 3 (MT techniques). Formula (12) is just formula (21) with a different expression for the remainder. Apart from the conditions required for f and h in Theorem 1 above, the MT technique requires also the integrability of $M[f; 1-c-y\cdot]$ or of $M[h; c+y\cdot]$ in order to write $I(x)$ in the form (11). Then, expansion (12) follows from calculating the poles and residues of $M[f; 1-z]$ and $M[h; z]$. But what we see in Theorem 1 is that it is not necessary to write $I(x)$ in the form (11) and therefore, the integrability of $M[f; 1-c-y\cdot]$ or of $M[h; c+y\cdot]$ is not necessary. In fact, the location of the poles and the value of the residue of $M[f; 1-z]$ and $M[h; z]$ are of fundamental importance to derive the asymptotic expansion of $I(x)$ in the MT technique. But from formulas (9)(b) and (10)(b) we see that that information is already contained in the expansions (7) and (8) of $f(t)$ and $h(t)$ and the expansion of $I(x)$ follows directly from (1).

Corollary 4 (M&W method). If $h \in \mathcal{S}[0,\infty) \subset \mathcal{C}^{(\infty)}[0,\infty)$ and $\alpha_k = k+\alpha$ then $\beta_k = k$ $(k = 0, 1, 2, ...)$, $m = n$ in Lemma 1 and

$$b_k = \frac{h^{(k)}(0)}{k!}.$$

Integrating by parts we have that

$$\int_0^\infty f_n(t)h_n(xt) = (-x)^n\int_0^\infty f_{n,n}(t)h_n^{(n)}(xt)dt = (-x)^n\int_0^\infty f_{n,n}(t)h^{(n)}(xt)dt,$$

where $f_{n,n}(t)$ is defined in (14). On the other hand, if $0 < \alpha < 1$:

$$M[h; 1-k-\alpha] = \int_0^\infty h_k(t)t^{-k-\alpha}dt = \frac{1}{(\alpha)_k}\int_0^\infty t^{-\alpha}h_k^{(k)}(t)dt = <\mathbf{t}_+^{-k-\alpha}, h>.$$

Therefore, from (21):

$$\begin{aligned}\int_0^\infty h(xt)f(t)dt &= \textstyle\sum_{k=0}^{n-1}\frac{a_k}{(\alpha)_k}x^{k+\alpha-1}\int_0^\infty t^{-\alpha}h^{(k)}(t)dt \\ &+ \textstyle\sum_{k=0}^{n-1}\frac{M[f;k+1]}{k!}h^{(k)}(0)x^k + (-1)^n x^{n-1}\int_0^\infty f_{n,n}\left(\frac{t}{x}\right)h^{(n)}(t)dt,\end{aligned} \tag{27}$$

which is a generalization of the expansions given in [[12], Chap. 6] for $0 < \alpha < 1$. When $\alpha = 1$ the proof is a little bit more cumbersome, but straightforward [8].

Corollary 5 (AC techniques). The expansions derived in [[6], Theorems 1 and 2] by means of analytic continuation are nothing but expansions (27) and its extension to the case $\alpha = 1$ [8]. The conditions required for $f(t)$ and $h(t)$ in [6] are more stringent than those of Theorem 1 above: $f \in \mathcal{F}$, $h \in \mathcal{C}^{(\infty)}[0, \infty) \subset \mathcal{H}$. Moreover, the expansions derived in [11] and [7] are just the expansion given in Theorem 1 for the particular case $\alpha_k = k + \alpha$ and $\beta_k = k + \beta$.

References

[1] A. Erdelyi and M. Wyman. The asymptotic evaluation of certain integrals. *Arch. Rational Mech. Anal.*, Vol. 14 (1963) pp. 217–260.

[2] R. A. Handelsman and J. S. Lew. Asymptotic expansions of a class of integral transforms via Mellin transforms. *Arch. Rational Mech. Anal.*, Vol. 35 (1969) pp. 382–396.

[3] R. A. Handelsman and J. S. Lew. Asymptotic expansions of Laplace transforms near the origin. *SIAM J. Math. Anal.*, Vol 1 (1970) pp. 118–129.

[4] R. A. Handelsman and J. S. Lew. Asymptotic expansions of a class of integral transforms with algebraically dominated kernels. *J. Math. Anal. Appl.* Vol 35 (1971), pp 405–433.

[5] J. L. Lopez. Asymptotic expansions of integrals: the term by term integration method. *J. Comput. Appl. Math.* Vol. 102 (1999) pp. 181–194.

[6] J.L. López. Asymptotic expansions of a class of Mellin convolution integrals by means of analytic continuation, Presented in the *Seventh International Symposioum on Orthogonal Polynomials, Special Functions and Applications*. Copenhaguen, 2003.

[7] J.L. López. Asymptotic expansions of Mellin convolution integrals by means of analytic continuation. *J. Comput. Appl. Math.* Vol. 100 no. 2 (2007) 628–636.

[8] J.L. López. Asymptotic expansions of Mellin convolution integrals. To be published in *SIAM. Rev.*

[9] J. P. McClure and R. Wong. Explicit error terms for asymptotic expansions of Stieltjes transforms. *J. Inst. Math. Appl.* Vol. 22 (1978) pp. 129–145.

[10] J. P. McClure and R. Wong. Exact remainders for asymptotic expansions of fractional integrals. *J. Inst. Math. Appl.* Vol. 24 (1979) pp. 139–147.

[11] R. Wong. Explicit error terms for asymptotic expansions of Mellin convolutions. *J. Math. Anal. Appl.* Vol. 72 no. 2 (1979) pp. 740–756.

[12] R. Wong. Asymptotic approximations of integrals. *Academic Press, New York*, 1989.

[13] M. Wyman. The method of Laplace. *Trans. Roy. Soc. Canada.* Vol. 2 (1963) pp. 227–256.

[14] A. I. Zayed. Handbook of Function and Generalized Function Transformations. *CRC Press, New York*, 1996.

Bol. Soc. Esp. Mat. Apl.
nº 42(2008), 103–115

TRANSFORMADAS DE DUNKL Y TEOREMAS DE MUESTREO

ÓSCAR CIAURRI Y JUAN LUIS VARONA

Departamento de Matemáticas y Computación,
Universidad de La Rioja,
26004 Logroño, Spain

oscar.ciaurri@unirioja.es, jvarona@unirioja.es

Resumen

La transformada de Dunkl en la recta real es una generalización de la transformada de Fourier, y muchos resultados en los que interviene la de Fourier se pueden adaptar a este nuevo contexto mucho más amplio. En este artículo mostramos un teorema de muestreo relacionado con la transformada de Dunkl; este teorema de muestreo generaliza al clásico de Whittaker-Shannon-Kotel'nikov. Por el camino, hay que construir, sirviéndonos de las funciones de Bessel, un sistema ortogonal que es completo en $L^2((-1,1),|x|^{2\alpha+1}\,dx)$. Cuando $\alpha=-1/2$, este nuevo sistema se reduce al clásico sistema trigonométrico (exponencial) que se usa en la definición de las series de Fourier.

Palabras clave: *Teorema de muestreo, WSK, transformada de Dunkl, sistema ortogonal, funciones de Bessel.*
Clasificación por materias AMS: *94A20, 42A38.*

1 Introducción

Comencemos con un tópico (en la acepción española de la palabra) a más no poder: de todos es sabido que el muestreo de señales constituye uno de los más importantes tópicos (con su habitual uso como falso amigo del inglés) de la matemática aplicada. De hecho, los autores hemos contribuido recientemente al tema, aportando nuestro pequeño grano de arena: en [5], hemos demostrado un nuevo teorema de muestreo que generaliza el clásico de Shannon. En ese resultado, la transformada de Fourier se sustituye por una más general, la denominada transformada de Dunkl sobre la recta real; en ella aparece un parámetro α que, en el caso particular $\alpha=-1/2$, da lugar a la transformada de Fourier, y al teorema de muestreo clásico.

La investigación de los autores está subvencionada por el proyecto de la DGI número MTM2006-13000-C03-03.

Nuestro propósito ahora va a ser mostrar estos resultados de una manera divulgativa, sin dar demostraciones rigurosas, que se pueden encontrar en el artículo antes citado. Comenzaremos esta sección introductoria explicando qué son los operadores de Dunkl; un poco más adelante, haremos una breve reseña sobre teoremas de muestreo. En la segunda sección mostraremos un sistema ortogonal (también dependiente de un parámetro α) que generaliza al trigonométrico clásico. Este nuevo sistema jugará un papel clave en la tercera sección: sirviéndonos de él y de la transformada de Dunkl obtendremos nuestro teorema de muestreo.

1.1 Operadores de Dunkl

Los operadores de Dunkl en $\mathbb{R}^n$ son operadores que tienen una parte diferencial y otra en diferencias (lo que en inglés se denomina *differential-difference operators*) asociados a un grupo de reflexión finito; fueron introducidos por Dunkl en 1989, en su artículo [7]. En la recta real, y con el grupo de reflexión $\mathbb{Z}_2$, el operador de Dunkl Λ_α se define como

$$\Lambda_\alpha f(x) = \frac{d}{dx} f(x) + \frac{2\alpha+1}{2} \left(\frac{f(x) - f(-x)}{x} \right).$$

Para $\alpha \geq -1/2$ y $\lambda \in \mathbb{C}$, el problema de valores iniciales

$$\begin{cases} \Lambda_\alpha f(x) = \lambda f(x), & x \in \mathbb{R}, \\ f(0) = 1 \end{cases} \tag{1}$$

tiene una única solución, $E_\alpha(\lambda x)$, dada por

$$E_\alpha(z) = \mathcal{I}_\alpha(z) + \frac{z}{2(\alpha+1)} \mathcal{I}_{\alpha+1}(z),$$

con

$$\mathcal{I}_\alpha(z) = 2^\alpha \Gamma(\alpha+1) \frac{J_\alpha(iz)}{(iz)^\alpha} = \Gamma(\alpha+1) \sum_{n=0}^{\infty} \frac{(z/2)^{2n}}{n!\,\Gamma(n+\alpha+1)}$$

(véase [8] y [12]); como es habitual, estamos usando J_α para denotar la función de Bessel de orden α (un amplio y clásico tratado sobre funciones de Bessel es [22]). La función E_α se denomina núcleo de Dunkl. Cuando $\alpha = -1/2$, se tiene $\Lambda_{-1/2} = d/dx$ y $E_{-1/2}(\lambda x) = e^{\lambda x}$.

De manera similar a la transformada de Fourier (que corresponde al caso $\alpha = -1/2$), podemos definir la transformada de Dunkl sobre la recta real como

$$\mathcal{F}_\alpha(f, y) = \int_{\mathbb{R}} E_\alpha(-ixy) f(x)\, d\mu_\alpha(x), \qquad y \in \mathbb{R}, \tag{2}$$

donde $d\mu_\alpha$ denota la medida

$$d\mu_\alpha(x) = \tfrac{1}{2^{\alpha+1}\Gamma(\alpha+1)} |x|^{2\alpha+1}\, dx.$$

Por ser $|E_\alpha(ix)| \leq 1$ para cada $x \in \mathbb{R}$, el operador $\mathcal{F}_\alpha$ está bien definido para funciones $f \in L^1(\mathbb{R}, d\mu_\alpha)$, y

$$\|\mathcal{F}_\alpha f\|_{L^\infty(\mathbb{R},d\mu_\alpha)} \leq \|f\|_{L^1(\mathbb{R},d\mu_\alpha)}.$$

Además, como en la transformada de Fourier, si S es la clase de Schwartz,

$$\begin{aligned} \mathcal{F}_\alpha \colon S &\longrightarrow S \\ f &\longmapsto \mathcal{F}_\alpha f \end{aligned}$$

es un isomorfismo y $\mathcal{F}_\alpha^2 f(x) = f(-x)$. Entonces, del teorema de Fubini se deduce la fórmula de multiplicación

$$\int_{\mathbb{R}} \mathcal{F}_\alpha(f,x) g(x)\, d\mu_\alpha(x) = \int_{\mathbb{R}} \mathcal{F}_\alpha(g,x) f(x)\, d\mu_\alpha(x), \quad f, g \in S.$$

Tomando $g(x) = \overline{\mathcal{F}_\alpha(f,x)}$ se obtiene

$$\|\mathcal{F}_\alpha f\|_{L^2(\mathbb{R},d\mu_\alpha)} = \|f\|_{L^2(\mathbb{R},d\mu_\alpha)}, \quad f \in S.$$

Y ahora, por densidad, $\mathcal{F}_\alpha$ se extiende a funciones en $L^2(\mathbb{R}, d\mu_\alpha)$.

Realmente, también se puede definir la transformada de Dunkl en $L^2(\mathbb{R}, d\mu_\alpha)$ cuando $-1 < \alpha \leq -1/2$, aunque ahora la expresión (2) ya no es siempre válida para funciones en $L^1(\mathbb{R}, d\mu_\alpha)$. Pero esta nueva $\mathcal{F}_\alpha$ sí que conserva las propiedades en $L^2(\mathbb{R}, d\mu_\alpha)$; los detalles se pueden ver en [15]. Esto nos permitirá extender nuestro estudio al caso $\alpha > -1$.

En los últimos años ha habido una gran actividad investigadora relacionada con la transformada de Dunkl. Al ser la transformada de Fourier un caso particular suyo, cada problema previamente estudiado para la transformada de Fourier proporciona un nuevo desafío en el contexto de la de Dunkl. Así, por ejemplo, se han obtenido resultados relacionados con multiplicadores ([18, 3]), teoría de Littlewood-Paley ([19]), teoremas de Paley-Wiener ([20, 2]), transplantación ([14]), incertidumbre ([17, 16, 6]), transformadas de Riesz ([21]), así como el teorema de muestreo que nos ocupa ahora ([5]).

1.2 El teorema de muestreo clásico

El teorema de muestreo de Shannon afirma que, si una señal $f(t)$ no tiene frecuencias mayores que w ciclos por segundo, la función f está completamente determinada por sus valores $f(k/(2w))$, y se puede reconstruir mediante la denominada *serie cardinal*:

$$f(t) = \sum_{k=-\infty}^{\infty} f\left(\frac{k}{2w}\right) \frac{\operatorname{sen}(\pi(2wt-k))}{\pi(2wt-k)}. \tag{3}$$

Cuando una señal $f(t)$ no tiene frecuencias mayores que w ciclos por segundo, se dice que es *banda-limitada* al intervalo $(-2\pi w, 2\pi w)$ (el valor $2w$ que nos aparece aquí se conoce con el nombre de *frecuencia de Nyquist*). Esto es

equivalente a decir que su transformada de Fourier F se anula fuera de ese intervalo, y que, por tanto,

$$f(t) = \frac{1}{\sqrt{2\pi}} \int_{-2\pi w}^{2\pi w} F(x) e^{ixt}\, dx.$$

Pensándolo de manera inversa, las funciones f que se pueden obtener de esta forma para alguna F constituyen el denominado *espacio de Paley-Wiener*.

El principio subyacente es que toda la información de una señal banda-limitada está contenida en sus *muestras* $f(k/(2w))$.

Sin preocuparnos de la convergencia de series e integrales, ni del intercambio entre sumatorios e integrales, es fácil dar una demostración informal de este resultado. Para ello, supongamos que $w = 1/2$ (esto no conlleva ninguna pérdida de generalidad, pues es un simple cambio de variable). Así, F está soportada en $[-\pi, \pi]$ y tiene el siguiente desarrollo de Fourier:

$$\begin{aligned} F(x) &= \sum_{k=-\infty}^{\infty} \frac{1}{2\pi} \left(\int_{-\pi}^{\pi} F(t) e^{-ikt}\, dt \right) e^{ikx} = \frac{1}{\sqrt{2\pi}} \sum_{k=-\infty}^{\infty} f(-k) e^{ikx} \\ &= \frac{1}{\sqrt{2\pi}} \sum_{k=-\infty}^{\infty} f(k) e^{-ikx}. \end{aligned}$$

En consecuencia,

$$\begin{aligned} f(t) &= \frac{1}{\sqrt{2\pi}} \int_{-\pi}^{\pi} F(x) e^{ixt}\, dx = \frac{1}{2\pi} \sum_{k=-\infty}^{\infty} f(k) \int_{-\pi}^{\pi} e^{ix(t-k)}\, dx \\ &= \sum_{k=-\infty}^{\infty} f(k) \frac{\operatorname{sen}(\pi(t-k))}{\pi(t-k)}. \end{aligned}$$

Éste es el teorema de muestreo clásico. Fue Whittaker quien, en 1915, lo mostró por primera vez; posteriormente, fue redescubierto, por separado, por Shannon (en EE.UU.) y Kotel'nikov (en la URSS). En consecuencia, a menudo se conoce como teorema de muestreo de Whittaker-Shannon-Kotel'nikov o, simplemente, WSK. No podemos dejar de destacar que la verdadera importancia de este resultado surgió cuando, alrededor de 1940, Shannon comienza a aplicar el teorema de muestreo a la teoría de comunicación, donde resulta extremadamente útil. Hoy en día, las técnicas de muestreo se aplican para todo tipo de señales; en particular, se usan para tratar digitalmente tanto el sonido como la imagen.

Existen otros teoremas de muestreo; entendemos por esto como la reconstrucción de una función f a partir de sus valores en un conjunto discreto:

$$f(x) = \sum_{n \in \Delta} f(t_n) G(x, t_n),$$

para alguna sucesión $\{t_n\}_{n \in \Delta}$. En algunos contextos, estas series se conocen como *series interpolatorias de tipo Lagrange*.

Usualmente, la herramienta principal para obtener estos resultados es la relación entre una transformada continua y su homóloga discreta. Por ejemplo:

1. Transformada de Fourier/series de Fourier: el teorema clásico.
2. Transformada de Hankel/series de Fourier-Bessel: [10].
3. q-transformada de Fourier/q-series de Fourier: [11].
4. q-transformada Hankel/q-series de Fourier-Bessel: [1].
5. Transformada de Dunkl/series de Fourier-Dunkl: el que expondremos a continuación.

El lector puede encontrar muchos detalles sobre gran cantidad de teoremas de muestreo en [23]. Y un precioso *survey* sobre el tema es [9].

2 Series de Fourier-Dunkl

Por comodidad, escribamos $E_\alpha(ix)$ mediante

$$E_\alpha(ix) = 2^\alpha \Gamma(\alpha+1)\left(\frac{J_\alpha(x)}{x^\alpha} + \frac{J_{\alpha+1}(x)}{x^{\alpha+1}}\, xi\right), \quad x \in \mathbb{R}$$

(del desarrollo en serie de las funciones de Bessel se sigue de manera inmediata que la parte real de $E_\alpha(ix)$ es una función par; y, la parte imaginaria, una impar). Es bien conocido que la función de Bessel $J_{\alpha+1}(x)$ tiene una sucesión creciente de raíces positivas, $\{s_j\}_{j=1}^\infty$; tomemos, además, $s_{-j} = -s_j$ y $s_0 = 0$. Así, $\{s_j\}_{j\in\mathbb{Z}}$ son las raíces de

$$\mathrm{Im}(E_\alpha(ix)) = 2^\alpha \Gamma(\alpha+1)\,\frac{J_{\alpha+1}(x)}{x^{\alpha+1}}\, x = \frac{x}{2(\alpha+1)}\,\mathcal{I}_{\alpha+1}(ix).$$

Con ellas, definamos las funciones

$$e_{\alpha,j}(r) = \frac{2^{\alpha/2}(\Gamma(\alpha+1))^{1/2}}{|\mathcal{I}_\alpha(is_j)|} E_\alpha(is_j r), \quad j \in \mathbb{Z}\setminus\{0\}, \quad r \in (-1,1),$$

y $e_{\alpha,0}(r) = 2^{(\alpha+1)/2}(\Gamma(\alpha+2))^{1/2}$.

En el caso $\alpha = -1/2$, esto es el sistema exponencial clásico del que surgen las series de Fourier, es decir, $e_{-1/2,j}(r) = e^{i\pi j r}$.

Nuestro objetivo ahora es probar que $\{e_{\alpha,j}\}_{j\in\mathbb{Z}}$ es un sistema ortonormal completo en $L^2((-1,1), d\mu_\alpha)$. Gran parte del trabajo descansa en el siguiente lema:

Lema 1 *Sea $\alpha > -1$ y $x, y \in \mathbb{C}$. Entonces, para $x \neq y$,*

$$\int_{-1}^{1} E_\alpha(ixr)\overline{E_\alpha(iyr)}\, d\mu_\alpha(r) = \frac{2^{-\alpha-1}}{\Gamma(\alpha+2)}\,\frac{x\mathcal{I}_{\alpha+1}(ix)\mathcal{I}_\alpha(iy) - y\mathcal{I}_{\alpha+1}(iy)\mathcal{I}_\alpha(ix)}{x-y}, \tag{4}$$

y, para $x = y$,

$$\int_{-1}^{1} |E_\alpha(ixr)|^2 \, d\mu_\alpha(r) = \frac{2^{-\alpha-1}}{\Gamma(\alpha+2)} \bigg(\frac{x^2}{2(\alpha+1)} \mathcal{I}_{\alpha+1}^2(ix) - (2\alpha+1)\mathcal{I}_{\alpha+1}(ix)\mathcal{I}_\alpha(ix) + 2(\alpha+1)\mathcal{I}_\alpha^2(ix) \bigg). \quad (5)$$

La demostración del lema no es complicada, aunque sí bastante técnica. Con el fin de lograr un artículo lo más autocontenido posible, y puesto que el lema es un ingrediente básico en este artículo, no queremos prescindir de ella. Pero, por otra parte, para hacer más llevadera la narración, la posponemos hasta después del teorema (y el lector puede saltarse esa parte si lo considera oportuno). Así pues, expongamos ya la ortogonalidad buscada:

Teorema 2 *Sea* $\alpha > -1$. *Entonces, la sucesión de funciones* $\{e_{\alpha,j}\}_{j\in\mathbb{Z}}$ *es un sistema ortonormal completo en* $L^2((-1,1), d\mu_\alpha)$. *En consecuencia, para cada* $f \in L^2((-1,1), d\mu_\alpha)$, *podemos escribir*

$$f(r) = \sum_{j=-\infty}^{\infty} a_j(f) e_{j,\alpha}(r), \qquad a_j(f) = \int_{-1}^{1} f(t)\overline{e_{j,\alpha}(t)} \, d\mu_\alpha(t).$$

Demostración de la ortogonalidad. Si tomamos $x = s_j$ e $y = s_k$ con $j \neq k$, de (4) se sigue que

$$\int_{-1}^{1} e_{\alpha,j}(r)\overline{e_{\alpha,k}(r)} \, d\mu_\alpha(r) = 0.$$

Para $j = k \neq 0$ tenemos

$$\int_{-1}^{1} |e_{\alpha,j}(r)|^2 \, d\mu_\alpha(r) = \frac{2^\alpha \Gamma(\alpha+1)}{\mathcal{I}_\alpha^2(is_j)} \int_{-1}^{1} |E_\alpha(is_j r)|^2 \, d\mu_\alpha(r) = 1,$$

sin más que usar (5). El caso restante $j = k = 0$ es similar, pero teniendo en cuenta que $s_0 = 0$ e $\mathcal{I}_{\alpha+1}(is_0) = 1$. □

Demostración de la completitud. Supongamos que, para alguna función $\phi \in L^2((-1,1), d\mu_\alpha)$, se cumple

$$\int_{-1}^{1} \phi(r)\overline{e_{\alpha,j}}(r) \, d\mu_\alpha(r) = 0, \quad j \in \mathbb{Z}.$$

Descompongamos ϕ en sus partes real e imaginaria, y estas, a su vez, en su partes par (*even*) e impar (*odd*):

$$\phi(r) = a_{\mathrm{e}}(r) + a_{\mathrm{o}}(r) + i(b_{\mathrm{e}}(r) + b_{\mathrm{o}}(r))$$

(si denotamos $a(r) = \mathrm{Re}(\phi(r))$, basta tomar $a_{\mathrm{e}}(r) = (a(r) + a(-r))/2$ y $a_{\mathrm{o}}(r) = (a(r) - a(-r))/2$; y análogamente con $b(r) = \mathrm{Im}(\phi(r))$). De aquí, obtendríamos

$$\int_0^1 a_{\mathrm{o}}(r)\mathcal{I}_\alpha(is_j r) r^{2\alpha+1} \, dr = \int_0^1 b_{\mathrm{o}}(r)\mathcal{I}_\alpha(is_j r) r^{2\alpha+1} \, dr = 0,$$

para $j = 0, 1, \dots$ y

$$\int_0^1 a_{\mathrm{e}}(r)(s_j r)\mathcal{I}_{\alpha+1}(is_j r)r^{2\alpha+1}\,dr = \int_0^1 b_{\mathrm{e}}(r)(s_j r)\mathcal{I}_{\alpha+1}(is_j r)r^{2\alpha+1}\,dr = 0,$$

para $j = 1, 2, \dots$.

Finalmente, caigamos en la cuenta de que

$$\mathcal{I}_\alpha(is_j r) = c_n\phi_n(r) \quad \text{y} \quad (s_j r)\mathcal{I}_{\alpha+1}(is_j r) = d_n\psi_n(r),$$

donde $\{\phi_n\}_{n\geq 0}$ y $\{\psi_n\}_{n\geq 1}$ son sistemas de Dini, ortogonales y completos en $L^2((0,1), r^{2\alpha+1}\,dr)$ (véase [22, p. 134]). Así, concluimos que

$$a_{\mathrm{e}}(r) = a_{\mathrm{o}}(r) = b_{\mathrm{e}}(r) = b_{\mathrm{o}}(r) = 0$$

luego $\phi \equiv 0$. □

2.1 Demostración del lema

Comencemos probando (4). Es claro que $\overline{E_\alpha(iyr)} = E_\alpha(-iyr)$. De (1), tenemos

$$E_\alpha(ixr)\Lambda_\alpha E_\alpha(-iyr) = -iyE_\alpha(ixr)E_\alpha(-iyr)$$

(a lo largo de la demostración, las derivadas en Λ_α lo son respecto a r), y la misma igualdad con el par (x, y) cambiado por $(-y, -x)$. Sumando ambas identidades, obtenemos

$$i(x-y)\int_{-1}^1 E_\alpha(ixr)E_\alpha(-iyr)\,d\mu_\alpha(r) = I(x,y) + I(-y,-x)$$

con

$$I(x,y) = \int_{-1}^1 E_\alpha(-iyr)\Lambda_\alpha E_\alpha(ixr)\,d\mu_\alpha(r).$$

Si usamos la definición de Λ_α, podemos escribir $I(x,y) = I_1(x,y) + I_2(x,y)$ con

$$I_1(x,y) = \int_{-1}^1 E_\alpha(-iyr)\frac{d}{dr}E_\alpha(ixr)\,d\mu_\alpha(r)$$

e

$$I_2(x,y) = (2\alpha+1)\int_{-1}^1 E_\alpha(-iyr)\frac{E_\alpha(ixr) - E_\alpha(-ixr)}{2r}\,d\mu_\alpha(r).$$

Integrando I_1 por partes, queda

$$I_1(x,y) = \frac{E_\alpha(ix)E_\alpha(-iy) - E_\alpha(-ix)E_\alpha(iy)}{2^{\alpha+1}\Gamma(\alpha+1)} - I_1(-y,-x) - (2\alpha+1)\int_{-1}^1 \frac{E_\alpha(ixr)E_\alpha(-iyr)}{r}\,d\mu_\alpha(r),$$

y por tanto

$$I(x,y)=\frac{E_\alpha(ix)E_\alpha(-iy)-E_\alpha(-ix)E_\alpha(iy)}{2^{\alpha+1}\Gamma(\alpha+1)}$$
$$-I_1(-y,-x)-(2\alpha+1)\int_{-1}^{1}E_\alpha(-iyr)\frac{E_\alpha(ixr)+E_\alpha(-ixr)}{2r}\,d\mu_\alpha(r).$$

En consecuencia,

$$I(x,y)+I(-y,-x)=\frac{E_\alpha(ix)E_\alpha(-iy)-E_\alpha(-ix)E_\alpha(iy)}{2^{\alpha+1}\Gamma(\alpha+1)}$$
$$-(2\alpha+1)\int_{-1}^{1}E_\alpha(-iyr)\frac{E_\alpha(ixr)+E_\alpha(-ixr)}{2r}\,d\mu_\alpha(r)+I_2(-y,-x)$$
$$=\frac{E_\alpha(ix)E_\alpha(-iy)-E_\alpha(-ix)E_\alpha(iy)}{2^{\alpha+1}\Gamma(\alpha+1)}$$
$$-(2\alpha+1)\int_{-1}^{1}\frac{E_\alpha(ixr)E_\alpha(iyr)+E_\alpha(-ixr)E_\alpha(-iyr)}{2r}\,d\mu_\alpha(r)$$
$$=\frac{E_\alpha(ix)E_\alpha(-iy)-E_\alpha(-ix)E_\alpha(iy)}{2^{\alpha+1}\Gamma(\alpha+1)},$$

donde en el último paso hemos usado que $\frac{E_\alpha(ixr)E_\alpha(iyr)+E_\alpha(-ixr)E_\alpha(-iyr)}{2r}$ es una función impar. Así, hemos mostrado que

$$\int_{-1}^{1}E_\alpha(ixr)\overline{E_\alpha(iyr)}\,d\mu_\alpha(r)=\frac{1}{2^{\alpha+1}\Gamma(\alpha+1)}\frac{E_\alpha(ix)E_\alpha(-iy)-E_\alpha(-ix)E_\alpha(iy)}{i(x-y)}.$$

Ahora, usando $E_\alpha(-ix)E_\alpha(iy)=\overline{E_\alpha(ix)}\,\overline{E_\alpha(-iy)}$ y que, para $a,b\in\mathbb{C}$, $ab-\overline{a}\overline{b}=2i\,\mathrm{Im}(ab)=2i(\mathrm{Re}(a)\,\mathrm{Im}(b)+\mathrm{Im}(a)\,\mathrm{Re}(b))$, aparece (4).

Para probar (5) basta con evaluar

$$\lim_{y\to x}\frac{1}{2^{\alpha+1}\Gamma(\alpha+2)}\frac{x\mathcal{I}_{\alpha+1}(ix)\mathcal{I}_\alpha(iy)-y\mathcal{I}_{\alpha+1}(iy)\mathcal{I}_\alpha(ix)}{x-y}.$$

Esto se consigue sin más que usar la regla de L'Hopital y las identidades

$$\frac{d\mathcal{I}_\alpha(iy)}{dy}=-\frac{y}{2(\alpha+1)}\,\mathcal{I}_{\alpha+1}(iy)$$

e

$$\mathcal{I}_{\alpha+2}(ix)=\frac{4(\alpha+1)(\alpha+2)}{x^2}(\mathcal{I}_{\alpha+1}(ix)-\mathcal{I}_\alpha(ix)),$$

con lo que concluye la demostración del lema.

3 El teorema de muestreo

Tal como se hace habitualmente en teoría de muestreo, tomamos el **espacio de tipo Paley-Wiener** que, en nuestro caso, se define como

$$PW_\alpha = \left\{ f \in L^2(\mathbb{R}, d\mu_\alpha) : f(x) = \int_{-1}^{1} u(y) E_\alpha(ixy)\, d\mu_\alpha(y), \right. \\ \left. u \in L^2((-1,1), d\mu_\alpha) \right\},$$

con la norma de $L^2(\mathbb{R}, d\mu_\alpha)$. Con esto, ya estamos en condiciones de establecer el teorema de muestreo:

Teorema 3 *Si $f \in PW_\alpha$, $\alpha > -1$, entonces f tiene la representación*

$$f(x) = f(s_0)\mathcal{I}_{\alpha+1}(ix) + \sum_{j \in \mathbb{Z}\setminus\{0\}} f(s_j) \frac{x\mathcal{I}_{\alpha+1}(ix)}{2(\alpha+1)\mathcal{I}_\alpha(is_j)(x - s_j)},$$

que converge en la norma de $L^2(\mathbb{R}, d\mu_\alpha)$. Además, la convergencia de la serie es uniforme en subconjuntos compactos de $\mathbb{R}$.

Antes de abordar su demostración, merece la pena que hagamos un par de comentarios.

En primer lugar, destaquemos que el caso $\alpha = -1/2$ da lugar al teorema WSK clásico. En efecto, no hay más que tener en cuenta que $J_{-1/2}(x) = \sqrt{2/(\pi x)}\cos(x)$, $J_{1/2}(x) = \sqrt{2/(\pi x)}\,\mathrm{sen}(x)$ y que los s_j son, ahora, $s_j = \pi j$. Con esto, obtener la fórmula clásica (3) (con frecuencia de Nyquist $2w = 1/\pi$) a partir de la del teorema 3 es un mero trámite.

En segundo lugar, también es fácil comprobar que, para funciones pares, el resultado se convierte en el teorema de Higgins ([10]) para el par transformada de Hankel/series de Fourier-Bessel antes citado.

3.1 Una demostración informal

Una demostración rigurosa, que además muestre la convergencia uniforme en subconjuntos compactos a la que alude el teorema, requeriría algo más de trabajo. Lo habitual en estos casos es recurrir al contexto de núcleos reproductores (siguiendo las técnicas que se detallan, por ejemplo, en [10] o [13]), que aquí no vamos a explicar. En todo caso, la demostración completa del teorema se puede consultar en [5].

Para $f \in PW_\alpha$, consideremos su correspondiente función $u \in$

$L^2((-1,1),d\mu_\alpha)$ y tomemos su desarrollo de Fourier-Dunkl:

$$\begin{aligned}
u(y) &= \sum_{j=-\infty}^{\infty} a_j(u)e_{\alpha,j}(y)\\
&= \sum_{j=-\infty}^{\infty} \left(\int_{-1}^{1} u(t)\overline{e_{\alpha,j}}(t)\,d\mu_\alpha(t)\right) e_{\alpha,j}(y)\\
&= 2^{\alpha/2}(\Gamma(\alpha+1))^{1/2} \sum_{j=-\infty}^{\infty} \frac{e_{\alpha,j}(y)}{|\mathcal{I}_\alpha(is_j)|}\int_{-1}^{1} u(t)\overline{E_\alpha}(is_jt)\,d\mu_\alpha(t)\\
&= 2^{\alpha/2}(\Gamma(\alpha+1))^{1/2} \sum_{j=-\infty}^{\infty} \frac{e_{\alpha,j}(y)}{|\mathcal{I}_\alpha(is_j)|}\int_{-1}^{1} u(t)E_\alpha(-is_jt)\,d\mu_\alpha(t)\\
&= 2^{\alpha/2}(\Gamma(\alpha+1))^{1/2} \sum_{j=-\infty}^{\infty} \frac{e_{\alpha,j}(y)}{|\mathcal{I}_\alpha(is_j)|}f(-s_j)\\
&= 2^{\alpha/2}(\Gamma(\alpha+1))^{1/2} \sum_{j=-\infty}^{\infty} \frac{\overline{e_{\alpha,j}}(y)}{|\mathcal{I}_\alpha(is_j)|}f(s_j).
\end{aligned}$$

De esta manera,

$$\begin{aligned}
f(x) &= \int_{-1}^{1} u(y)E_\alpha(ixy)\,d\mu_\alpha(y)\\
&= 2^{\alpha/2}(\Gamma(\alpha+1))^{1/2} \sum_{j=-\infty}^{\infty} \frac{f(s_j)}{|\mathcal{I}_\alpha(is_j)|}\int_{-1}^{1} E_\alpha(ixy)\overline{e_{\alpha,j}}(y)\,d\mu_\alpha(y)\\
&= 2^{\alpha}\Gamma(\alpha+1) \sum_{j=-\infty}^{\infty} \frac{f(s_j)}{(\mathcal{I}_\alpha(is_j))^2}\int_{-1}^{1} E_\alpha(ixy)E_\alpha(-is_jy)\,d\mu_\alpha(y)\\
&= f(s_0)\mathcal{I}_{\alpha+1}(ix) + \frac{1}{2(\alpha+1)} \sum_{j\in\mathbb{Z}\setminus\{0\}} f(s_j)\frac{x\mathcal{I}_{\alpha+1}(ix)}{\mathcal{I}_\alpha(is_j)(x-s_j)},
\end{aligned}$$

como queríamos demostrar.

3.2 Un ejemplo

Para $\alpha,\beta,\alpha+\beta>-1$, se cumple

$$\begin{aligned}
&\int_0^\infty \frac{J_{\alpha+\beta+2n+1}(t)}{t^{\alpha+\beta+1}}\frac{J_\alpha(xt)}{(xt)^\alpha}t^{2\alpha+1}\,dt\\
&\qquad = \frac{\Gamma(n+1)}{2^\beta\Gamma(\beta+n+1)}(1-x^2)^\beta P_n^{(\alpha,\beta)}(1-2x^2)\chi_{[0,1]}(x), \quad n\in\mathbb{N}.
\end{aligned}$$

donde $P_n^{(\alpha,\beta)}$ denota el n-ésimo polinomio de Jacobi de orden (α,β), y $\chi_{[0,1]}$ es la función característica del intervalo $[0,1]$ (véase [4]). A partir de esta expresión,

se sigue fácilmente que

$$x^{2n}E_{\alpha+\beta+2n+1}(ix) \in PW_\alpha.$$

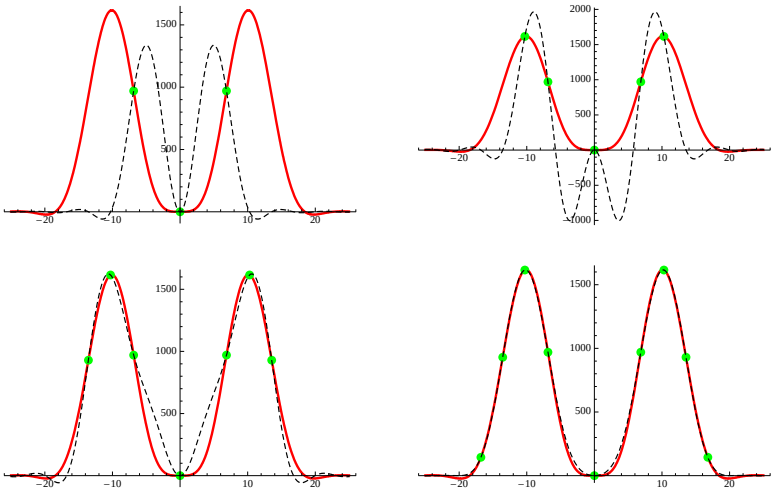
Entonces, aplicando el teorema de muestreo,

$$x^{2n}E_{\alpha+\beta+2n+1}(ix) = \sum_{j\in\mathbb{Z}\setminus\{0\}} s_j^{2n}E_{\alpha+\beta+2n+1}(is_j)\frac{x\mathcal{I}_{\alpha+1}(ix)}{2(\alpha+1)\mathcal{I}_\alpha(is_j)(x-s_j)},$$

válido para $\alpha, \beta, \alpha+\beta > -1$, y $n = 1, 2, \ldots$ Y, para $n = 0$,

$$E_{\alpha+\beta+1}(ix) = \mathcal{I}_{\alpha+1}(ix) + \sum_{j\in\mathbb{Z}\setminus\{0\}} E_{\alpha+\beta+1}(is_j)\frac{x\mathcal{I}_{\alpha+1}(ix)}{2(\alpha+1)\mathcal{I}_\alpha(is_j)(x-s_j)}.$$

Tomando $f(x) = x^{2n}E_{\alpha+\beta+2n+1}(ix)$ con $\alpha = 2.4$, $\beta = 7.2$ y $n = 2$, presentamos varios gráficos en los que representamos f (con trazo grueso), los puntos de muestreo, y (con líneas discontinuas) las sumas parciales $\sum_{j=-k}^{k}$ de la fórmula que recupera f, con $k = 1$, 2, 3 y 4, respectivamente:



Los figuras muestran que la serie truncada en seguida proporciona muy buenas aproximaciones.

Agradecimientos

Deseamos agradecer a Renato Álvarez Nodarse su invitación para impartir una charla en la sesión especial Approximation theory and special functions with applications del CEDYA-2007 en Sevilla, fruto de la cual son estas notas. Asimismo, le agredecemos el interés que se ha tomado en revisarlas.

Referencias

[1] L. D. Abreu, A q-sampling theorem related to the q-Hankel transform, *Proc. Amer. Math. Soc.* **133** (2005), 1197–1203.

[2] N. B. Andersen y M. de Jeu, Elementary proofs of Paley-Wiener theorems for the Dunkl transform on the real line, *Int. Math. Res. Not.* **30** (2005), 1817–1831.

[3] J. Betancor, Ó. Ciaurri y J. L. Varona, The multiplier of the interval $[-1, 1]$ for the Dunkl transform on the real line, *J. Funct. Anal.* **242** (2007), 327–336.

[4] Ó. Ciaurri, J. J. Guadalupe, M. Pérez y J. L. Varona, Mean and almost everywhere convergence of Fourier-Neumann series, *J. Math. Anal. Appl.* **236** (1999), 125–147.

[5] Ó. Ciaurri y J. L. Varona, A Whittaker-Shannon-Kotel'nikov sampling theorem related to the Dunkl transform, *Proc. Amer. Math. Soc.* **135** (2007), 2939–2947.

[6] Ó. Ciaurri y J. L. Varona, An uncertainty inequality for Fourier-Dunkl series, prepublicación.

[7] C. F. Dunkl, Differential-difference operators associated with reflections groups, *Trans. Amer. Math. Soc.* **311** (1989), 167–183.

[8] C. F. Dunkl, Integral kernels with reflections group invariance, *Canad. J. Math.* **43** (1991), 1213–1227.

[9] A. G. García, Orthogonal sampling formulas: a unified approach, *SIAM Rev.* **42** (2000), 499–512.

[10] J. R. Higgins, An interpolation series associated with the Bessel-Hankel transform, *J. Lond. Math. Soc.* **5** (1972), 707–714.

[11] M. E. Ismail y A. I. Zayed, A q-analogue of the Whittaker-Shannon-Kotel'nikov sampling theorem, *Proc. Amer. Math. Soc.* **131** (2003), 3711–3719.

[12] M. F. E. de Jeu, The Dunkl transform, *Invent. Math.* **113** (1993), 147–162.

[13] L. Máté, "Hilbert space methods in science and engineering", Adam Hilger, Bristol, 1989.

[14] A. Nowak y K. Stempak, Relating transplantation and multipliers for Dunkl and Hankel transforms, *Math. Nachr.* (en prensa). Prepublicación disponible en `http://www.im.pwr.wroc.pl/~anowak/research.html`

[15] M. Rosenblum, Generalized Hermite polynomials and the Bose-like oscillator calculus, *Oper. Theory Adv. Appl.* **73** (1994), 369–396.

[16] M. Rösler, An uncertainty principle for the Dunkl transform, *Bull. Austral. Math. Soc.* **59** (1999), 353–360.

[17] M. Rösler y M. Voit, An uncertainty principle for Hankel transforms, *Proc. Amer. Math. Soc.* **127** (1999), 183–194.

[18] F. Soltani, L^p-Fourier multipliers for the Dunkl operator on the real line, *J. Funct. Anal.* **209** (2004), 16–35. Corrigendum, *J. Funct. Anal.* **242** (2007), 672–673.

[19] F. Soltani, Littlewood-Paley operators associated with the Dunkl operator on $\mathbb{R}$, *J. Funct. Anal.* **221** (2005), 205–225.

[20] K. Trimèche, Paley-Wiener theorems for the Dunkl transform and Dunkl translation operators, *Integral Transforms Spec. Funct.* **13** (2002), 17–38.

[21] S. Thangavelu y Y. Xu, Riesz transform and Riesz potentials for Dunkl transform, *J. Comput. Appl. Math.* **199** (2007), 181–195.

[22] G. N. Watson, "A treatise on the theory of Bessel functions", Cambridge University Press, Cambridge, 1958.

[23] A. I. Zayed, "Advances in Shannon's sampling theory", CRC Press, Boca Raton, FL, 1993.

Bol. Soc. Esp. Mat. Apl.
nº 42(2008), 117–125

MODELIZACIÓN NUMÉRICA DEL FLUJO EN AGUAS POCO PROFUNDAS: APLICACIÓN A RÍAS Y ESTUARIOS

LUIS CEA

Grupo de Ingeniería del Agua y del Medioambiente GIAMA
Universidad de A Coruña.

lcea@udc.es

Resumen

Los modelos numéricos son una herramienta ampliamente utilizada para el estudio del flujo en rías y estuarios. Para este tipo de problemas, el coste computacional de un modelo tridimensional es en general excesivo, mientras que los modelos unidimensionales, tradicionalmente utilizados en hidráulica fluvial, no son adecuados debido a la compleja geometría de las regiones costeras. Debido a ello, los modelos bidimensionales de aguas poco profundas son habitualmente los más adecuados para el estudio de corrientes costeras. En este artículo se presenta un modelo en volúmenes finitos para el cálculo del flujo de marea en regiones costeras, centrándose en su aplicación a rías y estuarios. Se presentan las principales ventajas, inconvenientes y limitaciones del modelo para este tipo de aplicaciones, y se comparan algunos resultados numérico-experimentales.

Palabras clave: *Aguas someras, volúmenes finitos, frente seco-mojado, corrientes de marea, estuarios*

1 Introducción

La simulación numérica del flujo en estuarios es de gran importancia para entender, predecir y controlar los procesos físicos en zonas costeras. La dificultad de realizar ensayos de laboratorio, así como el coste económico de realizar mediciones experimentales en campo, hacen de los modelos numéricos una herramienta muy útil a la hora de estudiar la hidrodinámica de una zona costera.

Las ecuaciones de aguas someras promediadas en profundidad (2D-SWE) asumen una distribución de presión hidrostática, así como un perfil vertical de velocidades relativamente homogéneo. Ambas hipótesis se cumplen de manera razonable en estuarios y zonas costeras. Evidentemente, un modelo tridimensional proporcionaría unos resultados más precisos, pero a un coste computacional mucho más elevado, lo cual impide que en la actualidad se utilicen modelos tridimensionales para simular el flujo en estuarios extensos y complejos.

En este artículo se presenta la aplicación de un modelo en volúmenes finitos que resuelve las 2D-SWE al cálculo de la hidrodinámica de diversas zonas costeras. Se presentan 3 aplicaciones prácticas de modelización de corrientes de marea: el estuario Crouch (Reino Unido), la ría de O Barqueiro (Galicia) y la desembocadura del río Lérez en la ría de Pontevedra (Galicia).

2 Modelo numérico

2.1 Las ecuaciones de aguas someras promediadas en profundidad (2D-SWE)

Las 2D-SWE forman un sistema hiperbólico de 3 ecuaciones con 3 incógnitas, estando definidas sobre un dominio espacial bidimensional. En forma conservativa se pueden escribir como:

$$\begin{aligned} \frac{\partial h}{\partial t} + \frac{\partial q_j}{\partial x_j} &= 0 \\ \frac{\partial q_i}{\partial t} + \frac{\partial}{\partial x_j}\left(\frac{q_i q_j}{h} + \frac{g h^2}{2}\delta_{ij}\right) &= -gh\frac{\partial z_b}{\partial x_i} - gh\frac{n^2|q|q_i}{h^{10/3}} \end{aligned} \tag{1}$$

donde $q_i (i = 1, 2)$ son las dos componentes horizontales del caudal unitario, h es el calado, z_b es la altura del fondo, n es el coeficiente de Manning, y g es la aceleración de la gravedad. En la Ecuación (1) no se han considerado ni la aceleración de Coriolis, ni las variaciones de presión atmosférica, ni la fricción del viento, debido a que ninguna de estas fuerzas tiene un efecto apreciable en los casos presentados en este artículo. Tampoco se han incluido las tensiones de turbulentas. En todos los casos se ha realizado un análisis de sensibilidad de los resultados a la modelización de la turbulencia, utilizando un modelo de longitud de mezcla promediado en profundidad. Como suele ocurrir en la modelización del flujo en ríos y en zonas costeras, la influencia del modelo de turbulencia en los resultados de calado y velocidad es muy pequeña, en general inapreciable.

2.2 Esquema numérico

Para resolver las ecuaciones de aguas someras se ha utilizado un esquema numérico en volúmenes finitos para mallas no estructuradas. La discretización del dominio espacial se realiza con volúmenes finitos tipo arista. Una descripción detallada de este tipo de volúmenes se puede encontrar en [2]. Para la discretización del flujo convectivo se utiliza una extensión de orden 2 del esquema descentrado de Roe [9], con un limitador de pendiente (Superbee o Minmod) para evitar oscilaciones en regiones con gradientes elevados. Si se utiliza el esquema descentrado de Roe con una discretización centrada del término fuente pendiente del fondo, en problemas con batimetría irregular se generan oscilaciones espúrias en la superficie libre del agua, incluso en condiciones hidrostáticas [8, 1]. Para evitar estas oscilaciones debe utilizarse una discretización descentrada de la pendiente del fondo. Bermúdez y Vázquez proponen en [1] una discretización descentrada de la pendiente del fondo que

proporciona un balance exacto de las ecuaciones de flujo en el caso hidrostático cuando se utiliza con el esquema descentrado de Roe de primer orden. Sin embargo, cuando se utiliza la extensión de orden 2 del esquema de Roe, se generan oscilaciones de la superficie libre aunque se utilice la discretización del término fuente propuesta en [1]. Como posible solución, en [7] se propone utilizar la extensión de orden 2 únicamente para las dos componentes del caudal unitario (q_x, q_y), conservando una discretización de primer orden para el calado y la pendiente del fondo. Como resultado se obtiene un esquema híbrido de segundo orden en q_x y q_y, y de primer orden en h y z_b. De esta forma se elimina una gran parte de la difusión numérica, y se mantiene en gran medida la estabilidad del esquema.

El esquema descentrado de Roe genera oscilaciones espúrias en la superficie libre del agua si se aplica directamente en un frente seco-mojado. Para solucionar este problema Brufau [3] propone redefinir la elevación del fondo en el frente seco-mojado. En las simulaciones presentadas en este trabajo se ha utilizado la definición del fondo propuesta en [4], y se ha impuesto una condición de reflexión (flujo normal cero) en las aristas que definen el frente seco-mojado. La altura de agua nunca se fuerza a zero, con el fin de evitar pérdidas de masa en el interior del dominio de cálculo. Este tipo de tratamiento fue utilizado en [6] para simular la llegada de oleaje de onda larga a muros con pendiente elevada, proporcionando resultados aceptables, estables y con un frente no difusivo. Una descripción matemática del tratamiento de frentes seco-mojado similar se puede encontrar en [5].

3 Aplicación a rías y estuarios

3.1 El estuario Crouch

El estuario Crouch (Reino Unido) se caracteriza por tener una forma relativamente estrecha y alargada, con una extensión longitudinal de aproximadamente $25Km$, y una anchura de aproximadamente $1Km$ en la desembocadura (Figura 1). Existen numerosas zonas con una topografía relativamente plana e irregular que se anegan y drenan con cada ciclo de marea, generando bolsas de agua atrapada en cirtas depresiones del terreno en bajamar. Este tipo de topografía puede generar inestabilidades numéricas si el tratamiento de los términos fuente y del frente seco-mojado no es correcto.

La malla no estructurada utilizada en el modelo numérico consta de 48995 volúmenes finitos tipo arista, cubriendo una extensión espacial de aproximadamente $27{,}65Km^2$ con un tamaño medio de celda de $570m^2$.

No es necesario considerar las aportaciones externas de agua dulce en todo el estuario ya que estas son muy escasas. La única condición de contorno abierto a imponer es el nivel de marea en la desembocadura, el cual se ha obtenido directamente de una sonda de calado situada en la orilla norte de la desembocadura. En los contornos cerrados se utiliza una condición de deslizamiento libre. En todo caso, durante la mayor parte del tiempo los contornos están secos, y por lo tanto no intervienen en la solución. Debido

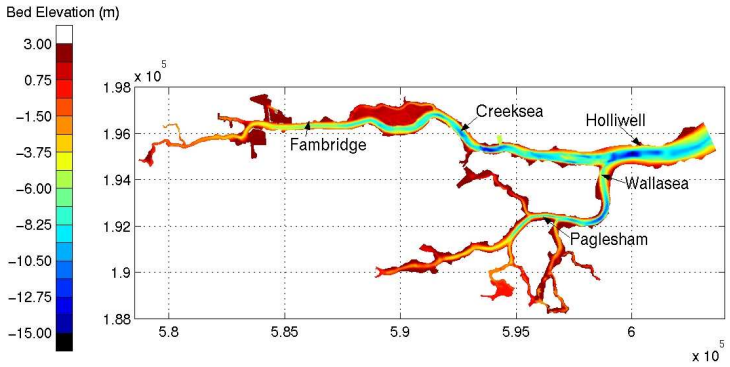


Figura 1: Batimetría $z_b(m)$. Relativa al nivel medio del mar en la desembocadura.

a ello se incrementa la importancia del frente seco-mojado, el cual define la extensión del fluido.

En general las velocidades inducidas por la marea en el estuario son relativamente elevadas, con valores superiores a $0{,}8m/s$ en una gran parte del estuario en marea entrante/saliente (Figura 2). Las velocidades máximas en el estuario se producen cerca de la desembocadura, con valores ligeramente superiores a $1{,}5m/s$. Las Figuras 1 y 2 muestran claramente que el campo de velocidad está muy determinado por la batimetría, siendo la velocidad mayor en las zonas más profundas.

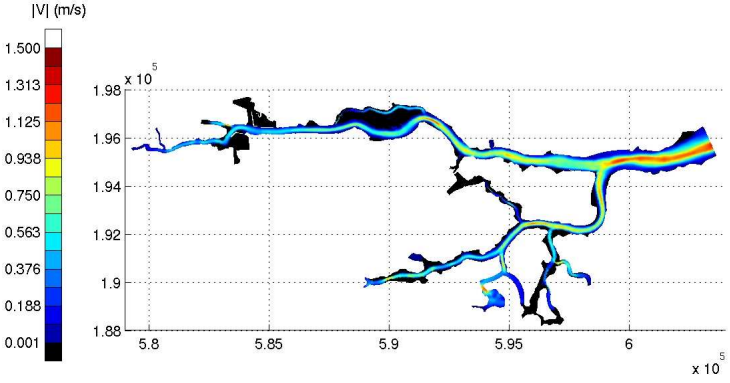


Figura 2: Campo de velocidad $|V|(m/s)$. $t = 45h$. Marea entrante.

La comparación entre resultados numéricos y experimentales en algunos de los puntos de medida (ver Figura 1) se muestra en la Figura 3. A fin de comparar con los resultados numéricos, la velocidad experimental se toma como representativa de la velocidad promediada en profundidad. Las predicciones del modelo numérico son muy satisfactorias.

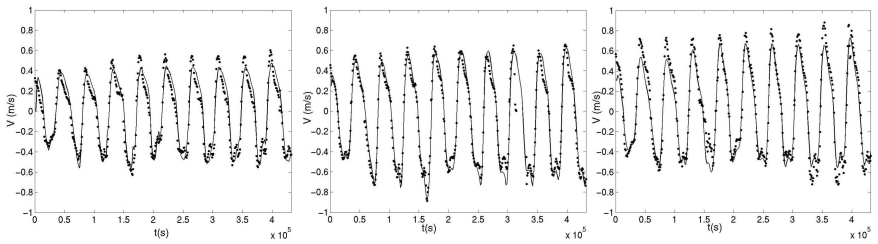


Figura 3: Series temporales de velocidad $V(m/s)$ en Holliwell (izquierda), Wallasea (centro) y Fambridge (derecha).

3.2 La ría de O Barqueiro

La forma de la ría de O Barqueiro es muy diferente a la del estuario Crouch. Con una longitud de aproximadamente $5Km$ y una anchura a la entrada de casi $3Km$, el flujo es mucho más bidimensional, produciéndose zonas de recirculación en el interior del estuario. El caudal medio aportado a la ría por el río Sor ($5{,}9m^3/s$) es varios órdenes de magnitud inferior al caudal de marea. Los efectos de la cuña salina son despreciables, pudiéndose considerar a nivel global que toda el agua en la ría es agua salada.

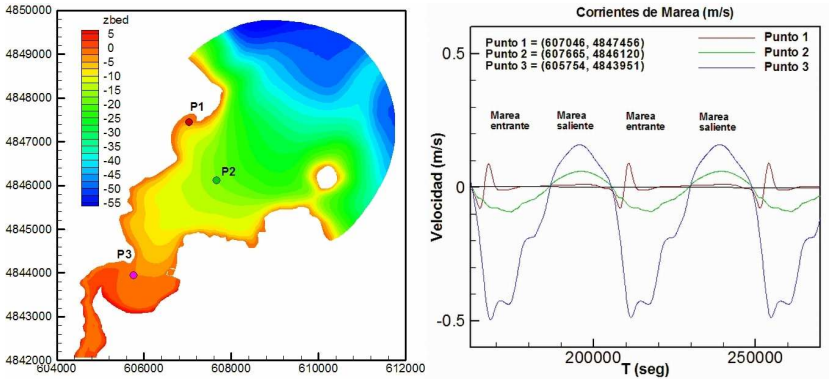


Figura 4: Batimetría $z_b(m)$ (izquierda). Series temporales de velocidad (derecha).

Se ha utilizado una malla no estructurada formada aproximadamente por 19000 volúmenes finitos tipo arista, la cual cubre una superficie de $24{,}7Km^2$. La malla comprende todo el interior de la ría, extendiéndose exteriormente a la ría en forma de semicircunferencia (Figura 4). La parte interior de la ría, con una extensión de $11{,}3Km^2$, comprende aproximadamente 14500 volúmenes finitos con un tamaño medio de $780m^2$.

Como condición de contorno en mar abierto se impone el nivel de marea, asumiéndolo constante en todo el contorno. Se toma un rango de marea de $4m$ con un período de 12 horas (marea semi-diurna). En el río Sor, se impone un caudal constante de $5{,}9m^3/s$. Se utiliza una condición de deslizamiento libre en

los contornos tipo pared.

Las velocidades inducidas por la marea (Figura 5) son mucho menores que en el estuario Crouch, manteniéndose en valores inferiores a $0{,}2m/s$ en casi toda la ría, excepto cerca de la desembocadura del río Sor, en donde existe una zona de bajos formada por la acumulación de arena, lo que provoca que se llegen a alcanzar valores de velocidad superiores a $1m/s$ con marea entrante. En esa zona existe una gran asimetría de velocidades entre marea entrante y marea saliente, produciéndose corrientes mucho más fuertes cuando sube la marea (Figuras 4 y 5). Esto se debe a la existencia de los mencionados bajos de arena, y contribuye asímismo a la acumulación de arena en dicha zona, produciéndose un efecto de *feedback*. Aunque no se dispone de medidas experimentales de velocidad, estos resultados concuerdan con observaciones visuales.

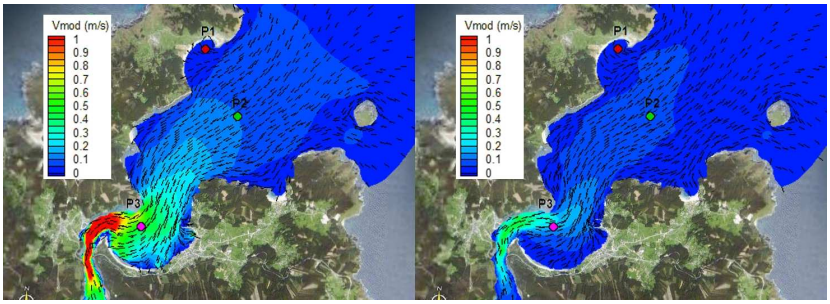


Figura 5: Campo de velocidad $|V|(m/s)$. Marea entrante $T = 3h$ (izquierda). Marea saliente $T = 9h$ (derecha).

3.3 Desembocadura del río Lérez

A su paso por la ciudad de Pontevedra el calado del río Lerez se encuentra condicionado por el nivel de marea en la ría de Pontevedra. Se ha modelado el flujo en un tramo del río Lerez de aproximadamente $1Km$ de longitud, a su paso por Pontevedra. La sección transversal del río en dicho tramo es variable, con anchura del cauce comprendida entre $65m$ y $125m$ (Figura 6). Se estudia la capacidad de arrastre de la corriente en condiciones de avenida. En dichas condiciones predomina el caudal del río sobre el caudal de marea. La velocidad de la corriente está dirigida hacia aguas abajo durante todo el ciclo de marea, pudiéndose considerar que toda el agua es dulce, i.e. no existe cuña salina en la zona estudiada.

La discretización espacial se realiza mediante una malla no estructurada de volúmenes finitos tipo arista compuesta por 12827 nodos de cálculo, cubriendo una superficie total de $105727{,}7m^2$. La Figura 6 muestra la batimetría utilizada en el modelo numérico.

En el contorno aguas arriba se impone el caudal total que entra en el tramo proveniente del río Lérez en condiciones de avenida máxima anual media ($Q_{max,medio} = 354{,}3m^3/s$). Dicho caudal se distribuye en toda la sección de entrada de manera proporcional a la profundidad en cada punto del contorno,

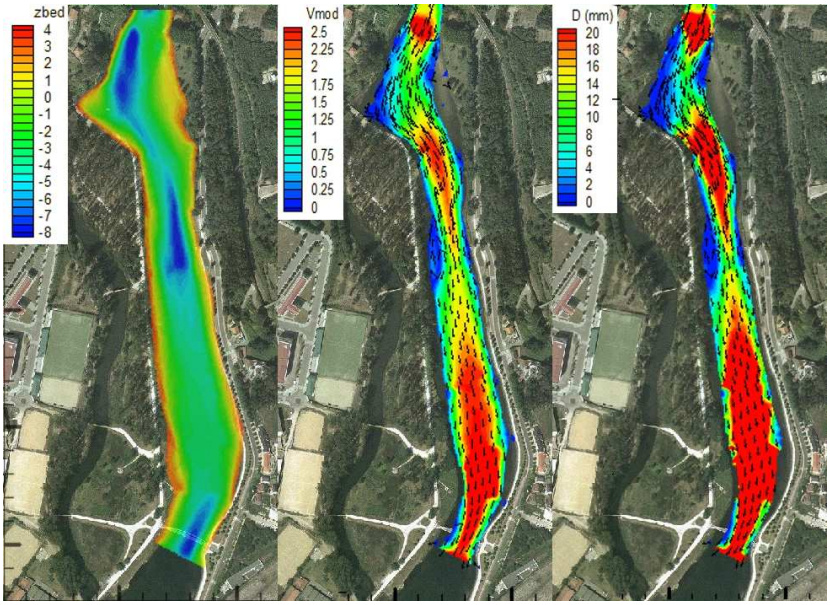


Figura 6: Río Lerez. Batimetría $z_b(m)$ (izquierda). Velocidad en bajamar $V_{mod}(m/s)$ (centro). Diámetro crítico de arrastre $D_{50}(mm)$ (derecha).

imponiendo un mayor caudal unitario en las zonas más profundas. En la sección de salida se impone la elevación de la superficie libre constante en todo el contorno, la cual viene dada por el nivel de marea. La situación pésima desde el punto de vista de arrastre de sedimentos se produce para el nivel de marea mínimo, ya que es en esa situación donde se producen velocidades máximas para desaguar el caudal de cálculo. Dicho nivel se ha determinado por métodos estadísticos extremales a partir de los registros de un mareografo. A falta de datos más precisos, se asume un período de marea de 12 horas (la marea en la ría de Pontevedra es fundamentalmente semidiurna) con una carrera de marea igual a la máxima observada (4.23m).

En la Figura 6 se muestran los campos de velocidad en condiciones de bajamar para el caudal de cálculo. Al contrario que en el estuario Crouch y en la ría de O Barqueiro, donde las corrientes máximas se producen con marea entrante/saliente, en este caso las máximas velocidades se producen en bajamar, cuando los calados y la sección mojada del río son mínimos. Se han calculado las tensiones de fondo mediante la fórmula de Manning y a partir de estas, en cada punto del modelo se evalúa mediante el ábaco de Shields el diámetro de sedimento que es capaz de soportar dicha tensión sin que se produzca transporte de fondo. Se calcula dicho diámetro crítico como:

$$D_{50}(mm) = \frac{1000\tau_b(N/m^2)}{0{,}05\ 1650\ 9{,}8} \tag{2}$$

En la Figura 6 se muestra, para condiciones de bajamar, el diámetro crítico calculado. Como puede apreciarse, para avenidas medias existen zonas del

tramo en las que se producirá transporte de sólidos incluso superiores a $20mm$, mientras que en otras zonas donde la sección es más ancha el sólido estable está entre 2 y $4mm$.

4 Conclusiones

Se ha presentado la simulación numérica del flujo en diferentes zonas costeras por medio de un modelo de volúmenes finitos que resuelve las ecuaciones de aguas someras promediadas en profundidad con un tratamiento del frente de marea estable y no difusivo. A pesar de que en este artículo no se ha incidido en la modelización de la turbulencia, en todos los casos presentados se ha realizado un análisis de sensibilidad de los resultados al modelo de turbulencia, utilizando un modelo de longitud de mezcla promediado en profundidad. Como suele ocurrir en la simulación del flujo en ríos y en zonas costeras, la influencia del modelo de turbulencia en los resultados de calado y velocidad es muy pequeña, en general inapreciable. A pesar de ello, es necesario remarcar que la turbulencia juega un papel fundamental en el tranporte de sustancias solubles y de sedimentos y por lo tanto, su correcta modelización es fundamental para la simulación de procesos de transporte y mezcla.

Cuando se estudian corrientes de marea es necesario realizar un *calentamiento* previo del modelo numérico, de forma que la masa de agua adquiera cierta inercia. En todos los casos presentados se ha realizado un calentamiento consistente en 2 ciclos de marea, lo cual suele ser suficiente, realizando el análisis de resultados a partir del tercer ciclo. En todos los casos estudiados se puede considerar flujo monofásico, ya sea de agua dulce o de agua salada. En otras situaciones puede ser necesario tener en cuenta en la modelización la existencia de una cuña salina mediante un modelo bicapa. La existencia, tamaño y forma de la cuña salina depende de la geometría del estuario, así como de la relación entre el caudal de marea y el caudal del río.

El ajuste numérico-experimental de velocidades y calados es bastante satisfactorio, confirmando la capacidad de los modelos de aguas someras bidimensionales de modelar los procesos de inundación y drenaje generados por el flujo de marea en estuarios con una geometría y topografía complejas.

Agradecimientos

Todos los resultados experimentales presentados en el estuario Crouch fueron obtenidos por el grupo CERU (University College of London).

Referencias

[1] A. Bermúdez and M.E. Vázquez-Cendón, *Upwind methods for hyperbolic conservation laws with source terms*. Comput. Fluids, Vol. **23**(8), pp. 1049, (1994).

[2] A. Bermúdez, A. Dervieux, J.A. Desideri and M.E. Vázquez-Cendón. Upwind schemes for the two-dimensional shallow water equations with variable depth using unstructured meshes, *Comput. Methods Appl. Mech. Eng.*, Vol. **155**, pp. 49–72, (1998).

[3] P. Brufau, *Simulación bidimensional de flujos hidrodinámicos transitorios en geometrías irregulares*, Tesis doctoral, Área de Mecánica de Fluidos, Universidad de Zaragoza, (2000).

[4] P. Brufau, M.E. Vázquez-Cendón y P. García-Navarro. A numerical model for the flooding and drying of irregular domains, *Int. J. Numer. Meth. Fluids*, Vol. **39**(3), pp. 247-275, (2002).

[5] M.J. Castro, J.M. González y C. Parés. Numerical treatment of wet/dry fronts in shallow flows with a modified Roe scheme, *Math. Mod. Meth. App. Sci.*, Vol. **16**(6), pp. 897-931, (2006).

[6] L. Cea, A. Ferreiro, M.E. Vázquez-Cendón y J. Puertas. Experimental and numerical analysis of solitary waves generated by bed and boundary movements, *Int. J. Numer. Meth. Fluids*, Vol. **46**(8), pp. 793-813, (2004).

[7] L. Cea, J. French y M.E. Vázquez-Cendón. Numerical modelling of tidal flows in complex estuaries including turbulence: an unstructured finite volume solver and experimental validation, *Int. J. Numer. Meth. Engineering*, Vol. **67**(13), pp. 1909-1932, (2006).

[8] E.D. Fernández-Nieto, *Aproximación numérica de leyes de conservación hiperbólicas no homogéneas*, Tesis doctoral, Universidad de Sevilla, (2003).

[9] P.L. Roe. Discrete models for the numerical analysis of time-dependent multidimensional gas dynamics, *J. Comput. Phys.*, Vol. **63**, pp. 458–476, (1986).

Bol. Soc. Esp. Mat. Apl.
nº 42(2008), 127–135

NUMERICAL MODELING OF BUOYANT TURBULENT MIXING LAYERS

AC. BENNIS*, T. CHACÓN REBOLLO†, M. GÓMEZ MÁRMOL† AND R. LEWANDOWSKI*

* IRMAR, Université de Rennes 1, Campus de Beaulieu, 35042 Rennes Cedex, France.

† Departamento de Ecuaciones Diferenciales y Análisis Numerico, Universidad de Sevilla. C/Tarfia, s/n.41080, Sevilla, Spain

Abstract

We introduce in this paper some elements for the mathematical and numerical analysis of turbulence models for oceanic surface mixing layers. In these models the turbulent diffusions are parameterized by means of the Richardson's number, that measures the balance between stabilizing buoyancy forces and un-stabilizing shearing forces. The well-possedness of these models is a difficult mathematical problem, due to the partial monotonic nature of the space operators involved. We analyze the existence and stability of equilibria state, and devise a conservative numerical scheme satisfying the maximum principle. We present some numerical tests for realistic flows in tropical seas that reproduce the formation of mixing layers, in agreement with the physics of the problem.

Key words: *Turbulent mixing layers, Richardson's number, First order closure models, Conservative numerical solution, Stability of steady states, Tests for tropical seas*

AMS subject classifications: *0123 1234*

1 Introduction

This paper is devoted to the mathematical and numerical analysis of turbulence models of surface oceanic mixing layers. The wind-stress generates intense mixing processes in a layer below the ocean surface. This layer has two parts, the upper one is an homogeneous layer, known as the mixed layer. This layer presents almost-constant temperature (and salinity). The bottom of the mixed layer corresponds to the top of the thermocline. In tropical seas a sharp thermocline is formed. Below this layer appears a thinner layer where still mixing processes do occur, but which has not a homogeneous structure. The

Research of T. Chacón and M. Gómez partially funded by Spanish DIG grants MTM2006-01275.

zone formed by the two layers is known as the mixing layer. Its thickness may vary between ten meters and a few hundred of meters, depending on the latitude. It also presents seasonal variations.

The parametrization of turbulence in the mixing layer must take into account the two forces that act in the momentum and mass exchange produced by mixing effects: Buoyancy and shear. This introduces additional complexities with respect to the usual modeling of turbulent flows with constant density, from both the physical and the mathematical standpoints. Closure terms are now parameterized in terms of the Richardson number (that measures the balance between stabilizing buoyancy forces and un-stabilizing shearing forces), that in this sense plays a role similar to that of the Reynolds number, used to parameterize closure terms for constant-density turbulence.

In this paper we introduce some mathematical and numerical elements for the analysis of the simplest turbulence models of mixing layers. These are first order closure models: Pacanowski and Philander model (called PP model, 1981, [6]) and the Large and Gent model (called KPP model, 1994, [3]) (Section 2). Let us mention that second order models have been developed by Mellor and Yamada (called MY model, 1982, [5]) and Gaspar et al. (1990,[1]). These models are widely used in physical oceanographic applications, but have received few attention from the mathematical community.

We observe that, in despite of their apparent simplicity, the well-possedness of first-order models is a difficult mathematical problem due to the partial monotonic nature of the space operators involved (Section 3). We analyze the existence of equilibria states, proving that these necessarily correspond to linear profiles of velocity and temperature (or salinity) (Section 4). We also analyze the stability of these equilibria, and prove that at least one is stable for vertical stable configurations. We introduce a new model that has just one equilibrium state (Section 5). We next devise a conservative numerical scheme for which we prove a maximum principle (Section 6). We finally present some numerical tests for realistic flows in tropical seas that reproduce the formation of mixing layers, in agreement with the physics of the problem. We stress that our new models produces results very close to the PP one, and in addition is able to handle unstable profiles (Section 7).

2 Setting of model problems

Typically, the variables used to describe the mixing layer are the statistical means of density and velocity (denoted by u and ρ). In the ocean, density =function(temperature, salinity) (State equation). We shall consider the density as an idealized thermodynamic variable.

We assume

$$U = (u(z,t), 0, w(z,t)), \quad p = p(z,t), \quad \rho = \rho(z,t)$$

and neglect Coriolis forces (hypothesis accurate for tropical oceans) and laminar diffusion (which will be absorbed by eddy diffusion). Then the averaged Navier-

Stokes equations reduce to

$$\begin{cases} \dfrac{\partial u}{\partial t} = -\dfrac{\partial}{\partial z}\left\langle u'\, w'\right\rangle, \\ \dfrac{\partial \rho}{\partial t} = -\dfrac{\partial}{\partial z}\left\langle \rho'\, w'\right\rangle, \end{cases} \tag{1}$$

To close these equations, we use the concept of eddy diffusion:

$$-\left\langle u'\, w'\right\rangle = \nu_1 \frac{\partial u}{\partial z}, \quad -\left\langle \rho'\, w'\right\rangle = \nu_2 \frac{\partial \rho}{\partial z}.$$

Coefficients ν_1 and ν_2 are expressed as functions of the **gradient Richardson number** R **defined as**

$$R = -\frac{g}{\rho_{ref}} \frac{\dfrac{\partial \rho}{\partial z}}{\left(\dfrac{\partial u}{\partial z}\right)^2}$$

Note that R is the ratio between the stabilizing vertical forces due to buoyancy and the un-stabilizing horizontal ones due to shear in a water column.

When $R >> 1$, a strongly stratified layer takes place. This correspond to a stable configuration. When $0 < R << 1$, a slightly stratified layer takes place. This correspond to a configuration with low stability. The case $R < 0$ corresponds to a configuration statically unstable ($\dfrac{\partial \rho}{\partial z} > 0$), that in fact we are not modeling. However, we must handle this situation for our numerical experiments. A simple way is to set large constant values for the turbulent diffusions in this case.

The set of equations, initial and boundary conditions governing the mixing layer can now be written

$$\begin{cases} \dfrac{\partial u}{\partial t} - \dfrac{\partial}{\partial z}\left(\nu_1 \dfrac{\partial u}{\partial z}\right) = 0, \\ \dfrac{\partial \rho}{\partial t} - \dfrac{\partial}{\partial z}\left(\nu_2 \dfrac{\partial \rho}{\partial z}\right) = 0, \text{ for } t \geqslant 0 \text{ and } -h \leqslant z \leqslant 0, \\ u = u_b,\ \rho = \rho_b \text{ at the depth } z = -h, \\ \nu_1 \dfrac{\partial u}{\partial z} = V,\ \nu_2 \dfrac{\partial \rho}{\partial z} = Q \text{ at the surface } z = 0, \\ u = u_0,\ \rho = \rho_0 \text{ at initial time } t = 0. \end{cases} \tag{2}$$

Here, V is the forcing exerced by the wind-stress ($V = \dfrac{\rho_{air}}{\rho_{ref}} C_{friction} \left|U^{air}\right|^2$), and Q represents the thermodynamical fluxes, heating or cooling, precipitations or evaporation.

To model the turbulent diffusions in terms of the Richardson number, a central idea is that a stable configuration due to buoyancy forces inhibits the turbulent exchange of mass and momentum. Pacanowski and Philander [6] propose

$$\nu_2(R) = \frac{Constant}{(1+\sigma R)^n}\ \nu_1(R).$$

This leads to the modeling $\quad \nu_1 = f_1(R), \quad$ and $\nu_2 = f_2(R)$, with

$$f_1(R) = \alpha_1 + \frac{\beta_1}{(1+5R)^2}, \quad f_2(R) = \alpha_2 + \frac{f_1(R)}{1+5R}, \quad \text{for PP model, and}$$

$$f_1(R) = \eta_1 + \frac{\gamma_1}{(1+10R)^2}, \quad f_2(R) = \eta_2 + \frac{\gamma_2}{(1+10R)^3} \quad \text{for KPP model.}$$

The constants are chosen to fit numerical results with experimental measurements, these are given by $\alpha_1 = 1.10^{-4}$, $\beta_1 = 1.10^{-2}$, $\alpha_2 = 1.10^{-5}$, and $\eta_1 = 1.10^{-4}$, $\gamma_1 = 1.10^{-1}$, $\eta_2 = 1.10^{-5}$, $\gamma_2 = 1.10^{-1}$ (units: $m^2 s^{-1}$).

3 Well-possedness

Some elements for the analysis of the well-possedness of problem (2) are deduced from the analysis of monotonicity of the space operator appearing in it.

Let us assume that the functions f_i are bounded C^1 functions, with

$$f_i'(R) \leq 0, \quad (i = 1, 2). \tag{3}$$

Denote $\mathbf{v} = (\rho, u)^T$, $\mathbf{V} = (Q, V)^T$, $M = M(R) = \begin{pmatrix} f_1(R) & 0 \\ 0 & f_2(R) \end{pmatrix}$. For any function $a = a(t, z)$, we shall denote $\partial_z a = \dfrac{\partial a}{\partial z}$, $\quad \partial_t a = \dfrac{\partial a}{\partial t}$.

Thus, our system can be written under the form (we assume homogeneous Dirichlet boundary conditions for simplicity),

$$\partial_t \mathbf{v} - \partial_z (M(R)\partial_z \mathbf{v}) = 0, \tag{4}$$

$$M(R)\partial_z \mathbf{v}|_{z=0} = \mathbf{V}, \quad \mathbf{v}|_{z=-h} = \mathbf{0}, \tag{5}$$

$$\mathbf{v}|_{t=0} = \mathbf{v}_0. \tag{6}$$

Let now $A = A(\mathbf{v})$ and $\mathbf{F}$ be defined by

$$(A(\mathbf{v}), \mathbf{w}) = \int_{-h}^{0} M(R)\partial_z \mathbf{v}.\partial_z \mathbf{w} = (M(R)\partial_z \mathbf{v}, \partial_z \mathbf{w}),$$
$$(\mathbf{F}, \mathbf{w}) = \mathbf{V}.\mathbf{w}(0).$$

Therefore system (4) − (5) − (6) is a system of the form

$$\frac{d\mathbf{v}}{dt} + A(\mathbf{v}) = \mathbf{F}, \quad \mathbf{v}(0) = \mathbf{v}_0,$$

in the sense that $\forall\, \mathbf{w} \in H^2$

$$\frac{d}{dt}(\mathbf{v}, \mathbf{w}) + (A(\mathbf{v}), \mathbf{w}) = (\mathbf{F}, \mathbf{w})$$

where the space H is defined by $H = \{u \in H^1([-h, 0]), \quad u(-h) = 0\}$.

We want to use the theory of monotonic operators to analyze the well-possedness of this equation. We intend to prove that the operator A is monotonic, in the sense that

$$\forall\,(\mathbf{v}_1,\mathbf{v}_2)\in H^{2\times 2},\quad (A(\mathbf{v}_1)-A(\mathbf{v}_2),\mathbf{v}_1-\mathbf{v}_2)\geq 0.$$

In the actual stage of our research, we are able to prove that under condition (3) indeed we have

$$\forall\,(\mathbf{v}_1,\mathbf{v}_2)\in H^2\times H^2,\quad (A(\mathbf{v}_1)-A(\mathbf{v}_2),\mathbf{v}_1-\mathbf{v}_2)\geq C_K\|\partial_z\mathbf{v}_1-\partial_z\mathbf{v}_2\|^2_{L^2(-h,0)}, \tag{7}$$

if $\mathbf{v}_1,\mathbf{v}_2$ belong to a neighborhood K of the origin. We hope that this will allow to prove a well-posseddness result for small data.

4 Equilibria states

Although we are not able to analyze the model system (2) in general, it is possible to study some properties of equilibria states. Let us consider the stationary model system:

$$\frac{\partial}{\partial z}\left(f_1(R)\,\frac{\partial u}{\partial z}\right)=0,\qquad \frac{\partial}{\partial z}\left(f_2(R)\,\frac{\partial \rho}{\partial z}\right)=0. \tag{8}$$

Integrating (8) with respect to z we obtain

$$\begin{cases} f_1(R)\,\dfrac{\partial u}{\partial z} &=& \text{constant}=V \quad \text{(momentum flux)},\\ f_2(R)\,\dfrac{\partial \rho}{\partial z} &=& \text{constant}=Q \quad \text{(heat flux)}. \end{cases} \tag{9}$$

Using the expression (2), we deduce an **implicit equation** for R

$$R=-\frac{g}{\rho_0}\,\frac{Q}{V^2}\,\frac{(f_1(R))^2}{f_2(R)}$$

If this equation has a solution R^e, this reads

$$\frac{\text{Potential energy}}{\text{Turbulent kinetic energy}}(\mathit{Equilibrium})=\frac{-Q}{V^2}\;\times\;\text{Constant}(R^e).$$

Note that from (9), the equilibria states are linear profiles for both velocity and density.

PP and KPP models present **several equilibria** R^e for a range $[r^*,+\infty)$ of fluxes ratio $r=-Q/V^2$, where r^* is negative. This corresponds to static instability. So, these model include as mathematical equilibria some physical static unstable configurations.

To avoid the multiplicity of steady states, we introduce a new model, given by

$$f_1(R) = \alpha_1 + \frac{\beta_1}{(1+5R)^2}, \quad f_2(R) = \alpha_2 + \frac{f_1(R)}{(1+5R)^2},$$

with the same constants as the PP model. This new model has **a unique equilibrium** R^e for any fluxes ratio r. This is a mathematically favorable property, still without physical meaning when $r < 0$.

5 Stability of equilibria states

We analyze the linear stability of equilibria states. To do it, we construct a model of time evolution of a small perturbation of a equilibrium state (u^e, ρ^e):

$$(u, \rho) = (u^e, \rho^e) + (u', \rho')$$

Set $\psi = \dfrac{\partial \rho}{\partial z}$ and $\theta = \dfrac{\partial u}{\partial z}$, and so $R = R(\theta, \psi)$, $\nu_i = \nu_i(\theta, \psi)$. The equations satisfied by the perturbation (u', ρ') are deduced from model equations :

$$\begin{cases} \dfrac{\partial u'}{\partial t} - \dfrac{\partial}{\partial z}\left(\nu_1(\theta,\psi)\ (\theta^e + \theta')\right) = 0, \\ \dfrac{\partial \rho'}{\partial t} - \dfrac{\partial}{\partial z}\left(\nu_2(\theta,\psi)\ (\psi^e + \psi')\right) = 0. \end{cases} \tag{10}$$

The linearized equations for (u', ρ') then are

$$\frac{\partial V}{\partial t} - A\frac{\partial^2 V}{\partial z^2} = 0, \text{with} \quad V = \begin{pmatrix} u' \\ \rho' \end{pmatrix}, \tag{11}$$

where A is the **amplification matrix**,

$$A = \begin{pmatrix} \nu_1^e + \theta^e \left(\dfrac{\partial \nu_1}{\partial \theta}\right)^e, & \theta^e \left(\dfrac{\partial \nu_1}{\partial \psi}\right)^e \\ \psi^e \left(\dfrac{\partial \nu_2}{\partial \theta}\right)^e, & \nu_2^e + \psi^e \left(\dfrac{\partial \nu_2}{\partial \psi}\right)^e \end{pmatrix}.$$

Linear stability of the equilibrium solution (u^e, ρ^e) follows if any perturbation (u_0', ρ_0') imposed at initial time $t = 0$ is damped as $t \to \infty$. This is verified if the eigenvalues λ_1, λ_2 of A are such that $Re(\lambda_1) > 0$ and $Re(\lambda_2) > 0$. After some algebra, we conclude that all models are linearly stable for $R^e > 0$. But strikingly also for a small range $[R^*, 0]$ with $R^* < 0$, which (we recall) corresponds to physically unstable configurations.

We have also investigated the non-linear stability of our models. To do it, we have solved numerically the full non-linear system (2) starting from small and even large perturbations of equilibria states. We have used the numerical scheme described in the next section. Our conclusions also are that for all models the equilibria states are non-linearly stable, and, even more, behave as strong attractors. The typical time that a given initial state takes to approach an equilibrium state is of the order of several months. This must be compared with the typical time of formation of the thermocline, which is of a few days.

6 Numerical discretization

We have performed a centered conservative semi-implicit discretization of the PDEs appearing in model (2)by finite differences. To describe it, assume that the interval $[-h, 0]$ is divided into N subintervals of length $\Delta z = h/(N-1)$, with nodes $z_i = -(i-1)\,h\,\Delta z,\ i = 1, \cdots, N$. We respectively approximate the values $u(z_i, t_n)$, $\rho(z_i, t_n)$ by u_i^n and ρ_i^n, where $t_n = n\Delta t$. The equation for u, for instance, is discretized at node z_i, with $i = 2, \cdots, N-1$ by

$$\frac{u_i^{n+1} - u_i^n}{\Delta t} -$$

$$\frac{f_1(R^n_{i-1/2})u^{n+1}_{i-1} - \left[f_1(R^n_{i-1/2}) + f_1(R^n_{i+1/2})\right] u_i^{n+1} + f_1(R^n_{i+1/2})u^{n+1}_{i+1}}{(\Delta z)^2} = 0,$$

where

$$R^n_{i-1/2} = -\frac{g}{\rho_{ref}}\,\frac{(\rho_i^n - \rho^n_{i-1})/\Delta z}{\left[(u_i^n - u^n_{i-1})/\Delta z\right]^2},$$

and a similar discretization for the equation for ρ. The boundary conditions have been discretized by

$$u_1^{n+1} = u_b^{n+1}, \quad \rho_1^{n+1} = \rho_b^{n+1};$$

$$f_1(R^n_{N-1/2})\,\frac{u_N^{n+1} - u^{n+1}_{N-1}}{\Delta z} = V_N^{n+1}.$$

This last equation allows to compute u_N^{n+1} from u^{n+1}_{N-1}. So we may construct our discretization in terms of the unknowns $U^{n+1} = (u_2^{n+1}, \cdots, u^{n+1}_{N-1})$ and similarly for ρ. In matrix form, this discretization reads

$$A^{n+1}\,U^{n+1} = B^{n+1},$$

where A^{n+1} and B^{n+1} respectively are the tridiagonal matrix and the vector array defined with obvious notation by

$$A^{n+1}_{i-1,i} = -\alpha^n_{i-1/2},\ A^{n+1}_{i,i} = 1+\alpha^n_{i-1/2}+\alpha^n_{i+1/2},\ A^{n+1}_{i+1,i} = -\alpha^n_{i+1/2}, \quad i = 2, \cdots, N-2;$$

$$A^{n+1}_{N-2,N-1} = -\alpha^n_{N-3/2}, \quad A^{n+1}_{N-1,N-1} = 1 + \alpha^n_{N-3/2};$$

$$B^{n+1} = (u_2^n + \alpha^n_{3/2}\,u_b^n, u_3^n, \cdots, u_i^n, \cdots, u^n_{N-2}, u^n_{N-1} + \frac{\Delta t}{\Delta z}\,V)^t,$$

where

$$\alpha^n_{i-1/2} = \frac{\Delta t}{(\Delta z)^2}\,f_1(R^n_{i-1/2}).$$

As $f_1 \geq 0$, A^{n+1} is an M-matrix and then $(A^{n+1})^{-1}$ has positive entries. Then, we deduce a *maximum principle:* If the initial data, u_b^n and V are positive, the u_i^n are all positive.

7 Numerical tests

We have simulated some realistic flows, corresponding to the Equatorial Pacific region called the West-Pacific Warm Pool, located at the equator between $120°E$ and $180°E$. In this region the sea temperature is high and almost constant along the year ($28-30°C$). The precipitations are intense and hence the salinity is low. We initialize the code with data from the TAO (Tropical Atmosphere Ocean)

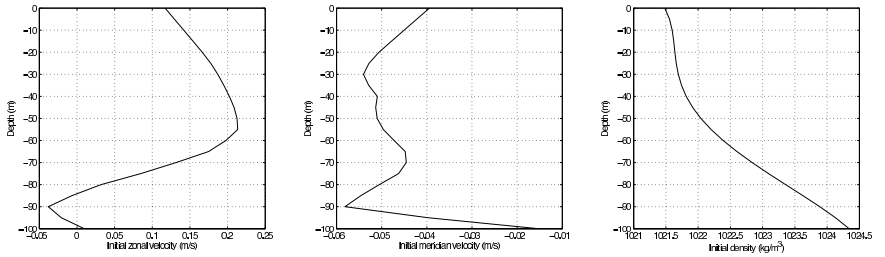


Figure 1: Initial zonal velocity, meridian velocity and density profiles (from left to right).

array (McPhaden [4]), which have been used in many numerical simulations.

Here, we present the results corresponding to a mixed layer induced by the wind stress, using initial velocity and density profiles measured at $0°N, 165°E$ for the time period between the 15th June 1991 and the 15th July 1991, displayed in Figure 1. Observe that the density profile does not present a mixed layer.

We used a two-dimensional version of model (2), with buoyancy flux equal to $-1.10^{-6}\ kg.m^{-2}.s^{-1}$ ($\simeq -11\ W/m^2$), which is realistic for this region (Cf. [2]). We have taken as boundary conditions, a zonal wind (u_1) equal to 8.1 m/s (eastward wind) and a meridional wind (u_2) equal to 2.1 m/s (northward wind). These values are larger than the measured ones, because we want to force the formation of a mixed layer. We have used $\Delta z = 1m$ and $\Delta t = 60s$. The results are grid-independent, in the sense that they remain practically unchanged when Δz and Δt decrease.

Figure 2 displays the results corresponding to $t = 48$ hours. On top we represent the whole mixing layer, and on bottom, the upper 40m of layer. The plots for density profiles show the formation of a pycnocline at $z = -30$, approximately. Velocity and density profiles are quite close for PP and the new model, while the velocity provided by KPP model is somewhat different, mainly near the surface. Also, the density profiles and the pycnocline simulated by the three models are quite similar.

Let us remark that our new model is the one that introduces the smallest levels of turbulent viscosity and diffusion. It is also able to simulate non-stable initial profiles, providing physically coherent results.

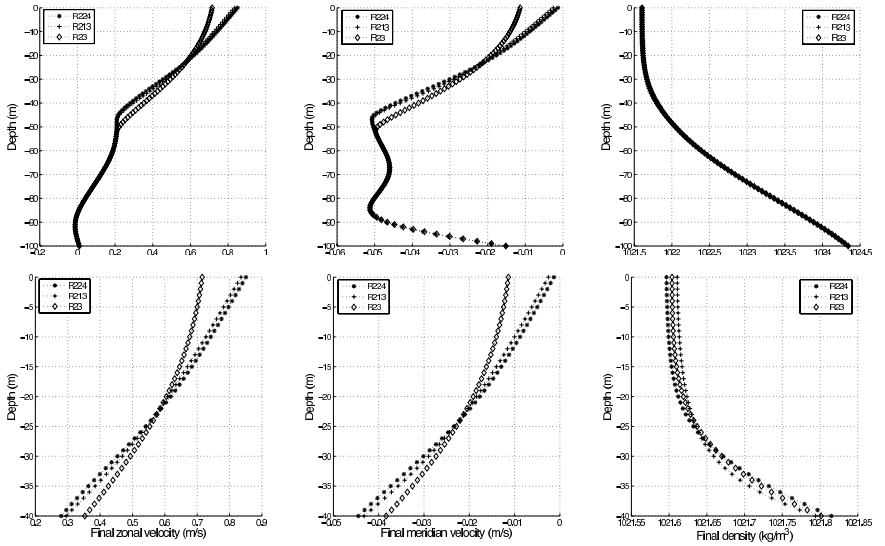


Figure 2: Comparison of three turbulence models: R213 (PP), R224 (KPP), R23 (new one) (In this notation Rijk, i, j and k represent the exponents of the denominators in the definition of the turbulent diffusions f_1 and f_2). Top: Full mixing layer. Bottom: Upper 40m of mixing layer.

References

[1] P. Gaspar, Y. Gregoris and L. J. M. A simple eddy kinetic energy model for simulations of oceanic vertical mixing: test at station papa and long-term upper ocean study site. *J. Geophys. Research.* Vol 16, 179-193. 2001.

[2] P. R. Gent. The heat budget of the toga-coare domain in an ocean model. *J. Geophys. Res.* Vol 96, 3323-3330. 1991.

[3] W. G. Large, C. McWilliams and S. C. Doney. Oceanic vertical mixing: a review and a model with a non-local boundary layer parametrization. *Rev. Geophys.* Vol. 32, 363-402. 1994

[4] M. McPhaden. The tropical atmosphere ocean (tao) array is completed. *Bull. Am. Meterorol. Soc.* Vol. 76, 739-741. 1995.

[5] G. Mellor and T. Yamada. Development of a turbulence closure model for geophysical fluid problems. *Reviews of Geophysics and Space Physics.* Vol. 20, 851-875. 1982.

[6] R. C. Pacanowski and S. G. H. Philander. Parametrization of vertical mixing in numerical models of the tropical oceans. *J. Phys. Oceanogr.* Vol 11, 1443-1451. 1981.

Bol. Soc. Esp. Mat. Apl.
nº 42(2008), 137–146

SIMULACIÓN DE CORRIENTES DE MAREA EN EL ESTRECHO DE GIBRALTAR MEDIANTE MODELOS BICAPA 2D DE AGUAS SOMERAS

J. M. GONZÁLEZ-VIDA, MANUEL J. CASTRO, J. A. GARCÍA-RODRÍGUEZ, J. MACÍAS, C. PARÉS

Grupo EDANYA.
Universidad de Málaga.

vida@anamat.cie.uma.es

Resumen

En este trabajo se muestran los últimos resultados que el grupo de investigación EDANYA de la Universidad de Málaga ha obtenido sobre la simulación de la hidrodinámica interna del Estrecho de Gibraltar. Inicialmente se presenta el problema físico, para después plantear un modelo numérico basado en volúmenes finitos bien adaptado para su resolución. Finalmente se presentan algunas simulaciones realizadas con dicho modelo, así como comparaciones con observaciones realizadas en la zona.

Palabras clave: *Ecuaciones de aguas someras 2D, métodos de volúmenes finitos, paralelización y vectorización, hidrodinámica del Estrecho de Gibraltar.*
Clasificación por materias AMS: *65M60 65Y05 65Y10 76B15 76B55 76B75*

1 Motivación.

En este trabajo, nuestro interés se centra en la obtención de un modelo numérico bien adaptado para representar la compleja hidrodinámica que genera el permanente intercambio de flujos existente en el Estrecho de Gibraltar.

El Estrecho de Gibraltar es el único enlace natural entre el océano Atlántico y el mar Mediterráneo. El mar Mediterráneo está sometido a una fuerte tasa de evaporación por lo que sus aguas son más densas que las del Atlántico. Esto, junto con los efectos de mareas, forzamiento atmosférico, la abrupta topografía del Estrecho, etc. produce una dinámica interna de intercambio a través del Estrecho de Gibraltar muy compleja: el agua atlántica, menos salina y más fría penetra en el Mediterráneo en superficie, mientras que el agua Mediterránea, más densa, sale hacia el Atlántico en profundidad. En la Figura 1 se muestra en forma esquemática la situación:

Trabajo subvencionado por el proyecto DGI número MTM2006-08075.

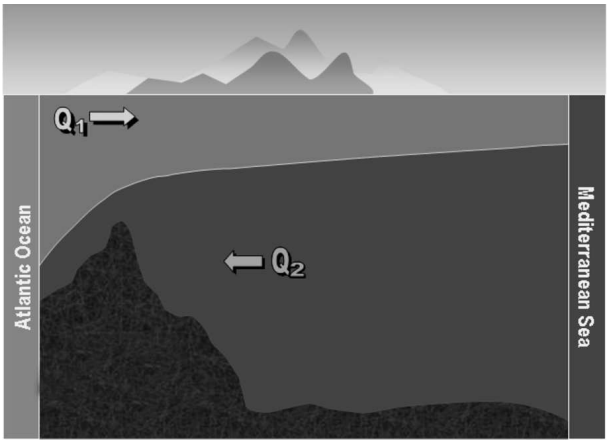


Figura 1: Esquema de la dinámica en el Estrecho de Gibraltar

En estas circunstancias, se puede llevar a cabo una aproximación de la dinámica del Estrecho de Gibraltar suponiendo dos capas de aguas someras de densidades constantes e inmiscibles que pueden ser modeladas usando un sistema 2D de ecuaciones de tipo Shallow Water acoplado con términos fuente y productos no conservativos que, a su vez, se pueden formular bajo la forma de dos sistemas de leyes de conservación acoplados.

2 Descripción del modelo.

2.1 Sistema de leyes de conservación.

Consideramos un problema general que consiste en un sistema de leyes de conservación con productos no conservativos y términos fuente que puede ser escrito bajo la forma:

$$\begin{aligned}\frac{\partial W}{\partial t}+\frac{\partial F_1}{\partial x_1}(W)+\frac{\partial F_2}{\partial x_2}(W) = B_1(W)\cdot\frac{\partial W}{\partial x_1}+B_2(W)\cdot\frac{\partial W}{\partial x_2}\\ +S_1(W)\frac{\partial H}{\partial x_1}+S_2(W)\frac{\partial H}{\partial x_2},\end{aligned}\tag{1}$$

donde $W(\boldsymbol{x},t)\colon D\times(0,T)\mapsto\Omega\subset\mathbb{R}^N$, siendo D un dominio acotado de $\mathbb{R}^2$; $\boldsymbol{x}=(x_1,x_2)$ es un punto arbitrario de D; Ω es un abierto convexo de $\mathbb{R}^N$. Finalmente $F_i\colon\Omega\mapsto\mathbb{R}^N$, $B_i\colon\Omega\mapsto\mathcal{M}_N$, $S_i\colon\Omega\mapsto\mathbb{R}^N$, $i=1,2$, son funciones regulares, y $H\colon D\mapsto\mathbb{R}$ es una función conocida. Obsérvese que si $B_1=B_2=0=S_1=S_2=0$, (1) es un sistema de leyes de conservación, y, en el caso de que $B_1=B_2=0$, (1) es un sistema de leyes de conservación con términos fuente.

$J_i(W) = \dfrac{\partial F_i}{\partial W}(W)$, $i = 1, 2$ denota a los Jacobianos de los flujos F_i, $i = 1, 2$. Dado un vector unitario $\boldsymbol{\eta} = (\eta_1, \eta_2) \in \mathbb{R}^2$, definimos la matriz

$$\mathcal{A}(W, \boldsymbol{\eta}) = J_1(W)\eta_1 + J_2(W)\eta_2 - (B_1(W)\eta_1 + B_2(W)\eta_2).$$

Suponemos que el problema (1) es estrictamente hiperbólico, esto es, para cada W in Ω y cada vector unitario $\boldsymbol{\eta} \in \mathbb{R}^2$, $A(W, \boldsymbol{\eta})$ tiene N autovalores reales y distintos y por tanto es diagonalizable:

$$\mathcal{A}(W, \boldsymbol{\eta}) = \mathcal{K}(W, \boldsymbol{\eta})\mathcal{D}(W, \boldsymbol{\eta})\mathcal{K}^{-1}(W, \boldsymbol{\eta}), \tag{2}$$

donde $\mathcal{D}(W, \boldsymbol{\eta})$ es la matriz diagonal cuyos coeficientes son los autovalores de $\mathcal{A}(W, \boldsymbol{\eta})$ y $\mathcal{K}(W, \boldsymbol{\eta})$ es una matriz cuyas columnas se corresponden con los autovectores asociados.

En general, cuando W presenta discontinuidades, los productos no conservativos $B_1(W)\partial_{x_1}W$, $B_2(W)\partial_{x_2}W$, carecen de sentido en el contexto de la teoría de distribuciones. Como consecuencia, no es fácil dar una definición rigurosa al concepto de solución débil (véanse [9], [10]).

2.2 Sistema de ecuaciones de dos capas de aguas someras 2D

El sistema de ecuaciones que gobierna el flujo de dos capas de aguas someras e inmiscibles de densidades constantes en un subdominio $D \subset \mathbb{R}^2$, se puede escribir bajo la forma (1) tomando:

$$W = \left[\ h_1, \quad q_{1,1}, \quad q_{1,2}, \quad h_2, \quad q_{2,1}, \quad q_{2,2}\ \right]^T, \tag{3}$$

$$F_1(W) = \begin{bmatrix} q_{1,1} \\ \dfrac{q_{1,1}^2}{h_1} + \dfrac{1}{2}gh_1^2 \\ \dfrac{q_{1,1}q_{1,2}}{h_1} \\ q_{2,1} \\ \dfrac{q_{2,1}^2}{h_2} + \dfrac{1}{2}gh_2^2 \\ \dfrac{q_{2,1}q_{2,2}}{h_2} \end{bmatrix}, \quad F_2(W) = \begin{bmatrix} q_{1,2} \\ \dfrac{q_{1,1}q_{1,2}}{h_1} \\ \dfrac{q_{1,2}^2}{h_1} + \dfrac{1}{2}g\ h_1^2 \\ q_{2,2} \\ \dfrac{q_{2,1}q_{2,2}}{h_2} \\ \dfrac{q_{2,2}^2}{h_2} + \dfrac{1}{2}gh_2^2 \end{bmatrix}, \tag{4}$$

Los términos de acoplamiento vienen dados por las siguientes matrices:

$$B_1(W) = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & -gh_1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ -rgh_2 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}, B_2(W) = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & -gh_1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ -rgh_2 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}. \tag{5}$$

Los términos fuente contienen los efectos debidos a la batimetría:

$$S_1(\boldsymbol{x}, W) = \left[\begin{array}{cccccc} 0, & gh_1, & 0, & 0, & gh_2, & 0 \end{array} \right]^T, \tag{6}$$

$$S_2(\boldsymbol{x}, W) = \left[\begin{array}{cccccc} 0, & 0, & gh_1, & 0, & 0, & gh_2 \end{array} \right]^T. \tag{7}$$

El índice 1 hace referencia a la capa superior y el 2 a la capa inferior. g es la gravedad y $H(\boldsymbol{x})$, la función profundidad medida desde un nivel de referencia fijado. $r = \dfrac{\rho_1}{\rho_2}$ es la razón de densidades entre las capas ($\rho_1 < \rho_2$) que, en aplicaciones oceanográficas realistas es próximo a 1, de hecho en nuestro caso es 0,99801. Finalmente, $h_i(\boldsymbol{x}, t)$ y $\boldsymbol{q}_i(\boldsymbol{x}, t)$ son, respectivamente, el espesor y el flujo de masa de la capa i-ésima en el punto $\boldsymbol{x}$ y en el instante t, y están relacionados con la velocidad $\boldsymbol{u}_i(\boldsymbol{x}, t) = (u_{i,1}(\boldsymbol{x}, t), u_{i,2}(\boldsymbol{x}, t))$, $i = 1, 2$ mediante las igualdades siguientes:

$$\boldsymbol{q}_i(\boldsymbol{x}, t) = \boldsymbol{u}_i(\boldsymbol{x}, t) h_i(\boldsymbol{x}, t), \quad i = 1, 2.$$

Por simplicidad no se presentan los términos de fricción entre capas y con el fondo.

2.3 Dificultades teóricas y numéricas

Este problema presenta numerosas dificultades tanto desde el punto de vista teórico como numérico. Ya se ha mencionado anteriormente la dificultad de dar sentido matemático a los productos no conservativos. Además aparecen dificultades relacionadas con el tratamiento de los términos fuente (véase [6]), el uso de métodos numéricos adecuados para capturar choques, el tratamiento de las situaciones seco-mojado producidas por el avance de un frente sobre una zona seca o por el afloramiento de una capa de agua en un sistema bicapa (véanse [8], [2], [5]), la aparición de autovalores complejos en el modelo bicapa debido a las inestabilidades de Kelvin-Helhmoltz (véase [4]), el tratamiento de la fuerza de Coriolis, etc.

2.4 El esquema numérico

Para la discretización de las ecuaciones se ha empleado un esquema explícito de tipo volúmenes finitos. Este tipo de esquemas es adecuado para capturar soluciones en las que aparecen discontinuidades de tipo choque, no necesita la adición de términos difusivos para garantizar la estabilidad del esquema numérico, se adapta bien a zonas donde la topografía es abrupta e irregular, presenta un bajo coste computacional y es fácil de paralelizar.

Inicialmente, el dominio computacional se descompone en celdas de discretización o volúmenes finitos, $V_i \subset \mathbb{R}^2$. Denotamos por $\mathcal{T}$ el conjunto de dichas celdas. Usaremos la siguiente notación: dado V_i un volumen finito, $N_i \in \mathbb{R}^2$ es el centro de V_i, $\mathcal{N}_i$ es el conjunto de índices j tales que V_j es un

vecino de V_i. Γ_{ij} es la arista común de las celdas V_i y V_j, y $|\Gamma_{ij}|$ representa su longitud. $\boldsymbol{\eta}_{ij} = (\eta_{ij,1}, \eta_{ij,2})$ es el vector unitario normal a la arista Γ_{ij} que apunta hacia la celda V_j (ver Figura 2).

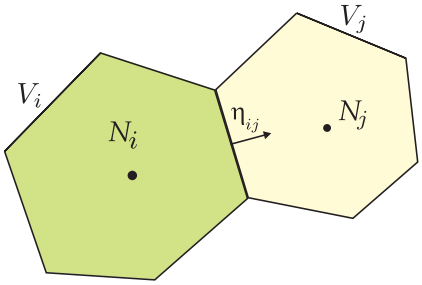


Figura 2: Volúmenes Finitos

Las aproximaciones de los promedios por celda de la solución exacta proporcionadas por el esquema numérico se denotarán como:

$$W_i^n \cong \frac{1}{|V_i|} \int W(x_1, x_2, t^n) dx_1 dx_2 \tag{8}$$

donde $|V_i|$ es el área de la celda y $t^n = n\triangle t$, siendo $\triangle t$ el paso de tiempo que podemos suponer constante sin pérdida de generalidad.

El esquema numérico resultante es:

$$W_i^{n+1} = W_i^n - \frac{\Delta t}{|V_i|} \sum_{j \in \mathcal{N}_i} |\Gamma_{ij}| F_{ij}^-, \tag{9}$$

donde

$$F_{ij}^- = \mathcal{P}_{ij}^-(\mathcal{A}_{ij}(W_j^n - W_i^n) - \mathcal{S}_{ij}(H_j - H_i)). \tag{10}$$

con $\mathcal{A}_{ij} = \mathcal{A}(W_{ij}^n, \boldsymbol{\eta}_{ij})$; donde W_{ij}^n es el "estado intermedio"de Roe correspondiente a W_i^n y W_j^n (véase [11]) y

$$P_{ij}^- = \frac{1}{2} \mathcal{K}_{ij} \cdot (I - \mathrm{sgn}(\mathcal{D}_{ij})) \cdot \mathcal{K}_{ij}^{-1}, \tag{11}$$

$$\mathcal{S}_{ij} = \eta_{ij,1} S_1(W_{ij}^n) + \eta_{ij,2} S_2(W_{ij}^n), \tag{12}$$

siendo I la matriz identidad, $\mathcal{D}_{ij}$ la matriz cuyos coeficientes son los autovalores de A_{ij}, y $\mathcal{K}_{ij}$ una matriz cuyas columnas son sus autovectores asociados. Finalmente $\mathrm{sgn}(\mathcal{D}_{ij})$ es la matriz diagonal cuyos coeficientes son el signo de los autovalores de la matriz $\mathcal{A}_{ij}$.

2.5 Paralelización y vectorización

Puesto que el esquema resultante es un esquema explícito es posible llevar a cabo la paralelización del mismo de una manera natural: partir el dominio

computacional en varios subdominios y enviar cada uno de ellos a un nodo de un cluster de ordenadores. En cada paso de tiempo los nodos correspondientes a los subdominios que posean aristas en común han de intercambiar información de manera eficiente. Siguiendo esta idea es posible "dividir" el tiempo de cálculo entre él número de nodos disponible. En [4] se pueden encontrar los detalles de esta técnica de paralelización.

Además de esta técnica, con el objetivo de conseguir un mejor rendimiento con un cluster estándar se ha estudiado cómo usar mejor el potencial de cálculo de cada nodo. Es conocido que los microprocesadores actuales disponen de unos registros de cálculo especiales, conectados a la memoria caché por buses de mayor tamaño, que permiten la vectorización de las operaciones en coma flotante. Es el caso de los registros SSE que incorporan los micros de Intel o registros similares que incorporan micros de otras marcas. Este tipo de registros e instrucciones dotan a los micros de una arquitectura paralela de tipo SIMD (véase [4]), ya que permiten realizar cálculos simultáneos de manera vectorial. Los resultados han sido espectaculares, ya que por ejemplo, para una simulación que en la versión paralela del esquema aquí presentado tarda 6m26.135s en 8 procesadores se consigue realizarla en 29.315s, es decir una reducción de tiempo cálculo del orden de 13 veces (véanse [4] y [3] para más detalles).

3 Experimentos numéricos.

En esta sección se muestran dos experimentos numéricos. Inicialmente llevamos a cabo un experimento de tipo lock-exchange con el objetivo de simular el intercambio secular a través del Estrecho de Gibraltar. El objetivo de este experimento es el de validar la capacidad del esquema implementado para simular flujos geofísicos a través de canales con geometrías complejas y con un coste computacional moderado. El segundo experimento consiste en la simulación de un experimento de mareas en el Estrecho de Gibraltar. Se emplea como condición inicial la solución de intercambio secular obtenida en el experimento lock-exchange y se imponen en las fronteras abiertas las cuatro componentes principales de la marea en la zona. Los resultados obtenidos han sido comparados con observaciones realizadas en campañas oceanográficas por García-Lafuente (1986) y Candela et al. (1990) y están siendo validados con medidas experimentales obtenidas por el Grupo de Oceanografía Física dirigido por Miguel Bruno Mejías de la Universidad de Cádiz en el marco del subproyecto CTM2005-08142-C03-02/MAR.

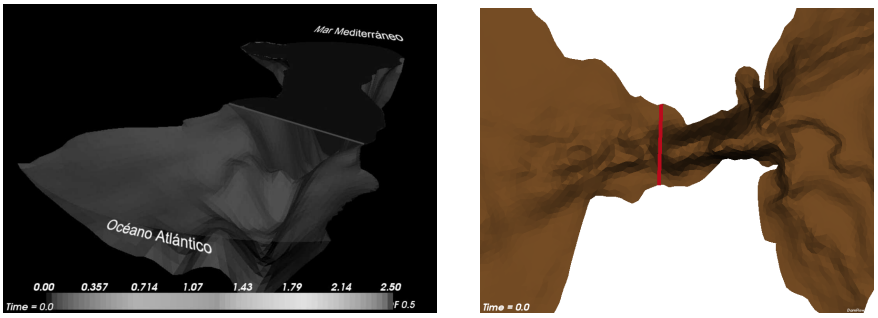
3.1 Experimento tipo lock-exchange

Partimos de una malla computacional de 32325 celdas y sobre las que está definida una función constante a trozos que aproxima la batimetría del Estrecho y que ha sido generada a partir de datos batimétricos reales (ver Figura 3(b)).

Para llevar a cabo un experimento de tipo lock-exchange se sitúa inicialmente una barrera artificial en la sección transversal de menor área del Estrecho de

modo que separa las aguas atlánticas y mediterráneas: las Figuras 3(a) y 3(b) muestran una vista 3D y una sección longitudinal de esta condición inicial. La razón de densidades $r = 0{,}99805$. En la frontera Γ_1 correspondiente a la línea de costa se impone la condición $\boldsymbol{q}_i \cdot \boldsymbol{\eta} = 0$, mientras que en las fronteras abiertas Γ_2 y Γ_3, se imponen las condiciones globales

$$\int_{\Gamma_i} (\boldsymbol{q}_1(\gamma) + \boldsymbol{q}_2(\gamma)) \cdot \boldsymbol{\eta}(\gamma) = 0, \quad i = 2, 3.$$



(a) Vista 3D del estado inicial (b) Sección Longitudinal

Figura 3: Experimento Lock-Exchange. Estado Inicial.

La simulación se lleva a cabo con un $CFL = 0{,}9$ hasta que se alcanza un estado estacionario. Las Figuras 4(a) y 4(b) muestran una sección longitudinal de la interfaz y la superficie libre así como una vista 3D correspondiente al estado estacionario. El flujo total en cada sección transversal Γ_T es aproximadamente

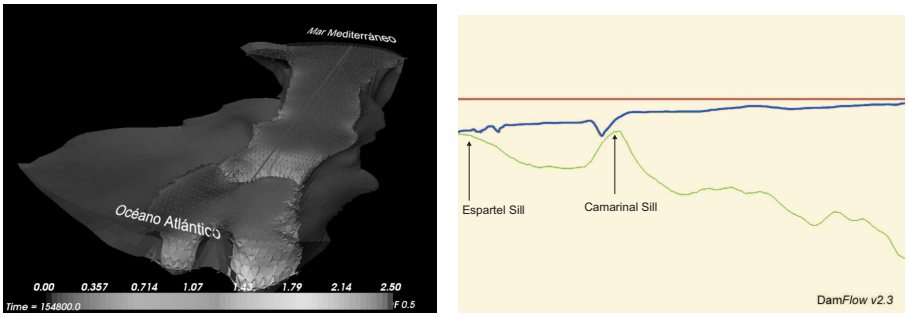
$$\int_{\Gamma_T} \boldsymbol{q}_1(\gamma) \cdot \boldsymbol{\eta}(\gamma)\, d\gamma \approx -\int_{\Gamma_T} \boldsymbol{q}_2(\gamma) \cdot \boldsymbol{\eta}(\gamma)\, d\gamma \approx 0{,}75 \text{ Sv}, \ (1Sv = 10^6\ m^3/s),$$

valor que está en buen acuerdo con las medidas experimentales (ver, por ejemplo, [1]).

3.2 Experimento de mareas

El experimento de mareas en el Estrecho de Gibraltar tiene como principal objetivo el estudio de los elementos esenciales de la respuesta de este modelo a un forzamiento de mareas. Para ello se ha diseñado el siguiente experimento: usando como condición inicial la solución estacionaria del experimento lock-exchange previo, el modelo se fuerza imponiendo en las fronteras abiertas Γ_2 y Γ_3 las cuatro principales ondas de marea de la zona, esto es, la M2 y S2 (semidiurnas) y la O1 y K1 (diurnas):

$$h_1(\boldsymbol{x}_B, t) + h_2(\boldsymbol{x}_B, t) = \bar{h}_B - \sum_{n=1}^{4} Z_n(\boldsymbol{x}_B) cos(\alpha_n t - \phi_n(\boldsymbol{x}_B)),$$



(a) Vista 3D del estado estacionario (b) Sección Longitudinal

Figura 4: Experimento Lock-Exchange. Estado Estacionario.

donde $\boldsymbol{x}_B$ representa un punto de las fronteras abiertas (Γ_2 o Γ_3); $Z_n(\boldsymbol{x}_B)$ y $\phi_n(\boldsymbol{x}_B)$ son la amplitud de la elevación y la fase de la n-sima componente de marea en las fronteras abiertas; α_n su frecuencia, $\bar{h}_B$ la elevación total de la columna de agua correspondiente al estado estacionario en esta frontera. Los datos de estas componentes de marea han sido extraídos del modelo de mareas oceánicas FES2004.

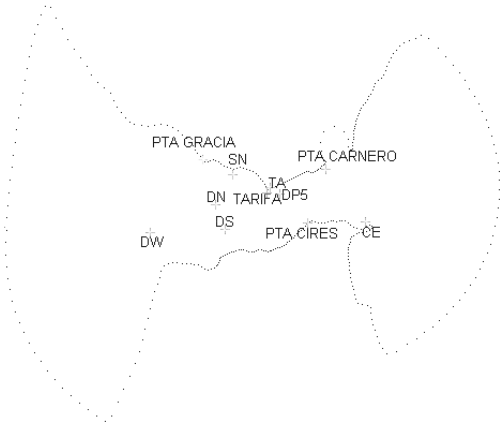


Figura 5: Puntos relevantes

El modelo es integrado hasta alcanzar un estado quasi-periódico en tiempo (aproximadamente 15 días). Seguidamente se integra durante un periodo de 29 días. A pesar de las simplificaciones del modelo, la dinámica que se obtiene coincide en gran medida con las observaciones: en cada ciclo de mareas se produce un choque al oeste del umbral de Camarinal en la interfaz que crece hasta tener una amplitud de más de 100m. Con la subida de la marea este choque llega a desaparecer generando ondas que se desplazan hacia el interior

del Mediterráneo (ver Figuras 6(a), 6(b)).

En la Figura 5 se muestran una serie de puntos significativos importantes en el Estrecho para los que se conocen (véase [7]) series de medidas experimentales de distintas variables (elevaciones, temperatura, salinidad, etc.). El proceso de validación del modelo consiste en comparar los resultados de un análisis armónico realizado sobre las elevaciones debidas a las mareas en dichos puntos con los datos experimentales.

Cuadro 1: Localización: SN (36°03' N 5°43' W)

	M2		S2		O1		K1	
	Obs	Pred	Obs	Pred	Obs	Pred	Obs	Pred
A (cm)	52.3	57.89	18.5	19.76	0.7	0.80	2.1	2.83
Fase	47.6	51.28	73.4	86.17	298	328.41	95.3	97.19

Los resultados son, en general muy satisfactorios, obteniéndose un margen de error pequeño tanto en las amplitudes de las elevaciones como en las fases. En las tablas 1 y 2 se muestran los resultados de dicho análisis para dos puntos significativos.

Cuadro 2: Localización: Tarifa (36°0.2' N 5°43' W)

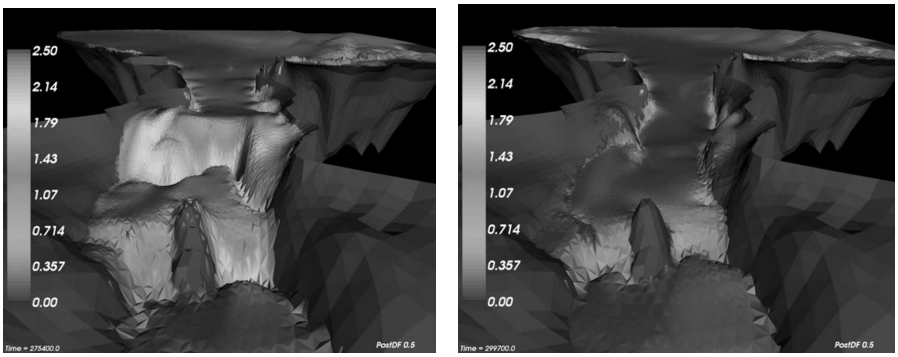
	M2		S2		O1		K1	
	Obs	Pred	Obs	Pred	Obs	Pred	Obs	Pred
A (cm)	41.7	42.31	14.2	14.23	0.7	0.34	2.2	2.79
Fase	57	47.07	85	78.75	165	169.62	131	140.80

Referencias

[1] H. Bryden, J. Candela, and T.H. Kinder. *Exchange through the Strait of Gibraltar*. Prog. Oceanogr, 33:201–248, 1994.

[2] M.J. Castro, A.M. Ferreiro, J.A. García, J.M. González, J. Macías, C. Parés and M.E. Vázquez. *On the numerical treatment of wet/dry fronts in shallow flows: applications to one-layer and two-layer systems*. Math. Comp. Model. 42 (3-4): 419–439, 2005.

[3] M.J. Castro, J.A. García-Rodríguez, J.M. González-Vida and C. Parés. *Solving shallow water systems in 2d domains using finite volume methods and multimedia SSE instructions.* Accepted in J. Comput. App. Math.



(a) Marea baja (b) Marea alta.

Figura 6: Exp. de mareas. Vista de la interfaz (velocidades en escala de colores).

[4] M.J. Castro, J.A. García, J.M. González and C. Parés. *A parallel 2d finite volume scheme for solving systems of balance laws with nonconservative products: application to shallow flows.* Comp. Meth. Appl. Mech. Eng. 196, 2788–2815, 2006.

[5] M.J. Castro, J.M. González and C. Parés. *Numerical treatment of wet/dry fronts in shallow flows with a modified Roe scheme.* Math. Mod. Meth. App. Sci. Vol. 16, No. 6, 897–931, 2006.

[6] M.J. Castro, J. Macías, C. Parés, J.A. García and M.E. Vázquez. *A two-layer finite volume model for flows through channels with irregular geometry: computation of maximal exchange solutions. Application to the Strait of Gibraltar.* Comm. Nonlinear Sci. Num. Simul. 9: 241–249, 2004.

[7] J. García Lafuente, J.L. Almazán, F. Castillejo, A. Khribeche and A. Hakimi. *Sea level in the Strait of Gibraltar: tides.* International Hydrographic Review, LXVII(1):111–130, 1990.

[8] J.M. González Vida. *Desarrollo de esquemas numéricos para el tratamiento de frentes seco-mojado en sistemas de aguas someras.* Phd. Thesis, Universidad de Málaga, 2003.

[9] G. Dal Maso, P.G. LeFloch and F. Murat. *Definition and weak stability of nonconservative products.* J. Math. Pures Appl. 74:483–548, 1995.

[10] C. Parés. *Numerical methods for nonconservative hyperbolic systems: a theoretical framework*, SIAM Journal of Numerical Analysis **44**(1), 300–321, 2006.

[11] C. Parés, M.J. Castro. *On the well-balance property of Roe's method for nonconservative hyperbolic systems. Applications to shallow-water systems.* ESAIM: M2AN, 38(5):821–852, 2004.

Bol. Soc. Esp. Mat. Apl.
nº 42(2008), 147–162

A SPACE-TIME ADAPTIVE SEMI DUAL WEIGHTED RESIDUAL FINITE ELEMENT METHOD

R. BERMEJO AND J. CARPIO

Dpto. Matemática Aplicada a la Ingeniería Industrial
Universidad Politécnica de Madrid.

rbermejo@etsii.upm.es, jaime.carpio@upm.es

Abstract

We propose in this paper a space-time adaptive algorithm based on the Dual Weighted Residual (DWR) idea in the framework of finite element method. Our algorithm consists of applying the DWR technique locally in each time interval $I_n := (t_{n-1}, t_n]$, thus, we control the local or truncation error for a functional of the solution J(u). That means that we can define a self-sufficient criterium that allows us to have control of the time step Δt and the mesh size h as time progresses. Another good feature of our algorithm is the extension of the spatial post-processing procedure of the traditional DWR method to unstructured meshed made of simplices.

Key words: *Finite elements, semi-Lagrangian method, a posteriori error estimator, Dual Weighted Residual method, unstructured triangular meshes.*
AMS subject classifications: *65N30 65M60*

1 Introduction

The efficient numerical treatment of multi-scales phenomena and problems with poor regularity of the solution must be carried out with an adaptive method based on a posteriori error estimator to make realistic computation feasible, specially in 3D. To treat time-dependent problems, we split the time interval $I := (0, T]$ into half-open subintervals $I_n =: (t_{n-1}, t_n]$ of length $\Delta t_n := t_n - t_{n-1}$, such that, $0 = t_0 < ... < t_n < ... < t_N = T$. In each time subinterval I_n, we generate a conforming triangulation $\mathbb{T}_h^n$ of the domain Ω and calculate with a time step size Δt_n the numerical solution $u_{h\Delta t}^n$. Then an adaptive finite element method in time could consist of successive loops as shown in Figure 1.

For each time level t_n we must make a discretization of the problem and solve an algebraic system of equations to get the numerical solution $u_{h\Delta t}^n$. Using the

Trabajo subvencionado por los proyectos REN2002-03276 y CGL2006-11264-C04-02/CLI del Ministerio de Educación y Ciencia de España

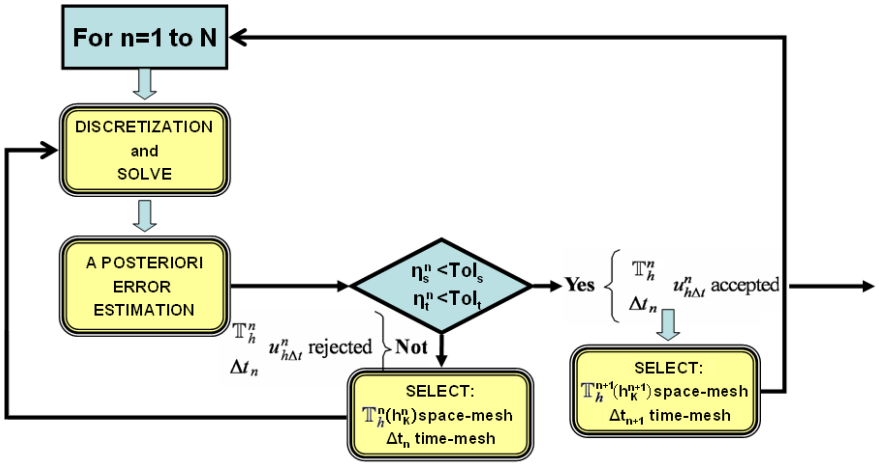


Figure 1: Scheme of the basic time-space adaptive algorithm for time-dependent PDE's.

numerical solution we perform a posteriori error analysis to estimate, both in time and in space, the error of the numerical solution. If the estimated errors are below given tolerances, then the solution $u_{h\Delta t}^n$, the mesh $\mathbb{T}_h^n$ and the time step size Δt_n are accepted; if not, all of them are rejected and the procedure is repeated after properly adjusting the size of the time step and changing the mesh.

The main purpose of this paper is to explain the 'a posteriori error estimation' procedure of the algorithm. Our method, formulated in the context of finite elements, is based on the philosophy of Dual Weighted Residual (hereafter, DWR) methodology to develop a posteriori error estimates to asses the accuracy of the numerical solution. For stationary problems, the application of DWR methodology yields efficient adaptive finite element algorithms; but for time dependent problems, it is still an open question how to achieve an efficient adaptive algorithm with such a methodology, since it needs large computational resources such as storage and CPU time. To this respect, it is worth mentioning the recent work of Schmich and Vexler [11] where and efficient DWR adaptive finite element algorithm for parabolic problems is presented. To alleviate the shortcomings of the DWR in time-dependent problems, we apply the DWR technique locally in each time subinterval I_n to the original problem with a perturbed initial condition given by the numerical solution at time t_{n-1}. Thus, rather than controlling the error for the whole interval $[0, T]$ what we have now is a good local control of the error in each interval I_n. Good properties of the algorithm presented in this paper are the following: (1) it is self-sufficient in providing a precise criterium for adaptation of both the time step Δt and the mesh size h; (2) it extends the idea of the space post-processing of the DWR

method to unstructured meshes made of simplices. However a weak point of this approach is that having a good local error control does not guarantee that the global error will be bounded as $\|J(u) - J(u_{h\Delta t})\| < GTOL$, $GTOL$ being a prescribed global tolerance, for it is well known a fact that the magnitude of the global error will depend on the stiffness of the problem. If the stiffness is low or moderate we will end up having a small global error if we control the local error well, but if the stiffness is large the global error will be large even if we control the local error with a reasonable tolerance.

2 The model problem: convection-diffusion-reaction equations

To make simple the presentation of the adaptive algorithm for convection-reaction-diffusion equations we shall consider in a bounded domain $\Omega \subset \mathbb{R}^2$ with sufficiently smooth boundary $\partial\Omega$ the model problem:

$$\begin{cases} \dfrac{Du}{Dt} = \epsilon\Delta u + f(u,x,t) \quad \text{in } \Omega \times (0,T], \\ u(x,0) = 0 \ \text{ in } \ \Omega, \\ u|_{\partial\Omega} = 0 \ \ t > 0. \end{cases} \tag{1}$$

Here, the diffusion parameter $\epsilon > 0$ is supposed to be constant, the reaction term $f : C^l(\mathbb{R}) \times \Omega \times (0,T] \to \mathbb{R}$, $l \geq 1$ integer, satisfies suitable growth conditions and the total derivative operator $D/Dt := \partial/\partial t + \mathbf{a}(x,t) \cdot \nabla$, where the velocity vector $\mathbf{a}(x,t)$ is such that

$$\mathbf{a}(x,t) \in L^\infty(0,T; W_0^{1,\infty}(\mathbb{R}^2)^2) \ \text{ and } \ \forall t > 0 \ \mathbf{a}(x,t) \cdot \mathbf{n}\,|_{\partial\Omega} = g(x,t) \in L^2(\partial\Omega). \tag{2}$$

$\mathbf{n}$ being the outward unit normal vector on $\partial\Omega$. Under these conditions there exists a unique weak solution to problem (1) $u \in L^2(0,T; H_0^1(\Omega)) \cap C((0,T; L^2(\Omega)) \cap L^p(\Omega \times (0,T))$, $p \geq 2$ integer, such that for all $v \in L^p(0,T; H_0^1(\Omega))$, $\dfrac{\partial v}{\partial t} \in L^q(0,T; H^{-1}(\Omega))$ and $v(0,T) = 0$, it holds

$$-\int_0^T \int_\Omega \ \frac{\partial v}{\partial t} + div(\mathbf{a}v) \ \ u d\Omega dt = \int_0^T \int_\Omega \epsilon \nabla u \cdot \nabla v d\Omega dt + \int_0^T \int_\Omega f(u,x,t) v d\Omega dt. \tag{3}$$

To calculate an approximate weak solution we shall break the interval $\overline{I} := [0,T]$ into subintervals $I_n := (t_{n-1}, t_n]$, $n = 1, 2, ..., N$, with $t_0 = 0$, $t_N = T$ and $\overline{I} = \cup_{n=1}^N \overline{I}_n$. We shall consider for each n the slab $S_n := \Omega \times I_n$, and for fixed integers $r \geq 1$ and s the trial and test spaces respectively:

$$\begin{aligned} V_{h\Delta t}^{(r)} = \{\varphi_{h\Delta t} : \overline{Q}_T \to \mathbb{R} : \ \forall n \text{ and } (x,t) \in S_n, \ \varphi_{h\Delta t} \in C(\overline{I}_n; V_h^n), \\ \varphi_{h\Delta t}(x,0) \in V_h^0 \ \text{ and } \ \varphi_{h\Delta t}(x,\cdot)|_{I_n} \in P_r\}, \end{aligned} \tag{4}$$

$$\begin{aligned} W_{h\Delta t}^{(s)} = \{\psi_{h\Delta t} : \overline{Q}_T \to \mathbb{R} : \ \forall n \text{ and } (x,t) \in S_n, \ \psi_{h\Delta t} \in L^p(I_n; V_h^n), \\ \psi_{h\Delta t}(x,0) \in V_h^0 \ \text{ and } \ \psi_{h\Delta t}(x,\cdot)|_{I_n} \in P_s\}; \end{aligned} \tag{5}$$

here, $\overline{Q}_T := \overline{\Omega} \times \overline{I}$, P_r and P_s are the set of polynomials of degrees at most r and s respectively defined on I_n, and

$$V_h^n = \{v_h : C^0(\overline{\Omega}) : v_h|_K \in P_m(K)\ \forall K \in \mathbb{T}_h^n\}$$

where $P_m(K)$ is the set of polynomials of degree at most m defined on $K \in \mathbb{T}_h^n$. Note that for $r = 0$ the space $V_{h\Delta t}^0$ coincides with $W_{h\Delta t}^0$.

The application of the semi-Lagrangian approach to calculate at any instant t_n an approximation to the weak solution requires the integration of the system

$$\begin{cases} \dfrac{dX(x,t_n;t)}{dt} = \mathbf{a}(X(x,t_n;t),t), \\ \\ X(x,t_n;t_n) = x, \end{cases} \tag{6}$$

for $x \in \Omega$ and $t \in \overline{I}_n$. $X(x,t_n;t)$ denotes the characteristics of the operator $\dfrac{D}{Dt}$ in the time subinterval $I_n := (t_{n-1}, t_n]$, in particular, $X(x,t_n;t_{n-1})$ are the feet of the characteristics at time t_n. By virtue of the assumptions on $\mathbf{a}(x,t)$, see (2), the unique solution of (6) is given as

$$X(x,t_n;t) = x - \int_t^{t_n} \mathbf{a}(X(x,t_n;\tau),\tau)d\tau. \tag{7}$$

Considering a fixed open bounded domain $\Omega^* \supset \overline{\Omega}$ and assuming the existence of an extension operator $E : H^1(\Omega) \to H^1(\Omega^*)$, such that $C^1(\overline{\Omega})$ functions are mapped into $C^1(\overline{\Omega}^*)$ functions and $E\,u|_\Omega = u$; we have that for all $\overline{I}_n$, given the function $u : \Omega \times \overline{I}_n \to \mathbb{R}$, we define $u^* : \Omega^* \times \overline{I}_n \to \mathbb{R}$ as $u^* = Eu$ and $\overline{u} : \Omega \times \overline{I}_n \to \mathbb{R}$ as $\overline{u}(x,t) = u^*(\cdot,t) \circ X(\cdot,t_n;t)(x)$. Then, (1) can be recast as

$$\begin{cases} \dfrac{\partial \overline{u}}{\partial t} = \epsilon \Delta \overline{u} + f(\overline{u}, X(x,t_n;t),t) \quad \text{in } S_n, \\ \\ \overline{u}(x,t_{n-1}) \ \text{ in } \Omega \text{ is a datum}, \\ \\ \overline{u}|_{\partial\Omega} = 0, \end{cases} \tag{8a}$$

noting that

$$\overline{u}(x,t)|_{t=t_n} = u(x,t_n), \quad x \in \Omega \tag{8b}$$

The adaptive algorithm we describe in this chapter computes the numerical solution to (1) in the finite dimensional spaces $V_{h\Delta t}^{(1)} \times W_{h\Delta t}^{(0)}$. A numerical solution to problem (8a) is then a function

$$\overline{u}_{h\Delta t}(x,t) = \overline{u}_{h\Delta t}^{n-1}(x) + \frac{t - t_{n-1}}{t_n - t_{n-1}}(u_{h\Delta t}^n - \overline{u}_{h\Delta t}^{n-1}(x)), \tag{9}$$

with

$$u_{h\Delta t}^n|_{\partial\Omega} = 0, \tag{10}$$

and such that for all $\psi_{h\Delta t} \in W^{(0)}_{h\Delta t}$

$$(u^n_{h\Delta t}-\overline{u}^{n-1}_{h\Delta t},\psi_{h\Delta t})_\Omega+\frac{\Delta t_n}{2}a(u^n_{h\Delta t}+\overline{u}^{n-1}_{h\Delta t},\psi_{h\Delta t})+\int_{I_n} f(u_{h\Delta t},X^{n-1}_{h\Delta t}(x),t),\psi_{h\Delta t})_\Omega\, dt. \tag{11}$$

Here, the following notations are used: for all n, $g(x,t_n)=g^n(x)$, $X^{n-1}_{\Delta t}(x)$ is an approximation to $X(x,t_n;t_{n+1})$,

$$a(u,v)=\epsilon\int_D \nabla u\cdot\nabla v d\Omega, \qquad u,v\in H^1(\Omega),$$

and

$$(u,v)_\Omega=\int_\Omega u\cdot v d\Omega, \quad u,v\in L^2(\Omega).$$

Thus, to calculate the solution $u^n_{h\Delta t}(x)$ for each I_n we perform the following two stages:

(1) *The semi-Lagrangian stage.* In this stage we calculate for each I_n the set of departure points $\{X^{n-1}(x_i)\}$, $x_i\in\Omega$, and then obtain $\overline{u}^{n-1}_{h\Delta t}(x)$ defined on the partition $\mathbb{T}^n_h$ via quasi-monotone quadratic interpolatory projection from the solution $u^{n-1}_{h\Delta t}(x)$ defined on the partition $\mathbb{T}^{n-1}_h$.

(2) *The parabolic stage.* Here, we calculate the solution $u^n_{h\Delta t}$ by solving (11).

3 A posteriori error estimator

As the numerical algorithm to obtain the approximate solution $u^n_{h\Delta t}$ is divided into two stages, we shall devise an a posteriori error estimator for both parts.

3.1 Semi-Lagrangian stage

Here, we propose a numerical procedure to calculate $\overline{u}_{h\Delta t}(x,t_{n-1})$ as a discrete approximation of $u_{h\Delta t}(X(x,t_n;t_{n-1}),t_{n-1})$. Hereafter we shall denote by $X^{n-1}(x)$ the departure point $X(x,t_n;t_{n-1})$ of the trajectories.

1. We approximate $X^{n-1}(x)$ by solving (6) with the embedded Runge-Kutta 2(3) algorithm:

$$\begin{cases} K_1=\mathbf{a}(x,t_n),\\ K_2=\mathbf{a}(x-\Delta t_n K_1,t_n-\Delta t_n),\\ K_3=\mathbf{a}\left(x-\dfrac{\Delta t_n K_1}{4}-\dfrac{\Delta t_n K_2}{4},t_n-\dfrac{\Delta t_n}{2}\right),\\ \\ X_{\Delta t}(x,t_n;t_{n-1})=x-\Delta t_n\left(\frac{K_1}{2}+\frac{K_2}{2}\right).\\ \\ X^*_{\Delta t}(x,t_n;t_{n-1})=x-\Delta t_n\left(\frac{K_1}{6}+\frac{K_2}{6}+\frac{4K_3}{6}\right). \end{cases} \tag{12a}$$

noting that $X^*_{\Delta t}(x)$ and $X_{\Delta t}(x)$ are second and third order approximations to $X^{n-1}(x)$, respectively.

2. Let $\{\varphi_i^n(x)\}$ and $\{\varphi_j^{n-1}(x)\}$ be the sets of global basis functions of V_h^n and V_h^{n-1}, respectively. Since $\overline{u}_{h\Delta t}^{n-1}(x) \in V_h^n$, then we set

$$\overline{u}_{h\Delta t}^{n-1}(x) = \sum_{i=1}^{N_h^n} \overline{U}_i^{n-1} \varphi_i^n(x), \tag{12b}$$

where N_h^n is the dimension of the set of mesh nodes $\{x_i\}$ of $\mathbb{T}_h^n$ and $\overline{U}_i^{n-1} = \overline{u}_{h\Delta t}^{n-1}(x_i) = u_{h\Delta t}^{n-1}(X_{\Delta t}^{n-1}(x_i))$. Moreover, given that $u_{h\Delta t}^{n-1}(x) \in V_h^{n-1}$, we calculate $u_{h\Delta t}^{n-1}(X_{\Delta t}^{n-1}(x_i))$ as

$$u_{h\Delta t}^{n-1}(X_{\Delta t}^{n-1}(x_i)) = \sum_{j=1}^{N_h^{n-1}} U_j^{n-1} \varphi_j^{n-1}(X_{\Delta t}^{n-1}(x_i)). \tag{12c}$$

where N_h^{n-1} is the dimension of the set of mesh nodes $\{x_j\}$ of $\mathbb{T}_h^{n-1}$. In general, $N_h^{n-1} \neq N_h^n$ and $X_{\Delta t}^{n-1}(x_i) \notin \{x_j\}$; in fact, there is an element $K \in \mathbb{T}_h^{n-1}$ where $X_{\Delta t}^{n-1}(x_i)$ is contained. To find such an element we use the search-locate algorithm presented in [1].

3. It is known that Lagrange interpolation of degree ≥ 2 leads to a result that exhibits an oscillatory behavior and does not satisfy a discrete maximum principle. To overcome these problems we use the mesh independent limiting procedure (or monotony procedure), specifically designed for semi-Lagrangian schemes. Then,

$$\overline{U}_i^{n-1} = \begin{cases} U^{n-1+} & \text{if} \quad U^{n-1+} > \overline{U}_i^{n-1}, \\ U^{n-1-} & \text{if} \quad U^{n-1-} < \overline{U}_i^{n-1}, \\ \overline{U}_i^{n-1} & \text{otherwise,} \end{cases} \tag{12d}$$

where

$$U^{n-1+} = \max_l \{U_l^{n-1}\}|_K \quad \text{and} \quad U^{n-1-} = \min_l \{U_l^{n-1}\}|_K. \tag{12e}$$

Now, we will define the a posteriori error for the quasi-monotone interpolatory semi-Lagrangian scheme introduced above. To have a coherent representation of the error following the ideas of goal-oriented adaptivity, we calculate the error in the output functional $J(u_{h\Delta t}^{n-1}(X))$ as

$$e_{convect} = J(u_{h\Delta t}^{n-1}(X^{n-1}(x))) - J(\overline{u}_{h\Delta t}^{n-1}(x)).$$

We use output functionals $J(\cdot)$ of the form $J(\cdot) = \int_\Omega j(\cdot) d\Omega$, where $j(\cdot)$ is a measurable function of u. Then, we can define a post-processing of the solution to obtain a spatial and temporal error estimators. For further details, see [7].

3.2 Diffusion-reaction stage

To apply the DWR methodology to control the local error of the numerical solution in each I_n, we follows the approach of [4] and consider (8a) with a perturbed initial condition given by the semi-Lagrangian stage. Thus, for each I_n we define the semi-linear form $A : V^n \times W^n \to \mathbb{R}$ as

$$A(\overline{U})(z) = \int_{I_n} \partial_t \overline{U}, z \ _\Omega + a(\nabla \overline{U}, \nabla z)_\Omega - \ f(\overline{U}), z \ _\Omega \ dt + ([\overline{U}]^{n-1}, z^{n-1+})_\Omega, \quad (13)$$

where $[\overline{U}]^{n-1} = \overline{U}^{n-1+} - \overline{u}_{h\Delta t}^{n-1}$ is the jump of the solution at time t_{n-1}, and V^n and W^n are the local restrictions to I_n of the spaces $L^2(0,T;H_0^1(\Omega)) \cap L^p(\Omega \times (0,T)) \cap C([0,T];L^2(\Omega))$ and $L^p(0,T;H_0^1(\Omega))$, respectively.

Next, we choose an output functional $J : V^n \to \mathbb{R}$, such that we can define the Lagrangian $\mathcal{L} : V^n \times W^n \to \mathbb{R}$ as

$$\mathcal{L}(\overline{U}; z) := J(\overline{U}) - A(\overline{U})(z).$$

Then, we calculate the stationary points $(\overline{U}, z) \in V^n \times W^n$ of $\mathcal{L}(\overline{U}; z)$ which are solution of

$$\mathcal{L}'(\overline{U}; z)(\varphi, \psi) = 0;$$

that is, we have to find the pair $(\overline{U}, z) \in V^n \times W^n$ that satisfies

$$-\int_{I_n} \partial_t \overline{U}, \psi \ _\Omega + (\nabla \overline{U}, \nabla \psi)_\Omega - \ f(\overline{U}), \psi \ _\Omega \ dt - ([\overline{U}]^{n-1}, \psi^{n-1+})_\Omega = 0 \ \ \forall \psi \in W^n \quad (14a)$$

and

$$J'(\overline{U})(\varphi) - \int_{I_n} - \langle \partial_t z, \varphi \rangle_\Omega + (\nabla \varphi, \nabla z)_\Omega - \ f'(\overline{U})\varphi, z \ _\Omega \ dt - (\varphi^{n-}, z^{n-})_\Omega = 0, \ \ \forall \varphi \in V^n. \quad (14b)$$

(14a) and (14b) are termed the primal and dual problems, respectively. The Galerkin approximation, $(\overline{u}_{h\Delta t}, z_{h\Delta t}) \in V_{h\Delta t} \times W_{h\Delta t}$, to such problems in each slab S_n satisfies for all $(\varphi_{h\Delta t}, \psi_{h\Delta t}) \in V_{h\Delta t} \times W_{h\Delta t}$ the equation

$$\mathcal{L}'(\overline{u}_{h\Delta t}; z_{h\Delta t})(\varphi_{h\Delta t}, \psi_{h\Delta t}) = 0;$$

that is, $(\overline{u}_{h\Delta t}, z_{h\Delta t})$ is the unique solution of the following numerical problem:

For each I_n, find the pair $(\overline{u}_{h\Delta t}, z_{h\Delta t}) \in V_{h\Delta t} \times W_{h\Delta t}$ such that for all $(\varphi_{h\Delta t}, \psi_{h\Delta t}) \in V_{h\Delta t} \times W_{h\Delta t}$

$$\begin{cases} -\int_{I_n} \{(\partial_t \overline{u}_{h\Delta t}, \psi_{h\Delta t})_\Omega + (\nabla \overline{u}_{h\Delta t}, \nabla \psi_{h\Delta t})_\Omega - (f(\overline{u}_{h\Delta t}), \psi_{h\Delta t})_\Omega\} \, dt - \\ -([\overline{u}_{h\Delta t}]^{n-1}, \psi_{h\Delta t}^{n-1+})_\Omega = 0, \end{cases} \quad (15a)$$

and

$$\begin{cases} J'(\overline{u}_{h\Delta t})(\varphi_{h\Delta t}) - \int_{I_n} (-\partial_t z_{h\Delta t}, \varphi_{h\Delta t})_\Omega + (\nabla z_{h\Delta t}, \nabla \varphi_{h\Delta t})_\Omega\} dt + \\ + \int_{I_n} (f'(\overline{u}_{h\Delta t})\varphi_{h\Delta t}, z_{h\Delta t})_\Omega \, dt - (\varphi_{h\Delta t}^{n-}, z_{h\Delta t}^{n-})_\Omega = 0. \end{cases} \quad (15b)$$

From (15a) and (15b) together with Proposition 6.1 of [2] is easy to obtain the following result.

Proposition 1. *For each I_n, let $(\overline{U}, z)$ and $(\overline{u}_{h\Delta t}, z_{h\Delta t})$ be solutions of ((14a), (14b)) and ((15a), (15b)) respectively. Assume that the functional $J : V^n \to \mathbb{R}$ and the semi-linear form $A : V^n \times W^n \to \mathbb{R}$ have directional derivatives up to order three. Then, we have the following error representation*

$$\begin{cases} e_{diff-react} = J(\overline{U}) - J(\overline{u}_{h\Delta t}), \\ \\ J(\overline{U}) - J(\overline{u}_{h\Delta t}) = \frac{1}{2}\rho(\overline{u}_{h\Delta t})(z - z_{h\Delta t}) + \frac{1}{2}\rho^*(\overline{u}_{h\Delta t}, z_{h\Delta t})(\overline{U} - \overline{u}_{h\Delta t}) + \mathcal{R}^{(3)}_{h\Delta t}, \end{cases} \tag{16}$$

where $\rho(\overline{u}_{h\Delta t})(\cdot)$ and $\rho^*(\overline{u}_{h\Delta t}, z_{h\Delta t})(\cdot)$ are the primal and dual residuals respectively, given by the formulas:

Primal residual:

$$\begin{cases} \rho(\overline{u}_{h\Delta t})(\cdot) = \sum\limits_{K\in\mathbb{T}^n_h} \int_{I_n} \ (\overline{R}_{h\Delta t}, \cdot)_K + (\overline{r}_{h\Delta t}, \cdot)_{\partial K} \ \ dt - ([\overline{u}_{h\Delta t}]^{n-1}, (\cdot)^{n-1+})_K, \\ \\ \overline{R}_{h\Delta t} = \overline{f}(\overline{u}_{h\Delta t}) - \partial_t \overline{u}_{h\Delta t} + \epsilon\Delta\overline{u}_{h\Delta t} \ \text{ and } \ \overline{r}_{h\Delta t} = \begin{cases} \frac{\epsilon}{2}\left[\partial_n \overline{u}_{h\Delta t}\right]_\Gamma \ \text{ if } \Gamma \subset \partial K \backslash \partial\Omega, \\ 0 \ \text{ if } \ \Gamma \subset \partial\Omega. \end{cases} \end{cases}$$

Dual residual:

$$\begin{cases} \rho^*(\overline{u}_{h\Delta t}, z_{h\Delta t})(\cdot) = \ \sum\limits_{K\in\mathbb{T}^n_h} \int_{I_n} \left\{ (\overline{R}^*_{h\Delta t}, \cdot)_K + (\overline{r}^*_{h\Delta t}, \cdot)_{\partial K} \right\} dt \\ \qquad\qquad\qquad\qquad -J'(\overline{u}_{h\Delta t})(\cdot)_K - (z^n_{h\Delta t}, (\cdot)^{n-})_K, \\ \\ R^*_{h\Delta t} = \overline{f}'(\overline{u}_{h\Delta t}) z_{h\Delta t} + \partial_t z_{h\Delta t} + \epsilon\Delta z_{h\Delta t} \ \text{ and } r^*_{h\Delta t} = \begin{cases} \frac{\epsilon}{2}\left[\partial_n z_{h\Delta t}\right]_\Gamma \ \text{ if } \Gamma \subset \partial K \backslash \partial\Omega, \\ 0 \ \text{ if } \Gamma \subset \partial\Omega. \end{cases} \end{cases}$$

The reminder term $\mathcal{R}^{(3)}_{h\Delta t}$ is the order three on the errors $e = \overline{U} - \overline{u}_{h\Delta t}$ and $e^* = z - z_{h\Delta t}$, usually small and can be neglected when Proposition 1 is used for mesh adaptation.

The terms $(z - z_{h\Delta t})$ and $(\overline{U} - \overline{u}_{h\Delta t})$ are the so called weights and in practice to make adaptation they must be estimated from their corresponding numerical solutions via a post-processing procedure (here, we use patch-wise higher order interpolation recovery). We also propose a post-processing procedure to separate the error contribution in two parts, namely, a part due to time discretization and another one due to space discretization. This splitting of errors is important because one wishes to adapt both time and space meshes and, therefore, one needs an estimator for space that, in general, will take different values that the estimator used for time. The post-processing procedure and examples of application for stationary and diffusion-reaction equation can be seen in the paper [4].

4 Error indicators

To design a practical adaptive algorithm is customary to use error indicators instead of the a posteriori error estimates. The space and time error indicators of the convective and diffusive-reaction stages are defined as

$$\begin{cases} \eta_s^n = \sum\limits_{K\in\mathbb{T}_h^n} \eta_{sK,convect}^n + \sum\limits_{K\in\mathbb{T}_h^n} \eta_{sK,diff-react}^n, \\ \eta_t^n = \sum\limits_{K\in\mathbb{T}_h^n} \eta_{tK,convect}^n + \sum\limits_{K\in\mathbb{T}_h^n} \eta_{tK,diff-react}^n. \end{cases}$$

where with $p = s$ or t

$$\eta_{pK,convect}^n = \frac{\left|e_{pK,convect}^n\right|}{\left|J(\overline{u}_{h\Delta t}^{n-1})\right|} \quad \text{and} \quad \eta_{pK,diff-react}^n = \frac{\left|e_{pK,diff-react}^n\right|}{|J(u_{h\Delta t}^n)|}.$$

Prescribing a tolerance $TOL = Tol_s + Tol_t$, where Tol_s and Tol_t are the space and time tolerances respectively, the adaptive algorithm will accept the numerical solution $u_{h\Delta t}^n$ if

$$\eta_s^n \le Tol_s \quad \text{and} \quad \eta_t^n \le Tol_t.$$

To balance the space and time errors one chooses $Tol_s \approx Tol_t$.

4.1 Mesh adaptation: mesh-optimization strategy.

The criterium to adapt the spatial mesh consists of calculating such a mesh with the minimum number of elements NE to satisfy $\eta_s^n \le Tol_s$. This yields (see [2] and [6]) the optimal size h_K of the element $K \in \mathbb{T}_h^n$ as

$$h_K^{opt} = h_K \left(\frac{Tol_s}{W}\right) (\eta_K^n)^{-1/(\alpha+d)}, \tag{17}$$

where $W := \sum\limits_{K\in\mathbb{T}_h^n} (\eta_K^n)^{\frac{d}{d+\alpha}} < \infty$; $\eta_K^n = \eta_s^n|_K$ is the spatial error indicator for the element K; α is the convergence rate of the spatial error, it is assumed that $\eta_s^n = O(h^\alpha)$ and $d = 2$ or 3 is the spatial dimension. Formula (17) gives a criterion to refine or coarsen the elements K. Specifically, comparing the size of our actual triangle h_K with the optimal size h_K^{opt} we obtain the number of times that this element needs to be refined or coarsened. Thus, we adopt the following refining and coarsening criteria:

Refining criterion:
If $\dfrac{h_K^{opt}}{h_K} \le 1$, mark the element K to refine n_r times,

$$n_r = Integer\ part\left[0.5 + 2\frac{\log(h_K/h_K^{opt})}{\log 2}\right].$$

Coarsening criterion:
If $\dfrac{h_K^{opt}}{h_K} > 1$, mark the element K to coarsen n_c times,

$$n_c = Integer\ part\left[2\frac{\log(h_K^{opt}/h_K)}{\log 2}\right].$$

The refinement of marked elements is made bisecting the largest edge by joining its mid-point with the opposite vertex and taking the vertices thus created as the vertices of a new refinement. To maintain the regularity of the mesh in the refining and coarsening procedure we follow the strategy of [12].

4.2 Adaptation of the time step size

The error indicator η_t^n is used to adjust the size of the new time step whether the solution $u_{h\Delta t}^n$ is accepted or not. Following the strategy of the numerical ODE community (see [10]) we adjust the time step by the formula

$$\begin{aligned}
&\Delta t_{new} = \min(fac_{\max}, \max(fac_{\min}, fac))\Delta t_{old},\\
&fac = \begin{cases} \left(\left(\dfrac{\eta_t^{n-1}}{\eta_t^n}\right)^{1/\beta}\dfrac{\Delta t_n}{\Delta t_{n-1}}\right)\left(\dfrac{\kappa\cdot Tol_t}{\eta_t^n}\right)^{1/\beta} & \text{when } u_{h\Delta t}^n \text{ is accepted,}\\ \left(\dfrac{\kappa\cdot Tol_t}{\eta_t^n}\right)^{1/\beta} & \text{when } u_{h\Delta t}^n \text{ is rejected,}\end{cases}
\end{aligned} \tag{18}$$

where β is an unknown coefficient, which is equal to the order of the time local truncation error, and is calculated by a recursive procedure (see [4]); $fac_{\max}$ and $fac_{\min}$ are factors limiting the maximum and minimum step sizes respectively, usually $fac_{\max} = 5$ and $fac_{\min} = 0.2$; and κ is a security factor to prevent unnecessary rejections because they cause recomputation and, therefore, loss of performance. In our examples $\kappa = 0.7$.

5 Numerical experiments

To illustrate the performance of the adaptive algorithm proposed in this paper we present the results of two hard numerical tests.

5.1 Example 1. A convection-diffusion problem

We consider a convection-diffusion problem with both internal and boundary layers taken from [8]. The equation of the problem is:

$$\begin{cases} \dfrac{\partial u}{\partial t} + \mathbf{a}(x,t)\cdot\nabla u = \epsilon\Delta u \quad \text{in } \Omega\times(0,T],\\ u(x,t) = g(x,t) \quad \text{on } \partial\Omega,\\ u(x,0) = u^0(x). \end{cases} \tag{19}$$

Here, $\Omega = (0,1)^2$, $T = 0.55$, the diffusion coefficient $\epsilon = 10^{-3}$ and $\mathbf{a}(x,t) = [2,\ 1]^T$. The initial condition $u^0(x)$is given as: $u^0(x) = 0$ for $x = (x_1,\ x_2) \in \Omega_\delta = (\delta, 1) \times (0, 1-\delta)$. For $x \in \Omega \backslash \Omega_\delta$, $u^0(x)$ is defined to be the linear function which satisfies the boundary conditions.

$$g(x,t) = \begin{cases} 1 & \text{for } x_1 = 0, \quad 0 \le x_2 \le 1, \\ 1 & \text{for } 0 \le x_1 \le 1, \quad x_2 = 1, \\ \dfrac{(\delta - x_1)^+}{\delta} & \text{for } 0 \le x_1 \le 1, \quad x_2 = 0, \\ \dfrac{(x_2 - 1 + \delta)^+}{\delta} & \text{for } x_1 = 1, \quad 0 \le x_2 \le 1, \end{cases}$$

where $(a)^+ = \max(0, a)$ and $\delta = 7.8125 \times 10^{-3}$.

In this example we take as output functional $J(u) = \int_\Omega (1 - u^n) d\Omega$ and the tolerances $Tol_s = Tol_t = 2 \cdot 10^{-4}$. Note that for δ small the initial condition exhibits two boundary layers along $x_1 = 0$ and $x_2 = 1$. As time progresses, the boundary layer along $x_1 = 0$ propagates into the interior and interacts with the outflow boundary at $x_1 = 1$ at time $t = 0.5$ developing a new boundary layer. Different details of the internal boundary layer at $t = 0.12$ are represented in Figure 2. The variation of the number of nodes and the size of Δt as functions of the number of time steps are shown in Figure 3.

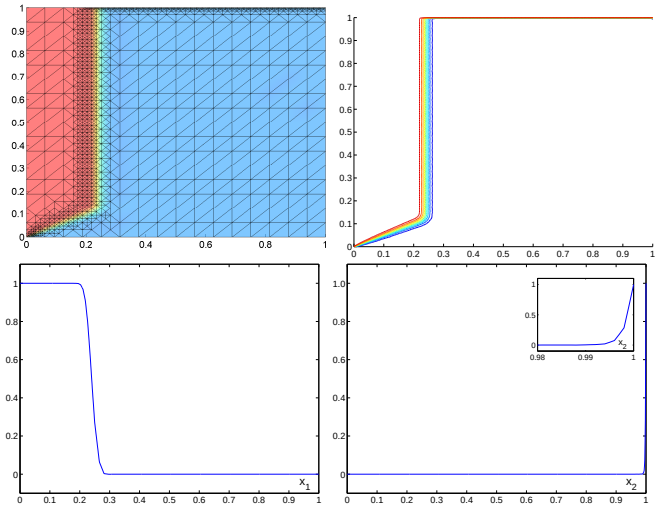


Figure 2: Solution at time t=0.12. On the top: mesh (on the left) and isolines of the solution (on the right) from $u = 0.1$ to $u = 0.9$ at intervals $\Delta u = 0.1$. At the bottom: cross-section at $x_2 = 0.75$ (on the left) and cross-section at $x_1 = 0.5$ (on the right).

We must remark that the total number of time steps to complete the integration is 148 with 4 steps being rejected. The CPU time of the execution

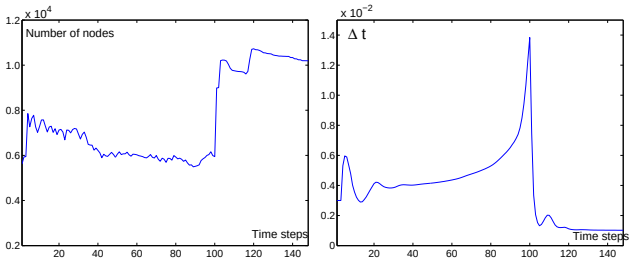


Figure 3: Number of nodes and time step size Δt against the number of time steps.

is 147 *seg.*, with the semi-Lagrangian stage and the mesh and time adaptation consuming a 10% of the CPU time, whereas solving the primal problem and calculating the a posteriori error estimators require 57% and 33% of the CPU time respectively.

5.2 Example 2. Lifted Flames Problem

As a final example we apply our algorithm to simulate the lift-off and blow-off of a diffusion flame generated in a stream of fuel (methane diluted with nitrogen) interacting with an air stream emerging from porous walls. To do so, we consider the systems of equations composed by the compressible Navier-Stokes equations at low Mach number and the convection-diffusion-reaction equations for temperature and species plus the state equation in a bounded domain $\Omega \subset \mathbb{R}^2$ with appropriately smooth boundary $\partial\Omega = \Gamma^D \cup \Gamma^N$, $\Gamma^D \cap \Gamma^N = \emptyset$, where Γ^D and Γ^N are the pieces of $\partial\Omega$ for Dirichlet and Neumann boundary conditions, respectively. The variables of the problem are the density of fluid ρ, the hydrodynamic correction of pressure p, the flow velocity $\mathbf{u} = (u_1, u_2)$, the temperature T and the species mass fractions $Y_{i=F,O_2,N_2,P}$, where F, O_2, N_2 and P stand for fuel, oxygen, nitrogen and products of combustion respectively. The system of equations of the model is

$$\left.\begin{array}{ll} \dfrac{\partial \rho}{\partial t} + \nabla \cdot (\rho \mathbf{u}) = 0, & \\ \rho \quad \dfrac{\partial \mathbf{u}}{\partial t} + \mathbf{u} \cdot \nabla \mathbf{u} & = \nabla \cdot (\mu \nabla \mathbf{u}) - \nabla p \end{array}\right\} \quad \mathbf{u} = \mathbf{u}_D \text{ en } \Gamma_{\mathbf{u}}^D \ , \ \mu \tfrac{\partial \mathbf{u}}{\partial \mathbf{n}} - p\mathbf{n} = \mathbf{0} \ \text{ en } \ \Gamma_{\mathbf{u}}^N$$

$$\rho \quad \frac{\partial T}{\partial t} + \mathbf{u} \cdot \nabla T \quad = \nabla \cdot (\rho D_T \nabla T) - H_F w_F \qquad T{=}T_D \text{ en } \Gamma_T^D \ , \ \rho D_T \tfrac{\partial T}{\partial \mathbf{n}}{=}0 \ \text{ en } \ \Gamma_T^N$$

$$\rho \quad \frac{\partial Y_i}{\partial t} + \mathbf{u} \cdot \nabla Y_i \quad = \nabla \cdot (\rho D_i \nabla Y_i) + w_i \qquad Y_i = Y_i^D \ \text{ en } \ \Gamma_{Y_i}^D \ , \ \rho D_{Y_i} \tfrac{\partial Y_i}{\partial \mathbf{n}}{=}0 \ \text{ en } \ \Gamma_{Y_i}^N$$

$$\begin{array}{l} Y_{N_2} = 1 - Y_F - Y_{O_2} - Y_P \\ \rho = \dfrac{M}{T} \dfrac{\rho_o T_o}{M_o} \end{array} \tag{20}$$

We take into account the effects of thermal expansion and assume a one-step overall Arrhenius reaction. The phenomenology of the lifted flame and the values of the coefficients and reaction rates can be seen in [5] and [3].

The output functional for this problem is $J(Y^n) = \int_\Omega \left(Y_{CH_4}^n \cdot Y_{O_2}^n\right) d\Omega$ and the tolerances are $Tol_s = Tol_t = 2 \cdot 10^{-4}$. This output functional is useful to capture well the main features of the mixing layer and the flame front.

In order to validate the solution achieved by our numerical method, we have compared our results with those provided by [9] in a mixing layer between two parallel streams of fuel and air case. In Figure 4 shows the lifted distance x_f/δ_L as function of the injection velocity U/S_L and the concentration of the fuel feed stream $Y_{CH_4,0}$, the solid lines represent the numerical results of [9] using asymptotic techniques and a non-adaptive finite difference scheme, whereas the results of our method are circles. We can observe that there is a good agreement for values of $Y_{CH_4,0} \geq 0.2$, although remarkable differences arise for $Y_{CH_4,0} = 0.1$.

Other configurations which there are few results in the literature can be studied with our fully adaptive procedure. One of them is the planar jet, where a fuel feed stream goes into the computational domain normal to an injector of width $2a$, with uniform velocity U/S_L; the air emerges from porous walls located above and below the injector, with the same velocity U/S_L. When we provoke the ignition symmetrically in both mixing layers, the flame fronts move together and reach an apparent symmetric steady solution, but that situation is not stable and the interaction of the flames breaks the symmetry of the flame and an asymmetric steady solution is reached (on the top of Figure 5 we show the two steady configurations for a planar jet). Symmetry breaking has been observed in laboratory experiments of coaxial jet flames, as we show in the photograph (at the bottom of Figure 5) taken by Pablo Martinez and Jean-Marie Truffaut.

Figure 6 shows the evolution of the number of nodes and the time step size versus the number of time steps. We can see the suddenly increase of the number of nodes and the drop of the time step size when the ignition

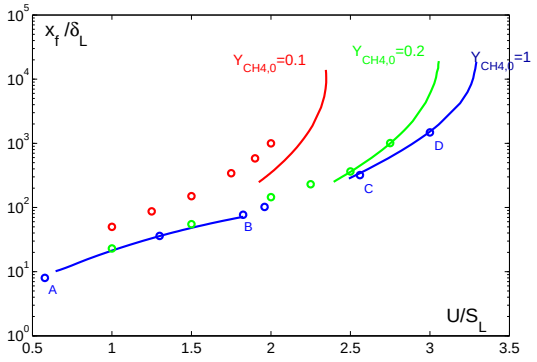


Figure 4: Lifted length x_f/δ_L versus velocity U/S_L and the concentration $Y_{CH_4,0}$.

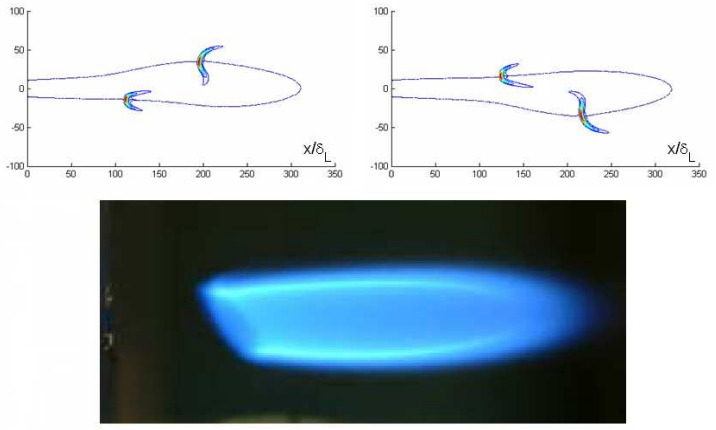


Figure 5: Top panel: Symmetry breaking of two stable solutions with $Y_{CH_4} = 0.1$ and $U/S_L = 1.5$. Bottom panel: Photograph of a laboratory experiment of a coaxial jet.

is provoked. The distribution of the CPU time is the following: The semi-Lagrangian adaptive stage 3%, the diffusion-reaction equation for temperature and chemical species 44%, the Stokes problem 31% and the calculation of the a posteriori error estimator consumes 22%.

References

[1] A. Allievi, R. Bermejo. 'A Generalized Particle Search-Locate Algorithm for Arbitrary Grid'. *J. Comp. Physics*, Vol. 132 (1997), pp. 157-166.

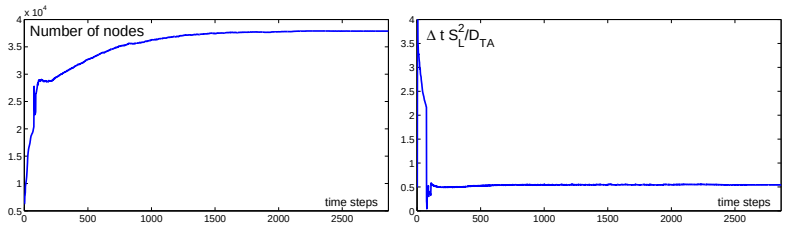


Figure 6: Number of nodes and time step size Δt against the number of time steps.

[2] W. Bangerth, R.Rannacher. 'Adaptive Finite Element Methods for Differential Equations'. *Birkhäuser, Basel*, (2003).

[3] R. Bermejo and J. Carpio. 'An adaptive finite element semi-Lagrangian implicit-explicit Runge-Kutta-Chebyshev method for convection dominated reaction-diffusion problems'. *Applied Numerical Mathematic* (2006), doi:10.1016/j.apnum 2006.10.008.

[4] R. Bermejo and J. Carpio. 'A space-time adaptive finite element algorithm based on dual weighted residual methodology for parabolic equations'. Submitted.

[5] R. Bermejo and J. Carpio. 'An adaptive finite element semi-Lagrangian Runge-Kutta-Chebyshev method for combustion problems'. *Applied and Industrial Mathematics in Italy II*, editted by V. Cutello, G. Fotia and L. Puccio. Series on Advances in Mathematics for Applied Sciences. World Scientific. Vol. 75 (2007), pp. 149-160.

[6] M. Braack. 'An adaptive finite element method for reactive flow problems'. *Doktorarbeit, Institut für Angewandte Mathematik, Universität Heidelberg* (1998).

[7] J. Carpio. 'Duality methods for time-space adaptivity to calculate the numerical solution of partial differential equations' *PhD Thesis. Universidad Politécnica de Madrid.* (2008).

[8] P. Houston, E. Süli. 'Adaptive Lagrange-Galerkin methods for unsteady convection-diffusion problems' *Math. of Computation*, Vol. 70 (2000), pp. 77-106.

[9] E. Fernandez-Tarrazo, M. Vera, A. Liñan. 'Lift-off and blow off of a diffusion flame between parallel streams of fuel and air'. *Combust. Flame* Vol. 144 (2006), pp. 261-276.

[10] K. Gustafsson, M. Lundh and G. Söderlind. 'A PI step size control for the numerical solution of ordinary differential equations' *BIT* Vol. 28 (1988), pp. 270-287.

[11] M. Schmich and B. Vexler. 'Adaptivity with dynamic meshes for space-time finite element discretizations of parabolic equations'. To appear in *SIAM J. Sci. comput.* (2006).

[12] A. Schmidt, K. G. Siebert. 'Designed of Adaptive Finite Element Software. the Finite Element Toolbox ALBERTA'. *Springer Lectures Notes in Computational Science and Engineering*, Springer Berlin (2005).

Bol. Soc. Esp. Mat. Apl.
nº 42(2008), 163–169

CONSTRUCCIÓN ALGEBRAICO-GEOMÉTRICA DE CÓDIGOS CONVOLUCIONALES

J.A. DOMÍNGUEZ PÉREZ, J.I. IGLESIAS CURTO,
J.M. MUÑOZ PORRAS, G. SERRANO SOTELO

Departamento de Matemáticas
Universidad de Salamanca.

jadoming@usal.es, joseig@usal.es, jmp@usal.es, laina@usal.es

Resumen

Las técnicas geométricas para construir los códigos de Goppa sobre una curva algebraica sobre un cuerpo finito $\mathbb{F}_q$ pueden extenderse a la construcción de códigos convolucionales sobre el cuerpo infinito de funciones racionales en una variable $\mathbb{F}_q(z)$. De este modo es posible constuir códigos convolucionales sobre la recta proyectiva o sobre curvas elípticas que alcanzan la cota de Singleton generalizada.

Palabras clave: *Curvas algebraicas, códigos algebraico-geométricos, códigos de Goppa, códigos convolucionales.*
Clasificación por materias AMS: *11T71 94B10*

1 Introducción

Los códigos de Goppa permiten incorporar las técnicas de la Geometría Algebraica de curvas a la construcción de códigos lineales, lo que supone disponer de potentes herramientas para el cálculo de los parámetros de estos códigos, así como la determinación de las condiciones de distancia óptima y el diseño de algoritmos de decodificación. A este respecto, presentaremos en la sección §2 un resumen de la teoría de códigos algebraico-geométricos.

Esta línea de trabajo puede trasladarse a la teoría algebraica de Forney de códigos convolucionales ([2], [4]), como veremos en la sección §3, lo que permitirá disponer de un método sistemático para construir códigos convolucionales cuyas propiedades se deducen de su estructura geométrica.

Así, daremos en la sección §4 ejemplos particulares en el caso de que la curva algebraica sea la recta proyectiva, donde los códigos alcanzan su distancia óptima (cota de Singleton generalizada), y otros ejemplos sobre curvas elípticas en la sección §5 donde la distancia se aproxima a dicha cota.

Trabajo subvencionado por el proyecto de investigación SA028A05 de la Junta de Castilla y León.

2 Códigos de Goppa

Para ampliar en contenido de esta sección puede consultarse [3] o [6].

Sea X una curva proyectiva lisa y geométricamente irreducible de género g sobre el cuerpo finito $\mathbb{F}_q$, y sea Σ_X el cuerpo de funciones racionales sobre X. Sea $P_1, \ldots, P_n$ un conjunto de n puntos $\mathbb{F}_q$-racionales distintos de X, y D el correspondiente divisor

$$D = P_1 + \cdots + P_n \, .$$

Considerando otro divisor $G = \sum n_i Q_i - \sum n'_j Q'_j$ con soporte disjunto a D, y su espacio de secciones globales $L(G)$,

$$L(G) = \left\{ f \in \Sigma_X / \begin{array}{c} \text{tiene ceros al menos en los puntos } Q'_j \text{ de orden} \geq n'_j \\ \text{tiene polos sólo en los puntos } Q_i \text{ de orden} \leq n_i \end{array} \right\} ,$$

cuyo grado $\operatorname{Gr} G = \sum n_i - \sum n'_j$ sea $\operatorname{Gr} G < n$, se tiene una aplicación $\mathbb{F}_q$-lineal inyectiva

$$\begin{aligned} \alpha \colon L(G) &\to \mathbb{F}_q \times \overset{n}{\cdots} \times \mathbb{F}_q \\ f &\mapsto (f(P_1), \ldots, f(P_n)) \end{aligned}$$

cuyo subespacio imagen determina el **código de Goppa** $\mathcal{C}(D, G) \subset \mathbb{F}_q^n$. Por construcción, la longitud de este código es n, y el resto de sus parámetros pueden determinarse a partir del teorema de Riemann-Roch: su dimensión k y su distancia mínima d verifican

$$k \geq \deg(G) + 1 - g, \;\; d \geq n - \deg(G) \, ,$$

de modo que si $2g-2 < \operatorname{Gr}(G)$, la dimensión es exactamente $k = \deg(G)+1-g$.

El correspondiente código dual puede construirse a partir de las formas diferenciales sobre la curva Ω_X. Denotando $\Omega(G)$ el espacio

$$\Omega(G) = \left\{ \omega \in \Omega_X / \begin{array}{c} \text{tiene ceros al menos en los puntos } Q_i \text{ de orden} \geq n_i \\ \text{tiene polos sólo en los puntos } Q'_j \text{ de orden} \leq n'_j \end{array} \right\} ,$$

por la dualidad de Serre el cálculo de residuos Res() produce una aplicación inyectiva

$$\begin{aligned} \beta \colon \Omega(G - D) &\to \mathbb{F}_q \times \overset{n}{\cdots} \times \mathbb{F}_q \\ \omega &\mapsto (\operatorname{Res}_{P_1}(\omega), \ldots, \operatorname{Res}_{P_n}(\omega)) \end{aligned}$$

cuyo subespacio imagen, de dimensión $n - k$, determina el **código de Goppa dual** $\mathcal{C}(D, G)^* \subset \mathbb{F}_q^n$, como puede comprobarse por el teorema de los residuos.

3 Códigos de Goppa convolucionales

Para ampliar en contenido de esta sección puede consultarse [1] o [5].

Sea $\mathbb{F}_q(z)$ el cuerpo (infinito) de funciones racionales en una variable, y X una curva proyectiva lisa sobre $\mathbb{F}_q(z)$. Las construcciones de la sección anterior siguen siendo válidas en esta situación, de modo que si $D = P_1 + \cdots + P_n$ es un

divisor de n puntos $\mathbb{F}_q(z)$-racionales distintos de X y G otro divisor con soporte disjunto a D, tal que

$$2g-2 < \mathrm{Gr}(G) < n\,,$$

sobre el $\mathbb{F}_q(z)$-espacio de secciones globales $L(G)$ se tiene una aplicación $\mathbb{F}_q(z)$-lineal inyectiva

$$\begin{aligned}\alpha\colon L(G) &\to \mathbb{F}_q(z)\times \overset{n}{\cdots}\times \mathbb{F}_q(z)\\ f &\mapsto (f(P_1),\ldots,f(P_n))\end{aligned}$$

cuya imagen es el **código de Goppa convolucional** $\mathcal{C}(D,G)\subset \mathbb{F}_q^n(n)$ de longitud n y dimensión $\mathrm{Gr}(G)+1-g$. Y análogamente puede realizarse también la construcción del **código de Goppa convolucional cual** $\mathcal{C}(D,G)^*\subset \mathbb{F}_q^n(n)$.

4 Ejemplos sobre $\mathbb{P}_1$

Consideremos la recta proyectiva sobre $\mathbb{F}_q(z)$

$$X=\mathbb{P}^1_{\mathbb{F}_q(z)}=\mathrm{Proj}\,\mathbb{F}_q(z)[x_0,x_1]\,,\text{con } t=x_1/x_0 \text{ la coordenada afín}\,.$$

Sea $P_0=(1,0)$ el punto origen, y $P_\infty=(0,1)$ el punto del infinito, y consideremos n puntos racionales $P_1,\ldots,P_n$ distintos, $P_i=(1,\alpha_i)\neq P_0,P_\infty$.

Sea $D=P_1+\cdots+P_n$ y $G=rP_\infty-sP_0$, con $0\le s\le r<n$, de modo que

$$L(G)=\langle t^s,t^{s+1},\ldots,t^r\rangle$$

y el código de Goppa convolucional $\mathcal{C}(D,G)$ es la imagen de la aplicación

$$\begin{aligned}\alpha\colon L(G) &\to \mathbb{F}_q(z)\times \overset{n}{\cdots}\times \mathbb{F}_q(z)\\ t^i &\mapsto (\alpha_1^i,\ldots,\alpha_n^i)\end{aligned}$$

Se obtiene de este modo un código convolucional $\mathcal{C}(D,G)$ de longitud n y dimensión $k=r-s+1$, cuya matriz generadora es

$$G=\begin{pmatrix}\alpha_1^s & \alpha_2^s & \ldots & \alpha_n^s\\ \alpha_1^{s+1} & \alpha_2^{s+1} & \ldots & \alpha_n^{s+1}\\ \vdots & \vdots & \ddots & \vdots\\ \alpha_1^r & \alpha_2^r & \ldots & \alpha_n^r\end{pmatrix}.$$

En cuanto al código dual, como

$$\Omega(G-D)=\left\langle \frac{dt}{t^s\prod_{i=1}^n(t-\alpha_i)},\frac{t\,dt}{t^s\prod_{i=1}^n(t-\alpha_i)},\ldots,\frac{t^{n-r+s-2}dt}{t^s\prod_{i=1}^n(t-\alpha_i)}\right\rangle$$

y calculando los residuos

$$\mathrm{Res}_{P_j}\left(\frac{t^m dt}{t^s\prod_{i=1}^n(t-\alpha_i)}\right)=\frac{\alpha_j^m dt}{\alpha_j^s\prod_{\substack{i=1\\ i\neq j}}^n(\alpha_j-\alpha_i)}$$

resulta que $\mathcal{C}(D,G)^*$ es un código convolucional de longitud n y dimensión $n-k=n-r+s-1$ que tiene como matriz generadora (= matriz de control de $\mathcal{C}(D,G)$)

$$H=\begin{pmatrix} h_1 & h_2 & \dots & h_n \\ h_1\alpha_1 & h_2\alpha_2 & \dots & h_n\alpha_n \\ \vdots & \vdots & \ddots & \vdots \\ h_1\alpha_1^{n-r+s-2} & h_2\alpha_2^{n-r+s-2} & \dots & h_n\alpha_n^{n-r+s-2} \end{pmatrix},$$

donde $h_j=\frac{1}{\alpha_j^s\prod_{\substack{i=1\\ i\neq j}}^{n}(\alpha_j-\alpha_i)}$.

Obsérvese que esa matriz generadora H de $\mathcal{C}(D,G)^*$ tiene la forma de un "codificador alternante" sobre $\mathbb{F}_q(z)$, lo que sugiere la posibilidad de emplear en este contexto convolucional los algoritmos algebraicos de decodificación conocidos para los códigos lineales alternantes.

Como casos particulares, vamos a calcular las matrices G y H en el caso de puntos $\mathbb{F}_q(z)$-racionales $P_i=(1,\alpha_i)$ cuando

$$\alpha_i=a^{i-1}z+b^{i-1}\,,\quad i=1,\dots,n\,,\quad n<q\,,$$

comprobando además que en estos casos la distancia mínima d de los códigos convolucionales que resultan alcanza la cota de Singleton generalizada

$$d\leq(n-k)(\lfloor\delta/k\rfloor+1)+\delta+1$$

donde δ denota el grado del código convolucional.

- Cuerpo $\mathbb{F}_3(z)=\{0,1,2\}$

  Caso $a=1,b=2,G=P_\infty-P_0$.
  $G=\begin{pmatrix} z+1 & z+2 \end{pmatrix}$
  $H=\begin{pmatrix} \frac{1}{2(z+1)} & \frac{1}{z+2} \end{pmatrix}$
  $(n,k,\delta,d)=(2,1,1,4)$

- Cuerpo $\mathbb{F}_4(z)=\{0,1,\xi,\xi^2\}$, con $\xi^2+\xi+1=0$

  Caso $a=\xi,b=\xi^2,G=P_\infty$
  $G=\begin{pmatrix} \frac{1}{z+1} & \frac{1}{\xi z+\xi^2} & \frac{1}{\xi^2 z+\xi} \end{pmatrix}$
  $H=\begin{pmatrix} \frac{1}{(\xi^2z+\xi)(\xi z+\xi^2)} & \frac{1}{(\xi^2z+\xi)(z+1)} & \frac{1}{(\xi z+\xi^2)(z+1)} \end{pmatrix}$
  $(n,k,\delta,d)=(3,2,1,3)\,.$

  Caso $a=1,b=\xi,G=P_\infty-P_0$
  $G=\begin{pmatrix} z+1 & z+\xi & z+\xi^2 \end{pmatrix}$

$$H = \begin{pmatrix} \frac{1}{z+1} & \frac{\xi}{z+\xi} & \frac{\xi^2}{z+\xi^2} \\ 1 & \xi & \xi^2 \end{pmatrix}$$
$$(n,k,\delta,d) = (3,1,1,6)\,.$$

- Cuerpo $\mathbb{F}_5(z)$

Caso $a=1, b=2, G = 2P_\infty - 2P_0$

$$G = \begin{pmatrix} (z+1)^2 & (z+2)^2 & (z+4)^2 \end{pmatrix}$$
$$H = \begin{pmatrix} \frac{2}{(z+1)^2} & \frac{2}{(z+2)^2} & \frac{1}{(z+4)^2} \\ \frac{2}{z+1} & \frac{2}{z+2} & \frac{1}{z+4} \end{pmatrix}$$
$$(n,k,\delta,d) = (3,1,2,9)\,.$$

Caso $a=2, b=3, G = 2P_\infty - P_0$

$$G = \begin{pmatrix} z+1 & 2z+3 & 4z+4 & 3z+2 \\ (z+1)^2 & (2z+3)^2 & (4z+4)^2 & (3z+2)^2 \end{pmatrix}$$
$$H = \begin{pmatrix} \frac{4}{(z+1)^2(z+2)(z+3)} & \frac{4}{(z+2)(z+3)(z+4)^2} & \frac{4}{(z+1)^2(z+2)(z+3)} & \frac{4}{(z+2)(z+3)(z+4)^2} \\ \frac{4}{(z+1)(z+2)(z+3)} & \frac{3}{(z+2)(z+3)(z+4)} & \frac{1}{(z+1)(z+2)(z+3)} & \frac{2}{(z+2)(z+3)(z+4)} \end{pmatrix}$$
$$(n,k,\delta,d_{free}) = (4,2,3,8)\,.$$

5 Ejemplos sobre curva elíptica

Consideremos el plano proyectivo sobre $\mathbb{F}_q(z)$

$$\mathbb{P}^2_{\mathbb{F}_q(z)} = \operatorname{Proj} \mathbb{F}_q(z)[x_0,x_1,x_2]\,, \text{con } x = x_1/x_0, y = x_2/x_0 \text{ coordenadas afines}\,.$$

Sea $X \subset \mathbb{P}^2_{\mathbb{F}_q(z)}$ una curva elíptica plana, expresada en la forma normal de Tate por la ecuación

$$y^2 + axy + by = x^3 + bx^2$$

Sea $P_0 = (1,0,0)$ el punto origen, $P_\infty = (0,1,0)$ el punto del infinito, y $P_1,\ldots,P_n$ puntos racionales distintos $P_i = (1,x_i,y_i)$, $P_i \neq P_0, P_\infty$. Considerando los divisores $D = P_1 + \cdots + P_n$ y $G = rP_\infty - sP_0$ con $0 < r-s < n$,

$$L(G) = \langle x^a y^b, \ldots, x^i y^j, \ldots \rangle\,, \quad a+2b = s, 2i+3j \leq r\,,$$

el código de Goppa convolucional $\mathcal{C}(D,G)$ que resulta es la imagen de la aplicación

$$\alpha\colon L(G) \to \mathbb{F}_q(z) \times \overset{n}{\cdots} \times \mathbb{F}_q(z)$$
$$x^i y^j \mapsto (x_1^i y_1^j, \ldots, x_n^i y_n^j)$$

$\mathcal{C}(D,G)$ es por tanto de longitud n y dimensión $r-s$, con matriz generadora

$$G = \begin{pmatrix} x_1^a y_1^b & x_2^a y_2^b & \cdots & x_n^a y_n^b \\ x_1^{a+1} y_1^b & x_2^{a+1} y_2^b & \cdots & x_n^{a+1} y_n^b \\ \vdots & \vdots & \ddots & \vdots \\ x_1^c y_1^d & x_2^c y_2^d & \cdots & x_n^c y_n^d \end{pmatrix}$$

Veamos algunos ejemplos, sobre la curva elíptica

$$X \equiv y^2 + (1+z)xy + (z+z^2)y = x^3 + (z+z^2)x^2\,,$$

variando el cuerpo y la elección de los puntos

- Cuerpo $\mathbb{F}_2(z)$

Caso $n = 4$, $r = 3$ y $s = 0$,
$P_1 = (z^3 + z^2, 0)$,
$P_2 = (0, z^3 + z^2)$,
$P_3 = (z^3 + z^2, z^5 + z^3)$,
$P_4 = (z^2 + z, z^4 + z^2)$
$G = 3P_\infty$, $L(G) = \langle 1, \frac{1}{z(z+1)}x, \frac{1}{z^2(z+1)y}\rangle$

$$G = \begin{pmatrix} 1 & 1 & 1 & 1 \\ z & 0 & z & 1 \\ 0 & 1 & z+z^2 & 1+z \end{pmatrix}$$

$(n, k, \delta, d) = (4, 3, 3, 3)$, siendo 4 la cota de la distancia.

- Cuerpo $\mathbb{F}_q(z)$, $q \neq 2^m$

Caso $n = 3$, $r = 2$ y $s = 0$,
$P_1 = (0, -z^3 - z^2)$,
$P_2 = (z^2 - z, -z^4 - 2z^3 + z^2)$,
$P_3 = (-z^2 - z, -z^3 + z)$,
$G = 2P_\infty$, $L(G) = \langle 1, \frac{1}{z}x\rangle$

$$G = \begin{pmatrix} 1 & 1 & 1 \\ 0 & z-1 & -z-1 \end{pmatrix}.$$

$(n, k, \delta, d) = (3, 2, 1, 3)$, alcanza la cota de la distancia.

Caso $n = 3$, $r = 3$ y $s = 2$,
$P_1 = (0, -z^3 - z^2)$,
$P_2 = (z^2 - z, -z^4 - 2z^3 + z^2)$,
$P_3 = (-z^2 - z, z^4 + 2z^3 + z^2)$
$G = 3P_\infty - 2P_0$, $L(G) = \langle \frac{1}{z^2}y\rangle$

$$G = \begin{pmatrix} -1-z & 1-2z-z^2 & 1+2z+z^2 \end{pmatrix}.$$

$(n, k, \delta, d) = (3, 1, 2, 8)$, siendo 9 la cota de la distancia.

Referencias

[1] J.A. Domínguez Pérez, J.M. Muñoz Porras, and G. Serrano Sotelo, Convolutional codes of Goppa type, *Appl. Algebra Engrg. Comm. Comput.*, **15** no. 1 (2004), 51–61.

[2] G.D. Forney Jr., Convolutional Codes I: Algebraic Structure, *IEEE Trans. Inform. Theory* **16** (1970) 720–738.

[3] T. Høholdt, J.H. van Lint and R. Pellikaan, Algebraic Geometric Codes, in: *Handbook of Coding theory*, Ed. by V.S. Pless and W.C. Huffman (Elsevier, Amsterdam, 1998) 871–962.

[4] R.J. McEliece, The Algebraic Theory of Convolutional Codes, in: *Handbook of Coding theory*, Ed. by V.S. Pless and W.C. Huffman (Elsevier, Amsterdam, 1998) 1065–1138.

[5] J.M. Muñoz Porras, J.A. Domínguez Pérez, J.I. Iglesias Curto, and G. Serrano Sotelo, Convolutional Goppa codes, *IEEE Trans. Inform. Theory* **52** (2006) 340–344.

[6] J.H. van Lint and G. van der Geer, *Introduction to Coding Theory and Algebraic Geometry* DMV Seminar, vol. 12, (Birkhäuser, Basel, 1998).

Bol. Soc. Esp. Mat. Apl.
nº 42(2008), 171–181

CONSTRUCCIÓN DE CÓDIGOS CONVOLUCIONALES UTILIZANDO LA TÉCNICA DE CONCATENACIÓN DESDE EL PUNTO DE VISTA DE SISTEMAS LINEALES

VICTORIA HERRANZ Y CARMEN PEREA

Centro de Investigación Operativa
Universidad Miguel Hernández de Elche.

mavi.herranz@umh.es, perea@umh.es

Resumen

En este trabajo construimos códigos convolucionales con parámetros fijos a partir de otros códigos convolucionales utilizando la concatenación en serie de dos códigos convolucionales. Establecemos condiciones para que los nuevos códigos obtenidos tengan una representación minimal y observable. Asímismo, presentamos cotas inferiores de la distancia libre de los nuevos códigos.

Palabras clave: *Código bloque, código convolucional, representación entrada-estado-salida, concatenación en serie.*
Clasificación por materias AMS: *93B05 93B07 93B20 94B05 94B10*

1 Introducción y resultados previos

La teoría de códigos correctores de errores se puede dividir, salvo pequeñas excepciones, en dos categorías: la teoría de códigos bloque y la teoría de códigos convolucionales. En los códigos bloque, en cada instante de tiempo una palabra de longitud K se codifica en una palabra de longitud N con $N > K$. Análogamente ocurre en los códigos convolucionales. Sin embargo, en este caso la palabra codificada en un instante de tiempo no sólo depende de la entrada en ese instante, sino que también depende de un número finito de entradas anteriores.

Los códigos convolucionales pueden considerarse como una extensión natural de los códigos bloque puesto que un (n, k)-código convolucional puede definirse como un subespacio vectorial k-dimensional de $\mathbb{F}((Z))$ con bases formadas completamente por vectores de $\mathbb{F}(Z)$ (para más detalle véanse [15] y [18]). Los códigos convolucionales nos permiten obtener buenos resultados desde la perspectiva de la detección y corrección de errores sin tener que recurrir a códigos bloque muy grandes, es decir, sin tener que trabajar con códigos bloques

Trabajo subvencionado por los proyectos MTM2005-05759, SA028A05 y E-GV06\078

sobre cuerpos finitos excesivamente grandes en ls que haya que introducir demasiada información redundante. Es importante señalar que, mientras que en la práctica es usual utilizar códigos bloque en los que $N = 1023$, en la literatura de códigos convolucionales encontramos códigos convolucionales de tasa $\frac{1}{2}$, es decir, $n = 2$.

Por otra parte, la teoría algebraica de códigos convolucionales no ha avanzado tanto como la teoría algebraica de códigos bloque. De hecho, en un principio, la mayor parte de códigos convolucionales fueron descubiertos por medio de cálculos computacionales. Posteriormente, se introdujeron construcciones basadas en la relación de las matrices generadoras de un código convolucional con las matrices generadores de un códigos bloque cíclico o quasi-cíclico [11, 13, 26]. Otras construcciones están basadas en técnicas algebraico-geométricas (véase por ejemplo [5]). Sin embargo, en este trabajo consideramos los códigos convolucionales desde otra perspectiva: desde el punto de vista de sistemas lineales.

Desde finales de los años sesenta se sabe que los códigos convolucionales y los sistemas discretos e invariantes en el tiempo con coeficientes en un cuerpo finito son el mismo objeto [14]. Consideraremos la representación entrada-estado-salida introducida por Rosenthal, Schumacher y York [20], muy empleada en los últimos años en el análisis y construcción de códigos convolucionales (véase por ejemplo [10, 19, 20, 21, 23, 25]). Sean $A \in \mathbb{F}^{\delta\times\delta}$, $B \in \mathbb{F}^{\delta\times k}$, $C \in \mathbb{F}^{(n-k)\times\delta}$ y $D \in \mathbb{F}^{(n-k)\times k}$. Un código convolucional $\mathcal{C}$ de tasa k/n y grado δ puede ser descrito por el sistema lineal gobernado por las ecuaciones

$$\begin{aligned} x_{t+1} &= Ax_t + Bu_t, \\ y_t &= Cx_t + Du_t, \\ v_t &= \begin{pmatrix} y_t \\ u_t \end{pmatrix}, \quad x_0 = 0, \end{aligned} \tag{1}$$

siendo $x_t \in \mathbb{F}^\delta$ el **vector de estados**, $u_t \in \mathbb{F}^k$ el **vector información**, $y_t \in \mathbb{F}^{n-k}$ el **vector de paridad** y $v_t \in \mathbb{F}^n$ **el vector código** o **palabra código**. Esta representación, como hemos mencionado anteriormente, es conocida como la **representación entrada-estado-salida**. Precisamente, el grado δ del código convolucional coincide con el grado de McMillan del sistema lineal (1), que es a su vez igual a la dimensión del espacio de estados $\mathbb{F}^\delta$. Por tanto, cuando $A \in \mathbb{F}^{\delta\times\delta}$, con δ el grado de McMillan, tenemos una representación minimal, caracterizada a través de la condición de que el par (A, B) es controlable, es decir, $\operatorname{rg}\begin{pmatrix} B & AB & \cdots & A^{\delta-2}B & A^{\delta-1}B \end{pmatrix} = \delta$.

Rosenthal y York [23] demuestran que, partiendo de una representación minimal de un código convolucional, dicho código es no catastrófico si y sólo si el par (A, C) es observable, es decir, si $\operatorname{rg}\begin{pmatrix} C & CA & CA^2 & \cdots & CA^{\delta-1} \end{pmatrix}^t = \delta$.

Uno de los parámetros más importantes de un código convolucional es su distancia libre, pues determina la capacidad detectora y correctora de errores. Tenemos la caracterización siguiente de la distancia libre en términos de la

representación entrada-estado-salida,

$$d_{free}(\mathcal{C}) = \text{mín}\left\{\sum_{t=0}^{\infty} \text{wt}(u_t) + \sum_{t=0}^{\infty} \text{wt}(y_t)\right\}, \tag{2}$$

en donde el mínimo se considera sobre todas las palabras código no nulas.

Debido a razones algebraicas, asumimos a lo largo del trabajo que las palabras código son de peso finito. El conjunto de palabras de peso finito tiene una estructura de módulo sobre el anillo de polinomios $\mathbb{F}[z]$ (véase [23]). Haciendo un abuso de notación, denotamos este módulo por $\mathcal{C}(A, B, C, D)$, que denominamos **código convolucional de peso finito** generado por las matrices A, B, C, D.

A lo largo de este trabajo adoptamos la notación introducida por McElliece [15] y nos referiremos a un código convolucional de tasa k/n y grado δ como un (n, k, δ)-código. Observemos que un código bloque es, en particular, un código convolucional con grado $\delta = 0$.

Al igual que en la teoría de códigos bloque, uno de los objetivos de la teoría de códigos convolucionales es construir códigos con la mayor distancia libre posible. Rosenthal y Smarandache [22] generalizaron la cota Singleton para códigos convolucionales. En concreto, dado un (n, k, δ)-código $\mathcal{C}$ sobre un cuerpo cualquiera $\mathbb{F}$, entonces

$$d_{free}(\mathcal{C}) \leq (n-k)\left(\left\lfloor \frac{\delta}{k} \right\rfloor + 1\right) + \delta + 1.$$

Dicha cota se conoce como **cota Singleton generalizada**. Decimos que un (n, k, δ)-código es **MDS** (*Maximum Distance Separable*) si su distancia libre alcanza la cota Singleton generalizada. Notemos que un (n, k, δ)-código MDS es, en particular, un código compacto, es decir, tiene $\delta \mod k$ índices de controlabilidad iguales a $\lceil \frac{\delta}{k} \rceil$ y $k - (\delta \mod k)$ índices de controlabilidad iguales a $\lfloor \frac{\delta}{k} \rfloor$.

La teoría desarrollada por Shannon [24], permite asegurar que si la longitud de un código es suficientemente grande, entonces es bueno. Ahora bien, la complejidad de decodificación aumenta con la longitud del código. Forney [6], en su búsqueda por encontrar una clase de códigos cuya probabilidad de error decreciera exponencialmente con la longitud del código mientras que la complejidad de decodificación aumentara sólo linealmente, llegó a una solución consistente en una estructura de código multinivel, conocida como código concatenado. En este trabajo, empleando la teoría de Forney y teniendo en cuenta que los sistemas multivariables pueden considerarse compuestos por varios subsistemas, de manera que las salidas de unos actúan como entradas de otros, introducimos cuatro modelos de concatenación en serie de códigos convolucionales. También en la sección siguiente establecemos condiciones necesarias y suficientes para que los códigos obtenidos sean controlables y observables, en términos de las propiedades de los códigos constituyentes, introduciendo finalmente cotas inferiores de la distancia libre. Finalmente, en la última sección presentamos las conclusiones y líneas futuras.

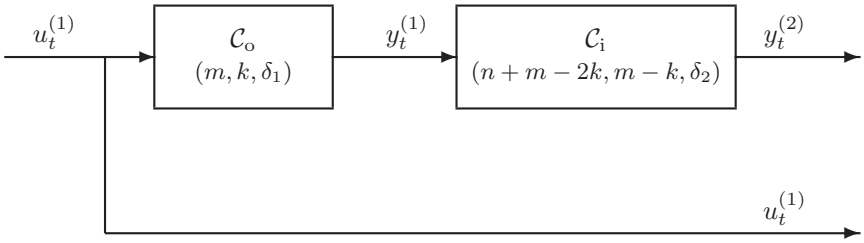


Figura 1: Código concatenado $\mathcal{SC}^{(1)}$

2 Concatenación en serie de códigos convolucionales

Los códigos concatenados se han empleado en aplicaciones en el espacio, realzado de datos en GSM (EDGE, en inglés, *Enhanced Data rates for GSM Evolution*) [17], sistemas de comunicación inalámbricos [12], por citar sólo algunos ejemplos.

Ahora bien, los códigos concatenados siempre han sido estudiados a partir de la matriz generadora. En esta sección, como ya hemos comentado en la introducción, estudiamos y caracterizamos diferentes tipos de conacatenación en serie de códigos convolucionales. En dicha concatenación, los códigos se ordenan en serie uno tras otro. En el caso de dos códigos constituyentes se habla de código externo y código interno. Denotamos por $\mathcal{C}_{\rm o}$ el **código externo** y por $\mathcal{C}_{\rm i}$ el **código interno**. También, denotamos por $x_t^{(1)}$, $u_t^{(1)}$, $y_t^{(1)}$ y $v_t^{(1)}$ el vector de estados, el vector información, el vector de paridad y la palabra código de $\mathcal{C}_{\rm o}$, respectivamente; del mismo modo, denotamos por $x_t^{(2)}$, $u_t^{(2)}$, $y_t^{(2)}$ y $v_t^{(2)}$ el vector de estados, el vector información, el vector de paridad y la palabra código de $\mathcal{C}_{\rm i}$, respectivamente. En los diferentes modelos de concatenación en serie, las palabras código $v_t^{(1)}$ y $v_t^{(2)}$ de $\mathcal{C}_{\rm o}$ y $\mathcal{C}_{\rm i}$, respectivamente, vienen dadas por las expresiones

$$v_t^{(1)} = \begin{pmatrix} y_t^{(1)} \\ u_t^{(1)} \end{pmatrix} \qquad \text{y} \qquad v_t^{(2)} = \begin{pmatrix} y_t^{(2)} \\ u_t^{(2)} \end{pmatrix}. \tag{3}$$

Análogamente, x_t, u_t, y_t y v_t serán el vector de estados, el vector información, el vector de paridad y la palabra código, respectivamente, del correspondiente modelo de código concatenado que estemos tratando.

Una representación entrada-estado-salida de los cuatro modelos presentados en las Figuras 1–4 son las dadas en la tabla 1.

Una vez tenemos una representación del nuevo código obtenido al concatenar dos códigos convolucionales siguiendo alguno de los cuatro modelos introducidos en esta sección, nuestro objetivo es establecer condiciones que nos aseguren que dichas representaciones son minimales y observables. Tal y como comentamos en la sección anterior, el concepto de controlabilidad del par (A, B) está relacionado con el de minimalidad de una representación entrada-estado-salida (A, B, C, D).

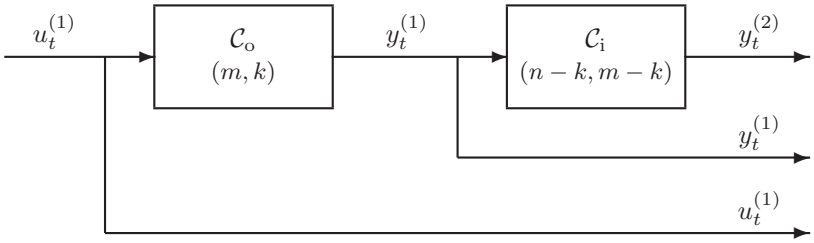


Figura 2: Código concatenado $\mathcal{SC}^{(2)}$

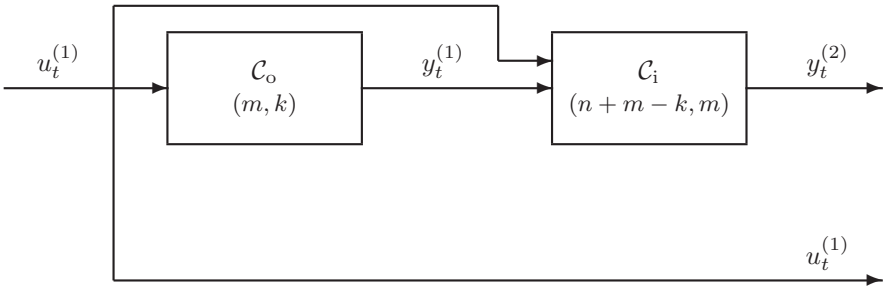


Figura 3: Código concatenado $\mathcal{SC}^{(3)}$

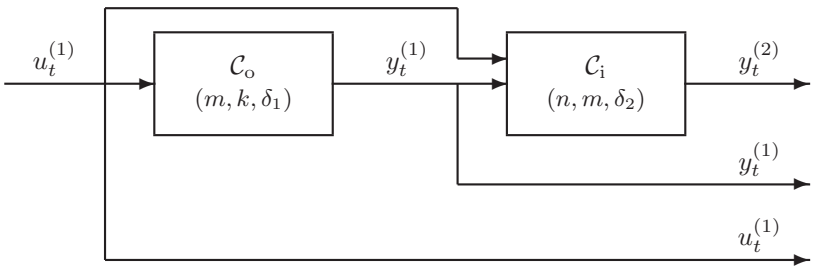


Figura 4: Código Concatenado $\mathcal{SC}^{(4)}$

Modelo	Representación entrada-estado-salida
$\mathcal{SC}^{(1)}$	$A = \begin{pmatrix} A_2 & B_2C_1 \\ O & A_1 \end{pmatrix}$, $B = \begin{pmatrix} B_2D_1 \\ B_1 \end{pmatrix}$, $C = \begin{pmatrix} C_2 & D_2C_1 \end{pmatrix}$ $D = D_2D_1$.
$\mathcal{SC}^{(2)}$	$A = \begin{pmatrix} A_2 & B_2C_1 \\ O & A_1 \end{pmatrix}$, $B = \begin{pmatrix} B_2D_1 \\ B_1 \end{pmatrix}$, $C = \begin{pmatrix} C_2 & D_2C_1 \\ O & C_1 \end{pmatrix}$, $D = \begin{pmatrix} D_2D_1 \\ D_1 \end{pmatrix}$.
$\mathcal{SC}^{(3)}$	$A = \begin{pmatrix} A_2 & B_{21}C_1 \\ O & A_1 \end{pmatrix}$, $B = \begin{pmatrix} B_{21}D_1 + B_{22} \\ B_1 \end{pmatrix}$, $C = \begin{pmatrix} C_2 & D_{21}C_1 \end{pmatrix}$, $D = D_{21}D_1 + D_{22}$
$\mathcal{SC}^{(4)}$	$A = \begin{pmatrix} A_2 & B_{21}C_1 \\ O & A_1 \end{pmatrix}$, $B = \begin{pmatrix} B_{21}D_1 + B_{22} \\ B_1 \end{pmatrix}$, $C = \begin{pmatrix} C_2 & D_{21}C_1 \\ O & C_1 \end{pmatrix}$, $D = \begin{pmatrix} D_{21}D_1 + D_{22} \\ D_1 \end{pmatrix}$,

Cuadro 1: Representación entrada-estado-salida de los modelos de concatenación

Así pues, a continuación analizamos las condiciones que deben cumplir las matrices A_l, B_l, C_l y D_l, para $l = 1, 2$, de los códigos constituyentes para que el código concatenado tenga una representación minimal y observable. Dado que las matrices A y B de los códigos $\mathcal{SC}^{(1)}$ y $\mathcal{SC}^{(2)}$ tienen la misma expresión, los resultados que presentamos, serán válidos para ambos tipos de concatenación. Por tanto, a lo largo de esta sección, $\mathcal{SCC}_{\text{sys}}$ denota el código concatenado $\mathcal{SC}^{(1)}$ así como el código $\mathcal{SC}^{(2)}$. Siguiendo un razonamiento análogo, denotamos por $\mathcal{SCC}$ a los códigos concatenados $\mathcal{SC}^{(3)}$ y $\mathcal{SC}^{(4)}$.

El ejemplo siguiente pone de manifiesto que no es suficiente con que los pares (A_l, B_l), para $l = 1, 2$, de los códigos constituyentes sean controlables, para que el par (A, B) del código concatenado $\mathcal{SCC}_{\text{sys}}$ sea controlable.

Ejemplo 1 *Sea α un elemento primitivo de $\mathbb{F} = GF(8)$ y consideremos el $(6, 4, 2)$-código externo $\mathcal{C}_{\text{o}}(A_1, B_1, C_1, D_1)$, donde*

$$A_1 = \begin{pmatrix} \alpha^3 & \alpha \\ \alpha & \alpha^4 \end{pmatrix}, \qquad B_1 = \begin{pmatrix} 1 & \alpha & \alpha^3 & \alpha^4 \\ \alpha^2 & \alpha^3 & \alpha^4 & \alpha^5 \end{pmatrix}$$

$$C_1 = \begin{pmatrix} 1 & \alpha \\ \alpha & 1 \end{pmatrix} \qquad y \qquad D_1 = \begin{pmatrix} \alpha^3 & \alpha^4 & \alpha^3 & \alpha^4 \\ 1 & 1 & 1 & 1 \end{pmatrix}.$$

Observemos que (A_1, B_1) es controlable.

Sea $\mathcal{C}_{\rm i}(A_2, B_2, C_2, D_2)$ *el* $(5, 2, 1)$*-código interno descrito por las matrices*

$$A_2 = \begin{pmatrix}\alpha\end{pmatrix}, \quad B_2 = \begin{pmatrix}\alpha^3 & 0\end{pmatrix}, \quad C_2 = \begin{pmatrix}0\\1\\0\end{pmatrix} \quad y \quad D_2 = \begin{pmatrix}1 & 0\\0 & 1\\\alpha & 0\end{pmatrix}.$$

Claramente, el par (A_2, B_2) *es controlable.*

Por otro lado, teniendo en cuenta la tabla 1, las matrices A *y* B *de una representación entrada-estado-salida de los código concatenados* $\mathcal{SC}^{(1)}$ *y* $\mathcal{SC}^{(2)}$ *vienen dadas por*

$$A = \begin{pmatrix}\alpha & \alpha^3 & \alpha^4\\0 & \alpha^3 & \alpha\\0 & \alpha & \alpha^4\end{pmatrix} \qquad y \qquad B = \begin{pmatrix}\alpha^6 & 1 & \alpha^6 & 1\\1 & \alpha & \alpha^3 & \alpha^4\\\alpha^2 & \alpha^3 & \alpha^4 & \alpha^5\end{pmatrix}.$$

Pero el par (A, B) *no es controlable, ya que*

$$\operatorname{rg}\begin{pmatrix}\alpha I - A & B\end{pmatrix} = \operatorname{rg}\begin{pmatrix}0 & \alpha^3 & \alpha^4 & \alpha^6 & 1 & \alpha^6 & 1\\0 & 1 & \alpha & 1 & \alpha & \alpha^3 & \alpha^4\\0 & \alpha & \alpha^2 & \alpha^2 & \alpha^3 & \alpha^4 & \alpha^5\end{pmatrix} = 2 \neq 3.$$

Análogamente, podemos encontrar ejemplos que ponen de manifiesto que el hecho de que los códigos constituyentes tengan una representación minimal, no implica que las representaciones entrada-estado-salida de los códigos concatenados $\mathcal{SC}^{(3)}$ y $\mathcal{SC}^{(4)}$ dadas por la tabla 1 también lo sean. Así pues, necesitamos más condiciones sobre las matrices que describen los códigos externo e interno para obtener la controlabilidad del par (A, B) del código convolucional concatenado.

El resultado siguiente introduce condiciones que aseguran dicha controlabilidad para los códigos $\mathcal{SC}^{(1)}$ y $\mathcal{SC}^{(2)}$.

Teorema 1 *Sean* $\mathcal{C}_{\rm o}(A_1, B_1, C_1, D_1)$ *un* (m, k, δ_1)*-código,* $\mathcal{C}_{\rm i}(A_2, B_2, C_2, D_2)$ *un* $(n, m - k, \delta_2)$*-código y* $\mathcal{SCC}_{sys}(A, B, C, D)$ *el código concatenado correspondiente. Si* $\operatorname{rg}(B) = \delta_1 + \delta_2$*, entonces* (A, B, C, D) *es una representación minimal de* $\mathcal{SCC}_{sys}(A, B, C, D)$ *con complejidad* $\delta_1 + \delta_2$*.*

Para el caso de los códigos $\mathcal{SC}^{(3)}$ y $\mathcal{SC}^{(4)}$, tenemos el resultado siguiente.

Teorema 2 *Sean* $\mathcal{C}_{\rm o}(A_1, B_1, C_1, D_1)$ *un* (m, k, δ_1)*-código,* $\mathcal{C}_{\rm i}(A_2, B_2, C_2, D_2)$ *un* (n, m, δ_2)*-código y* $\mathcal{SCC}(A, B, C, D)$ *el código concatenado correspondiente.*

Si $\operatorname{rg}(B) = \delta_1 + \delta_2$*, entonces* (A, B, C, D) *es una representación minimal de* $\mathcal{SCC}(A, B, C, D)$ *con complejidad* $\delta_1 + \delta_2$*.*

Podemos encontrar otros resultados para cada uno de los modelos basados en casos particulares de la tasa y la complejidad de los códigos constituyentes en [3].

De forma análoga, podemos encontrar condiciones para que la representación entrada-estado-salida de cada uno de los modelos sea observable. En concreto, para el modelo $\mathcal{SC}^{(1)}$ tenemos el resultado siguiente.

Teorema 3 *Sean* $\mathcal{C}_{\mathrm{o}}(A_1, B_1, C_1, D_1)$ *un* (m, k, δ_1)*-código,* $\mathcal{C}_{\mathrm{i}}(A_2, B_2, C_2, D_2)$ *un* $(n, m-k, \delta_2)$*-código y* $\mathcal{SC}^{(1)}(A, B, C, D)$ *el código concatenado de tasa* $k/(n-m+2k)$ *descrito por las matrices dadas por la tabla 1. Si* $\mathrm{rg}(C) = \delta_1 + \delta_2$*, entonces el par* (A, C) *es observable.*

Una vez tenemos asegurada una representación entrada-estado-salida minimal y observable del nuevo código, nuestro siguiente objetivo es establecer cotas de la distancia libre. Para el modelo $\mathcal{SC}^{(2)}$ obtenemos las cotas siguientes.

Teorema 4 *Sea* $\mathcal{SC}^{(2)}$ *el código descrito por las matrices dadas por la tabla 1, obtenido al concatenar el código externo* $\mathcal{C}_{\mathrm{o}}$ *y el código interno* $\mathcal{C}_{\mathrm{i}}$*. Entonces*

1. $d_{free}(\mathcal{SC}^{(2)}) \geq d_{free}(\mathcal{C}_{\mathrm{o}})$.
2. *Si* $\mathrm{rg}(D_1) = k$*, entonces* $d_{free}(\mathcal{SC}^{(2)}) \geq d_{free}(\mathcal{C}_{\mathrm{i}}) + 1$.

Finalmente, para el modelo $\mathcal{SC}^{(4)}$ obtenemos las siguientes cotas de la distancia libre.

Teorema 5 *Sea* $\mathcal{SC}^{(4)}$ *el código convolucional descrito por las matrices dadas por la tabla 1, obtenido al concatenar el código externo* $\mathcal{C}_{\mathrm{o}}$ *y del código interno* $\mathcal{C}_{\mathrm{i}}$*. Entonces,*

1. $d_{free}(\mathcal{SC}^{(4)}) \geq \max\{d_{free}(\mathcal{C}_{\mathrm{o}}), d_{free}(\mathcal{C}_{\mathrm{i}})\}$.
2. *Si además* $\mathrm{rg}(D_2) = m$*, entonces* $d_{free}(\mathcal{SC}^{(4)}) \geq \max\{d_{free}(\mathcal{C}_{\mathrm{o}}) + 1, d_{free}(\mathcal{C}_{\mathrm{i}})\}$.

3 Conclusiones y líneas futuras

En este trabajo hemos visto que la concatenación en serie de códigos convolucionales da lugar a nuevos códigos convolucionales de parámetros prescritos. Además, en [3] y en [8] se pueden ver ejemplos que ponen de manifiesto que los códigos obtenidos tienen distancia libre muy próxima a la cota Singleton e incluso pueden alcanzarla en algunas ocasiones. Por tanto, un trabajo futuro consiste en mejorar las cotas de las distancias libres de cada uno de los modelos. Por otra parte, en [4] podemos encontrar la modelización desde el punto de vista de sistemas de la concatenación de un código bloque con un código convolucional, obteniendo incluso condiciones para que el código obtenido sea óptimo. Asímismo, en [8] podemos ver una primera aproximación a la modelización matemática mediante sistemas lineales de concatenación en paralelo. Estos modelos son una primera aproximación a la modelización entrada-estado-salida de los conocidos turbo códigos. Por consiguiente, otra línea futura de trabajo será el refinamiento de la modelización matemática desde el punto de vista de sistemas de los turbo códigos, así como el uso de la representación entrada-estado-salida de los modelos concatenados para introducir algoritmos algebraicos de decodificación. Por otra parte, es importante resaltar que a partir de la representación entrada-estado-salida de un código convolucional se podrían plantear otro tipo de construcciones además

de las aquí presentadas y de las introducidas por [23], utilizando resultados de teoría de control como la completación por columnas de pares de matrices [1] o bien empleando matrices con filas prescritas y factores invariantes [27].

Referencias

[1] I. Baragaña y I. Zaballa. Column completation of a pair of matrices *Linear and Multilinear Algebra*, vol. 27, pp. 243-273, 1990.

[2] R. S. Chernock, R. J. Crinon, M. A. Dolan y J. R. Mick. *Data Broadcasting: Understanding the Atsc Data Broadcast Standard*. Springer, Berlin, 1994.

[3] J. J. Climent, V. Herranz y C. Perea. A first approximation of concatenated convolutional codes from linear systems theory viewpoint. *Linear Algebra and its Applications*. vol. 425, pp. 673–699. 2007.

[4] J. J. Climent, V. Herranz y C. Perea. Linear System Modelization of Concatenated Block and Convolutional Codes. *Linear Algebra and its Applications*. To appear.

[5] J. A. Domínguez, J. M. Muñoz, J. I. Iglesias y G. Serrano. Convolutional goppa codes. *IEEE Transactions on Information Theory*, vol. 52, pp. 340–344. 2006.

[6] G. D. Forney, Jr. *Concatenated Codes*. MA: MIT Press, Cambridge, 1966.

[7] M. Hatori y T. Shiomi. *Digital Broadcasting*. Ohmsha, Ltd, Tokyo, 2000.

[8] V. Herranz. *Estudio y construcción de códigos convolucionales: códigos perforados, códigos concatenados desde el punto de vista de sistemas*. Tesis Doctoral, Universidad Miguel Hernández, 2007.

[9] W. C. Huffman y V. Pless. *Fundamentals of Error-Correcting Codes*. Cambridge University Press, 2003.

[10] R. Hutchinson, J. Rosenthal y R. Smarandache. Convolutional codes with maximum distance profile. *Systems and Control Letters*, vol. 54, no. 1, pp. 53–63. 2005.

[11] J. Justesen. New convolutional code constructions and a class of asymptotically good time-varying codes. *IEEE Transactions on Information Theory*, vol. 19, no. 2, pp. 220–225. 1973.

[12] W.G. Kim, B.J. Ku, L.H. Baek, H.Y. Yang y C.Kang. Serially concatenated space-time code (SCSTC) for high rate wireless communication systems. *Electronics Letters*, vol. 36, no. 7, pp. 646–648. 2000.

[13] Y. Levy y D. J. Costello, Jr. An algebraic approach to constructing convolutional codes from quasicyclic codes. *DIMACS Ser. in Discr. Math and Theor. Comp. Sci.*, vol. 14, pp. 189–198. 1993.

[14] J. L. Massey y M. Sain. Codes, automata, and continuous systems: explicit interconnections. *IEEE Transactions on Automatic Control*, vol. 12, no. 6, pp. 644–650. 1967.

[15] R. J. McEliece. The algebraic theory of convolutional codes. En V. Pless y W. Huffman (editores). *Handbook of Coding Theory*, pp. 1065–1138. Elsevier, 1998.

[16] McWilliams, F. and Sloane, N. *The Theory of Error-Correcting Codes* . North Holand, Amsterdam, 1977.

[17] N. Nefedov. Application of low complexity serially concatenated codes for edge circuit switched data. *Proceedings of the IEEE International Symposium on Personal, Indoor and Mobile Communications (PIMRC'99)*, pp. 573–577. IEEE, Osaka, Japan, 1999.

[18] P. Piret. *Convolutional Codes, an Algebraic Approach*. MIT Press, 1988.

[19] J. Rosenthal. Connections between linear systems and convolutional codes. En B. Marcus y J. Rosenthal (editores). *Codes, Systems and Graphical Models*, vol. 123 de *IMA*, pp. 39–66. Springer-Verlag, Berlin, 2001.

[20] J. Rosenthal, J. Schumacher y E. V. York. On behaviors and convolutional codes. *IEEE Transactions on Information Theory*, vol. 42, no. 6, pp. 1881–1891. 1996.

[21] J. Rosenthal y R. Smarandache. Construction of convolutional codes using methods from linear systems theory. *Proccedings of the 35th Allerton Conference on Communications, Control and Computing*, pp. 953–960. Allerton House, Monticello, IL, septiembre 1997.

[22] J. Rosenthal y R. Smarandache. Maximum distance separable convolutional codes. *Applicable Algebra in Engineering, Communication and Computing*, vol. 10, pp. 15–32. 1999.

[23] J. Rosenthal y E. V. York. BCH convolutional codes. *IEEE Transactions on Information Theory*, vol. 45, no. 6, pp. 1833–1844. 1999.

[24] C. E. Shannon. A mathematical theory of communication. *The Bell System Technical Journal*, vol. 27, pp. 379–423 y 623–656. 1948.

[25] R. Smarandache y J. Rosenthal. A state space approach for constructing MDS rate $1/n$ convolutional codes. *Proceedings of the 1998 IEEE Information Theory Workshop on Information Theory*, pp. 116–117. Killarney, Kerry, Ireland, Junio 1998.

[26] R. M. Tanner. Convolutional codes from quasicyclic codes: A link between the theories of block and convolutional codes. *Informe Técnico*. Univ. California, Santa Cruz, CA, USC-CRL-87-21, noviembre. 1987.

[27] I. Zaballa Matrices with prescribed rows and invariants factors *Linear Algebra and its Applictions*, vol. 87, pp. 113-146, 1987.

Bol. Soc. Esp. Mat. Apl.
nº 42(2008), 183–193

CODE DECOMPOSITION IN THE ANALYSIS OF A CONVOLUTIONAL CODE

E. FORNASINI*, R. PINTO†

*Department of Information Engineering, University of Padua, 35131 Padova, ITALY. E-mail: fornasini@dei.unipd.it.
†Department of Mathematics, University of Aveiro, 3810-193 Aveiro, PORTUGAL. E-mail: raquel@ua.pt.

Abstract

A convolutional code can be decomposed into smaller codes if it admits decoupled encoders. In this paper, we show that if a code can be decomposed into smaller codes (subcodes) its column distances are the minimum of the column distances of its subcodes. Moreover, the j-th column distance of a convolutional code $\mathcal{C}$ is equal to the j-th column distance of the convolutional codes generated by the truncation of the canonical encoders of $\mathcal{C}$ to matrices which entries have degree smaller or equal than j. We show that if one of such codes can be decomposed into smaller codes, so can be all the other codes.

Key words: *Convolutional codes, decoupled encoders, code decomposition, free distance, column distance*
AMS subject classifications:

1 Introduction

Some convolutional codes can be decomposed into smaller codes (subcodes). This happens if they admit decoupled encoders among its encoders [2]. The free distance and the column distances are the most common distance measures for a convolutional code. It was shown in [1] that the free distance of a code is equal to the minimum of the free distances of its subcodes. In this paper, we will show that similarly to the free distance, the j-th column distance of a convolutional code $\mathcal{C}$ is equal to the minimum of the j-th column distances of its subcodes. Moreover, to calculate the j-th column distance of $\mathcal{C}$ we can consider the truncation of the entries of a canonical encoder of $\mathcal{C}$ to polynomials of degree smaller or equal than j (truncation to degree j). Such obtained matrix generates a different convolutional code with same i-th column distances than $\mathcal{C}$, for $i \leq j$. So, if $\mathcal{C}$ admits a canonical encoder which truncation to degree j is

† Supported in part by the Portuguese Science Foundation (FCT) through the Unidade de Investigação Matemática e Aplicações of the University of Aveiro, Portugal.

a decoupled encoder, $G_d(d)$, we have that the j-th column distance of $\mathcal{C}$ is equal to the minimum of the column distances of the subcodes of the convolutional code generated by $G_d(d)$. We will see that if a convolutional code has such a canonical encoder, all the convolutional codes generated by the truncation to degree j of canonical encoders of $\mathcal{C}$ also admit decoupled encoders.

2 Convolutional codes

We will consider convolutional codes constituted by left compact sequences of $(\mathbb{F}^p)^{\mathbb{Z}}$, where $p \in \mathbb{N}$ and $\mathbb{F}$ is a finite field. Such sequences are naturally represented by Laurent power series $\hat{\mathbf{w}}(d) = \sum \mathbf{w}_t d^t \in \mathbb{F}((d))^p$, and we are allowed to multiply any left compact support sequence by a scalar Laurent series $s(d) = \sum s_t d^t \in \mathbb{F}((d))$. In fact, $\mathbb{F}((d))^p$ is a vector space over the field $\mathbb{F}((d))$. $\mathbb{F}[d]$ and $\mathbb{F}(d)$ will denote, as usually, the ring of polynomials and the field of rational functions with coefficients in $\mathbb{F}$, respectively.

A $[p, m]$-*convolutional code* is an m-dimensional subspace of $\mathbb{F}((d))^p$, which has a rational (and polynomial) basis, i.e., that is generated by a full row rank rational matrix $G(d) \in \mathbb{F}(d)^{m \times p}$,

$$\mathcal{C} = Im\, G(d) = \{\hat{\mathbf{w}}(d) : \hat{\mathbf{w}}(d) = \hat{\mathbf{u}}(d)G(d),\ \hat{\mathbf{u}}(d) \in \mathbb{F}((d))^m\}.$$

$\mathcal{C}$ is said to have *rate* $\frac{m}{p}$ and $G(d)$ is called an *encoder* of $\mathcal{C}$. $G(d)$ produces a codeword $\hat{\mathbf{w}}(d) = \hat{\mathbf{u}}(d)G(d)$ corresponding to each information sequence $\hat{\mathbf{u}}(d) \in \mathbb{F}((d))^m$. A convolutional code admits many encoders. Two encoders that generate the same code are called *equivalent encoders* and are related by a nonsingular rational left factor in $\mathbb{F}(d)^{m \times m}$. The encoders that can be realized by a physical device are called *causal encoders*. A causal encoder induces a "non-anticipatory" input-output map, i.e., produces codewords that start at the same time or after the corresponding information sequences. Among the causal encoders of a convolutional code we have the polynomial encoders and in the class of the polynomial encoders we distinguish the *canonical encoders* which are the left prime and row reduced ones. Two canonical encoders of a convolutional code $\mathcal{C}$ have the same row degrees $\phi_i,\ i = 1, \ldots, m$, up to a row permutation, and these row degrees are called *Forney indices* of $\mathcal{C}$. The maximum of the Forney indices is the *memory* of $\mathcal{C}$ and is represented by ν [3, 5].

The *free distance* of a convolutional code $\mathcal{C}$ [5] is defined as

$$d_{free}(\mathcal{C}) := \min\{w_H(\hat{\mathbf{w}}(d)) = \sum w_H(\mathbf{w}_t) : \hat{\mathbf{w}}(d) = \sum \mathbf{w}_t d^t \in \mathcal{C} \backslash \{0\}\},$$

where $w_H(\mathbf{w}_t)$ represents the Hamming weight of $\mathbf{w}_t$, and is bounded by

$$d_{free}(\mathcal{C}) \leq (p - m)(\lfloor \frac{\nu}{m} \rfloor + 1) + \nu + 1.$$

Such a bound is called the *generalized Singleton bound*. If the free distance of $\mathcal{C}$ is equal to the corresponding generalized Singleton bound, then $\mathcal{C}$ is said to be an *MDS-code* [6].

A distance measure associated to each causal encoder of a convolutional code $\mathcal{C}$ is the *column distance* [5].

Definition 2.1 *The j-th order column distance of a causal encoder $G(d) \in \mathbb{F}(d)^{m\times p}$ is the minimum of the Hamming weights of $\hat{\mathbf{w}}(d)|_{[0,j]}$ [1] where $\hat{\mathbf{w}}(d)$ is a codeword corresponding to an information sequence $\hat{\mathbf{u}}(d) = \sum_{t\geq 0} \mathbf{u}_t d^t$ such that $\mathbf{u}_0 \neq 0$.*

The column distance is an encoder property and two equivalent causal encoders can have different column distances. However the causal encoders of a convolutional code which are delay-free (i.e., that produce codewords that start at the same time as the corresponding information sequences) have the same column distances, which leads to the definition of column distance of the code. In this definition we only refer to polynomial encoders for simplicity. A polynomial encoder $G(d)$ is delay-free if and only if $G(0)$ has full row rank. Observe that a convolutional code always admit such encoders, being the canonical encoders an example of delay-free polynomial encoders.

Definition 2.2 *The j-th order column distance of a convolutional code is the j-th order column distance of any polynomial encoder $G(d)$ of $\mathcal{C}$ such that $G(0)$ is full row rank.*

Let $G(d) = G_0 + G_1 d + \cdots + G_\ell d^\ell$, $G_i \in \mathbb{F}^{m\times p}$, $i = 1, \ldots, \ell$, be a polynomial encoder of degree ℓ [2] of the convolutional code $\mathcal{C}$, with $G(0) = G_0$ full row rank, and

$$\mathbf{M}(G(d)) = \begin{bmatrix} G_0 & G_1 & \cdots & \cdots & G_\ell & & \\ & G_0 & G_1 & \cdots & \cdots & G_\ell & \\ & & \ddots & \ddots & & & \ddots \end{bmatrix}$$

the corresponding semi-infinite matrix. Denote by $\mathbf{M}_j^c(G(d))$ the truncation of $\mathbf{M}(G(d))$ after $j+1$ (block) columns

$$\mathbf{M}_j^c(G(d)) = \begin{bmatrix} G_0 & G_1 & G_2 & \cdots & G_j \\ & G_0 & G_1 & \cdots & G_{j-1} \\ & & G_0 & & G_{j-2} \\ & & & \ddots & \vdots \\ & & & & G_0 \end{bmatrix} \tag{1}$$

where $G_i = 0$ for $i > \ell$. Then the j-th order column distance of $\mathcal{C}$ is

$$d_j^c(\mathcal{C}) = \min_{\mathbf{u}_0 \neq 0} \{ w_H([\mathbf{u}_0\, \mathbf{u}_1 \ldots \mathbf{u}_j]\mathbf{M}_j^c(G(d))) \},$$

[1] If $\hat{\mathbf{w}}(d) = \sum_{t\geq k} \mathbf{w}_t d^t$ then $\hat{\mathbf{w}}(d)|_{[0,j]} = \sum_{t=0}^{j} \mathbf{w}_t d^t$ where $\mathbf{w}_t = 0$ for $t < k$, if $k > 0$.

[2] We consider the degree of a polynomial matrix as the maximum of the degrees of its entries. If $G(d)$ is an $m \times p$ polynomial matrix of degree ℓ, we can write $G(d) = G_0 + G_1 d + \cdots + G_\ell d^\ell$, with $G_i \in \mathbb{F}^{m\times p}$, $i = 1, \ldots, \ell$, and $G_\ell \neq 0$.

with $\mathbf{u}_i \in \mathbb{F}^m$, $i = 0, 1, \ldots, j$, for all $j \in \mathbb{N}$ [5]. For every $j \geq 0$, the j-th column distance of a $[p, m]$-convolutional code $\mathcal{C}$ is bounded by [4]

$$d_j^c(\mathcal{C}) \leq (p - m)(j + 1) + 1.$$

3 Decoupled encoders and code decomposition

A convolutional code is decomposable into smaller codes if it admits encoders in block diagonal form, called *decoupled encoders*. In this section we present a brief introduction to such encoders. For more details see [2].

Definition 3.1 *Let $p_1, \ldots, p_k$ be positive integers such that $\sum_{i=1}^k p_i = p$ and P a permutation matrix. An encoder $G(d)$ of $\mathcal{C}$ is said to be a $(p_1, \ldots, p_k)$-decoupled encoder of $\mathcal{C}$ associated with P if there exist positive integers $m_1, \ldots, m_k$ with $\sum_{i=1}^k m_i = m$ such that*

$$G(d)P = \operatorname{diag}\{G^{(1)}(d), \ldots, G^{(k)}(d)\}, \tag{2}$$

with $G^{(i)}(d) \in \mathbb{F}(d)^{m_i \times p_i}$, $i = 1, \ldots, k$.

If $G(d)$ is a $(p_1, \ldots, p_k)$-decoupled encoder that satisfies (2) and $\hat{\mathbf{u}}(d) = [\hat{\mathbf{u}}_1(d) \cdots \hat{\mathbf{u}}_k(d)] \in \mathbb{F}((d))^m$, with $\hat{\mathbf{u}}_i(d) \in \mathbb{F}((d))^{m_i}$, an information sequence, then its corresponding codeword $\hat{\mathbf{w}}(d) = \hat{\mathbf{u}}(d)G(d)$ is of the form

$$\hat{\mathbf{w}}(d) = [\hat{\mathbf{w}}_1(d) \cdots \hat{\mathbf{w}}_k(d)]P,$$

where $\hat{\mathbf{w}}_i(d) = \hat{\mathbf{u}}_i(d)G^{(i)}(d)$, $i = 1, \ldots, k$. Consequently, up to a permutation of the components of the codewords of $\mathcal{C}$,

$$\mathcal{C} = \mathcal{C}^{(1)} \times \cdots \times \mathcal{C}^{(k)},$$

where $\mathcal{C}^{(i)}$ is the $[p_i, m_i]$-convolutional code generated by $G^{(i)}(d)$, $i = 1, \ldots, k$, and we say that $\mathcal{C}$ is *decomposable* into $\mathcal{C}^{(1)}, \ldots, \mathcal{C}^{(k)}$. If $\mathcal{C}$ does not have a (p_1, p_2)-decoupled encoder, for some $p_1, p_2 \in \mathbb{N}$, then $\mathcal{C}$ is said to be an *undecomposable code*.

Definition 3.2 *A $(p_1, \ldots, p_k)$-decoupled encoder $G(d)$ of $\mathcal{C}$ associated with a permutation matrix P,*

$$G(d) = \operatorname{diag}\{G^{(1)}(d), \ldots, G^{(k)}(d)\}P^{-1},$$

with $G^{(i)}(d) \in \mathbb{F}(d)^{m_i \times p_i}$, $i = 1, \ldots, k$ and $\sum_{i=1}^k m_i = m$, is said to be maximally-decoupled if $\mathcal{C}^{(i)} = Im\, G^{(i)}(d)$ is undecomposable, $i = 1, \ldots, k$.

The determination of a decoupled encoder of a $[p, m]$-convolutional code $\mathcal{C}$ is directly related with a partition of the columns of the encoders of $\mathcal{C}$. We

will consider that the columns of any encoder of $\mathcal{C}$ constitute a set of nonzero generators of $\mathbb{F}((d))^m$. [3]

Definition 3.3 *A set of nonzero generators of* $\mathbb{F}((d))^m$,

$$\mathcal{G} = \{\hat{\mathbf{v}}_1(d), \hat{\mathbf{v}}_2(d), \ldots, \hat{\mathbf{v}}_p(d)\}$$

and a decomposition of $\mathbb{F}((d))^m$ *in direct sum*

$$\mathbb{F}((d))^m = V_1 \oplus V_2 \oplus \cdots \oplus V_k, \tag{3}$$

are compatible if every vector of $\mathcal{G}$ *belongs to a summand of (3) (and, obviously, to only one).*

If a generator set $\mathcal{G}$ is compatible with (3), then

(i) $\mathcal{G}_1 \dot{\cup} \mathcal{G}_2 \dot{\cup} \ldots \dot{\cup} \mathcal{G}_k$ with $\mathcal{G}_i := V_i \cap \mathcal{G}$, $i = 1 \ldots, k$, is a partition of $\mathcal{G}$ and $V_i = \text{span}(\mathcal{G}_i)$, $i = 1, \ldots, k$.

(ii) if $\mathrm{B} := \{\hat{\mathbf{v}}_{i_1}(d), \ldots, \hat{\mathbf{v}}_{i_m}(d)\} \subset \mathcal{G}$ is a basis of $\mathbb{F}((d))^m$, $\mathrm{B}_i := \mathcal{G}_i \cap \mathrm{B}$ is a basis of $\text{span}(\mathcal{G}_i)$.

(iii) there exists a unique finest direct sum decomposition

$$V = \bar{V}_1 \oplus \bar{V}_2 \oplus \cdots \oplus \bar{V}_h \tag{4}$$

compatible with $\mathcal{G}$. Each summand of any other compatible decomposition of $\mathbb{F}((d))^m$ can be expressed as a suitable sum of some $\bar{V}_i$s in (4).

The following algorithm determines the partition of $\mathcal{G} = \{\hat{\mathbf{v}}_1(d) \ldots \hat{\mathbf{v}}_p(d)\}$ associated with (4).

Algorithm 1:
Input: $G(d) = [\hat{\mathbf{v}}_1(d) \ldots \hat{\mathbf{v}}_p(d)]$.
Step 1: Select an $m \times m$ nonsingular submatrix $B(d)$ of $G(d)$ and put

$$X(d) = B(d)^{-1} G(d).$$

Step 2: Construct the $m \times p$ boolean matrix A defined by

$$A_{ij} = \begin{cases} 1 & if \quad X_{ij} \neq 0 \\ 0 & if \quad X_{ij} = 0 \end{cases}.$$

Step 3: Compute $(A^T A)^{p-1}$ and determine a permutation matrix $P \in \mathbb{F}^{p \times p}$ such that

$$P^T (A^T A)^{p-1} P = \text{diag}\{N^{(1)}, \ldots, N^{(h)}\},$$

[3] If the i-th column of an encoder of $\mathcal{C}$ is zero, the same happens for all equivalent encoders and, moreover, the i-th component of all codewords of $\mathcal{C}$ is also zero. Therefore to determine the decoupled encoders of $\mathcal{C}$ it is sufficient to consider the subcode of $\mathcal{C}$ constituted by its codewords without the i-th component, which encoders are the encoders of $\mathcal{C}$ without the i-th column.

where $N^{(i)} = \begin{bmatrix} 1 & \dots & 1 \end{bmatrix}^T \begin{bmatrix} 1 & \dots & 1 \end{bmatrix} \in \mathbb{F}^{p_i \times p_i}$, $i = 1, \dots, h$.

Step 4: Partitionate $P = [P_1 | \dots | P_h]$ where $P_i \in \mathbb{F}^{p \times p_i}$, $i = 1, \dots, h$ and define $\mathcal{P} := [GP_1 | GP_2 | \dots | GP_h]$.

Output: $\mathcal{P}$ and P.

Let $G(d)$ be an encoder of $\mathcal{C}$, $\mathcal{P} = [G_1(d) | \dots | G_h(d)]$, with $G_i(d) \in \mathbb{F}(d)^{m \times p_i}$, $i = 1, \dots, h$, be the partition of the columns of $G(d)$ obtained by applying Algorithm 1 and P be the corresponding permutation matrix. Then $[G_1(d) | \dots | G_h(d)] = G(d)P$ with

$$\mathbb{F}((d))^m = span\, G_1(d) \oplus \cdots \oplus span\, G_h(d).$$

Let also $[B_1(d) | \dots | B_h(d)]$ be an $m \times m$ nonsingular matrix such that $B_i(d) \in \mathbb{F}(d)^{m \times m_i}$ is a submatrix of $G_i(d)$, with $m_i = rank\, G_i(d)$, $i = 1, \dots, h$. Then

$$\bar{G}(d) := [B_1(d) | \dots | B_h(d)]^{-1} G(d) = \mathrm{diag}\{\bar{G}^{(1)}(d), \dots, \bar{G}^{(h)}(d)\} P^{-1} \qquad (5)$$

with $\bar{G}^{(i)}(d) \in \mathbb{F}(d)^{m_i \times p_i}$, $i = 1, \dots, h$. $\bar{G}(d)$ is a $(p_1, \dots, p_h)$-decoupled encoder of $\mathcal{C}$ which is *maximally-decoupled* since $[G_1(d) | \dots | G_h(d)]$ is the partition of the columns of $G(d)$ associated with the finest direct sum decomposition of $\mathbb{F}((d))^m$, which implies that the $[p_i, m_i]$-convolutional codes $\mathcal{C}^{(i)} = Im\, \bar{G}^{(i)}(d)$, $i = 1, \dots, h$, are undecomposable. Moreover, any other maximally-decoupled encoder of $\mathcal{C}$, $\tilde{G}(d)$, is such that $\tilde{G}(d)P = \mathrm{diag}\{\tilde{G}_1(d), \dots, \tilde{G}_h(d)\}$, with $\tilde{G}_i(d) \in \mathbb{F}(d)^{m_i \times p_i}$, $i = 1, \dots, h$.

Proposition 3.1 *If $\mathcal{C}$ admits a $(p_1, \dots, p_k)$-decoupled encoder associated with a permutation matrix P then it also admits a $(p_1, \dots, p_k)$-decoupled encoder associated with P which is canonical.*

The following result is immediate.

Corollary 3.1 *A convolutional code admits maximally-decoupled canonical encoders.*

4 Code decomposition in the analysis of a convolutional code

Let $\mathcal{C}$ be a $[p, m]$-convolutional code with free distance $d_{free}(\mathcal{C})$. Suppose that $\mathcal{C}$ can be decomposed into smaller codes, i.e., that admits a decoupled encoder

$$G(d) = \mathrm{diag}\{G^{(1)}(d), \dots, G^{(k)}(d)\} P,$$

with $k \geq 2$, $G^{(i)}(d) \in \mathbb{F}(d)^{m_i \times p_i}$, $i = 1, \dots, k$, $\sum_{i=1}^{k} m_i = m$, $\sum_{i=1}^{k} p_i = p$ and P a permutation matrix. Let $\mathcal{C}^{(i)}$ be the $[p_i, m_i]$-convolutional code generated by $G^{(i)}(d)$, $i = 1, \dots, k$. It is easy to see that [1]

$$d_{free}(\mathcal{C}) = \min_{1 \leq i \leq k} d_{free}(\mathcal{C}_i). \qquad (6)$$

So, if we have a code $\mathcal{C}$ which can be decomposed into smaller codes with different free distances, we obtain a better code just by considering the smaller subcode with better free distance. If all the subcodes have the same free distance but different rates, $\mathcal{C}$ will have rate smaller than the subcode with higher rate. Moreover, if $\mathcal{C}$ is an MDS-code, it can not be decomposable into smaller codes, as stated in the following proposition.

Proposition 4.1 *If $\mathcal{C}$ is an MDS-code then $\mathcal{C}$ is undecomposable.*

Proof: Assume by contradiction that $\mathcal{C}$ is an MDS-code that admits a (p_1, p_2)-decoupled encoder $G(d) \in \mathbb{F}(d)^{m\times p}$ for some positive integers p_1, p_2 such that $p_1 + p_2 = p$. Then $\mathcal{C}$ also admits a canonical (p_1, p_2)-decoupled encoder $G_c(d)$ such that

$$G_c(d)P = \begin{bmatrix} G^{(1)}(d) & 0 \\ 0 & G^{(2)}(d) \end{bmatrix},$$

for some permutation matrix P and $G^{(i)}(d) \in \mathbb{F}(d)^{m_i\times p_i}$, $i = 1, 2$, with $m_1 + m_2 = m$.

Let $\mathcal{C}^{(i)} = \text{Im } G^{(i)}(d)$ and represent $\nu_1 = \deg(\mathcal{C}^{(1)})$, $\nu_2 = \deg(\mathcal{C}^{(2)})$ and $\nu = \deg(\mathcal{C})$. Observe that $\nu_1 + \nu_2 = \nu$. Since $\mathcal{C}$ is an MDS-code,

$$d_{free}(\mathcal{C}) = (p - m)(\lfloor\frac{\nu}{m}\rfloor + 1) + \nu + 1.$$

Let us consider two cases: $\nu_2 m_1 \geq \nu_1 m_2$ and $\nu_2 m_1 < \nu_1 m_2$.

Case 1: $\nu_2 m_1 \geq \nu_1 m_2$. Since $p_1 + p_2 = p$, $m_1 + m_2 = m$ and $\nu_1 + \nu_2 = \nu$, we have that

$$d_{free}(\mathcal{C}) = (p_1 + p_2 - m_1 - m_2)(\lfloor\frac{\nu_1 + \nu_2}{m_1 + m_2}\rfloor + 1) + \nu_1 + \nu_2 + 1 =$$

$$= (p_1 - m_1)(\lfloor\frac{\nu_1}{m_1}\rfloor + 1) + \nu_1 + 1 + (p_1 - m_1)(\lfloor\frac{\nu_1 + \nu_2}{m_1 + m_2}\rfloor - \lfloor\frac{\nu_1}{m_1}\rfloor)+$$

$$+(p_2 - m_2)(\lfloor\frac{\nu_1 + \nu_2}{m_1 + m_2}\rfloor + 1) + \nu_2.$$

But $d_{free}(\mathcal{C}^{(1)}) \leq (p_1 - m_1)(\lfloor\frac{\nu_1}{m_1}\rfloor + 1) + \nu_1 + 1$ which implies that

$$d_{free}(\mathcal{C}) \geq d_{free}(\mathcal{C}^{(1)}) + (p_1 - m_1)(\lfloor\frac{\nu_1 + \nu_2}{m_1 + m_2}\rfloor - \lfloor\frac{\nu_1}{m_1}\rfloor)+$$

$$+(p_2 - m_2)(\lfloor\frac{\nu_1 + \nu_2}{m_1 + m_2}\rfloor + 1) + \nu_2$$

and therefore $d_{free}(\mathcal{C}) > d_{free}(\mathcal{C}^{(1)})$ since $(p_2 - m_2)(\lfloor\frac{\nu_1+\nu_2}{m_1+m_2}\rfloor + 1) + \nu_2 \geq 1$ which contradicts (6).

Case 2: $\nu_2 m_1 < \nu_1 m_2$. Proceeding the same way we conclude that $d_{free}(\mathcal{C}) > d_{free}(\mathcal{C}^{(2)})$ which also contradicts (6), and we conclude that $\mathcal{C}$ is undecomposable. □

Observe that the converse of the above lemma is not true as it is shown in the next example.

Example 4.1 *Consider the* $[4,2]$*-convolutional code* $\mathcal{C}$ *over the binary field such that*

$$G_c(d) = \begin{bmatrix} 1 & 0 & d & 0 \\ 0 & 1 & 1 & 1 \end{bmatrix}$$

is a canonical encoder of $\mathcal{C}$*. We can easily see that* $\mathcal{C}$ *is an undecomposable code which is not an MDS-code since it has free distance* 2 *but the corresponding generalized Singleton bound is* 4*.*

A similar result to (6) holds for the column distances of a convolutional code that can be decomposed into smaller codes, as stated in the following proposition.

Proposition 4.2 *Let* $G(d)$ *be a* $(p_1, \ldots, p_k)$*-decoupled encoder of* $\mathcal{C}$ *associated with a permutation matrix* P*,*

$$G(d)P = \operatorname{diag}\{G^{(1)}(d), \ldots, G^{(k)}(d)\}, \;\; G^{(i)}(d) \in \mathbb{F}(d)^{m_i \times p_i},$$

and $\mathcal{C}^{(i)}$ *be the* $[p_i, m_i]$*-convolutional code generated by* $G^{(i)}(d)$*,* $i = 1, \ldots, k$*. Then* $d_j^c(\mathcal{C}) = min_{1 \le i \le k}\, d_j^c(\mathcal{C}^{(i)})$*.*

Proof: By Proposition 3.1 we can assume without loss of generality that $G(d)$ is canonical. Representing $G^{(i)}(d) = G_0^{(i)} + G_1^{(i)} d + \cdots + G_\nu^{(i)} d^\nu$, $G_r^{(i)} \in \mathbb{F}^{m_i \times p_i}$, $i = 1, \ldots, k$, $r = 0, \ldots, \nu$, we have that

$$G(d) = \operatorname{diag}\{G_0^{(1)}, \ldots, G_0^{(k)}\}P + \operatorname{diag}\{G_1^{(1)}, \ldots, G_1^{(k)}\}Pd +$$

$$+ \cdots + \operatorname{diag}\{G_\nu^{(1)}, \ldots, G_\nu^{(k)}\}Pd^\nu$$

and then for all $j \ge 0$ there exist permutation matrices P_1 and P_2 such that

$$\mathbf{M}_j^c(G(d)) = P_1 \operatorname{diag}\{\mathbf{M}_j^c(G^{(1)}(d)), \ldots, \mathbf{M}_j^c(G^{(k)}(d))\}\, P_2,$$

where P_1 is such that if $\mathbf{u}_n = [\mathbf{u}_n^{(1)} \ldots \mathbf{u}_n^{(k)}]$, $\mathbf{u}_n^{(i)} \in \mathbb{F}^{m_i}$, $i = 1, \ldots, k$, $n = 0, \ldots, j$, then

$$[\mathbf{u}_0 \; \cdots \; \mathbf{u}_j]P_1 = [\mathbf{u}_0^{(1)} \; \cdots \; \mathbf{u}_j^{(1)} | \cdots | \mathbf{u}_0^{(k)} \; \cdots \; \mathbf{u}_j^{(k)}].$$

Consequently, for $\mathbf{u}_n \in \mathbb{F}^m$, $n = 0, \ldots, j$,

$$[\mathbf{u}_0 \; \cdots \; \mathbf{u}_j]\mathbf{M}_j^c(G(d)) =$$

$$= [\mathbf{u}_0^{(1)} \; \cdots \; \mathbf{u}_j^{(1)} | \cdots | \mathbf{u}_0^{(k)} \; \cdots \; \mathbf{u}_j^{(k)}] \operatorname{diag}\{\mathbf{M}_j^c(G^{(1)}(d)), \ldots, \mathbf{M}_j^c(G^{(k)}(d))\}\, P_2,$$

which implies that $d_j^c(\mathcal{C}) = \min_{1 \le i \le k_j} d_j^c(\mathcal{C}^{(i)})$. □

Let $G_c(d) = G_0 + G_1 d + \cdots + G_\nu d^\nu$, with $G_i \in \mathbb{F}^{m \times p}$, $i = 1, \ldots, \nu$, be a canonical encoder of $\mathcal{C}$ and define

$$G_c(d)|_{[0,j]} = G_0 + G_1 d + \cdots + G_j d^j, \tag{7}$$

for $j = 0, 1, \ldots, \nu$. Since $G_c(d)$ is left prime, G_0 is full row rank and then so it is $G_c(d)|_{[0,j]},\ j = 0, 1, \ldots, \nu$. Therefore we can define $\mathcal{C}_{[j]}$ to be the $[p, m]$-convolutional code generated by $G_c(d)|_{[0,j]},\ j = 0, 1, \ldots, \nu$. It is immediate to see that

$$d_j^c(\mathcal{C}) = d_j^c(\mathcal{C}_{[j]}),\ j = 0, 1, \ldots, \nu.$$

Observe that if $G_c(d)$ and $\tilde{G}_c(d)$ are equivalent canonical encoders and $\mathcal{C}_{[j]}$ and $\tilde{\mathcal{C}}_{[j]}$ are the convolutional encoders generated by $G_c(d)|_{[0,j]}$ and $\tilde{G}_c(d)|_{[0,j]}$, respectively, it is not true that $\mathcal{C}_{[j]}$ and $\tilde{\mathcal{C}}_{[j]}$ are the same as it is shown in the following example.

Example 4.2 *Consider the* $[5, 3]$*-convolutional code* $\mathcal{C}$ *generated by the equivalent canonical encoders*

$$G_c(d) = \begin{bmatrix} 1 & 0 & d^2 & d^3 & 0 \\ 0 & d & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1+d \end{bmatrix}$$

and

$$\tilde{G}_c(d) = \begin{bmatrix} 1 & d^2 & d^2+d & d^3+d^2 & d^3+d^2 \\ 0 & d & 1 & 1 & 1+d \\ 0 & d & 1 & -1 & -1-d \end{bmatrix}$$

and let $j = 1$*. It is easy to see that the convolutional codes*

$$\mathcal{C}_{[1]} = Im\, G_c(d)|_{[0,1]} = Im \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & d & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1+d \end{bmatrix}$$

and

$$\tilde{\mathcal{C}}_{[1]} = Im\, \tilde{G}_c(d)|[0, 1] = Im \begin{bmatrix} 1 & 0 & d & 0 & 0 \\ 0 & d & 1 & 1 & 1+d \\ 0 & d & 1 & -1 & -1-d \end{bmatrix}$$

are distinct.

However, these codes $\mathcal{C}_{[j]}$ and $\tilde{\mathcal{C}}_{[j]}$ have similar properties of decoupling as stated in the following proposition.

Proposition 4.3 *Let* $G_c(d)$ *and* $\tilde{G}_c(d)$ *in* $\mathbb{F}[d]^{m\times p}$ *be equivalent canonical encoders of degree* ν *and let* $\mathcal{C}_{[j]}$ *and* $\tilde{\mathcal{C}}_{[j]}$ *be the convolutional codes generated by* $G_c(d)|_{[0,j]}$ *and* $\tilde{G}_c(d)|_{[0,j]}$*, respectively, for* $j = 0, 1, \ldots, \nu$*. Then if* $G_c(d)|_{[0,j]}P = \mathrm{diag}\{G^{(1)}(d), \ldots, G^{(k)}(d)\}$*,* $G^{(i)}(d) \in \mathbb{F}[d]^{m_i\times p_i}$*,* $i = 1, \ldots, k$*, with* $\sum_{i=1}^{k} m_i = m$*,* $\sum_{i=1}^{k} p_i = p$ *and* P *a permutation matrix, then* $\mathrm{diag}\{G^{(1)}(d), \ldots, G^{(k)}(d)\}P^{-1}$ *is also an encoder of* $\tilde{\mathcal{C}}_{[j]}$*.*

Proof: Since $G_c(d)$ and $\tilde{G}_c(d)$ are equivalent canonical encoders, there exists a unimodular matrix $U(d) \in \mathbb{F}[d]^{m\times p}$ such that $G_c(d) = U(d)\tilde{G}_c(d)$. Therefore $\bar{G}(d) := U(d)\tilde{G}_c(d)|_{[0,j]}$ is an encoder of $\tilde{\mathcal{C}}|_{[0,j]}$ such that $\bar{G}(d)|_{[0,j]}P = (U(d)\tilde{G}_c(d)|_{[0,j]})|_{[0,j]}P = (U(d)\tilde{G}_c(d))|_{[0,j]}P = G_c(d)|_{[0,j]}P = \mathrm{diag}\{G^{(1)}(d),\dots,G^{(k)}(d)\}$. □

Propositions 4.2 and 4.3 immediately imply the following result.

Corollary 4.1 *Let $\mathcal{C}$ be a $[p,m]$-convolutional code. If $d_j^c(\mathcal{C}) = (p-m)(j+1)+1$ for some $j \geq 0$ then $\mathcal{C}$ does not have a canonical encoder $G_c(d)$ such that $\mathcal{C}_{[j]} := Im\, G_c(d)|_{[0,j]}$ is decomposable into $k \geq 2$ smaller codes.*

5 Conclusions

We have showed that, similarly to the free distance, a convolutional code that can be decomposed into smaller codes has j-th column distance equal to the minimum of the j-th column distances of its subcodes. Moreover, a convolutional code $\mathcal{C}$ can be undecomposable and admit a canonical encoder $G_c(d)$ such that $G_c(d)_{[0,j]}$ is decoupled for some j (as in Example 4.2 where $\mathcal{C}$ is undecomposable and $G_c(d)|_{[0,1]}$ is a $(3,2)$-decoupled encoder). The j-th column distance of such code $\mathcal{C}$ is equal to the j-th column distance of $\mathcal{C}_{[j]} := Im\, G_c(d)_{[0,j]}$ which can be decomposed into smaller codes and, consequently, the j-th column distance of $\mathcal{C}$ is equal to the minimum of the j-th column distances of the subcodes of $\mathcal{C}_{[j]}$. A subject of future investigation is the study of these codes. Although they seem not to be the best codes in terms of their distances, they seem to present good performance in terms of decoding.

References

[1] J-J. Climent, V. Herrandez, C. Perea, *New convolutional codes from old convolutional codes, Electronic Proceedings of the 16th International Symposium on Mathematical Theory and Systems (MTNS2004)* (2004).

[2] E. Fornasini, R. Pinto, *Matrix fraction descriptions in convolutional coding, Linear Algebra and its Applications* (2004), 392, 119-158.

[3] G.D. Forney Jr., R. Johannesson, Z. Wan, *Minimal and canonical rational generator matrices for convolutional codes, IEEE Trans. Inform. Theory* (1996), 42:6, 1865-1880.

[4] H. Gluesing-Luerssen, J. Rosenthal, R. Smarandache, *Strongly-MDS convolutional codes, IEEE Trans. Inform. Theory* (2006) 52:2, 584-598.

[5] R. Johannesson, K. Zigangirov, *Fundamentals of convolutional coding, Piscataway. NJ: IEEE Press* (1999).

[6] J. Rosenthal, R. Smarandache, *Maximum distance separable convolutional codes, Applicable Algebra in Engineering, Communication and Computing* (1999), 10(1), 15-32.

Título:	Control y pluma térmica: dos perspectivas de un problema de convección.
Doctorando:	María Cruz Navarro Lérida.
Director/es:	Henar Herrero Sanz, Ana María Mancho Sánchez.
Defensa:	18 de diciembre de 2007, Universidad de Castilla–La Mancha.
Calificación:	Sobresaliente cum laude por unanimidad.

Resumen:

En el presente trabajo, consideramos un problema de Rayleigh–Bénard en un cilindro con calentamiento localizado en la tapa inferior mediante un perfil Gaussiano de temperatura. Así, aparte de los gradientes de temperatura horizontal y vertical, se introduce un parámetros que permite considerar diferentes perfiles de temperatura localizados en el centro del cilindro aproximando a una pluma térmica. El principal objetivo del trabajo se centra en el estudio del tipo de inestabilidades desarrolladas dependiendo de los parámetros relacionados con el calor, que definen el perfil de temperatura en la tapa inferior. Dependiendo de los valores de dichos parámetros una gran variedad de estructuras tridimensionales aparecen: espirales gigantes de un solo brazo o de varios brazos localizadas en el centro del cilindro o extendidas a lo largo de toda la celda, estructuras estacionarias localizadas en la parte externa del cilindro, etc. El método numérico utilizado en el estudio está basado en un método de Chebyshev colocación mostrando ser una herramienta muy eficiente para el estudio de estos problemas termoconvectivos. El estudio numérico de las inestabilidades desarrolladas en el sistema dependiendo de los parámetros físicos requiere un método numérico eficiente para el cálculo de autovalores del problema de autovalores generalizado derivado. Este ha sido otro de los objetivos de la tesis. El uso de la transformación de Cayley y una segunda transformación del problema de autovalores para el tratamiento de los autovalores infinitos nos ha permitido diseñar un algoritmo rápido y eficiente basado en el método de Arnoldi para determinar el autovalor crítico que describe si el flujo es estable o inestable. Los flujos termoconvectivos aparecen, además de en la naturaleza, en numerosas aplicaciones industriales. Por ejemplo, inestabilidades termoconvectivas son responsables de estados convectivos no deseados en procesos de formación de aleaciones, etc. En estos procesos es importante evitar estructuras convectivas para conseguir materiales homogéneos y resistentes. En la última parte del presente trabajo, abordamos un problema de control óptimo para el problema de Rayleigh–Bénard con calentamiento localizado estudiado en la primera parte. En concreto, buscamos un flujo de calor en la tapa superior del

cilindro que minimice la enstrofía del flujo. Más específicamente, el problema de control es formulado como un problema de optimización con restricciones que minimiza un funcional que involucra la vorticidad del flujo y el flujo de calor en la frontera superior del dominio. Las condiciones de optimalidad derivadas son resueltas usando de nuevo un método de Chebyshev colocación. Del estudio se obtienen tres resultados importantes. En primer lugar, hemos comprobado que las técnicas de control óptimo pueden ser usadas para evitar la formación de estructuras convectivas en los estados básicos. Segundo, los estados básicos controlados encontrados presentan una fuerte reducción en el valor de la enstrofía y tercero, podemos concluir que flujos de calor adecuados en la frontera superior conducen a nuevos estados controlados muy estables para los que las inestabilidades termoconvectivas son evitadas.

Título:	Non-periodic Γ-convergence.
Doctorando:	Hélia Serrano.
Director/es:	Pablo Pedregal.
Defensa:	4 de diciembre 2007, Madrid.
Calificación:	Sobresaliente cum laude.

Resumen:

This Thesis focuses on the study of Γ-convergence of integral functionals, in the non-periodic setting, and homogenization of second-order elliptic equations (in divergence form) and p–laplacian equations. Precisely, the explicit representation of the Γ–limit of sequences of functionals of the form

$$I_\varepsilon(u) = \int_\Omega W(a_\varepsilon(x), \nabla u(x))\, dx, \qquad u \in W^{1,p}(\Omega),$$

where $W : \mathbb{R}^m \times \mathbb{R}^n \to \mathbb{R}$ is continuous and satisfies a growth condition of order p in the second variable, is studied in the general non–periodic case. A new sufficient condition on the sequence $\{a_\varepsilon\}$ is introduced, which is called CGP (Composition Gradient Property), in order to characterize the sequences $\{a_\varepsilon\}$ for which is possible to represent the limit energy density through the Young measure associated with $\{a_\varepsilon\}$. Basically, this condition asks $\{a_\varepsilon\}$ to be "essentially a sequence of gradients", when composed with an one-to-one continuous function.

On the other hand, the Γ–limit of sequences of quadratic functionals with oscillatory linear perturbations of type

$$I_\varepsilon(u) \;=\; \int_\Omega \left[\nabla u(x)^T \frac{A_\varepsilon(x)}{2} \nabla u(x) \;+\; b_\varepsilon(x) \cdot \nabla u(x) \right] dx$$

is studied in four different situations. In the general one–dimensional case, a full characterization of the density of the Γ–limit is achieved in terms of the Young measure associated with $\{A_\varepsilon\}$, and the joint Young measure associated with $\{(A_\varepsilon, b_\varepsilon)\}$. In the periodic multidimensional case, two situations are considered: when $\{A_\varepsilon\}$ and $\{b_\varepsilon\}$ oscillate at the same family of separated length scales, and when do not. It is pointed out that the linear term of the limit energy density does not depend on $\{A_\varepsilon\}$, whenever $\{A_\varepsilon\}$ and $\{b_\varepsilon\}$ oscillate at different length scales. In the non–periodic multidimensional case, the CGP condition on the sequence of pairs $\{(A_\varepsilon, b_\varepsilon)\}$ is re–introduced and some interesting examples are explored.

Besides, the Γ–convergence (with respect to the weak topology in $W^{1,p}(\Omega)$) of sequences of functionals with non–standard growth conditions of type

$$I_\varepsilon(u) \;=\; \int_\Omega f(a_\varepsilon(x))\, |\nabla u(x)|^{a_\varepsilon(x)}\, dx,$$

where $\{a_\varepsilon\}$ stands for a first order laminate taking the values p and q, with $1 \leq p \leq q < \infty$, is also studied. A characterization of the limit energy density is obtained through a finite minimization problem, and the Γ-limit is defined in an intermediate space between $W^{1,p}(\Omega)$ and $W^{1,q}(\Omega)$.

Tipo de evento:	Congreso
Nombre:	Congreso en honor de Mikel Bilbao
Lugar:	Facultad de Ciencia y Tecnología, Universidad del País Vasco–Euskal Herriko Unibertsitatea, Leioa (Vizcaya)
Fecha:	del 5 al 9 de mayo de 2008
Organiza:	R. Clement, L. Escauriaza, M. Escobedo, M. A. de Prada, J. I. Maeztu, F. Vadillo, J. M. Zarzuelo
Información:	
E-mail:	miguel.escobedo@ehu.es
WWW:	www.ehu.es/miguelescobedo/congresoMB

Tipo de evento:	Curso
Nombre:	Course on nanotechnology and mathematics
Lugar:	Santiago de Compostela
Fecha:	del 6 al 8 de mayo de 2008
Organiza:	
Información:	
E-mail:	fpena@usc.es (Francisco José Pena Brage)
WWW:	www.usc.es/nanomath

Tipo de evento:	Workshop
Nombre:	Workshop on Nonlinear Processes in Oceanic and Atmospheric Flows
Lugar:	Castro Urdiales (Cantabria)
Fecha:	del 2 al 8 de julio de 2008
Organiza:	
Información:	
E-mail:	A.M.Mancho@imaff.cfmac.csic.es, nloa2008@ifisc.uib.es
WWW:	http://ifisc.uib.es/public/nloa2008

Tipo de evento:	Congreso
Nombre:	8th International Conference on Computational and Mathematical Methods in Science and Engineering (CMMSE-2008)
Lugar:	Murcia
Fecha:	del 13 al 16 de julio de 2008
Organiza:	
Información:	
E-mail:	cmmse@usal.es, jvigo@usal.es
WWW:	www.usal.es/ CMMSE

Tipo de evento:	Congreso
Nombre:	SIAG/LA-SIMUMAT International Summer School on Numerical Linear Algebra
Lugar:	Centro Internacional de Encuentros Matemáticos (CIEM), Castro Urdiales, Cantabria
Fecha:	del 21 al 25 de julio de 2008
Organiza:	
Información:	
E-mail:	dopico@math.uc3m.es (Froilán M. Dopico)
WWW:	www.simumat.es/SIAGLA2008

Tipo de evento:	Workshop
Nombre:	Workshop on Reliable Modelling and Scientific Computing
Lugar:	University of Jyväskylä, Department of Mathematical Information Technology, Jyväskylä, Finland
Fecha:	15 y 16 de agosto de 2008
Organiza:	University of Jyväskylä
Información:	Prof. Olli Mali
E-mail:	olli.mali@jyu.fi
WWW:	www.mit.jyu.fi/scoma/Worms2008

Tipo de evento:	Congreso
Nombre:	XIII Escuela Jacques-Louis Lions Hispano-Francesa sobre Simulación Numérica en Física e Ingeniería
Lugar:	Valladolid
Fecha:	del 15 al 19 de septiembre de 2008
Organiza:	Universidad de Valladolid, SEMA, SMAI
Información:	Mari Paz Calvo
E-mail:	ehf2008@uva.es
WWW:	http://hermite.mac.cie.uva.es/ehf2008

Tipo de evento:	Congreso
Nombre:	Mathematical Models in Engineering, Biology and Medicine. Conference on Boundary Value Problems
Lugar:	Santiago de Compostela
Fecha:	del 16 al 19 de septiembre de 2008
Organiza:	
Información:	
E-mail:	bvp2008@gmail.com
WWW:	www.usc.es/congresos/bvp2008

Tipo de evento:	Congreso
Nombre:	International Conference of Numerical Analysis and Applied Mathematics 2008 (ICNAAM 2008)
Lugar:	Hotel Kypriotis Village-Kypriotis Panorama-Kypriotis International Conference Center, Psalidi, Kos, Greece
Fecha:	del 16 al 20 de septiembre de 2008
Organiza:	Prof. T. E. Simos, Dr. Ch. Tsitouras, Dr. G. Psihoyios
Información:	Secretary ICNAAM (Mrs. Eleni Ralli–Simou. Postal Address: 10 Konitsis Street, Amfithea Paleon Faliro, GR-175 64, Athens, Greece. Fax: +30210 94 20 091, +30 2710 237 397
E-mail:	tsimos@mail.ariadne-t.gr, tsimos.conf@gmail.com
WWW:	www.icnaam.org

Colección SMAI

La Sociedad francesa de Matemáticas Aplicadas e Industriales (SMAI) publica dos colecciones de libros científicos. La más antigua, *Mathématiques et Applications*, publicada por Springer desde 1992, publica cursos de nivel posgrado M2, final de estudios de escuelas de ingenieros o escuelas doctorales, y monografías de investigación a un nivel introductorio. La colección más reciente se llama *Mathématiques appliquées pour le Master/SMAI*, es publicada por Dunod desde 2006, y publica obras de enseñanza de nivel posgrado M1. El objetivo de este artículo es presentar estas dos colecciones y explicar sus diferencias de objetivos.

El objetivo de la colección *Mathématiques et Applications* es pues publicar, en francés o en inglés, cursos avanzados de posgrado M2, de escuelas doctorales o las introducciones pedagógicas a ámbitos de investigación. Los lectores potenciales son estudiantes de nivel predoctorado o doctorado, pero también los investigadores e ingenieros de todos horizontes que quieren iniciarse en los métodos y en los resultados de las matemáticas aplicadas. Los temas abordados cubren tanto los ámbitos clásicos de las matemáticas aplicadas (análisis numérico y ecuaciones en derivadas parciales, probabilidad y estadística, optimización e investigación operativa...) como de las aplicaciones más específicas (en ciencias naturales y físicas, economía, informática, ingeniería, tratamiento de señales e imágenes...). Algunas obras tendrán así una vocación puramente pedagógica mientras que otras podrán constituir textos de referencia. La vocación de esta colección es de ser distribuida ampliamente en todo el mundo gracias a su editor internacional, Springer. Las obras publicadas están escritas mayoritariamente en lengua francesa pero algunas se escriben en inglés. El Comité editorial, nombrado por la SMAI para períodos de 4 años, garantiza la calidad científica y pedagógica de las obras.

La colección *Mathématiques appliquées pour le Master/SMAI* se inscribe en el marco de la nueva organización LMD de la enseñanza superior en Europa (reforma de Bolonia). Su objetivo es promover una nueva generación de obras de nivel Curso de posgrado M1, mejor adaptados a los cursos actuales. Se inscribe en la prolongación de la antigua colección *Mathématiques appliquées pour la maîtrise*, dirigida por Ph. Ciarlet y J.-L. Lions en la editora Masson, retomada por Dunod. Permite a estudiantes, científicos o ingenieros adquirir las bases teóricas de un nuevo ámbito matemático proponiendo llaves para una explotación aplicada. Así pues, su mayor legibilidad está al servicio de la calidad científica. La SMAI garantiza la dirección editorial gracias a un Comité renovado periódicamente, y ampliamente representativo de los distintos temas de las matemáticas aplicadas. Su ambición es constituir un conjunto de obras de referencia. Las obras se publican exclusivamente en lengua francesa pero Dunod puede ayudar a una publicación posterior en inglés para una difusión fuera de Francia u otros países de lengua francesa.

Los lectores o autores potenciales encontrarán más información con respecto

a estas dos colecciones en la página web de la SMAI `smai.emath.fr`, en el capítulo *publicaciones*.

Artículo redactado por

Grégoire Allaire y Michel Benaïm, directores de *Mathématiques et Aplications*, y por Monique Dauge y Olivier Pironneau, directores de *Mathématiques Appliquées pour le Master/SMAI*.

Últimas obras publicadas en la colección *Mathématiques et Aplications*:

- Volumen 58, G. Allaire, *Conception optimale de structures*, 2007, 278 p.
- Volumen 59, M. Elkadi, B. Mourrain, *Introduction à la résolution des systèmes polynomiaux*, 2007, 307 p.
- Volumen 60, N. Caspard, B. Monjardet, B. Leclerc, *Ensembles ordonnés finis : concepts, résultats et usages*, 2007, 340 p.
- Volumen 61, H. Pham, *Optimisation et contrôle stochastique appliqués à la finance*, 2007, 188 p.

Obras publicadas en la colección *Mathématiques Appliquées pour le Master/SMAI*:

- F. Comets et T. Meyre, *Calcul stochastique et modèles de diffusion*, 2006, 336 p.
- F. Bonnans, *Optimisation continue*, 2006, 336 p.
- E. Pardoux, *Processus de Markov et applications*, 2007, 336 p.
- B. Bercu et D. Chafaï, *Modélisation stochastique et simulation*, 2007, 352 p.

Abderramán Marrero, Jesús C.
Prof. Ayudante. *Líneas de investigación:* Física matemática, ecuaciones en diferencias, algoritmos genéticos – Univ. Politécnica de Madrid – Fac. de Informática – Dpto. de Matemática Aplicada – Campus Montegancedo - Boadilla. 28660 Madrid.
Tlf.: 913.365.014. *Fax:* 913.367.426.
e-mail: jc.abderraman@fi.upm.es.
http://www.dma.fi.upm.es/jesus

Alarcón Cotillas, Begoña
Estudiante. *Líneas de investigación:* Dinámica topológica – Univ. de Valencia – Facultad de Matemáticas – Depto. de Matemática Aplicada – Avda. Vicente Andrés Estellés, 1. 46100 - Burjassot (Valencia).
Tlf.: 963.543.233.
e-mail: bego.alarcon@uv.es.

Baeza Manzanares, Antonio
Líneas de investigación: Dinámica de fluidos computacional – IMDEA Matemáticas – Fac. de Ciencias – Dpto. de Matemática Aplicada a la Aeronáutica – Campus de Cantoblanco, C-IX. 28049 Madrid.
Tlf.: 914.976.830.
e-mail: tbaeza@gmail.com.

Cabada Fernández, Alberto
Prof. Titular de Universidad. *Líneas de investigación:* Ecuaciones diferenciales ordinarias – Univ. de Santiago de Compostela – Fac. de Matemáticas – Depto. de Análisis Matemático – Campus Universitario Sur, s/n. 15782 - Santiago de Compostela.
Tlf.: 981.563.100, Ext. 13206. *Fax:* 981.597.054.
e-mail: cabada@usc.es.
http://web.usc.es/~cabada

Casoni Rero, Eva
Estudiante (Becario). *Líneas de investigación:* Elementos finitos, discontinuos Galerkin, problemas de transporte, flujo compresible – Univ. Politécnica de Cataluña – E. T.S. de Ingenieros de Caminos, Canales y Puertos – Dpto. de Matemática Aplicada III – Jordi Girona Salgado, 1-3. 08034 Barcelona.
Tlf.: 934.017.959. *Fax:* 934.011.825.
e-mail: eva.casoni@upc.edu.
http://www-lacan.upc.edu

Costa Romero, Sara
Prof. Asociado. *Líneas de investigación:* Sistemas dinámicos discretos – Univ. Autónoma de Barcelona – Facultad de Ciencias – Depto. de Matemáticas – Campus Universitario. 08193 - Cerdanyola del Vallés.
Tlf.: 935.811.886. *Fax:* 935.812.790.
e-mail: scosta@mat.uab.cat.
http://www.gsd.uab.cat/personal/scosta

Ferreiro Darriba, Juan Bosco
Prof. Titular de Escuela Universitaria. *Líneas de investigación:* Ecuaciones en diferencias, ecuaciones diferenciales ordinarias, ecuaciones funcionales, principios del máximo, estabilidad – Univ. de Santiago de Compostela – Escuela Politécnica Superior – Depto. de Matemática Aplicada – C/ Benigno Ledo. Campus Universitario. 27002 - Lugo.
Tlf.: 982.285.900, Ext. 23230. *Fax:* 982.285.926.
e-mail: jbosco@lugo.usc.es.

González Peñas, David
Estudiante.
Univ. de Vigo
e-mail: DGPenhas@gmail.com.

López Pouso, Rodrigo
Prof. Titular de Universidad. *Líneas de investigación:* Ecuaciones diferenciales ordinarias – Univ. de Santiago de Compostela – Fac. de Matemáticas – Depto. de Análisis Matemático – C/ Lope Gómez de Marzoa, s/n. Campus Sur. 15782 - Santiago de Compostela.
Tlf.: 981.563.100, Ext. 13166. *Fax:* 981.597.054.
e-mail: rodrigolp@usc.es.

Martínez Dopico, Froilán
Prof. Titular de Universidad. *Líneas de investigación:* Álgebra lineal numérica, teoría de perturbaciones de matrices – Univ. Carlos III de Madrid – Escuela Politécnica Superior – Dpto. de Matemáticas – Avda. de la Universidad, 30. 28911 Leganés (Madrid).
Tlf.: 916.249.446. *Fax:* 916.249.129.
e-mail: dopico@math.uc3m.es.

Meis Fernández, Marcos
Estudiante. *Líneas de investigación:* Dinámica computacional en mecánica de fluidos – Univ. de Vigo – E. T. S. de Ingenieros de Telecomunicación – Dpto. de Matemática Aplicada II – Campus de Lagoas-Marcosende. 36200 Vigo.
Tlf.: 986.813.613.
e-mail: marcos@dma.uvigo.es.

Mora Giné, Xavier
Prof. Titular de Universidad. *Líneas de investigación:* EDP's, métodos de votación – Univ. Autónoma de Barcelona – Fac. de Ciencias – Dpto. de Matemáticas – Campus Universitari. 08193 Bellaterra.
Tlf.: 935.811.302. *Fax:* 935.812.790.
e-mail: xmora@mat.uab.cat.

Morales de Luna, Tomás
Técnico de Investigación. *Líneas de investigación:* – Univ. de Málaga – Fac. de Ciencias – Depto. de Análisis Matemático – Campus de Teatinos, s/n. 29071 - Málaga.
Tlf.: 952.132.016.
e-mail: morales@anamat.cie.uma.es.

Novo Martín, Sylvia
Prof. Titular de Universidad. *Líneas de investigación:* Sistemas dinámicos no autónomos – Univ. de Valladolid – E. T. S. de Ingenieros Industriales – Depto. de Matemática Aplicada – Paseo del Cauce, s/n. 47011 - Valladolid.
Tlf.: 983.423.393. *Fax:* 983.423.406.
e-mail: sylnov@wmatem.eis.uva.es.

Pérez Fernández, Teresa E.
Prof. Titular de Universidad. *Líneas de investigación:* Teoría de aproximación. Polinomios ortogonales – Univ. de Granada – E. T. S. de Ingenierías Informática y Telecomunicación – Dpto. de Matemática Aplicada – C/ Daniel Saucedo Aranda, s/n. 18071 Granada.
Tlf.: 958.243.192. *Fax:* 958.248.596.
e-mail: tperez@ugr.es.
http://www.ugr.es/local/tperez

Requena Arévalo, Verónica
Estudiante (Becario). *Líneas de investigación:* Criptografía, teoría de códigos y seguridad informática – Univ. de Alicante – Escuela Politécnica Superior – Depto. de Ciencia de la Computación e Inteligencia Artificial – Campus de Sant Vicent. 03080 Alicante.
Tlf.: 965.903.400, ext. 2068. *Fax:* 965.903.902.
e-mail: vrequena@dccia.ua.es.

Tomás Estevan, Virtudes
Estudiante (Becario). *Líneas de investigación:* Teoría de códigos, criptografía – Univ. de Alicante – Escuela Politécnica Superior – Depto. de Ciencia de la Computación e Inteligencia Artificial – Crta. San Vicente del Respeig, s/n. 03690 Alicante.
Tlf.: 965.903.400, ext. 2068. *Fax:* 965.903.464.
e-mail: vtomas@dccia.ua.es.
http://www.ua.es

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 SëMA

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 SëMA

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 S$\ddot{e}$MA
   Dpto. E.D.A.N., Facultad de Matemáticas
   Aptdo. 1160, 41080 SEVILLA

   También podrán ser enviados por correo electrónico a la dirección

   `boletin.sema@uclm.es`

   En ambos casos, el/los autor/es deberán enviar por correo certificado una carta a la dirección precedente mencionando explícitamente que el artículo es sometido a publicación e indicando el nombre y dirección del autor corresponsal. En esta carta, podrán sugerirse nombres de miembros del Comité Científico que, a juicio de los autores, sean especialmente adecuados para juzgar el trabajo.

   La decisión final sobre aceptación del trabajo será precedida de un procedimiento de revisión anónima.

2. Las contribuciones serán preferiblemente de una longitud inferior a 24 páginas y se deberán ajustar al formato indicado en los ficheros a tal efecto disponibles en la página web de la Sociedad (`http://www.sema.org.es/`).

3. El contenido de los artículos publicados corresponderá a un área de trabajo preferiblemente conectada a los objetivos propios de la Matemática Aplicada. En los trabajos podrá incluirse información sobre resultados conocidos y/o previamente publicados. Se anima especialmente a los autores a presentar sus propios resultados (y en su caso los de otros investigadores) con estilo y objetivos divulgativos.

Ficha de Inscripción Individual

Sociedad Española de Matemática Aplicada 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