SëMA

BOLETÍN NÚMERO 37

Diciembre 2006

sumario

Boletín de la Sociedad Española de Matemática Aplicada SëMA

Página web de SëMA
`http://www.sema.org.es/`

Dirección Editorial: Boletín de SëMA. Dpto. de Matemática Aplicada. Universidad de Salamanca. Plaza de la Merced, s/n. 37008. Salamanca. `boletin_sema@usal.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: Luis Ferragut Alonso.

Estimados compañeros:

Finalizamos el año con el que será el último Boletín que realice el actual Grupo Editor. Un grupo de compañeros de la Universidad de Castilla la Mancha ha accedido con gran amabilidad a tomar el relevo. Desde aquí, nuestro agradecimiento y nuestros mejores deseos. Sin ninguna duda, sabrán mejorar y corregir los errores que hayamos podido cometer.

En este número podéis seguir disfrutando con la segunda parte del trabajo de Roland Glowinski y sus colaboradores sobre el método de los elementos finitos para las ecuaciones de Navier-Stokes. Tomás Caraballo, desde la Universidad de Sevilla, nos presenta algunos resultados recientes de los efectos de estabilización por ruido estocástico de ecuaciones en derivadas parciales. Seguidamente, encontramos un interesante trabajo sobre simulación de reservas petrolíferas enviado por Zhangxin Chen, de la S. M. University de Dallas. Finalizamos la sección de artículos científicos con un trabajo de Oscar Ciaurri y Juan Luis Varona, de la Universidad de La Rioja, donde nos invitan a reflexionar sobre la fiabilidad de los cálculos hechos con ordenador.

En la sección de Matemática e Industria contamos con unas reflexiones de Barbara Lee Keyfitz, Directora del Fields Institute for Research in Mathematical Sciences de Canadá, sobre el futuro de la investigación matemática en la industria, que ha traducido para el Boletín nuestro compañero Mikel Lezaun, responsable de esta sección.

También contamos con una nota de Soledad Rodríguez Salazar, sobre las ideas que se lanzaron en el pasado ICM 2006 sobre la enseñanza de las Matemáticas, en general, y de la Matemática Aplicada en particular.

Este Boletín también incluye, como es habitual, algunos resúmenes de Tesis Doctorales recientemente defendidas por jóvenes investigadores en nuestro país, así como algunas reseñas de nuevos libros que pueden ser de vuestro interés.

Finalizamos con algunas noticias y anuncios, destacando la convocatoria del VIII Premio S$\vec{e}$MA$\tilde{d}$e Divulgación de la Matemática Aplicada.

Nuestro agradecimiento a todos los que habéis colaborado con nosotros, por vuestra generosidad y vuestra paciencia: a los autores por vuestras aportaciones y la comprensión mostrada para facilitar nuestra labor, a todos los responsables de secciones y en especial a Enrique Fernández-Cara, sin cuya incansable labor el Boletín de S$\vec{e}$MA no sería lo que es hoy.

Nuestra felicitación al Presidente saliente, Juan Ignacio Montijano, por su dedicación al frente de la Sociedad, y al nuevo Presidente de S$\vec{e}$MA, Carlos Vázquez Cendón, a quien auguramos una fructífera etapa al frente de la misma. Igualmente al resto de miembros, antiguos y nuevos, del Comité Ejecutivo y a todas aquellas personas que en un momento u otro han dedicado o dedican parte de su tiempo y esfuerzo para hacer avanzar la Sociedad Española de Matemática Aplicada.

No quisiéramos despedirnos sin antes agradecer y recordar a todos los

Grupos Editores que nos han precedido, en Málaga, Zaragoza, Córdoba y Oviedo, aportando su buen hacer y sus iniciativas y perfilando la actual estructura de nuestro Boletín, al que a buen seguro los compañeros de Castilla La Mancha sabrán dar un renovado empuje.

Un cordial saludo,

Grupo Editor
boletin_sema@usal.es

Queridos amigos y compañeros de la Sociedad:

Este periodo que he pasado al frente de SeMA ha sido realmente apasionante tanto desde el punto de vista personal como profesional. He tenido que enfrentarme a retos y tareas con los que, habitualmente, un profesor no trata, y me ha permitido una visión global de nuestra sociedad, e incluso de la investigación en matemática aplicada, que difícilmente hubiera tenido de no haber ocupado este cargo. El poder representar, promocionar y defender la querida SeMA ha sido una de las labores más interesantes de mi vida profesional. Al mismo tiempo, he podido hacer en estos años buenos amigos, hecho que me hace sentir enormemente agradecido.

Con estas líneas quiero despedirme de vosotros, dando las gracias a todos los que habéis colaborado en nuestras iniciativas. En particular a los miembros del Consejo Ejecutivo que han compartido conmigo las tareas de gestión de la Sociedad. A José Javier, Javier, José, Ireneo, que ya terminaron su mandato, y a Javier Chavarriga que desgraciadamente falleció; a Enrique, Pablo, Rafael, Inmaculada y José Antonio, que continúan en estos menesteres; y de manera especial al Vicepresidente Mikel Lezaun, cuya presencia me ha transmitido siempre una gran seguridad, al Secretario Carlos Castro, de ayuda inestimable, y a la Tesorera Pilar Laburta, que ha realizado una labor impecable. Ha sido un placer y un honor trabajar con todos vosotros. También quiero mencionar a quienes han hecho posible que el Boletín nos llegue regularmente cada trimestre. En particular a Enrique Fernández Cara y al grupo editor, de Salamanca, con Luis Ferragut a la cabeza y, por supuesto, a Mabel Asensio.

Sin la colaboración de todos, difícilmente sería SeMA la gran sociedad que es. No somos muchos, pero podemos presumir de una rica historia, de actividades propias de gran relevancia y de compartir sociedad con investigadores de enorme prestigio, algunos premiados nacional e internacionalmente, algunos con responsabilidades importantes en el contexto autonómico, nacional e internacional, que además son socios muy activos. SeMA mantiene excelentes relaciones con las sociedades más representativas, tiene una presencia destacada en los eventos importantes, como el reciente ICM2006, y tiene mucho que decir en el futuro de la investigación y el desarrollo de las matemáticas. Es tarea de todos conseguir que continúe en esta línea.

No tengo ninguna duda de que la presidencia pasa a buenas manos, las de Carlos Vázquez, a quien le deseo la colaboración desinteresada y el consejo amigable que tan a menudo me han sido brindados durante estos años. Mi agradecimiento por asumir esta responsabilidad, para la que me consta que cuenta con un excelente Consejo Ejecutivo. Suerte a todos ellos en esta andadura.

Muchas gracias a todos.

Juan Ignacio Montijano

Queridos socios de S$\bar{e}$MA, deseo aprovechar la ocasión que me brinda este número del Boletín para dirigirme a todos vosotros. En primer lugar, quiero agradeceros vuestra confianza en mi persona para presidir la Sociedad. Por mi parte, asumo esta nueva responsabilidad con ganas, orgullo y satisfacción, comprometiéndome a trabajar para que S$\bar{e}$MA continúe su trayectoria ascendente y a representar a la Sociedad allí donde sea necesario.

Desde su creación en 1991, S$\bar{e}$MA ha experimentado un crecimiento continuo en lo que se refiere a su presencia y relevancia en la sociedad científica y a su contribución a la Matemática Aplicada. Una parte importante del éxito se debe atribuir a la excelente labor de su Comisión Gestora inicial, sus sucesivos Presidentes, sus Consejos Ejecutivos y el resto de las personas que han venido dedicando su tiempo a distintas tareas relacionadas con la Sociedad. Aunque quiero aprovechar la ocasión para hacer extensivo el agradecimiento a todos ellos, deseo mostrarlo en especial a Juan Ignacio Montijano, nuestro Presidente saliente y un generoso colaborador en la etapa de transición, a Ireneo Peral, miembro saliente del Consejo Ejecutivo que ha realizado importantes aportaciones, y a Pilar Laburta, que termina su estupenda labor de Tesorera. También concluye sus labores de edición del Boletín el grupo de Salamanca, formado por Mabel Asensio, Luis Ferragut, María Teresa de Bustos, Antonio Fernández y Francisco Andrés. Su dedicación y la calidad indiscutible del trabajo realizado merecen nuestro agradecimiento.

Por otra parte, quiero felicitar por su elección a los nuevos miembros del Consejo Ejecutivo, Rosa Donat (Universidad de Valencia) y Carlos Parés (Universidad de Málaga). A partir de ahora, Iñigo Arregui (Universidad de La Coruña) se encargará de la Tesorería y, con el nuevo año, tomará el relevo en la edición del Boletín el grupo de la Universidad de Castilla-La Mancha, formado por Ernesto Aranda, José Carlos Bellido, Alfonso Bueno, Alberto Donoso y Pablo Pedregal. Todos ellos se incorporan a sus tareas con una excelente predisposición y ganas de trabajar, por lo que sus aportaciones serán muy beneficiosas para la Sociedad.

En esta nueva etapa es nuestra intención continuar las líneas de actuación de nuestros predecesores, promoviendo y estimulando la investigación y la divulgación de los conocimientos en Matemática Aplicada, potenciando el acercamiento de ésta a la sociedad y fomentando la cooperación con otras sociedades e instituciones nacionales y extranjeras.

Para ello, será muy importante seguir impulsando y patrocinando foros y congresos científicos en el ámbito de la Matemática Aplicada. En este aspecto, debemos recordar la excelente organización de la recientemente celebrada Escuela Hispano-Francesa de Castro-Urdiales (gracias a Mikel Lezaun y a los demás organizadores del evento). En el curso académico anterior, han destacado el ICM y la Asamblea de la IMU como acontecimientos matemáticos mundiales celebrados en España. Deseo personalizar en Manuel de León y

en Juan Viaño nuestro agradecimiento por su trabajo y nuestra felicitación por los resultados obtenidos en la organización de sendas actividades, gracias que hacemos extensivas a todos nuestros socios encargados de distintas tareas relacionadas con ambos eventos.

En el curso que se inicia se celebrarán distintas actividades con patrocinio de S$\vec{e}$MA, como RTNS 2007, Zaragoza Numérica, el Congreso Hispano-Francés de Matemáticas y Benasque 2007, a los que os animamos a asistir. En especial, esperamos vuestra participación en *nuestro congreso* CEDYA-CMA, que está siendo cuidadosamente organizado por el Departamento de Ecuaciones Diferenciales y Análisis Numérico de la Universidad de Sevilla. Sin duda, la calidad de los conferenciantes invitados, las interesantes temáticas de las sesiones especiales, la posibilidad de escuchar y presentar comunicaciones en una amplia variedad de temas de Matemática Aplicada y el atractivo de la ciudad de Sevilla, deberían convertir el XX CEDYA-X CMA en una cita ineludible. En el plano internacional, os recuerdo que el ICIAM 2007 se celebrará en Zurich en el mes de julio, siendo S$\vec{e}$MA una de las sociedades del consorcio constituido en CICIAM, organizador del evento. Creo que es muy importante asistir para incrementar nuestra presencia en este congreso mundial de Matemática Aplicada. Aunque es una cita más lejana, también os informo de que la Escuela Hispano-Francesa de 2008 se celebrará en Valladolid, encargándose de ella el Departamento de Matemática Aplicada, con la novedad de involucrar también en la organización a la SMAI francesa.

Por otra parte, para mantener la difusión de la información de S$\vec{e}$MA, se ha modificado la página web de la Sociedad, incluyendo las secciones de Bolsa de Trabajo y Anuncios a las que os invito a que hagáis llegar toda la información de interés, pues de ello depende el éxito de la iniciativa. También, semanalmente recibiréis un correo electrónico *Alerta S$\vec{e}$MA*, que (al margen del posible susto de la primera vez que lo recibisteis) pretende informaros de las últimas novedades. Cualquier comentario para mejorar el funcionamiento de estos servicios es bienvenido y redundará en beneficio de todos.

En el apartado de Premios, ya publicamos en este número del Boletín las bases de la convocatoria del VIII Premio a la Divulgación de la Matemática Aplicada, en el que os animo especialmente a participar. También S$\vec{e}$MA está involucrada en el Premio Addlink de Software Científico, dedicado en esta edición a trabajos realizados en entorno MATHEMATICA. Es buen momento para que los que desarrolláis software en este entorno aprovechéis la ocasión para presentaros.

Finalmente, me gustaría utilizar esta oportunidad para invitaros a todos a participar activamente en S$\vec{e}$MA, aportando ideas y posibles mejoras, pues ello será de gran ayuda para el progreso y desarrollo de nuestra Sociedad.

Un abrazo,

Carlos Vázquez Cendón
Presidente de S$\vec{e}$MA

carlosv@udc.es

Bol. Soc. Esp. Mat. Apl. n°37(2006), 11–46

Finite Element Methods for the Numerical Simulation of Incompressible Viscous Fluid Flow Modeled by the Navier-Stokes Equations. Part II

Roland Glowinski*, Tsorng-Whay Pan*,
L. Héctor Juárez V.+ and Edward Dean*

*Department of Mathematics, University of Houston, Houston, Texas 77204-3008, USA
+Departamento de Matematicas, Universidad Autonoma Metropolitana-Iztapalapa, Iztaplapa, D. F. 09340, MEXICO

5 Finite Element Approximation of the Navier-Stokes Equations

We have discussed in Section 2 the *time discretization* by operator-splitting of the *Navier-Stokes equations* modeling *incompressible viscous flow*, these equations being completed by convenient *initial* and *boundary conditions*. In order to implement on computers the solution methods described in Sections 2, 3, and 4, we still have to address the *space discretization* issue; in this note we will focus on *finite element methods*. There exists a quite large literature concerning the finite element approximation of the Navier-Stokes equations; concentrating on books, let us mention Temam 1977 [19] (Chapter 3), Thomasset 1981 [65], Peyret and Taylor 1982 [66] (Chapter 7), Glowinski 1984 [8] (Chapter 7), Girault and Raviart 1986 [59], Cuvelier, Segal and Van Steenhoven 1986 [67], Gunzburger 1989 [69], Pironneau 1989 [10], Fletcher 1991 [70, 71], Gunzburger and Nicolaides 1993 [72], Fortin 1993 [68], Quartapelle 1993 [73], Hebeker, Rannacher and Wittum 1994 [74], Quarteroni and Valli 1994 [42] (Chapter 13), Brenner and Scott 1994 [26] (Chapter 11), Marion and Temam 1998 [23], Gresho and Sani 1998 [75]; the above list is far from complete. The basic reference on the mathematical analysis of finite element approximations for the *steady Navier-Stokes* equations is still Girault and Raviart 1986 [59], to be completed by Fortin 1993 [68], where finite element approximations not available in Girault and Raviart 1986 [59] are discussed. To our knowledge,

Fecha de recepción: 21/10/2005

there is no book form analogue of Girault and Raviart 1986 [59], concerning the finite element approximation of the *time dependent Navier-Stokes equations*.

What about the mathematical analysis of solution methods for the Navier-Stokes equations, combining finite element approximations and time discretization by operator-splitting?

There is clearly an abundance of such methods and, indeed, most modern Navier-Stokes solvers use some form of time discretization by operator-splitting in order to treat the *incompressibility condition*. These splitting methods can be roughly divided in two families:

The *first family* of splitting methods for the Navier-Stokes equations is related to those methods described in Section 2. The convergence and stability properties of these methods are discussed in Fernandez-Cara and Beltran 1989 [76] and Kloucek and Rys 1994 [77], the last article discussing mainly the θ-scheme introduced in Section 2.

The *second family* is related to the splitting methods of Marchuk and Yanenko - also known as *fractional step methods* - for which basic references are Yanenko 1971 [27], Marchuk 1975 and 1990 [28, 29]. These methods have been applied to the solution of the Navier-Stokes equations for incompressible viscous fluid flow by, e.g., Chorin 1967 and 1968 [38, 39] and Temam 1969 [40, 41], the space discretization being by *finite differences* in the above references. A thorough discussion of fractional step methods for the Navier-Stokes equations can be found in Temam 1977 [19] (Chapter 3) and Marion and Temam 1998 [23] (Chapter 3) (see also the references therein).

In the present section, we shall focus on implementation issues when the θ-schemes of Section 2 are combined to *low-order finite element approximations* à la Bercovier-Pironneau (see Bercovier and Pironneau 1979 [79]) and *Hood-Taylor* (see Hood and Taylor 1973 [78]). In the later section, we will focus on implementation issues related to the splitting methods of Marchuk and Yanenko with the Bercovier-Pironneau finite element method. We are giving a special attention to the Hood and Taylor finite element methods for the following reasons:

(i) They are easy to implement, particularly in combination with the *time discretizations by operator-splitting* described in Section 2, the *least-squares/conjugate gradient algorithms* described in Section 3 and the *Stokes solvers* discussed in Section 4.

(ii) They are at the basis of some production codes for the simulation of incompressible viscous fluid flow, such as *N3S* developed by *Electricité de France* (EDF) and *FASTFLO* developed by the *CSIRO*, in Australia (N3S and FASTFLO are distributed by *SIMULOG* and *NAG*, respectively).

5.1 Finite element methods for the Stokes problem.

5.1.1 some observations.

It is a fairly general opinion that the main difficulty related to the space approximation of the Navier-Stokes equations, in the *pressure-velocity* formulation, is the treatment of the *incompressibility condition*

$$\boldsymbol{\nabla} \cdot \mathbf{u} = 0. \tag{280}$$

In order to show that the *boundary conditions* play also a role in these difficulties, let us consider first the *periodic Stokes problem*,

$$\begin{cases} \alpha \mathbf{u} - \nu \Delta \mathbf{u} + \boldsymbol{\nabla} p = \mathbf{f} \ in \ \Omega, \\ \boldsymbol{\nabla} \cdot \mathbf{u} = 0 \ in \ \Omega, \\ \mathbf{u}, \ \boldsymbol{\nabla} \mathbf{u} \ and \ p \ periodic \ at \ \Gamma, \end{cases} \tag{281}$$

with $\alpha > 0, \ \nu > 0, \ \Omega = (0,1)^d$ and $\Gamma = \partial\Omega$; in the present context, we say that a function v is *periodic at* Γ if

$$\begin{cases} v(x_1, ...x_{i-1}, 0, x_{i+1}, ...x_d) = v(x_1, ...x_{i-1}, 1, x_{i+1}, ...x_d), \\ \forall i = 1, ...d, \ \forall x_j \in (0,1), \ \forall j = 1, ...d, \ j \neq i. \end{cases} \tag{282}$$

Solving problem (281) is quite easy; we compute first the *pressure* p from

$$\begin{cases} \Delta p = \boldsymbol{\nabla} \cdot \mathbf{f} \ in \ \Omega, \\ p, \ \boldsymbol{\nabla} p \ periodic \ at \ \Gamma, \end{cases} \tag{283}$$

and then the velocity $\mathbf{u}$ from

$$\begin{cases} \alpha \mathbf{u} - \nu \Delta \mathbf{u} = \mathbf{f} - \boldsymbol{\nabla} p \ in \ \Omega, \\ \mathbf{u}, \ \boldsymbol{\nabla} \mathbf{u} \ periodic \ at \ \Gamma. \end{cases} \tag{284}$$

Suppose that $\mathbf{f}$ is *sufficiently smooth* and is also *periodic* at Γ; then, problems (283) and (284) are *well-posed* in $H^1(\Omega)/\mathbb{R}$ and $(H^1(\Omega))^d$, respectively. Now, denote $\boldsymbol{\nabla} \cdot \mathbf{u}$ by φ; it follows from (283), (284) that φ verifies

$$\begin{cases} \alpha \varphi - \nu \Delta \varphi = 0 \ in \ \Omega, \\ \varphi \ and \ \boldsymbol{\nabla} \varphi \ periodic \ at \ \Gamma, \end{cases} \tag{285}$$

whose unique solution is $\varphi = 0$, i.e., $\boldsymbol{\nabla} \cdot \mathbf{u} = 0$ *on* Ω. We have thus shown that problem (281) has a unique solution in $(H^1(\Omega))^d \times (H^1(\Omega)/\mathbb{R})$; this solution can be obtained via the solution of problems (283), (284) which are quite classical *elliptic* problems. Variational formulations for problems (283), (284) are given by

$$\begin{cases} p \in H^1_P(\Omega), \\ \displaystyle\int_\Omega \boldsymbol{\nabla} p \cdot \boldsymbol{\nabla} q dx = \int_\Omega \mathbf{f} \cdot \boldsymbol{\nabla} q dx, \ \forall q \in H^1_P(\Omega), \end{cases} \tag{286}$$

$$\begin{cases} \mathbf{u} \in (H^1_P(\Omega))^d, \\ \alpha \displaystyle\int_\Omega \mathbf{u} \cdot \mathbf{v} dx + \nu \int_\Omega \boldsymbol{\nabla} \mathbf{u} : \boldsymbol{\nabla} \mathbf{v} dx = \int_\Omega \mathbf{f} \cdot \mathbf{v} dx + \int_\Omega p \boldsymbol{\nabla} \cdot \mathbf{v} dx, \ \forall \mathbf{v} \in (H^1_P(\Omega))^d, \end{cases} \tag{287}$$

respectively, with, in (286), (287), H^1_P defined by

$$H^1_P(\Omega) = \{q | q \in H^1(\Omega), \ q \ periodic \ at \ \Gamma\}. \tag{288}$$

Solving problem (281), by *Galerkin type methods*, via the equivalent variational formulation (286), (287) is quite easy. We introduce first two families $\{P_h\}_h$ and $\{V_h\}_h$ of *finite dimensional spaces*; we suppose that these families verify

$$P_h \subset H^1_P(\Omega), \ \forall h, \ V_h \subset (H^1_P(\Omega))^d, \ \forall h, \tag{289}$$

$$\forall q \in H^1_P(\Omega), \ \exists \{q_h\}_h \ s.\ t.\ q_h \in P_h, \ \forall h, \lim_{h \to 0} ||q_h - q||_{H^1(\Omega)} = 0, \tag{290}$$

$$\forall \mathbf{v} \in (H^1_P(\Omega))^d, \ \exists \{\mathbf{v}_h\}_h \ s.\ t.\ \mathbf{v}_h \in V_h, \ \forall h, \lim_{h \to 0} ||\mathbf{v}_h - \mathbf{v}||_{(H^1(\Omega))^d} = 0. \tag{291}$$

Starting from the variational formulation (286), (287), we approximate problem (281) by

$$\begin{cases} p_h \in P_h, \\ \displaystyle\int_\Omega \boldsymbol{\nabla} p_h \cdot \boldsymbol{\nabla} q_h dx = \int_\Omega \mathbf{f}_h \cdot \boldsymbol{\nabla} q_h dx, \ \forall q_h \in P_h, \end{cases} \tag{292}$$

$$\begin{cases} \mathbf{u}_h \in V_h, \\ \displaystyle\int_\Omega (\alpha \mathbf{u}_h \cdot \mathbf{v}_h + \nu \boldsymbol{\nabla} \mathbf{u}_h : \boldsymbol{\nabla} \mathbf{v}_h) dx = \int_\Omega \mathbf{f}_h \cdot \mathbf{v}_h dx + \int_\Omega p_h \boldsymbol{\nabla} \cdot \mathbf{v}_h dx, \ \forall \mathbf{v}_h \in V_h, \end{cases} \tag{293}$$

where, in (292), (293), $\mathbf{f}_h$ is an approximation of $\mathbf{f}$ such that $\lim_{h \to 0} ||\mathbf{f}_h - \mathbf{f}||_{(L^2(\Omega))^d} = 0$.

It is a fairly easy exercise to prove that problems (292) and (293) are well-posed in $P_h/\mathbb{R}$ and V_h, respectively, and also that

$$\lim_{h \to 0} \{\mathbf{u}_h, p_h\} = \{\mathbf{u}, p\} \ in \ (H^1(\Omega))^{d+1}, \tag{294}$$

where, in (294), $\{\mathbf{u}, p\}$ is a solution of problem (281); to prove the convergence result (294) we can use the techniques discussed in, e.g., Strang and Fix 1973 [80], Ciarlet 1978 [24], Raviart and Thomas 1983 [37], Glowinski 1984 [8] (Appendix 1), Ciarlet 1991 [81] (Chapter 3) and Brenner and Scott 1994 [26] (Chapter 5).

From the above results, it appears that approximating the "periodic" Stokes problem (281) is a rather simple issue. Indeed, we can combine any *pressure* approximation to any *velocity* one, as long as properties (289)-(291) are verified. Thus, pressure and velocity approximations can be of different nature, use different meshes and/or basis functions, etc. On the other hand, as we shall

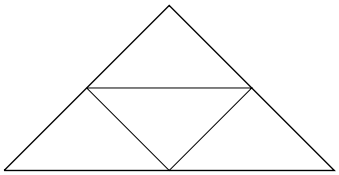


Figure 5.1: Dividing $T \in \mathcal{T}_h/$ to define $\mathcal{T}_{h/2}$

see in the following section, approximating the *Stokes-Dirichlet* problem

$$\begin{cases} \alpha \mathbf{u} - \nu \Delta \mathbf{u} + \boldsymbol{\nabla} p = \mathbf{f} \ in \ \Omega, \\ \boldsymbol{\nabla} \cdot \mathbf{u} = 0 \ in \ \Omega, \\ \mathbf{u} = \mathbf{g} \ on \ \Gamma \ (with \int_{\Gamma} \mathbf{g} \cdot \mathbf{n} d\Gamma = 0), \end{cases} \tag{295}$$

or the *Stokes-Neumann* problem

$$\begin{cases} \alpha \mathbf{u} - \nu \Delta \mathbf{u} + \boldsymbol{\nabla} p = \mathbf{f} \ in \ \Omega, \\ \boldsymbol{\nabla} \cdot \mathbf{u} = 0 \ in \ \Omega, \\ \nu \dfrac{\partial \mathbf{u}}{\partial n} - \mathbf{n} p = \mathbf{g} \ on \ \Gamma, \end{cases} \tag{296}$$

is a much more complicated matter, since *compatibility conditions* between the velocity and pressure approximations seem to be required if one wants to avoid spurious oscillations. In Glowinski 1991 [82] (Section 5.2), the mechanism producing numerical instabilities has been investigated on a particular case of the *Stokes-Dirichlet* problem (295) where $\Omega = (0,1) \times (0,1)$ via *Fourier Analysis*. To overcome these numerical instabilities we can use one of the following approaches

(a) *Use different type of approximations for pressure and velocity*

(b) *Use the same type of approximation for pressure and velocity, combined with a regularization procedure.*

Approach (a) is well known and will be further discussed in this section. The main idea here is to construct pressure spaces which are "poor" in high frequency modes, compared to the velocity space. Figure 5.1 suggests an obvious remedy to spurious oscillations which is to use a pressure grid which is *twice coarser* than the velocity one, and then use approximations of the same type on both grids. This observation makes sense for finite difference, finite element, spectral, pseudo-spectral, and wavelet approximations of problem (295); the well-known (and converging) finite element method (introduced in Bercovier and Pironneau 1979 [79]) obtained by using a *continuous piecewise linear* approximation of the *pressure* (resp., of the *velocity*) on a triangulation $\mathcal{T}_h$ (resp., $\mathcal{T}_{h/2}$, obtained from $\mathcal{T}_h$ by joining as shown in Figure 5.1 the midpoints in any $T \in \mathcal{T}_h$)

definitely follows the above rule. Beside the above reference, this method is discussed in, e.g., Glowinski 1984 [8] (Chapter 7), Glowinski 1985, 1986, 1991 [32, 33, 82], Bristeau, Glowinski, Mantel, Periaux, and Perrier 1985 [9], Girault and Raviart 1986 [59], Bristeau, Glowinski, and Periaux 1987 [83], Dean, Glowinski and Li 1989 [84], Pironneau 1989 [10], Gunzburger 1989 [69], Brezzi and Fortin 1991 [25], Glowinski and Pironneau 1992 [85], Fortin 1993 [68] (some of the above references show also numerical results obtained with it). Actually, the *Bercovier-Pironneau method* is a simple variation (easier to implement but less accurate) of the celebrated *Hood-Taylor method* (introduced in Hood and Taylor 1973 [78]) where *pressure* and *velocity* are approximated on the same triangulation by continuous approximations which are *piecewise linear* and *piecewise quadratic*, respectively.

Approach (b), introduced in Hughes, Franca and Balestra 1986 [86] (see also Douglas and Wang 1989 [87], Fortin 1993 [68], Cai and Douglas 1997 [88] and the references therein) leads essentially to *Tychonoff regularization procedures*, an obvious one being to "regularize" (one also says "stabilize") equation (216) by the following problem (written in *variational* form)

$$\begin{cases} p_\varepsilon \in H^1(\Omega), \\ \varepsilon \displaystyle\int_\Omega \boldsymbol{\nabla} p_\varepsilon \cdot \boldsymbol{\nabla} q dx + \int_\Omega (Ap_\varepsilon) q dx = -\int_\Omega \boldsymbol{\nabla} \cdot \mathbf{u}_0 q dx, \ \forall q \in H^1(\Omega), \end{cases} \tag{297}$$

where, in (297), ε is a *positive* parameter. Very good results have been obtained with approach (b) (see, e.g., Hughes, Franca and Balestra 1986 [86]), however, we prefer approach (a) for the following reasons:

(i) It is *parameter free*, unlike the second approach which requires the adjustment of the regularization parameter.

(ii) Quite often, the mesh size is adjusted, globally or locally, on the basis of the velocity behavior (boundary and shear layer thickness, for example). Therefore, compared to approach (a), approach (b) will be four times more costly (eight times for three-dimensional problems) from the pressure point of view, without further gains in accuracy.

(iii) Tychonoff regularization procedures are systematic methods for stabilizing ill-posed problems; in most cases, the adjustment of the regularization parameter is a delicate problem in itself, therefore, if there exist alternatives which are parameter free, we definitely think that the latter are preferable, particularly if they are based on an analysis of the mechanism producing the unwanted oscillations. Actually, the author of this article is a strong believer of Tychonoff regularization procedures when there is no alternative available to stabilize an ill-posed problem; indeed, we have been using such a procedure to solve *boundary control problems* for the *wave equation* (see Dean, Glowinski and Li 1989 [84], Glowinski, Li, and Lions 1990 [90]); however, as a consequence of our investigations concerning the Stokes problem, we have introduced, in

Glowinski and Li 1990 [89], new solution methods for the above control problems which are more efficient than those discussed in Dean, Glowinski and Li 1989 [84], and Glowinski, Li, and Lions 1990 [90] (results obtained with the new method are also shown in Glowinski and Lions 1995 [91]).

5.1.2 Discrete spaces.

We suppose that Ω is a bounded polygonal domain of $\mathbb{R}^2$ (cases where domain has curved boundary have been discussed in Glowinski 2003 [4] (Chapter 5)). With $\mathcal{T}_h$ a standard finite element *triangulation* of Ω (see, e.g., Ciarlet 1978 and 1991 [24, 81], Raviart and Thomas 1983 [37], Glowinski 1984 [8] (Appendix 1) for this notion) and h the maximal length of the edges of $\mathcal{T}_h$, we introduce the following discrete spaces (with P_k the space of the polynomials in two variables of degree $\leq k$):

$$P_h = \{q_h | q_h \in C^0(\overline{\Omega}),\ q_h|_T \in P_1,\ \forall T \in \mathcal{T}_h\}, \tag{298}$$

$$V_h = \{\mathbf{v}_h | \mathbf{v}_h \in (C^0(\overline{\Omega}))^2, \mathbf{v}_h|_T \in (P_2)^2,\ \forall T \in \mathcal{T}_h\}. \tag{299}$$

If the *boundary conditions* imply $\mathbf{u} = \mathbf{g}_0$ on Γ_0, we shall need the space V_{0h} defined by

$$V_{0h} = \{\mathbf{v}_h | \mathbf{v}_h \in V_h,\ \mathbf{v}_h = \mathbf{0}\ on\ \Gamma\},\ if\ \Gamma_0 = \Gamma, \tag{300}$$

and by

$$V_{0h} = \{\mathbf{v}_h | \mathbf{v}_h \in V_h,\ \mathbf{v}_h = \mathbf{0}\ on\ \Gamma_0\},\ if\ \int_{\Gamma_0} d\Gamma > 0, \Gamma_0 \neq \Gamma; \tag{301}$$

if we are in the situation associated with (300) it is of fundamental importance to have the points at the interface of Γ_0 and $\Gamma_1 (= \Gamma \backslash \Gamma_0)$ as vertices of $\mathcal{T}_h$.

Another useful variant of V_h (and then V_{0h}), the *Bercovier-Pironneau* velocity space, is obtained as follows:

$$V_h = \{\mathbf{v}_h | \mathbf{v}_h \in (C^0(\overline{\Omega}))^2,\ \mathbf{v}_h|_T \in (P_1)^2,\ \forall T \in \mathcal{T}_{h/2}\}. \tag{302}$$

In (302), $\mathcal{T}_{h/2}$ is (as in previous subsection) the triangulation of Ω obtained from $\mathcal{T}_h$ by joining the mid-points of the edges of $T \in \mathcal{T}_h$ (see Figure 5.1); for the same triangulation $\mathcal{T}_h$, we have the same global number of degrees of freedom if we use V_h defined by either (299) or (302), space P_h being the same; however, the matrices encountered in the second case are more compact and sparse.

Remark 1 *For obvious reasons, the finite element approximations of the Stokes problem based on the pair $\{P_h, V_h\}$ defined by (298), (299) (resp., (298), (302)) is called a P_1/P_2 approximation (resp., a P_1-iso-P_2/P_1 approximation).*

5.1.3 Approximation of the boundary conditions.

If the boundary conditions are defined by

$$\mathbf{u} = \mathbf{g}\ on\ \Gamma,\ with\ \int_\Gamma \mathbf{g} \cdot \mathbf{n} d\Gamma = 0, \tag{303}$$

it is of fundamental importance to approximate $\mathbf{g}$ by $\mathbf{g}_h$ so that

$$\int_\Gamma \mathbf{g}_h \cdot \mathbf{n} d\Gamma = 0. \tag{304}$$

We shall discuss the simple case where Ω is *polygonal* domain. For the curved boundary cases, please follow the methods discussed in Glowinski 2003 [3] (Chapter 5).

We suppose that $\mathbf{g}$ is *continuous* on Γ. Then we have that $\mathbf{n}$ will be *piecewise constant* on Γ. Starting from V_h defined by either (299) or (302), we define the *boundary space* γV_h by

$$\gamma V_h = \{\boldsymbol{\mu}_h | \boldsymbol{\mu}_h = \mathbf{v}_h|_\Gamma,\ \mathbf{v}_h \in V_h\}, \tag{305}$$

i.e., γV_h is the space of the traces on Γ of the functions $\mathbf{v}_h$ belonging to V_h. Actually, if V_h is defined by (299), γV_h is also the space of the functions *continuous* over Γ, taking their values in $\mathbb{R}^2$ and *quadratic* over the edges of $\mathcal{T}_h$ contained in Γ; similarly, if V_h is defined by (302) we have

$$\gamma V_h = \{\boldsymbol{\mu}_h | \boldsymbol{\mu}_h \in (C^0(\Gamma))^2,\ \boldsymbol{\mu}_h \textit{ is affine over the edges of } \mathcal{T}_{h/2} \textit{ contained in } \Gamma\}.$$

Our problem is to construct an approximation $\mathbf{g}_h$ of $\mathbf{g}$ such that

$$\mathbf{g}_h \in \gamma V_h,\ \int_\Gamma \mathbf{g}_h \cdot \mathbf{n} d\Gamma = 0. \tag{306}$$

If $\pi_h \mathbf{g}$ is the unique element of γV_h, obtained by piecewise linear or piecewise quadratic interpolation of $\mathbf{g}$ over Γ, i.e., obtained from the values taken by $\mathbf{g}$ at those vertices of $\mathcal{T}_h$ (or $\mathcal{T}_{h/2}$) belonging to Γ, we usually have $\int_\Gamma \pi_h \mathbf{g} \cdot \mathbf{n} d\Gamma \neq 0$. To overcome this difficulty we may proceed as follows:

(i) We define an approximation $\mathbf{n}_h$ of $\mathbf{n}$ as the solution of the following *linear variational problem* in γV_h

$$\begin{cases} \mathbf{n}_h \in \gamma V_h, \\ \displaystyle\int_\Gamma \mathbf{n}_h \cdot \boldsymbol{\mu}_h d\Gamma = \int_\Gamma \mathbf{n} \cdot \boldsymbol{\mu}_h d\Gamma,\ \forall \boldsymbol{\mu}_h \in \gamma V_h. \end{cases} \tag{307}$$

Problem (307) is *equivalent* to a linear system whose matrix is *sparse, symmetric positive definite*, very *well-conditioned* and *easy to compute* (also, problem (307) needs to be solved *only once* if the finite element mesh is *fixed*). Observe also that the fact that $\mathbf{n}$ is constant, on each edge of $\mathcal{T}_h$ contained in Γ, makes the calculation of the right hand side of the above equivalent linear system very easy (the details can be found in Glowinski 2003 [4] (Chapter 5)).

(ii) Define $\mathbf{g}_h$ by

$$\mathbf{g}_h = \pi_h \mathbf{g} - \left(\int_\Gamma \pi_h \mathbf{g} \cdot \mathbf{n} d\Gamma / \int_\Gamma \mathbf{n} \cdot \mathbf{n}_h d\Gamma \right)\ \mathbf{n}_h. \tag{308}$$

It is easy to check that (307), (308) imply that $\mathbf{g}_h$ verifies the *flux condition* (306).

5.1.4 Formulation of the discrete Stokes problem.

In the following we shall denote by Ω_h the *computational domain* and by Γ_h its boundary even though we have considered here that Ω is *polygonal* and hence $\Omega_h = \Omega$ and $\Gamma_h = \Gamma$.

The Dirichlet case.

The *Stokes problem*, considered here, has the following formulation:

$$\begin{cases} \alpha \mathbf{u} - \nu \Delta \mathbf{u} + \boldsymbol{\nabla} p = \mathbf{f} \ in \ \Omega, \\ \boldsymbol{\nabla} \cdot \mathbf{u} = 0 \ in \ \Omega, \\ \mathbf{u} = \mathbf{g} \ on \ \Gamma, \end{cases} \tag{309}$$

with $\mathbf{f} \in (H^{-1}(\Omega))^d$ and $\mathbf{g} \in (H^{1/2}(\Gamma))^d$, with $\displaystyle\int_\Gamma \mathbf{g} \cdot \mathbf{n} d\Gamma = 0$. It follows from Section 4, that problem (309) has a *unique solution* in $V_g \times (L^2(\Omega)/\mathbb{R})$, with

$$V_g = \{\mathbf{v} | \mathbf{v} \in (H^1(\Omega))^d, \ \mathbf{v} = \mathbf{g} \ on \ \Gamma\}. \tag{310}$$

Problem (309) can also be formulated as

$$\begin{cases} \mathbf{u} \in V_g, \ p \in L^2(\Omega), \\ \alpha \displaystyle\int_\Omega \mathbf{u} \cdot \mathbf{v} dx + \nu \int_\Omega \boldsymbol{\nabla} \mathbf{u} : \boldsymbol{\nabla} \mathbf{v} dx - \int_\Omega p \boldsymbol{\nabla} \cdot \mathbf{v} dx = < \mathbf{f}, \mathbf{v} >, \ \forall \mathbf{v} \in V_0, \\ \displaystyle\int_\Omega q \boldsymbol{\nabla} \cdot \mathbf{u} dx = 0, \ \forall q \in L^2(\Omega), \end{cases} \tag{311}$$

where, in (311), the *test function space* V_0 is defined by

$$V_0 = (H_0^1(\Omega))^d, \tag{312}$$

and where $< \cdot, \cdot >$ denotes the duality pairing between $(H^{-1}(\Omega))^d$ and V_0.

Next, let us define V_{0h} and V_{g_h} by

$$V_{0h} = \{\mathbf{v}_h | \mathbf{v}_h \in V_h, \ \mathbf{v}_h = \mathbf{0} \ on \ \Gamma_h\} \tag{313}$$

$$V_{g_h} = \{\mathbf{v}_h | \mathbf{v}_h \in V_h, \ \mathbf{v}_h = \mathbf{g}_h \ on \ \Gamma_h\} \tag{314}$$

with, in (313) and (314), V_h and $\mathbf{g}_h$ defined as in Sections 5.1.2 and 5.1.3, respectively; we have in particular $\displaystyle\int_{\Gamma_h} \mathbf{g}_h \cdot \mathbf{n} \, d\Gamma_h = 0$. We approximate the *Stokes-Dirichlet problem* (309) by

$$\begin{cases} \mathbf{u}_h \in V_{g_h}, \ p_h \in P_h, \\ \alpha \displaystyle\int_{\Omega_h} \mathbf{u}_h \cdot \mathbf{v}_h dx + \nu \int_{\Omega_h} \boldsymbol{\nabla} \mathbf{u}_h : \boldsymbol{\nabla} \mathbf{v}_h dx - \int_{\Omega_h} p_h \boldsymbol{\nabla} \cdot \mathbf{v}_h dx \\ \qquad = < \mathbf{f}_h, \mathbf{v}_h >_h, \ \forall \mathbf{v}_h \in V_{0h}, \\ \displaystyle\int_{\Omega_h} q_h \boldsymbol{\nabla} \cdot \mathbf{u}_h dx = 0, \ \forall q_h \in P_h \end{cases} \tag{315}$$

with the space P_h as in Sections 5.1.2; in (315) $\mathbf{f}_h$ is an approximation of $\mathbf{f}$ and $< \cdot, \cdot >_h$ denotes the duality pairing between $(H^{-1}(\Omega_h))^d$ and $(H_0^1(\Omega_h))^d$. The *well-posedness* of problem (315) will be addressed in next subsection, which will contain also some comments on the *convergence* of the pair $\{\mathbf{u}_h, p_h\}$ as $h \to 0$.

The case of the mixed boundary conditions.

The *Stokes problem*, considered now, has the following formulation

$$\begin{cases} \alpha\mathbf{u} - \nu\Delta\mathbf{u} + \boldsymbol{\nabla} p = \mathbf{f} \ in \ \Omega, \\ \boldsymbol{\nabla} \cdot \mathbf{u} = 0 \ in \ \Omega, \\ \mathbf{u} = \mathbf{g}_0 \ on \ \Gamma_0, \ \nu \dfrac{\partial \mathbf{u}}{\partial n} - \mathbf{n}p = \mathbf{g}_1 \ on \ \Gamma_1. \end{cases} \tag{316}$$

In order to avoid too many technicalities we shall assume that, in (316), we have $\mathbf{g}_0 = \tilde{\mathbf{g}}_0|_{\Gamma_0}$ with $\tilde{\mathbf{g}}_0 \in (H^1(\Omega))^d$, $\mathbf{g}_1 \in (L^2(\Gamma_1))^d$, and $\mathbf{f} \in (L^2(\Omega))^d$. A *variational formulation* of problem (316) is given by

$$\begin{cases} \mathbf{u} \in V_{g_0}, \ p \in L^2(\Omega), \\ \alpha \displaystyle\int_\Omega \mathbf{u} \cdot \mathbf{v} dx + \nu \int_\Omega \boldsymbol{\nabla}\mathbf{u} : \boldsymbol{\nabla}\mathbf{v} dx - \int_\Omega p \boldsymbol{\nabla} \cdot \mathbf{v} dx \\ \qquad = \displaystyle\int_\Omega \mathbf{f} \cdot \mathbf{v} dx + \int_{\Gamma_1} \mathbf{g}_1 \cdot \mathbf{v} d\Gamma, \forall \mathbf{v} \in V_0, \\ \displaystyle\int_\Omega q \boldsymbol{\nabla} \cdot \mathbf{u} dx = 0, \ \forall q \in L^2(\Omega), \end{cases} \tag{317}$$

where, in (317), V_{g_0} and V_0 are defined by

$$V_{g_0} = \{\mathbf{v} | \mathbf{v} \in (H^1(\Omega))^d, \ \mathbf{v} = \mathbf{g}_0 \ on \ \Gamma_0\}, \tag{318}$$

$$V_0 = \{\mathbf{v} | \mathbf{v} \in (H^1(\Omega))^d, \ \mathbf{v} = \mathbf{0} \ on \ \Gamma_0\}, \tag{319}$$

respectively; if $\Gamma_0 = \emptyset$, then $V_0 = V_{g_0} = (H^1(\Omega))^d$ and $\Gamma_1 = \Gamma$.

Following (317) we approximate the *Stokes problem* (316) by

$$\begin{cases} \mathbf{u}_h \in V_{g_{0h}}, \ p_h \in P_h; \ \forall \mathbf{v}_h \in V_{0h} \ and \ q_h \in P_h, \ we \ have \\ \alpha \displaystyle\int_{\Omega_h} \mathbf{u}_h \cdot \mathbf{v}_h dx + \nu \int_{\Omega_h} \boldsymbol{\nabla}\mathbf{u}_h : \boldsymbol{\nabla}\mathbf{v}_h dx - \int_{\Omega_h} p_h \boldsymbol{\nabla} \cdot \mathbf{v}_h dx \\ \qquad = \displaystyle\int_{\Omega_h} \mathbf{f}_h \cdot \mathbf{v}_h dx + \int_{\Gamma_{1h}} \mathbf{g}_{1h} \cdot \mathbf{v}_h d\Gamma_h, \\ \displaystyle\int_{\Omega_h} q_h \boldsymbol{\nabla} \cdot \mathbf{u}_h dx = 0. \end{cases} \tag{320}$$

In (320) the space P_h is defined as in Section 5.1.2, while

$$V_{g_{0h}} = \{\mathbf{v}_h | \mathbf{v}_h \in V_h, \ \mathbf{v}_h = \mathbf{g}_{0h} \ on \ \Gamma_{0h}\}, \tag{321}$$

$$V_{0h} = \{\mathbf{v}_h | \mathbf{v}_h \in V_h, \ \mathbf{v}_h = \mathbf{0} \ on \ \Gamma_{0h}\} \tag{322}$$

with V_h defined as in Section 5.1.2. The functions $\mathbf{f}_h$, $\mathbf{g}_{0h}$ and $\mathbf{g}_{1h}$ are approximations of $\mathbf{f}$, $\mathbf{g}_0$ and $\mathbf{g}_1$, respectively; Γ_{ih} approximates $\Gamma_i, \forall \ i = 0, 1$.

5.1.5 On the convergence of the finite element approximations of the Stokes problem.

In this subsection, we are going to discuss the *convergence* - as $h \to 0$ - of the *finite element approximations* of the Stokes problems, introduced in the preceding paragraphs. *Convergence* is, at the same time, a delicate and well-documented issue. It is our opinion that the celebrated article by Crouzeix and Raviart 1973 [92] was really the first one to address the convergence issues in a systematic, rigorous and general way; also, this article introduced novel (at the time) approximations of the Stokes problem which are still used nowadays by some practitioners. A very complete discussion of the convergence properties of various finite element approximations to the Stokes and steady Navier-Stokes equations can be found in the book by Girault and Raviart 1986 [59], which is still a basic (if not *the* basic) reference on the subject. However, the reader should also consult Brezzi and Fortin 1991 [25] (Chapter 6) and the review article by Fortin 1993 [68] which discusses - among other things - finite element approximations of the Stokes and Navier-Stokes equations not available in the mid-eighties (the following references are also worth consulting: Temam 1977 [19] (Chapter 1), Glowinski 1984 [8] (Chapter 7 and Appendix 3), Gunzburger 1989 [69] (Part 1), Pironneau 1989 [10] (Chapter 4), Brenner and Scott 1994 [26] (Chapter 10)).

For simplicity, in the following we shall consider only the *Stokes-Dirichlet problem* with $\mathbf{g} = \mathbf{0}$ *on* Γ; we have then, from (309),

$$\begin{cases} \alpha\mathbf{u} - \nu\Delta\mathbf{u} + \boldsymbol{\nabla} p = \mathbf{f} \ in \ \Omega, \\ \boldsymbol{\nabla} \cdot \mathbf{u} = 0 \ on \ \Omega, \\ \mathbf{u} = \mathbf{0} \ on \ \Gamma. \end{cases} \tag{323}$$

A *variational formulation* of problem (323) is given by

$$\begin{cases} \mathbf{u} \in V_0, \ p \in L^2(\Omega), \\ \alpha \displaystyle\int_\Omega \mathbf{u} \cdot \mathbf{v} dx + \nu \int_\Omega \boldsymbol{\nabla}\mathbf{u} : \boldsymbol{\nabla}\mathbf{v} dx - \int_\Omega p \boldsymbol{\nabla} \cdot \mathbf{v} dx = < \mathbf{f}, \mathbf{v} >, \ \forall \mathbf{v} \in V_0, \\ \displaystyle\int_\Omega q \boldsymbol{\nabla} \cdot \mathbf{u} dx = 0, \ \forall q \in L^2(\Omega), \end{cases} \tag{324}$$

with $V_0 = (H_0^1(\Omega))^d$ and $< \cdot, \cdot >$ the duality pairing between $(H^{-1}(\Omega))^d$ and $(H_0^1(\Omega))^d$. We know from Section 4, that problem (323), (324) is equivalent to the following saddle-point problem in $V_0 \times L^2(\Omega)$:

$$\begin{cases} Find \ \{\mathbf{u}, p\} \in V_0 \times L^2(\Omega), \ so \ that \\ \mathcal{L}(\mathbf{u}, q) \leq \mathcal{L}(\mathbf{u}, p) \leq \mathcal{L}(\mathbf{v}, p), \ \forall \{\mathbf{v}, q\} \in V_0 \times L^2(\Omega), \end{cases} \tag{325}$$

with the *Lagrangian functional* $\mathcal{L}$ defined, $\forall \{\mathbf{v}, q\} \in (H^1(\Omega))^d \times L^2(\Omega)$, by

$$\mathcal{L}(\mathbf{v}, q) = \frac{1}{2} \int_\Omega (\alpha |\mathbf{v}|^2 + \nu |\boldsymbol{\nabla}\mathbf{v}|^2) dx - \int_\Omega q \boldsymbol{\nabla} \cdot \mathbf{v} dx - < \mathbf{f}, \mathbf{v} > . \tag{326}$$

The saddle-point problem (324), (325) is a member of the following family of *generalized linear saddle-point problems*

$$\begin{cases} Find\ \{u,\lambda\} \in X \times \Lambda\ so\ that \\ a(u,v) + b(v,\lambda) =< l, v >,\ \forall v \in X, \\ b(u,\mu) =< \mathcal{X}, \mu >,\ \forall \mu \in \Lambda, \end{cases} \tag{327}$$

where, in (327):

- X and Λ are two *real Hilbert spaces*, with X' and Λ' their respective *dual spaces*;
- $a : X \times X \to \mathbb{R}$ is *bilinear* and *continuous* (possibly non-symmetric);
- $b : X \times \Lambda \to \mathbb{R}$ is *bilinear* and *continuous*,
- $< \cdot, \cdot >$ denotes the *duality pairing* between either X' and X or Λ' and Λ,
- $l \in X'$ and $\mathcal{X} \in \Lambda'$.

Using the *Riesz Theorem* we can associate to the bilinear functionals $a(\cdot,\cdot)$ and $b(\cdot,\cdot)$ two operators A and B so that

$$\begin{cases} A \in \mathcal{L}(X, X'), \\ < Av, w >= a(v,w),\ \forall v, w \in X, \end{cases}$$

$$\begin{cases} B \in \mathcal{L}(\Lambda, \Lambda'), \\ < Bv, \mu >= b(v,\mu),\ \forall v \in X,\ \forall \mu \in \Lambda. \end{cases}$$

The generalized saddle-point problem (327) takes then the equivalent operator formulation

$$\begin{cases} Au + B'\lambda = l, \\ Bu = \mathcal{X}, \end{cases} \tag{328}$$

where, in (328), $B' (\in \mathcal{L}(\Lambda, X'))$ is the dual (transpose) operator of B, i.e.

$$< Bv, \mu >=< B'\mu, v >,\ \forall \{v,\mu\} \in X \times \Lambda.$$

Remark 2 *If the bilinear functional $a(\cdot,\cdot)$ is symmetric problem (327), (328) is equivalent to the genuine saddle-point problem*

$$\begin{cases} \{u,\lambda\} \in X \times \Lambda, \\ L(u,\mu) \le L(u,\lambda) \le L(v,\lambda),\ \forall \{v,\mu\} \in X \times \Lambda, \end{cases} \tag{329}$$

with, in (329), the Lagrangian L defined by

$$L(v,\mu) = \frac{1}{2}a(v,v) + b(v,\mu) - < \mathcal{X}, \mu > - < l, v >,\ \forall \{v,\mu\} \in X \times \Lambda.$$

We can also easily show (using the techniques employed in Section 4) that the vector u in (329) is also the solution of the following *constrained minimization* problem

$$\begin{cases} u \in V(\mathcal{X}), \\ j(u) \leq j(v), \ \forall v \in V(\mathcal{X}), \end{cases} \tag{330}$$

with, in (330), the functional $j(\cdot)$ and the space $V(\mathcal{X})$ defined by

$$\begin{aligned} j(v) &= \frac{1}{2}a(v,v) - < l, v >, \ \forall v \in X, \\ V(\mathcal{X}) &= \{v | v \in X, \ b(v,\mu) = < \mathcal{X}, \mu >, \ \forall \mu \in \Lambda\}, \end{aligned}$$

respectively (we clearly have, for $V(\mathcal{X})$, the alternative definition

$$V(\mathcal{X}) = \{v | v \in X, Bv = \mathcal{X}\}).$$

Vector λ can be seen as a *Lagrange multiplier* associated with the linear relation $Bv = \mathcal{X}$.

Remark 3 *We can easily show that the component u of the solution of problem (327), (328) is also a solution of the following linear variational problem in $V(\mathcal{X})$ ($V(\mathcal{X})$ has been defined in the above remark):*

$$\begin{cases} u \in V(\mathcal{X}), \\ a(u,v) = < l, v >, \ \forall v \in V_0, \end{cases} \tag{331}$$

where $V_0 = \ ker(B)$, i.e.

$$V_0 = \{v | v \in X, \ b(v,\mu) = 0, \ \forall \mu \in \Lambda\}.$$

□

With space V_0 still being the *kernel* of operator B, let us define $\pi \in \mathcal{L}(X', V_0')$ by

$$< \pi f, v > = < f, v >, \ \forall f \in X', \ \forall v \in V_0.$$

Concerning the *uniqueness* and the *existence* of a solution to problem (327), (328) we have the following

Theorem 1 *Problem (327), (328) is well-posed (i.e., operator* $\begin{pmatrix} A & B' \\ B & 0 \end{pmatrix}$ *is an isomorphism from $X \times \Lambda$ onto $X' \times \Lambda'$) if and only if the following conditions hold:*

(i) operator πA is an isomorphism from V_0 onto V_0';

(ii) there exists a constant $\beta > 0$ such that

$$\inf_{\mu\in\Lambda\setminus\{0\}} \sup_{v\in X\setminus\{0\}} \frac{b(v,\mu)}{||v||_X||\mu||_\Lambda} \geq \beta. \tag{332}$$

(Condition (332) is known as an inf-sup condition).

For a proof of Theorem 1 see, e.g., Girault and Raviart 1986 [59] (Chapter 1, Section 4); actually in the above reference one can also find a proof of the following

Corollary 2 *Suppose that the bilinear functional $a(\cdot,\cdot)$ is V-elliptic, i.e., there exists a constant $\alpha > 0$ such that*

$$a(v,v) \geq \alpha||v||_X^2, \ \forall v \in X.$$

Then, problem (327), (328) is well-posed if and only if the bilinear functional $b(\cdot,\cdot)$ satisfies the inf-sup condition (332).

Before going further, we think that it may be worthwhile to check if either Theorem 1 or Corollary 2 apply to the solution of the Stokes-Dirichlet problem (323), (324); it is indeed the case as shown by the following

Corollary 3 *It follows from Corollary 2 that problem (323), (324) has a unique solution in $(H_0^1(\Omega))^d \times L_0^2(\Omega)$, where*

$$L_0^2(\Omega) = \{q | q \in L^2(\Omega), \ \int_\Omega q dx = 0\}.$$

PROOF: The above result has been shown already in Section 4. The other proof can be obtained as a direct consequence of Corollary 2 (see Glowinski 2003 [4] (Chapter 5)).

Let us discuss now the *approximation* of the *generalized saddle-point problem* (327). With h a *discretization parameter*, we introduce two finite-dimensional spaces X_h and Λ_h, so that

$$X_h \subset X \text{ and } \Lambda_h \subset \Lambda. \tag{333}$$

Next, to each $\mathcal{X} \in \Lambda'$ we associate $V_h(\mathcal{X})$ - a discrete analogue of $V(\mathcal{X})$ - defined by

$$V_h(\mathcal{X}) = \{v_h | v_h \in X_h, \ b(v_h,\mu_h) =< \mathcal{X}, \mu_h >, \ \forall \mu_h \in \Lambda_h\}, \tag{334}$$

and we denote $V_h(0)$ by V_{0h}, i.e.

$$V_{0h} = \{v_h | v_h \in X_h, b(v_h,\mu_h) = 0, \ \forall \mu_h \in \Lambda_h\}. \tag{335}$$

We observe that, in general, $V_h(\mathcal{X}) \not\subset V(\mathcal{X})$ and $V_{0h} \not\subset V_0$ (with V_0 as in Remark 3).

We approximate, then, problem (327) by

$$\begin{cases} Find \ \{u_h,\lambda_h\} \in X_h \times \Lambda_h \ so \ that \\ a(u_h,v_h) + b(v_h,\lambda_h) =< l, v_h >, \ \forall v_h \in X_h, \\ b(u_h,\mu_h) =< \mathcal{X}, \mu_h >, \ \forall \mu_h \in \Lambda_h. \end{cases} \tag{336}$$

If $\{u_h, \lambda_h\}$ is a solution of problem (336), we can easily show that u_h is also a solution of the following finite dimensional linear variational problem

$$\begin{cases} u_h \in V_h(\mathcal{X}), \\ a(u_h, v_h) =< l, v_h >, \ \forall v_h \in V_{0h}; \end{cases} \tag{337}$$

problem (337) is clearly a discrete analogue of problem (331). Define now the norms $||a||$ and $||b||$ of the bilinear functionals $a(\cdot,\cdot)$ and $b(\cdot,\cdot)$ by

$$||a|| = \sup \frac{|a(v,w)|}{||v||_X ||w||_X}, \ \{v,w\} \in (X\backslash\{0\})^2 \tag{338}$$

and

$$||b|| = \sup \frac{|b(v,\mu)|}{||v||_X ||\mu||_\Lambda}, \{v,\mu\} \in (X\backslash\{0\}) \times (\Lambda\backslash\{0\}), \tag{339}$$

respectively; concerning the approximation of the solutions $\{u, \lambda\}$ of problem (327) by the solutions $\{u_h, \lambda_h\}$ of problem (336), we have then the following

Theorem 4 *1. Assume that the following conditions are verified*

(i) space $V_h(\mathcal{X})$ is not empty;

(ii) there exists a positive constant α^* such that

$$a(v_h, v_h) \geq \alpha^* ||v_h||_X^2, \ \forall v_h \in V_{0h}. \tag{340}$$

Then, problem (337) has a unique solution $u_h \in V_h(\mathcal{X})$ and there exists a constant C_1 depending only of $\alpha^*, ||a||$ and $||b||$ such that the following error estimate holds:

$$||u - u_h||_X \leq C_1 \left(\inf_{v_h \in V_h(\mathcal{X})} ||u - v_h||_X + \inf_{\mu_h \in \Lambda_h} ||\lambda - \mu_h||_\Lambda \right). \tag{341}$$

2. Assume that hypothesis (ii) holds and, in addition, that:

(iii) there exists a positive constant β^* such that

$$\inf_{\mu_h \in \Lambda_h \backslash \{0\}} \sup_{v_h \in X_h \backslash \{0\}} \frac{b(v_h, \mu_h)}{||v_h||_X ||\mu_h||_\Lambda} \geq \beta^*. \tag{342}$$

Then, $V_h(\mathcal{X}) \neq \emptyset$ and there exists a unique λ_h in Λ_h such that $\{u_h, \lambda_h\}$ is the unique solution of problem (336). Furthermore, there exists a constant C_2, depending only of $\alpha^*, \beta^*, ||a||$ and $||b||$, such that

$$||u - u_h||_X + ||\lambda - \lambda_h||_\Lambda \leq C_2 (\inf_{v_h \in X_h} ||u - v_h||_X + \inf_{\mu_h \in \Lambda_h} ||\lambda - \mu_h||_\Lambda). \tag{343}$$

For a proof of the above theorem, see Girault and Raviart 1986 [59] (pp. 114-116) (see also Roberts and Thomas 1991 [93] (Chapter 3) and Brezzi and Fortin 1991 [25] (Chapter 2); actually, the two above references contain a discussion of the effects of *numerical integration* on the error estimates, a most important practical issue).

Before discussing the convergent results, we have to introduce some (fairly classical) definitions, namely:

Definition 5.1: A family $\{\mathcal{T}_h\}_h$ of *triangulations* of Ω is said to be *regular* if there exists $\theta_0, 0 < \theta_0 \leq \pi/3$, such that

$$\theta_T \geq \theta_0, \ \forall T \in \mathcal{T}_h, \ \forall h, \tag{344}$$

where, in (344), θ_T is the *smallest angle* of triangle T.

Definition 5.2: A family $\{\mathcal{T}_h\}_h$ of *triangulations* of Ω is said to be *uniformly regular* if it is *regular* and if there exists $\sigma, \ \sigma \geq 1$, such that

$$\max_{T \in \mathcal{T}_h} h_T / \min_{T \in \mathcal{T}_h} h_T \leq \sigma, \ \forall h, \tag{345}$$

where, in (345), h_T is the length of the largest edge(s) of triangle T.

Remark 4 *In Definitions 5.1 and 5.2, we have been assuming that Ω is a polygonal domain of $\mathbb{R}^2$ such that $\overline{\Omega} = \bigcup_{T \in \mathcal{T}_h} T, \forall h$. Actually, the above two definitions can be generalized to two-dimensional domains with curved boundaries and also to three-dimensional domains with curved or polyhedral boundaries, as shown in, e.g., Ciarlet 1991 [81] (Chapter 6, Section 37). If Ω is a polyhedral domain of $\mathbb{R}^3$ and $\mathcal{T}_h$ a "triangulation" of Ω (i.e., $T \in \mathcal{T}_h \Rightarrow T$ is a tetrahedron) so that $\overline{\Omega} = \bigcup_{T \in \mathcal{T}_h} T$, we say that the family $\{\mathcal{T}_h\}_h$ is regular if there exists $\sigma_1 > 0$ such that*

$$h_T / \rho_T \leq \sigma_1, \ \forall T \in \mathcal{T}_h, \ \forall h, \tag{346}$$

with h_T the length of the largest edge(s) of tetrahedron T, and ρ_T the diameter of the sphere inscribed in T. Similarly, we say that the family $\{\mathcal{T}_h\}_h$ is uniformly regular if it is regular and if there exists $\sigma_2, \sigma_2 \geq 1$, so that

$$\max_{T \in \mathcal{T}_h} h_T / \min_{T \in \mathcal{T}_h} h_T \leq \sigma_2, \ \forall T \in \mathcal{T}_h, \ \forall h, \tag{347}$$

with h_T as just above.

Following Girault and Raviart 1986 [59] (Chapter 2, Section 4), and Brezzi and Fortin 1991 [25] (Chapter 6), we are going to provide convergence results for *finite element approximations* of the *Stokes-Dirichlet problem* (323). We shall focus our attention on the *Hood-Taylor* and *Bercovier-Pironneau* approximations described in Section 5.1.2; convergence results concerning other finite element approximations of the Stokes problem can be found in, e.g., the two above references and in Fortin 1993 [68] (see also the references therein).

Since Ω is polygonal it follows from Section 5.1.2 that the Stokes problem (323) is approximated by

$$\begin{cases} \{\mathbf{u}_h, p_h\} \in V_{0h} \times P_h;\ \forall \{\mathbf{v}, q_h\} \in V_{0h} \times P_h \text{ we have} \\ \alpha \displaystyle\int_\Omega \mathbf{u}_h \cdot \mathbf{v}_h dx + \nu \int_\Omega \boldsymbol{\nabla} \mathbf{u}_h : \boldsymbol{\nabla} \mathbf{v}_h dx - \int_\Omega p_h \boldsymbol{\nabla} \cdot \mathbf{v}_h dx = < \mathbf{f}, \mathbf{v}_h >, \\ \displaystyle\int_\Omega \boldsymbol{\nabla} \cdot \mathbf{u}_h q_h dx = 0, \end{cases} \tag{348}$$

with

$$P_h = \{q_h | q_h \in C^0(\overline{\Omega}), q_h|_T \in P_1,\ \forall T \in \mathcal{T}_h\} \tag{349}$$

and

$$V_{0h} = \{\mathbf{v}_h | \mathbf{v}_h \in (C^0(\overline{\Omega}))^2,\ \mathbf{v}_h|_T \in (P_2)^2,\ \forall T \in \mathcal{T}_h,\ \mathbf{v}_h = 0 \text{ on } \Gamma\}. \tag{350}$$

In Girault and Raviart 1986 [59] (Chapter 2, Section 4.2), it is shown that the approximate Stokes-Dirichlet problem (348) has a *unique* solution in $V_{0h} \times P_{0h}$ if *no triangle of* $\mathcal{T}_h$ *has more than one edge contained* in Γ and that the following convergence theorem holds:

Theorem 5 *Let Ω be a bounded polygonal domain of $\mathbb{R}^2$ and suppose that the solution $\{\mathbf{u}, p\}$ of the Stokes problem (323) verifies*

$$\mathbf{u} \in (H^{k+1}(\Omega) \cap H_0^1(\Omega))^2,\ p \in H^k(\Omega) \cap L_0^2(\Omega),\ k = 1 \text{ or } 2.$$

If the family $\{\mathcal{T}_h\}_h$ is regular and if, $\forall h$, no triangle of $\mathcal{T}_h$ has more than one edge contained in Γ, the solution $\{\mathbf{u}_h, p_h\}$ of problem (348), with P_h and V_{0h} defined by (349) and (350), respectively, verifies,

$$||\mathbf{u}_h - \mathbf{u}||_{(H_0^1(\Omega))^2} + ||p - p_h||_{L^2(\Omega)} \leq C_1 h^k (||\mathbf{u}||_{(H^{k+1}(\Omega))^2} + ||p||_{H^k(\Omega)}). \tag{351}$$

If Ω is convex, we also have

$$||\mathbf{u}_h - \mathbf{u}||_{(L^2(\Omega))^2} \leq C_2 h^{k+1} (||\mathbf{u}||_{(H^{k+1}(\Omega))^2} + ||p||_{H^k(\Omega)}). \tag{352}$$

Finally, if the family $\{\mathcal{T}_h\}_h$ is uniformly regular (but Ω not necessarily convex) we also have

$$||p_h - p||_{H^1(\Omega)} \leq C_3 h^{k-1} (||\mathbf{u}||_{(H^{k+1}(\Omega))^2} + ||p||_{H^k(\Omega)}). \tag{353}$$

In (351)-(353), C_1, C_2 and C_3 are positive constants.

A first proof of the above theorem was given by Bercovier and Pironneau 1979 [79]; this proof was improved by Verfurth 1984 [94] and further improved by Girault and Raviart 1986 [59] (Chapter 2, Section 4.2) (see also Brezzi and Fortin 1991 [25] (Chapter 6, Section 6)). We shall conclude this paragraph with the *Bercovier-Pironneau* approximation of the Stokes problem (323); from

Section 5.1.2, this approximation is defined by (348), with P_h given by (349) and V_{0h} by

$$V_{0h} = \{\mathbf{v}_h | \mathbf{v}_h \in (C^0(\overline{\Omega}))^2, \mathbf{v}_h|_T \in (P_1)^2, \ \forall T \in \mathcal{T}_{h/2}, \ \mathbf{v}_h = \mathbf{0} \ on \ \Gamma\}, \quad (354)$$

with, in (354), $\mathcal{T}_{h/2}$ obtained from $\mathcal{T}_h$ by dividing each triangle T of $\mathcal{T}_h$ in four similar triangles, by joining the mid-points of the edges of T (as shown in Figure 5.1). It follows from Girault and Raviart 1986 [59] (Chapter 2, Section 4.2) that if no triangle of $\mathcal{T}_h$ has more than one edge contained in Γ, then problem (348) has a unique solution in $V_{0h} \times P_{0h}$ and the following convergence theorem holds:

Theorem 6 *Let Ω and $\{\mathcal{T}_h\}_h$ be as in Theorem 5 and suppose that the solution $\{\mathbf{u}, p\}$ of problem (323) verifies*

$$\mathbf{u} \in (H^2(\Omega) \cap H_0^1(\Omega))^2, \ p \in H^1(\Omega) \times L_0^2(\Omega).$$

Then, the solution $\{\mathbf{u}_h, p_h\}$ of problem (348), with P_h and V_{0h} defined by (349) and (354), respectively, verifies

$$||\mathbf{u}_h - \mathbf{u}||_{(H_0^1(\Omega))^2} + ||p_h - p||_{L^2(\Omega)} \leq C_1 h(||\mathbf{u}||_{(H^2(\Omega))^2} + ||p||_{H^1(\Omega)}). \quad (355)$$

Moreover, if Ω is convex, we have the following L^2-error estimate

$$||\mathbf{u}_h - \mathbf{u}||_{(L^2(\Omega))^d} \leq C_2 h^2(||\mathbf{u}||_{(H^2(\Omega))^2} + ||p||_{H^1(\Omega)}). \quad (356)$$

In (355), (356), C_1 and C_2 are two positive constants.

Remark 5 *We have discussed several finite element approximations of the Stokes problems (309) and (316) . Once a formulation such as (315) (or (320)) has been obtained, several practical issues still have to be addressed, among them the derivation of the linear systems equivalent to the discrete Stokes problems, and then the numerical solution of these systems. Those issues have been discussed in details in Glowinski 2003 [4] (Chapter 5), especially when deriving the equivalent linear systems how to obtain the accurate evaluation of multiple integrals over the elements of $\mathcal{T}_h$ (or $\mathcal{T}_{h/2}$), or over the element of reference $\hat{T}$. The discussion starts from the finite element approximation of the Stokes-Dirichlet problem (295) by the Hood-Taylor and Bercovier-Pironneau methods, assuming that Ω is a bounded polygonal domain of $\mathbb{R}^2$, and then the mini-element of Arnold-Brezzi-Fortin, the case of curved boundaries and finally the Stokes problem with other boundary conditions than Dirichlet.*

5.2 Finite element implementation of the θ-scheme.

We are going to discuss in this section the *full discretization* of the *Navier-Stokes equations*

$$\frac{\partial \mathbf{u}}{\partial t} - \nu \triangle \mathbf{u} + (\mathbf{u} \cdot \boldsymbol{\nabla})\mathbf{u} + \boldsymbol{\nabla} p = \mathbf{f} \ in \ \Omega \times (0, T), \quad (357)$$

$$\boldsymbol{\nabla} \cdot \mathbf{u} = 0 \ in \ \Omega \times (0, T), \quad (358)$$

$$\mathbf{u}(0) = \mathbf{u}_0 (with \ \boldsymbol{\nabla} \cdot \mathbf{u}_0 = 0), \quad (359)$$

$$\mathbf{u} = \mathbf{g}_0 \ on \ \Gamma_0 \times (0, T), \nu \frac{\partial \mathbf{u}}{\partial n} - \mathbf{n} p = \mathbf{g}_1 \ on \ \Gamma_1 \times (0, T). \quad (360)$$

To approximate problem (357) – (360) we shall combine the finite element methods discussed in the previous subsection with the θ-scheme described by relations (77) - (85). We have seen in Section 2 that a "good" choice for θ, α, β is given by

$$\theta = 1 - 1/\sqrt{2}, \alpha = (1-2\theta)/(1-\theta), \beta = \theta/(1-\theta). \tag{361}$$

A safe way to achieve the *full discretization* of the time dependent Navier-Stokes equations (357) – (360) is to proceed as follows (this approach applies, obviously, to other problems):

(i) Keeping time continuous we shall use the *finite element spaces* introduced in the previous subsection to space discretize the Navier-Stokes equations. We obtain then a system coupling *ordinary differential equations* and *algebraic equations*.

(ii) We shall apply the *operator splitting-methods* of Section 2, to the time-discretization of the above system of algebraic and ordinary differential equations.

At first we shall consider the *pure Dirichlet* case (i.e., the particular case of (360) where $\Gamma_1 = \emptyset$) and, then, mixed boundary conditions such as (360).

5.2.1 Space approximation of the time dependent Navier-Stokes equations

The Dirichlet case.

The problem that we consider is defined by (357) – (359), completed by

$$\mathbf{u} = \mathbf{g} \; on \; \partial\Omega \times (0,T). \tag{362}$$

To have a *well-posed* problem we assume that

$$\int_{\partial\Omega} \mathbf{g}(t) \cdot \mathbf{n} d\Gamma = 0 \; on \; (0,T), \tag{363}$$

and also, in principle,

$$\mathbf{u}_0 \cdot \mathbf{n} = \mathbf{g}(0) \cdot \mathbf{n} \; on \; \partial\Omega \tag{364}$$

(we say "in principle" since some of the test problems does not verify (364), without too much damage on the computational procedure and on the computed solution).

Assuming that $\Omega \subset I\!R^2$, we *space-approximate* problem (357) – (359), (362) by

$$Find \; \{\mathbf{u}_h(t), p_h(t)\} \in V_h \times P_h, \forall \; t \in (0,T), \; such \; that$$

$$\begin{cases} \int_{\Omega_h} \dot{\mathbf{u}}_h \cdot \mathbf{v}_h dx + \nu \int_{\Omega_h} \boldsymbol{\nabla}\mathbf{u}_h : \boldsymbol{\nabla}\mathbf{v}_h dx + \int_{\Omega_h} (\mathbf{u}_h \cdot \nabla)\mathbf{u}_h \cdot \mathbf{v}_h dx \\ -\int_{\Omega_h} p_h \boldsymbol{\nabla} \cdot \mathbf{v}_h dx = \int_{\Omega_h} \mathbf{f}_h \cdot \mathbf{v}_h dx, \forall\ \mathbf{v}_h \in V_{0h}, \end{cases} \tag{365}$$

$$\int_{\Omega_h} \boldsymbol{\nabla} \cdot \mathbf{u}_h q_h dx = 0, \forall\ q_h \in P_h, \tag{366}$$

$$\mathbf{u}_h(t) = \mathbf{g}_h(t)\ on\ \partial\Omega_h (with\ \mathbf{g}_h(t) \in \gamma V_h), \tag{367}$$

$$\mathbf{u}_h(0) = \mathbf{u}_{0h} (with\ \mathbf{u}_{0h} \in V_h). \tag{368}$$

In (365) - (368):

- We have $\Omega_h = \Omega$ and $\partial\Omega_h = \partial\Omega$ if Ω is polygonal. For the cases where Ω is not polygonal, see the discussion in Glowinski 2003 [4] for the *isoparametric* generalization of the *Hood-Taylor* (resp., *Bercovier-Pironneau*) approximation.

- The finite element *velocity* and *pressure* spaces V_h and P_h are as in Section 5.1.2 and, here,

$$V_{0h} = V_h \cap (H_0^1(\Omega_h))^2 = \{\mathbf{v}_h | \mathbf{v}_h \in V_h, \mathbf{v}_h = \mathbf{0}\ on\ \partial\Omega_h\}.$$

- We have used the notation $\dot{\mathbf{u}}_h$ for $\dfrac{\partial \mathbf{u}_h}{\partial t}$.

- The functions $\mathbf{f}_h, \mathbf{u}_{0h}$ and $\mathbf{g}_h$ are convenient approximations of $\mathbf{f}, \mathbf{u}_0$ and $\mathbf{g}$, respectively. Function $\mathbf{g}_h$ has to verify

$$\int_{\partial\Omega_h} \mathbf{g}_h(t) \cdot \mathbf{n} d\Gamma_h = 0\ on\ (0, T); \tag{369}$$

  to construct, from $\mathbf{g}$, an approximation $\mathbf{g}_h$ verifying (369) we shall use the methods discussed in Section 5.1.3.

- The boundary space γV_h is defined as in Section 5.1.3.

The case of the mixed boundary conditions (360).

In this case the boundary conditions are given by

$$\mathbf{u} = \mathbf{g}_0\ on\ \Gamma_0 \times (0, T), \nu \frac{\partial \mathbf{u}}{\partial n} - \mathbf{n}p = \mathbf{g}_1\ on\ \Gamma_1 \times (0, T),$$

leading to the following approximate problem:

$$Find\ \{\mathbf{u}_h(t), p_h(t)\} \in V_h \times P_h, \forall\ t \in (0, T),\ such\ that$$

$$\begin{cases} \int_{\Omega_h} \dot{\mathbf{u}}_h \cdot \mathbf{v}_h dx + \nu \int_{\Omega_h} \boldsymbol{\nabla}\mathbf{u}_h : \boldsymbol{\nabla}\mathbf{v}_h dx + \int_{\Omega_h} (\mathbf{u}_h \cdot \boldsymbol{\nabla})\mathbf{u}_h \cdot \mathbf{v}_h dx \\ -\int_{\Omega_h} p_h \boldsymbol{\nabla} \cdot \mathbf{v}_h dx = \int_{\Omega_h} \mathbf{f}_h \cdot \mathbf{v}_h dx + \int_{\Gamma_{1h}} \mathbf{g}_{1h} \cdot \mathbf{v}_h d\Gamma_h, \forall\ \mathbf{v}_h \in V_{0h}, \end{cases} \tag{370}$$

$$\int_{\Omega_h} \boldsymbol{\nabla} \cdot \mathbf{u}_h q_h dx = 0, \forall\ q_h \in P_h, \tag{371}$$

$$\mathbf{u}_h(t) = \mathbf{g}_{0h}(t)\ on\ \Gamma_{0h}, \tag{372}$$

$$\mathbf{u}_h(0) = \mathbf{u}_{0h}(with\ \mathbf{u}_{0h} \in V_h); \tag{373}$$

in (370) the space V_{0h} is defined as in Section 5.1.2, the other notation being self-explanatory.

Expanding $\mathbf{u}_h$ and p_h on vector bases of V_h and P_h, respectively, and taking for the test functions $\mathbf{v}_h$ and q_h *all* the elements of the vector bases of V_{0h} and P_h, formulations (365) – (368) and (370) – (373) will produce a system of *ordinary differential equations* with respect to t coupled to the linear relations associated to the discrete incompressibility condition. Applying to these *algebraic-differential* problems the time discretization methods by operator splitting of Section 2 is straightforward, as we shall see hereafter, where we shall focus on the θ-scheme of Section 2 in order to derive the fully discrete analogs of schemes (77) - (85) including the particular case where $\Gamma_1 = \emptyset$ (pure Dirichlet boundary conditions).

Full discretization by the θ-scheme : Case of the Dirichlet boundary conditions.

The *algebraic-differential system* to time-discretize is (365) – (368). We obtain then

$$\mathbf{u}_h^0 = \mathbf{u}_{0h}; \tag{374}$$

then, for $n \geq 0, \mathbf{u}_h^n$ *being known, we compute* $\{\mathbf{u}_h^{n+\theta}, p_h^{n+\theta}\} \in V_h \times P_h$, *then* $\mathbf{u}_h^{n+1-\theta} \in V_h$, *and finally* $\{\mathbf{u}_h^{n+1}, p_h^{n+1}\} \in V_h \times P_h$ *by solving the following discrete elliptic systems*

$$\begin{cases} \int_{\Omega_h} \dfrac{\mathbf{u}_h^{n+\theta} - \mathbf{u}_h^n}{\theta \Delta t} \cdot \mathbf{v}_h dx + \alpha\nu \int_{\Omega_h} \boldsymbol{\nabla}\mathbf{u}_h^{n+\theta} : \boldsymbol{\nabla}\mathbf{v}_h dx - \int_{\Omega_h} p_h^{n+\theta} \boldsymbol{\nabla} \cdot \mathbf{v}_h dx = \\ \int_{\Omega_h} \mathbf{f}_h{}^{n+\theta} \cdot \mathbf{v}_h dx - \beta\nu \int_{\Omega_h} \boldsymbol{\nabla}\mathbf{u}_h^n : \boldsymbol{\nabla}\mathbf{v}_h dx - \int_{\Omega_h} (\mathbf{u}_h^n \cdot \boldsymbol{\nabla})\mathbf{u}_h^n \cdot \mathbf{v}_h dx, \forall \mathbf{v}_h \in V_{0h}, \\ \int_{\Omega_h} \boldsymbol{\nabla} \cdot \mathbf{u}_h^{n+\theta} q_h dx = 0, \forall q_h \in P_h, \\ \mathbf{u}_h^{n+\theta} = \mathbf{g}_h^{n+\theta}\ on\ \partial\Omega_h, \end{cases} \tag{375}$$

then

$$\begin{cases} \displaystyle\int_{\Omega_h} \frac{\mathbf{u}_h^{n+1-\theta} - \mathbf{u}_h^{n+\theta}}{(1-2\theta)\Delta t} \cdot \mathbf{v}_h dx + \beta\nu \int_{\Omega_h} \boldsymbol{\nabla}\mathbf{u}_h^{n+1-\theta} : \boldsymbol{\nabla}\mathbf{v}_h dx \\ \displaystyle + \int_{\Omega_h} (\mathbf{u}_h^{n+1-\theta} \cdot \boldsymbol{\nabla})\mathbf{u}_h^{n+1-\theta} \cdot \mathbf{v}_h dx = \int_{\Omega_h} \mathbf{f}_h^{n+\theta} \cdot \mathbf{v}_h dx \\ \displaystyle -\alpha\nu \int_{\Omega_h} \boldsymbol{\nabla}\mathbf{u}_h^{n+\theta} : \boldsymbol{\nabla}\mathbf{v}_h dx - \int_{\Omega_h} p_h^{n+\theta} \cdot \boldsymbol{\nabla}\mathbf{v}_h dx, \forall \mathbf{v}_h \in V_{0h}, \\ \mathbf{u}_h^{n+1-\theta} = \mathbf{g}_h^{n+1-\theta} \text{ on } \partial\Omega_h, \end{cases} \tag{376}$$

and finally

$$\begin{cases} \displaystyle\int_{\Omega_h} \frac{\mathbf{u}_h^{n+1} - \mathbf{u}_h^{n+1-\theta}}{\theta\Delta t} \cdot \mathbf{v}_h dx + \alpha\nu \int_{\Omega_h} \boldsymbol{\nabla}\mathbf{u}_h^{n+1} : \boldsymbol{\nabla}\mathbf{v}_h dx \\ \displaystyle -\int_{\Omega_h} p_h^{n+1} \boldsymbol{\nabla} \cdot \mathbf{v}_h dx = \int_{\Omega_h} \mathbf{f}_h^{n+1} \cdot \mathbf{v}_h dx - \beta\nu \int_{\Omega_h} \boldsymbol{\nabla}\mathbf{u}_h^{n+1-\theta} : \boldsymbol{\nabla}\mathbf{v}_h dx \\ \displaystyle -\int_{\Omega_h} (\mathbf{u}_h^{n+1-\theta} \cdot \boldsymbol{\nabla})\mathbf{u}_h^{n+1-\theta} \cdot \mathbf{v}_h dx, \forall \mathbf{v}_h \in V_{0h}, \\ \displaystyle \int_{\Omega_h} \boldsymbol{\nabla} \cdot \mathbf{u}_h^{n+1} q_h dx = 0, \forall q_h \in P_h, \\ \mathbf{u}_h^{n+1} = \mathbf{g}_h^{n+1} \text{ on } \partial\Omega_h, \end{cases} \tag{377}$$

respectively. In (374) – (377), the finite element spaces V_h, V_{0h}, and P_h are as in Section 5.1.4 for the Dirichlet case. For θ, α, β we advocate the values given by (361).

Full discretization by the θ-scheme : Case of the mixed boundary conditions.

The time discretization of problem (370) – (373) leads to the following scheme:

$$\mathbf{u}_h^0 = \mathbf{u}_{0h}; \tag{378}$$

then, for $n \geq 0, \mathbf{u}_h^n$ *being known, we compute* $\{\mathbf{u}_h^{n+\theta}, p_h^{n+\theta}\} \in V_h \times P_h$, *then* $\mathbf{u}_h^{n+1-\theta} \in V_h$, *and finally* $\{\mathbf{u}_h^{n+1}, p_h^{n+1}\} \in V_h \times P_h$ *by solving the following discrete elliptic systems*

$$\begin{cases} \displaystyle\int_{\Omega_h} \frac{\mathbf{u}_h^{n+\theta} - \mathbf{u}_h^n}{\theta\Delta t} \cdot \mathbf{v}_h dx + \alpha\nu \int_{\Omega_h} \boldsymbol{\nabla}\mathbf{u}_h^{n+\theta} : \boldsymbol{\nabla}\mathbf{v}_h dx - \int_{\Omega_h} p_h^{n+\theta} \boldsymbol{\nabla} \cdot \mathbf{v}_h dx = \\ \displaystyle \int_{\Omega_h} \mathbf{f}_h^{n+\theta} \cdot \mathbf{v}_h dx + \int_{\Gamma_{1h}} \mathbf{g}_{1h}^{n+\theta} \cdot \mathbf{v}_h d\Gamma_h - \beta\nu \int_{\Omega_h} \boldsymbol{\nabla}\mathbf{u}_h^n : \boldsymbol{\nabla}\mathbf{v}_h dx \\ \displaystyle -\int_{\Omega_h} (\mathbf{u}_h^n \cdot \boldsymbol{\nabla})\mathbf{u}_h^n \cdot \mathbf{v}_h dx, \forall \mathbf{v}_h \in V_{0h}, \\ \displaystyle \int_{\Omega_h} \boldsymbol{\nabla} \cdot \mathbf{u}_h^{n+\theta} q_h dx = 0, \forall q_h \in P_h, \\ \mathbf{u}_h^{n+\theta} = \mathbf{g}_{0h}^{n+\theta} \text{ on } \Gamma_{0h}, \end{cases} \tag{379}$$

then

$$\begin{cases} \displaystyle\int_{\Omega_h} \frac{\mathbf{u}_h^{n+1-\theta} - \mathbf{u}_h^{n+\theta}}{(1-2\theta)\Delta t} \cdot \mathbf{v}_h dx + \beta\nu \int_{\Omega_h} \boldsymbol{\nabla}\mathbf{u}_h^{n+1-\theta} : \boldsymbol{\nabla}\mathbf{v}_h dx \\ \displaystyle + \int_{\Omega_h} (\mathbf{u}_h^{n+1-\theta} \cdot \boldsymbol{\nabla})\mathbf{u}_h^{n+1-\theta} \cdot \mathbf{v}_h dx = \int_{\Omega_h} \mathbf{f}_h^{n+\theta} \cdot \mathbf{v}_h dx \\ \displaystyle + \int_{\Gamma_{1h}} \mathbf{g}_{1h}^{n+\theta} \cdot \mathbf{v}_h d\Gamma_h - \alpha\nu \int_{\Omega_h} \boldsymbol{\nabla}\mathbf{u}_h^{n+\theta} : \boldsymbol{\nabla}\mathbf{v}_h dx - \int_{\Omega_h} p_h^{n+\theta} \boldsymbol{\nabla} \cdot \mathbf{v}_h dx, \forall \mathbf{v}_h \in V_{0h}, \\ \mathbf{u}_h^{n+1-\theta} = \mathbf{g}_{0h}^{n+1-\theta} \text{ on } \Gamma_{0h}, \end{cases} \tag{380}$$

and finally

$$\begin{cases} \displaystyle\int_{\Omega_h} \frac{\mathbf{u}_h^{n+1} - \mathbf{u}_h^{n+1-\theta}}{\theta\Delta t} \cdot \mathbf{v}_h dx + \alpha\nu \int_{\Omega_h} \boldsymbol{\nabla}\mathbf{u}_h^{n+1} : \boldsymbol{\nabla}\mathbf{v}_h dx \\ \displaystyle - \int_{\Omega_h} p_h^{n+1} \boldsymbol{\nabla} \cdot \mathbf{v}_h dx = \int_{\Omega_h} \mathbf{f}_h^{n+1} \cdot \mathbf{v}_h dx + \int_{\Gamma_{1h}} \mathbf{g}_{1h}^{n+1} \cdot \mathbf{v}_h d\Gamma_h \\ \displaystyle - \beta\nu \int_{\Omega_h} \boldsymbol{\nabla}\mathbf{u}_h^{n+1-\theta} : \boldsymbol{\nabla}\mathbf{v}_h dx - \int_{\Omega_h} (\mathbf{u}_h^{n+1-\theta} \cdot \boldsymbol{\nabla})\mathbf{u}_h^{n+1-\theta} \cdot \mathbf{v}_h dx, \forall \mathbf{v}_h \in V_{0h}, \\ \displaystyle\int_{\Omega_h} \boldsymbol{\nabla} \cdot \mathbf{u}_h^{n+1} q_h dx = 0, \forall q_h \in P_h, \\ \mathbf{u}_h^{n+1} = \mathbf{g}_{0h}^{n+1} \text{ on } \Gamma_{0h}, \end{cases} \tag{381}$$

respectively. In (378)–(381), the finite element spaces V_h, V_{0h}, and P_h are as in Section 5.1.4 for the case of mixed boundary conditions and for θ, α, β we advocate the values given by (361).

Remark 6 *In order to solve the discrete Stokes problems (375), (377), (379), (381) and the discrete-advection diffusion problems (376), (380), one can use discrete variants of the conjugate gradient algorithms discussed in Sections 3 and 4. The implementation of these algorithms, which boils down to the solution of sequences of linear systems for symmetric and positive definite matrices, will be further discussed later.*

Remark 7 *If we replace the* nonlinear problem *(376) by the following (*linearized*) one*

$$\begin{cases} \displaystyle\int_{\Omega_h} \frac{\mathbf{u}_h^{n+1-\theta} - \mathbf{u}_h^{n+\theta}}{(1-2\theta)\Delta t} \cdot \mathbf{v}_h dx + \beta\nu \int_{\Omega_h} \boldsymbol{\nabla}\mathbf{u}_h^{n+1-\theta} : \boldsymbol{\nabla}\mathbf{v}_h dx \\ \displaystyle + \int_{\Omega_h} (\mathbf{u}_h^{n+\theta} \cdot \boldsymbol{\nabla})\mathbf{u}_h^{n+1-\theta} \cdot \mathbf{v}_h dx = \int_{\Omega_h} \mathbf{f}_h^{n+\theta} \cdot \mathbf{v}_h dx \\ \displaystyle - \alpha\nu \int_{\Omega_h} \boldsymbol{\nabla}\mathbf{u}_h^{n+\theta} : \boldsymbol{\nabla}\mathbf{v}_h dx - \int_{\Omega_h} p_h^{n+\theta} \cdot \boldsymbol{\nabla}\mathbf{v}_h dx, \forall \mathbf{v}_h \in V_{0h}, \\ \mathbf{u}_h^{n+1-\theta} = \mathbf{g}_h^{n+1-\theta} \text{ on } \partial\Omega_h, \end{cases}$$

the new scheme is essentially as stable and accurate as the original scheme (374) – (377); on the other hand, it is less costly to solve the linearized one

than problem (376) (for the same value of Δt, at least). Similar replacement can be done in (380).

The numerical integration of the advection term in (374)–(377) and (378)–(381) (also in the linearized ones) for the Hood-Taylor, Bercovier-Pironneau and Arnold-Brezzi-Fortin approximations of the Navier-Stokes equations have been discussed in Glowinski 2003 [4] (Chapter 5, Section 27).

5.3 Finite element implementation of the L^2-projection/wave-like equation method

This section is dedicated to the numerical solution of the Navier-Stokes equations modeling incompressible viscous fluid flow by a methodology combining time discretization by a first order accurate operator-splitting, Stokes solvers à la Uzawa and a wave-like equation treatment of the advection. The goal is to apply this approach to simulate more complicated flow problems, such as rigid bodies moving freely in the fluid (see, e.g., Glowinski et al. 1999, 2000, and 2001 [95, 96, 97, 98].

Following Chorin 1967 and 1968 [38, 39], most "modern" Navier-Stokes solvers are based on operator splitting algorithms (see, e.g., in Marchuk 1990 [29] and Turek 1996 [99]) in order to force the incompressibility condition via a Stokes solver or a L^2-projection method.

Applying scheme à la Marchuk–Yanenko discussed in Section 2.3, we have the following scheme for the Dirichlet case (365) - (368) (after dropping some of the subscripts h and applying the backward Euler's method for time discretization):

$$\mathbf{u}^0 = \mathbf{u}_{0h} \ \textit{is given};\tag{382}$$

for $n \geq 0$, $\mathbf{u}^n$ being known,

$$\begin{cases} \displaystyle\int_\Omega \frac{\mathbf{u}^{n+1/3} - \mathbf{u}^n}{\triangle t} \cdot \mathbf{v}\, d\mathbf{x} - \int_\Omega p^{n+1/3} \boldsymbol{\nabla} \cdot \mathbf{v}\, d\mathbf{x} = 0, \ \forall \mathbf{v} \in V_{0h}, \\ \displaystyle\int_\Omega q \boldsymbol{\nabla} \cdot \mathbf{u}^{n+1/3}\, d\mathbf{x} = 0, \ \forall q \in L^2_h; \\ \mathbf{u}^{n+1/3} \in V^{n+1}_{\mathbf{g}_h}, \ p^{n+1/3} \in L^2_{0h}, \end{cases}\tag{383}$$

$$\begin{cases} \displaystyle\int_\Omega \frac{\partial \mathbf{u}(t)}{\partial t} \cdot \mathbf{v}\, d\mathbf{x} + \int_\Omega (\mathbf{u}^{n+1/3} \cdot \boldsymbol{\nabla})\mathbf{u}(t) \cdot \mathbf{v}\, d\mathbf{x} = 0 \ \textit{on} \ (t^n, t^{n+1}), \ \forall \mathbf{v} \in V^{n+1,-}_{0h}, \\ \mathbf{u}(t^n) = \mathbf{u}^{n+1/3}, \\ \mathbf{u}(t) \in V_h, \ \mathbf{u}(t) = \mathbf{g}_h(t^{n+1}) \ \textit{on} \ \Gamma^{n+1}_- \times (t^n, t^{n+1}), \end{cases}\tag{384}$$

$$\mathbf{u}^{n+2/3} = \mathbf{u}(t^{n+1}),\tag{385}$$

$$\begin{cases} \displaystyle\int_\Omega \frac{\mathbf{u}^{n+1} - \mathbf{u}^{n+2/3}}{\triangle t} \cdot \mathbf{v}\, d\mathbf{x} + \nu \int_\Omega \boldsymbol{\nabla}\mathbf{u}^{n+1} : \boldsymbol{\nabla}\mathbf{v}\, d\mathbf{x} = \int_\Omega \delta \cdot \mathbf{v} d\mathbf{x}, \\ \forall \mathbf{v} \in V_{0h}; \ \mathbf{u}^{n+1} \in V^{n+1}_{\mathbf{g}_h}, \end{cases}\tag{386}$$

with:

(a) $V_{\mathbf{g}_h}^{n+1} = V_{\mathbf{g}_h(t^{n+1})}$,

(b) $\Gamma_-^{n+1} = \{\mathbf{x} \mid \mathbf{x} \in \Gamma,\ \mathbf{g}_h(\mathbf{x}, t^{n+1}) \cdot \mathbf{n}(\mathbf{x}) < 0\}$,

(c) $V_h = \{\mathbf{v}_h \mid \mathbf{v}_h \in (C^0(\overline{\Omega}))^2,\ \mathbf{v}_h|_T \in P_1 \times P_1,\ \forall T \in \mathcal{T}_{h/2}\}$,

(d) $V_{0h}^{n+1,-} = \{\mathbf{v} \mid \mathbf{v} \in V_h, \mathbf{v} = \mathbf{0}\ on\ \Gamma_-^{n+1}\}$.

Problem (383) can be viewed as a degenerated (zero viscosity) discrete Stokes problem for which efficient solution methods already exist (e.g., the discrete analogue of the preconditioned conjugate gradient algorithm for the generalized Stokes problems discussed in Section 4.7). Problem (384) can be solved by a wave-like equation method discussed in Section 3.3. Similarly problem (386) is a discrete elliptic system whose iterative or direct solution is quite a classical problem.

5.4 On the numerical solution of the discrete subproblems.

The solution of the subproblems, encountered at each time step of the operator splitting schemes described in Sections 5.2 and 5.3, can be computed by iterative methods which are the discrete analogues of the conjugate gradient methods discussed in Sections 3 and 4. In particular, we shall have to solve quite systematically the linear systems approximating the elliptic systems associated to the *Helmholtz operator* $\alpha I - \nu\triangle$. Also, some of the *Stokes solvers* discussed in Section 4 require the solution of *Poisson problems* for preconditioning purposes. From the above observations, it makes sense to discuss with some detail the numerical solution of the discrete Helmholtz and Poisson problems encountered at each step of the operator splitting schemes.

On the solution of the discrete Helmholtz equations.

If the boundary conditions are of the *Dirichlet type* only (i.e., if $\Gamma_0 = \Gamma(= \partial\Omega)$), we shall have to solve problems like

$$\alpha\mathbf{u} - \nu\triangle\mathbf{u} = \mathbf{f}\ in\ \Omega, \mathbf{u} = \mathbf{g}\ on\ \Gamma. \tag{387}$$

Paradoxically, solving problem (387) is not very expensive for flow at *high Reynolds numbers*. Why? Because for such flow, the viscosity ν is small, and their fast dynamics requires small $\triangle t$, i.e., large values of α. Suppose for simplicity that $\Omega = (0,1)^2$, and also that one uses over Ω a regular triangulation like the one in Figure 5.2 where $h = 1/(I+1)$ (I a positive integer). Suppose also that one uses continuous and piecewise linear approximations of the velocity over the above triangulation, and that integrals like $\int_\Omega \mathbf{v}\cdot\mathbf{w}dx$ are approximated using the trapezoidal rule. One obtains then the approximation of problem (387)

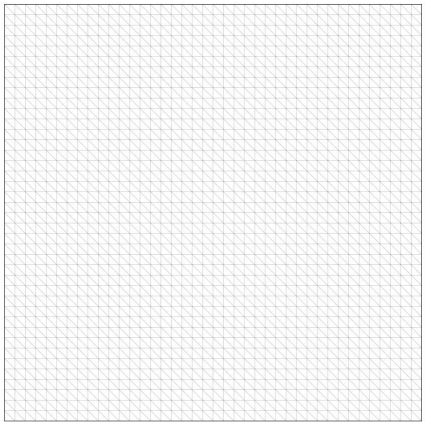


Figure 5.2: An example of a regular triangulation.

associated to the traditional *five point finite difference scheme*, namely (with obvious notation)

$$\begin{cases} \alpha \mathbf{u}_{ij} + \dfrac{\nu}{h^2}(4\mathbf{u}_{ij} - \mathbf{u}_{i+1j} - \mathbf{u}_{i-1j} - \mathbf{u}_{ij+1} - \mathbf{u}_{ij-1}) = \mathbf{f}_{ij}, \\ 1 \leq i, j \leq I, \\ \mathbf{u}_{kl} = \mathbf{g}_{kl} \ if \ \{kh, lh\} \in \Gamma. \end{cases} \tag{388}$$

It is well known that the matrix in (388) has for smallest and largest eigenvalues

$$\lambda_{min} \ = \alpha + \frac{8\nu}{h^2} \ sin^2 \frac{\pi h}{2}, \quad \lambda_{max} \ = \alpha + \frac{8\nu}{h^2} sin^2 \frac{I\pi h}{2},$$

respectively. For *small values* of h, we clearly have

$$\lambda_{min} \ \approx \alpha + 2\pi^2\nu, \lambda_{max} \ \approx \alpha + 8\nu/h^2,$$

implying that the *condition number* $\mathcal{N}$ of the above matrix verifies

$$\mathcal{N} = \lambda_{max}/\lambda_{min} \ \approx (\alpha + 8\nu/h^2)/(\alpha + 2\pi^2\nu).$$

Suppose now that $\nu = 10^{-3}, h = 10^{-2}, \triangle t = 10^{-2} (\Rightarrow \alpha = 10^2)$; we have then

$$\mathcal{N} \approx 1.8. \tag{389}$$

Suppose now that we solve the linear system (388) by a *nonpreconditioned conjugate gradient algorithm*. It follows then from (389) and from (123) that the distance, between the solution of problem (388) and the n^{th} iterate, converges to zero at least as fast as

$$\left(\frac{\sqrt{1.8} - 1}{\sqrt{1.8} + 1}\right)^n = (.145898\ldots)^n,$$

which corresponds to a high speed of convergence. A similar conclusion would hold for the *successive over-relaxation* method with optimal parameter. Actually, the convergence of the above methods is sufficiently fast (in that particular case, at least) that it makes useless further speeding up (by a multigrid method for example).

Remark 8 *Suppose now that the finite element mesh used to solve problem (387) is unstructured (or at least less structured than the mesh shown on Figure 5.2). We advocate, then, to solve the discrete analogue of (387), namely (with obvious notation)*

$$\mathbf{A}_h \mathbf{U}_h = \mathbf{F}_h, \tag{390}$$

by a conjugate gradient algorithm, preconditioned by the diagonal $\mathbf{D}_h$ *of matrix* $\mathbf{A}_h$.

On the solution of the pressure related discrete Poisson problems.

The solution of the *discrete Stokes problems* (375), (377), (379), (381), by the discrete analogues of the *preconditioned conjugate gradient algorithms* discussed in Section 4, requires – at each iteration – the solution of a linear system approximating *Poisson problems* of the following types

$$\begin{cases} -\triangle\varphi = f \text{ in } \Omega, \dfrac{\partial\varphi}{\partial n} = 0 \text{ on } \Gamma, \displaystyle\int_\Omega \varphi dx = 0, \\ if\ \Gamma_0 = \Gamma(Stokes - Dirichlet\ case), \end{cases} \tag{391}$$

and

$$\begin{cases} -\triangle\varphi = f \text{ in } \Omega, \dfrac{\partial\varphi}{\partial n} = 0 \text{ on } \Gamma_0, \varphi = 0 \text{ on } \Gamma_1, \\ if \displaystyle\int_{\Gamma_i} d\Gamma > 0, \forall i = 0, 1 (Stokes\ problem\ with\ mixed\ boundary\ conditions). \end{cases} \tag{392}$$

The matrices approximating the Laplace operators occurring in (391) and (392) do not enjoy the nice properties of the elliptic operator $\alpha I - \nu\triangle$ discussed above, concerning their condition number, and therefore the approximate solution of problems (391) and (392) may be costly (for three-dimensional problems, particularly). For $two - dimensional\ problems$, we advocate *direct methods* (à la *Cholesky*, for example) for solving these *discrete Poisson problems*. For $three - dimensional$ flow problems, *multigrid methods* seem to be well-suited to solve problems (391) and (392); the multigrid solution of problems such as (391) and (392) has been discussed in, e.g., Glowinski 2003 [4] (Chapter 5).

Remark 9 *The condition number of the finite element matrices approximating the Laplace operator in (391) and (392) behaves like* h^{-2}.

Remark 10 *To solve the linear system approximating (391), by the method of Cholesky, we shall proceed as follows:*

(i) *We delete one equation and set to zero the corresponding unknown.*

(ii) *We solve the remaining system by the method of Cholesky.*

(iii) *Let* φ_h^* *be the element of the pressure space* P_h *associated to the solution of the above linear system. Compute (via the trapezoidal rule)* $m_h = \int_{\Omega_h} \varphi_h^* dx / meas.(\Omega_h)$ *and denote by* φ_h *the function defined by*

$$\varphi_h = \varphi_h^* - m_h;$$

we clearly have $\int_{\Omega_h} \varphi_h dx = 0.$

Remark 11 *The discrete Poisson problems, approximating problems (391) and (392), have to be solved in the discrete pressure space* P_h*; if one uses the approximations defined by (298), (299) (Hood − Taylor), (298), (302) (Bercovier − Pironneau), we have 8 times more unknowns for velocity than for pressure (16 times more for three-dimensional flow).*

References

[1] C. Canuto, M.Y. Hussaini, A. Quarteroni, and T.A. Zang, *Spectral Methods in Fluid Dynamics*, Springer–Verlag, New York, 1988.

[2] M. Lesieur, *Turbulence in Fluids*, Kluwer, Dordrecht, 1990.

[3] E. Guyon, J.P. Hulin, and L. Petit, *Hydrodynamique Physique*, Intereditions/Editions du CNRS, Paris, 1991.

[4] R. Glowinski, Finite element methods for incompressible viscous flow, in *Handbook of Numerical Analysis*, **Vol. IX**, P.G. Ciarlet, J.-L. Lions (deceased) eds., North-Holland, Amsterdam, 2003.

[5] W. Prager, *Introduction to Mechanics of Continua*, Ginn and Company, Boston, MA, 1961.

[6] G.K. Batchelor, *An Introduction to Fluid Mechanics*, Cambridge University Press, Cambridge, U.K., 1967.

[7] A.J. Chorin and J.E. Marsden, *A Mathematical Introduction to Fluid Mechanics*, Springer–Verlag, New York, 1990.

[8] R. Glowinski, *Numerical Methods for Nonlinear Variational Problems*, Springer–Verlag, New York, 1984.

[9] M.O. Bristeau, R. Glowinski, B. Mantel, J. Periaux, P. Perrier, Numerical Methods for incompressible and compressible Navier-Stokes problems, in *Finite Element in Fluids*, **Vol. 6**, R.H. Gallagher, G. Carey, J.T. Oden, and O.C. Zienkiewicz eds., J. Wiley, Chicester, 1985, 1–40.

[10] O. Pironneau, *Finite Element Methods for Fluids*, J. Wiley, Chicester, 1989.

[11] J. Leray, Sur le mouvement d'un liquide visqueux emplissant l'espace, *Acta Mathematica*, **63**(1934), 193–248.

[12] J. Leray, Essai sur les mouvements d'un liquide visqueux que limitent des parois, *J. Math. Pures et Appl.*, **13**(1934), 331–418.

[13] E. Hopf, Uber die Anfangswertaufgabe fur die hydrodynamischen Grundgleichungen, *Math. Nachrichten*, **4**(1951), 213–231.

[14] J. Leray, Aspects de la mécanique théorique des fluides, *La Vie des Sciences*, Comptes Rendus de l'Académie des Sciences, Paris, Série Générale, **11**(1994), 287–290.

[15] J.L. Lions and G. Prodi, Un théorème d'existence et d'unicité dans les équations de Navier-Stokes en dimension2, *C.R. Acad. Sci., Paris* **248**, 3519–3521.

[16] J.L. Lions, *Equations Différentielles Opérationnelles et Problémes aux Limites*, Springer–Verlag, Berlin, 1961.

[17] J.L. Lions, *Quelques Méthodes de Résolution des Problémes aux Limites Non Linéaires*, Dunod, Paris, 1969.

[18] O. Ladysenskaya, *Theory and Numerical Analysis of the Navier-Stokes Equations*, Gordon and Breach, New York, NY, 1969.

[19] R. Temam, *The Mathematical Theory of Viscous Incompressible Flow*, North-Holland, Amsterdam, 1977.

[20] L. Tartar, *Topics in Nonlinear Analysis*, Publications Mathématiques d'Orsay, Université Paris-Sud, Départment de Mathématiques, Paris, 1978.

[21] H.O. Kreiss and J. Lorenz, *Initial-Boundary Value Problems and the Navier-Stokes Equations*, Academic Press, Boston, MA, 1989.

[22] P.L. Lions, *Mahtematical Topics in Fluid Mechanics, Vol I: Incompressible Models*, Oxford University, Oxford, UK, 1996.

[23] M. Marion and R. Temam, *Navier-Stokes Equations*, in *Handbook of Numerical Analysis*, **Vol. VI**, P.G. Ciarlet, J.-L. Lions (deceased) eds., North-Holland, Amsterdam, 1998, 503–689.

[24] P.G. Ciarlet, *The Finite Element Method for Elliptic Problems*, North-Holland, Amsterdam, 1978.

[25] F. Brezzi and M. Fortin, *Mixed and Hybrid Finite Element Methods*, Springer-Verlag, New York, NY, 1991.

[26] S.C. Brenner and L.R. Scott *The Mathematical Theory of Finite Element Methods*, Springer-Verlag, New York, NY, 1994.

[27] N.N. Yanenko, *The Method of Fractional Steps*, Springer-Verlag, Berlin, 1971.

[28] G.I. Marchuk, *Methods of Numerical Mathematics*, Springer-Verlag, New York, NY, 1975.

[29] G.I. Marchuk, Splitting and alternating direction methods. In Ciarlet, P.G., and Lions, J.L. (eds.) *Handbook of Numerical Analysis*, **Vol I**, North-Holland, Amsterdam, 1990, 197–462.

[30] M. Crouzeix and A. Mignot, *Analyse Numérique des Equations Différentielles Ordiuaires*, Masson, Paris, 1984.

[31] R. Glowinski and P. Le Tallec, *Augmented Lagrangians and Operator Splitting Methods in Nonlinear Mechanics*, SIAM, Philadelphia, PA, 1989.

[32] R. Glowinski, Viscous flow simulation by finite element methods and related numerical techniques. In Murman, E.M., and Abarbanel, S.S. (eds.) *Progress and Supercomputing in Computational Fluid Dynamics*, Birkhauser, Boston, MA, 1985, 173–210.

[33] R. Glowinski, Splitting methods for the numerical solution of the incompressible Navier-Stokes equations. In Balakrishnan, A.V., Dorodnitsyn, A.A., and Lions, J.L. (eds.) *Vistas in Applied Mathematics*, Optimization Software, New York, NY, 1986, 57–95.

[34] G. Strang, On the construction and comparison of difference schemes, *SIAM J. Num. Anal.*, **5**(1968), 506–517.

[35] J.T. Beale and A. Majda, Rates of convergence for viscous splitting of the Navier-Stokes equations, *Math. Comp.*, **37**(1981), 243–260.

[36] R. Leveque and J. Oliger, Numerical methods based on additive splitting for hyperbolic partial differential equations, *Math. Comp.*, **37** (1983), 243–260.

[37] P.A. Raviart and J.M. Thomas, *Introduction à l'Analyse Numérique des Equations aux Dérivées Partielles*, Masson, Paris, 1983.

[38] A.J. Chorin, A numerical method for solving incompressible viscous flow problems, *J. Comp. Phys.*, **2** (1967), 12–26.

[39] A.J. Chorin, Numerical solution of the Navier-Stokes equations, *Math. Comp.*, **23** (1968), 341–354.

[40] R. Temam, Sur l'approximation des équations de Navier-Stokes par la méthode des pas fractionnaires (I), *Arch. Rat. Mech. Anal.*, **32** (1969), 135–153.

[41] R. Temam, Sur l'approximation des équations de Navier-Stokes par la méthode des pas fractionnaires (II), *Arch. Rat. Mech. Anal.*, **33** (1969), 377–385.

[42] A. Quarteroni and A. Valli, *Numerical Approximation of Partial Differential Equations*, Springer-Verlag, Berlin, 1994.

[43] J. Daniel, *The Approximate Minimization of Functionals*, Prentice Hall, Englewood Cliffs, NJ, 1970.

[44] E. Polak, *Computational Methods in Optimization*, Academic Press, New York, NY, 1971.

[45] M.R. Hestenes and E.L. Stiefel, Methods of conjugate gradients for solving linear systems, *J. Res. Bureau National Standards*, Section B, **49** (1952), 409–436.

[46] R.W. Freund, G.H. Golub, and N.M. Nachtigal, Iterative solution of linear systems, *Acta Numerica 1992*, Cambridge University Press, 1992, 57–100.

[47] J. Nocedal, Theory of algorithms for unconstrained optimization, *Acta Numerica 1992*, Cambridge University Press, 1992, 199–242.

[48] C.T. Kelley, *Iterative Methods for Linear and Nonlinear Equations*, SIAM, Philadelphia, PA, 1995.

[49] Y. Saad, *Iterative Methods for Sparse Linear Systems*, PWS, Boston, MA, 1995.

[50] G.H. Golub and D.P. O'Leary, Some history of the conjugate gradient and Lanczos algorithms: 1948-1976, *SIAM Review*, **31** (1989), 50–102.

[51] E. Zeidler, *Nonlinear Functional Analysis and its Applications. Volume I: Fixed-Point Theorems*, Springer-Verlag, New York, NY, 1986.

[52] I. Ekeland and R. Teman, *Convex Analysis and Variational Problems*, North-Holland, Amsterdam, 1976.

[53] J.M. Ortega and W.C. Rheinboldt, *Iterative Solution of Nonlinear Equations in Several Variables*, Academic Press, New York, NY, 1970.

[54] J.M. Ortega, and W.C. Rheinboldt, Local and global convergence of generalized linear iterations. In Ortega, J.M., and Rheinboldt, W.C. (eds.) *Numerical Solution of Nonlinear Problems*, SIAM, Philadelphia, PA, 1970.

[55] J.M. Ortega and W.C. Rheinboldt, A general convergence result for unconstrained minimization methods, *SIAM J. Num. Anal.*, **9** (1972), 40–43.

[56] M. Avriel, *Nonlinear Programming: Analysis and Methods*, Prentice-Hall, Englewood Cliffs, NJ, 1976.

[57] M.J.D. Powell, Some convergence properties of the conjugate gradient method, *Math. Program.*, **11** (1976), 42–49.

[58] M.J.D. Powell, Restart procedures of the conjugate gradient method, *Math. Program.*, **12** (1977), 148–162.

[59] V. Girault and P.A. Raviart, *Finite Element Methods for Navier-Stokes Equations: Theory and Algorithms*, Springer-Verlag, Berlin, 1986.

[60] J.B. Hiriart-Urruty and C. Lemarechal, *Convex Analysis and Minimization Algorithms*, Springer-Verlag, Berlin, 1993.

[61] M. Crouzeix, Etude d'une méthode de linéarisation. Résolution numérique des équations de Stokes stationaires. In *Approximations et Méthodes Itératives de Résolution d'Inéquations Variationelles et de Problèmes Non Linéaires,* Cahiers de l'IRIA, **12** (1974), 139-244.

[62] M. Crouzeix, On an operator related to the convergence of Uzawa's algorithm for the Stokes equation. In Bristeau, M.O., Etgen, G., Fitzgibbon, W., Lions, J.L., Périaux, J., and Wheeler, M.F. (eds.) *Computational Science for the 21st Century,* Wiley, Chichester, 1997, 242–259.

[63] J. Cahouet and J.P. Chabard, Some fast 3-D solvers for the generalized Stokes problem, *Int. J. Numer. Meth. in Fluids*, **8** (1988), 269–295.

[64] J.E. Dennis and R.B. Schnabel, A view of unconstrained optimization. In Newhauser, G.L., Rinnooy Kan, A.H.G., and Todd, M.J. (eds.) *Handbook in Operations Research and Management Science,* **Vol. 1***: Optimization,* North-Holland, Amsterdam, 1989, 1–66.

[65] F. Thomasset, *Implementation of Finite Element Methods for Navier-Stokes Equations*, Springer-Verlag, New York, NY, 1981.

[66] R. Peyret and T.D. Taylor, *Computational Methods for Fluid Flow*, Springer-Verlag, New York, NY, 1982.

[67] C. Cuvelier, A. Segal, and A. Van Steenhoven, *Finite Element Methods and Navier-Stokes Equations*, Reidel, Dordrecht, 1986.

[68] M. Fortin, Finite element solution of the Navier-Stokes equations, *Acta Numerica 1993*, Cambridge University Press, 1993, 239–284.

[69] M.D. Gunzburger, *Finite Element Methods for Viscous Incompressible Flows*, Academic Press, Boston, MA, 1989.

[70] C.A.J. Fletcher, *Computational Techniques for Fluid Dynamics, Volume 1: Fundamental and General Techniques*, Springer-Verlag, Berlin, 1991.

[71] C.A.J. Fletcher, *Computational Techniques for Fluid Dynamics, Volume 2: Specific Techniques for Different Flow Categories*, Springer-Verlag, Berlin, 1991.

[72] M.D. Gunzburger and R.A. Nicolaides (eds.), *Incompressible Computational Fluid Dynamics,* Cambridge University Press, New York, NY, 1993.

[73] L. Quartapelle, *Numerical Solution of the Incompressible Navier-Stokes Equations*, Birkhauser, Basel, 1993.

[74] F.K. Hebeker, R. Rannacher, and G. Wittum (eds.), *Numerical Methods for the Navier-Stokes Equations*, Vieweg, Braunschweig/Wiesbaden, 1994.

[75] P.M. Gresho and R.L. SANI, *Incompressible Flow and the Finite Element Method: Advection-Diffusion and Isothermal Laminar Flow,* J. Wiley, Chichester, 1998.

[76] E. Fernandez-Cara and M.M. Beltran, The convergence of two numerical schemes for the Navier-Stokes equations, *Numerische Mathematik*, **55** (1989), 33–60.

[77] P. Kloucek and F.S. Rys, On the stability of the fractional step-θ-scheme for the Navier-Stokes equations, *SIAM J. Num. Anal.*, **31** (1994), 1312–1335.

[78] P. Hood and C. Taylor, A numerical solution of the Navier-Stokes equations using the finite element technique, *Computers and Fluids*, **1** (1973), 73–100.

[79] M. Bercovier and O. Pironneau, Error estimates for finite element method solution of the Stokes problem in the primitive variables, *Numer. Math.*, **33** (1979), 211–224.

[80] G. Strang and G. Fix, *An Analysis of the Finite Element Method*, Prentice Hall, Englewood Cliffs, NJ, 1973.

[81] P.G. Ciarlet, Basic error estimates for elliptic problems. In Ciarlet, P.G., and Lions, J.L. (eds.) *Handbook of Numerical Analysis*, **Vol. II**, North-Holland, Amsterdam, 1991, 17–351.

[82] R. Glowinski, Finite element methods for the numerical simulation of incompressible viscous flow. Introduction to the control of the Navier-Stokes equations. In Anderson, C.R., and Greengard, C. (eds.) *Vortex Dynamics and Vortex Methods*, Lecture in Applied Mathematics, Vol. 28, American Mathematical Society, Providence, RI, 1991, 219–301.

[83] M.O. Bristeau, R. Glowinski, and J. Periaux, Numerical methods for the Navier-Stokes equations. Applications to the simulation of compressible and incompressible viscous flow, *Computer Physics Reports*, **6** (1987), 73–187.

[84] E.J. Dean, R. Glowinski, and C.H. Li, Supercomputer solution of partial differential equation problems in Computational Fluid Dynamics and in Control, *Computer Physics Communications*, **53** (1989), 401–439.

[85] R. Glowinski and O. Pironneau, Finite element methods for Navier-Stokes equations, *Annual Review of Fluid Mechanics*, **24** (1992), 167–204.

[86] T.J.R. Hughes, L.P. Franca, and M. Balestra, A new finite element formulation for Computational Fluid Dynamics: V. Circumventing the Babaska-Brezzi Condition; A stable Petrov-Galerkin formulation of the Stokes problem accomodating equal-order interpolation, *Comp. Meth. Appl. Mech. Eng.*, **59** (1986), 85–100.

[87] J. Douglas and J. Wang, An absolutely stabilized finite element method for the Stokes problem, *Math. Comp.*, **52** (1989), 495–508.

[88] Z. Cai and J. Douglas, An analytic basis for multigrid methods for stabilized finite element methods for the Stokes problem. In Bristeau, M.O., Etgen, G., Fitzgibbon, W., Lions, J.L., Periaux, J., and Wheeler, M.F. (eds.) *Computational Science for the 21st Century*, Wiley, Chichester, 1997, 113–118.

[89] R. Glowinski and C.H. Li, On the numerical implementation of the Hilbert Uniqueness Method for the exact boundary controllability of the wave equation, *C.R. Acad. Sc., Paris*, t. 311 (1990), Série I, 135-142.

[90] Glowinski, R., C.H. Li, and J.L. Lions, A numerical approach to the exact boundary controllability of the wave equation (I) Dirichlet controls: Description of the numerical methods, *Japan J. Applied Math.*, **7** (1990), 1–76.

[91] R. Glowinski and J.L. Lions, Exact and approximate controllability for distributed parameter systems, Part II, *Acta Numerica 1995*, Cambridge University Press, 1995, 159–333.

[92] M. Crouzeix and P.A. Raviart, Conforming and nonconforming finite element methods for solving the stationary Stokes equations, *Revue*

Française d'Automatique, Informatique et Recherche Opérationnelle, **R3** (1973), 33–76.

[93] J.E. Roberts and J.M. Thomas, Mixed and hybrid methods. In Ciarlet, P.G., and Lions, J.L. (eds.) *Handbook of Numerical Analysis*, **Vol. II**, North-Holland, Amsterdam, 1991, 523–639.

[94] R. Verfurth, Error estimates for a mixed finite element approximation of the Stokes problem, *Revue Française d'Automatique, Informatique et Recherche Opérationelle, Anal. Numer.*, **18** (1984), 175–182.

[95] R. Glowinski, T.W. Pan, T.I. Hesla, and D.D. Joseph, A distributed Lagrange multiplier/fictitious domain method for particulate flow, *Int. J. Multiphase Flow*, **25** (1999), 755–794.

[96] R. Glowinski, T.W. Pan, T.I. Hesla, D.D. Joseph, and J. Periaux , A distributed Lagrange multiplier/fictitious domain method for flows around moving rigid bodies: Application to particulate flow, *Int. J. Numer. Meth. in Fluids,* **30** (1999), 1043–1066.

[97] R. Glowinski, T.W. Pan, T.I. Hesla, D.D. Joseph, and J. Periaux, A distributed Lagrange multiplier/fictitious domain method for the simulation of flow around moving rigid bodies: Application to particulate flow, *Comp. Meth. Appl. Mech. Eng.,* **184** (2000), 241–267.

[98] R. Glowinski, T.W. Pan, T.I. Hesla, D.D. Joseph, and J. Periaux, A fictitious domain approach to the direct numerical simulation of incompressible viscous fluid flow past moving rigid bodies: Application to particulate flow, *J. Comp. Phys.,* **169** (2001), 363–426.

[99] S. Turek, A comparative study of time-stepping techniques for the incompressible Navier-Stokes equations: from fully implicit non-linear schemes to semi-implicit projection methods, *Int. J. Num. Math. in Fluids*, **22**(1996), 987-1011.

[100] E.J. Dean, R. Glowinski. A wave equation approach to the numerical solution of the Navier-Stokes equations for incompressible viscous flow, *C.R. Acad. Sci. Paris*, t. 325, Série I, (1997), 783–791.

[101] E. Dean, R. Glowinski, T.-W. Pan. A wave equation approach to the numerical simulation of incompressible viscous fluid flow modeled by the Navier-Stokes equations, in *Mathematical and numerical aspects of wave propagation*, J. De Santo ed., SIAM, Philadelphia, 1998, 65–74.

[102] R. Glowinski, O. Pironneau, Finite Element Methods for Navier-Stokes Equations, *Annu. Rev. Fluid Mech.*, **24**(1992), 167–204.

[103] C. Johnson, Streamline diffusion methods for problems in fluid mechanics, in *Finite Element in Fluids 6*, R. Gallagher ed., Wiley, 1986.

[104] U. Ghia, K.N. Ghia, and C.T. Shin, High-Reynolds solutions for incompressible flow using Navier-Stokes equations and a multigrid method, *J. Comp. Phys.*, **48**(1982), 387–411.

[105] R. Schreiber, H.B. Keller, Driven cavity flow by efficient numerical techniques, *J. Comp. Phys.*, **40**(1983), 310–333.

[106] C.H. Bruneau and C. Jouron, Un nouveau schéma décentré pour le probléme de la cavité entraînée, *C.R. Acad. Sci. Paris*, t. 307, Série I (1988), 359–362.

[107] J. Shen, Hopf bifurcation of the unsteady regularized driven cavity flow, *J. Comp. Phys.*, **95**(1991), 228–245.

[108] O. Goyon, High-Reynolds number solutions of Navier-Stokes equations using incremental unknowns, *Comput. Methods Appl. Mech. Engrg.*, **130** (1996), 319–335.

[109] S. Fujima, M. Tabata, Y. Fukasawa, Extension to three-dimensional problems of the upwind finite element scheme based on the choice up- and downwind points, Comp. Meth. Appl. Mech. Eng., **112**(1994), 109–131.

[110] H.C. Ku, R.S. Hirsh, T.D. Taylor, A pseudospectral method for solution of the three-dimensional incompressible Navier-Stokes equations, *J. Comp. Phys.*, **70**(1987), 439–462.

[111] T.P. Chiang, W.H. Sheu, R.R. Hwang, Effect of Reynolds number on the eddy structure in a lid–driven cavity, *Int. J. Numer. Meth. in Fluids,* **26**(1998), 557–579.

[112] A.K. Prasad and J.R. Koseff, Renolds number and end-wall efftects on a lid-driven cavity flow, *Phys. Fluids*, **A 1** (1989), 208–218.

Bol. Soc. Esp. Mat. Apl. nº 37(2006), 47–70

Recent results on stabilization of PDEs by noise

Tomás Caraballo

Dpto. Ecuaciones Diferenciales y Análisis Numérico, Universidad de Sevilla, Apdo. Correos 1160, 41080–Sevilla, Spain.

caraball@us.es

Dedicated to Ana María Heredia[1] with my utmost Gratitude and Affection

Abstract

This paper is intended to be a brief review on some recent results on the stabilization effect produced by noise in phenomena modelled by partial differential equations. We emphasyse the different effects that distinct interpretations of the noise may cause on the same system, and we focus on two classical and canonical interpretations (Itô versus Stratonovich). Finally, we comment on some open problems.

Key words: *exponential stability stabilization Ito's noise Stratonovich's noise stochastic PDE.*

AMS subject classifications: *35R10 35B40 47H20 58F39 73K70*

1 Introduction: Why is Stratonovich noise more significant than Itô noise?

The use of stochastic partial differential equations in physics, chemistry, biology, economics, engineering, etc. is widespread. The addition of random elements (noise) is based on the assumption that such equations are a better model of reality than their deterministic counterparts. Depending on the situation, one can find arguments justifying either of the canonical choices of noise (Itô or Stratonovich). We will not discuss this is detail here, but will emphasyse that these different types of noise can

[1]Quiero dedicar este trabajo a Doña Ana María Heredia, como muestra de mi gratitud y cariño. Tuve la oportunidad de disfrutar de sus enseñanzas de verdadera MAESTRA DE ESCUELA durante los dos últimos años de mis estudios de primaria, y desde entonces tengo la suerte de poder contar con su amistad y aprecio. Ella es una de las personas "culpables" de que hoy pueda estar dedicándole este modesto homenaje.

produce solutions with very different long-time behaviours (see Caraballo & Langa [11]).

A fundamental question is the following: assuming that the real world is actually non-deterministic, are deterministic models good approximations? If the answer is affirmative, then the use of such models would be justified. Otherwise, then in some situations the addition of noise could produce dramatic changes in the behaviour. Here, we analyse the long-time behaviour of the solutions and investigate the potential stabilization (or destabilization) effect of the addition of noise.

In the finite-dimensional context, there is a wide literature about such problems; see Arnold [4], Arnold et al. [6], Mao [34], Scheutzow [38], etc. Many results on the stabilization and destabilization produced by both Itô and Stratonovich noise have been obtained, and these have also been applied to construct feedback stabilisers (an important tool in control problems). Although in each particular situation one or other choice of the noise may be more appropriate, it is stabilization by Stratonovich noise that might be more significant. To explain this in more detail, let us consider the linear n-dimensional ordinary differential equation

$$\dot{x} = Ax, \tag{1}$$

where A could have unstable directions, and the following stochastic versions of this equation, corresponding to the two different interpretations of the stochastic integral

$$\mathrm{d}x = Ax\,\mathrm{d}t + \sigma x \circ \mathrm{d}W(t) \qquad \text{(Stratonovich)} \tag{2}$$

$$\mathrm{d}y = Ay\,\mathrm{d}t + \sigma y\,\mathrm{d}W(t), \qquad \text{(Itô)} \tag{3}$$

where $W(t)$ is a standard Wiener process on the complete probability space $(\Omega, \mathcal{F}, \mathbb{P})$ (see e.g. Arnold [2] for the constructions and properties of both types of differential equations). The initial value problems for (2) and (3) corresponding to the data $x(0) = x_0$, $y(0) = y_0$, can be solved explicitly, and their solutions are given by

$$x(t) = \mathrm{e}^{\sigma W(t)} \exp(tA)x_0 \quad \text{and} \quad y(t) = \mathrm{e}^{-\frac{\sigma^2}{2}t + \sigma W(t)} \exp(tA)y_0.$$

Taking into account the properties of the Wiener process (see again Arnold [2]), it follows that, for σ large enough, the zero solution is exponentially stable for the Itô equation (with probability one), while the same is not true for the Stratonovich equation. This seems to imply that Itô noise has a more profound stabilising effect than Stratonovich noise.

However, this argument is somewhat misleading. Indeed, (1) can be obviously stabilised by using a simple deterministic feedback control, i.e., the new equation

$$\dot{x} = Ax + \lambda x \tag{4}$$

becomes exponentially stable provided $\lambda < 0$ and $|\lambda|$ is large enough. Similarly, it could be stabilised with the periodic control $\lambda(t)x$,

$$\dot{x} = Ax + \lambda(t)x,$$

where $\lambda(t) = \lambda_0 + \sin t$, with $\lambda_0 < 0$ and $|\lambda_0|$ large enough. The function $\sin t$ is, in some sense, a mean-zero function. The same is true with faster mean-zero periodic fluctuations: stabilization takes place because of the systematic dissipative term $\lambda_0 x$ in

the equation, while the mean-zero property means that the other term has no influence on the asymptotic behaviour. We can write such an equation in the form

$$\dot{x} = Ax + \lambda_0 x + x\dot{W}_\varepsilon(t)$$

where $\dot{W}_\varepsilon(t)$ denotes the zero-mean periodic term and may be considered as a physically realistic approximation of the ideal white noise $\dot{W}(t)$. It is therefore not surprising that the same result is true in the limit when the regular mean-zero fluctuations tend to a mean-zero white noise, and the systematic term $\lambda_0 x$ is still present. Now, it is well known that in such a limit (more precisely in all cases where there are rigorous results concerning the Wong-Zakai [40] approximation of a stochastic equation by a random equation with regularised noise), the correct stochastic interpretation for the equation is the Stratonovich one (see, e.g., Sussmann [39] for a more detailed analysis):

$$\mathrm{d}x = Ax\,\mathrm{d}t + \lambda_0 x\,\mathrm{d}t + \sigma x \circ \mathrm{d}W(t),$$

where $\sigma > 0$ describes the intensity of the noise. While it should be intuitively clear that this equation is exponentially stable when $\lambda_0 < 0$ is sufficiently small, a rigorous proof follows from Itô's formula, since as we previously mentioned an explicit form for the solution is

$$x(t) = \mathrm{e}^{\lambda_0 t + \sigma W(t)} \exp(At)x(0).$$

Notice that this property is independent of σ. However, the previous Stratonovich equation can be rewritten in its equivalent Itô form:

$$\mathrm{d}x = Ax\,\mathrm{d}t + \lambda_0 x\,\mathrm{d}t + \sigma x\,\mathrm{d}W(t) + \frac{\sigma^2}{2} x\,\mathrm{d}t.$$

If we choose a white noise with intensity such that $\frac{\sigma^2}{2} = -\lambda_0$, we arrive at the Itô equation

$$\mathrm{d}x = Ax\,\mathrm{d}t + \sigma x\,\mathrm{d}W(t). \tag{5}$$

As a consequence of this elementary analysis, it is clear that this equation is exponentially stable with probability one for σ large enough.

In general, what this example means is that Itô equations with multiplicative noise correspond to the limit of deterministic equations with a mean-zero fluctuating control plus a stabilising systematic control. So, the fact that an Itô equation such as (5) is exponentially stable, even if the equation (1) is not, should not be much more surprising than the fact that (4) with sufficiently small $\lambda < 0$ is exponentially stable; it is only that the mathematics required for the proof are more elaborate.

There is a non-trivial literature on stabilization by *Stratonovich* noise, with both mathematical and engineering contributions (see [4],[6] and the references therein). Since such a noise behaves like a periodic zero-mean feedback control, its stabilising effect is unexpected and very intriguing. In the finite-dimensional case, Arnold and his collaborators have proved that the linear differential system (1) can be stabilised by the addition of a collection of multiplicative noisy terms,

$$\mathrm{d}x = Ax\,\mathrm{d}t + \sum_{i=1}^{d} B_i x \circ \mathrm{d}W_i(t), \tag{6}$$

where the W_i are mutually independent Wiener processes and the B_i are suitable skew-symmetric matrices, if and only if

$$\operatorname{tr} A < 0. \tag{7}$$

Note that the form of the noise is more complex than just a single multiplicative term of the form $\sigma x \circ \mathrm{d}W_t$.

The corresponding problem for linear partial differential equations has remained open for a long time, perhaps because one avenue would be to follow a similar approach but in the infinite-dimensional case, e.g. by proving a version of the celebrated Oseledec Multiplicative Ergodic Theorem for infinite-dimensional spaces. Fortunately, in [19] a very simple argument is successfully used to show that one can obtain a stabilization result for a linear PDE

$$\frac{\mathrm{d}u}{\mathrm{d}t} = Au$$

with a finite sum of Stratonovich terms as in (6).

However, to the best of our knowledge, the problem of stabilization of nonlinear PDEs by Stratonovich noise is still an unsolved problem in general, while it has already been solved, in various interesting applications, by using Itô's noise (see [9] ,[11],[16] ,[14],[17] ,[30],[32] amongst many others).

On the other hand, it is worth mentioning that the analysis carried out in this field is only concerned with the stabilization of the trivial solution. Nevertheless, going deeper in the investigation of the nonlinear models, we can find in the stochastic models some special solutions called *stationary* but which are not stationary in the deterministic sense. These solutions sometimes become random attractors for some systems, so their existence and properties are very important. Some preliminary results have been obtained in [10].

The aim of this paper is to review on the recent results obtained in this field, to outline the basic techniques used in this area, and to comment on some open questions. To be more precise, we consider a linear evolution equation on a separable Hilbert space H given by

$$\frac{\mathrm{d}u}{\mathrm{d}t} = Au, \tag{8}$$

where A is a linear (unbounded) operator, i.e., $A : D(A) \subset H \mapsto H$, and the stochastically perturbed evolution equation

$$\mathrm{d}u = Au\,\mathrm{d}t + \sum_{i=1}^{N} B_i u \circ \mathrm{d}W_i, \tag{9}$$

where the $B_i : D(B_i) \subset H \mapsto H$ are linear operators and the W_i are mutually independent Wiener process on the same probability space $(\Omega, \mathcal{F}, \mathbb{P})$. In Section 2 we first prove that the stability of (8) and (9) are equivalent if the operators B_i and A are mutually commuting and satisfy other additional assumptions (we call this situation "fully commuting"). It is remarkable that, if the noise is considered in the Itô sense, it may produce stabilization or even destabilization under the same assumptions (see Caraballo & Langa [11] for a detailed analysis). Then, we prove that (9) becomes exponentially stable with probability one (w.p.1) for suitable operators B_i if and only if the trace of A is negative, an infinite-dimensional analogue of the results of Arnold et al. [6]. In Section 3 we consider the nonlinear framework and split our analysis into two cases. First, we show how the theory of stabilization has been widely developed by considering the noise in the Itô sense, having been applied to several interesting examples arising in applications. Next, we consider the stabilization problem for nonlinear PDEs by using Stratonovich noise and we show how to prove this fact in a canonical example as the Chafee-Infante equation in one spatial dimension. Also, for this example, it can be shown a "super-stabilization" effect produced by a rich enough

additive noise. We end the paper with some additional comments and stating some open problems.

2 Stabilization and destabilization of a linear PDE

In this section we first establish some results concerning the exponential stability of the null solution to a linear stochastic PDE. Our main purposes are, on the one hand, to point out the different effects that the interpretation of the noise may produce in the final results, and on the other, to characterise the stabilization of linear PDEs by Stratonovich noise.

To start with, we can consider the problem in the Itô formulation, so that we can apply a result due to Da Prato & Zabczyk [25] which ensures the equivalence of the stochastic PDE to a nonautonomous deterministic equation depending on a random parameter, i.e. a random PDE. Then, we will transform our Stratonovich model to an equivalent[2] Itô model and will apply this result.

Let us consider the Cauchy problem

$$\begin{cases} \mathrm{d}u = Au\,\mathrm{d}t + \sum_{k=1}^{d} B_k u\,\mathrm{d}W_k, \\ u(0) = u_0 \in H, \end{cases} \tag{10}$$

where $A : D(A) \subset H \to H$, $B_k : D(B_k) \subset H \to H$, $k = 1, 2, \cdots, d$ are generators of C_0-semigroups $S_A(t)$ and S_k respectively, and the $W_1, \cdots, W_d$ are independent real Wiener processes on the same probability space $(\Omega, \mathcal{F}, \mathbb{P})$.

We need the following additional assumptions:

- (A1) The operators $B_1, \cdots, B_d$ generate mutually commuting C_0-groups S_k.
- (A2) $D(B_k^2) \supset D(A)$ for $k = 1, \cdots, d$ and $\bigcap_{k=1}^{d} D((B_k^*)^2)$ is dense in H, where B_k^* denotes the adjoint operator of B_k.
- (A3) $C = A - \frac{1}{2}\sum_{k=1}^{d} B_k^2$ generates a C_0-semigroup S_C.

Given a fixed realisation of our Wiener processes $W_k(t, \omega)$, $\omega \in \Omega$, in order to solve (10) we define

$$U_\omega(t) = \prod_{k=1}^{d} S_k(W_k(t, \omega)) \quad \text{and} \quad v(t) = U_\omega^{-1}(t)u(t), \qquad t \geq 0, \tag{11}$$

and we consider the auxiliary system

$$\begin{cases} \dot{v}(t) = U_\omega^{-1}(t) C U_\omega(t) v(t) \\ v(0) = u_0, \end{cases} \tag{12}$$

which is a deterministic Cauchy problem depending on the parameter ω. The following result, along with the definition of a strong solution, can be found in Da Prato & Zabczyk [25]:

[2] This equivalence has been proved by Kunita [28] for suitable partial differential operators. We implicitly assume that we are considering this case. It is undoubtedly an important task to develop a general theory of stochastic PDEs in the Stratonovich sense.

Proposition 1 *Let assumptions* (A1)–(A3) *be satisfied. Then, if u is a strong solution to* (10)*, the process* $v(t,\omega)$ *defined by* (11) *satisfies* (12)*. Conversely, if v is a predictable process whose trajectories are continuously differentiable and satisfy* (12) $\mathbb{P}$*-a.s., then the process* $u(t,\omega) = U_\omega(t)v(t,\omega)$ *takes values in* $D(C)$ $\mathbb{P}$*-a.s and for almost all t, and is a strong solution to* (10)*.*

Remark 1 One can also find in Da Prato & Zabczyk [25] a sufficient condition ensuring the solvability of (12) which could be useful in applications (see [25, pp. 177–179] for more details).

Now, we consider the Stratonovich version of the problem:

$$\begin{cases} \mathrm{d}u = Au\,\mathrm{d}t + \sum\limits_{k=1}^{d} B_k u \circ \mathrm{d}W_k, \\ u(0) = u_0 \in H. \end{cases} \tag{13}$$

To ensure existence of solutions to this problem, we can consider its equivalent Itô version

$$\begin{cases} \mathrm{d}u = (A + \frac{1}{2}\sum\limits_{k=1}^{d} B_k^2)u\,\mathrm{d}t + \sum\limits_{k=1}^{d} B_k u\,\mathrm{d}W_k, \\ u(0) = u_0 \in H. \end{cases} \tag{14}$$

If we now assume (A1)–(A2) and, instead of (A3), the following

- (A3') $C = A + \frac{1}{2}\sum_{k=1}^{d} B_k^2$, generates a C_0-semigroup S_C,

then, thanks to Proposition 1, problem (13) can be equivalently rewritten as

$$\begin{cases} \dot{v}(t) = U_\omega^{-1}(t)AU_\omega(t)v(t) \\ v(0) = u_0. \end{cases} \tag{15}$$

We can now proceed with our stability analysis.

2.1 Itô vs Stratonovich in the fully commuting case

Now we establish a result which characterises the asymptotic stability of the Stratonovich model (13) under the assumption that all the operators involved in the equation mutually commute (we call this the "fully commuting" case) and weshow how, under the same assumptions, the Itô equation (10) may exhibit very different asymptotic behaviour.

The next result can be found in [19] (see also [11]).

Theorem 2 *In addition to assumptions* (A1)–(A2) *and* (A3')*, suppose that A commutes with each* $S_k(t)$*. Then the strongly continuous semigroup* $S_A(t)$ *generated by A is exponentially stable, i.e. there exist* $M_0, \gamma > 0$ *such that* $|S_A(t)| \leq M_0 e^{-\gamma t}$ *for all* $t > 0$*, if and only if there exist* $\alpha, C > 0$ *and* $\Omega_0 \subset \Omega$ *with* $\mathbb{P}(\Omega_0) = 0$ *such that for any* $\omega \notin \Omega_0$ *there exists* $T(\omega) > 0$ *such that the following holds for the solution of* (13)*:*

$$|u(t)| \leq C|u_0|e^{-\alpha t} \quad \textit{for} \;\; t \geq T(\omega)\ \mathbb{P} - a.s.$$

Proof. (Sketch) Denote by $u(t) = u(t, \omega; 0, u_0)$ the solution of (13) for $u_0 \in D(A)$, and by $v(t) = v(t, \omega; 0, u_0)$ the corresponding solution to (15), i.e. $v(t) = U_\omega^{-1}(t)u(t)$. Owing to the commutativity of the operator A and the operators S_k, the problem (15) can be written as

$$\begin{cases} \dot{v}(t) = Av(t), \\ v(0) = u_0 \in D(A), \end{cases} \tag{16}$$

whose solution is given by $v(t) = S_A(t)u_0$, so we have an explicit expression for our solution $u(t)$:

$$u(t) = u(t, \omega; 0, u_0) = U_\omega(t)S_A(t)u_0. \tag{17}$$

Taking into account now the properties of strongly continuous semigroups and the Wiener processes (especially that $\lim_{t\to+\infty} |W_k(t, \omega)|/t = 0$, for all $k = 1, \cdots, d$), it is not difficult to conclude the proof. □

Consequently, in this fully commuting case the stability properties of the deterministic problem (8) and the stochastic (13) are equivalent, so we can ensure the adequacy of the deterministic model to the stochastic real phenomenon. However, if we interpret the noise in the sense of Itô, we may have very different results. It may happen that (8) is stable and (10) remains stable (persistence of stability from the deterministic to the stochastic model), or (8) is unstable and (10) becomes stable (stabilization produced by the noise), or (8) is stable and (10) becomes unstable (destabilization), etc. Let us now illustrate these facts.

2.1.1 Persistence of stability and stabilization by Itô noise

To illustrate these features, we will analyse the following example. Let $\mathcal{O}$ be a bounded domain in $\mathbb{R}^d$ ($d \leq 3$) with C^∞-boundary, and consider the reaction-diffusion equation

$$\begin{cases} \mathrm{d}u(t, x) = (\Delta u(t, x) + \alpha u(t, x))\,\mathrm{d}t + \gamma u(t, x)\,\mathrm{d}W(t), \\ u(t, x) = 0, \quad t > 0, \quad x \in \partial\mathcal{O}, \\ u(0, x) = u_0(x), \quad x \in \mathcal{O}, \end{cases} \tag{18}$$

where, as usual, Δ denotes the Laplacian operator and $W(t)$ is a scalar Wiener process. To set this problem in our framework, we take $H = L^2(\mathcal{O})$, $A = \Delta + \alpha I$ and $B = \gamma I$. It then holds that $D(A) = H_0^1(\mathcal{O}) \cap H^2(\mathcal{O})$. Let $\lambda_1 > 0$ denote the first eigenvalue of $-\Delta$. Then, as a consequence of Theorem 3 in Subsection 2.2 (see also Kwiecinska [29]), it is easy to check that the null solution of (18) is exponentially stable $\mathbb{P}$-a.s. if the parameters in the equation satisfy

$$2(\alpha - \lambda_1) - \gamma^2 < 0.$$

Let us now discuss what this condition means.

First, notice that when $\alpha < \lambda_1$, the deterministic equation (i.e. Eq. (18) with $\gamma = 0$) is exponentially stable. Then, for any $\gamma \in \mathbb{R}$ (in other words, no matter how large or small the intensity of the noise might be), the stochastic equation (18) remains exponentially stable $\mathbb{P}$-a.s. So, the persistence of stability takes place in the presence of noise.

However, if $\alpha > \lambda_1$, then the deterministic equation is not stable (see, for instance, Example 2.1.2 below for a more detailed analysis in a case of one spatial dimension). But, if we choose γ large enough so that $2(\alpha - \lambda_1) - \gamma^2 < 0$, then the stochastic equation becomes exponentially stable $\mathbb{P}$-a.s. Consequently, a large intensity of the noise has produced a stabilization effect on the system.

2.1.2 Destabilization produced by Itô noise

Consider the deterministic heat equation but now in one spatial dimension:

$$\begin{cases} \dfrac{\partial u(t,x)}{\partial t} = \dfrac{\partial^2 u(t,x)}{\partial x^2} + \alpha u(t,x), \quad t>0, \quad 0<x<\pi, \\ u(t,0) = u(t,\pi) = 0, \quad t>0, \\ u(0,x) = u_0(x), \quad x \in [0,\pi]. \end{cases} \tag{19}$$

Set again $H = L^2([0,\pi])$ and let $A = \frac{\partial^2}{\partial x^2} + \alpha$. It then follows that $D(A) = H_0^1([0,\pi]) \cap H^2([0,\pi])$. Notice that this system can be explicitly solved and its solution is given by

$$u(t,x) = \sum_{n=1}^{\infty} a_n e^{-(n^2-\alpha)t} \sin nx,$$

where $u_0(x) = \sum_{n=1}^{\infty} a_n \sin nx$. Hence, exponential stability holds if and only if $\alpha < n^2$ for all $n \in \mathbb{N}$, i.e. if and only if $\alpha < 1$.

Consider now the problem

$$\begin{cases} \mathrm{d}u(t,x) = Au(t,x)\,\mathrm{d}t + Bu(t,x)\,\mathrm{d}W(t), \\ u(0,x) = u_0(x), \end{cases} \tag{20}$$

where B is defined by $Bu(x) = \delta \frac{\partial u(x)}{\partial x}$, for any $u \in H_0^1([0,\pi])$, $\delta \in \mathbb{R}$.

We will show that, if we choose δ such that

$$\frac{\delta^2}{2} < 1 \quad \text{and} \quad \frac{\delta^2}{2} - 1 + \alpha \geq 0,$$

then the stochastic problem becomes unstable.

Indeed, denoting by $C = A - \frac{1}{2}B^2$, the stability of problem (20) is equivalent to the stability of

$$\begin{cases} \mathrm{d}u(t,x) = Cu(t,x)\,\mathrm{d}t + Bu(t,x) \circ \mathrm{d}W(t) \\ u(0,x) = u_0(x). \end{cases} \tag{21}$$

But, due to the commutativity property of the operators involved in the equation, Theorem 2 ensures that the stability of (21) is equivalent to the stability of the deterministic problem

$$\begin{cases} \dfrac{\partial u(t,x)}{\partial t} = (1 - \dfrac{\delta^2}{2}) \dfrac{\partial^2 u(t,x)}{\partial x^2} + \alpha u(t,x), \\ u(t,0) = u(t,\pi) = 0, t>0, \\ u(0,x) = u_0(x), x \in [0,\pi]. \end{cases}$$

This is exponentially stable if and only if $\alpha < 1 - \frac{\delta^2}{2}$. Since our constants satisfy the opposite inequality, we have that the noise has destabilised the deterministic exponentially stable system.

As a conclusion in this fully commuting case, it is thus evident that we should be very careful with the interpretation given to the noise since, depending on that, the behaviours of the deterministic and stochastic models may be completely different. More precisely, the Stratonovich noise does not modify the stability properties of the deterministic model, while Itô noise can produce very different effects.

2.2 Stabilization of a linear PDE in the non-fully commuting case

Notice that under our fully commuting assumptions in the previous subsection, the deterministic problem is exponentially stable if and only if the stochastically perturbed equation has the same property. However, an immediate question arises. What happens if no commutativity holds between A and some B_k? In this case, one can find in [19] some sufficient conditions ensuring the persistence of exponential stability from the deterministic to the stochastic model.

2.2.1 Straightforward stabilization produced by a simple multiplicative Itô noise

First, we point out that a simple multiplicative noise in the sense of Itô can always stabilise the deterministic linear partial differential equation (8) in a lot of cases. So, if we are interested in finding appropriate types of Itô noise to produce stabilization, we do not need to worry too much about looking for a very complicate expression of the noise. Just a term like

$$\sigma u \dot{W}(t)$$

can produce that effect. However, this stabilization can be produced for more general terms and, moreover, we can determine in some cases even the decay rate of the solutions (exponential, sub- or super-exponential, etc).

We now include a result which is a particular situation of a much more general nonlinear theorem (see Section 3 for more details).

First, recall that a linear operator A generates a strongly continuous semigroup $S_A(t)$ satisfying $|S_A(t)| \leq \mathrm{e}^{\alpha t}$, $\alpha \in \mathbb{R}$, if and only if $(Au, u) \leq \alpha |u|^2$, for all $u \in D(A)$.

Theorem 3 *Assume that A generates a strongly continuous semigroup $S_A(t)$ satisfying $|S_A(t)| \leq \mathrm{e}^{\alpha t}$, $\alpha \in \mathbb{R}$, and $B : D(B) \subset H \mapsto H$ is a linear (bounded or unbounded) operator with $D(A) \subset D(B)$. Suppose that the two following hypotheses hold:*

i) There exists $\beta \in \mathbb{R}$ such that

$$(Au, u) + \frac{1}{2}|Bu|^2 \leq \beta |u|^2, \forall u \in D(A) \tag{22}$$

(which is immediately fulfilled for $\beta = \alpha + \frac{1}{2}\|B\|^2$, if B is bounded).

ii) There exists $b, \widetilde{b} \in \mathbb{R}$, with $0 \leq b \leq \widetilde{b}$, such that

$$b|u|^2 \leq (u, Bu) \leq \widetilde{b}|u|^2 \quad \forall u \in D(B). \tag{23}$$

Then, for every $u_0 \in D(A), u_0 \neq 0$, the solution $u(t) = u(t, \omega; 0, u_0)$ to the problem

$$\begin{cases} \mathrm{d}u = Au\,\mathrm{d}t + Bu\,\mathrm{d}W, \\ u(0) = u_0 \in H, \end{cases}$$

satisfies

$$\limsup_{t \to +\infty} \frac{1}{t} \log |u(t; u_0)|^2 \leq -(b^2 - \beta), \quad \mathbb{P} - a.s.$$

Notice that, in the particular case in which B is defined by $Bu = bu$ with $b \in \mathbb{R}$, then $\beta = \alpha + \frac{1}{2}b^2$ and therefore $b^2 - \beta = \frac{1}{2}b^2 - \alpha$, which is positive when $|b|$ is large enough (i.e. large intensity of the noise produces stabilization on the null solution).

2.2.2 Not so straightforward stabilization produced by Stratonovich noise

However, to obtain the same effect using Stratonovich noise turns out to be a completely different and much more difficult problem as commented in the Introduction. But, surprisingly, a very simple trick (discovered long time after the results in the finite-dimensional case were obtained) will allow to prove that the negative trace assumption (7) is a necessary and sufficient condition for the stabilization of a linear PDE by using a suitable Stratonovich noise (see [19] for a detailed exposition on this problem). Instead of establishing and proving this stabilization result, we will motivate the problem with an example on which the proof of the theorem is based.

Consider the following one-dimensional heat equation

$$\begin{cases} \dfrac{\partial u(t,x)}{\partial t} = \dfrac{\partial^2 u(t,x)}{\partial x^2} + 2u(t,x), \quad t>0, \quad 0<x<\pi, \\ u(t,0) = u(t,\pi) = 0, t>0, \\ u(0,x) = u_0(x), \quad x \in [0,\pi]. \end{cases} \tag{24}$$

This problem can be formulated in our framework by setting $H = L^2([0,\pi])$ and $A = \frac{\partial^2}{\partial x^2} + 2I$. It follows that $D(A) = H_0^1([0,\pi]) \cap H^2([0,\pi])$. Recall that this problem can be solved explicitly as in Section 2.1.2, yielding

$$u(t,x) = \sum_{n=1}^{\infty} a_n e^{-(n^2-2)t} \sin nx,$$

where $u_0(x) = \sum_{n=1}^{\infty} a_n \sin nx$. Hence, it is clear that the zero solution of our problem (24) is not stable. But we will choose appropriate operators $B_k : H \to H,\ k = 1, 2, ...d$, such that

$$\begin{cases} \mathrm{d}u(t,x) = Au(t,x)\,\mathrm{d}t + \sum\limits_{k=1}^{d} B_k u(t,x) \circ \mathrm{d}W_k(t) \\ u(0,x) = u_0(x), \end{cases} \tag{25}$$

becomes exponentially stable with probability one. It is worth pointing out that the operators B_k cannot commute with A.

Notice that A possesses a sequence of eigenvalues given by $\lambda_n = 2 - n^2, n \geq 1$, with associated eigenfunctions $e_n = \sqrt{\frac{2}{\pi}} \sin nx$, which form an orthonormal basis of the Hilbert space H. This means that any $u \in H$ can be represented in the form

$$u = \sum_{k\geq 1} (u, e_k) e_k = \sum_{k \geq 1} u_k e_k.$$

Now we define $B : H \mapsto H$ as $Be_1 = -\sigma e_2$, $Be_2 = \sigma e_1$ and $Be_n = 0$ for any $n \geq 3$, which is a linear operator (and does not commute with A). Then, using the Fourier representation for the solution $u(t)$ to (25), our problem can be re-written as

$$\begin{cases} \sum_{k\geq 1} \mathrm{d}u_k(t)\, e_k = \sum_{k\geq 1} \lambda_k u_k(t) e_k\, \mathrm{d}t + (\sigma u_2(t) e_1 - \sigma u_1(t) e_2) \circ \mathrm{d}W(t) \\ u(0) = u_0 = \sum_{k\geq 1} u_{0k} e_k. \end{cases} \tag{26}$$

Identifying the coefficients, we get two coupled problems.The first one is a 2-dimensional stochastic ordinary differential system, and the second one is an infinite-

dimensional system which is exponentially stable since $\lambda_n < 0$ for all $n \geq 3$):

$$\begin{cases} \begin{pmatrix} \mathrm{d}u_1(t) \\ \mathrm{d}u_2(t) \end{pmatrix} = \begin{pmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{pmatrix} \begin{pmatrix} u_1(t) \\ u_2(t) \end{pmatrix} \mathrm{d}t + \begin{pmatrix} 0 & \sigma \\ -\sigma & 0 \end{pmatrix} \begin{pmatrix} u_1(t) \\ u_2(t) \end{pmatrix} \circ \mathrm{d}W(t) \\ \begin{pmatrix} u_1(0) \\ u_2(0) \end{pmatrix} = \begin{pmatrix} u_{01} \\ u_{02} \end{pmatrix}, \end{cases} \tag{27}$$

and

$$\begin{cases} \sum_{k\geq 3} \mathrm{d}u_k(t)e_k = \sum_{k\geq 1} \lambda_k u_k(t) e_k \, \mathrm{d}t, \\ \sum_{k\geq 3} u_k(0)e_k = \sum_{k\geq 3} u_{0k} e_k. \end{cases} \tag{28}$$

Now, since the matrix

$$\begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix}$$

is a basis for the linear space of skew symmetric 2×2 matrices, the results in Arnold *et al.* [6] prove that the leading Lyapunov exponent of solutions to (27) tends to $(\lambda_1 + \lambda_2)/2 = -1/2$ as the parameter σ grows to $+\infty$. As it easily follows that the leading Lyapunov exponent for the solutions to (28) is $\lambda_3 = -7$, we can ensure that the top Lyapunov exponent for the solutions of (26) is negative.

Thus, the main idea for the stabilization is to decompose the problem into two new problems: a finite-dimensional one which can be stabilised by using previously available methods from the finite dimensional framework, and another infinite-dimensional system which is already exponentially stable. This idea can be extended in a general way to solve the stabilization problem for a class of deterministic PDEs which appears very frequently in applications.

Consider the infinite-dimensional linear system

$$\frac{\mathrm{d}u}{\mathrm{d}t} = Au, \tag{29}$$

where $A : D(A) \subset H \mapsto H$ is a linear operator which has a sequence of eigenvalues λ_j with associated eigenfunctions e_j. We assume that these eigenfunctions form an orthonormal basis of the separable Hilbert space H, and that the eigenvalues λ_j are bounded above (but not necessarily below), so that they can be ordered in the form $\lambda_1 \geq \lambda_2 \geq \ldots$. We denote by $|\cdot|$ the norm in H and by $(\cdot, \cdot)$ its associated scalar product.

Now, we can establish our main stabilization result:

Theorem 4 *(Caraballo & Robinson [19]) Assume that the trace of A is negative, i.e.*

$$\sum_{j=1}^{\infty} \lambda_j < 0. \tag{30}$$

Then there exist linear operators $B_k : H \mapsto H$, $k = 1, 2, \ldots, d$, such that, for

$$\mathrm{d}u = Au \, \mathrm{d}t + \sum_{j=1}^{d} B_k u \circ \mathrm{d}W_k, \tag{31}$$

the zero solution is exponentially stable $\mathbb{P}$-a.s. The operators B_k are such that for some $N > 0$, the $N \times N$ matrices $D_1, ..., D_k$ defined as

$$D_k = \begin{pmatrix} (B_k e_1, e_1) & (B_k e_2, e_1) & \cdots & (B_k e_N, e_1) \\ (B_k e_1, e_2) & (B_k e_2, e_2) & \cdots & (B_k e_N, e_2) \\ \vdots & \vdots & \ddots & \vdots \\ (B_k e_1, e_N) & (B_k e_2, e_N) & \cdots & (B_k e_N, e_N) \end{pmatrix}$$

are skew-symmetric.

Conversely, if there exist linear operators $B_k : H \mapsto H$, $k = 1, 2, ..., d$, *with the above properties, for which the zero solution of (31) is exponentially stable with probability one then the trace of* A *is negative.*

3 Stabilization of nonlinear PDEs

The objectives of this section are the following. First, we will show that there exists a well developed theory concerning the stabilization of nonlinear PDEs by Itô noise with applications to several interesting examples. On the other hand, since not much is known about the same topic but involving Stratonovich noise, we will analyse a particular example (which, somehow, can be considered as canonical) in which the previous Theorem 4, jointly with some order preserving properties will allow to prove stabilization for the Chafee-Infante equation by using Stratonovich noise. A more complete study for more general nonlinear equations is to be done.

3.1 Some representative results concerning the stabilization of nonlinear PDEs by Itô noise

First, we introduce the framework where our analysis is going to be carried out.

Let H be a real, separable Hilbert space and V a real, reflexive and separable Banach space such that

$$V \hookrightarrow H \equiv H' \hookrightarrow V',$$

where the injections are continuous and dense. In particular, we also assume that both V and V' are uniformly convex.

We denote by $\|\cdot\|$, $|\cdot|$ and $\|\cdot\|_*$ the norms in V, H and V', respectively; by $\langle\cdot,\cdot\rangle$ the duality product between V, V', and by $(\cdot,\cdot)$ the scalar product in H. Let a_1 be the constant of the injection $V \hookrightarrow H$, i.e.

$$a_1|u|^2 \leq \|u\|^2 \quad \forall u \in V.$$

Consider the following problem

$$\begin{cases} \dfrac{\mathrm{d}u}{\mathrm{d}t} = F(t,u), \\ u(0) = u_0 \in H, \end{cases} \tag{32}$$

where $F(t,\cdot) : V \mapsto V'$, $t \in \mathbb{R}_+$, is a family of (nonlinear) operators satisfying $F(t,0) = 0$ and the following hypothesis:

There exist a continuous function $\nu(\cdot)$ and a real number $\nu_0 \in \mathbb{R}$ such that

$$2\langle u, F(t,u)\rangle \leq \nu(t)|u|^2 \quad \forall u \in V, \tag{33}$$

where

$$\limsup_{t\to\infty} \frac{1}{t}\int_0^t \nu(s)\,ds \leq \nu_0. \tag{34}$$

Assume that for each $u_0 \in H$ there exists a unique strong solution $u = u(t;u_0)$ to (32), with $u(t;u_0) \in L^2(0,T;V) \cap C^0([0,T];H)$. Observe that, when $F(t,\cdot)$ satisfies a coercivity condition of the type

$$2\langle u, F(t,u)\rangle \leq -\varepsilon\|u\|^p + \alpha|u|^2, \ \forall u \in V, \ \varepsilon > 0, \ \alpha \in \mathbb{R}, \ p > 1$$

and a monotonicity hypothesis, there exists a unique strong solution $u = u(t; u_0)$ to (32) in $L^p(0,T;V) \cap C^0([0,T];H)$ (see, for instance, Lions [33]).

Note that this coercivity assumption obviously implies (33).

Now, we will see that (32) can be stabilised by using a stochastic perturbation of the kind $g(t,u(t))\mathrm{d}W(t)$. Here, $W(t)$ is (for simplicity) a standard real Wiener process defined on a certain complete probability space $(\Omega, \mathcal{F}, \mathbb{P})$ with filtration $(\mathcal{F}_t)_{t\geq 0}$ and $g(t,\cdot) : H \to H$ satisfies $g(t,0) = 0$ and the following condition

$$|g(t,u) - g(t,v)|^2 \leq \lambda(t)|u-v|^2 \quad \forall t \in \mathbb{R}_+, \ \forall u, v \in H, \tag{35}$$

where $\lambda(\cdot)$ is a nonnegative continuous function such that

$$\limsup_{t\to\infty} \frac{1}{t} \int_0^t \lambda(s)\, ds \leq \lambda_0 \in \mathbb{R}_+. \tag{36}$$

Indeed, let us consider the following perturbed problem

$$\begin{cases} \mathrm{d}u(t) = F(t,u(t))\ \mathrm{d}t + g(t,u(t))\ \mathrm{d}W(t), \quad t > 0, \\ u(0) = u_0 \in H. \end{cases} \tag{37}$$

We suppose that for each $u_0 \in H$ there exists a unique strong solution to (37) in $I^p(0,T;V) \cap L^2(\Omega; C^0([0,T];H))$ for all $T > 0$ and certain $p > 1$,where $I^p(0,T;V)$ denotes the space of all V-valued measurable processes $u = u(t)$ satisfying $E\int_0^T \|u(t)\|^p\, dt < +\infty$ (see for instance Pardoux [36] for conditions under which there exists a unique solution for each $u_0 \in L^2(\Omega, \mathcal{F}_0, P; H)$).

Assume that $V : \mathbb{R}_+ \times H \to \mathbb{R}_+$ is a $C^{1,2}$-positive functional such that, for any $u \in V$ and $t \in \mathbb{R}_+$, $V'_u(t,u) \in V$. We define the operators L and Q as follows: for each $u \in V$, $t \in \mathbb{R}_+$,

$$LV(t,u) = V'_t(t,u) + < V'_u(t,u), F(t,u) > + \frac{1}{2}\big(V''_{u,u}(t,u)g(t,u), g(t,u)\big)$$

and

$$QV(t,u) = \big(V'_u(t,u), g(t,u)\big)^2.$$

Theorem 5 *(Caraballo et al. [16]) Assume that the solution of* (37) *satisfies that* $|u(t)| \neq 0$ *for all* $t \geq 0$ $\mathbb{P}$*-a.s. provided* $|u_0| \neq 0$ $\mathbb{P}-$*a.s. Let* $V \in C^2(H;\mathbb{R}_+)$ *and let* ψ_1 *and* ψ_2 *be two real-valued continuous functions on* $\mathbb{R}_+$*, with* $\psi_2 \geq 0$*. Assume that there exist* $p > 0$*,* $\gamma \geq 0$ *and* $\theta \in \mathbb{R}$ *such that*

(*a*). $|u|^p \leq V(u) \quad \forall u \in V$;

(*b*). $LV\big(t,u\big) \leq \psi_1(t)V(u), \quad \forall u \in V \quad \forall t \in \mathbb{R}_+$;

(*c*). $QV\big(t,u\big) \geq \psi_2(t)V^2(u), \quad \forall u \in V \quad \forall t \in \mathbb{R}_+$;

(*d*). $\limsup_{t\to\infty} \frac{\int_0^t \psi_1(s)ds}{t} \leq \theta, \qquad \liminf_{t\to\infty} \frac{\int_0^t \psi_2(s)ds}{t} \geq 2\gamma.$

Then the unique strong solution of (37) *satisfies*

$$\limsup_{t\to\infty} \frac{\log\big|u(t,u_0)\big|}{t} \leq -\frac{\gamma-\theta}{p} \qquad \mathbb{P}-a.s.,$$

whenever $u_0 \in H$ *is an* $\mathcal{F}_0$*-measurable random vector such that* $|u_0| \neq 0$ *a.s. In particular, if* $\gamma > \theta$*, the solution is* $\mathbb{P}$*-a.s. exponentially stable.*

Proof. (Sketch) The proof relies on Itô's formula, the exponential martingale inequality and the Borel-Cantelli lemma. To be more precise, let us fix $u_0 \in H$ such that $|u_0| \neq 0$ $\mathbb{P}-$a.s. By Itô's formula,

$$\log V(u(t,u_0)) \leq \log V(u(0)) + M(t) + \int_0^t \Big(\frac{LV(s,u(s))}{V(u(s))} - \frac{1}{2}\frac{QV(s,u(s))}{V^2(u(s))} \Big) ds, \tag{38}$$

where $M(t) = \int_0^t \frac{1}{V(u(s))}(V_x'(u(s)), g(s,u(s)))dw(s)$.

From the exponential martingale inequality, we can deduce that

$$\mathbb{P}\{\, \omega : \sup_{0\leq t\leq w} \Big[M(t) - \int_0^t \frac{u}{2}\frac{1}{V^2(u(s))} QV(s,u(s))ds \Big] > v \,\} \leq e^{-uv}$$

for any positive u, v and w. Assigning $\epsilon > 0$ arbitrarily, taking

$$u = \alpha, \quad v = 2\alpha^{-1}\log k, \quad w = k\epsilon \qquad (k \geq 1)$$

where $0 < \alpha < 1$ and applying the well-known Borel-Cantelli lemma, we see that there exists an integer $k_0(\epsilon,\omega) > 0$ for almost all $\omega \in \Omega$ such that

$$M(t) \leq 2\alpha^{-1}\log k + \frac{\alpha}{2}\int_0^t \frac{QV(s,u(s))}{V^2(u(s))} ds$$

for all $0 \leq t \leq k\epsilon$, $k \geq k_0(\epsilon,\omega)$. Replacing this in (38) and using conditions (b) and (c), we deduce that there exists a positive random integer $k_1(\epsilon)$ such that

$$\log V(u(t)) \leq \log V(u(0)) + 2\alpha^{-1}\log k + \int_0^t \psi_1(s)ds - \frac{1}{2}(1-\alpha)\int_0^t \psi_2(s)ds$$

$\mathbb{P}$-a.s. for all $(k-1)\epsilon \leq t \leq k\epsilon$ and $k \geq \max(k_0(\epsilon,\omega) \vee k_1(\epsilon)$. Now, assumption (d) implies that

$$\frac{\log|u(t)|}{t} \leq \frac{\log V(u(t))}{pt} \leq \frac{1}{pt}\Big(\log V(u(0)) + 2\alpha^{-1}\log k + \big(\theta+\epsilon\big)t - \frac{1}{2}(1-\alpha)\big(2\gamma+\epsilon\big)t \Big).$$

Therefore,

$$\limsup_{t\to\infty} \frac{\log|u(t)|}{t} \leq \frac{1}{p}\Big[\big(\theta+\epsilon\big) - (1-\alpha)\big(\gamma + \frac{\epsilon}{2}\big) \Big] \qquad a.s.$$

Letting $\alpha \to 0$ and $\epsilon \to 0$, we obtain:

$$\limsup_{t\to\infty} \frac{\log|u(t,u_0)|}{t} \leq -\frac{\gamma-\theta}{p} \qquad a.s.$$

□

As a direct consequence of Theorem 5, by using the function $V(t,u) = |u|^2$, we can prove the following:

Theorem 6 *Assume that the solution of* (37) *satisfies* $|u(t,u_0)| \neq 0$ *for all* $t \geq 0$ $\mathbb{P}$*-a.s. provided* $|u_0| \neq 0$ $\mathbb{P}$*-a.s. In addition to hypotheses* (33) – (36)*, assume that*

$$(g(t,u),u)^2 \geq \rho(t)|u|^4 \quad \forall u \in H, \tag{39}$$

where $\rho(\cdot)$ *is a nonnegative continuous function such that*

$$\liminf_{t\to\infty} \frac{1}{t}\int_0^t \rho(s)\,ds \geq \rho_0, \quad \rho_0 \in \mathbb{R}_+. \tag{40}$$

Then, if $u = u(t,u_0)$ *denotes the solution to* (37)*, it follows that*

$$\limsup_{t\to\infty} \frac{1}{t}\log|u(t,u_0)|^2 \leq -(2\rho_0 - \nu_0 - \lambda_0) \quad P-a.s. \tag{41}$$

for any $u_0 \in H$*. In particular, if* $2\rho_0 > \nu_0 + \lambda_0$*, the equation* (37) *is* $\mathbb{P}$*-a.s. exponentially stable.*

3.1.1 Stability properties of a general nonlinear example

Now, we are going to apply Theorem 6 to analyse the pathwise stability of a nonlinear stochastic partial differential equation. Consider the following problem previously studied, among others, by Pardoux [36], Caraballo and Liu [15]:

$$\begin{cases} \mathrm{d}u(t) = A(t,u(t))\ \mathrm{d}t + B(t,u(t))\ \mathrm{d}W(t), \quad t>0, \\ u(0) = u_0 \in H. \end{cases} \tag{42}$$

where $A(t,\cdot) : V \mapsto V'$ is a family of nonlinear operators defined for almost every t (t-a.e. for short), satisfying $A(t,0) = 0$ for $t \in \mathbb{R}_+$; $B(t,\cdot) : V \mapsto H$, satisfies

(b.1) $B(t,0) = 0$;

(b.2) There exists $k > 0$ such that

$$|B(t,y) - B(t,x)| \leq k\|y-x\|, \quad \forall x,y \in V,\ t\text{-a.e.}$$

In [15], the following result is proved:

Theorem 7 *In addition to* (b.1)–(b.2)*, assume the following coercivity condition: there exist* $\alpha > 0$*,* $p > 1$ *and* $\lambda \in \mathbb{R}$ *such that, for almost all* $t \in \mathbb{R}_+$ *and for all* $x \in V$*, one has*

$$2\langle x, A(t,x)\rangle + |B(t,x)|^2 \leq -\alpha\|x\|^p + \lambda|x|^2. \tag{43}$$

Then there exists $r > 0$ *such that*

$$E|u(t;u_0)|^2 \leq E|u_0|^2 e^{-rt} \quad \forall t \geq 0,$$

if at least one of the following hypotheses holds:

(i) $\lambda < 0$*;*

(ii) $\lambda\beta^2 - \alpha < 0$ *and* $p = 2$.

Furthermore, under the same assumptions the solution is $\mathbb{P}$*-a.s. exponentially stable. That is, there exist positive constants* ξ*,* η *and a subset* $\Omega_0 \subset \Omega$ *with* $\mathbb{P}(\Omega_0) = 0$ *such that, for each* $\omega \notin \Omega_0$*, there exists a positive random number* $T(\omega)$ *satisfying*

$$|u(t,\omega;u_0)|^2 \leq \eta|u_0|^2 e^{-\xi t}, \quad \forall t \geq T(\omega).$$

Observe that, in many applications, conditions (i) and (ii) mean that the term containing B must be small enough with respect to A. For example, let $\mathcal{O}$ be an open, bounded subset in $\mathbb{R}^N$ with regular boundary and let $2 \leq p < +\infty$. Consider the Sobolev spaces $V = W_0^{1,p}(\mathcal{O})$, $H = L^2(\mathcal{O})$ with their usual inner products, and the operator $A\colon V \mapsto V'$ defined as

$$\langle v, Au\rangle = -\sum_{i=1}^{N} \int_{\mathcal{O}} \left|\frac{\partial u(x)}{\partial x_i}\right|^{p-2} \frac{\partial u(x)}{\partial x_i} \frac{\partial v(x)}{\partial x_i}\, dx + \int_{\mathcal{O}} au(x)v(x)\, dx \ \ \forall u, v \in V,$$

where $a \in \mathbb{R}$. We also introduce B, with $B(t,u) \equiv bu$, where $b \in \mathbb{R}$. Finally, let $W(t)$ be a standard real Wiener process.

Then,

$$2\langle x, A(t,x)\rangle + |B(t,x)|^2 = -2\|x\|^p + 2a|x|^2 + b^2|x|^2 \ \ \forall x \in V, \tag{44}$$

so (43) holds with equality for $\alpha = 2$ and $\lambda = 2a + b^2$. Now, condition (i) requires $2a + b^2 < 0$, so $a < 0$ and $b^2 < -2a$. On the other hand, (ii) holds whenever $(2a + b^2)a_1^{-1} - 2 < 0$, that is, $b^2 < 2a_1 - 2a$. Therefore, Theorem 7 guarantees the exponential stability of paths $\mathbb{P}$-a.s. only for these values of a and b, which means that the deterministic system $\mathrm{d}u(t) = A(t, u(t))\,\mathrm{d}t$ is exponentially stable and the random perturbation is small enough. However, Theorem 6 ensures exponential stability for sufficiently large perturbations although the deterministic system is unstable. Note that, in this case, it is not difficult to see that

$$2\langle x, A(t,x)\rangle = -2\|x\|^p + 2a|x|^2 \leq \begin{cases} 2a|x|^2, & \text{if } \ p > 2, \\ (2a - 2a_1)|x|^2, & \text{if } \ p = 2, \end{cases} \tag{45}$$

so that

$$\nu(t) = \nu_0 = \begin{cases} 2a, & \text{if } \ p > 2, \\ 2a - 2a_1, & \text{if } \ p = 2, \end{cases}$$
$$\lambda_0 = \rho_0 = b^2$$

and thus Theorem 6 yields

$$\limsup_{t\to\infty} \frac{1}{t} \log |u(t; u_0)|^2 \leq \begin{cases} -(b^2 - 2a), & \text{if } \ p > 2, \\ -(b^2 - 2a + 2a_1), & \text{if } \ p = 2. \end{cases}$$

Consequently, we get pathwise exponential stability $\mathbb{P}$-a.s. if

$$b^2 > \begin{cases} 2a, & \text{if } \ p > 2, \\ 2a - 2a_1, & \text{if } \ p = 2. \end{cases}$$

In general, we have the following result:

Theorem 8 *Assume* (b.1)–(b.2), (43) *and that there exists a nonnegative continuous function $b = b(t)$ such that*

$$(B(t,x), x)^2 \geq b(t)|x|^4 \quad \forall x \in V, \tag{46}$$

with

$$\liminf_{t\to+\infty} \frac{1}{t} \int_0^t b(s)\, ds \geq b_0 \in \mathbb{R}_+. \tag{47}$$

Then, $\mathbb{P}$-a.s. it follows that

$$\limsup_{t\to+\infty} \frac{1}{t} \log |u(t; u_0)|^2 \leq \begin{cases} -(2b_0 - \lambda) & \text{if } \ p > 1, \\ -(2b_0 - \lambda + \alpha a_1) & \text{if } \ p = 2, \end{cases} \tag{48}$$

for any $u_0 \in L^2(\Omega, \mathcal{F}_0, P; H)$ such that $|u_0| \neq 0$, $\mathbb{P}$-a.s.

Remark 2 Observe that if $\lambda < 0$ then (42) is pathwise exponentially stable $\mathbb{P}$-a.s. for all $p > 1$ and all $b_0 \in \mathbb{R}_+$; while when $\lambda > 0$, (42) is stable if $2b_0 > \lambda$ (for $p \neq 2$) or $2b_0 > \lambda - \alpha a_1$ (for $p = 2$). Now, taking into account our previous theorems, we can summarize the analysis for the preceding example:

- *Case 1: The nonlinear problem, i.e., $p > 2$.* Observe that in this case, the problem is exponentially stable for all $b \in \mathbb{R}$ when $a \leq 0$. However, if $a > 0$ Theorem 7 gives stability provided $b^2 > 2a$. Note that we do not know what happens if $a > 0$ and $b^2 \leq 2a$.
- *Case 2: The linear problem, i.e., $p = 2$.* As in the preceding case, when $a \leq 0$ the system is $\mathbb{P}$-a.s. exponentially stable for all $b \in \mathbb{R}$. But if $a > 0$ we need to check (ii), which requires $b^2 < 2a_1 - 2a$, or it should hold $b^2 > 2a - 2a_1$. So, if $a \leq a_1$ exponential stability $\mathbb{P}$-a.s. follows for all $b \in \mathbb{R}$. But, when $a > a_1$, we only can ensure stability for $b^2 > 2a - 2a_1$ and we do not know what happens for $b^2 \leq 2a - 2a_{1.}$.

In conclusion, our results guarantee exponential stability for a wide range of values of a and b. Of course, this also means that, given the deterministic system $\mathrm{d}x(t) = A(t, x(t))\,\mathrm{d}t$, if a stochastic perturbation of the type $bx(t)\,\mathrm{d}W(t)$ appears, the perturbed system becomes exponentially stable when the parameter of the noise satisfies the conditions above. But when this does not happen, that is, when we do not know whether the system is stable or not, what could we say? Is it possible to add another stochastic term in order to stabilise the stochastic PDE? The answer is positive and some results on this direction can be found, for instance, in [16].

Remark 3 It is remarkable that Theorem 7 is a particular case of a more general result which ensures stabilization with general decay rate (super- or sub-exponential). Also, these results can be used to construct stabilisers of PDEs (see [9] for more details on these topics).

Remark 4 The technique used to prove our previous theorems can be adapted to study interesting examples arising in applications. Two important cases are the 2D Navier-Stokes equations (see [14]) and the 3D Lagrangian averaged Navier-Stokes α-model, also called Camassa-Holm equation (see [18, 17]).

3.2 A first stabilization result of a nonlinear PDEs by Stratonovich noise

As far as we know, there are no general results concerning the stabilization of nonlinear PDEs by Stratonovich noise. The problem seems to be very difficult and challenging. The unique work in this direction involves a canonical model whose dynamics is very well known in the deterministic case. This is the Chafee-Infante equation

$$\frac{\partial u}{\partial t} = \Delta u + \beta u - u^3, \quad \text{for} \quad x \in D \quad \text{and} \quad u|_{\partial D} = 0, \tag{49}$$

where D is a smooth bounded domain in $\mathbb{R}^m$.

In [8] it is proved that the *nonlinear* equation (49) can be stabilised by adding a collection of noisy terms similar to the linear case in Section 2:

$$\mathrm{d}u = [\Delta u + \beta u - u^3]\,\mathrm{d}t + \sum_{i=1}^{d} B_i u \circ \mathrm{d}W_i. \tag{50}$$

Essentially it is shown that solutions of (50) can be bounded using appropriate *positive* solutions of the linear equation

$$\mathrm{d}u = [\Delta u + \beta u]\,\mathrm{d}t + \sum_{i=1}^{d} B_i u \circ \mathrm{d}W_i. \tag{51}$$

Since (51) can be stabilised via a suitable choice of $\{B_i\}$, so can (50). The proof makes essential and continual use of the order-preserving properties of (50).

To set this problem in a suitable context, we choose $H = L^2(D)$; denote by $-\mathbb{A}$ the linear operator in H associated to the Laplacian. We then take $A = \mathbb{A} + \beta I$, which clearly satisfies the conditions of Theorem 4, and let N be the smallest integer such that $\sum_{j=1}^{N}(\beta - \lambda_j) < 0$. It follows that there exist linear operators $B_k : H \to H$ such that the zero solution of

$$\mathrm{d}u = [-\mathbb{A}u + \beta u]\,\mathrm{d}t + \sum_{k=1}^{d} B_k u \circ \mathrm{d}W_k(t) \tag{52}$$

is exponentially stable $\mathbb{P}$-a.s.

This fact can be used to deduce the stabilization of the nonlinear equation via the addition of the same noisy terms. The next result can be found in [8]:

Theorem 9 *There exist bounded linear operators $B_k : H \mapsto H$ and independent real Wiener processes W_k , $k = 1, \dots d$, such that the zero solution of*

$$\mathrm{d}u = (-\mathbb{A}u + \beta u - u^3)\,\mathrm{d}t + \sum_{j=1}^{d} B_k u \circ \mathrm{d}W_k(t) \tag{53}$$

is exponentially stable $\mathbb{P}$-a.s.

This simple, but illustrative, example may help to solve the stabilization problem for more general nonlinear PDEs appearing in applications. To the best of our knowledge, this is still an open problem.

4 "Super"-stabilization to a non-trivial stationary solution (random fixed point)

In the previous sections, we have exhibited a collection of results and techniques to stabilise the null solution of some classes of partial differential equations by using either Stratonovich or Itô noisy terms. However, there exists a very important effect of the noise not matter the interpretation we could give that can be regarded as a "super"-stabilization effect, since somehow the instabilities of the problem "dissapear" when the noise is added to the equation. It is worth mentioning that, in many situations, the noise may appear in an additive way, so that the Itô and Stratonovich interpretations coincide. In [8] it is shown that the addition of an additive noise that is rich enough may produce such a "super-stabilization" effect on the system. In this case, the dynamics of the model will be conduced to a stationary process (random fixed point). We first need some preliminaries.

4.1 Random dynamical systems and random attractors

In the interest of brevity we only state the definitions here: for more background on random dynamical systems, see Arnold [4].

Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and let $\{\vartheta_t : \Omega \to \Omega,\ t \in \mathbb{R}\}$ be a family of measure preserving transformations such that $(t, \omega) \mapsto \vartheta_t\omega$ is measurable, $\vartheta_0 = \mathrm{id}$, and $\vartheta_{t+s} = \vartheta_t\vartheta_s$ for all $s, t \in \mathbb{R}$. The flow ϑ_t together with the corresponding probability space $(\Omega, \mathcal{F}, \mathbb{P}, (\vartheta_t)_{t\in\mathbb{R}})$ is called a *(measurable) dynamical system*.

A continuous *random dynamical system* (RDS) on a Polish space (X, d) with Borel σ-algebra $\mathcal{B}$ over ϑ on $(\Omega, \mathcal{F}, \mathbb{P})$ is a measurable map

$$\begin{aligned}\varphi : \mathbb{R}^+ \times \Omega \times X &\to X\\ (t, \omega, x) &\mapsto \varphi(t, \omega)x\end{aligned}$$

such that $\mathbb{P}$-a.s.

i) $\varphi(0, \omega) = \mathrm{id}$ on X

ii) $\varphi(t + s, \omega) = \varphi(t, \vartheta_s\omega) \circ \varphi(s, \omega)$ for all $t, s \in \mathbb{R}^+$ (cocycle property)

iii) $\varphi(t, \omega) : X \to X$ is continuous.

A *random attractor* for an RDS φ is a random set $\omega \mapsto \mathcal{A}(\omega)$ such that

(i) $\mathcal{A}$ is a random compact set, that is, $\mathbb{P}$-a. s., $\mathcal{A}(\omega)$ is compact, and for all $x \in X$ the map $\omega \mapsto \mathrm{dist}\big(x, \mathcal{A}(\omega)\big)$ is measurable with respect to $\mathcal{F}$.

(ii) $\mathbb{P}$-a. s. $\varphi(t, \omega)\mathcal{A}(\omega) = \mathcal{A}(\vartheta_t\omega)$ for all $t \geq 0$, and

(iii) for every $D \subset H$ bounded, $\mathbb{P}$-a. s.,

$$\lim_{t\to\infty} \mathrm{dist}\big(\varphi(t, \vartheta_{-t}\omega)D, \mathcal{A}(\omega)\big) = 0.$$

To set our equation in the framework of random dynamical systems, we let $(\Omega, \mathcal{F}, \mathbb{P})$ denote the probability space generating the two-sided Wiener process $W(t)$, and define a shift ϑ_t on Ω by

$$W(t, \vartheta_s\omega) = W(t + s, \omega) - W(s, \omega),$$

the additional subtracted term ensuring that $W_\cdot(\vartheta_s\omega)$ is still a Brownian motion.

From Pardoux [36] we deduce that for any initial condition $u_0 \in L^2(D)$ and $T > 0$, there exists a unique strong solution

$$u(t; u_0) \in L^2(\Omega \times (0, T); H_0^1(D)) \cap L^4(\Omega \times (0, T) \times D) \cap L^2(\Omega; C(0, T; L^2(D)),$$

which generates a random dynamical system φ on the phase space $L^2(D)$ by setting $\varphi(t, \omega)u_0 = u(t; u_0)$.

In [12], it is proved that

$$\mathrm{d}u = [\Delta u + \beta u - u^3]\,\mathrm{d}t + \sigma u \circ \mathrm{d}W \tag{54}$$

has a random attractor. We then showed that the Hausdorff dimension of the attractor is bounded by d when

$$\beta < \frac{1}{d}\sum_{j=1}^{d} \lambda_j,$$

where λ_j are the eigenvalues of the Laplacian arranged in increasing order. Recently, Langa & Robinson proved in [31] the same upper bound for the upper box-counting (fractal) dimension of the attractor. Since $\lambda_n \sim n^{2/m}$ this implies that $d(\mathcal{A}(\omega)) \leq c\beta^{m/2}$.

In [13] we show that, provided $m \leq 5$, the unstable manifold near the origin has dimension at least d when $\beta > \lambda_d$. This leads to a lower bound on the dimension of the same order as the upper bound, and hence together show that the dimension of the random attractor is of the same order as its deterministic counterpart, namely

$$d(\mathcal{A}(\omega)) \sim \beta^{m/2}.$$

In this sense, the addition of a single multiplicative Stratonovich noise has no effect on the asymptotic complexity of the dynamics.

4.2 Collapse of the random attractor produced by additive noise

In this final section we show that the addition of a sufficiently rich additive white noise will reduce the random attractor of the equation to a single (random) point.

Such behaviour was originally demonstrated for the one-dimensional ordinary differential equation

$$\mathrm{d}x = [\alpha x - x^3]\,\mathrm{d}t + \epsilon\,\mathrm{d}W, \qquad \text{with} \qquad \alpha > 0$$

by Crauel & Flandoli [22], and has recently been shown by Robinson & Tearne [37] for a general gradient ODE of the form

$$\mathrm{d}x = -\nabla V(x) + \epsilon\,\mathrm{d}W$$

where $x \in \mathbb{R}^m$, W_t is an m-dimensional Brownian motion, and ϵ is sufficiently small (note that this is not in general an order-preserving system).

In [8] we prove a similar result for the equation

$$\mathrm{d}u = [\Delta u + \beta u - u^3]\,\mathrm{d}t + \sqrt{C}\,\mathrm{d}W, \qquad x \in D = [0, L], \tag{55}$$

where W, $t \in \mathbb{R}$, is a two-sided Q-cylindrical Wiener process on $H = L^2(D)$ (see Da Prato & Zabczyk [25] for the description and properties of the cylindrical Wiener process) and C is a bounded linear operator with bounded inverse on H. Here, we restrict ourselves to a one-dimensional domain.

The argument used in this case could be generalised to treat more abstract problems (cf. Chueshov & Scheutzow [20]) but the underlying idea is simple: Results of Arnold & Chueshov [5] on the structure of random attractors in order-preserving systems guarantee the existence of two random fixed points $\underline{a}$ and $\overline{a}$ that are contained in the attractor and are such that

$$\underline{a}(\omega) \leq u \leq \overline{a}(\omega), \quad \forall u \in \mathcal{A}(\omega).$$

Corresponding to these random fixed points there are invariant measures $\delta_{\underline{a}(\omega)}$ and $\delta_{\overline{a}(\omega)}$. Since the noise in (55) is sufficiently rich to guarantee that the equation has a *unique* invariant measure (e.g. Da Prato, Debussche, & Goldys [24]), it follows that the laws of $\underline{a}(\omega)$ and $\overline{a}(\omega)$ must coincide. It is only a small step from this, using the fact that $\underline{a}(\omega) \leq \overline{a}(\omega)$, to the deduction that $\underline{a}(\omega) = \overline{a}(\omega) = a(\omega)$, and hence that $\mathcal{A}(\omega) = \{a(\omega)\}$, i.e. the attractor is a single point.

Let us now recall the formal existence and uniqueness results for (55), and let us establish that the random attractor is a point.

We take $(\Omega, \mathcal{F}, \mathbb{P})$ to be the probability space that generates the Q-cylindrical Wiener process $W(t)$, and define a shift ϑ_t on Ω by $W(t, \vartheta_s\omega) = W(t+s, \omega) - W(s, \omega)$.

Under these assumptions (see Da Prato & Zabczyk [25]), for each $u_0 \in L^2(D)$ and $T > 0$ there exists a unique solution $u(t; u_0)$ for (55), with

$$u(t; u_0) \in L^2\big(\Omega \times (0,T); H_0^1(D)\big) \cap L^4\big(\Omega \times (0,T) \times D\big) \cap L^2\big(\Omega; C(0,T; L^2(D))\big).$$

It follows that the solutions of (55) generate a random dynamical system on $L^2(D)$ if we define

$$\varphi(t, \omega)u_0 = u(t; \omega, u_0),$$

where $u(t; \omega, u_0)$ is the solution of (55) with noise ω and initial condition $u(0) = u_0$.

Then, we have the following result (see [8] for the proof):

Theorem 10 *The random attractor for* (55) *consists of a single point, i.e. there exists a random variable* $a : \Omega \mapsto H$ *with*

$$\varphi(t, \omega)a(\omega) = a(\vartheta_t\omega) \qquad \textit{for every } t \geq 0, \qquad \mathbb{P} - a.s.$$

such that $\mathcal{A}(\omega) = \{a(\omega)\}$.

The above result would extend to m-dimensional domains if the existence of a unique invariant measure could be guaranteed in this case (cf. Hairer [26]).

5 Final remarks and some open problems

Needless to say that what we have included in the previous section does not cover all the aspects involving the asymptotic behaviour of stochastic partial differential equations. We have only emphasised some facts related to the stabilising effect that can be produced by the appearance of noise in deterministic PDEs. Therefore, many other topics could have been considered. For instance, we have only mentioned the effects produced by the noise in the structure of the global attractor in a very particular example, so it is very interesting to analyse these potential effects that different classes of noise can produce on the global attractors for other interesting models from applications.

Going deeper in this direction, there are many situations in which the uniqueness of solution is not known or cannot be guaranteed, or even that the evolution of the system can be better modelled by a differential inclusion. To our knowledge, there are no results on the stabilization effect of the noise in multivalued dynamical systems.

Sometimes, the consideration of delay terms in the equations of some models is fully justified. The problem of stabilization of delay (ordinary or partial) differential equations is therefore an important task. In the finite dimensional context, there are only a few results on the stabilization by the Itô noise when the delay is small enough (see Appleby and Mao [1]), but nothing is known for systems with arbitrary delay (finite or infinite) neither by using Itô nor Stratonovich noise.

Another problem that may be even closer to reality is related to the effect produced by the noise when it acts only on (part of) the boundary of the domain and not in the forcing term in the equation. For instance, if we are considering an oceanic model, the stochastic disturbances may appear on the ocean surface and not in the equations driving the system. Some preliminary results are to appear in the future (see [7]).

Of course, we could list some more interesting and challenging situations but we content ourselves with the previous ones since this area is still in its infancy, and too much work is to be done in the future. We hope we can contribute to solve some of these problems.

Acknowledgements. I would like to thank all those colleagues who have collaborated with me and who have taught me so much in this field. Especially, my sincere gratitude and thanks go to José Real, José A. Langa, María J. Garrido-Atienza, James Robinson, Kai Liu, Xuerong Mao, Takeshi Taniguchi, Peter Kloeden, Björn Schmalfuss and Hans Crauel.

This work has been partly supported by D.G.E.S. (Ministerio de Ciencia y Tecnología, Spain) under the grant BFM2002-03068.

References

[1] J. Appleby and X. Mao, Stochastic stabilization of functional differential equations, *Systems and Control Letters* 54(11) (2005), 1069–1081.

[2] L. Arnold, *Stochastic Differential Equations: Theory and Applications*, Wiley & Sons, New York, (1974).

[3] L. Arnold, *Random Dynamical Systems*. Springer, New York, 1998.

[4] L. Arnold, Stabilization by noise revisited, *Z. Angew. Math. Mech.* 70(1990), 235-246.

[5] L. Arnold and I. Chueshov, Order-preserving random dynamical systems: Equilibria, attractors, applications, *Dyn. Stab. Sys.* 13 (1998), 265–280.

[6] L. Arnold, H. Crauel and V. Wihstutz, Stabilization of linear systems by noise, *SIAM J. Control Optim.* 21(1983), 451-461.

[7] T. Caraballo, D. Blömker, and J. Duan, Stabilization produced by noise in the boundary, in preparation.

[8] T. Caraballo, H. Crauel, J.A. Langa and J.C. Robinson, The effect of noise on the Chafee-Infante equation: a nonlinear case study, *Proc. Amer. Math. Soc.*, in press.

[9] T. Caraballo, M.J. Garrido-Atienza and J. Real, Stochastic stabilization of differential systems with general decay rate, *Systems & Control Letters* 48(5) (2003), 397-406.

[10] T. Caraballo, P.E. Kloeden and B. Schmalfuss, Exponentially stable starionary solutions for stochastic evolution equations and their perturbations, *Appl. Math. Optim.* 20(2004), 183–207

[11] T. Caraballo and J.A. Langa, Comparison of the long-time behavior of linear Ito and Stratonovich partial differential equations, *Stoch. Anal. Appl.* 19(2) (2001), 183-195.

[12] T. Caraballo, J.A. Langa and J.C. Robinson, Stability and random attractors for a reaction-diffusion equation with multiplicative noise, *Discrete Cont. Dyn. Sys.* 6 (2000), 875–892.

[13] T. Caraballo, J.A. Langa and J.C. Robinson, A stochastic pitchfork bifurcation in a reaction-diffusion equation, *R. Soc. Lond. Proc. Ser. A* 457 (2001), 2041–2061.

[14] T. Caraballo, J.A. Langa and T. Taniguchi, The exponential behaviour and stabilizability of stochastic 2D-Navier-Stokes equations, *J. Diff. Eqns.* 179(2002), 714-737.

[15] T. Caraballo and K. Liu, On exponential stability criteria of stochastic partial differential equations, *Stoch. Proc. & Appl.* 83 (1999), 289-301.

[16] T. Caraballo, K. Liu and X.R. Mao, On stabilization of partial differential equations by noise, *Nagoya Math. J.* 161(2) (2001), 155-170.

[17] T. Caraballo, A.M. Márquez-Durán and J. Real, On the asymptotic behaviour of a stochastic 3D-Lans-alpha model, *Appl. Math. Optim.* 53(2006), 141-161.

[18] T. Caraballo, J. Real and T. Taniguchi, On the existence and uniqueness of solutions to stochastic 3-dimensional Lagrangian averaged Navier-Stokes equations, *R. Soc. Lond. Proc. Ser. A* 462(2006), 459-479.

[19] T. Caraballo and J.C. Robinson, Stabilization of linear PDEs by Stratonovich noise, *Systems & Control Letters* 53(2004), 41-50.

[20] I. D. Chueshov and M. Scheutzow, On the structure of attractors and invariant measures for a class of monotone random systems, *Dyn. Sys.* 19 (2004), 127–144.

[21] H. Crauel, White noise eliminates instability, *Archiv der Mathematik* 75 (2000), 472–480.

[22] H. Crauel and F. Flandoli, Additive noise destroys a pitchfork bifurcation, *J. Dyn. Diff. Eqn.* 10 (1998), 259–274.

[23] H. Crauel and F. Flandoli, Hausdorff dimension of invariant sets for random dynamical systems, *J. Dyn. Diff. Eqn.* **10** (1998), 449–474.

[24] G. Da Prato, A. Debussche, and B. Goldys, Some properties of invariant measures of non symmetric dissipative stochastic systems, *Prob. Theor. Relat. Fields* **123** (2002), 355-380.

[25] G. Da Prato and J. Zabczyk, *Stochastic Equations in Infinite Dimensions*, Cambridge University Press, (1992).

[26] M. Hairer, Exponential mixing properties of stochastic PDEs through asymptotic coupling, *Prob. Theory Relat. Fields* 124 (2002), 345–380.

[27] R. has'minskii, Stochastic Stability of Differential Equations, Sijthoff and Noordhoff, Netherlands, (1980).

[28] H. Kunita, *Stochastic Partial Differential Equations connected with Non-Linear Filtering*, in Lecture Notes in Mathematics 972, pp. 100-169, (1981)

[29] A.A. Kwiecinska, Stabilization of partial differential equations by noise, *Stoch. Proc.& Appl.* 79 (1999), no. 2, 179–184

[30] A.A. Kwiecinska, Stabilization of evolution equations by noise, *Proc. Amer. Math. Soc.* 130(2002), No. 10, 3067-3074.

[31] J.A. Langa and J.C. Robinson, Upper box-counting dimension of a random invariant set, *J. Math. Pures App.*, to appear.

[32] G. Leha, B. Maslowski and G. Ritter, Stability of solutions to semilinear stochastic evolution equations, *Stoch. Anal. Appl.* 17(1999), No. 6, 1009-1051.

[33] J.L. LIONS, *Quelque méthodes de résolution des problèmes aux limites non linéaires*, Dunod, Gauthier- Villars, Paris, 1969.

[34] X.R. MAO, Stochastic stabilization and destabilization, *Systems & Control Letters* 23 (1994), 279-290.

[35] A. PAZY, *Semigroups of Linear Operators and Applications to Partial Differential Equations*, Springer-Verlag, New York Inc., 1983.

[36] E. PARDOUX, Équations aux Dérivées Partielles Stochastiques non Linéaires Monotones, Thesis Univ. Paris XI, 1975.

[37] J.C. ROBINSON AND O.M. TEARNE, Collapse of attractors of gradient ODEs under small random perturbations, in preparation.

[38] M. SCHEUTZOW, Stabilization and destabilization by noise in the plane, *Stoch. Anal. Appl.* 11(1) (1993), 97-113.

[39] H.J. SUSSMANN, On the gap between deterministic and stochastic ordinary differential equations, *The Annals of Probability* 6(1978) , No. 1, 19-41.

[40] E. WONG & M. ZAKAI, On the relationship between ordinary and stochastic differential equations and applications to stochastic problems in control theory, Proc. Third IFAC Congress, paper 3B, 1966.

Bol. Soc. Esp. Mat. Apl. n°37(2006), 71–92

Reservoir Simulation

Zhangxin (John) Chen

Center for Scientific Computation.
Southern Methodist University
Box 750156, Dallas, TX 75275-0156, USA.

1. INTRODUCTION

Mathematical models of petroleum reservoirs have been utilized since the late 1800s. A mathematical model consists of a set of differential equations that describe the flow of fluids in petroleum reservoirs, together with an appropriate set of boundary and/or initial conditions. The reliability of predictions from a petroleum model depends on how well the model describes the field situation. To develop a model, in general, simplifying assumptions need to be made because the field situation is too complicated to be described exactly. The assumptions needed to solve a model analytically are very restrictive; many analytical solutions require that the reservoir be homogeneous and isotropic, for example. It is usually necessary to solve a mathematical model approximately using numerical methods. Since the 1950s, when high-speed digital computers became widely available, numerical models have been used to predict, understand, and optimize complex physical fluid flow processes in petroleum reservoirs. Moreover, the emergence of complex nhanced recovery techniques in the field of oil production has emphasized the need for sophisticated mathematical and numerical tools, capable of modeling intricate physical phenomena and sharply changing fluid interfaces, for example. The objectives of this three-hour course are to provide researchers in the area of porous media flow, especially in petroleum reservoirs, with a review of single phase, two phase, black oil, and compositional flows and with the current, state-of-the-art numerical methods used in solution of these flows. For more information, the reader should refer to Chen *et al.* (2006).

A petroleum reservoir is a porous medium which contains hydrocarbons. The primary goal of reservoir simulation is to predict future performance of a reservoir and find ways and means of optimizing the recovery of some of the hydrocarbons.

There are two important characteristics of a petroleum reservoir, the nature of the rock and that of the fluids filling it. A reservoir is usually heterogeneous;

its properties heavily depend on the space location. The so-called fractured reservoir is heterogeneous, for example. It consists of a set of blocks of porous media (the matrix) and a net of fractures. The rock properties in such a reservoir dramatically change; its permeability may vary from one in the matrix to thousands in the fractures, for example. While the governing equations for the fractured reservoir are similar to those for an ordinary reservoir, they possess additional difficulties to solve. The mathematical models presented take into account the heterogeneity of a porous medium, but the numerical methods are primarily presented for an ordinary one.

The nature of the fluids filling a reservoir strongly depend on the stage of oil recovery. In the very early stage, the reservoir usually contains a single fluid like gas or oil. Often the pressure at this stage is so high that the gas or oil is produced by simple natural decompression without pumping effort at wells. This stage is referred to as *primary recovery*, and it ends when the pressure equilibrium between the oil field and the atmosphere occurs. Often primary recovery leaves $70 - 85\%$ of hydrocarbons in the reservoir.

To recover part of the remaining oil, a fluid (usually water) is injected into some wells (*injection wells*) while oil is produced through other wells (*production wells*). This process serves maintaining the high reservoir pressure and flow rates. It also displaces some of the oil and pushes it toward the production wells. This stage of oil recovery is called *secondary recovery*.

In the secondary recovery, if the pressure is above the bubble pressure of the oil phase, the flow is the two-phase immiscible flow, one phase being water and the other being oil, without mass transfer between the phases. If the pressure drops below the bubble pressure, then the oil (more precisely, the hydrocarbon phase) is split into a liquid phase and a gaseous phase at thermodynamical equilibrium. In this case, the flow is of the so-called *black-oil* type; the water phase does not exchange mass with the other phases, and the liquid and gaseous phases exchange mass between them.

Water flooding is still not very effective and 50% or more of hydrocarbons often remain in the reservoir. Due to strong surface tension, a large amount of oil is trapped in small pores and cannot be washed out with this technique. Also, when the oil is heavy and viscous, the water is extremely mobile. If the flow rate is sufficiently high, instead of producing oil, the production wells primarily produce water.

To recover more of the hydrocarbons, several enhanced recovery techniques have been developed. These techniques involve complex chemical and thermal effects and are termed *tertiary recovery* or *enhanced recovery*. There are many different versions of enhanced recovery techniques, but one of the main objectives of these techniques is to achieve miscibility and thus eliminate the residual oil saturation. The miscibility is achieved by increasing temperature (in-situ combustion) or by injecting other chemical species like CO_2. One typical flow in enhanced recovery is the *compositional flow*, where only the number of chemical species is a-priori given, and the number of phases and the composition of each phase in terms of the given species depend on the thermodynamical conditions and the overall concentration of each species. All these types of

flows in petroleum reservoirs will be considered.

In general, the equations governing a mathematical model of a reservoir cannot be solved by analytical methods. Instead, a numerical model can be produced in a form that is amenable to solution by digital computers. Since the 1950s, when high-speed digital computers became widely available, numerical models have been used to predict, understand, and optimize complex physical fluid flow processes in petroleum reservoirs. Moreover, the emergence of complex enhanced recovery techniques in the field of oil production has emphasized the need for sophisticated mathematical and numerical tools, capable of modeling intricate physical phenomena and sharply changing fluid interfaces, for example.

The standard finite element method and two nonstandard ones, the mixed finite element method and the characteristics-based finite element method are covered. The reason for the development of the mixed method is that in many applications a vector variable (e.g., the velocity field in petroleum simulation) is the primary variable in which we are interested, and then the mixed formulation is utilized to approximate both this variable and a scalar variable (e.g., the pressure) simultaneously and give a high order approximation of both variables. The characteristics-based finite element method is suitable for advection-dominated problems. This method takes reasonably large time steps, captures sharp solution fronts, and conserves mass.

Since the fluid flow models in porous media involve large, coupled systems of nonlinear, time-dependent partial differential equations, an important problem in the numerical simulation is to develop stable, efficient, robust, accurate, and self-adaptive time stepping techniques. Explicit methods like forward Euler methods require that a Courant-Friedrichs-Lewy (CFL) time step constraint be satisfied, while implicit methods such as backward Euler and Crank-Nicolson methods are reasonably stable. On the other hand, the explicit methods are computationally efficient and the implicit methods require the solution of large systems of nonlinear equations at each time step. Explicit methods, together with linearization by some Newton-like scheme, have been frequently used in reservoir simulation. Due to the CFL condition, enormous computations are needed to simulate a long time period (several years) problem in a field-scale model, and thus fully explicit methods cannot be efficiently exploited for problems with strong nonlinearities.

A variation to achieve better stability without suffering too much in computation is the IMPES (implicitly in pressure and explicitly in saturation) technique. This technique works well for problems of intermediate difficulty and is still widely used in the oil industry. However, it is not efficient for problems of difficult nonlinearities.

The other basic method for solving multiphase equations is the simultaneous solution (SS) method, which solves all of the coupled nonlinear equations simultaneously and implicitly. This technique is stable and can take very large time steps while stability is maintained. However, for complex problems where many equations are coupled, the size of matrices to be solved is too large, even for a super computer. Thus its robustness suffers.

A variety of sequential methods for solving the equations in an implicit

fashion without a full coupling have been developed. They are less stable but more computational efficient than the SS method, and more stable but less efficient than the IMPES method. The majority of the research in the solution methods has concentrated on the stability of time-stepping methods, and the efficient linearization and iterative solution of the resulting equations. The accuracy of these methods also needs to be addressed. All these solution techniques will be addressed.

2. FLOW AND TRANSPORT EQUATIONS IN POROUS MEDIA

2.1. Single Phase Flow

In this section we consider the transport of a Newtonian fluid that occupies the entire void space in a porous medium under the isothermal condition.

2.1.1. Single phase flow in a rigid medium

The governing equations for the single phase flow of a fluid (a single component or a homogeneous mixture) in a rigid porous medium are given by the conservation of mass, Darcy's law, and an equation of state. We make the assumptions that the mass fluxes due to dispersion and diffusion are so small (relative to the advective mass flux) that they are negligible and the fluid-solid interface is a material surface with respect to the fluid mass so no mass of this fluid can cross it.

Denote by ϕ the porosity of the porous medium (the fraction of a representative elementary volume available for the fluid), by ρ the density of the fluid per unit volume, by $\mathbf{u}$ the superficial Darcy velocity, and by q the external sources and sinks. The mass conservation is expressed as follows:

$$\frac{\partial(\phi\rho)}{\partial t} = -\nabla \cdot (\rho \mathbf{u}) + q, \tag{2.1}$$

where $\nabla\cdot$ is the *divergence* operator:

$$\nabla \cdot \mathbf{u} = \frac{\partial u_1}{\partial x_1} + \frac{\partial u_2}{\partial x_2} + \frac{\partial u_3}{\partial x_3},$$

with $\mathbf{u} = (u_1, u_2, u_3)$ (in three dimensions).

In addition to (2.1), we state the momentum conservation in the form of Darcy's law (Darcy, 1856). This law indicates a linear relationship between the flow rate and the pressure head gradient:

$$\mathbf{u} = -\frac{1}{\mu}\mathbf{k}\left(\nabla p - \rho g \nabla Z\right), \tag{2.2}$$

where p is the fluid pressure, $\mathbf{k}$ is the absolute permeability tensor of the porous medium, μ is the fluid viscosity, g is the magnitude of the gravitational acceleration, Z is the depth, and ∇ is the *gradient* operator:

$$\nabla p = \left(\frac{\partial p}{\partial x_1}, \frac{\partial p}{\partial x_2}, \frac{\partial p}{\partial x_3}\right).$$

An equation of state is expressed in terms of the fluid compressibility c_f:

$$c_f = -\frac{1}{V}\frac{\partial V}{\partial p}\Big|_T = \frac{1}{\rho}\frac{\partial \rho}{\partial p}\Big|_T, \tag{2.3}$$

at a fixed temperature T. Combine equations (2.1)–(2.3) to have a closed system for the main unknown p or ρ. Simplified expressions such as a linear relationship between p and ρ for a slightly compressible fluid can be used for (2.3) (Aziz-Settari, 1979).

In general, there is not a distributed mass source or sink in single phase flow in a three dimensional medium. However, as an approximation, we may consider the case where sources and sinks of a fluid are located at isolated points $\mathbf{x}^{(i)}$. Then these point sources and sinks can be surrounded by small spheres which are excluded from the medium. The surface of these spheres is treated as part of the boundary of the medium, and the mass flow rate per unit volume of each source or sink specifies the total flux through its surface.

Another approach to handle point sources and sinks is to insert them in the mass conservation equation as in (2.1). That is, for point sinks, we define q in (2.1) by

$$q = -\sum_i \rho q^{(i)} \delta(\mathbf{x} - \mathbf{x}^{(i)}), \tag{2.4}$$

where $q^{(i)}$ indicates the volume of the fluid produced per unit time at $\mathbf{x}^{(i)}$ and δ is the Dirac delta function. For point sources, q is given by

$$q = \sum_i \rho^{(i)} q^{(i)} \delta(\mathbf{x} - \mathbf{x}^{(i)}), \tag{2.5}$$

where $q^{(i)}$ and $\rho^{(i)}$ denote the volume of the fluid injected per unit time and its density (which is known) at $\mathbf{x}^{(i)}$, respectively.

2.1.2. Single phase flow in a deformable medium

We consider a deformable porous medium whose solid skeleton has compressibility and shearing rigidity. The medium is assumed to be composed of a linear elastic material and its deformation to be small.

Let $\mathbf{w_s}$ and $\mathbf{w}$ be the displacements of the solid and fluid, respectively. For a deformable medium, the Darcy law in (2.2) is generalized as follows (Biot, 1955; Chen *et al.*, 2004):

$$\dot{\mathbf{w}} - \dot{\mathbf{w}}_\mathbf{s} = -\frac{1}{\mu}\mathbf{k}\left(\nabla p - \rho g \nabla Z\right), \tag{2.6}$$

where $\dot{\mathbf{w}} = \partial \mathbf{w}/\partial t$. Note that $\mathbf{u} = \dot{\mathbf{w}}$, so (2.6) just introduces a new dependent variable $\mathbf{w_s}$. Additional equations are needed to have a closed system.

Let $\mathbf{I}$ be the identity matrix. The total stress tensor of the bulk material is denoted by

$$\boldsymbol{\sigma} + \sigma \mathbf{I} \equiv \begin{pmatrix} \sigma_{11} + \sigma & \sigma_{12} & \sigma_{13} \\ \sigma_{21} & \sigma_{22} + \sigma & \sigma_{23} \\ \sigma_{31} & \sigma_{32} & \sigma_{33} + \sigma \end{pmatrix},$$

with the symmetry property $\sigma_{ij} = \sigma_{ji}$. To see the meaning of this tensor, let us consider a cube of the bulk material with unit size. Then σ represents the total normal tension force applied to the fluid part of the faces of the cube, while the remaining components σ_{ij} are the forces applied to the portion of the cube faces occupied by the solid. This stress tensor satisfies the equilibrium relation

$$\nabla \cdot \left(\boldsymbol{\sigma} + \sigma \mathbf{I}\right) + \rho_t g \nabla Z = 0, \tag{2.7}$$

where $\rho_t = \phi\rho + (1-\phi)\rho_s$ is the mass density of the bulk material and ρ_s is the solid density. To relate $\boldsymbol{\sigma}$ to $\mathbf{w_s}$, we need a constitutive relationship between the stress and strain tensors.

Denote the strain tensors of the solid and fluid by $\boldsymbol{\epsilon}_s$ and $\boldsymbol{\epsilon}$, respectively. They are defined by

$$\epsilon_{s,ij} = \frac{1}{2}\left(\frac{\partial w_{s,i}}{\partial x_j} + \frac{\partial w_{s,j}}{\partial x_i}\right), \quad \epsilon_{ij} = \frac{1}{2}\left(\frac{\partial w_i}{\partial x_j} + \frac{\partial w_j}{\partial x_i}\right), \quad i,\ j = 1,2,3.$$

Also, define $\epsilon = \epsilon_{11} + \epsilon_{22} + \epsilon_{33}$. The stress-strain relationship is given by

$$\begin{pmatrix} \sigma_{11} \\ \sigma_{22} \\ \sigma_{33} \\ \sigma_{23} \\ \sigma_{31} \\ \sigma_{12} \\ \sigma \end{pmatrix} = \begin{pmatrix} c_{11} & c_{12} & c_{13} & c_{14} & c_{15} & c_{16} & c_{17} \\ & c_{22} & c_{23} & c_{24} & c_{25} & c_{26} & c_{27} \\ & & c_{33} & c_{34} & c_{35} & c_{36} & c_{37} \\ & & & c_{44} & c_{45} & c_{46} & c_{47} \\ & & & & c_{55} & c_{56} & c_{57} \\ & & & & & c_{66} & c_{67} \\ & & & & & & c_{77} \end{pmatrix} \begin{pmatrix} \epsilon_{s,11} \\ \epsilon_{s,22} \\ \epsilon_{s,33} \\ \epsilon_{s,23} \\ \epsilon_{s,31} \\ \epsilon_{s,12} \\ \epsilon \end{pmatrix},$$

where $c_{ij} = c_{ji}$ (i.e., the coefficient matrix $\mathbf{c}$ is symmetric). Now substitute this relationship into (2.7) to have three equations for the three unknows $w_{s,1}$, $w_{s,2}$, and $w_{s,3}$.

To see an example of the stress-strain relationship, we consider the case where the solid matrix is isotropic. In this case, with $\epsilon_s = \epsilon_{s,11} + \epsilon_{s,22} + \epsilon_{s,33}$, this relationship is given by

$$\begin{aligned} \sigma_{ii} &= 2G\left(\epsilon_{s,ii} + \frac{\nu\epsilon_s}{1-2\nu}\right) - Hp, \quad && i = 1,2,3, \\ \sigma_{ij} &= 2G\epsilon_{s,ij}, && i,j = 1,2,3,\ i \neq j, \end{aligned}$$

where G and ν are Young's modulus and Poisson's ratio for the solid skeleton, and H is a physical constant whose value must be determined by experiments or by numerical methods (Biot, 1955, Chen *et al.*, 2004).

2.1.3. Single phase flow in a fractured medium

A fractured porous medium is a medium which is intersected by a network of interconnected fractures, or solution channels. Such a medium could be modeled by allowing the porosity and permeability to vary rapidly and discontinuously

over the whole domain. Both of these quantities are much larger in the fractures than in the porous rock. However, the data requirement and computational cost for simulating such a single porosity model would be to great to approximate the flow in the entire medium. Instead, it is more convenient to regard the fluid in the void space as made up of two parts: one in the fractures and the other in the porous blocks and to treat each part as a continuum that occupies the entire domain. These two overlapping continua are allowed to interact with each other. Such a model is often termed a dual porosity model.

Since fluid flows more rapidly in the fractures than in the porous blocks, we assume that it does not flow directly from one block to another block. Rather, it first flows into the fractures, and then it flows into another block or remain in the fractures. Also, the equations that describe the flow in the fracture continuum contains a source term that represents the flow of fluid from the porous blocks to the fractures; this term is assumed to be distributed over the entire medium. Finally, we assume that the external sources and sinks interact only with the fracture system. This is reasonable since flow is much faster in this system than in the blocks. Based on these assumptions, flow through each block in a fractured (rigid) porous medium is given by

$$\frac{\partial(\phi\rho)}{\partial t} = -\nabla\cdot(\rho\mathbf{u}). \tag{2.8}$$

The flow in the fractures is described by

$$\frac{\partial(\phi_{fr}\rho_{fr})}{\partial t} = -\nabla\cdot(\rho_{fr}\mathbf{u}_{fr}) + q_{fr} + q_{ext}, \tag{2.9}$$

where the subscript fr represents the fracture quantities, q_{fr} denotes the flow from the porous blocks to the fractures, and q_{ext} indicates the external sources and sinks. The velocities $\mathbf{u}$ and $\mathbf{u}_{fr}$ are determined by Darcy's law as in (2.2).

The source term q_{fr} is defined as follows (Pirson, 1953; Barenblatt *et al.*, 1960; Douglas-Arbogast, 1990). The total mass of fluid leaving the ith block Ω_i per unit time is

$$\int_{\partial\Omega_i} \rho\mathbf{u}\cdot\boldsymbol{\nu}\, da(\mathbf{x}),$$

where $\boldsymbol{\nu}$ is the outward unit normal to the surface $\partial\Omega_i$ of Ω_i and the dot product $\mathbf{u}\cdot\boldsymbol{\nu}$ is defined by

$$\mathbf{u}\cdot\boldsymbol{\nu} = u_1\nu_1 + u_2\nu_2 + u_3\nu_3.$$

The divergence theorem and (2.8) imply that

$$\int_{\partial\Omega_i} \rho\mathbf{u}\cdot\boldsymbol{\nu}\, da(\mathbf{x}) = \int_{\Omega_i} \nabla\cdot(\rho\mathbf{u})\, d\mathbf{x} = -\int_{\Omega_i} \frac{\partial(\phi\rho)}{\partial t}\, d\mathbf{x}.$$

Now we define q_{fr} by

$$q_{fr} = -\sum_i \chi_i(\mathbf{x})\frac{1}{|\Omega_i|}\int_{\Omega_i} \frac{\partial(\phi\rho)}{\partial t}\, d\mathbf{x}, \tag{2.10}$$

where $|\Omega_i|$ denotes the volume of Ω_i and $\chi_i(\mathbf{x})$ is its characteristic function, i.e.,

$$\chi_i(\mathbf{x}) = \begin{cases} 1 & \text{if } \mathbf{x} \in \Omega_i, \\ 0 & \text{otherwise.} \end{cases}$$

2.1.4. Non-Darcy's law

Strictly speaking, the Darcy law holds only for a Newtonian fluid over a certain range of flow rates. As the flow rate increases, a deviation from this law has been noticed (Dupuit, 1863; Forchheimer, 1901). It has been experimentally and mathematically observed that this deviation is due to inertial, turbulence, and other high velocity effects (Fancher-Lewis, 1933; Hubbert, 1956; Mei-Auriault, 1991; Chen *et al.*, 2001). Hubbert (1956) observed a deviation from the usual Darcy law at a Reynolds' number of flow of about one (based on the grain diameter of a unconsolidated medium), whereas turbulence was not noticed until the Reynolds' number approached 600 (Aziz-Settari, 1979).

The correction to Darcy's law for high flow rates can be described by a quadratic term (Forchheimer, 1901; Katz *et al.*, 1959; Ward, 1964; Chen *et al.*, 2001)

$$\big(\mu\mathbf{I} + \beta\rho|\mathbf{u}|\mathbf{k}\big)\mathbf{u} = -\mathbf{k}\left(\nabla p - \rho g \nabla Z\right),$$

where β indicates the inertial or turbulence factor and

$$|\mathbf{u}| = \sqrt{u_1^2 + u_2^2 + u_3^2}.$$

This equation is generally called the Forchheimer law and incorporates laminar, inertial, and turbulence effects. It has been the subject of many experimental and theoretical investigations. These investigations have centered on the issue of providing a physical or theoretical basis for the derivation of Forchheimer's law. Many approaches have been developed and analyzed for this purpose such as empiricism fortified with dimensional analysis (Ward, 1964), experimental study (MacDonald *et al.*, 1979), averaging methods (Chen *et al.*, 2001), and variational principles (Knupp-Lage, 1995).

2.2. Two Phase Immiscible Flow

In reservoir simulation, we are often interested in the simultaneous flow of two or more fluid phases within a porous medium. We now develop basic equations for multi-phase flow in a porous medium. In this section we consider two phase flow where the fluids are immiscible and there is no mass transfer between the phases. One phase (e.g., water) wets the porous medium more than the other (e.g., oil) and is called the wetting phase and indicated by a subscript w. The other phase is termed the nonwetting phase and indicated by o.

2.2.1. Basic equations

Since we assume that there is no mass transfer between phases in the immiscible flow, mass is conserved within each phase, as in (2.1):

$$\frac{\partial(\phi\rho_\alpha S_\alpha)}{\partial t} = -\nabla\cdot(\rho_\alpha \mathbf{u}_\alpha) + q_\alpha, \qquad \alpha = w,\, o, \tag{2.11}$$

where each phase has its own density ρ_α, saturation S_α (the fraction of the void volume of the medium filled by phase α), Darcy velocity $\mathbf{u}_\alpha$, and mass flow rate q_α. The Darcy law for single phase flow can be directly extended to multi-phase flow:

$$\mathbf{u}_\alpha = -\frac{1}{\mu_\alpha}\mathbf{k}_\alpha\left(\nabla p_\alpha - \rho_\alpha g \nabla Z\right), \qquad \alpha = w,\, o, \tag{2.12}$$

where $\mathbf{k}_\alpha$, p_α, and μ_α are the effective permeability, pressure, and viscosity for phase α. Since the simultaneous flow of the two fluids causes each to interfere with the other, the effective permeabilities are not greater than the absolute permeability of the porous medium $\mathbf{k}$. The relative permeabilities $k_{r\alpha}$ are widely used in reservoir simulation:

$$\mathbf{k}_\alpha = k_{r\alpha}\mathbf{k}, \qquad \alpha = w,\, o.$$

A couple of new equations peculiar to multi-phase flow are needed. The fact that the two fluids jointly fill the voids means the relation

$$S_w + S_o = 1. \tag{2.13}$$

Also, due to the curvature and surface tension of the interface between the two phases, the pressure in the wetting fluid is smaller than that in the nonwetting fluid. The pressure difference is given by the capillary pressure p_c:

$$p_c = p_o - p_w. \tag{2.14}$$

Typical functions of p_c and $k_{r\alpha}$ wil be described in the next chapter. When q_w and q_o represent a finite number of point sinks or sources, they can be defined as in (2.4) or (2.5). Also, the densities ρ_w and ρ are functions of their respective pressures. Thus, after substituting (2.12) into (2.11), we have a complete set of equations in (2.11), (2.13), and (2.14) for the four main unknowns p_α and S_α, $\alpha = w,\, o$. Other mathematical formulations will be discussed in the next subsection. The development of single phase flow in deformable and fractured porous medium is applicable to two phase flow. We do not pursue this similar development.

2.2.2. Alternative differential equations

In this subsection we derive several alternative formulations of the differential equations in (2.11)–(2.14).

2.2.2.1. Formulation in phase pressures

Assume that the capillary pressure p_c has a unique inverse function:

$$S_w = p_c^{-1}(p_o - p_w).$$

Then it follows from (2.11)–(2.14) that

$$\begin{aligned}
\nabla\cdot\left(\frac{\rho_w}{\mu_w}\mathbf{k}_w\left(\nabla p_w - \rho_w g\nabla Z\right)\right) &= \frac{\partial(\phi\rho_w p_c^{-1})}{\partial t} - q_w,\\
\nabla\cdot\left(\frac{\rho_o}{\mu_o}\mathbf{k}_o\left(\nabla p_o - \rho_o g\nabla Z\right)\right) &= \frac{\partial\big(\phi\rho_o(1-p_c^{-1})\big)}{\partial t} - q_o.
\end{aligned} \tag{2.15}$$

System (2.15) for p_w and p_o was used in the so-called *simultaneous solution* (SS) method in petroleum reservoirs (Douglas *et al.*, 1959). The equations in this system are strongly nonlinear and coupled.

2.2.2.2. Formulation in a phase pressure and saturation

Using (2.12)–(2.14), (2.11) can be rewritten as follows:

$$\begin{aligned} &\nabla\cdot\left(\frac{\rho_w}{\mu_w}\mathbf{k}_w\left(\nabla p_o-\frac{dp_c}{dS_w}\nabla S_w-\rho_w g\nabla Z\right)\right)=\frac{\partial(\phi\rho_w S_w)}{\partial t}-q_w,\\ &\nabla\cdot\left(\frac{\rho_o}{\mu_o}\mathbf{k}_o\left(\nabla p_o-\rho_o g\nabla Z\right)\right)=\frac{\partial\big(\phi\rho_o(1-S_w)\big)}{\partial t}-q_o. \end{aligned} \tag{2.16}$$

This system can be solved for p_o and S_w. Again, the two equations in (2.16) are strongly coupled. For their finite difference or finite element solution, they can be decoupled under appropriate assumptions.

Carrying out the time differentiation in (2.16), dividing the first and second equations by ρ_w and ρ_o, respectively, and adding the resulting equations, we obtain

$$\begin{aligned} &\frac{1}{\rho_w}\nabla\cdot\left(\frac{\rho_w}{\mu_w}\mathbf{k}_w\left(\nabla p_o-\frac{dp_c}{dS_w}\nabla S_w-\rho_w g\nabla Z\right)\right)\\ &\quad+\frac{1}{\rho_o}\nabla\cdot\left(\frac{\rho_o}{\mu_o}\mathbf{k}_o\left(\nabla p_o-\rho_o g\nabla Z\right)\right)\\ &\quad=\frac{S_w}{\rho_w}\frac{\partial(\phi\rho_w)}{\partial t}+\frac{1-S_w}{\rho_o}\frac{\partial(\phi\rho_o)}{\partial t}-\frac{q_w}{\rho_w}-\frac{q_o}{\rho_o}. \end{aligned} \tag{2.17}$$

Note that if the saturation S_w in (2.17) is taken explicitly, we can use this equation to solve for p_o. After the computation of this pressure, the second equation in (2.16) can be utilized to calculate S_w. This is the known *implicit pressure explicit saturation* (IMPES) approach and has been widely exploited in petroleum reservoirs.

2.2.2.3. Formulation in a global pressure

The equations in (2.15) and (2.16) are strongly coupled, as discussed above. To reduce the coupling, we now write them in a different formulation, where a so-called global pressure is used. For simplicity, we assume that the densities are constant; the formulation extends to variable densities (Chen-Ewing, 1997A, 1997B). Introduce the phase mobilities

$$\lambda_\alpha=\frac{k_{r\alpha}}{\mu_\alpha},\quad \alpha=w,o,$$

and the total mobility

$$\lambda=\lambda_w+\lambda_o.$$

Also, define the *fractional flow* functions

$$f_\alpha=\frac{\lambda_\alpha}{\lambda},\quad \alpha=w,o.$$

With $S = S_w$, we now define the *global pressure* (Antoncev, 1972; Chavent-Jaffré, 1978)

$$p = p_o - \int^{p_c(S)} f_w\big(p_c^{-1}(\xi)\big)d\xi \tag{2.18}$$

and the total velocity

$$\mathbf{u} = \mathbf{u_w} + \mathbf{u_o}. \tag{2.19}$$

It followss from (2.12), (2.14), and (2.18) that the total velocity is expressed as follows:

$$\mathbf{u} = -\mathbf{k}\lambda\big(\nabla p - (\rho_w f_w + \rho_o f_o)g\nabla Z\big). \tag{2.20}$$

Also, carry out the differentiation in (2.11), divide by ρ_α, add the resulting equations with $\alpha = w$ and o, and use (2.18) to see that

$$\nabla \cdot \mathbf{u} = -\frac{\partial \phi}{\partial t} + \frac{q_w}{\rho_w} + \frac{q_o}{\rho_o}. \tag{2.21}$$

Substituting (2.20) into (2.21), we have a pressure equation for p:

$$-\nabla \cdot \big(\mathbf{k}\lambda\big(\nabla p - (\rho_w f_w + \rho_o f_o)g\nabla Z\big)\big) = -\frac{\partial \phi}{\partial t} + \frac{q_w}{\rho_w} + \frac{q_o}{\rho_o}. \tag{2.22}$$

The phase velocities are related to the total velocity by

$$\begin{aligned} \mathbf{u}_w &= f_w\mathbf{u} + \mathbf{k}\lambda_o f_w \nabla p_c + \mathbf{k}\lambda_o f_w(\rho_w - \rho_o)g\nabla Z, \\ \mathbf{u}_o &= f_o\mathbf{u} - \mathbf{k}\lambda_w f_o \nabla p_c + \mathbf{k}\lambda_w f_o(\rho_o - \rho_w)g\nabla Z. \end{aligned} \tag{2.23}$$

¿From the first equation of (2.23) and (2.11) with $\alpha = w$, we have a saturation equation for $S = S_w$:

$$\begin{aligned} \phi\frac{\partial S}{\partial t} + \nabla \cdot \left(\mathbf{k}\lambda_o f_w \left(\frac{dp_c}{dS}\nabla S - (\rho_o - \rho_w)g\nabla Z\right) + f_w\mathbf{u}\right) \\ = -S\frac{\partial \phi}{\partial t} + \frac{q_w}{\rho_w}. \end{aligned} \tag{2.24}$$

While the phase mobilities can be zero, the total mobility is always positive, so the pressure equation is elliptic. If one of the densities varies, this equation becomes parabolic. Note that $-\mathbf{k}\lambda_o f_w dp_c/dS$ is generally semi-positive definite, so the saturation equation is a parabolic equation, which is degenerate in the sense that this diffusion can be zero. This equation becomes hyperbolic if the capillary pressure is ignored. The total velocity is used in the global pressure formulation. This velocity is smoother than the phase velocities. It can be used in the phase formulation (2.15) or (2.16) as well (Chen-Ewing, 1997B). We remark that the coupling between equations (2.22) and (2.24) is much less than those in (2.15) and (2.16). Finally, with $p_c = 0$, equation (2.24) becomes

the known Buckley-Leverett equation where the flux function f_w is generally nonconvex over the range of saturation values where this function is nonzero.

2.3. Three Phase Flow

We now develop basic equations for the simultaneous flow of three immiscible fluids through a porous medium. The three phases can be water, oil, and gas, for example, and we use the subscripts w, o, and g to refer to them, respectively. We assume that no mass transfer between phases occurs. This unrealistic assumption will be removed in a later section where the so-called compositional flow is considered.

The development of basic equations for three phase flow parallels that for two phases. Mass is conserved within each phase:

$$\frac{\partial(\phi\rho_\alpha S_\alpha)}{\partial t} = -\nabla\cdot(\rho_\alpha \mathbf{u}_\alpha) + q_\alpha, \qquad \alpha = w,\, o,\, g. \tag{2.28}$$

Darcy's law for each phase is written in the usual form

$$\mathbf{u}_\alpha = -\frac{1}{\mu_\alpha}\mathbf{k}_\alpha\left(\nabla p_\alpha - \rho_\alpha g \nabla Z\right), \qquad \alpha = w,\, o,\, g. \tag{2.29}$$

The fact that the three fluids jointly fill the void space is given by the equation

$$S_w + S_o + S_g = 1. \tag{2.30}$$

Finally, the phase pressures are related by capillary pressures

$$p_{cow} = p_o - p_w, \quad p_{cgo} = p_g - p_o. \tag{2.31}$$

It is not necessary to define a third capillary pressure since it can be defined in terms of the two independent ones p_{cow} and p_{cgo}.

The alternative differential equations developed for two phases can be done for three phases in a similar fashion (Chen-Ewing, 1997B). Namely, equations (2.28)–(2.31) can be rewritten in the three phase pressure formulation, in a phase pressure and two phase saturation formulation, or in a global pressure and two phase saturation formulation. Other formulations are possible. In the global formulation, the pressure equation is elliptic or parabolic depending on the effects of densities. The two saturation equations are parabolic if the capillary pressure effects dominate; otherwise, they are hyperbolic or near-hyperbolic (Chen-Ewing, 1997B).

2.4. Transport of a Component in a Fluid Phase

We now consider the transport of a component (e.g., a solute) in a fluid phase that occupies the entire void space in a (rigid) porous medium. We do not consider the effects of chemical reactions among components in the fluid phase, radioactive decay, biodegradation, and growth due to bacterial activities, that

cause the quantity of this component to increase or decrease. Conservation of mass of the component in the fluid phase is then given by

$$\begin{aligned}\frac{\partial(\phi c)}{\partial t} &= -\nabla\cdot(c\mathbf{u}-\mathbf{D}\nabla c)\\ &\quad -\sum_i q_1^{(i)}(\mathbf{x}^{(i)},t)\delta(\mathbf{x}-\mathbf{x}^{(i)})c(\mathbf{x},t)\\ &\quad +\sum_j q_2^{(j)}(\mathbf{x}^{(j)},t)\delta(\mathbf{x}-\mathbf{x}^{(j)})c^{(j)}(\mathbf{x},t),\end{aligned} \tag{2.32}$$

where c is the (volumetric) concentration of the component, $\mathbf{D}$ is the diffusion-dispersion tensor, and $q_1^{(i)}$ and $q_2^{(j)}$ are the rates of production and injection (in terms of volume per unit time) at point $\mathbf{x}^{(i)}$ and $\mathbf{x}^{(j)}$, respectively, and $c^{(j)}$ is the specified concentration at source points.

The Darcy law for the fluid is expressed as in (2.2); namely,

$$\mathbf{u} = -\frac{1}{\mu}\mathbf{k}\left(\nabla p - \rho g \nabla Z\right). \tag{2.33}$$

The mass balance of the fluid is written as follows:

$$\begin{aligned}\frac{\partial(\phi\rho)}{\partial t} + \nabla\cdot(\rho\mathbf{u}) &= -\sum_i q_1^{(i)}(\mathbf{x}^{(i)},t)\delta(\mathbf{x}-\mathbf{x}^{(i)})\\ &\quad +\sum_j q_2^{(j)}(\mathbf{x}^{(j)},t)\delta(\mathbf{x}-\mathbf{x}^{(j)}).\end{aligned} \tag{2.34}$$

The diffusion-dispersion tensor $\mathbf{D}$ in (2.32) in three space dimensions is represented by

$$\mathbf{D}(\mathbf{u}) = \phi\left\{d_m\mathbf{I} + |\mathbf{u}|\left(d_l\mathbf{E}(\mathbf{u}) + d_t\mathbf{E}^{\perp}(\mathbf{u})\right)\right\}, \tag{2.35}$$

where d_m is the molecular diffusion coefficient, d_l and d_t are, respectively, the longitudinal and transverse dispersion coefficients, $|\mathbf{u}|$ is the Euclidean norm of $\mathbf{u}$: $|\mathbf{u}| = \sqrt{u_1^2+u_2^2+u_3^2}$, $\mathbf{u} = (u_1,u_2,u_3)$, $\mathbf{E}(\mathbf{u})$ is the orthogonal projection along the velocity:

$$\mathbf{E}(\mathbf{u}) = \frac{1}{|\mathbf{u}|^2}\begin{pmatrix} u_1^2 & u_1u_2 & u_1u_3\\ u_2u_1 & u_2^2 & u_2u_3\\ u_3u_1 & u_3u_2 & u_3^2\end{pmatrix},$$

and $\mathbf{E}^{\perp}(\mathbf{u}) = \mathbf{I} - \mathbf{E}(\mathbf{u})$.

Physically, the tensor dispersion is more significant than the molecular diffusion; also, d_l is usually considerably larger than d_t. The density and viscosity are known functions of p and c:

$$\rho = \rho(p,c), \quad \mu = \mu(p,c).$$

After the substitution of (2.33) into (2.32) and (2.34), we have a coupled system of two equations in c and p. Boundary and initial conditions for this system can

be developed as in the earlier sections. Note that the equations described here apply to the miscible displacement problem of one fluid by another in a porous medium.

2.5. Transport of Multicomponents in a Fluid Phase

The equation used to model the transport of multicomponents in a fluid phase in a porous medium takes the form

$$\frac{\partial(\phi c_i)}{\partial t} = -\nabla \cdot (c_i \mathbf{u} - \mathbf{D}\nabla c_i) + q_i, \qquad i = 1, \ldots, N_c, \tag{2.36}$$

where c_i and q_i are the (volumetric) concentration and the source and sink term of the ith component, respectively, and N_c is the number of the components in the fluid. Sources and sinks of a component can result from injection and production of this component from external means. They can also stem from various processes within the fluid phase such as chemical reactions among components, radioactive decay, biodegradation, and growth due to bacterial activities, that cause the quantity of this component to increase or decrease. In this section we focus only on the chemical reactions; i.e., we consider a reactive flow problem.

When a component participates in chemical reactions which cause its concentration to increase or decrease, q_i can be expressed as

$$q_i = Q_i - L_i c_i, \tag{2.37}$$

where Q_i and L_i represent the chemical production and loss rates of the ith component, respectively. To see their expressions in terms of concentrations, we consider unimolecular, bimolecular, and termolecular reactions among the chemical components. These cases can be generally written as follows:

$$s_1 \rightleftharpoons s_2 + s_3,$$
$$s_1 + s_2 \rightleftharpoons s_3 + s_4,$$
$$s_1 + s_2 + s_3 \rightleftharpoons s_4 + s_5,$$

where the s_i's denote generic chemical components. Corresponding to these reactions, Q_i and L_i can be expressed by

$$Q_i = \sum_{j=1}^{N_c} k^f_{i,j} c_j + \sum_{j,l=1}^{N_c} k^f_{i,jl} c_j c_l + \sum_{j,l,m=1}^{N_c} k^f_{i,jlm} c_j c_l c_m,$$
$$L_i = k^r_i + \sum_{j=1}^{N_c} k^r_{i,j} c_j + \sum_{j,l=1}^{N_c} k^r_{i,jl} c_j c_l,$$

where k^f and k^r are forward and reverse chemical rates, respectively. These rates are functions of pressure and temperature.

2.6. Compositional Flow

In the earlier sections, we have assumed that no mass transfer occurs between phases. In this section we consider the multicomponent, multiphase compositional flow in a porous medium, which involves mass interchange between phases. In a model for this type of flow, a finite number of hydrocarbon components is used to represent the composition of reservoir fluids. These components associate as phases in the reservoirs. Here we describe a compositional model under the assumptions that the flow process is isothermal (i.e., the constant temperature), the components form at most three phases (e.g., gas, oil, and water), and there is no mass interchange between the water phase and the hydrocarbon phases (i.e., the oil and gas phases).

Instead of using the concentration, it will be convenient to employing the number of moles for each component in the compositional flow. Because of mass interchange between phases, mass is not conserved within each phase; the total mass of each component is conserved:

$$\begin{aligned} &\frac{\partial(\phi n_w)}{\partial t} + \nabla \cdot (\xi_w \mathbf{u}_w) = q_w, \\ &\frac{\partial(\phi n_i)}{\partial t} + \nabla \cdot (x_{ig}\xi_g \mathbf{u}_g + x_{io}\xi_o \mathbf{u}_o) = q_i, \qquad i = 1, \ldots, N_c, \end{aligned} \tag{2.38}$$

where g, o, and w refer to gas, oil, and water phases, i is the component index, N_c is the number of hydrocarbon components, n_w and n_i denote the number of overall moles per pore volume of the water and ith hydrocarbon component, x_{ig} and x_{io} are the mole fraction of the ith component in gas and oil phases, ξ_α is the molar density of the α phase, and q_w and q_i stand for the molar flow rate of the water and the ith component, respectively, $\alpha = g,\ o,\ w$. In (2.38), the volumetric velocity $\mathbf{u}_\alpha$ in multiphase flow is given by Darcy's law as in (2.2):

$$\mathbf{u}_\alpha = -\frac{1}{\mu_\alpha}\mathbf{k}_\alpha(\nabla p_\alpha - \rho_\alpha g \nabla Z), \qquad \alpha = g,\ o,\ w. \tag{2.39}$$

In addition to the differential equations (2.38) and (2.39), there are also algebraic constraints. Assume that the fluid volume completely fills the available pore volume as defined by

$$n_w v_w + \sum_{i=1}^{N_c} n_i v_i = 1, \tag{2.40}$$

where n_w and v_i represent the partial molar volumes of water and component i, respectively. Equation (2.40) is referred to as the pore volume constraint. Previously, the equation of state for the phase saturations s_α has been used:

$$S_g + S_o + S_w = 1.$$

It turns out that (2.40) is more appropriate in the numerical simulation of compositional flow. The phase pressures are related by capillary pressures:

$$p_{cow} = p_o - p_w, \quad p_{cgo} = p_g - p_o. \tag{2.41}$$

where p_{cow} and p_{cgo} represent the water and gas phase capillary pressures, respectively. They are assumed to be known functions of the saturations. The relative permeabilities $k_{r\alpha}$ are also assumed to be known in terms of the saturations. The viscosities μ_α, molar densities ξ_α, and mass densities ρ_α are functions of their respective phase pressure and compositions.

Other algebraic relations are stated as follows. The mass balance implies that

$$n_i = n_{ig} + n_{io}, \qquad i = 1, \ldots, N_c, \tag{2.42}$$

where n_{ig} and n_{io} represents the number of moles per pore volume of the ith hydrocarbon component in the oil and gas phases, respectively. Also, the mole fractions x_{ig} and x_{io} are defined by

$$x_{i\alpha} = n_{i\alpha} / \sum_{j=1}^{N_c} n_{j\alpha}, \qquad i = 1, \ldots, N_c, \quad \alpha = g,\, o. \tag{2.43}$$

Finally, the saturations are expressed in terms of the phase compositions:

$$S_w = \frac{n_w}{\xi_w}, \quad S_\alpha = \sum_{i=1}^{N_c} n_{i\alpha}/\xi_\alpha, \qquad \alpha = g,\, o. \tag{2.44}$$

It should be noted that there are more dependent variables than there are differential and algebraic relations; there are formally $5N_c + 10$ dependent variables: n_w, n_i, n_{ig}, n_{io}, x_{ig}, x_{io}, $\mathbf{u}_\alpha$, p_α, and S_α, $\alpha = g, o$, $i = 1, \ldots, N_c$. It is then necessary to have $5N_c + 10$ independent relations to determine a solution of the system. Equations (2.38)–(2.44) provide $4N_c + 10$ independent relations, differential or algebraic; the additional N_c relations are provided by the equilibrium relations needed to relate the numbers of moles, which will be addressed in the next chapter.

2.7. Black Oil Flow

The black oil model is a simplified compositional model. For this model, it is assumed that the hydrocarbon components can be divided into methane and a heavy oil component in a stock tank at the standard pressure. It is also assumed that no mass transfer occurs between the water phase and the other two phases (oil and gas) and no volatile oil exits.

Let B_α and R_{so} be the formation volume factor of the α phase, $\alpha = w, o, g$, and the gas solubility. Then the mass conservation equations of the black oil model are

$$-\nabla \cdot \left(\frac{\rho_{WS}}{B_w} \mathbf{u_w} \right) + q_w = \frac{\partial}{\partial t} \left(\phi \frac{\rho_{WS}}{B_w} S_w \right) \tag{2.45}$$

for the water component,

$$-\nabla \cdot \left(\frac{\rho_{OS}}{B_o} \mathbf{u_o} \right) + q_o = \frac{\partial}{\partial t} \left(\phi \frac{\rho_{OS}}{B_o} S_o \right) \tag{2.46}$$

for the oil component,

$$-\nabla\cdot\left(\frac{\rho_{GS}}{B_g}\mathbf{u_g}+\frac{R_{so}\rho_{GS}}{B_o}\mathbf{u_o}\right)+q_g=\frac{\partial}{\partial t}\left[\phi\left(\frac{\rho_{GS}}{B_g}S_g+\frac{R_{so}\rho_{GS}}{B_o}S_o\right)\right] \tag{2.47}$$

for the gas component, where $\rho_{\beta S}$ and q_β are the density at standard conditions (stock tank) and the mass flow rate at wells of the β component, $\beta = W, O, G$, respectively.

2.8. Non-Isothermal Flow

The differential equations have so far developed under the condition that flow is isothermal. This condition can be removed by adding a conservation of energy equation. This equation introduces an additional dependent variable, temperature, to the system. Unlike the case of mass transport, where the solid itself is assumed impervious to mass flux, the solid matrix does conduct heat. The average temperature of the solid and fluids in a porous medium may not be the same. Furthermore, heat may exchange between phases. For simplicity, we invoke the requirement of local thermal equilibrium that the temperature be the same in all phases.

As in the previous section, we consider a multicomponent, multiphase flow in a porous medium. The mass balance and other equations can be presented as in (2.38)–(2.44). Under the nonisothermal condition, some of the variables such as the porosity, densities, and viscosities may depend on temperature. The overall energy balance equation is stated as follows:

$$\begin{aligned}&\frac{\partial}{\partial t}\left(\rho_t U+\frac{1}{2}\sum_{\alpha=w}^{g}\rho_\alpha|\mathbf{u}_\alpha|^2\right)\\&\quad+\nabla\cdot\left(\sum_{\alpha=w}^{g}\rho_\alpha\mathbf{u}_\alpha\left(H_\alpha+\frac{1}{2}|\mathbf{u}_\alpha|^2\right)\right)\\&\quad-\nabla\cdot(k_T\nabla T)+\sum_{\alpha=w}^{g}\rho_\alpha\mathbf{u}_\alpha\cdot g\nabla Z=q_T,\end{aligned} \tag{2.48}$$

where $\rho_t U$ is the total internal energy, H_α is the enthalpy of the α-phase (per unit mass), k_T is the total thermal conductivity, T is the temperature, and q_T is the external heating source term. The total internal energy is defined by

$$\rho_t U=\phi\sum_{\alpha=w}^{g}\rho_\alpha S_\alpha U_\alpha+(1-\phi)\rho_s U_s, \tag{2.49}$$

where U_α and U_s are the internal energies per unit mass of phase α and the solid, respectively, and ρ_s is the density of the solid. The overall density ρ_t is determined by

$$\rho_t=\phi\sum_{\alpha=w}^{g}\rho_\alpha S_\alpha+(1-\phi)\rho_s.$$

The enthalpies are related to the pressures by

$$H_\alpha = U_\alpha + \frac{p_\alpha}{\rho_\alpha}, \quad \alpha = w, o, g.$$

Note that equations (2.45) and (2.46) are two equations for the nine unknows U, T, U_s, U_α, and H_α, $\alpha = w, o, g$, so more relations are needed to have a complete specification of the nonisothermal flow problem. The phase internal energies U_α and U_s and enthalpies H_α can be determined in terms of their respective partial molar quantities:

$$\begin{aligned} U_\alpha &= \sum_{i=1}^{N_c} x_{i\alpha} U_{i\alpha}, \qquad U_s = \sum_{i=1}^{N_c} x_{is} U_{is}, \\ H_\alpha &= \sum_{i=1}^{N_c} x_{i\alpha} H_{i\alpha}, \qquad \alpha = w,\ o,\ g. \end{aligned} \tag{2.50}$$

These partial molar quantities $U_{i\alpha}$, U_{is}, and $H_{i\alpha}$ are functions of temperature, phase pressure p_α, and composition $x_{i\alpha}$:

$$\begin{aligned} U_{i\alpha} &= U_{i\alpha}(T, p_w, p_o, p_g, x_{1w}, \ldots, x_{N_c w}, \ldots, x_{1g}, \ldots, x_{N_c g}), \\ U_{is} &= U_{is}(T, p_w, p_o, p_g, x_{1w}, \ldots, x_{N_c w}, \ldots, x_{1g}, \ldots, x_{N_c g}), \\ H_{i\alpha} &= H_{i\alpha}(T, p_w, p_o, p_g, x_{1w}, \ldots, x_{N_c w}, \ldots, x_{1g}, \ldots, x_{N_c g}), \\ &\qquad \alpha = w,\ o,\ g, \quad i = 1, \ldots, N_c. \end{aligned} \tag{2.51}$$

Now it can be checked that (2.45)–(2.48) gives $7N_c + 9$ equations for the same number of dependent variables U, T, U_s, U_α, H_α, $U_{i\alpha}$, U_{is}, and $H_{i\alpha}$, $\alpha = w, o, g$, $i = 1, \ldots, N_c$.

3. NUMERICAL METHODS

As an example, we consider the flow of two incompressible, immiscible fluids in a porous medium $\Omega \subset \Re^3$. The mass balance equation for each of the fluid phases is

$$\phi \frac{\partial(\rho_\alpha s_\alpha)}{\partial t} + \nabla \cdot (\rho_\alpha \mathbf{u}_\alpha) = \rho_\alpha q_\alpha, \quad \alpha = w, o, \tag{3.1}$$

where $\alpha = w$ denotes the wetting phase (e.g., water), $\alpha = o$ indicates the nonwetting phase (e.g., oil), ϕ is the porosity of the medium, and ρ_α, s_α, $\mathbf{u}_\alpha$, and q_α are, respectively, the density, saturation, volumetric velocity, and external volumetric flow rate of the α-phase. The volumetric velocity $\mathbf{u}_\alpha$ is given by the Darcy law

$$\mathbf{u}_\alpha = -\frac{K_{r\alpha}}{\mu_\alpha} \mathbf{K} \nabla (p_\alpha - \rho_\alpha g Z), \quad \alpha = w, o, \tag{3.2}$$

where $\mathbf{K}$ is the absolute permeability of the porous medium, p_α, μ_α, and $K_{r\alpha}$ are the pressure, viscosity, and relative permeability of the α-phase, respectively,

g denotes the gravitational constant, Z is the depth, and the z-coordinate is in the vertical downward direction. In addition to (3.1) and (3.2), the constraint for the saturations is

$$s_w + s_o = 1,$$

and the two pressures are related by the capillary pressure function

$$p_c(s_w) = p_o - p_w.$$

We introduce the phase mobility functions

$$\lambda_\alpha(\mathbf{x}, s_\alpha) = K_{r\alpha}/\mu_\alpha, \quad \alpha = w, o,$$

and the total mobility

$$\lambda(\mathbf{x}, s) = \lambda_w + \lambda_o,$$

where $s = s_w$. The fractional flow functions are defined by

$$f_\alpha(\mathbf{x}, s) = \lambda_\alpha/\lambda, \quad \alpha = w, o.$$

The model is completed by specifying boundary and initial conditions. In this paper we consider no flow boundary conditions

$$\mathbf{u}_\alpha \cdot \boldsymbol{\nu} = 0, \qquad \alpha = w, o, \ \mathbf{x} \in \partial\Omega,$$

where $\boldsymbol{\nu}$ is the outer unit normal to the boundary $\partial\Omega$ of Ω. The initial condition is given by

$$s(\mathbf{x}, 0) = s_0(\mathbf{x}), \qquad \mathbf{x} \in \Omega.$$

For a theoretical study of the model in this section, the reader may refer to the paper by Chen (2001), for example.

3.1. The Classical IMPES Method

We use the oil phase pressure as the pressure variable

$$p = p_o,$$

and define the total velocity

$$\mathbf{u} = \mathbf{u}_w + \mathbf{u}_o.$$

Under the assumption that the fluids are incompressible, we see that

$$\nabla \cdot \mathbf{u} = q(p, s) \equiv q_w(p, s) + q_o(p, s),$$

and

$$\mathbf{u} = -\mathbf{K}\left(\lambda(s)\nabla p - \lambda_w(s)\nabla p_c - \big(\lambda_w\rho_w + \lambda_o\rho_o\big)g\nabla Z\right).$$

Similarly, we have

$$\phi\frac{\partial s}{\partial t} + \nabla \cdot \left\{\mathbf{K}f_w(s)\lambda_o(s)\left(\frac{dp_c}{ds}\nabla s + (\rho_o - \rho_w)g\nabla Z\right) + f_w(s)\mathbf{u}\right\} = q_w(p, s).$$

The pressure equation is now given by

$$-\nabla \cdot \big(\mathbf{K}\lambda\nabla p\big) = q - \nabla \cdot \big(\mathbf{K}(\lambda_w \nabla p_c + (\lambda_w \rho_w + \lambda_o \rho_o) g \nabla Z\big). \tag{3.3}$$

Let $J = (0, T]$ $(T > 0)$ be the time interval of interest, and for a positive integer N, let $0 = t^0 < t^1 < \ldots < t^N = T$ be a partition of J. In the pressure computation in the IMPES method, the saturation s in (3.3), is supposed to be known, and (3.3) is solved implicitly for p. That is, for each $n = 0, 1, \ldots$, p^n satisfies

$$-\nabla \cdot \big(\mathbf{K}\lambda(s^n)\nabla p^n\big) = F(p^n, s^n), \tag{3.4}$$

where $F(p, s)$ denotes the right-hand side of (3.3) and s^n is supposed to be given.

It follows that

$$\phi \frac{\partial s}{\partial t} = q_w - \nabla \cdot \left\{ \mathbf{K} f_w(s) \lambda_o(s) \left(\frac{dp_c}{ds} \nabla s + (\rho_o - \rho_w) g \nabla Z \right) + f_w(s) \mathbf{u} \right\}. \tag{3.5}$$

In this classical IMPES method, (3.5) is explicitly solved for s; i.e., for each $n = 0, 1, 2, \ldots$, s^{n+1} satisfies

$$\phi \frac{\partial s^{n+1}}{\partial t} = G(p^n, \mathbf{u}^n, s^n), \tag{3.6}$$

where $G(p, \mathbf{u}, s)$ represents the right-hand side of (3.4).

Now, the standard IMPES method goes as follows: After startup, for $n = 0, 1, \ldots$, we use (3.4) and s^n to evaluate p^n and then to evaluate $\mathbf{u}^n$; next, we exploit s^n, p^n, $\mathbf{u}^n$, and (3.6) to compute s^{n+1}. As noted, the time step $\Delta t^n = t^n - t^{n-1}$ must be sufficiently small for this method to be stable.

3.2. An Improved IMPES Method

As discussed in the previous section, most of the computational time in the classical IMPES method is spent on the implicit calculation of the pressure. Also, it follows from the mechanics of fluid flow in porous media that the pressure changes less rapidly in time than the saturation. Furthermore, the constraint for time steps is primarily used for the explicit calculation of the saturation. For all these reasons, it is appropriate to take a much larger time step for the pressure than for the saturation.

Again, for a positive integer N, let $0 = t^0 < t^1 < \ldots < t^N = T$ be a partition of J for the pressure into subintervals $J^n = (t^{n-1}, t^n]$, with length $\Delta t_p^n = t^n - t^{n-1}$. Each subinterval J^n is divided into sub-subintervals $J^{n,m} = (t^{n-1,m-1}, t^{n-1,m}]$ for the saturation:

$$t^{n-1,m} = t^{n-1} + m\Delta t_p^n / M^n, \quad m = 1, \ldots, M^n.$$

The length of $J^{n,m}$ is denoted by $\Delta t_s^{n,m} = t^{n-1,m} - t^{n-1,m-1}$, $m = 1, \ldots, M^n$, $n = 0, 1, \ldots$. The number of steps, M^n, can depend on n. Below we simply write $t^{n-1,0} = t^{n-1}$, and set $v^{n,m} = v(\cdot, t^{n,m})$.

We denote the right-hand side of the equation for $\mathbf{u}$ by $\mathbf{H}(p, s)$. Now, the improved IMPES method is defined as follows: For each $n = 0, 1, \ldots$, find p^n such that

$$-\nabla \cdot \big(\mathbf{K}\lambda(s^n)\nabla p^n\big) = F(p^n, s^n),$$

and $\mathbf{u}^n$ such that

$$\mathbf{u}^n = \mathbf{H}(p^n, s^n).$$

Next, for $m = 1, \ldots, M^n$, $n = 0, 1, \ldots$, find $s^{n+1,m}$ such that

$$\phi \frac{\partial s^{n+1,m}}{\partial t} = G(p^n, \mathbf{u}^n, s^{n+1,m-1}).$$

The time step $\Delta t_s^{n+1,m}$ is chosen as follows: Set

$$\left(\frac{\partial s^{n+1,m}}{\partial t}\right)_{\max} = \left(\frac{G(p^n, \mathbf{u}^n, s^{n+1,m-1})}{\phi}\right)_{\max},$$

and then calculate

$$\Delta t_s^{n+1,m} = \frac{DS_{\max}}{\left(\dfrac{\partial s^{n+1,m}}{\partial t}\right)_{\max}}, \quad m = 1, \ldots, M^n, \ n = 0, 1, \ldots.$$

For numerical results of these two methods as well as comparisons between them, please refer to Chen *et al.*, 2003.

References

[1] S. N. Antontsev (1972), On the solvability of boundary value problems for degenerate two-phase porous flow equations, *Dinamika Sploš̆noĭ Sredy Vyp.* **10**, 28–53, in Russian.

[2] K. Aziz and A. Settari (1979), Petroleum Reservoir Simulation, Applied Science Publishers Ltd, London.

[3] G. I. Barenblatt, Iu. P. Zheltov, and I. N. Kochina (1960), Basic concepts in the theory of seepage of homogeneous liquids in fissured rocks [strata], *Prikl. Mat. Mekh.* **24**, 852–864.

[4] M. A. Biot (1955), Theory of elasticity and consolidation for a porous anisotropic solid, *J. Appl. Phys.* **26**, 182–185.

[5] G. Chavent and J. Jaffré (1978), Mathematical Models and Finite Elements for Reservoir Simulation, North-Holland, Amsterdam.

[6] Z. Chen and R. E. Ewing (1997A), Fully-discrete finite element analysis of multiphase flow in groundwater hydrology, *SIAM J. Numer. Anal.* **34**, 2228–2253.

[7] Z. Chen and R. E. Ewing (1997B), Comparison of various formulations of three-phase flow in porous media, *J. Comp. Physics* **132**, 362–373.

[8] Z. Chen, G. Huan, and B. Li (2003), An improved IMPES method for two-phase flow in porous media, *Transport in Porous Media*, **54** (2004), 361–376.

[9] Z. Chen, G. Huan, and Y. Ma (2006), Computational Methods for Multiphase Flows in Porous Media, in the Computational Science and Engineering Series, Vol. 2, SIAM, Philadephia, PA, 2006.

[10] Z. Chen, S. Lyons, and G. Qin (2001), Derivation of the Forchheimer law via homogenization, *Transport in Porous Media* **44**, 325–335.

[11] Z. Chen, S. Lyons, and G. Qin (2004), The mechanical behavior of poroelastic media saturated with a Newtonian fluid derived via homogenization, *International Journal of Numerical Analysis and Modeling* **1**, 75–98.

[12] J. Douglas, Jr. and T. Arbogast (1990), Dual-porosity models for flow in naturally fractured reservoirs, in Dynamics of Fluids in Hierarchical Porous Media, J. H. Cushman, ed., 177–221.

[13] J. Douglas, Jr., D. W. Peaceman, and H. H. Rachford, Jr. (1959), A method for calculating multi-dimensional immiscible displacement, *Trans. SPE of AIME* **216**, 297–306.

[14] J. Dupuit (1863), Estudes Theoriques et Pratiques sur le Mouvement des Eaux, Dunod.

[15] P. Forchheimer (1901), Wasserbewegung durch Boden, *VDIZ.* **45**, 1782–1788.

[16] P. M. Knupp and J. L. Lage (1995), Generalization of the Forchheimer-extended Darcy flow model to the tensor permeability case via a variational principle, *J. Fluid Mech.* **299**, 97–104.

[17] I. F. MacDonald, M. S. El-Sayed, K. Mow, and F. A. L. Dullien (1979), Flow through porous media: the Ergun equation revisited, *Indust. Chem. Fundam.* **18**, 199–208.

[18] C. C. Mei and J.-L. Auriault (1991), The effect of the weak inertia on flow through a porous medium, *J. Fluid Mech.* **222**, 647–663.

[19] J. C. Ward (1964), Turbulent flow in porous media, *J. Hydr. Div. ASCE* **90**, 1–12.

Bol. Soc. Esp. Mat. Apl. n° 37(2006), 93–121

¿Podemos fiarnos de los cálculos efectuados con ordenador?*

Óscar Ciaurri y Juan Luis Varona

oscar.ciaurri@dmc.unirioja.es, jvarona@dmc.unirioja.es

Departamento de Matemáticas y Computación
Universidad de La Rioja
Calle Luis de Ulloa s/n
26004 Logroño, Spain

Resumen

En este artículo analizamos la fiabilidad de los cálculos hechos con ordenador. Es bien conocido que, si hemos usado algoritmos numéricos y aritmética de coma flotante, los resultados pueden estar afectados por diversos problemas de estabilidad y pérdida de precisión. Pero, ¿qué ocurre con los cálculos simbólicos efectuados por medio de los potentes programas de álgebra computacional de los que ahora disponemos? Por medio de numerosos ejemplos, y particularizando en el programa Mathematica, vemos que abundan los fallos, así como los comportamientos distintos de lo razonablemente esperado. Estos problemas son difíciles de detectar, y sólo los conocimientos matemáticos del usuario pueden ayudar a paliar sus efectos.

A todos nos han puesto alguna vez excusas del tipo «no ha podido ser por culpa del ordenador». Así que, antes de comenzar, aclaremos que no vamos aquí a caer en ello. Los pobres ordenadores no tienen la culpa de nada; son las personas que han escrito los programas quienes consiguen que funcionen de una manera u otra. Algunas veces, a nuestra entera satisfacción; otras, no tanto.

Pero el caso es que ahí tenemos a nuestro ordenador y a sus programas, que son los que interactúan con nosotros. Generalmente, el programador es, para el usuario, totalmente desconocido. Como si no existiera. No podemos plantearle ninguna aclaración. Tenemos que confiar (o no) en lo que el ordenador nos dice.

*Este artículo ha sido previamente publicado en La Gaceta de la RSME, **9** (2006), no. 2, 483–514. Agradecemos su permiso para reproducirlo.

La investigación de los autores está subvencionada, en parte, por el proyecto BFM2003-06335-C03-03 de la DGI.

Y la realidad es que los ordenadores se introducen cada vez más en nuestras vidas. Comprobar que el *software* funciona como estaba previsto que lo hiciera es una tarea ardua, y su verificación formal es prácticamente imposible. Los programas que no incorporan ningún *bug*[1] son muy raros. Fallos tontos pueden dar lugar a catástrofes. Muy conocido es el mal funcionamiento de ciertos misiles defensivos Patriot durante la Guerra del Golfo por no usar números de suficiente precisión; esto supuso, el 25 de febrero de 1991, la muerte de 28 soldados norteamericanos en Arabia Saudí. O la explosión, el 4 de junio de 1996, del primer cohete Ariane 5 que se lanzaba; el cohete pertenecía a la Agencia Espacial Europea, y en su desarrollo se habían empleado 10 años y $7 \cdot 10^9$ dólares. El problema fue similar: una confusión entre números enteros y reales en el programa informático que controlaba el lanzamiento.[2] En [11] podemos encontrar muchos otros ejemplos, así como un análisis de las medidas que se intentan tomar para evitar todo tipo de fallos informáticos (lamentablemente, muchas veces las medidas sólo fueron propuestas, pero no llevadas a la práctica). También son muy interesantes las reflexiones de [9]; pese al tiempo transcurrido y a la rápida evolución de la informática, no han perdido nada de actualidad.

No sólo el software nos puede dar quebraderos de cabeza. También pueden fallar los circuitos electrónicos, el denominado *hardware*. El ejemplo más conocido es el del error en los primeros microprocesadores Pentium, que dividían mal determinadas combinaciones de números. No todo el mundo sabe que fue un matemático —Thomas R. Nicely— quien descubrió el fallo. En 1994, con ayuda de varios ordenadores, Nicely estaba intentando estimar la suma de los inversos de los números primos gemelos,[3] pero le aparecían resultados inconsistentes. Tras arduos esfuerzos, logró darse cuenta de que el microprocesador no dividía bien (véase [10]). Como consecuencia, el 20 de diciembre de 1994, Intel, la multinacional fabricante del Pentium, ofreció reemplazar, gratis, todos los chips defectuosos.

En lo que a nosotros —matemáticos— concierne, los ordenadores ya desempeñan un papel fundamental para ayudarnos a realizar cálculos, tanto numéricos como simbólicos. Tal es así que, entre otros colectivos, hay gente

[1] Se acostumbra a denominar *bugs* (en inglés, 'bichos') a los errores en los programas. La historia cuenta que, en los primeros tiempos de la informática, cierto fallo de funcionamiento en un ordenador se debía a que en sus circuitos se había introducido un insecto. Este hecho popularizó la palabra. En la entrada '*Computer bug*' de Wikipedia (`http://en.wikipedia.org/wiki/Computer_bug`) se pueden encontrar más detalles sobre la etimología del término (¡incluso una foto del insecto!), junto con información adicional y otras anécdotas.

[2] No siempre se puede achacar a la informática el fracaso de las misiones espaciales: el 23 de septiembre 1999, la sonda Mars Climate Orbiter se estrelló contra Marte. La explicación que dio la NASA es que una empresa involucrada en la construcción de la sonda había suministrado los datos requeridos en unidades británicas (millas, libras...), no en unidades del sistema internacional (kilómetros, kilogramos...) como la NASA esperaba. Contrariamente a nuestra opinión, hay quienes también echan la culpa a la informática de esto, y lo ponen como ejemplo de bug desastroso; pero, ¿qué podía hacer el programa que controlaba el descenso de la sonda —por muy bien hecho que estuviera— si los datos que tenía que usar estaban equivocados de antemano?

[3] No se sabe si hay o no infinitos de ellos, pero sí que la suma de sus inversos es finita (por el contrario, la serie de los inversos de los números primos es divergente).

tentada a opinar que aprender matemáticas es cada vez menos importante, pues los ordenadores hacen «todo» lo necesario. La vieja polémica de las calculadoras se ha hecho mayor. De hecho, habitualmente, los ordenadores operan o representan gráficas mejor que cualquier matemático experto; y en las tareas repetitivas no se aburren. Pero no hay que alarmarse; nuestra profesión no va a desaparecer. Es realmente difícil que alguien logre indicarle al ordenador que le resuelva algún problema si ni siquiera sabe plantear lo que quiere, ni darle las instrucciones oportunas. Por el contrario, si sabe hacerlo, es que *sabe matemáticas*; en ese caso, hará muy bien en servirse del ordenador para que le ayude.

Aún así, todavía debemos estar preocupados. Los programas de cálculo actuales son impresionantes. Sus posibilidades son grandiosas, y el tiempo que nos ahorran podemos emplearlo —con gran satisfacción— en realizar las múltiples e imprescindibles tareas burocráticas a las que nuestra profesión nos obliga, y en rellenar el currículum en diecisiete formatos diferentes. Aunque hay un pero: ¿hasta qué punto podemos fiarnos de las respuestas que nos da un ordenador? Lamentablemente, nuestra experiencia nos hace ser algo pesimistas.

¿O quizás tenemos que mostrarnos contentos? En realidad, lo que observamos es que debemos ser muy precavidos con las respuestas de un ordenador. Ante un problema numérico, hay que tener mucho cuidado con aceptar lo que un programa informático afirma sin pensar qué puede estar haciendo por dentro: quizás ha podido emplear un algoritmo que no es estable, por ejemplo. Pero no sólo las respuestas numéricas son propensas a errores; también los hemos observado en las simbólicas. En general, de estas últimas suele ser más difícil darse cuenta, por inesperadas. Hace falta saber bastantes matemáticas, conocer el funcionamiento interno de los ordenadores, el tipo de algoritmos que suelen usar para responder a lo que le hayamos planteado, dónde un programador puede estar pasando algo por alto, si la respuesta es o no razonable... También son necesarias ciertas dosis de intuición. Todo esto sólo se consigue con conocimientos matemáticos y experiencia. Definitivamente, las máquinas no nos suplantan.

Pero —hablando más en serio— no deja de ser desalentador; por muy expertos y cuidadosos que seamos, a veces es muy difícil detectar errores imprevisibles. Los fallos de hardware parecen menos peligrosos; hay mucho menos hardware diferente que software, luego el control de calidad es mucho mayor (además, los errores serios son fácilmente detectables pues el ordenador falla estrepitosamente). Por si queda alguno, hay una solución relativamente sencilla: podemos ejecutar el mismo programa en un ordenador con una circuitería distinta.[4]

Otro tipo de problemas que muchos considerarán de ciencia ficción es la interacción de los rayos cósmicos. Al colisionar con los chips de memoria del ordenador, pueden alterar un *bit*, que quizás tenga influencia en lo que el ordenador está haciendo.[5] Esto es muy improbable, pero casi seguro que un

[4]En particular, esto —y muchos otros argumentos similares— es una razón de peso en contra de los monopolios informáticos.

[5]Todavía ningún sistema operativo lo ha usado como excusa para justificar sus cuelgues,

proyecto que involucre varios años de uso de CPU se habrá topado con ello. Normalmente, esto no afecta a nuestro quehacer diario pero, si fuese necesario, repitiendo cálculos se solventa el problema.

Los fallos más dañinos son los de software. Si un paquete destinado a una determinada tarea la hace mal por estar mal diseñado, la hará igual de mal en todo tipo de ordenadores (y esto, si es que existen versiones para más de una plataforma informática, claro). Chequear lo que estamos obteniendo por medio de otro paquete distinto no es fácil, y requiere mucho esfuerzo; en particular, habrá que aprender otra sintaxis para lograr plantear nuestro problema en el ordenador.

Todos estamos —o deberíamos estarlo— prevenidos contra los posibles errores que cualquier método numérico acarrea. En primer lugar, la aritmética real finita que usan tales métodos hace que muchos números debamos manejarlos de manera aproximada; hasta tal punto es así que esto es causa de que las operaciones básicas de sumar y multiplicar no conserven sus propiedades habituales (conmutatividad, asociatividad y distributividad). Además, los procesos infinitos del análisis matemático hay que abreviarlos, lo cual contribuye a introducir nuevos errores. También aparecen problemas de inestabilidad, que no siempre son fáciles de detectar ni analizar. No seguiremos por esta senda; simplemente, somos conscientes de las dificultades y esto nos hace estar alerta. No ocurre así cuando estamos tratando con los actuales programas que incluyen cálculo simbólico —también llamados sistemas de álgebra computacional—. Al fin y al cabo, no usan una aritmética finita, y tampoco deberían emplear algoritmos aproximados, sino exactos. Cualquiera podría pensar que, ahora, los errores no deberían existir. Si los hay, es fácil que pillen al usuario con la guardia baja.

Por otra parte, existen programas numéricos libres[6] de calidad contrastada, y también magníficos paquetes simbólicos especializados en diversos campos de la matemática. Pero no es así con los programas simbólicos de propósito general, que —salvo honrosas excepciones que no llegan al nivel de desarrollo de los líderes— están comercializados por diversas empresas. Esto quiere decir dos cosas: En primer lugar, lo habitual es que sean carísimos.[7] En segundo lugar, no podemos ver su código fuente, y por tanto no podemos saber cómo funcionan. Para nosotros, son «cajas negras», les metemos unos datos y nos proporcionan otros; si algo no va bien, es muy difícil averiguar el motivo, por

pero llegará el día...

[6]En inglés, la palabra *free* que describe a estos programas significa tanto 'libre' (en el sentido de que podemos coger el código fuente del programa y hacer con él lo que queramos, dentro de ciertos límites razonables que nos impone la licencia), como 'gratis' (¡no nos cuestan dinero!). Aunque a gran parte de los usuarios no les preocupa, la accesibilidad al código es fundamental dentro del concepto de «software libre»; quizás incluso prevaleciendo sobre el precio.

[7]Esto no sólo es un problema para el investigador que quiere usarlos, que quizás tenga dinero para ello, sino sobre todo para el profesor que quiere emplearlos en clase y pretende que los alumnos los puedan instalar en sus ordenadores personales para utilizarlos como herramienta habitual. Nunca hemos encontrado una solución apropiada para solventar esta dificultad.

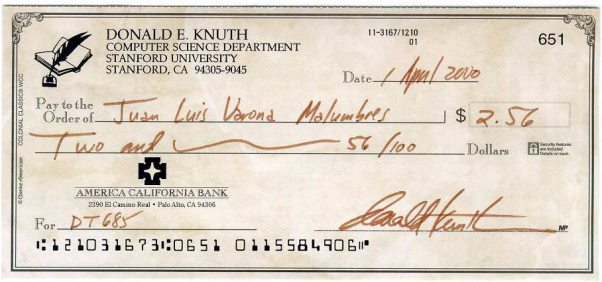


Figura 1: Un cheque firmado por Donald Knuth.

muy expertos que seamos. Más aún, si descubrimos un fallo, no podemos hacer nada por arreglarlo, ni podemos conseguir que el fabricante repare su producto defectuoso y nos suministre uno nuevo que funcione.[8] Si somos afortunados, podemos lograr que el fabricante atienda a nuestras comunicaciones sobre el fallo; e incluso que intente arreglarlo en una versión posterior del producto. Eso sí, la nueva versión nos la cobrará de nuevo, y es fácil que el arreglo de una cosa haya estropeado otras trece (algo exclusivamente achacable a la mala suerte, naturalmente).

Posiblemente, los paquetes de cálculo simbólico genéricos son los únicos programas no libres que los matemáticos nos vemos forzados a utilizar,[9] pues no tienen contrapartida libre. Al menos, no la tienen con una potencia y facilidad de uso similares. No sería mala idea que todas las mañanas implorásemos a Dios para que nos proporcionara el tan deseado sistema de álgebra computacional libre de propósito general; bueno, quizás alguien tenga un día una idea mejor.[10]

En este artículo vamos a mostrar varios fallos de funcionamiento, o comportamiento distinto al esperado, de uno de los paquetes de cálculo simbólico más extendidos: Mathematica [15]. Que nadie entienda que esto es una crítica a este programa. En realidad, es el manipulador algebraico que más usamos los autores, y por tanto el que conocemos con más profundidad, y el que más nos

[8] Es sorprendente que las leyes permitan esto a la industria del software. Tal como ya hemos comentado, no obró así Intel cuando se descubrió su fallo en el Pentium; ni lo hacen los fabricantes de coches si se detecta que un modelo que está en el mercado tiene un defecto, por citar otro ejemplo.

[9] Bueno, salvo cuando las diversas administraciones públicas nos obligan a usar..., y mejor no seguir para evitar cabrearnos (¡uy!, perdón por la palabrota).

[10] Por supuesto, el paradigma de software libre que los matemáticos usamos a todas horas es TEX. Hace ya bastantes años que Donald Knuth —su autor— decidió dejar de actualizarlo, salvo para corregir posibles errores en el programa. Sus motivos eran que esto iba a permitir librarse de ellos y, además, aseguraría la compatibilidad futura de los documentos escritos con TEX. Así mismo, decidió ofrecer recompensas monetarias —cuyas cuantías crecían con potencias de 2— a los que encontraran algún bug en TEX. Actualmente, si alguien descubriera un nuevo error, recibiría 327,68 dólares. Donald Knuth también gratifica (con 256 céntimos de dólar o, como él lo llama, un «dolar hexadecimal») a quien halla erratas en sus libros. Numerosos «gazapólogos» y coleccionistas de cheques incobrados hemos detectado bastantes de ellos (véase la figura 1), que serán corregidos en posteriores ediciones de los libros.

gusta; incluso hemos escrito algunos artículos en los que mostramos su utilidad para diversas tareas (véanse [12, 13]). El tipo de problemas que aquí presentamos se da en todos los paquetes informáticos similares (en [14] puede encontrarse una excelente comparativa de las habilidades matemáticas de siete de estos programas). Pero, como Mathematica es el que mejor conocemos, ha tenido la «mala suerte» de que nos cebemos con él; sirva esto de estímulo para que Wolfram Research —la empresa que lo desarrolla— lo continúe mejorando. Insistimos en que nuestra pretensión es concienciar al usuario de que, en cualquier caso, debe permanecer alerta, y proporcionarle ideas que le pueden resultar útiles para detectar los posibles cálculos erróneos efectuados por este tipo de software.

Eso sí, no vamos a llegar al nivel de alarma de Householder, uno de los pioneros del álgebra lineal numérica y destacada figura de la matemática aplicada, quien declaró en cierta ocasión que nunca volaría en un aeroplano que hubiera sido diseñado con la aritmética de punto flotante, insistiendo en que no se podía confiar en la exactitud de tales cálculos.

Varios de los fallos que aquí presentamos los han descubierto quienes esto escriben. Algunos nos han aparecido en situaciones de nuestro trabajo como matemáticos, y los hemos aislado hasta encontrar un ejemplo sencillo en el que se manifestara el funcionamiento incorrecto. En numerosas ocasiones los hemos comentado con diversos colegas; su reacción ha sido siempre una combinación de sorpresa y desencanto. Del mismo modo, otros ejemplos nos han sido proporcionados por otros matemáticos, o los hemos leído en algún foro de internet. Tanto a los que nos los han comunicado personalmente como a los que los han divulgado en la red se lo agradecemos, pero seguro que errábamos si intentábamos dar una relación de todos ellos (lo cual sería, por supuesto, un «metafallo»).

Hemos analizado diversas versiones no muy antiguas de Mathematica; en concreto, las versiones 3.0, 4.1, 4.2, 5.0 y 5.1 (para ser más precisos, 3.0.1, 4.1.5, 4.2.1, 5.0.0 y 5.1.0; pero no hay ningún tipo de información de qué es lo que cambia cuando varía el tercer dígito, ni existe ninguna publicidad de tales lanzamientos, ni conocemos forma alguna de actualizar el programa cuando esto ocurre; así que no hemos tenido en cuenta este tercer dígito en el número de versión). No siempre todas las versiones se comportan igual cuando se enfrentan al mismo problema. Es más, veremos que, en ocasiones, las más nuevas lo hacen peor. A menudo usaremos la notación de Mathematica sin explicarla explícitamente. No hay ninguna necesidad de que el lector la conozca; ni siquiera de que haya manejado Mathematica alguna vez. Con un poco de buena voluntad, todo lo que aquí aparece será perfectamente comprensible por cualquiera que tenga conocimientos matemáticos de nivel preuniversitario o —para algunos ejemplos— de un primer curso universitario de cálculo.

1 Cálculo de límites

Desde la versión 5.0, Mathematica ya conoce la equivalencia de Stirling. Así, ya sabe que el límite

```
Limit[n^(n+1/2)*E^(-n)/n!, n -> Infinity]
```

vale $1/\sqrt{2\pi}$. Una buena mejora. Pero no queremos con esto decir que las versiones anteriores fueran malas porque no supieran calcularlo. No es ése el problema; si no saben y nos devuelven el límite sin hacer (como así ocurría), no nos engañan. Lo grave no es eso, sino que nos proporcione resultados erróneos.

Y es que eso también sucede. Hasta el punto que nos atrevemos a decir que el tratamiento que hace Mathematica de los límites es una chapuza. Sin paliativos. Si pensamos en funciones de variable real, todos sabemos que, para que exista el límite, tiene que haber límite por la derecha y por la izquierda. En cambio, Mathematica parece ignorarlo; así, por ejemplo, afirma alegremente que

```
Limit[Sin[x]/Abs[x], x -> 0]
```

vale 1. Y no acertamos a comprender por qué lo hace así, pues es capaz de calcular los dos límites laterales, que son 1 por la derecha y -1 por la izquierda. Esto se consigue, respectivamente, con la sintaxis

```
Limit[Sin[x]/Abs[x], x -> 0, Direction -> -1]
Limit[Sin[x]/Abs[x], x -> 0, Direction -> 1]
```

Aun así, podemos ser benevolentes. Si estamos advertidos de este fallo, no es un problema muy grave; basta que nos acordemos de que debemos calcular ambos límites laterales. ¿Pero qué pasa si no lo sabíamos?

2 Gráficos ficticios

Representando funciones, Mathematica es bastante bueno.[11] Por ejemplo, es capaz de dibujar la gráfica de $\operatorname{sen}(x)/x$ sin dificultad. Sin embargo, cuando la función tiene una discontinuidad de salto —como $\operatorname{sen}(x)/|x|$ o $\operatorname{tg}(x)$—, suele incluir en el salto una línea vertical inexistente, incluso aunque los saltos no sean finitos. Se lo podemos perdonar. Lo normal es que cualquier programa que dibuja funciones lo haga «dando valores» (como muchos alumnos); así, es difícil que detecte las discontinuidades. Quizás sería exigir demasiado pensar que un programa puede desenvolverse ante un concepto matemático como puede ser el de discontinuidad; podemos pues asumir que se lo teníamos que haber indicado de alguna manera.

Además, dar valores lo hace con una cierta «inteligencia». Por defecto, utiliza cierta cantidad de puntos equidistantes (su número depende de la versión de Mathematica que estemos usando); pero incluye un algoritmo de muestreo adaptativo que muchas veces le hace tomar puntos adicionales. Esto suele

[11] Aunque podría mejorar; otros paquetes son mucho más útiles para dibujar funciones implícitas en dos e incluso tres dimensiones.

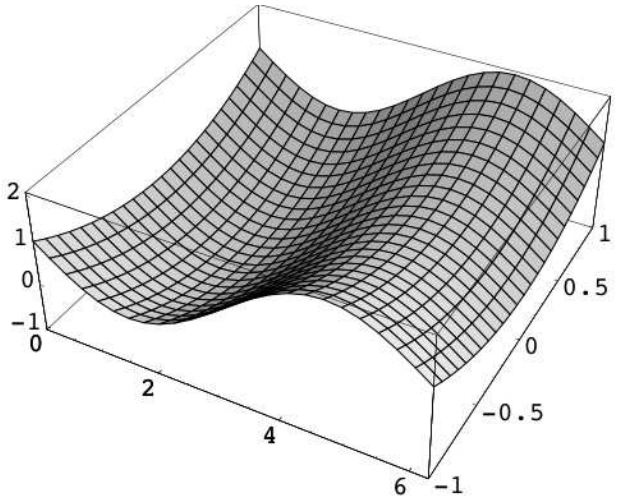


Figura 2: La supuesta gráfica de la función $z = y^2 + \operatorname{sen}(23x)$ con $x \in [0, 2\pi]$, $y \in [-1, 1]$, según Mathematica 5.

funcionar bastante bien, y ha ido mejorando con las versiones. Por ejemplo, el libro [5, § 4.2.2] nos muestra que Mathematica 1.2 y 2.0 dibujaban

```
Plot[x + Sin[2*Pi*x], {x, 0, 24}]
```

exactamente igual que la recta $y = x$. Internamente, para hacer ese gráfico, Mathematica tomaba los 25 puntos $(x_j, x_j + \operatorname{sen}(2\pi x_j))$, $x_j = 0, 1, 2, \ldots, 24$, y los unía; y da la casualidad de que todos ellos están sobre la recta $y = x$. Se puede solucionar el problema dando un número distinto de puntos de muestreo, aunque para eso tenemos que sospechar que algo falla. Funciona tanto asignar un número mayor (`PlotPoints -> 50`) como menor (`PlotPoints -> 20`). Pero los métodos de muestreo adaptativo de las versiones posteriores han mejorado mucho, y no hemos observado dibujos erróneos de ninguna función $y = f(x)$ razonable.

Lamentablemente, estos algoritmos adaptativos son bastante más difíciles de usar para funciones $z = f(x, y)$; y, sobre todo, consumirían mucho tiempo de cálculo. Así que los desarrolladores de Mathematica han optado por no aplicarlos. Como consecuencia, si la función que representamos tiene oscilaciones, existe la posibilidad de que obtengamos un dibujo que no se corresponde con la realidad. Así, por ejemplo, Mathematica 5.0 y 5.1 representan

```
Plot3D[y^2 + Sin[23*x], {x, 0, 2Pi}, {y, -1, 1}]
```

tal como aparece en la figura 2. La suavidad del dibujo no nos hace pensar que sea erróneo; pero si reflexionamos nos damos cuenta de que la parte `Sin[23*x]` debe forzosamente producir una oscilación que se ha perdido (y que se puede ver

añadiendo un `PlotPoints -> 50`). Ha dado la casualidad de que los puntos de muestreo que ha usado Mathematica han estado todos ellos sobre una superficie bastante suave, y ésa es la que ha dibujado. En versiones anteriores (la 4.2, por ejemplo), Mathematica usaba un número distinto de puntos de muestreo, y el problema se observaba poniendo `13` o `15` en el lugar del `23`.

Descubramos el truco: este fallo lo hemos encontrado adrede. Nuestro conocimiento de lo que son capaces de hacer este tipo de programas nos hacía pensar que se tenía que producir. Pero sólo sucede cuando los puntos de muestreo están en determinados lugares estratégicos. La mejor manera de ayudar a la casualidad a que esto ocurra es usar un bucle:

```
For[k=1, k<=30, k++,
  Plot3D[Sin[k*x], {x, 0, 2Pi}, {y, -1, 1}]
]
```

Así es fácil encontrar un dibujo ficticio (en el ejemplo de la figura 2, el sumando y^2 no tiene importancia, pero hace que la función tenga realmente dos variables).

Como contrapartida, aprendemos una lección: cuando hagamos una gráfica, ante la menor sospecha conviene variar el número de `PlotPoints` asignados por defecto, y probar con varios valores. De este modo, podremos detectar una posible cancelación casual de la oscilación. No olvidemos que, aunque hemos confesado que esta vez el fallo lo hemos buscado aposta, puede producirse en un caso real que no nos resulte tan evidente.

3 Simplificar tiene su miga

Cuando le pedimos que simplifique una expresión, Mathematica es muy cuidadoso. Veámoslo por medio de un par de ejemplos. Si no lo pensamos demasiado, es habitual que nosotros mismos simplifiquemos $\operatorname{arc\,sen}(\operatorname{sen}(x)) = x$; pero esto no es cierto en general, sólo cuando $x \in [-\pi/2, \pi/2]$. Mathematica no pasa esto por alto, de tal manera que `Simplify[ArcSin[Sin[x]]]` devuelve $\operatorname{arc\,sen}(\operatorname{sen}(x))$, no x. Para conseguir que realmente simplifique, debemos especificar

```
Simplify[ArcSin[Sin[x]], -Pi/2 <= x <= Pi/2]
```

De manera similar, la orden `Simplify` tampoco hace nada si se la aplicamos a `Sin[x + 2*n*Pi]` (muchos matemáticos estamos tan acostumbrados a que n denote números enteros que damos por hecho que siempre es así, y posiblemente hubiéramos afirmado que $\operatorname{sen}(x + 2n\pi) = \operatorname{sen}(x)$ sin pestañear; pero Mathematica es prudente, y no cae en esa trampa). Para que simplifique, hay que indicarle que que n es un entero, lo que se consigue, por ejemplo, así:

```
Simplify[Sin[x + 2*n*Pi], Element[n, Integers]]
```

Además, es muy importante tener en cuenta que no siempre está claro qué queremos decir con «simplificar». ¿Qué es más simple, $(x^{100} - 1)/(x - 1)$ o los 100 sumandos que aparecen al dividir? ¿Y $\cos^2(x)$ o $(1 + \cos(2x))/2$?

Pues depende de para qué lo queramos, y eso no lo puede saber la máquina. En consecuencia, todos los programas de cálculo simbólico tienen bastantes comandos destinados a efectuar muy diversas manipulaciones. No importa aquí cuáles son esos comandos, pero sí que conviene caer en la cuenta de que, si queremos que los cálculos que estemos efectuando avancen en la dirección que nos interesa, no nos queda otro remedio que conocerlos; no basta con servirse de una sola orden como `Simplify`.

Por otra parte, simplificar puede suponer muchísimo tiempo de cálculo. Para simplificar una expresión, Mathematica intenta aplicar lo que denomina reglas de transformación. Tiene que ver cuáles, de las muchas que tiene almacenadas, se ajustan a la expresión en cuestión, y adónde llevan. Es habitual que, ante una expresión medianamente complicada, haya multitud de caminos que se pueden explorar, y muy pocos conducen a buen puerto. Así, Mathematica dispone de dos órdenes, `Simplify` y `FullSimplify`. La segunda permite que el tiempo empleado sea mayor, y usa más reglas de transformación. Aunque, en nuestra opinión, hay veces que es incomprensible que el comando básico no se dé cuenta de algunas cosas que a cualquier matemático le parecen obvias. Por ejemplo, `Simplify` no se percata de que $\log(8)/\log(2)$ vale 3; si queremos conseguir esa simplificación, hay que hacerlo con `FullSimplify[Log[8]/Log[2]]`.

A veces, ni siquiera `FullSimplify` hace un trabajo que pudiéramos considerar satisfactorio. Es evidente que $\sum_{k=1}^{n}\sqrt{k}-\sum_{k=1}^{n-1}\sqrt{k}=\sqrt{n}$; sin embargo,

```
FullSimplify[
  Sum[Sqrt[k], {k, 1, n}] - Sum[Sqrt[k], {k, 1, n-1}]
]
```

no nos proporciona la esperada respuesta $\sqrt{n}$. Ni siquiera si le indicamos que n es un entero positivo, no vayamos a pensar que ése es el problema. Hay quien diría, en defensa de Mathematica, que el usuario puede añadir sus propias reglas de transformación, y de ese modo podríamos conseguir que simplificara, por ejemplo, la expresión anterior. No es tarea fácil y, además, ¡para ese viaje no hacían falta esas alforjas!

Otro ilustrativo ejemplo con el que las rutinas de simplificación no saben enfrentarse es

```
FullSimplify[12345678901!/12345678900!]
```

La respuesta debería haber sido 12345678901, pero Mathematica devuelve ese cociente sin simplificar. Y eso que sí que evalúa `FullSimplify[(n+1)!/n!]` como $n+1$. Vale la pena analizar lo que está pasando. Mathematica puede calcular factoriales, y eso intenta; si sabe calcularlo tal como está, ¿para qué simplificar?, debe pensar. Pero se da cuenta de que esos factoriales son realmente demasiado grandes (imagínese el lector cuantos dígitos podría tener 12345678901!), así que no los hace. Y ahí se acabó la historia; se percata de que la primera estrategia no vale, pero no usa la alternativa: simplificar antes de operar. ¿Cuántas veces le habremos dicho a nuestros alumnos que es mejor simplificar antes de ponerse a «echar cuentas» como locos?

4 Un par de simplificaciones erróneas

Pongámonos de nuevo del lado de Mathematica, y volvamos a mostrar ejemplos de la cautela que emplea al simplificar. Así, `Simplify[Sqrt[x^2]]` no devuelve x, pues se da cuenta de que x puede ser un número real negativo. Tampoco con `Simplify[Sqrt[x^4]]` obtenemos x^2 como respuesta, que no sería cierta si x es un complejo. Si no es éste el caso y queremos conseguir que simplifique, debemos indicárselo expresamente. A decir verdad, los investigadores que trabajan en temas relacionados con el cálculo simbólico son muy conscientes de las dificultades inherentes a la aritmética de números complejos (véase [6]), y procuran tenerlas en cuenta.

Esto no impide que, de vez en cuando, pueda surgir el habitual desliz. Aunque aquí no nos importa cómo están definidas, digamos que hay ciertas funciones que se conocen con el nombre de integrales elípticas (si al lector le ha picado la curiosidad de qué es eso, puede ojear [1, capítulo 17]; también hay algo de información en la propia ayuda de Mathematica, o en su manual [15]). Nosotros vamos ahora a usar dos de ellas, $F(\phi, m)$ y $K(m)$, que Mathematica representa con `EllipticF` y `EllipticK`. Pese a que la nota explicativa «qué es nuevo en la versión 5.1» no dice nada sobre ello, parece que esa versión ha añadido algunas rutinas para el tratamiento de estas funciones. Así, por ejemplo, Mathematica 5.1 afirma que

```
FullSimplify[EllipticF[I*Log[1 + Sqrt[2]], -1]]
```

(donde `I` es la unidad imaginaria) vale $-iK(-1)$, mientras que las versiones anteriores no transforman la función F en la K. Además, Mathematica (todas las versiones) sabe que `FunctionExpand[EllipticK[-1]]` vale $\Gamma(1/4)^2/(4\sqrt{2\pi}) \approx 1{,}31103$ (`FunctionExpand` es uno de los diversos comandos dedicados a simplificar a los que antes aludíamos). Como consecuencia de todo esto, Mathematica 5.1, ante la evaluación numérica

```
FullSimplify[EllipticF[I*Log[1 + Sqrt[2]], -1]] // N
```

nos proporciona el resultado $-1{,}31103i$. Por el contrario, si no hubiéramos simplificado, lo que obtenemos es que

```
EllipticF[I*Log[1 + Sqrt[2]], -1] // N
```

vale $1{,}31103i$. ¿Qué es lo que está pasando? Pues que $F(i * \log(1 + \sqrt{2}), -1)$ no es $-iK(-1)$, sino $iK(-1)$. Aparentemente, al programador de Mathematica que estaba incluyendo las mejoras se le ha deslizado un signo «$-$» en alguna línea de código. Parece fácil de arreglar en futuras versiones. Veremos...

Finalicemos esta sección analizando un error muy «original». Comencemos recordando que la serie $\zeta(s) = \sum_{n=1}^{\infty} 1/n^s$ es convergente cuando $s > 1$. Hay expresiones exactas de $\zeta(2k)$, con $k = 1, 2, 3, \ldots$, y Mathematica nos las proporciona con órdenes del tipo `Zeta[2k]`; obsérvese que, tal como se deduce de la figura 3, el valor de $\zeta(2)$ es realmente muy conocido.[12] Pero no hay

[12] Existen numerosas demostraciones diferentes de que $\sum_{n=1}^{\infty} 1/n^2 = \pi^2/6$ (véase [7], donde se muestran catorce), aunque ninguna de ellas cabía en la pared fotografiada.

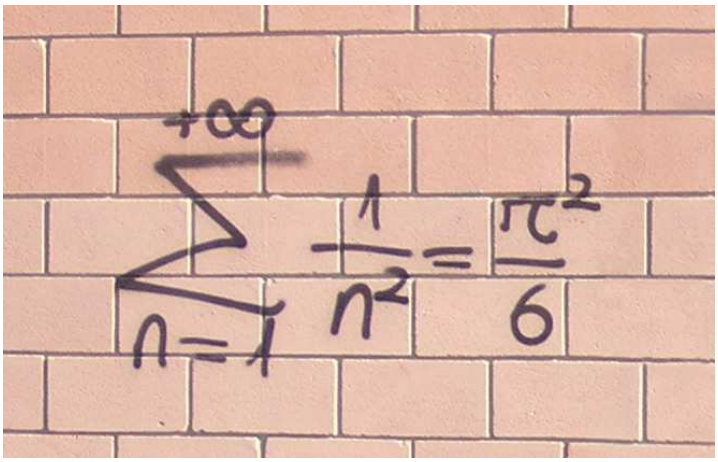


Figura 3: Fotografía tomada en el trayecto de un tren de cercanías valenciano.

nada parecido para $\zeta(2k+1)$; en consecuencia, Mathematica deja `Zeta[3]` —o similares— sin evaluar.[13] Menos habitual es la función $\mathrm{Li}_k(z) = \sum_{n=1}^{\infty} z^n/n^k$, que Mathematica denota como `PolyLog[k,z]`.

Centrándonos en el caso $k = 2$, y si tomamos un complejo z tal que $|z| = 1$, la serie $\sum_{n=1}^{\infty} z^n/n^2$ es absolutamente convergente, ya que está mayorada por $\sum_{n=1}^{\infty} 1/n^2$. Elijamos un complejo de módulo 1, como $z = e^i$. Mathematica no conoce ninguna expresión sencilla para `PolyLog[2,E^I]` (no la hay, hasta donde nosotros sabemos), así que lo devuelve sin evaluar; y lo mismo sucede si se lo pedimos junto con `Simplify`, `FullSimplify` o `FunctionExpand`. Eso sí, si le indicamos que nos dé su valor numérico aproximado lo hace, y nos dice que es $0{,}324138+1{,}01396i$. Hasta aquí, todo correcto. Lo sorprendente llega cuando, de manera simbólica, utilizamos `PolyLog` y `FunctionExpand` junto con la función `Zeta` (en un argumento s en el que Mathematica no sepa evaluar $\zeta(s)$). Da lo mismo que sea en una suma, en dos componentes de un vector, etc. Así, por ejemplo, supongamos que ejecutamos la orden

```
FunctionExpand[{Zeta[3], PolyLog[2, E^I]}]
```

Como cabía esperar, la primera parte queda sin evaluar; pero en la segunda obtenemos una respuesta totalmente falsa: `ComplexInfinity` (ya hemos

[13] Para valores impares de la variable, es muy poco lo que se sabe: está demostrado que $\zeta(3)$ es irracional, pero se desconoce si es o no un número algebraico; de $\zeta(5)$, $\zeta(7)$, etc., ni siquiera se sabe si son o no números racionales. La función $\zeta(z)$ se denomina zeta de Riemann, y adquiere toda su importancia cuando se extiende al campo complejo. Sorprendentemente, tener información sobre las raíces complejas de $\zeta(z)$ resulta ser fundamental en teoría de números. En particular, se usa para probar resultados sobre la distribución de números primos. La hipótesis de Riemann afirma que todos los ceros «no triviales» de $\zeta(z)$ se encuentran sobre esta recta del plano complejo: $z = \frac{1}{2} + ti$, $t \in (-\infty, \infty)$. De momento, esto no deja de ser una conjetura. Hay un premio de un millón de dólares esperando a quien la demuestre o la refute.

comentado que la serie involucrada es convergente). ¿Por qué esta conjunción estropea las cosas, si por separado iban bien? ¡Quién sabe! Lo que es seguro es que muchos usuarios estarían deseando ver el código interno de Mathematica para averiguar a qué se debe tan llamativo fallo y poder repararlo. Lamentablemente, es imposible, pues sólo Wolfram Research tiene acceso a él. Abundando más en esto, diremos que este error no estaba presente en la versión 3.0 de Mathematica, sino que apareció con la versión 4, y ha sido divulgado en los foros de internet adecuados (aunque no tal como aquí lo hemos presentado); siendo así, ¿por qué no lo corrigen los únicos que pueden hacerlo?

5 Tratamiento de casos

En las secciones anteriores hemos comentado, en varias ocasiones, que Mathematica acostumbra a ser muy cuidadoso cuando tiene que simplificar. Misteriosamente, otras veces parece pasar por alto toda precaución. Todos los que están habituados a tratar con sistemas ortogonales —y a cualquier otro matemático le resultará muy sencillo comprobarlo— saben que, cuando n y m son enteros,

$$\int_0^\pi \cos(nt)\cos(mt)\,dt = \begin{cases} 0, & \text{si } n \neq m, \\ \pi/2, & \text{si } n = m \neq 0, \\ \pi, & \text{si } n = m = 0. \end{cases} \tag{1}$$

Sin embargo, Mathematica yerra clamorosamente al calcular esta integral. Las respuestas concretas dependen de la versión que estemos usando pero, en esencia, podemos afirmar que Mathematica no se da cuenta de que n y m pueden ser iguales; y si evaluamos $\int_0^\pi \cos^2(nt)\,dt$, pasa por alto la posibilidad $n = 0$. Ninguna versión calcula correctamente la integral propuesta en (1). Así, por ejemplo, Mathematica 5.0 y 5.1 evalúan

```
Integrate[Cos[n*t]*Cos[m*t], {t, 0, Pi},
  Assumptions -> {Element[{n, m}, Integers]}]
```

como 0; y si hacemos la correspondiente integral con $n = m$, obtenemos $\pi/2$. (Las versiones anteriores proporcionan respuestas más oscuras, pero también equivocadas.)

Mathematica tampoco anda muy fino cuando lo enfrentamos a algo tan cotidiano como calcular las soluciones de un sistema lineal. La orden `Solve`, destinada a resolver ecuaciones o sistemas, no tiene en cuenta los posibles valores de los parámetros. Así, ante el sistema lineal

```
Solve[{2x + a*y == 3, 4x + 8y == 6}, {x, y}]
```

sólo nos proporciona la solución $x = 3/2$, $y = 0$; no se percata de que, si $a = 4$, hay muchas más soluciones. Si le proponemos un sistema que, para algunos valores del parámetro, es incompatible, tampoco lo nota. Es más, todos estos fallos persisten si intentamos resolver estos sistemas por medio del comando

`LinearSolve`, que es específico para sistemas lineales. En realidad, se puede lograr que Mathematica analice adecuadamente estos problemas si, en lugar de `Solve`, usamos `Reduce`; pero, ¿quién se lo podía esperar? No parece razonable que los creadores de Mathematica esperen que el usuario se haya estudiado concienzudamente su voluminoso manual [15] antes de manejarlo.

6 Una pifia espectacular

Algunas veces, el patinazo es morrocotudo. Mathematica 5.0 afirma que

```
Integrate[(1 - Exp[-t])^2 * t^(-3/2), {t, 0, Infinity}]
```

vale $-2\sqrt{2\pi}$. Esto es claramente imposible, ya que el integrando es positivo; es más, si no lo hubiera sido, difícilmente hubiéramos sospechado que la respuesta es errónea, pues nada parece indicar que la integral tenga alguna complicación oculta (obsérvese que $\lim_{t\to 0}(1-e^{-t})^2 t^{-3/2} = 0$). Sí que podíamos habernos dado cuenta comparando con el valor aproximado de la integral calculado mediante un procedimiento numérico (lo que se consigue con `NIntegrate`), pues esto sí proporciona una respuesta satisfactoria.

Afortunadamente, parece que sólo ha sido un problema momentáneo. El resto de las versiones de Mathematica que hemos evaluado, tanto anteriores como posteriores, proporcionan el valor correcto

$$\int_0^\infty (1-e^{-t})^2 t^{-3/2}\,dt = (4-2\sqrt{2})\sqrt{\pi}.$$

Que la versión 5.1 haya arreglado el problema de la 5.0 es bueno, pero no nos hace muy felices. Es evidente que, cuando aparece una nueva versión de un programa, contiene rutinas nuevas; en lo que ahora nos concierne, las dedicadas a la integración. Aunque sin dar detalles, esto lo suelen indicar los creadores de Mathematica, pues —lógicamente— les gusta destacar las mejoras que introducen. Pero, ¿por qué no nos señalan igualmente los fallos que han corregido? Sólo podemos pensar una cosa: nos quieren hacer creer que sus programas no tienen fallos, inducirnos a confiar ciegamente en ellos. Este oscurantismo es, desde todo punto de vista, lamentable.

Analicemos ahora otro ejemplo; esta vez —todo hay que decirlo— no tan grave. De nuevo tiene como protagonista principal a Mathematica 5.0, que afirma que la integral

$$\int_0^\infty \frac{\log(t)}{\sqrt{t}(1+t)}\,dt \tag{2}$$

no converge en $(0,\infty)$. Bastan unos instantes de reflexión para convencernos de que eso no es cierto. Además, resulta sencillo calcular su valor; para ello, sin más que efectuar el cambio $t = 1/s$, es inmediato comprobar que

$$\int_0^1 \frac{\log(t)}{\sqrt{t}(1+t)}\,dt = -\int_1^\infty \frac{\log(s)}{\sqrt{s}(1+s)}\,ds,$$

de donde se sigue que (2) vale 0. Sorprendentemente, Mathematica 5.0 sí que sabe calcular la integral sobre $(0,1)$, y nos indica —correctamente— que su valor es `-4Catalan` (como el lector habrá podido adivinar, `Catalan` alude a cierta constante, cuya definición concreta y valor aproximado no tienen ahora importancia; pero, para satisfacer su curiosidad, aclaremos que la denominada constante de Catalan es $\sum_{n=0}^{\infty}(-1)^n(2n+1)^{-2} = 0{,}915966\ldots$). Mathematica 5.0 sí que logra calcular (2) si le indicamos que «preste atención» al punto $t = 1$, lo que se consigue con

```
Integrate[Log[t]/(Sqrt[t]*(1 + t)), {t, 0, 1, Infinity}]
```

En realidad, esto está pensado para los puntos singulares, y $t = 1$ no lo es, pero lo reseñamos pues esta misma idea puede ser de ayuda en otras ocasiones. Hasta ahora hemos relatado lo que hace la versión 5.0; ¿cómo se comportan otras versiones cuando las enfrentamos con (2)? Las anteriores calculan esa integral sin problemas, y con rapidez. También la versión 5.1 sabe evaluar (2) pero, misteriosamente, emplea mucho tiempo (medio minuto, en un ordenador fabricado en 2004). Esto nos hace pensar: ¿por qué, si los programadores han detectado un fallo que antes no estaba, no lo han arreglado hasta dejarlo tal como funcionaba cuando lo hacía bien? ¿Quizás lo han corregido por casualidad, sin darse cuenta? Misterios.

7 Despistes y complicaciones innecesarias al integrar

En general, si prescindimos de fallos como los que acabamos de relatar, Mathematica es muy potente en cuanto a integración simbólica se refiere. Podemos dar por sentado que es mejor que cualquier matemático experto, y que muchos libros de tablas de integración. Los ejemplos de la sección anterior hacían todos referencia a integrales definidas; vamos ahora a dar algunas muestras de su comportamiento frente a integrales indefinidas —cálculo de primitivas—. Antes de seguir, conviene que reflexionemos sobre la dificultad e imprecisión de esta tarea, siendo éste otro tema que preocupa a los teóricos del cálculo simbólico (véase [8]). Hay que darse cuenta de que no siempre está bien planteado el significado de $\int f(x)\,dx$ (`Integrate[f[x], x]`, con la sintaxis de Mathematica). Siendo rigurosos, deberíamos aclarar en qué rango de valores de la variable x pretendemos que sea válido el resultado. Dependiendo del dominio, serán unos u otros los cambios de variable que se puedan efectuar, por ejemplo. Usualmente, ni siquiera nos preocupamos por ello cuando calculamos integrales «a mano»: lo que obtenemos tendrá sentido en algún dominio, que seríamos capaces de precisar si fuera necesario, pero no solemos hacerlo. Además, diversas estrategias de abordar el problema pueden dar lugar a expresiones aparentemente muy distintas. Por último, tampoco se especifica en qué términos se desea expresar la integral: es obvio que $\int_a^x f(t)\,dt$ es una respuesta cierta, pero no muy útil. Afortunadamente, ni que decir tiene que el método para detectar posibles errores es ahora relativamente sencillo: ¡basta derivar lo que hemos obtenido al integrar!

Aunque no podemos saber exactamente cuál es su funcionamiento interno, lo razonable es pensar que Mathematica incorpora una amplia tabla de integrales

conocidas, y mecanismos de sustitución, cambio de variable, etc., para reducir otras integrales que le propongamos a las de la tabla.

Hay veces en las que, inexplicablemente, no logra hacerlo. Por ejemplo, todas las versiones de Mathematica que hemos usado (la última, recordemos, la 5.1) saben calcular

$$\int \frac{x^2 e^{\operatorname{arc\,tg}(x)}}{\sqrt{1+x^2}}\,dx = \frac{e^{\operatorname{arc\,tg}(x)}(-1+x-x^2+x^3)}{2\sqrt{1+x^2}};$$

sin embargo, devuelven

$$\int \frac{x^2 e^{-\operatorname{arc\,tg}(x)}}{\sqrt{1+x^2}}\,dx$$

sin evaluar. Mathematica no se da cuenta de que, con el cambio de variable $x=-t$, esta integral se transforma en la anterior. Bueno, no saber hacer algo no es una respuesta errónea. ¡Pero parece mentira!, con tanta potencia y tanto algoritmo interno.

En algunas ocasiones, y aunque Mathematica proporciona una respuesta correcta, lo hace por medio de una expresión más complicada que la que podría obtener —sin dificultad y de manera directa— cualquier matemático. Veamos algunos ejemplos. Esta vez, las respuestas no van a ser unánimes, sino que dependerán de la versión de Mathematica que estemos usando. Resulta curioso observar que, unas veces, son las versiones más antiguas las que proporcionan respuestas más naturales; otras, al contrario.

Para calcular

$$\int \frac{\operatorname{sen}(x)}{\sqrt{2}\cos(x)-\sqrt{\cos(x)}}\,dx,$$

en el denominador sacamos $\sqrt{\cos(x)}$ factor común, y efectuamos el cambio $u=\sqrt{2\cos(x)}-1$. Así,

$$\int \frac{\operatorname{sen}(x)\,dx}{\sqrt{2}\cos(x)-\sqrt{\cos(x)}} = -\sqrt{2}\int \frac{du}{u} = -\sqrt{2}\log\left(-1+\sqrt{2}\sqrt{\cos(x)}\right).$$

Mathematica 5.0 y 5.1 obtienen directamente este resultado, pero las versiones anteriores nos daban una primitiva mucho más enrevesada:

$$\int \frac{\operatorname{sen}(x)\,dx}{\sqrt{2}\cos(x)-\sqrt{\cos(x)}} = -\frac{\sqrt{2}\left(-1+\sqrt{2}\sqrt{\cos(x)}\right)\sqrt{\cos(x)}\log(-1+\sqrt{2}\sqrt{\cos(x)})}{-\sqrt{\cos(x)}+\sqrt{2}\cos(x)}.$$

A decir verdad, si empleamos `Simplify` también llegamos a la expresión sencilla. En todo caso, esto es esperanzador; parece que las nuevas versiones son un avance.

Pero veamos qué ocurre con otra integral trigonométrica:

$$\int \frac{\operatorname{sen}(x)\,dx}{\sqrt{2\cos(x)-(\cos(x))^2}}.$$

El radicando del denominador se puede escribir como $1-(1-\cos(x))^2$, y entonces hacer el cambio $u = 1 - \cos(x)$. De este modo, es claro que

$$\int \frac{\operatorname{sen}(x)\,dx}{\sqrt{2\cos(x) - (\cos(x))^2}} = \int \frac{du}{\sqrt{1-u^2}} = \operatorname{arc\,sen}(u) = \operatorname{arc\,sen}(1-\cos(x)).$$

Esta vez, son Mathematica 3.0, 4.1 y 4.2 los que proporcionan directamente la primitiva $\operatorname{arc\,sen}(1-\cos(x))$. Sin embargo, Mathematica 5.0 y 5.1 evalúan la integral como

$$\int \frac{\operatorname{sen}(x)\,dx}{\sqrt{2\cos(x) - (\cos(x))^2}} = -\frac{2\sqrt{-2+\cos(x)}\sqrt{\cos(x)}\log(\sqrt{-2+\cos(x)}+\sqrt{\cos(x)})}{\sqrt{-(-2+\cos(x))\cos(x)}}. \qquad (3)$$

En este ejemplo, simplificar tampoco ayuda; no hemos logrado recuperar la respuesta $\operatorname{arc\,sen}(1-\cos(x))$ que hemos calculado nosotros mismos. Así pues, no da la impresión de que las nuevas versiones sean siempre mejores, como sería deseable. Alguien podría argumentar que la primitiva (3) no es tan mala; que, simplemente, estamos hallando otra expresión, y que nunca está claro qué se quiere decir con «simplificar». Quien piense de manera tan benévola es que no se ha fijado en la fórmula que nos han proporcionado Mathematica 5.0 y 5.1: en ella aparece varias veces $\sqrt{-2+\cos(x)}$, ¡un radicando que siempre es negativo!

Somos matemáticos, y no hemos podido resistirnos a analizar un poco más qué estaba pasando. ¿Será la respuesta (3) correcta si la desarrollamos como números complejos?; ¿o quizás es, simplemente, un error? Siempre cabe la posibilidad, incluso, de que las varias expresiones complejas que, forzosamente, aparecen en (3), se cancelen unas con otras. La realidad es una mezcla de todo esto. Hemos logrado comprobar que, en un intervalo real centrado en $x = 0$, la función (3) se diferencia de $\operatorname{arc\,sen}(1-\cos(x))$ en $2i\log(1+i)$, siendo $i = \sqrt{-1}$. Así pues, rigurosamente hablando, la primitiva hallada por Mathematica 5.0 y 5.1 no es errónea, ya que se diferencia en una constante (¡aunque compleja!) de la hallada por nosotros (y por las versiones anteriores de Mathematica). Pero nadie podrá negar que $\operatorname{arc\,sen}(1-\cos(x))$ es una primitiva mucho más apropiada.

8 Tres ejemplos numéricos llamativos

En algunas ocasiones, Mathematica dispone de dos comandos similares para realizar la misma tarea; uno para hacerla de manera simbólica y otro de forma numérica. Así, por ejemplo, existen los comandos `Integrate` y `NIntegrate`, `Sum` y `NSum`, `Solve` y `NSolve`. En las «órdenes simbólicas» (entre las anteriores, las tres que no comienzan por `N`), si la expresión que estamos evaluando no contiene números decimales, Mathematica efectúa las operaciones de manera exacta (por ejemplo, muestra fracciones o raíces en lugar de su correspondiente valor numérico aproximado); en caso contrario, lleva a cabo aproximaciones

con un determinado número de decimales (pero no acude al correspondiente comando con `N`). Este tipo de órdenes puede dar lugar a problemas inesperados, como mostramos a continuación. Veremos que unos métodos son más apropiados que otros; pero, ¿cómo saberlo previamente si no tenemos suficiente dominio de la cuestión?

En primer lugar, observemos que, cuando t es negativo, $e^t = \sum_{k=0}^{\infty} t^k/k!$ es una serie alternada. El teorema de Leibniz asegura que el error que se comete al aproximar e^t por $\sum_{k=0}^{n} t^k/k!$ está acotado por $|t|^{n+1}/(n+1)!$. Así, si tomamos $t = -40$ y $n = 300$, debe ser

$$\sum_{k=0}^{300} \frac{t^k}{k!} \approx e^{-40} \approx 4{,}24835 \cdot 10^{-18},$$

con un error realmente despreciable (mucho menor que las cifras significativas que hemos mostrado). ¿Qué hace Mathematica? Pues depende de cómo lo usemos. Imaginemos que estamos intentado obtener el valor numérico de esa suma y que, ingenuamente, escribimos

```
t = -40.0; Sum[t^k/k!, {k, 0, 300}]
```

Al ejecutarlo, Mathematica nos proporciona una respuesta totalmente equivocada. El valor concreto que nos da puede depender de la versión que estemos usando;[14] por ejemplo, Mathematica 5.0 y 5.1 calculan esa suma como $-3{,}52833$, que no se parece al valor $4{,}24835 \cdot 10^{-18}$ que debería haber tenido. Podemos adivinar lo que ha pasado. En $\sum_{k=0}^{300}(-40)^k/k!$, los primeros sumandos son, en valor absoluto, bastante grandes (luego, el denominador crece mucho más rápido que el numerador, y por eso la serie converge). El número de cifras significativas que está manejando Mathematica es inferior al que necesitaría almacenar para que todas las cancelaciones entre términos de distinto signo no se perdieran. Eso hace que acumule un error que no es significativo frente al tamaño de los sumandos grandes, pero sí frente al resultado final, que es casi cero.

Como entendemos el funcionamiento de la máquina, logramos comprender qué está pasando, e incluso podíamos haber previsto que éste no era un buen método de cálculo. Mucho más provechoso hubiera sido haber tomado $t = -40$ (exacto, sin ser un número decimal), haber efectuado la suma (que hubiera salido una fracción enorme), y luego tomado su aproximación decimal, así:

```
t = -40; Sum[t^k/k!, {k, 0, 300}] // N
```

Eso ya proporciona sin problemas el valor $4{,}24835 \cdot 10^{-18}$. Pero —y perdónenos el lector por nuestra insistencia— esto requiere tener claro qué es lo que estamos haciendo; eso sólo se consigue con conocimientos matemáticos —en

[14] Como se trata de un proceso puramente numérico, que posiblemente descanse sobre el hardware y las rutinas de cálculo de la máquina de una manera bastante directa, no sería de extrañar que también dependiera del microprocesador y del sistema operativo del ordenador que hayamos empleado. Pero, a decir verdad, esto no lo hemos observado.

particular, de cálculo numérico y su interacción con los procesos internos de los ordenadores—. Como moraleja, conviene recordar que, cuando se pueda, suele ser conveniente operar de manera exacta, y al final hacer aproximaciones numéricas. Aunque, como a veces el fallo está en alguna rutina simbólica (¡ya hemos visto algunos ejemplos!), no está de más intentar siempre ambos caminos.

La orden `NSum` está especializada en llevar a cabo procesos numéricos. Así, uno podría esperar que, internamente, Mathematica tomara precauciones y que, como consecuencia,

```
t = -40.0; NSum[t^k/k!, {k, 0, 300}]
```

se iba a comportar mejor. Algo mejor sí que lo hace: se da cuenta de que quizás la suma no la ha calculado bien y, antes de proporcionarnos la supuesta suma, nos da un aviso que indica que el resultado puede ser incorrecto. Tras la advertencia, Mathematica 3.0 aventura que la suma vale 4261,26. Por el contrario, Mathematica 4.1 y 4.2 continúan calculando durante bastante rato; al final, nos cansamos de esperar y abortamos el proceso. Mathematica 5.0 y 5.1 también se atreven a dar un resultado, aunque disparatado: $4261{,}32 - 9{,}61518 \cdot 10^{-149} i$, con $i = \sqrt{-1}$. Lo mismo ocurre si, con `NSum`, ponemos `t = -40` en lugar de `t = -40.0`.

Vamos ahora a ver un ejemplo similar, pero con integrales en vez de con series. Como e^{-x^2} es despreciable para valores grandes de $|x|$, se tiene

$$\int_{-1000}^{1000} e^{-x^2}\,dx \approx \int_{-\infty}^{\infty} e^{-x^2}\,dx = \sqrt{\pi} \approx 1{,}77245.$$

Si intentamos evaluar numéricamente la expresión de la izquierda mediante

```
NIntegrate[Exp[-x^2], {x, -1000.0, 1000.0}]
```

de nuevo obtenemos un aviso, y luego aproximaciones desafortunadas. Tampoco es siempre satisfactorio lo que conseguimos si, manteniendo `-1000.0` y `1000.0`, usamos `Integrate` en vez de `NIntegrate`. Esta vez, hay más variación de resultados dependiendo de la versión usada. Comentemos algunos de ellos:

- Con `NIntegrate`, Mathematica 3.0, 4.1, 4.2 y 5.0 afirman que la integral vale $1{,}349346 \cdot 10^{-26}$. Eso sí, con la correspondiente advertencia de que posiblemente esté mal evaluada.
- Usando `Integrate`, Mathematica 5.0 calcula la integral correctamente.
- Tanto con `Integrate` como con `NIntegrate`, Mathematica 5.1 efectúa los cálculos de forma impecable (¡bravo por él!; esta vez ha sido el mejor).
- Resulta curioso observar que, con `Integrate`, Mathematica 3.0 es capaz de evaluar la integral; pero para ello emplea alrededor de cinco minutos (medido en un ordenador no muy moderno), y muestra el resultado con cientos de miles de decimales. Mathematica 4.1 hace lo mismo (y tarda unos 20 segundos en un ordenador de 2004). Tras una larga espera, no hemos tenido paciencia para ver a qué llegaba Mathematica 4.2, y hemos abortado los cálculos.

Como ocurría con las series, lo recomendable es hacer

```
Integrate[Exp[-x^2], {x, -1000, 1000}] // N
```

Pero no siempre es factible recurrir a esto. Si nos hubiéramos topado con una función que Mathematica no sabe cómo tratar de manera exacta, este método no hubiera servido de nada.

Una chapuza gorda de Mathematica 5.0 se pone de manifiesto ante la integral compleja

```
NIntegrate[1/(1+z^2), {z, 0, I}]
```

(recordemos que `I` denota la unidad imaginaria). Es claro que es divergente, pues en $z = \sqrt{-1}$ hay un polo. Y de ello se dan cuenta todas las versiones de Mathematica si hacemos la integral de manera simbólica (con `Integrate`). Al efectuarla numéricamente, casi todas las versiones se comportan de manera aceptable: proporcionan un valor —erróneo—, pero previamente advierten de que dudan de la respuesta. Salvo Mathematica 5.0, que sin pestañear afirma que esa integral vale $0{,}0 + 1{,}77949i$.

Por el contrario, no pensemos que, en cuanto a integración se refiere, Mathematica 5.0 sólo sirvió para empeorar cosas que funcionaban bien. Esta vez, veamos algo que ha arreglado. Las versiones de Mathematica hasta la 4.2 incluida afirmaban que

```
Integrate[ArcSec[z], {z, 0, 1}]
```

vale 0. Pero, al calcular esa integral con `NIntegrate`, encontraban el valor $1{,}5708i$ (obsérvese que $\sec(x) = 1/\cos(x) \notin (-1, 1)$ cuando x es un número real, luego el integrando de la expresión anterior no es una función real, sino compleja). Mathematica 5.0 y 5.1 sí que calculan la integral simbólica correctamente, y nos dan el valor exacto, $i\pi/2$, que se ajusta al obtenido numéricamente.

9 Cálculos numéricos inesperados

Hay veces que no está claro si la orden que estamos usando va a dar lugar a un comportamiento simbólico o numérico. Quizás se pueda considerar un fallo de diseño que esto no esté mejor explicitado (aunque siempre se pueden consultar las instrucciones, por supuesto).

Algunas veces, esto da lugar a situaciones que —simplemente— se pueden considerar un poco ridículas. Por ejemplo, imaginemos que queremos aproximar la tabla de datos `{{0, 0}, {1, 1}, {2, 4}}` por medio de una combinación lineal de las funciones 1, x y x^2. Esto se hace mediante la orden

```
Fit[{{0, 0}, {1, 1}, {2, 4}}, {1, x, x^2}, x]
```

Es obvio que, sea cual sea el tipo de aproximación que se pretenda efectuar —si leemos el manual o la ayuda del programa nos enteramos de que `Fit`

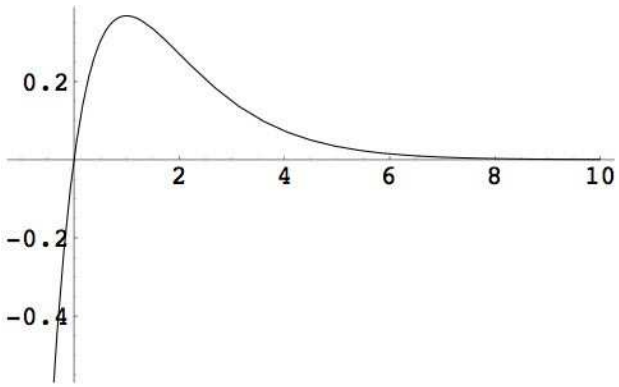


Figura 4: Gráfica de la función $y = xe^{-x}$.

efectúa aproximación por mínimos cuadrados—, la respuesta lógica es que la función que mejor lo hace es x^2. Sin embargo, Mathematica no nos proporciona esa función, sino algo similar a $2{,}52912 \cdot 10^{-16} + 3{,}97205 \cdot 10^{-16}x + 1{,}0x^2$ (ésa es la respuesta concreta de Mathematica 5.0 y 5.1; otras versiones nos dan coeficientes ligeramente diferentes). ¿Qué es lo que ha sucedido? Es fácil adivinarlo: Mathematica ha hecho un tratamiento numérico del problema y, como es habitual en todo proceso numérico, ha introducido pequeños errores de redondeo. La verdad es que no tiene demasiada importancia práctica (y, además, Mathematica dispone de la orden `Chop` para librarse de esos «presuntos ceros» en los coeficientes), pero no podemos negar que la situación resulta graciosa.

En otras ocasiones, los métodos numéricos pueden fallar por completo. Es claro que la función xe^{-x} sólo tiene una raíz: $x = 0$ (véase, además, la figura 4). Sin embargo, imaginemos que queremos encontrarla mediante el comando `FindRoot`. Tras leer las instrucciones, nos enteramos de que `FindRoot` utiliza un proceso iterativo para resolver ecuaciones, y que hay que proporcionarle un punto inicial (aunque escuetamente, esta vez Mathematica nos suministra algo de información de lo que va a hacer internamente: nos dice que «utiliza los métodos de Newton, de la secante o de Brent»). Si ejecutamos la orden

```
FindRoot[x*Exp[-x], {x, 3}]
```

tanto Mathematica 3.0 como 4.1 y 4.2 nos proporcionan la supuesta raíz $x = 16{,}9273$. A nuestro entender, podemos considerar que esto es un fallo garrafal. Posiblemente se haya producido porque no se ha utilizado una condición de parada adecuada. Ante un método iterativo que, para intentar resolver $f(x) = 0$, genera una serie de valores $\{x_n\}_{n=0}^{\infty}$, hay dos tipos de condiciones que se suelen usar: comprobar que, para ciertos valores de $\varepsilon, \varepsilon'$ suficientemente pequeños, se verifica $|f(x_n)| < \varepsilon$ (es decir, x_n es una raíz aproximada) y $|x_{n+1} - x_n| < \varepsilon'$ (x_n es, aproximadamente, un punto fijo del método que busca raíces). Si, sobre la figura 4, nos imaginamos cuál va a ser el comportamiento del proceso iterativo,

los valores de x_n que parten de $x_0 = 3$ van a ir creciendo, luego el método no va a poder encontrar ninguna raíz verdadera. Parece que Mathematica sólo ha comprobado si $|f(x_n)|$ es muy pequeño; esto seguro que ocurre alguna vez, pues xe^{-x} decrece muy rápido cuando $x \geq 3$. Es verdad que $x = 3$ es un mal punto de partida para encontrar una raíz de $xe^{-x} = 0$; pero eso no vale como excusa: ¿por qué —quizás en otro ejemplo menos obvio— el usuario debería saberlo? Puede ser razonable que el programa no encuentre una raíz, pero no que nos engañe.

Mathematica 5.0 y 5.1 tampoco hallan la raíz, e incluso aventuran una solución más alejada de la verdadera: $x = 106{,}795$. Aunque hay que reconocer que, esta vez, su comportamiento es aceptable, pues previamente avisan de que no han podido comprobar la convergencia del método.

10 Raíces de polinomios

Hace años, nos encontrábamos inmersos en una investigación teórica sobre ciertas propiedades asintóticas de los ceros de polinomios ortogonales sobre la circunferencia unidad del plano complejo (véase [2]). Habíamos demostrado un nuevo teorema, y queríamos ilustrarlo gráficamente con algunos ejemplos; para ello necesitábamos construir ciertos polinomios de grado alto (alrededor de 100, por ejemplo), y hallar sus raíces. En ese momento, Mathematica iba por la versión 3.0. Allí hicimos nuestros dibujos, que se ajustaban sin problemas a lo que nos decían los resultados teóricos. Es más; previamente, Mathematica nos había sido muy útil como herramienta experimental. Las numerosas variaciones que habíamos ensayado nos habían llevado a conjeturar cuál tenía que ser el resultado teórico que perseguíamos, para así saber qué teníamos que intentar probar.

Como todos los que nos dedicamos a la investigación matemática sabemos, los procesos de publicación en esta ciencia no son tan rápidos como en otras. Esto se juntó con otros avatares que no merece la pena recordar. El caso es que, entre la primera versión del artículo, la inclusión de los cambios sugeridos por los *referees*, y su aceptación definitiva para ser publicado, pasaron unos cuantos años. En ese tiempo, Mathematica había cambiado varias veces de versión. Es más, incluso los ordenadores que habíamos usado para hacer las gráficas estaban ya obsoletos, y teníamos otros mucho más rápidos.

Quisimos hacer alguna modificación sin importancia en las gráficas (por ejemplo, eliminar el color y cambiarlo por tonos de gris, pues las revistas matemáticas no suelen publicar en colores). El caso es que, tras efectuar estas pequeñas modificaciones, y conservando toda la parte esencial del código en Mathematica que habíamos usado previamente, repetimos los cálculos con los nuevos ordenadores y sus versiones actualizadas de Mathematica. ¡Qué desastre! El código se ejecutaba sin problemas aparentes, pero los dibujos que obteníamos eran muy distintos. Ya no eran útiles para ilustrar los teoremas. ¿Qué estaba pasando?; ¿qué funcionaba mal? Quizás nuestros ejemplos no servían para nada; cuando nos topamos por primera vez con este problema éramos más ingenuos

que ahora, y pensábamos que las respuestas proporcionadas por las versiones nuevas debían ser las más ajustadas a la realidad.

Nuestros programas tenían cierta componente numérica, quizás propensa a errores —tal como ya hemos comentado varias veces, desconocemos qué algoritmos ha usado Mathematica internamente, luego no podemos saber cuál va a ser su comportamiento—. Sin embargo, al ejecutarlos con Mathematica 3.0, producían resultados razonables, y eran acordes con lo que los teoremas que estábamos enunciando preveían. Nada hacía pensar que lo que obteníamos pudiera ser ficticio.

El caso es que nosotros estábamos dibujando configuraciones de ceros de polinomios ortogonales sobre la circunferencia unidad. Una importante propiedad de estos ceros es que siempre deben estar dentro de la circunferencia. En seguida nos dimos cuenta de que lo que estaban haciendo las versiones nuevas de Mathematica tenía que ser forzosamente erróneo, pues estaban encontrando supuestas raíces en el exterior de la circunferencia. Nos parecía mentira, pero algo que funcionaba bien, ahora ya no lo hacía. Eso sí, todo era mucho más rápido; no sólo por el cambio de ordenador, sino que una gran parte de la mejora de velocidad era manifiestamente imputable a los algoritmos que estaba usando Mathematica. Cada vez estaba más claro. Las nuevas versiones de Mathematica anunciaban espectaculares mejores de rendimiento. Lo que no advertían es que, al parecer, lo habían logrado a costa de utilizar algoritmos mucho más inestables.

No vamos a repetir ahora los cálculos que hacíamos entonces, y que tenían cierta sofisticación que no tiene aquí ningún interés. En su lugar, hemos buscado cómo aislar el problema, mostrando el tipo de comportamiento que hemos relatado en un ejemplo sencillo. Lo hemos logrado de una manera que —quizás— puede resultar incluso sorprendente: vamos a pedir a Mathematica que, para diversos valores concretos de n, halle todas las raíces del polinomio $z^n - 1 = 0$. Sabemos que las raíces son $e^{2k\pi i/n} = \cos(2k\pi/n) + i\,\mathrm{sen}(2k\pi/n)$, $k = 0, 1, \ldots, n-1$, pero le vamos a proponer a Mathematica que las calcule mediante un algoritmo numérico. Obviamente, esto podía haber sido resuelto de manera simbólica; pero no hubiera sido así con polinomios más complicados, como ocurría con los que teníamos cuando nos topamos con el problema. Además, nuestro principal interés radica ahora en ver si las nuevas versiones lo hacen mejor o peor (reiteramos que los algoritmos numéricos presentan muchas dificultades, así que quizás podríamos haber aceptado que, para n grande, Mathematica nunca hubiera sido capaz de encontrar las raíces).

Así pues, veamos cómo calculan las raíces de $z^n - 1$ las diversas versiones de Mathematica que estamos analizando, siempre de manera numérica. Para ello, tras asignar un valor a n, le indicamos a Mathematica que ejecute la orden

```
NSolve[z^n - 1 == 0, z]
```

Tras ello, representamos las raíces (en el plano complejo) y la circunferencia unidad. Si el cálculo estuviera bien hecho, deberíamos obtener n puntos ($e^{2k\pi i/n}$, con $k = 0, 1, \ldots, n-1$) regularmente distribuidos sobre la circunferencia. Hemos hecho bastantes experimentos, con n tomando valores hasta 500. Hay tanta

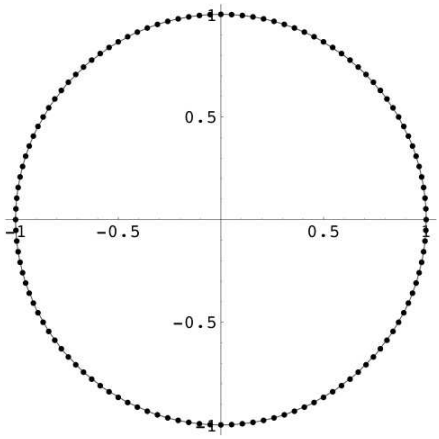


Figura 5: Las raíces de $z^{120} - 1 = 0$, según Mathematica 3.0 (y 4.1).

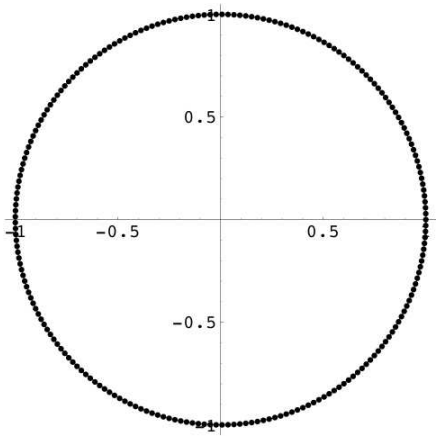


Figura 6: Las raíces de $z^{217} - 1 = 0$, según Mathematica 3.0 (y 4.1).

casuística que es difícil resumir las respuestas más significativos que aparecen, pero vamos a intentarlo.

Con Mathematica 3.0 siempre encontramos resultados aceptables, del estilo que podemos ver en las figuras 5 y 6. Y los tiempos de cálculo son siempre breves. Al menos en lo que nosotros hemos observado, Mathematica 4.1 halla las mismas gráficas, pero para ello emplea un tiempo decenas de veces mayor.

A partir de $n = 75$, Mathematica 4.2 comienza a dar muestras de pérdida de precisión al hallar las raíces. Así, por ejemplo, representa las raíces de $z^{120} - 1 = 0$ como en la figura 7, y las de $z^{217} - 1 = 0$ como en la figura 8; eso sí, es muy rápido. Misteriosamente, y según vamos aumentando el valor de n, de repente nos topamos con algunos n para los que Mathematica emplea mucho más tiempo de cálculo del que nos tenía acostumbrado (típicamente, 100 veces mayor), y nos proporciona un dibujo impecable. Así ocurre, por ejemplo con $n = 219$ (pese a que 218 y 220 sí que daban problemas). Éste también es el caso con $n = 250$, 400 y 500; pero el gráfico correspondiente a 300 es erróneo (y lo obtiene muy rápidamente). Da la impresión de que Mathematica está usando, al menos, dos algoritmos diferentes; uno rápido pero inestable, y otro lento y preciso; desconocemos el motivo que le induce a emplear uno u otro.

Al principio, Mathematica 5.0 parece comportarse de manera similar a como lo hacía la versión 4.2; aunque, cuando falla al encontrar las raíces, no obtiene exactamente lo mismo. Pero, a partir de $n = 219$, parece que siempre obtiene dibujos acordes a lo que se esperaba (al menos, en los experimentos que hemos hecho no hemos encontrado ninguno incorrecto); por ejemplo, dibuja bien los gráficos correspondientes a $n = 220$ y 300, con los que Mathematica 4.2 tenía problemas.

Mathematica 5.1 de nuevo empeora las cosas. Sus dificultades también comienzan a observarse con valores de n próximos a 75. A partir de ahí, muchos

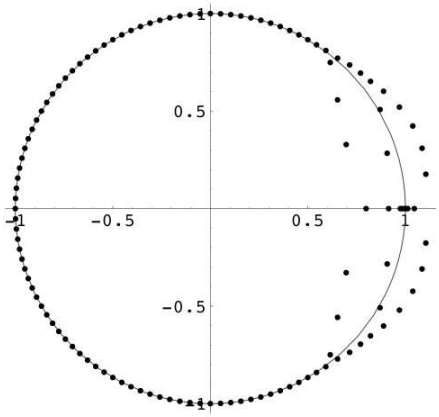


Figura 7: Las raíces de $z^{120} - 1 = 0$, según Mathematica 4.2.

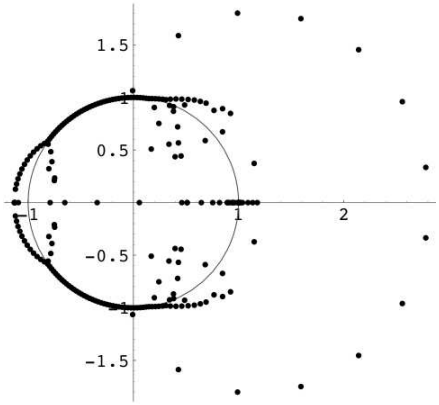


Figura 8: Las raíces de $z^{217} - 1 = 0$, según Mathematica 4.2.

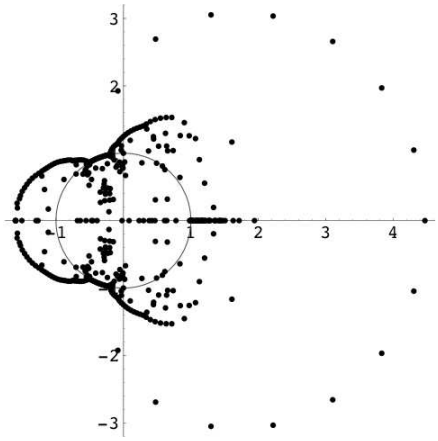


Figura 9: Las raíces de $z^{400} - 1 = 0$, según Mathematica 5.1.

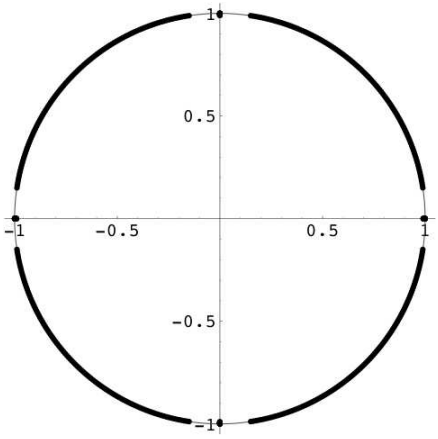


Figura 10: Las raíces de $z^{500} - 1 = 0$, según Mathematica 5.1.

de los dibujos que obtiene son similares a los de las versiones 4.2 y 5.0; aunque, al contrario que estas, logra dibujar correctamente el correspondiente a $n = 200$. Así hasta llegar al enigmático $n = 219$ que, como es habitual, y tras tomarse su tiempo, representa correctamente. Luego, parece no volver a acertar. Como ejemplo, mostramos lo que obtiene en relación con $n = 400$ (figura 9) y 500 (figura 10); ambos eran tratados correctamente por Mathematica 4.2 y 5.0 (y también por las versiones 3.0 y 4.1). Vale la pena comentar que, por el tiempo empleado, parece que con $n = 500$ ha usado el supuesto «algoritmo lento»; y que el tipo de dibujo que se obtiene —en el que no aparecen los ceros cercanos a 1, i, -1 y $-i$ (o están concentrados justo sobre esos puntos, a simple vista no se puede asegurar qué ocurre)— no lo hemos observado en ninguna otra versión de Mathematica.

Curioso, ¿no? Concluyamos la sección comentando que en los ejemplos que estábamos usando en [2] aún obteníamos más variedad de resultados, que no hemos logrado reproducir con polinomios de tipo $z^n - 1$:

- Aparecían ejemplos en los que Mathematica 3.0 no tenía dificultades, pero Mathematica 4.1 fallaba.

- A veces, Mathematica 4.1 nos daba mensajes de error incomprensibles, y encontraba muchos menos ceros que el grado del polinomio.

- Surgían circunstancias bajo las que Mathematica 5.0 se comportaba peor que 5.1, algo que no hemos observado con $z^n - 1$.

- En algunos polinomios con los que otras versiones no tenían dificultades, Mathematica 5.1 concentraba —inexplicablemente— todos los ceros en el origen.

Y todo este caos, sin explicación alguna. ¿Cómo podemos fiarnos? Es más, ¿cómo puede, una mente tan racional como la de un matemático, no pensar que nos están tomando el pelo? Para no ser tan negativos, comentemos que, tras ardua búsqueda, descubrimos un truco que parece útil (y que posiblemente pueda resultar provechoso en numerosas ocasiones). La idea es forzar a Mathematica a que, al resolver las ecuaciones, use aritmética de mayor precisión que la que emplea por defecto. Por ejemplo, si queremos que utilice números con 128 dígitos, recurrimos a la sintaxis `NSolve[z^n - 1 == 0, z, 128]`. Por lo que hemos podido experimentar, da la impresión de que no hacen falta muchos dígitos, pero sí que es útil evitar que use la precisión que tiene por defecto (aunque esto produce una pérdida de rendimiento, pues la trasmisión de los cálculos al microprocesador ya no es tan directa). Bueno..., lo mejor es ser siempre cauteloso y, en consecuencia, recurrir a chequeos adicionales; los aquí descritos, o los que se le ocurran al lector ante el problema concreto que esté abordando.

11 Conclusión

Todo lo que hemos expuesto a lo largo de estas páginas han sido ejemplos académicos. No han ocasionado que un puente se derrumbe. A lo más, contratiempos al publicar un artículo de investigación, o al explicar algo en clase. ¿Pero qué hubiera pasado si uno de estos cálculos erróneos que hemos mostrado se hubiera cruzado con el cálculo de estructuras de una obra de ingeniería?

Las dificultades y consiguientes inexactitudes del cálculo numérico están asumidas desde hace tiempo. Cualquiera sabe que no se puede confiar a la ligera en un resultado numérico; hay que tomar precauciones. Pero el cálculo simbólico aún tiene poca tradición, y sus errores son inesperados. En muchas ocasiones, no hay ninguna razón que explique la aparición de resultados falsos. Son, simplemente, errores de programación; y, lamentablemente, existen. Si sabemos lo que estamos haciendo, es mucho más difícil que una respuesta errónea nos pase desapercibida. Los ordenadores no tienen malicia; si se equivocan, no saben disimular. Es habitual que las respuestas equivocadas carezcan de toda lógica para quien domina el tema.

Recomendamos al lector que sea precavido, pero no alarmista. Al fin y al cabo, todos podemos recordar el fallo informático más publicitado de los últimos tiempos: muchos programas sólo tenían en cuenta los dos últimos dígitos de cada año; al entrar el año 2000, éste se confundiría con el 1900, lo que podría ocasionar el mal funcionamiento de muchos sistemas informáticos. Se auguraban catástrofes pero, al final, no fue para tanto (aunque quizás debido a todo el dinero que se invirtió en prevenirlo).

Así pues, y pese a lo expuesto hasta ahora, queremos finalizar rompiendo una lanza a favor del uso de las potentes herramientas informáticas de las que ahora disponemos. No hay que explicar a nadie que los ordenadores han llegado a ser imprescindibles en todo tipo de trabajos técnicos y científicos. Sencillamente, muchas cosas no se podrían hacer sin su ayuda. Es más, ya nos hemos acostumbrado a usarlos en situaciones en las que no son estrictamente necesarios. No es dependencia; es que realmente son muy útiles.

Hasta tal punto es así que hace ya años que la informática ha venido contribuyendo a la matemática teórica. Queremos recordar que el famoso teorema de los cuatro colores —cada mapa se puede colorear, de modo que dos regiones adyacentes nunca estén pintadas igual, con sólo cuatro colores— fue demostrado, en 1977, con la ayuda imprescindible de un programa informático (véanse [3, 4]). Esto ocasionó una considerable controversia, pero no creemos que alguien piense ahora que tal prueba no es válida (aunque tampoco se puede considerar que sea una demostración sencilla y elegante, claro).

Sin llegar al extremo de pretender usar herramientas informáticas genéricas para demostrar teoremas —algo que quizás será más usual en el futuro—, lo que sí es habitual es que nos sirvamos de los ordenadores para efectuar cálculos de manera rápida y precisa. Incluso nos han proporcionado un laboratorio experimental, del que los matemáticos carecíamos. Los paquetes de cálculo simbólico están resultando ser muy útiles en todo esto; la ayuda que nos prestan es magnífica. Pero no son la panacea universal; hay que saber manejarlos y

conocer sus limitaciones, a menudo inherentes a las matemáticas involucradas en el proceso. Además, este tipo de programas son aún muy nuevos, y no están adecuadamente depurados. Cometen errores achacables exclusivamente a sus programadores. Sólo los adecuados conocimientos matemáticos del usuario pueden ayudarle a detectar y mitigar estos problemas. De momento, los sistemas de álgebra computacional cuentan con un amplio margen para mejorar, y estamos convencidos de que así ocurrirá en el futuro; las numerosas y muy competentes personas que trabajan en este campo se ocuparán de ello. Ojalá que su precio y su opacidad disminuyan en la misma medida. Mientras, no queda otro remedio que desconfiar, ¡y rascarse el bolsillo!

Agradecimientos

Nuestra gratitud al profesor Julio Rubio, de la Universidad de La Rioja. Sus valiosos comentarios y sugerencias han logrado, sin duda, mejorar este artículo. No obstante, si algún *bug* queda, es exclusivamente achacable a los autores.

Referencias

[1] M. Abramowitz y I. A. Stegun (editores), *Handbook of mathematical functions with formulas, graphs, and mathematical tables*, United States Government Printing Office, 1964. Reimpreso por Dover, 1972 (10.ª reimpresión).

[2] M. P. Alfaro, M. Bello, J. M. Montaner y J. L. Varona, Some asymptotic properties for orthogonal polynomials with respect to varying measures, *J. Approx. Theory* **135** (2005), 22–34.

[3] K. Appel y W. Haken, The solution of the four-color-map problem, *Sci. Amer.* **237** (1977), 108–121.

[4] K. Appel y W. Haken, The four color proof suffices, *Math. Intelligencer* **8** (1986), n.º 1, 10–20.

[5] N. Blachman, *Mathematica: Un enfoque práctico*, Ariel informática, 1993.

[6] R. Bradford, R. M. Corless, J. H. Davenport, D. J. Jeffrey y S. M. Watt, Reasoning about the elementary functions of complex analysis, *Ann. Math. Artif. Intell.* **36** (2002), 303–318.

[7] R. Chapman, Evaluating $\zeta(2)$, 2003. Disponible en `http://www.maths.ex.ac.uk/~rjc/rjc.html`.

[8] J. H. Davenport, The difficulties of definite integration. Conferencia plenaria en el *11th Symposium on the integration of symbolic computation and mechanized reasoning "Calculemus 2003"* (Roma, 2003). Disponible en `http://www-calfor.lip6.fr/~rr/Calculemus03/`.

[9] C. W. H. Lam, How reliable is a computer-based proof?, *Math. Intelligencer* **12** (1990), n.º 1, 8–12.

[10] T. R. Nicely, Enumeration to 10^{14} of the twin primes and Brun's constant, *Virginia J. Sci.* **46** (1995), 195–204.

[11] I. Peterson, *Error fatal: A la caza de fallos informáticos*, Alianza Editorial, 1999.

[12] J. L. Varona, Graphic and numerical comparison between iterative methods, *Math. Intelligencer* **24** (2002), n.º 1, 37–46.

[13] J. L. Varona, Representación gráfica de fractales mediante un programa de cálculo simbólico, *Gac. R. Soc. Mat. Esp.* **6** (2003), 213–230.

[14] M. Wester, A critique of the mathematical abilities of CA systems. Capítulo 3 de *Computer algebra systems: A practical guide* (M. Wester, editor), John Wiley & Sons, 1999. También disponible en `http://math.unm.edu/~wester/cas_review.html`.

[15] S. Wolfram, *The Mathematica book*, 5.ª ed., Wolfram Media, 2003.

Bol. Soc. Esp. Mat. Apl. nº 37(2006), 123–132

Matemáticas e industria: una perspectiva interdisciplinar*

Barbara Lee Keyfitz

Director of the Fields Institute for Research in Mathematical Sciences, Toronto, Canada
John and Rebecca Moores Professor of Mathematics at the University of Houston, Houston, Texas, USA.

Traducido por Mikel Lezaun [1]

"....the keen
Unpassioned beauty of a great machine."
Rupert Brooke

En un mundo en el que la calidad, y puede ser que incluso el futuro, de la vida en el planeta depende cada vez más de la tecnología, las oportunidades para los matemáticos de incidir en nuestro entorno son ilimitadas. El mundo se está dirigiendo a nosotros en busca de asesoramiento, y la actitud que adoptemos ante este asesoramiento va a influir en la cultura de nuestra disciplina, en el reconocimiento de nuestra labor, y en el estatus de nuestra profesión. Los matemáticos de todo el mundo están examinando el potencial de las conexiones con la industria. Este breve informe intenta mostrar el espíritu de un conjunto de actividades, advirtiendo de entrada que las iniciativas actuales son muy diversas y que un resumen no puede hacer justicia a todas ellas. Junto con las oportunidades vienen los riesgos, y la necesidad de liderazgo. Este trabajo debería ser leído como un alegato para definir un modo justo de recompensar los esfuerzos y evaluar los resultados, y mostrar a los colegas más jóvenes alguna orientación en la búsqueda de nuevas oportunidades.

Las siguientes secciones tratan de la naturaleza de la investigación industrial, de las diferentes culturas del entorno industrial y del académico, de algunas comparaciones entre diversos países, y de los desafíos a las actuales tradiciones académicas.

*Publicado en *Madrid Intelligencer for ICM 2006*. Volumen especial de Intelligencer publicado con ocasión del International Congress of Mathematicians 2006, editado por Fernando Chamizo y Adolfo Quirós, ambos de la Universidad Autónoma de Madrid, Springer 2006.

[1]M. Lezaun, Universidad del País Vasco-Euskal Herriko Unibertsitatea.

La naturaleza de la investigación industrial

En un interesante capítulo sobre la investigación industrial, el reciente informe de la National Academic of Sciences titulado *Facilitating Interdisciplinary Research* [NA] resalta los distintos fines de los esfuerzos que se realizan en la investigación académica, en la de los laboratorios gubernamentales y en la de la industria (disciplina motivada, centrada en las necesidades nacionales, o en las necesidades del mercado respectivamente) y sus diferentes culturas (individual, guiada por la curiosidad, "de abajo hacia arriba" en el caso de la académica; dirigida, con objetivos fijados, "de arriba hacia abajo" en laboratorios gubernamentales y en la industria). Estos dos distintos modos de proceder hacen que la experiencia de los matemáticos en la industria no resulte familiar a los matemáticos académicos. Los matemáticos que trabajan en la industria encuentran la investigación académica, con su focalización en la excelencia individual y su estructura de reconocimiento competitiva, tan intimidadora como encuentran los matemáticos académicos la cultura de la industria. Debemos reconocer que nuestras percepciones están teñidas por nuestra cultura.

"Matemáticas en la industria" generalmente significa una de estas dos cosas: el trabajo de matemáticos empleados en la industria, o los esfuerzos de matemáticos académicos para establecer relaciones con la industria. Publicaciones como el *SIAM Report on Mathematics in Industry* [SI] han resaltado el valor que los empleadores adjudican a la formación matemática. Los empleadores reconocen la capacidad de los matemáticos para abstraer, para analizar, y para buscar las herramientas apropiadas para resolver un problema. Cada vez más, los potenciales empleadores observan que una educación matemática es más que un conjunto de técnicas y que la satisfacción por resolver problemas, la cual lleva a muchos estudiantes a las matemáticas, puede ser un potente motor de productividad en el sector industrial.

En el mismo informe de SIAM, y también en muchas conversaciones informales, se pone énfasis en la importancia que tiene para un matemático que aspira a una carrera exitosa en la industria, o para un académico con aspiraciones a interaccionar con la industria, desarrollar habilidades en comunicación y entender el trabajo en equipo. El contenido de pura investigación de una tesis parece menos importante que la capacidad para explicarla a un profano. Un atributo añadido valorado por los empleadores industriales es la diversidad de conocimientos, o un interés en adquirir esa diversidad. *No toda la investigación interdisciplinar es industrial, pero toda investigación industrial es interdisciplinar*, y muchas de las consideraciones que se hacen para desarrollar la conexiones interdisciplinares siguen siendo válidas para quienes quieran conectar con la industria, o animen a sus estudiantes a que lo hagan. El estudio NAS, [NA], aunque no está centrado principalmente en las matemáticas, pone énfasis en algunos de los retos que plantea proporcionar a los estudiantes una mejor experiencia interdisciplinar, y advierte de algunos de los riesgos, como son el alargamiento del periodo de los estudios de posgrado y el retraso de la entrada de los estudiantes en la investigación.

Mientras que hay acuerdo en que una mejora de la capacidad de comunicación es beneficiosa para todos, puede que para el tutor de un estudiante de posgrado no esté claro a la hora de aconsejarle dónde trazar la línea de separación entre amplitud y profundidad. Para los matemáticos empleados en la industria, los beneficios de una mayor diversidad, de una mayor variedad de problemas y experiencias, lleva consigo le desventaja de una pérdida de identidad como matemáticos. Incluso asociaciones como SIAM e ICIAM, que tratan de representar a las matemáticas "industriales y aplicadas", encuentran que proporcionalmente pocos matemáticos empleados en la industria consideran beneficioso ser miembros de un grupo cuya función es producir revistas y conferencias de estilo académico. Se puede discutir si el beneficio principal de la comunicación entre el sector profesional y el académico corresponde al sector académico o al sector industrial, pero lo que está claro es que la falta de comunicación, cuando la sociedad se enfrenta a enormes retos tecnológicos de los cuales tenemos noticia todos los días, es un perjuicio para todos. Recientemente, SIAM ha estado experimentando con pequeñas conferencias organizadas de forma que las presentaciones de matemáticos empleados en la industria sean mayoritarias. Las actas de estas conferencias proporcionan una muestra interesante de perspectivas y problemas industriales, [MF].

El ECMI (European Consortium for Mathematics in Industry) juega un papel importante en el diseño de programas de posgrado, implementados cooperativamente en Europa, en matemáticas para la industria, [EC]. La página web de ECMI contiene un listado de programas, caracterizados por combinar análisis matemático, conocimientos computacionales que incluyen el uso de herramientas software, y experiencias de aprendizaje en una industria, los cuales sirven como modelos para un nuevo tipo de estudios de posgrado.

Educar estudiantes para trabajar en la industria resulta difícil debido a la gran variedad de recorridos y de tipos de profesión que se dan en la industria, al hecho de que pocos científicos industriales son contratados como "matemáticos" (o incluso como investigadores), y a los cambios en la investigación y desarrollo industrial. En Estados Unidos, los esfuerzos en investigación básica en laboratorios de investigación industrial como los de ATT, Boeing o IBM, que para toda una generación han sido la vanguardia de las matemáticas en la industria, han sufrido una gran reducción. Es cierto que nuevos actores, como Microsoft, se han apuntado al juego. Sin embargo, es difícil imaginar a una gran universidad abandonando tan rápidamente parte de su investigación. Los cambios en la investigación industrial pueden ser un reflejo de los cambios en la industria: la globalización de la industria química y manufacturera, la consolidación y concentración de la propiedad, y la emergencia de nuevos polos industriales y mercados en Europa del Este y en Asia. Al menos en Estados Unidos y en Canadá, no hay indicios de que la industria esté respondiendo al reto tecnológico contratando a más doctores que hace cinco o diez años. ¿Es que las industrias ya no perciben que este nivel de capacitación investigadora es una parte deseable de la formación matemática?

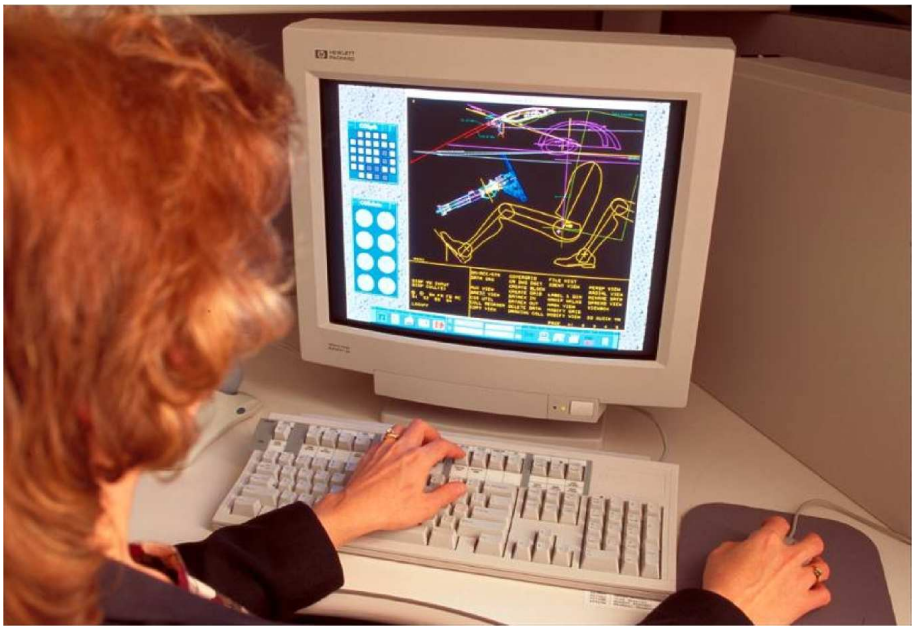


Figura 1: *La investigación en seguridad en General Motors incluye modelización matemática además de investigación experimental. Aquí son importantes las herramientas de geometría y de mecánica. Otras partes del análisis incluyen simulación, optimización y visualización. De un investigador que participe en un proyecto como éste se espera que domine muchos temas de matemática clásica y de matemática aplicada. Fotografía cortesía de General Motors Corporation*

La torre de marfil y el mundo real

Puede ser que haya una gran distancia incluso entre las contribuciones más valiosas hechas por matemáticos implicados en investigación industrial y los temas que realmente importan a los altos directivos de las grandes empresas. Quizás resulte tonto declarar que las matemáticas pueden "resolver" los problemas de la industria. Pero actualmente hay una sensación generalizada de que muchos aspectos de la investigación y del desarrollo industrial involucran en mayor o menor grado a las matemáticas. Una nueva generación de áreas comerciales, como la criptografía, la minería de datos y la matemática financiera, utilizan campos de la matemáticas - teoría de números, combinatoria, teoría de probabilidades - que tradicionalmente no eran vistas como parte de la matemática aplicada. Esto ha estimulado un nuevo respeto por la matemática aplicada como disciplina, y una conmoción en todo el espectro de las áreas de investigación en matemáticas en lo que respecta a los usos industriales de las mismas. James R.Schatz, jefe del grupo de investigación de matemáticas en la National Security Agency, afirma que "Nunca ha habido una época mejor que ésta para ser matemático". Algunos ejemplos de éxito económico atribuidos a las matemáticas están recogidos en *Unleashing Mathematics* [NE]: National Grid Transco (UK) estima en 400 millones de libras esterlinas el valor anual de la modelización y simulación, mientras que Arjo Wiggins, un fabricante internacional de papel de seguridad, obtuvo 15 millones de euros por la explotación de secuencias de de Brujin en matemática discreta. Ahora bien, ¿los matemáticos sirven al mundo del comercio e industria de la misma forma que los empleados en la entrada de datos y los servicios de mensajería? ¿Qué necesitan saber de negocios los científicos que trabajan en la industria para servir adecuadamente a la empresa, y cómo lo aprenderán?

Nuevas conexiones con la industria

En muchos países se está haciendo muy popular un nuevo modo de establecer contactos con la industria. Aunque en épocas anteriores algunos científicos han conseguido de forma individual acuerdos de asesoramiento con empresas, esos contratos no han tenido mucho impacto fuera de la experiencia inmediata del grupo participante, debido a que en el contrato habitualmente existen clausulas de confidencialidad y prohibición de publicación. Sin embargo, recientemente grupos de matemáticos han tenido éxito al suscribir contratos más abiertos por medio de *Grupos de estudio* o *Worshops de problemas industriales*. Los nombres de éstos varían y también sus detalles. Un departamento o un instituto matemático puede poner en marcha un seminario en el cual investigadores provenientes de la industria, generalmente no matemáticos, presentan problemas, y miembros académicos de la audiencia seleccionan aspectos de esos problemas para trabajar en ellos, habitualmente en un grupo que incluye estudiantes, a menudo en colaboración con el industrial que ha presentado el problema. El concepto de grupo de estudio empezó en Oxford,

se ha desarrollado por todo Europa, y en Estados Unidos fue adoptado por el Rensselaer Polytechnic Institute y ha sido extendido en diferentes formatos por el Institute for Mathematics and its Applications, y por otras muchas instituciones. La lista de ejemplos es demasiado larga para incluirla en este corto trabajo. Mediante acuerdo, el trabajo se hace en modo abierto, sin confidencialidad, y cualquiera que contribuya de forma significativa a su solución está autorizado a publicarla, de la misma forma que lo haría con una investigación académica. Un formato de workshop que se ha hecho muy popular radica en un esfuerzo concentrado, típicamente una semana, de un equipo organizado alrededor de un problema particular; los organizadores de grupos de estudio pueden recopilar media docena de problemas y ofrecérselos a grupos de entre 30 y 60 personas. El objetivo es obtener una solución al final de la semana. Como a menudo es el caso en la industria, el resultado no es un producto acabado, pero es lo mejor que pueden hacer los investigadores bajo las exigencias de un plazo de tiempo ajustado.

Por otro lado, esos ejercicios pueden servir como muestra del formato, de la idea, para los directivos e ingenieros de la industria que se muestren escépticos. Una vez que estén dispuestos a dedicar unos pocos días de su tiempo (o del de sus empleados), pueden llegar a convencerse de que la inversión en matemáticas puede contribuir a resolver un problema. La inversión por parte de la industria es pequeña, y si los resultados son buenos, puede dar lugar a una colaboración continuada, a acuerdos de asesoramiento, a prácticas de estudiantes en la empresa, o incluso a un trabajo permanente para un graduado en matemáticas. El arte de dirigir, de animar un workshop de este tipo se adquiere y transmite con la práctica y actualmente algunos departamentos los realizan regularmente. Cuando tienen éxito, estas iniciativas sirven para conseguir la valiosa finalidad de acercar grupos industriales y académicos, para que unos aprendan de los otros sus métodos y culturas, y para desarrollar un estilo que les posibilite trabajar juntos. Este tipo de ejercicios todavía está en desarrollo, y numerosas cuestiones están sin resolver: ¿Cómo medir el éxito de un grupo de estudio o de un workshop? ¿Cual es el lugar apropiado para publicar los resultados? ¿Cuál es el impacto en los estudiantes? ¿Qué tipos de departamentos son los que deben intentar poner en marcha un workshop y con qué frecuencia? La falta de un acuerdo general sobre cómo conceder reconocimiento al profesorado por la participación en seminarios industriales y worshops está afectando a algunos departamentos, de forma que les está resultando difícil involucrar a sus profesores, más preocupados por su carrera académica. Aún así, el número y la variedad de tales workshops continuan creciendo, como lo atestigua la lista que mantiene Oxford, [MI]. Por el momento no hay estándares sobre el rendimiento aceptados por toda la comunidad. A este respecto, el estatus de los workshops industriales puede ser algo similar a lo que es la enseñanza universitaria en departamentos con una investigación intensiva. Puede ser importante, interesante, incluso excitante, pero no es la vía para alcanzar una reputación nacional o internacional, salvo para una pocas "estrellas".

En Canada, la red MITACS está jugando un interesante papel en la promoción de las conexiones con la industria. MITACS (Mathematics of

Information Technology and Complex Structures) es una Network of Centres of Excellence (NCE) financiada por el gobierno federal, con un amplio presupuesto y una garantía de 14 años subvencionando a proyectos NCE. Originalmente se creó para hacer los departamentos de matemáticas más empresariales. Ahora MITACS opera de manera similar a un grupo de estudio a gran escala: se dan subvenciones a equipos de investigadores y estudiantes, típicamente de más de un departamento, los cuales trabajan en proyectos, ya sea con socios industriales o con el objetivo de comercializar los resultados de un proyecto de investigación. Como los grupos de estudio, o algunas organizaciones de investigación europeas como la Fraunhofer-Gesellschaft (que se ocupa de todas las ciencias e ingenierías, no sólo de las matemáticas) en Alemania o el Smith Institute en el Reino Unido, MITACS patrocina un conjunto diverso de actividades, algunas más comerciales que otras. Una novedad introducida por MITACS consiste en que lo típico es que los proyectos sean iniciados por investigadores académicos. Todo aquél que tenga una idea sobre cómo podría tener uso comercial su investigación es animado a ponerla en práctica. Canadá, a diferencia Estados Unidos y muchos países de Europa, no tiene un gran centro de investigación industrial, y la utilización de la energía de la investigación académica para estimular la investigación industrial ha encontrado una respuesta entusiasta. En los MITACS Interchanges se suelen exponer unos 100 posters.

Comparaciones globales

En todo el mundo, las relaciones entre matemáticas e industria son tan diversas como lo son los campos de las matemáticas y los ámbitos de las empresas. El grupo regional ECMI [EC] sirve para encauzar esas relaciones en Europa. En algunos países, actividades que en Estados Unidos pueden ser llevadas a cabo por la industria (como la producción de electricidad) son casi gubernamentales, y la investigación correspondiente puede ser realizada por el gobierno. Líneas arriba, citando el informe de la National Academy [NA], hemos dibujado una distinción entre la industria, que opera por presiones del mercado, y los laboratorios gubernamentales, organizados para satisfacer necesidades nacionales y sociales. Las tensiones entre las fuerzas del mercado y la regulación gubernamental del mercado de acuerdo con los intereses de un amplia capa social se manifiestan de diferente forma según los países, lo mismo que los conflictos por reclamaciones acerca de la propiedad intelectual. En algunos países, las universidades no están totalmente exentas de la necesidad de justificar los gastos de investigación por relevancia industrial y social, mientras en otros la investigación básica sigue estando de moda. En cualquier caso, en la batalla por mejorar las conexiones con la industria, los departamentos académicos deben tener cuidado en no olvidar su principal cometido: la enseñanza de conocimientos fundamentales y la formación en investigación básica. Mike Lizaridis, presidente de Research in Motion, que creó el *Blackberry*, habla apasionadamente a favor de que las universidades conserven su fortaleza en investigación fundamental. Lizaridis alerta contra la valoración de la investigación universitaria mediante criterios comerciales como

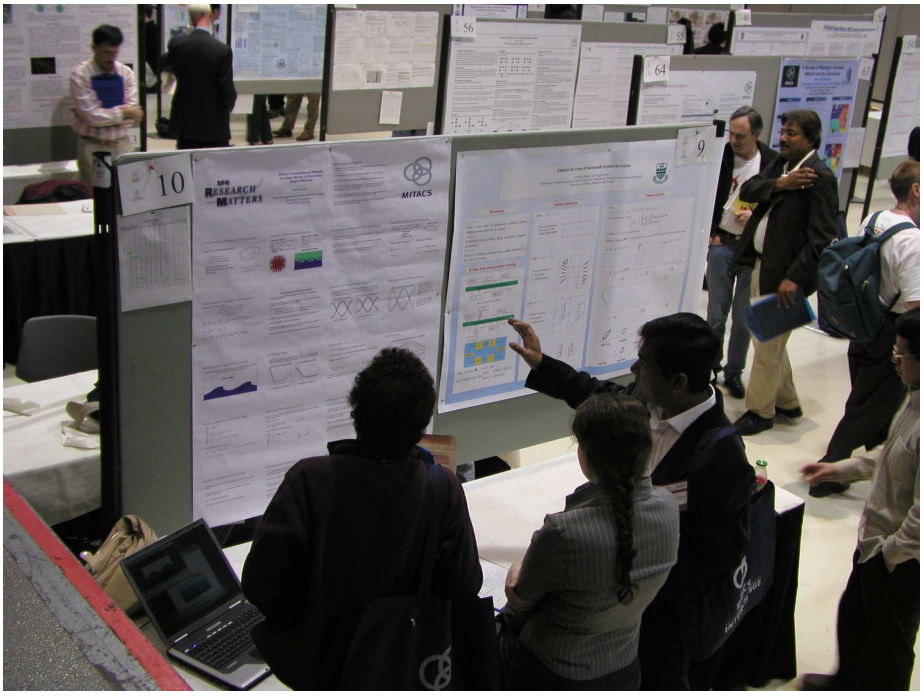


Figura 2: *Participantes en un reciente MITACS Interchange en una sesión de posters de los estudiantes. En los proyectos MITACS se pone especial énfasis en preparar a los estudiantes para entender las prioridades de los directivos de las empresas cuando se enfrenten a problemas de cierta envergadura. Los estudiantes trabajan codo con codo con los profesores y con ingenieros de la industria; también participan en workshops que desarrollan habilidades de presentación y comunicación. Fotografía cortesía de MITACS.*

el número de patentes generadas. En sus palabras "El sistema actual funciona".

Conclusiones

Las interacciones entre matemáticas e industria se están promocionando de numerosas formas: preparando estudiantes para ejercer en la industria (este artículo no tiene espacio para analizar el número creciente de programas de máster profesional en matemáticas, como un ejemplo de esta tendencia), atendiendo a las prioridades de los matemáticos que actualmente trabajan en la industria, y desarrollando avenidas para la interacción entre matemáticos académicos y profesionales (a menudo no matemáticos) de la industria. Actualmente los intentos surgen de centros espontáneos de interés entre individuos y pequeños grupos en departamentos e institutos matemáticos. Esos intentos son bienvenidos. Apelan al deseo de los matemáticos - y de los jóvenes en particular - de jugar un papel en la configuración del mundo en el que viven. El entusiasmo que interviene en la puesta en marcha de actividades y comunicaciones con personas ajenas a los departamentos académicos de matemáticas puede ayudar a invertir el decrecimiento gradual del interés por las matemáticas entre los estudiantes de facultades, los cuales se están dirigiendo hacia los negocios, la medicina y el derecho. Tenemos todos los motivos para recibir esas iniciativas como positivas.

El siguiente paso es encontrar respuestas adecuadas a las cuestiones que plantean estas nuevas aventuras. El estímulo de lo novedoso debe estar compensado con el apoyo al núcleo de la misión académica: la investigación fundamental y la educación de la siguiente generación de investigadores matemáticos. Los departamentos de matemáticas, las agencias financiadoras y las industrias deben desarrollar maneras para medir la calidad de las interacciones matemáticas-industria, deben determinar su valor y deben recompensar a los participantes. Los incentivos pecuniarios siempre son atractivos, pero también son necesarios el prestigio y una sensación de excelencia compartida. No es nada sorprendente que las diferentes culturas de las universidades y de la industria hayan dado origen a diferentes modos de evaluar y reconocer el mérito. Si está naciendo un nuevo paradigma que proclama beneficios para las dos, para la empresa industrial y para la académica, entonces ambas partes deberán encontrar una forma común de mostrar su reconocimiento. Será interesante observar este desarrollo.

Agradecimientos

Agradezco a Chandler Davis, de la University of Toronto, sus sugerencias al escribir este artículo, y a James Crowley, Executive Director of SIAM; David Field of General Motors; and Arvin Gupta, Scientific Director of MITACS, con los que he mantenido conversaciones muy valiosas, y que me han proporcionado fotografías y referencias.

La Sociedad Española de Matemática Aplicada (SEMA) agradece a la autora

y a Springer su autorización para publicar esta traducción.

Referencias

[BW] Business Week, January 23, 2006.

[EC] ECMI-European Consortium for Mathematics in Industry, http://www.ecmi-indmath.org/.

[MF] Mathematics for Industry: Challenges and Frontiers, edited by David R. Ferguson and Thomas J. Peters, SIAM, Philadelphia, 2005.

[MI] Mathematics in Industry Information Service, http://www.maths-in-industry.org/past/, Website maintained by OCIAM (Oxford Centre for Industrial and Applied Mathematics).

[NA] Facilitating Interdisciplinary Research, National Academies Press, Washington, 2005.

[NE] Unleashing Mathematics. Report of the NETIAM project, available from the Smith Institute at http://www.smithinst.ac.uk/Projects/NETIAM.

[SI] The SIAM Report on Mathematics in Industry, SIAM, Philadelphia, 1998.

[UA] Michael Smith, "Commercialization: the system works", University Affairs, February 2005.

Bol. Soc. Esp. Mat. Apl. nº 37(2006), 133–133

Breve nota sobre el ICM 2006 y la enseñanza de las matemáticas

S. Rodríguez Salazar

Departamento de Matemática Aplicada
Facultad de Ciencias Químicas
Universidad Complutense de Madrid
28040 MADRID, España

Soledad_Rodriguez@mat.ucm.es

Se acaba de celebrar en Madrid, con un éxito indiscutible, el Congreso Internacional de Matemáticos. Como no podía ser de otra forma, también se habló de la enseñanza de las matemáticas. En la ceremonia de clausura, Sir John Ball, presidente de la IMU, hizo mención explícita de la importancia que tiene la enseñanza de las matemáticas a distintos niveles tanto para el futuro de la investigación matemática y para el desarrollo de la ciencia y la técnica en el siglo XXI, como para la formación y el bienestar de toda la humanidad, y por todo ello nos invitó a todos los matemáticos y, muy en particular, a los más sabios de los allí presentes, a ocuparse y responsabilizarse de la educación matemática.

A mí me pareció muy importante esa reflexión en ese momento.

Por lo que se refiere a la enseñanza de la Matemática Aplicada, el profesor Lennart Carleson, premio Abel 2006, en la mesa redonda de clausura que se celebró el martes 29 de agosto dijo, entre otras cosas, que todo estudiante de ciencias o técnicas aplicadas debe estudiar un poco de ecuaciones en derivadas parciales, aunque no se le cuenten los teoremas de existencia y unicidad de las ecuaciones diferenciales ordinarias. También me pareció muy importante esa frase, dicha en en ese lugar por un matemático de la categoría del prof. Carleson.

S. Rodríguez Salazar es representante de S$\vec{e}$MA en la comisión de enseñanza del CEMAT.

Título:	Cálculo de la escisión de separatrices usando técnicas de matching complejo y resurgencia aplicadas a la ecuación de Hamilton-Jacobi.
Doctorando:	Carme Olivé Farré.
Director/es:	Tere Martínez-Seara.
Defensa:	10 de julio de 2006, Barcelona.
Calificación:	Excelente cum laude.

Resumen:

El objetivo de nuestro estudio es aplicar las técnicas de matching complejo y de la Teoría de la Resurgencia al problema de la medida de zonas caóticas en sistemas dinámicos hamiltonianos.

Nos hemos centrado en un ejemplo test usado ya por diversos autores anteriormente, como es el sistema del péndulo con una perturbación rápidamente oscilante $\mu(1-\cos q)\sin(t/\varepsilon)$ con $\varepsilon \in (0,1)$ y μ independiente de ε. La metodología y las técnicas utilizadas son pero aplicables a un conjunto más amplio de sistemas.

Aunque la perturbación no sea pequeña, este tipo de sistemas se comportan como próximos a integrables en el sentido que las zonas caóticas son exponencialmente pequeñas cuando ε tiende a cero. Para tener una idea de la magnitud de estas zonas caóticas, estudiamos la rotura de las separatrices, asociadas a la órbita homoclínica del sistema no perturbado, y medimos la separación entre las variedades invariantes estable e inestable asociadas a la órbita periódica hiperbólica del sistema perturbado. Estas variedades bidimensionales pueden representarse como grafos de las diferenciales de unas funciones analíticas T^+ y T^-, que son dos soluciones particulares de la ecuación en derivadas parciales de Hamilton-Jacobi.

Después de un cambio de variables adecuado en el plano complejo, que nos lleve cerca de la singularidad de la órbita homoclínica del sistema no perturbado, es posible obtener la parte dominante de la ecuación de Hamilton-Jacobi independente del parámetro singular ε, llamada Ecuación Inner. A partir de dos soluciones particulares de esta ecuación, y mediante la Teoría de la Resurgencia, calculamos en primer orden de ε la separación entre las variedades invariantes.

La diferencia total $T^+ - T^-$ es solución de una ecuación en derivadas parciales lineal homogénea, de la cual, por enderezamiento del flujo, se demuestra que sus soluciones acotadas en una cierta banda vertical compleja son exponencialmente pequeñas en el campo real. Usando técnicas de matching complejo, obtenemos tanto la cota de $T^+ - T^-$, como el cambio de variables que endereza el flujo.

Si $\mu = \varepsilon^p$, los resultados que hemos obtenido confirman, en los casos $0 < p < 2$ que aún se mantenían sin resolver, el término dominante de la distancia entre variedades que prevé el método perturbativo de Poincaré-Melnikov. En cualquier

caso, obtenemos una fórmula asintótica de esta distancia para ε pequeño, con el parámetro μ independiente de ε.

Título:	An unstructured finite volume model for unsteady turbulent shallow water flow with wet-dry fronts: Numerical solver and experimental validation.
Doctorando:	Luis Cea Gómez.
Director/es:	Jerónimo Puertas Agudo y María-Elena Vázquez Cendón.
Defensa:	23/06/2005.
Calificación:	Sobresaliente cum Laude[1].

Resumen:

El principal objetivo de la tesis es aplicar las ecuaciones de aguas someras bidimensionales a diferentes flujos en lámina libre, centrándose especialmente en aquellos problemas en los que o bien el tratamiento de la turbulencia o bien el tratamiento del frente seco-mojado son de especial relevancia. Para ello se ha desarrollado un código de volúmenes finitos que resuelve las ecuaciones de aguas someras acopladas con diferentes modelos de turbulencia. El código incluye un modelo parabólico de viscosidad turbulenta, un modelo algebraico de longitud de mezcla y 3 versiones del modelo k-e para aguas someras. Al mismo tiempo se ha propuesto e incluido un modelo de tensiones algebraicas para aguas someras. El código desarrollado se ha utilizado para simular el flujo en 4 aplicaciones prácticas diferentes que incluyen el oleaje de onda larga, el flujo inducido por la marea en un estuario, el flujo en canal con un codo de 90º, y el flujo en escalas de peces de hendidura vertical. En todos los casos los resultados numéricos se han comparado con datos experimentales.

[1]Esta tesis ha sido propuesta este año por SëMApara el premio Eccomas

Título:	Modelos de aguas poco profundas obtenidos mediante la técnica de desarrollos asintóticos.
Doctorando:	Raquel Taboada Vázquez.
Director/es:	José Manuel Rodríguez Seijo.
Defensa:	27 de septiembre de 2006, Universidade da Coruña.
Calificación:	Sobresaliente Cum Laude por unanimidad.

Resumen:

Clásicamente las ecuaciones de aguas poco profundas se obtienen a partir de las ecuaciones de Euler o Navier-Stokes mediante ciertas hipótesis simplificadoras. Dichas hipótesis no siempre están debidamente justificadas, lo que conduce a gran variedad de modelos, sin que resulte claro cuál de ellos es el "mejor". En esta tesis se obtienen, utilizando el método de desarrollos asintóticos, diferentes modelos de aguas someras con y sin viscosidad, unidimensionales y bidimensionales de una forma rigurosa y sin realizar las usuales hipótesis a priori.

Para aplicar el método de desarrollos asintóticos, identificamos un pequeño parámetro adimensional (relacionado con la profundidad del dominio) y realizamos un cambio de variable a un dominio independiente de dicho parámetro. Suponemos entonces que la solución de las ecuaciones de Euler o de Navier-Stokes admite un desarrollo en serie de potencias del pequeño parámetro y calculamos los primeros términos de dicho desarrollo. Construimos una aproximación de la solución a partir de los términos del desarrollo en serie de potencias calculados y deshacemos el cambio de variable, con lo que obtenemos un modelo de aguas poco profundas.

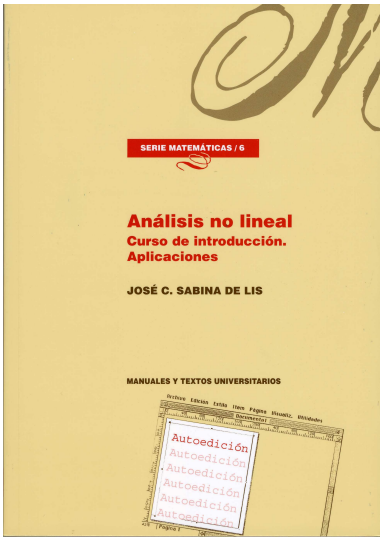
En el modelo obtenido a partir de las ecuaciones de Euler la velocidad horizontal depende de la profundidad si la vorticidad es no nula, lo que supone una interesante novedad respecto a los modelos que se encuentran en la literatura. El modelo de aguas someras sin viscosidad que proponemos generaliza al modelo clásico y permite calcular de forma exacta soluciones de las ecuaciones de Euler lineales en z, mientras que el modelo clásico tan sólo lo hace con soluciones constantes en z.

Si partimos de las ecuaciones de Navier-Stokes, el modelo al que llegamos incluye un nuevo término de viscosidad. Hemos comparado nuestro modelo analítica y numéricamente con otros modelos que se pueden encontrar en la literatura, obteniendo que el modelo que proponemos mejora (o en el peor de los casos iguala) los resultados de los otros modelos.

En definitiva, los modelos propuestos suponen una mejora respecto a los modelos que se encuentran en la literatura en el sentido de que el modelo sin viscosidad incorpora una dependencia de la profundidad que permite aproximar mejor las ecuaciones de Euler y el modelo viscoso incorpora un nuevo término de viscosidad justificado mediante el método de desarrollos asintóticos y numéricamente.

Análisis no lineal. Curso de introducción. Aplicaciones
José C. Sabina de Lis
Manuales y textos universitarios. Serie Matemáticas/6. Servicio de Publicaciones de la Universidad de La Laguna
ISBN: 84–7756–653–5 (248 páginas) – 2005

Por J. M. Arrieta
Univ. Complutense de Madrid



En este obra, José C. Sabina de Lis, Catedrático de Matemática Aplicada de la Universidad de La Laguna, nos brinda un tratado de análisis no lineal y aplicaciones con énfasis en las ecuaciones diferenciales, que refleja en gran medida la visión e interés particular en este tema y el claro dominio de las herramientas y conceptos por parte del autor en el campo.

Este libro tiene dos partes diferenciadas. La primera, que abarca los tres primeros capítulos, recoge los elementos esenciales del análisis funcional lineal en espacios de Banach y lo que podríamos denominar la teoría de la linealización, entendiendo que esta última engloba los distintos conceptos de derivada (Gateaux y Fréchet), polinomios de Taylor y las funciones analíticas en espacios de funciones. En aproximadamente ochenta páginas, se pasa revista sobre conceptos y herramientas básicas (que todo estudiante de doctorado en el área de ecuaciones diferenciales debe aprender) y que se usan para el desarrollo posterior de los contenidos del libro. El autor ha conseguido realmente resumir en tres capítulos de fácil lectura un material que, además de ser interesante en sí mismo, sirve de referencia para adentrase en aguas más profundas.

En la segunda parte del libro, Capítulos IV a VII, se desarrollan las herramientas para abordar el problema fundamental del análisis no lineal que es la obtención de soluciones de ecuaciones no lineales en espacios funcionales y el estudio de sus propiedades. El teorema estrella de esta parte, sobre todo de los

capítulos IV y V es el Teorema de la Función Implícita, que se obtiene a partir del Teorema de la Contracción Uniforme. La importancia de este teorema se ve tanto en las aplicaciones inmediatas e importantes que se describen en el libro: problemas de valor inicial para ecuaciones diferenciales ordinarias y su dependencia con respecto a parámetros, la construcción de la variedad estable e inestable para puntos silla así como en las elaboraciones mucho menos triviales de este teorema, como son los teoremas de Bifurcación de Hopf con su aplicación al teorema centro de Lyapunov, el teorema de Crandall–Rabinowitz sobre bifurcación desde un autovalor simple y el método de reducción a dimensión finita de Lyapunov–Schmidt.

Ya que con el Teorema de la Funcion Implícita se obtiene solo información local de las soluciones, se hace necesario obtener resultados de inversión global para analizar la existencia y posible unicidad de soluciones alejados de los puntos de bifurcación. En el Capítulo VI se aborda este problema y en particular se incluye un resultado de inversión de Hadamard y su versión dada por Schwartz.

Finalmente, y como no podía faltar en un libro sobre análisis no lineal, se desarrolla en el último capítulo una introducción a la teoría del grado topológico, que constituye, tanto en su versión finito dimensional (Brower) como en su versión infinito dimensional (Leray–Schauder) una segunda herramienta, más elaborada pero también fundamental para el estudio de las soluciones de ecuaciones no lineales. Como aplicación del grado topológico a la teoría de bifurcación se obtienen resultados de bifurcación desde autovalores de multiplicidad impar, que no se pueden obtener a partir del teorema de la funcion implícita. A lo largo de la segunda parte del libro, la teoría desarrollada se va aplicando a ejemplos concretos e importantes de las ecuaciones diferenciales. Hay uno de estos ejemplos que brilla con luz propia y es el estudio de problemas elípticos semilineales con condiciones de tipo Dirichlet. En total se dedican cinco secciones (notas las denomina el autor) a este problema importante.

En este libro se vislumbra la influencia que sobre el autor y su obra científica han ejercido distintas escuelas como la de J. Hale, P. Rabinowitz, H. Amann y H. Brezis. Ahora bien, lejos de ser una mera recopilación de teoremas, esta obra, que el autor dedica a la memoria de su maestro y amigo, José M. Fraile, constituye una versión original y propia, con contenidos de profundidad pero al mismo tiempo escrito con transparencia y delicadeza, poniendo el énfasis en los puntos importantes y con una vocación de acompañar al lector a través de las herramientas y resultados fundamentales del análisis no lineal.

Estimados socios:

Os comento algunas noticias que pueden ser de vuestro interés. Recordad que podéis consultar esta información desde la página web de SEMA (`www.sema.org.es`).

1. Renovación de los cargos de SEMA. El pasado mes de septiembre, durante la Asamblea Ordinaria de SEMA se procedió a la renovación del Presidente y dos miembros del Comité Ejecutivo de SEMA. El Profesor Carlos Vázquez Cendón fue elegido nuevo Presidente de SEMA. La Profesora Rosa Donat Beneito y el Profesor Carlos Parés Madroñal fueron elegidos nuevos miembros del Consejo Ejecutivo. El nuevo Tesorero de SEMA es el Profesor Iñigo Arregui.

2. CEDYA 2007. Recientemente hemos recibido por parte de la organización la segunda circular sobre el próximo CEDYA. En particular ya está abierto el plazo para la inscripción, presentación de propuestas de comunicaciones y solicitud de beca. Tenéis más información en la página web del Congreso `www.congreso.us.es/cedya2007`.

3. Congreso conjunto RSME-SMF-SEMA. Del 9 al 13 de julio del próximo año 2007 se celebrará el primer Congreso conjunto entre las sociedades españolas RSME y SEMA, y la sociedad francesa SMF. Podéis encontrar más información en la página web del Congreso: `http://www.unizar.es/ICHFM07/`

4. Web SEMA. Como ya sabéis tenemos una nueva página web con un tablón de anuncios y una bolsa de trabajo que está a vuestra disposición. Os animamos a que uséis este servicio de publicidad. Los anuncios los podéis incluir vosotros mismos a través de la web desde las secciones correspondientes.

Un cordial saludo, Carlos Castro.

VIII PREMIO S$\vec{e}$MA A LA "DIVULGACIÓN DE LA MATEMÁTICA APLICADA"

SOCIEDAD ESPAÑOLA DE MATEMÁTICA APLICADA

(PATROCINADO POR IBERDROLA)

PREÁMBULO

La Sociedad Española de Matemática Aplicada (S$\vec{e}$MA), en cumplimiento de su objetivo de contribuir al desarrollo en nuestro país de las Matemáticas y sus aplicaciones, consciente del notable desarrollo que las Matemáticas están experimentando, del incremento de su influencia sobre todos los aspectos de la vida en las sociedades desarrolladas, de la conveniencia de promover el interés de los investigadores por este punto de vista sin por ello ocultar sus peligros o dilemas, de la necesidad no menos acuciante de estimular el interés del público por la cultura científica y, finalmente continuando con una tradición honrosa y habitual tanto en las Artes como en las Ciencias, convoca el "VIII Premio S$\vec{e}$MA de Divulgación de la Matemática aplicada", según las bases que se adjuntan.

S$\vec{e}$MA busca ante todo promover la divulgación de las Matemáticas, su relevancia y su eficacia. Dada la enorme variedad de intereses aplicados de las Matemáticas, las Bases del concurso pretenden dar preferencia a los temas que tradicionalmente han estado ligados a S$\vec{e}$MA de una u otra manera. Muy en especial, deben ser mencionados el análisis teórico y numérico, el control y los aspectos computacionales de sistemas que permiten modelizar fenómenos con origen en otras Ciencias.

BASES DE LA CONVOCATORIA

1. La Sociedad Española de Matemática aplicada (S$\vec{e}$MA) convoca el "Premio S$\vec{e}$MA a la Divulgación de la Matemática aplicada", que se concederá anualmente.

2. Son posibles candidatos todos los ciudadanos del mundo que sometan un texto de acuerdo con los puntos 4 y 9 de estas Bases.

3. El Premio está destinado a promover los valores de la belleza, relevancia y eficacia de las Matemáticas como instrumento indispensable del funcionamiento de la sociedad y cultura modernas. El Premio tomará en especial consideración los temas que incidan en la realidad de la Matemática aplicada en la sociedad española.

4. Los candidatos habrán de presentar dentro del plazo fijado en el punto 10 un texto original de una longitud mínima de 20 páginas mecanografiadas a un espacio y con el formato que juzguen conveniente. Los requisitos básicos son que el texto contribuya a la divulgación de algún aspecto relevante de la Matemática Aplicada y que su contenido esté pensado para un público no exclusivamente formado por profesionales de las Matemáticas. El trabajo será presentados

bajo un seudónimo, incluyendo con el mismo un sobre cerrado en el que figuren el nombre y dirección del autor. El autor no podrá formar parte del Comité Científico que habrá de juzgar los trabajos por lo que en caso de ser propuesto para el mismo, deberá indicar al Presidente su incompatibilidad. 5. Los méritos serán juzgados por un Comité Científico de cinco miembros nombrados por el Comité Ejecutivo de la Sociedad, personalidades de probado prestigio en la Ciencia Matemática y la cultura científica. Este Comité tendrá su propio reglamento de funcionamiento, pudiendo quedar desierto el Premio. En todo caso, el Comité será presidido por el Presidente de la Sociedad u otro miembro del Comité Ejecutivo en quien delegue, no pudiendo ser miembros del Comité Científico más de dos miembros del Comité Ejecutivo. 6. El galardonado con el Premio recibirá de la Sociedad un Diploma acreditativo y una cuantía de 1500 euros. Además quedará eximido del pago de las cuotas como socio de S$\vec{e}$MA correspondientes a los años 2008 y 2009. En caso de no ser miembro de S$\vec{e}$MA, pasaría automáticamente a serlo.

7. El fallo del concurso es irrevocable. El Comité acompañará la concesión del Premio de una exposición de los méritos hallados en el candidato galardonado. 8. La Sociedad publicará la obra premiada en su Boletín.

9. Si el texto original no estuviera escrito en castellano, el jurado podrá solicitar al autor su traducción si así lo estimase necesario.

10. La fecha límite de presentación de originales es el 15 de mayo de 2007.

11. La documentación, compuesta del texto por quintuplicado, su traducción si ha lugar, así como los datos identificativos, debe ser dirigida a

Prof. Carlos Vázquez Cend'on
VIII Premio S$\vec{e}$MA a la Divulgación de la Matemática Aplicada
Departamento de Matemáticas
Facultad de Informática
Campus de Elviña s/n
Universidad de La Coruña
15071- A Coruña

12. El Premio será fallado antes del 31 de agosto del año 2007 y será entregado con ocasión de la Asamblea anual de la Sociedad, en el marco del XX CEDYA / X CMA que tendrá lugar en Sevilla del 24 al 28 de Septiembre de 2007.

Tipo de evento:	Congreso
Nombre:	Recent Trends in Nonlinear Science (RTNS 2007)
Lugar:	Granada
Fecha:	del 5 al 9 de febrero de 2007
E-mail:	`rtns2007@dance-net.org`
WWW:	`www.dance-net.org/rtns/index.php?id_evento=7`

Tipo de evento:	Congreso
Nombre:	6th International Conference on Applied Mathematics (APLIMAT 2007)
Lugar:	Bratislava (Slovakia)
Fecha:	del 6 al 9 de febrero de 2007
E-mail:	`aplimat@aplimat.com`
WWW:	`www.aplimat.com`

Tipo de evento:	Congreso
Nombre:	Ninth Workshop on Optimal Control, Dynamic Games and Nonlinear Dynamics
Lugar:	Montréal (Canada)
Fecha:	del 7 al 9 de mayo de 2007
Organiza:	Michèle Breton, Sihem Taboubi, Georges Zaccour, Groupe d'études et de recherche en analyse des décisions (GERAD), Chair in Game Theory and Management
Información:	Carole Dufour (carole.dufour@gerad.ca), Georges Zaccour (georges.zaccour@gerad.ca)
WWW:	`www.gerad.ca/colloques/9thWorkshop2007/index.htm`

Tipo de evento:	Congreso
Nombre:	Congreso de Educación Matemática (XIII JAEM)
Lugar:	Granada
Fecha:	del 4 al 7 de julio de 2007
Organiza:	Sociedad Andaluza de Educación Matemática "THALES"
E-mail:	xiiijaem@fespm.org
WWW:	http://thales.cica.es/jaem

Tipo de evento:	Congreso
Nombre:	INTERNATIONAL CONFERENCE ON APPROXIMATION METHODS AND NUMERICAL MODELLING IN ENVIRONMENT AND NATURAL RESOURCES (MAMERN07)
Lugar:	Granada
Fecha:	del 11 al 13 de julio de 2007
Organiza:	Universidad de Granada, Universidad Mohammed Premier of Oujda y CNRST (Marruecos), Universidad de Pau y CNRS (Francia)
Información:	Domigo Barrera Rosillo (dbarrera@ugr.es), Brain Amaziane (brahim.amaziane@univ-pau.fr), Driss Sbibih (mamern@sciences.univ-oujda.ac.ma)
E-mail:	mamern07@ugr.es
WWW:	www.ugr.es/local/mamern07

Tipo de evento:	Congreso
Nombre:	7th European Conference on Numerical Mathematics and Advanced Applications (ENUMATH 2007 conference)
Lugar:	Graz (Austria)
Fecha:	del 10 al 14 de septiembre de 2007
Organiza:	Institute for Mathematics and Scientific Computing (University of Graz), Institute for Computational Mathematics, Graz University of Technology
E-mail:	enumath07@uni-graz.at
WWW:	www.uni-graz.at/enumath07

Aguilar Baltar, Adolfo Luis
Prof. Titular de Universidad. *Líneas de investigación:* Distribución y abundancia de poblaciones. Restos asociados a yacimientos arqueológicos – Univ. Autónoma de Madrid – Fac. de Ciencias – Dpto. de Biología – Crta. de Colmenar, Km 15; 28049 Madrid.
Tlf.: 914978287. *Fax:* 914978344.
e-mail: adolfo.aguilar@uam.es

Cendán Verdes, José Jesús
Prof. Titular de Escuela Universitaria. *Líneas de investigación:* Lubricación elastohidrodinámica – Univ. de La Coruña – Fac. de Informática – Dpto. de Matemáticas – Campus de Elviña, s/n; 15071 La Coruña.
Tlf.: 981167000 ext. 1334. *Fax:* 981167160.
e-mail: suceve@udc.es
http://www.udc.es/dep/mate/jcv/index.htm

Sánchez Sánchez, Angel
Prof. Titular de Universidad. *Líneas de investigación:* Sistemas complejos. EDP's con solitones – Univ. Carlos III de Madrid – Escuela Politécnica Superior – Dpto. de Matemáticas – Avda. de la Universidad, 30; 28911 Leganés (Madrid).
Tlf.: 916249411. *Fax:* 916249129.
e-mail: anxo@math.uc3m.es

Silva Santos, Antonio
˙ *Líneas de investigación:* .*Tlf.:*
Fax: .
e-mail: asisanto@clix.pt

Tapiador Fernández, Bárbara
Estudiante. *Líneas de investigación:* – Univ. Complutense de Madrid – Fac. de CC. Matemáticas – Dpto. de Matemática Aplicada – Ciudad Universitaria, s/n; 28040 Madrid.
Tlf.: . *Fax:* .
e-mail: Barbara.Tapiador@gmail.com

Direcciones útiles

Consejo Ejecutivo de SëMA

Presidente:

Carlos Vázquez Cendón. (carlosv@udc.es).
Dpto. de Matemáticas.Facultad de Informática. Univ. de La Coruña.Campus de Elviña, s/n. 15071 La Coruña. *Tel:* 981 16 7000-1335.

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.

Tesorero:

Íñigo Arregui Álvarez. (arregui@udc.es).
Dpto. de Matemáticas. Fac. de Informática. Univ. de La Coruña. Campus de Elviña, s/n. 15071 La Coruña. *Tel:* 981 16 7000-1327.

Vocales:

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:* 93 581 2413.

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:* 96 354 4727.

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 13 2017.

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.

José Javier Valdés García. (valdes@orion.ciencias.uniovi.es).
Dpto. de Matemáticas. Fac. de Ciencias. Univ. de Oviedo. Avda. de Calvo Sotelo, s/n. 33007 Oviedo. *Tel:* 985 103 340.

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:* 914 974 368.

Comité Científico del Boletín de SëMA

Enrique Fernández Cara. (cara@us.es).
Dpto. de Ecuaciones Diferenciales y An. Numérico. Fac. de Matemáticas. Univ. de Sevilla. Tarfia, s/n. 41012 Sevilla. *Tel:* 954 557 992.

Alfredo Bermúdez de Castro. (mabermud@usc.es).
Dpto. de Matemática Aplicada. Fac. de Matemáticas. Univ. de Santiago de Compostela. Campus Univ.. 15706 Santiago (A Coru na) *Tel:* 981 563 100.

Eduardo Casas Rentería. (eduardo.casas@unican.es).
Dpto. de Matemática Aplicada y C.C.. E.T.S.I. Ind. y Telec. Univ. de Cantabria. Avda. de Los Castros s/n. 39005 Santander. *Tel:* 942 201 427.

José Luis Cruz Soto. (jlcruz@uco.es).
Dpto. de Informática y An. Numérico. Univ. de Córdoba. Campus de Rabanales. Edificio C-2. 14071 Córdoba. *Tel:* 957 218 629.

José Manuel Mazón Ruiz. (Jose.M.Mazon@uv.es).
Dpto. de Análisis Matemático. Fac. de Matemáticas. Univ. de Valencia. Dr. Moliner, 50. 46100 Burjassot (Valencia) *Tel:* 963 664 721.

Ireneo Peral Alonso. (ireneo.peral@uam.es).
Dpto. de Matemáticas, C-XV. Fac. de Ciencias. Univ. Aut. de Madrid. Cantoblanco, Ctra. de Colmenar, km. 14. 28049 Madrid. *Tel:* 913 974 204.

Luis Ferragut Canals. (ferragut@usal.es).
Dpto. de Matemática Aplicada. Fac. de Ciencias. Univ. de Salamanca. Plaza de la Merced, s/n. 37006 Salamanca. *Tel:* 923 294 400 ext. 1522.

Juan Luis Vázquez Suárez. (juanluis.vazquez@uam.es).
Dpto. de Matemáticas, C-XV. Fac. de Ciencias. Univ. Aut. de Madrid. Cantoblanco, Crta. de Colmenar, km. 14. 28049 Madrid. *Tel:* 913 974 935.

Luis Vega González. (mtpvegol@lg.ehu.es).
Dpto. de Matemáticas. Fac. de Ciencias. Univ. del País Vasco. Aptdo. 644. 48080 Bilbao (Vizcaya). *Tel:* 944 647 700.

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

Luis Ferragut Canals. (ferragut@usal.es).
Dpto. de Matemática Aplicada. Fac. de Ciencias. Univ. de Salamanca. Plaza de la Merced, s/n. 37006 Salamanca. *Tel:* 923 294 400, ext. 1522.

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.

Francisco Andrés Pérez. (franc@usal.es).
Dpto. de Matemática Aplicada. Fac. de Ciencias. Univ. de Salamanca. Plaza de la Merced, s/n. 37006 Salamanca. *Tel:* 923 294 400 ext. 1537.

M. Isabel Asensio Sevilla. (mas@usal.es).
Dpto. de Matemática Aplicada. Fac. de Ciencias Químicas. Univ. de Salamanca. Plaza de la Merced, s/n. 37006 Salamanca. *Tel:* 923 294 400 ext. 1537.

M. Teresa de Bustos Muñoz. (tbustos@usal.es).
Dpto. de Matemática Aplicada. E.T.S. Ing. Ind. de Béjar. Univ. de Salamanca. Avda. Fernando Ballesteros, 2. 37700 Béjar, Salamanca. *Tel:* 923 408 080 ext. 2263.

Antonio Fernández Martínez. (anton@usal.es).
Dpto. de Matemática Aplicada. E. Politécnica Superior Zamora. Univ. de Salamanca. Avda. Requejo, 33. Campus Viriato. 49022 Zamora. *Tel:* 980 545 000 ext. 4459.

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 SëMA:

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 SëMA

Gestión de Socios:

María Pilar Laburta Santamaría. (laburta@unizar.es).
Dpto. de Matemática Aplicada. Centro Politécnico Superior. Univ. de Zaragoza. Edificio Torres Quevedo. C/ María de Luna 3. 50018 Zaragoza. *Tel:* 976 762 006.

Página web: www.sema.org.es/**:**

J. Rafael Rodríguez Galván. (rafael.rodriguez@uca.es).
Dpto. de Matemáticas. Fac. de CC. EE. y Empresariales. Univ. de Cádiz. C/ Duque de Nájera, 6. 11002 Cádiz. *Tel:* 956 015 478.

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 SëMA

Dpto. E.D.A.N., Facultad de Matemáticas

Aptdo. 1160, 41080 SEVILLA

También podrán ser enviados por correo electrónico a la dirección

`boletin_sema@usal.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 SēMA

Remitir a: SEMA, Despacho 520, Facultad de Matemáticas,
Universidad Complutense. 28040 Madrid.
Fax: 913 944 607. 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 2005

☐ Socio ordinario: 30 EUR. ☐ Socio de reciprocidad con la RSME: 12 EUR.
☐ Socio estudiante: 15 EUR. ☐ Socio extranjero: 25 EUR.

Datos bancarios

. . . de . . . . . de 200. .

Muy Sres. Míos:

Ruego a Uds. que los recibos que emitan a mi cargo en concepto de cuotas de inscripción y posteriores cuotas anuales de S$\vec{e}$MA (Sociedad Espa nola de Matemática Aplicada) sean pasados al cobro en la cuenta cuyos datos figuran a continuación

Entidad (4 dígitos)	Oficina (4 dígitos)	D.C. (2 dígitos)	Número de cuenta (10 dígitos)

- Entidad bancaria: . . . . . .
- Domicilio: . . . . . .
- C.P.: . . . . . Población: . . . . . .

Con esta fecha, doy instrucciones a dicha entidad bancaria para que obren en consecuencia.

Atentamente,

Fdo. . . . . .

Para remitir a la entidad bancaria

. . . de . . . . . de 200. .

Muy Sres. Míos:

Ruego a Uds. que los recibos que emitan a mi cargo en concepto de cuotas de inscripción y posteriores cuotas anuales de S$\vec{e}$MA (Sociedad Espa nola 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: SEMA, Despacho 520, Facultad de Matemáticas,
Universidad Complutense. 28040 Madrid.
Fax: 913 944 607. 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 2005 como Socio Institucional es de 150 EUR.
El pago se realiza mediante transferencia bancaria a

SEMA: 0128 - 0380 - 03 - 0100034244
Bankinter
C/ Hernán Cortés, 63
39003 Santander