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

Página web de S$\vec{e}$MA

`http://www.uca.es/sema/`

Dirección Editorial: Boletín de S$\vec{e}$MA. Dpto. de Matemáticas. Universidad de Oviedo. Avda. de Calvo Sotelo, s/n. 33007-Oviedo. `boletin_sema@orion.ciencias.uniovi.es`

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

Imprime: Grupo Bitácora. C/ Instituto, 17, Entresuelo. 33201 Gijón (Asturias)

Diseño de portada: Ana Cernea

Estimados amigos:

Como ya es habitual, incluimos en este mensaje nuestro más cordial saludo, seguido de unas palabras de introducción a un nuevo número de nuestro Boletín. En este caso, junto a las secciones habituales de anuncios, noticias, resúmenes de Tesis Doctorales, etc., hemos incluido tres interesantes artículos de investigación, el trabajo recientemente galardonado con el IV Premio S$\vec{e}$MA de Divulgación de la Matemática Aplicada y, finalmente, un nuevo artículo de Educación Matemática.

En la sección de Noticias se recogen dos hechos relevantes en el ámbito de la Matemática Aplicada que se han producido en los meses de septiembre y octubre. Por una parte, informaremos de la concesión del Premio Nacional de Chile en Ciencias Exactas 2003 al Prof. Carlos Conca, de la Universidad de Chile, viejo conocido y colaborador de muchos socios de S$\vec{e}$MA. Por otro lado, recordaremos que (como probablemente ya sabéis) el Premio Nacional de Investigación 2003 Julio Rey Pastor de Matemáticas y Tecnologías de la Información y Comunicaciones ha sido concedido a nuestro compañero y ex-Presidente Juan Luis Vázquez, de la Universidad Autónoma de Madrid. Es para nosotros un verdadero placer y orgullo que sean reconocidos de esta forma los méritos de dos investigadores de máximo prestigio en nuestro ámbito.

Muy en particular, nos congratulamos de lo acaecido en el caso de un ex-Presidente de nuestra sociedad y miembro muy activo de la misma. Las aportaciones científicas de Juan Luis, su decidido trabajo en favor del desarrollo de las Matemáticas de calidad y sus actividades de política científica son dignas de elogio y le hacen merecedor de este prestigioso galardón. ¡Enhorabuena!

Una vez más, ha sido para nosotros un placer elaborar este volumen. Por supuesto, seguimos estando a vuestra disposición para cualquier sugerencia que permita mejorar el resultado.

Un cordial saludo,

Grupo Editor
`boletin_sema@orion.ciencias.uniovi.es`

Bol. Soc. Esp. Mat. Apl. n°26(2003), 7–34

Simulación numérica de las deformaciones termomecánicas de una placa de aluminio durante el proceso de colada*

P. Barral y P. Quintela

Departamento de Matemática Aplicada. Universidade de Santiago de Compostela

patribr@usc.es, mapere@usc.es

Resumen

En este trabajo se resume la investigación realizada por las autoras en el campo de la simulación termomecánica de coladas de aluminio. Se propone un modelo matemático para simular las deformaciones de la placa durante su solidificación y se presentan resultados de existencia y regularidad de solución. A continuación, se aborda la resolución numérica del problema. Para implementar numéricamente la presión metalostática ejercida por el aluminio líquido sobre la placa se utiliza un método de dominio ficticio, que se justifica mediante un análisis asintótico del problema. Para el tratamiento de las no linealidades del problema se aplica el método de Bermúdez-Moreno combinado con un método de Newton generalizado. El problema se discretiza en espacio mediante un método de elementos finitos y en tiempo con un esquema implícito. Finalmente, se presentan resultados numéricos para una colada real.

Palabras clave: *colada, contacto, viscoelasticidad, dominio ficticio, desarrollos asintóticos, operadores maximales monótonos, elementos finitos.*
Clasificación por materias AMS: *35, 65, 68, 74.*

1 El proceso de producción de aluminio

El proceso de producción de aluminio, esquematizado en la Figura 1, es muy complejo y se puede dividir en las siguientes etapas:

* Investigación financiada por: ALCOA-INESPAL S.A. (A Coruña), CICYT (proyecto MAT96-0715), CYCIT-FEDER (proyecto DPI2001-29098), IGAPE y Xunta de Galicia (proyectos PGIDIT02PXIC20701PN, PGIDIT00PXI20701PR).

Fecha de recepción: 24 de marzo de 2003

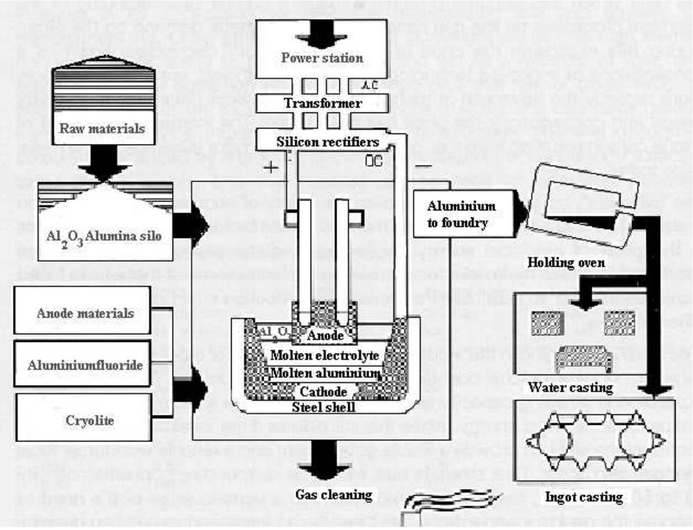


Figura 1: Esquema del proceso de producción de aluminio

Extracción de bauxita

La bauxita se extrae directamente de distintos yacimientos y se transporta hasta la estación de trituración. Allí se traslada hasta el molino que tiene como función reducir el material a una granulometría menor de 100 mm para su transporte y mejor manejo.

Obtención de alúmina

El proceso de obtención de alúmina de grado metalúrgico a partir de la bauxita consiste en la reducción del tamaño de partículas de bauxita, la extracción de la alúmina contenida en ella por medio de la digestión con soda cáustica y la separación de las impurezas que acompañan a la alúmina a fin de prevenir la contaminación del producto final. Después de varios filtrados, el líquido, rico en alúmina, se somete a una fase de enfriamiento que lo prepara para la fase de precipitación, donde se obtienen los cristales de alúmina hidratada. El proceso concluye en los calcinadores, constituidos por grandes hornos que eliminan la humedad de la alúmina hidratada para obtener la alúmina de grado metalúrgico, producto final dispuesto para ser utilizado en las cubas electrolíticas.

Electrolisis

La planta de reducción o de cubas electrolíticas es el corazón del proceso de producción del aluminio. Allí se disuelve la alúmina en un medio electrolítico de criolita fundida a 950°C aproximadamente, descomponiéndola en sus dos elementos básicos: oxígeno y aluminio. El

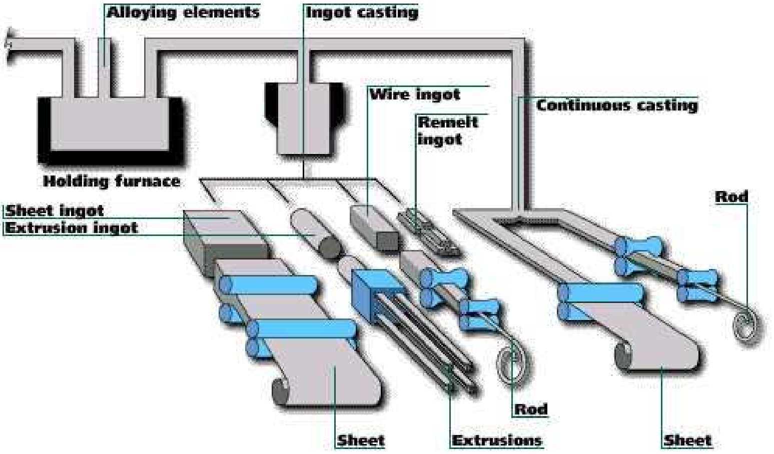


Figura 2: El aluminio líquido viaja desde el horno hasta la mesa de colada, donde solidifica. El lingote obtenido sufre posteriormente un proceso de extrusión, trefilado o laminación.

oxígeno es atraído por los ánodos hacia la parte superior de la celda, quemado y convertido en dióxido de carbono en el ánodo. El aluminio, a su vez, se precipita hacia el fondo y se extrae fundido por succión hacia el crisol, para ser enviado a la planta de fundición.

Colada

En la planta de fundición se preparan las aleaciones de aluminio atendiendo a criterios comerciales. El metal, procedente de las cubas de electrolisis, se vacía en los hornos de retención donde se le añaden aditivos como titanio, magnesio, cobre o hierro, para preparar las distintas aleaciones. El metal líquido es sometido a diversos análisis y controles de calidad dentro de los hornos, para luego ser transferido a la mesa de colada. En el proceso de colada se vierte el metal líquido a través de canales a los diferentes moldes que son enfriados por agua (ver Figura 5). El producto final es aluminio primario en forma de lingote para trefilado, cilindros para extrusión y placas para laminación.

Trefilado, extrusión y laminación

Finalmente, mediante trefilado, extrusión o laminación, se obtienen bobinas de alambre, barras o láminas, que son enviadas a los clientes, quienes las transformarán en recipientes, ventanas, piezas de automóviles, etc. (ver Figura 2).

2 Proceso de colada

En este trabajo nos centramos en el proceso de colada. Como ya hemos mencionado, al llegar a la mesa de colada el aluminio se vierte en el molde o lingotera, donde un sistema de refrigeración por agua lo hace solidificar. Se forma así una corteza, correspondiente al aluminio ya solidificado, que sirve de receptáculo al aún líquido. Al mismo tiempo, una plataforma móvil, situada en la parte inferior y llamada falso fondo, va descendiendo de forma continua a una velocidad aproximada de 1mm/s, mientras sobre la piscina de metal sigue llegando nuevo aluminio procedente del horno. Para más información sobre el proceso industrial de solidificación del aluminio se puede consultar Díaz [15] o Drezet-Plata [17].

Existen dos tipos de tecnologías para la colada de aluminio: la colada clásica, en la cual el aluminio líquido se vierte en un molde físico y la electromagnética, en la que el aluminio se confina mediante un campo magnético. En la colada clásica el contacto con el molde provoca procesos de cristalización del aluminio durante la solidificación. En la electromagnética, puesto que el aluminio solidifica sin estar en contacto con un molde físico, la placa tiene una calidad superficial mayor que en la clásica.

El proceso de colada es muy complejo, pues en él se desarrollan fenómenos electromagnéticos, térmicos, hidrodinámicos y mecánicos. Además, todos estos fenómenos están acoplados en algún sentido. Así, por ejemplo, el campo magnético, debido a la corriente, produce movimientos en la piscina de metal. Estos movimientos perturban a su vez el campo magnético y producen transporte de calor por convección en el aluminio aún líquido. Por otra parte, los grandes gradientes térmicos que surgen en el proceso de solidificación provocan tensiones térmicas dentro de la placa.

Los parámetros fundamentales que regulan el proceso de solidificación son:

- La velocidad de descenso del falso fondo, también llamada velocidad de colada.
- El caudal y la temperatura del agua de refrigeración, así como la ubicación de los puntos de refrigeración.
- El tipo de aleación.
- La geometría del falso fondo.
- La geometría del inductor.
- La intensidad de corriente que atraviesa el inductor.
- La posición de la pantalla.

La regulación de estos parámetros, mediante experiencias y extrapolaciones de las mismas, es un proceso laborioso, lento y muy caro pues el bloque obtenido puede quedar inutilizado, siendo necesaria su refundición. Por otra parte, un ensayo de modificaciones cualitativamente distintas en los parámetros puede llegar a provocar un accidente e inutilizar la instalación de colada. De aquí surge la necesidad de disponer de un modelo válido con el cual simular el proceso previamente y así adelantar las incidencias de tales modificaciones en la colada real, reduciendo el número de experimentos necesarios para optimizar el rendimiento de la colada y evitar ensayos peligrosos.

En la investigación realizada hasta el momento por el equipo de investigación al que pertenecen las autoras de este trabajo, se han seguido las siguientes fases:

1. Modelización térmica bi y tridimensional del bloque (Bermúdez-Otero [13, 30]).

2. Cálculo de las tensiones termomecánicas producidas en la fase sólida y geometría final de la placa (Barral-Quintela [1], [4]-[9]).

3. Cálculo del campo magnético, diseño de nuevas geometrías de inductores y obtención de la forma de la superficie superior del aluminio líquido, conocida como menisco (Bermúdez-Muñiz [12]).

4. Cálculo del campo de velocidades en la piscina de metal líquido debido a la fuerza de Lorentz (Bermúdez-Salgado [2]).

Este trabajo se centra en el segundo punto, que conlleva el estudio del submodelo mecánico, en el que surgen dos problemas bien diferenciados:

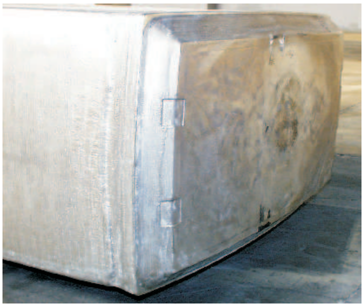


Figura 3: Deformación del talón

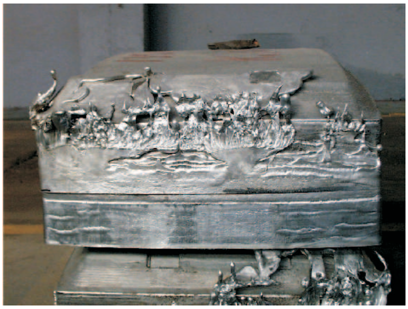


Figura 4: Derramamiento de aluminio en las bases de las placas

- *La deformación del talón.* Al comenzar el arranque, el agua de refrigeración incide directamente sobre el falso fondo y la superficie de la placa, produciéndose un fuerte incremento del calor evacuado. Estos gradientes térmicos provocan fuertes tensiones térmicas que, combinadas con el aumento de la densidad del aluminio en el proceso de solidificación, se traducen en una deformación de las paredes laterales y del pie de la placa, conocida como deformación del talón y caracterizada por un levantamiento del fondo del bloque de metal, más pronunciado en las caras menores del mismo (ver Figura 3). El hueco entre el pie de la placa y el falso fondo puede llegar a medir 50 mm de altura, dependiendo, entre otros factores, del tamaño de la lingotera, la velocidad de colada, el tipo de aleación y la geometría del falso fondo.

Experimentalmente puede observarse que la velocidad de esta deformación alcanza un máximo poco después de que el agua de refrigeración entre en contacto con la placa y decrece muy rápidamente hasta un valor próximo a cero, de forma que, después de unos segundos de iniciarse la deformación, el proceso ha concluido, independientemente del tamaño de la placa.

La simulación numérica de este fenómeno permite mejorar el funcionamiento de la colada y así prevenir la aparición de complicaciones, como son:

- La disminución de la estabilidad de la placa, debida a la reducción de la superficie de apoyo del bloque de aluminio sobre el falso fondo.
- La elevación del nivel de metal líquido, que podría provocar un derrame del mismo (ver Figura 4).
- El recalentamiento de la corteza sólida encargada de contener el aluminio aún líquido; esto impediría su posterior extrusión o laminación o, en el peor de los casos, podría provocar la aparición de fisuras por donde se derramaría el líquido.
- La necesidad de cortar y refundir el aluminio de la zona deformada (ver Figura 4).

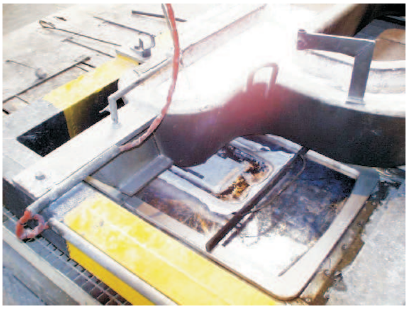


Figura 5: Vista superior de una colada clásica.

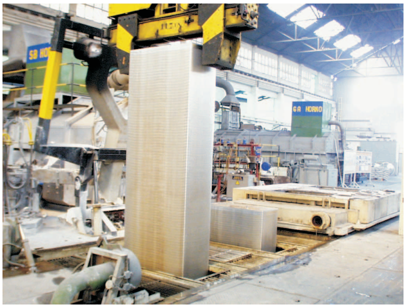


Figura 6: Solidificado el lingote, se desplaza la mesa de colada y se extrae.

- *La contracción de las paredes laterales de la placa.* En la fase posterior al arranque, el aluminio sólido se contrae hacia el interior de la piscina líquida (ver Figuras 5 y 6). Este estrechamiento de la placa puede llegar a ser de 40 mm. Así, si se emplea un molde o inductor rectangular, las paredes del bloque resultante son cóncavas. Esto representa un problema para comercializar las placas, pues el tamaño y el formato de las mismas está preestablecido.

3 Modelización matemática de las deformaciones termomecánicas de la colada

3.1 Dominio de cálculo

Sea $[0, t_f]$ el intervalo de tiempo en el que se realiza la modelización mecánica; se denota por $\overline{\Omega}(t)$ la región tridimensional ocupada por la placa en el instante $t \in (0, t_f]$. $\overline{\Omega}(0)$ representa el aluminio dentro del falso fondo antes de que éste empiece a descender.

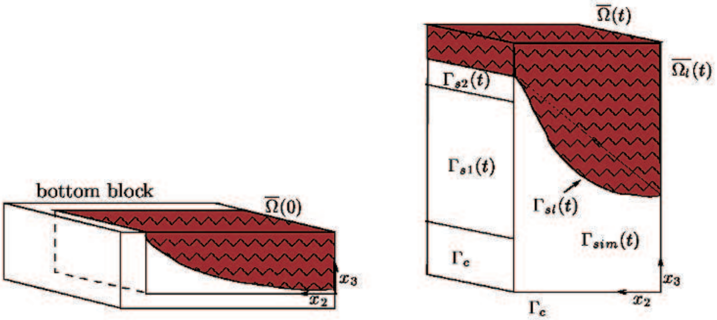


Figura 7: Dominio de cálculo.

El dominio considerado en la modelización mecánica en el instante t es la parte de la placa ya solidificada, $\overline{\Omega_s}(t)$, y se corresponde con el dominio no sombreado de la Figura 7. El molde y el falso fondo no forman parte del dominio de cálculo. Suponemos que $\Omega_s(t)$ es un conjunto abierto, acotado y conexo de $\mathbb{R}^3$ con frontera suficientemente regular.

El origen de coordenadas se sitúa en el centro de la cara superior del falso fondo; el descenso de éste, puesto que conlleva sólo un movimiento de traslación, se modela con un desplazamiento hacia arriba del molde. Debido a la simetría sólo se simula la cuarta parte de la placa.

3.2 Campo de temperaturas

El campo de temperaturas $T(x,t)$ en cada punto $x \in \Omega(t)$ y en cada instante $t \in (0, t_f]$ se calcula antes e independientemente de la deformación mecánica. Para ello, se usa el código desarrollado por Bermúdez y Otero [13], donde la evolución térmica se predice resolviendo un modelo de conducción de calor. El líquido, la zona pastosa y el sólido se tratan como un cuerpo continuo que se mueve en la dirección x_3 (a la velocidad de la colada). El enfriamiento de la superficie de la placa se representa por coeficientes de transferencia de calor que dependen del tiempo y del espacio. El problema térmico se resuelve utilizando algoritmos iterativos de multiplicadores de Lagrange para tratar la no linealidad del término de entalpía; en cada iteración se utiliza un método de elementos finitos de Lagrange de grado uno, para la discretización en espacio, combinado con el método de las características para la discretización en tiempo. Para obtener una descripción más precisa del modelo matemático pueden consultarse los trabajos de Ciavaldini [14], Otero [30] o El-Raghy *et al.* [31]. Las isotermas, transcurridos 1000 s de colada, se muestran en la Figura 8.

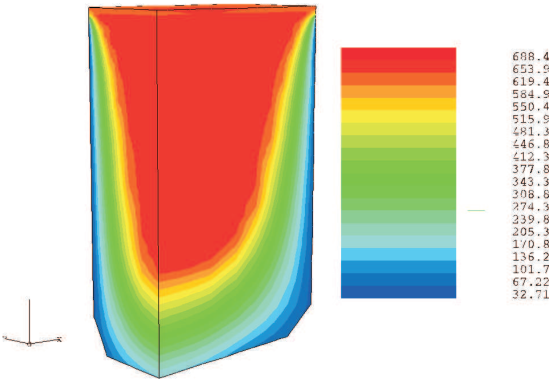


Figura 8: Isotermas.

Las Figuras 9-10 muestran un buen ajuste entre las temperaturas medidas y calculadas en puntos próximos al falso fondo y en puntos de una misma sección transversal de la fase cuasi-estacionaria, respectivamente. Estas representaciones muestran la existencia de grandes gradientes térmicos debidos al contacto directo de la placa con los chorros del agua de enfriamiento. Esto obligará a considerar en el modelo mecánico mallas adaptadas a esos gradientes.

3.3 Modelo mecánico

El problema mecánico consiste en determinar el campo de desplazamientos $\mathbf{u}(x,t)$ y el tensor de tensiones $\boldsymbol{\sigma}(x,t)$ en cada punto $(x,t) \in \Omega_s(t) \times (0, t_f]$.

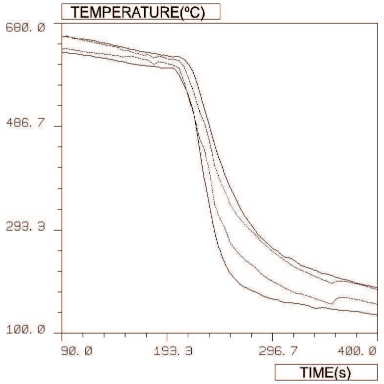


Figura 9: Evolución de la temperatura con el tiempo en dos puntos próximos al falso fondo.
(—- medida y $\cdots$ calculada).

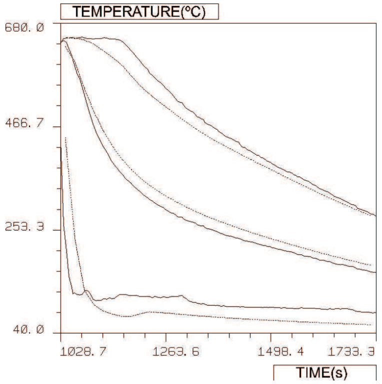


Figura 10: Evolución de la temperatura en tres puntos del eje corto de la sección transversal situada a un metro del pie de la placa.

- *Condiciones de contorno*

La frontera $\Gamma(t)$ de $\Omega_s(t)$ se descompone en

$$\Gamma(t) = \overline{\Gamma}_{sl}(t) \cup \overline{\Gamma}_c \cup \overline{\Gamma}_s(t) \cup \overline{\Gamma}_{sim}(t),$$

(ver Figura 7), donde $\Gamma_{sl}(t)$ es la superficie superior de la placa solidificada, definida por la isoterma correspondiente a la temperatura del *liquidus*, T_l; Γ_c denota la frontera entre la placa y el falso fondo; $\Gamma_{sim}(t)$ es la frontera lateral del dominio correspondiente a los planos de simetría $[x_1 = 0]$ y $[x_2 = 0]$; $\Gamma_s(t)$ corresponde a las fronteras laterales exteriores.
Sobre $\Gamma_{sl}(t)$ se considera la presión correspondiente al peso ejercido por el aluminio líquido sobre el ya solidificado:

$$\boldsymbol{\sigma}\mathbf{n} = p_r\mathbf{n}, \text{ sobre } \Gamma_{sl}(t),$$

donde $p_r(x,t) = \rho(T)g(x_3 - h(t))$ es la presión metalostática, siendo $h(t)$ la longitud de la placa en el instante t, ρ la densidad de masa de la aleación, que se supone dependiente de la temperatura y g la aceleración de la gravedad; $\mathbf{n}$ denota el vector normal unitario exterior a $\Gamma(t)$.
Sobre Γ_c, siguiendo los trabajos de Drezet y Plata [17] y Mariaux *et al.* [26], se considera una condición de contacto sin fricción sobre el falso fondo, que se supone rígido:

$$\boldsymbol{\sigma}_t = \mathbf{0}, \;\; \sigma_n \leq 0, \;\; u_n \leq 0, \;\; \sigma_n u_n = 0 \text{ sobre } \Gamma_c,$$

donde u_n es la componente normal del vector desplazamiento $\mathbf{u}$, y σ_n, $\boldsymbol{\sigma}_t$ son, respectivamente, la componente normal y tangencial del vector de esfuerzos $\boldsymbol{\sigma}\mathbf{n}$.

Sobre $\Gamma_{sim}(t)$ se consideran las condiciones de simetría usuales

$$\boldsymbol{\sigma}_t = \mathbf{0}, \ \ u_n = 0 \text{ sobre } \Gamma_{sim}(t).$$

Por último, en $\Gamma_s(t)$ se distinguen dos zonas: $\Gamma_s(t) = \Gamma_{s1}(t) \cup \Gamma_{s2}(t)$, donde, si T_s denota la temperatura del solidus,

$$\Gamma_{s2}(t) = \{x \in \Gamma_s(t);\, T_s < T(x,t) < T_l\},$$

representa las caras exteriores de la zona pastosa, que deben estar confinadas por el molde en el caso de la colada clásica o por el campo magnético en la electromagnética, es decir:

$$\boldsymbol{\sigma}_t = \mathbf{0}, \ u_n = 0 \text{ sobre } \Gamma_{s2}(t).$$

En el resto de la frontera lateral exterior, $\Gamma_{s1}(t)$, se supone que las fuerzas de superficie son nulas:

$$\boldsymbol{\sigma}\mathbf{n} = \mathbf{0} \text{ sobre } \Gamma_{s1}(t).$$

- *Ecuaciones de equilibrio*

Bajo la hipótesis de pequeños desplazamientos, la deformación de la placa se rige por la ecuación local de equilibrio:

$$-\mathrm{div}(\boldsymbol{\sigma}) = \mathbf{f} \text{ en } \Omega_s(t),$$

donde $\mathbf{f} = (0, 0, -\rho(T)g)$ es el campo de fuerzas gravitacional.

- *Ley de comportamiento*

La elección de la ley de comportamiento fue sugerida por los trabajos de Drezet *et al.* [18], Kristiansson-Zetterlund [24] y Mariaux *et al.* [26]. En ellos, se supone que el aluminio es un material termoviscoelástico y, por tanto, el tensor de velocidad de deformación es la suma de las componentes viscoelástica y térmica:

$$\dot{\boldsymbol{\varepsilon}} = \dot{\boldsymbol{\varepsilon}}^{\mathbf{v}} + \dot{\varepsilon}^{\mathbf{T}}\mathbf{I},$$

donde el punto indica diferenciación respecto al tiempo, $\mathbf{I}$ el tensor identidad y $\boldsymbol{\varepsilon}$ el tensor de deformación linealizado; además:

- $\varepsilon^{\mathbf{T}}$ denota la expansión térmica, relacionada con la temperatura por la expresión

$$\varepsilon^{\mathbf{T}} = \int_{T_l}^{T} \alpha_s(r)dr,$$

siendo α_s un coeficiente que incluye los cambios de volumen debidos a posibles cambios de fase (ver [1]).

- $\boldsymbol{\varepsilon}^{\mathbf{v}}$ es el tensor viscoelástico de Maxwell-Norton, que se puede expresar como suma de una parte elástica lineal, $\boldsymbol{\varepsilon}^{\mathbf{e}}$, y una parte viscoplástica, $\boldsymbol{\varepsilon}^{\mathbf{p}}$.

  El tensor elástico viene dado por la ley de Hooke con coeficientes dependientes de la temperatura,

$$\boldsymbol{\varepsilon}^{\mathbf{e}} = \Lambda_s(T)\boldsymbol{\sigma},$$

  donde Λ_s está definido en la forma usual en términos del módulo de Young E y del coeficiente de Poisson ν.

  El tensor viscoplástico está relacionado con el tensor de tensiones por la ley de Norton-Hoff clásica

$$\dot{\boldsymbol{\varepsilon}}^{\mathbf{p}}(\mathbf{u}) = \nabla\Phi_q(\boldsymbol{\sigma}),$$

  donde $\nabla\Phi_q(\boldsymbol{\sigma})$ denota el gradiente de la función polar del potencial de disipación

$$\Phi_q(\boldsymbol{\sigma}) = \frac{\theta_o}{q}|\boldsymbol{\sigma}^D|^q, \tag{1}$$

  siendo θ_o y q parámetros que dependen de la aleación y $\boldsymbol{\sigma}^D$ el desviatorio del tensor $\boldsymbol{\sigma}$, definido por $\boldsymbol{\sigma}^D = \boldsymbol{\sigma} - \frac{1}{3}\mathrm{tr}(\boldsymbol{\sigma})\mathbf{I}$.

  Si se realizan experimentos de fluencia con un metal a distintas temperaturas, se observa que la tensión generada por una deformación viscoplástica determinada disminuye al aumentar la temperatura (ver Lemaitre-Chaboche [25]). Una forma de tener en cuenta esta dependencia es considerando la fluencia como un fenómeno térmicamente activo; para ello, se introduce en la expresión (1) un coeficiente tipo Arrhenius:

$$e^{\frac{-G}{R(T+273)}},$$

  siendo G la energía de activación del proceso y R la constante de los gases. La ley viscoplástica resultante es, por tanto,

$$\dot{\boldsymbol{\varepsilon}^{\mathbf{p}}} = \kappa(T) \mid \boldsymbol{\sigma}^D \mid^{q-2} \boldsymbol{\sigma}^D,$$

  donde

$$\kappa(T) = \theta_o e^{\frac{-G}{R(T+273)}}.$$

  En la práctica la ley de comportamiento es no lineal, pues para las aleaciones de aluminio más usuales $q > 2$.

En consecuencia, la ecuación constitutiva es

$$\dot{\boldsymbol{\varepsilon}}(\mathbf{u}) = \dot{\widehat{\Lambda_s(T)\boldsymbol{\sigma}}} + \kappa(T) \mid \boldsymbol{\sigma}^D \mid^{q-2} \boldsymbol{\sigma}^D + \alpha(T)\dot{T}\mathbf{I} \text{ en } \Omega_s(t).$$

- *Condiciones iniciales*

$$\mathbf{u}(0) = \mathbf{u}_0, \ \ \boldsymbol{\sigma}(0) = \boldsymbol{\sigma}_0 \text{ en } \Omega_s(0).$$

En resumen, para modelar el comportamiento mecánico durante el proceso de solidificación, se debe resolver el siguiente problema de evolución cuasiestática asociado a un sólido tridimensional viscoelástico con contacto:

$$-\mathrm{div}(\boldsymbol{\sigma}) = \mathbf{f}, \text{en } \Omega_s(t), \tag{2}$$

$$\boldsymbol{\sigma}\mathbf{n} = p_r\mathbf{n}, \text{sobre } \Gamma_{sl}(t), \tag{3}$$

$$\boldsymbol{\sigma}\mathbf{n} = \mathbf{0}, \text{sobre } \Gamma_{s1}(t), \tag{4}$$

$$\boldsymbol{\sigma}_t = \mathbf{0},\ u_n = 0, \text{sobre } \Gamma_{s2}(t) \cup \Gamma_{sim}(t), \tag{5}$$

$$\boldsymbol{\sigma}_t = \mathbf{0},\ u_n \leq 0,\ \sigma_n \leq 0,\ \sigma_n u_n = 0, \text{sobre } \Gamma_c, \tag{6}$$

$$\dot{\boldsymbol{\varepsilon}}(\mathbf{u}) = \dot{\widehat{\Lambda_s(T)\boldsymbol{\sigma}}} + \kappa(T) \mid \boldsymbol{\sigma}^D \mid^{q-2} \boldsymbol{\sigma}^D + \alpha_s(T)\dot{T}\mathbf{I}, \text{en } \Omega_s(t), \tag{7}$$

$$\mathbf{u}(0) = \mathbf{u}_0, \quad \boldsymbol{\sigma}(0) = \boldsymbol{\sigma}_0, \text{en } \Omega_s(0), \tag{8}$$

con $t \in (0, t_f]$.

4 Formulación débil

Se consideran los siguientes espacios funcionales de Banach, estudiados por Geymonat y Suquet [21]:

$$\begin{aligned}
\mathbf{V}^p(t) &= \{\mathbf{v} \in [W^{1,p}(\Omega_s(t))]^3;\ \mathrm{div}(\mathbf{v}) = \mathrm{tr}(\boldsymbol{\varepsilon}(\mathbf{v})) \in L^2(\Omega_s(t))\},\\
\mathbf{X}^q(t) &= \{\boldsymbol{\xi} = (\xi_{ij});\ \xi_{ij} = \xi_{ji},\ \xi_{ij}^D \in L^q(\Omega_s(t)),\ \mathrm{tr}(\boldsymbol{\xi}) \in L^2(\Omega_s(t))\},\\
\mathbf{H}^q(t) &= \{\boldsymbol{\xi} \in \mathbf{X}^q(t);\ \mathrm{div}(\boldsymbol{\xi}) \in [L^q(\Omega_s(t))]^3\},
\end{aligned} \tag{9}$$

donde q es el exponente de la ley viscoplástica y $p = q/(q-1)$ su exponente conjugado. Los conjuntos de desplazamientos y tensiones admisibles a considerar son:

$$\begin{aligned}
\mathbf{U}_{ad}^p(t) &= \{\mathbf{v} \in \mathbf{V}^p(t);\ v_n = 0 \text{ en } \Gamma_{sim}(t) \cup \Gamma_{s2}(t);\ v_n \leq 0 \text{ en } \Gamma_c\}, \qquad (10)\\
\mathbf{H}_{ad}^q(t) &= \{\boldsymbol{\tau} \in \mathbf{H}^q(t);\ -\mathrm{div}(\boldsymbol{\tau}) = \mathbf{f}(t) \text{ en } \Omega_s(t),\ \boldsymbol{\tau}\mathbf{n} = \mathbf{0} \text{ en } \Gamma_{s1}(t),\\
&\quad \boldsymbol{\tau}\mathbf{n} = p_r(t)\mathbf{n} \text{ en } \Gamma_{sl}(t),\ \boldsymbol{\tau}_t = \mathbf{0} \text{ en } \Gamma_c \cup \Gamma_{sim}(t) \cup \Gamma_{s2}(t),\ \tau_n \leq 0 \text{ en } \Gamma_c\}.
\end{aligned}$$

En [4] se introduce la siguiente formulación débil del problema (2)-(8) como inecuación variacional:

Encontrar $\mathbf{u} \in W^{1,\infty}(0, t_f; \mathbf{U}_{ad}^p(t))$ y $\boldsymbol{\sigma} \in W^{1,\infty}(0, t_f; \mathbf{H}^q(t))$ verificando *c.p.d.* en $(0, t_f)$:

$$\begin{aligned}
\int_{\Omega_s(t)} \boldsymbol{\sigma}(t) : \boldsymbol{\varepsilon}(\mathbf{v} - \mathbf{u}(t))\, dx &\geq \int_{\Omega_s(t)} \mathbf{f}(t) \cdot (\mathbf{v} - \mathbf{u}(t))\, dx\\
&+ \int_{\Gamma_{sl}(t)} p_r(t)\mathbf{n} \cdot (\mathbf{v} - \mathbf{u}(t))\, d\gamma, \forall \mathbf{v} \in \mathbf{U}_{ad}^p(t),
\end{aligned} \tag{11}$$

$$\dot{\boldsymbol{\varepsilon}}(\mathbf{u})(t) = (\dot{\widehat{\Lambda_s(T)\boldsymbol{\sigma}}})(t) + (\kappa(T) \mid \boldsymbol{\sigma}^D \mid^{q-2} \boldsymbol{\sigma}^D)(t) + \left(\alpha_s(T)\dot{T}\right)(t)\mathbf{I}, \tag{12}$$

$$\mathbf{u}(0) = \mathbf{u}_0,\ \ \boldsymbol{\sigma}(0) = \boldsymbol{\sigma}_0,\ \text{en } \Omega_s(0). \tag{13}$$

Las hipótesis consideradas sobre los datos del problema (11)-(13) son:

(H1) El campo de temperaturas $T \in W^{1,\infty}(0, t_f; L^\infty(\Omega_s(t)))$.

(H2) La presión metalostática $p_r \in L^\infty(0, t_f; W^{-\frac{1}{q},q}(\Gamma(t)) \cap L^q(\Gamma_{sl}(t)))$, las fuerzas de volumen $\mathbf{f} \in L^\infty(0, t_f; [L^q(\Omega_s(t)]^3)$ y la densidad es tal que $\rho(s) \geq \rho_0 > 0$, $\forall s \in \mathbb{R}$.

(H3) El tensor de elasticidad $\Lambda_s \in [W^{1,\infty}(\mathbb{R})]^{81}$ es tal que:

- $(\Lambda_s)_{ijkh} = (\Lambda_s)_{khij} = (\Lambda_s)_{jikh}$.
- $\exists \alpha > 0$ tal que $\Lambda_s \boldsymbol{\tau} : \boldsymbol{\tau} \geq \alpha \mid \boldsymbol{\tau} \mid^2$, $\forall \boldsymbol{\tau} \in \mathcal{S}_3$, *c.p.d.* en $\mathbb{R}$, siendo $\mathcal{S}_3$ el espacio de tensores simétricos de orden dos sobre $\mathbb{R}^3$.

(H4) El coeficiente de expansión térmica $\alpha_s \in L^\infty(\mathbb{R})$ y el coeficiente de la ley viscoplástica $\kappa \in L^\infty(\mathbb{R})$.

(H5) El exponente de la ley de Norton-Hoff verifica $q \geq 2$ y, por tanto, $1 < p \leq 2$.

(H6) Las condiciones iniciales son tales que $\boldsymbol{\sigma}_0 \in \mathbf{H}^q_{ad}(0)$, $\mathbf{u}_0 \in \mathbf{U}^p_{ad}(0)$ y $(u_0)_n(\sigma_0)_n = 0$ sobre Γ_c.

4.1 Existencia y regularidad de solución de un submodelo viscoelástico con contacto

Con vistas a estudiar la existencia de solución del modelo mecánico, debido a la complejidad del mismo, se considera en este apartado el submodelo obtenido al suponer que el dominio no depende del tiempo y eliminar la componente térmica de la ley de comportamiento; el problema resultante consiste en el análisis de las deformaciones de un sólido viscoelástico con condición de contacto.

La frontera de Ω_s se divide en tres partes disjuntas:

$$\Gamma = \overline{\Gamma}_D \cup \overline{\Gamma}_N \cup \overline{\Gamma}_C,$$

con $\text{mes}(\Gamma_D) > 0$ (ver Figura 11). Γ_D denota el centro de la base de la placa, donde se impone una condición Dirichlet homogénea; Γ_C denota la parte restante de la frontera entre la placa y el falso fondo, donde se considera una condición de contacto de Signorini; Γ_N denota las superficies lateral y superior, sobre las que actúan fuerzas de superficie de densidad $\mathbf{h}$.

El problema que se considera es encontrar $\mathbf{u}$ y $\boldsymbol{\sigma}$ verificando:

$$-\text{div}(\boldsymbol{\sigma}) = \mathbf{f} \quad \text{en} \quad (0, t_f] \times \Omega_s, \tag{14}$$

$$\boldsymbol{\sigma}\mathbf{n} = \mathbf{h} \quad \text{sobre} \quad (0, t_f] \times \Gamma_N, \tag{15}$$

$$\mathbf{u} = \mathbf{0} \quad \text{sobre} \quad (0, t_f] \times \Gamma_D, \tag{16}$$

$$u_n \leq 0,\ \sigma_n \leq 0,\ \boldsymbol{\sigma}_t = \mathbf{0},\ \sigma_n u_n = 0 \quad \text{sobre} \quad (0, t_f] \times \Gamma_C, \tag{17}$$

$$\boldsymbol{\varepsilon}(\dot{\mathbf{u}}) - \Lambda_s \dot{\boldsymbol{\sigma}} = \nabla \Phi_q(\boldsymbol{\sigma}) \quad \text{en} \quad (0, t_f] \times \Omega_s, \tag{18}$$

$$\mathbf{u}(0) = \mathbf{u}_0,\ \boldsymbol{\sigma}(0) = \boldsymbol{\sigma}_0 \quad \text{en} \quad \Omega_s. \tag{19}$$

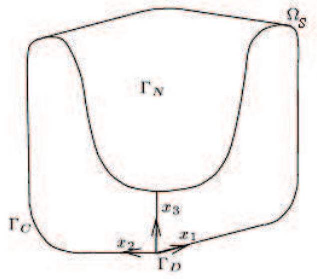


Figura 11: Un cuarto del dominio Ω_s considerado en el estudio de existencia de solución.

En Sofonea [34], Han-Sofonea [22] y Rochdi-Sofonea [32] se demuestra la existencia de solución de algunos problemas viscoelásticos y viscoplásticos con condición de contacto. En todos ellos la función de plasticidad considerada cumple una condición de Lipschitz, no verificada por la ley de Norton-Hoff; por ello, las técnicas de punto fijo aplicadas a esas leyes de comportamiento no son trasladables al estudio de existencia de solución del problema (14)-(19). En la bibliografía también se pueden encontrar resultados de existencia de solución para materiales viscoelásticos con una ley plástica tipo Norton-Hoff, pero asociados a las condiciones de contorno clásicas desplazamiento-tracción (ver Bensoussan-Frehse [10], Djaoua-Suquet [16], Friaâ [20] y Le Tallec [35]). La metodología empleada en estos trabajos no se adapta fácilmente a condiciones de contacto tipo Signorini para modelar el despegue del talón; la dificultad surge en que la formulación variacional natural para problemas viscoelásticos se realiza en velocidades, mientras que la condición de contacto de Signorini se expresa en desplazamientos. En Barral-Quintela [9] se demuestra el siguiente resultado para el problema (14)-(19).

Teorema 1 *Bajo las hipótesis:*

(HE1) *el tensor de elasticidad Λ_s es independiente del tiempo, simétrico y satisface*

$$\Lambda_s \in [L^\infty(\Omega_s)]^{81} \text{ y } \exists \alpha > 0 \text{ tal que } \Lambda_s \boldsymbol{\tau} : \boldsymbol{\tau} \geq \alpha |\boldsymbol{\tau}|^2_{\mathcal{S}_3}, \ \forall \boldsymbol{\tau} \in \mathcal{S}_3, \ \text{c.p.d. en } \Omega_s;$$

(HE2) *las fuerzas externas verifican*

$$\mathbf{f} \in W^{2,\infty}(0, t_f; [L^q(\Omega_s)]^3) \text{ y } \mathbf{h} \in W^{2,\infty}(0, t_f; [W^{-\frac{1}{q},q}(\Gamma)]^3 \cap [L^q(\Gamma_N)]^3);$$

(HE3) *la tensión y el desplazamiento iniciales satisfacen las condiciones naturales de compatibilidad;*

existe una solución $(\mathbf{u}, \boldsymbol{\sigma})$ del problema (14)-(19) tal que

$$\mathbf{u} \in W^{1,2}(0, t_f; \mathbf{V}_0^p) \text{ y } \boldsymbol{\sigma} \in W^{1,2}(0, t_f; \mathbf{X}^2) \cap L^\infty(0, t_f; \mathbf{X}^q).$$

La demostración consiste, en primer lugar, en formular el problema mediante una inecuación variacional análoga a (11)-(13). Posteriormente, se discretiza el problema en tiempo utilizando un esquema de Euler implícito. Para demostrar la convergencia de las soluciones de los problemas discretizados a la del problema inicial se adaptan las técnicas de monotonía utilizadas en [16] a la condición de contacto. Debido a la poca regularidad de los desplazamientos admisibles, para obtener la ley de comportamiento en el límite son necesarias técnicas de compacidad por compensación (ver [28]).

Nota 1 *En la simulación de coladas de aluminio* $\mathbf{f}$ *representa las fuerzas debidas a la gravedad y* $\mathbf{h}$ *coincide con la presión metalostática en la interfase y es nula en las caras exteriores, por tanto, se puede comprobar que cumplen la hipótesis* **(HE2)** *sobre las fuerzas aplicadas. Además, en la práctica,* $\mathbf{u}_0 = \mathbf{0}$ *y* $\boldsymbol{\sigma}_0 = \mathbf{0}$*, consecuentemente,* **(HE3)** *también se verifica.*

En Frehse y Málek [19] se estudia la regularidad de solución de un problema estático para materiales elastoplásticos de Norton-Hoff, obteniéndose que, bajo ciertas hipótesis sobre las fuerzas de volumen, $\boldsymbol{\sigma} \in [H^1_{\mathrm{loc}}(\Omega_s)]^9$. Esta demostración se puede adaptar al problema evolutivo aquí considerado para demostrar el siguiente resultado.

Teorema 2 *Si* $\mathbf{f} \in W^{2,\infty}(0, t_f; [W^{2,q}(\Omega_s)]^3)$ *y* $\boldsymbol{\sigma}_0 \in [H^1_{\mathrm{loc}}(\Omega_s)]^9$*, entonces*

$$\boldsymbol{\sigma}(t) \in [H^1_{\mathrm{loc}}(\Omega_s)]^9, \; t \in (0, t_f].$$

Además, en Bensoussan y Frehse [10] se demuestra que, para una ley de Maxwell-Norton formulada en velocidades, $\boldsymbol{\sigma} \in [H^2_{\mathrm{loc}}(\Omega_s)]^9$. No es difícil comprobar que la demostración es válida también para la ley formulada en desplazamientos.

5 Solución numérica del modelo termoviscoelástico con contacto

Si se analiza la formulación variacional del modelo completo (11)-(13), las principales dificultades que plantea su resolución numérica son:

- *La implementación de la presión metalostática sobre la frontera* $\Gamma_{sl}(t)$, que es la frontera libre del problema térmico y, por tanto, debe calcularse numéricamente en cada instante de tiempo. Para implementar la condición de presión sobre esta frontera caben, al menos, tres posibilidades: implementar la presión directamente sobre la frontera obtenida numéricamente, resolver también la fase líquida definiendo un problema de interacción fluido estructura, o definir una perturbación del problema mecánico resolviendo, en la fase líquida, un problema ficticio sencillo que permita aproximar la presión metalostática del líquido sobre la interfase. La primera posibilidad se ha ensayado sin éxito, debido a las singularidades que presenta la aproximación de la superficie

$\Gamma_{sl}(t)$, obtenida numéricamente a partir de la resolución tridimensional del problema térmico, lo que impide recuperar correctamente el peso del líquido. La segunda opción se ha desechado por su elevado coste computacional al tratarse de una simulación tridimensional. Finalmente, inspirados por los trabajos de Hannart-Cialti-Schalkwijk [23], se ha optado por utilizar un método de dominio ficticio que se justifica mediante un análisis asintótico.

- *La no linealidad de la ley viscoplástica.* Para tratar esta no linealidad se utilizan las técnicas de operadores maximales monótonos desarrolladas por Bermúdez y Moreno [11] para la resolución de inecuaciones variacionales. El multiplicador asociado es un punto fijo de la ecuación

$$\mathbf{q}^{n+1} = (\nabla\Phi_q)_{\lambda_p}\left(\left(\boldsymbol{\sigma}^{\boldsymbol{D}}\right)^{n+1} + \lambda_p\mathbf{q}^{n+1}\right), \lambda_p > 0,$$

  que se aproxima mediante un método de Newton generalizado.

- *El tratamiento de la condición de contacto con el falso fondo.* Para implementar esta condición de contorno se aplican de nuevo las técnicas de Bermúdez y Moreno [11]. El multiplicador asociado en este caso es un punto fijo de una ecuación multívoca no lineal que se calcula combinando una estrategia de penalización con un método tipo Newton.

- *La implementación de la condición de confinamiento* sobre $\Gamma_{s2}(t)$. Debido a que la sección de las placas de aluminio no es rectangular (ver Figura 5), la condición $u_n = 0$ en $\Gamma_{s2}(t)$ acopla las dos primeras componentes del desplazamiento; por ello, esta condición no se puede tratar como Dirichlet y se implementa mediante un método de penalización.

5.1 Implementación de la presión metalostática: Problema perturbado

Para implementar el término integral sobre $\Gamma_{sl}(t)$ de la formulación débil (11) se considera la siguiente aproximación asintótica: se incluye la fase líquida en el dominio de cálculo, y se supone que el líquido es un dominio ficticio constituido por un material muy elástico sometido a las fuerzas de gravedad. Para ello, se considera que los coeficientes de Lamé en la zona líquida -denotada por $\Omega_l(t)$- dependen de un pequeño parámetro ϵ de la siguiente forma:

$$\boldsymbol{\sigma}^{\epsilon} = (\Lambda^{\epsilon})^{-1}(T)\boldsymbol{\varepsilon}^{\mathbf{e}} = \begin{cases} \lambda(T)\mathrm{tr}(\boldsymbol{\varepsilon}^{\mathbf{e}})\mathbf{I} + 2\mu(T)\boldsymbol{\varepsilon}^{\mathbf{e}} & \text{en} \quad \Omega_s(t), \\ \epsilon^{\beta}\overline{\lambda}\mathrm{tr}(\boldsymbol{\varepsilon}^{\mathbf{e}})\mathbf{I} + 2\epsilon^{\alpha}\overline{\mu}\boldsymbol{\varepsilon}^{\mathbf{e}} & \text{en} \quad \Omega_l(t), \end{cases} \tag{20}$$

donde $\overline{\lambda}$, $\overline{\mu}$, α y β son números reales independientes de ϵ. Sobre $\Omega_l(t)$ no se consideran deformaciones plásticas ni térmicas, por tanto, se extiende la ley viscoplástica y el coeficiente de expansión térmica por cero; se denota por $Y(\boldsymbol{\sigma}^{\epsilon D})$ y $\alpha_T(T)$, respectivamente, estas leyes extendidas a todo el dominio

$\Omega(t)$. Se obtiene así la siguiente formulación débil del problema extendido, que se denominará perturbado:
Encontrar $\mathbf{u}^\epsilon \in W^{1,\infty}(0,t_f;\mathbf{U}^p_{ad}(t))$ y $\boldsymbol{\sigma}^\epsilon \in W^{1,\infty}(0,t_f;\mathbf{H}^q(t))$ verificando *c.p.d.* en $(0,t_f)$:

$$\int_{\Omega(t)} \boldsymbol{\sigma}^\epsilon(t) : \boldsymbol{\varepsilon}(\mathbf{v}-\mathbf{u}^\epsilon(t))\,dx \geq \int_{\Omega(t)} \mathbf{f}(t)\cdot(\mathbf{v}-\mathbf{u}^\epsilon(t))\,dx,\ \forall \mathbf{v}\in \mathbf{U}^p_{ad}(t), \tag{21}$$

$$\dot{\boldsymbol{\varepsilon}}(\mathbf{u}^\epsilon)(t) = (\dot{\widehat{\Lambda^\epsilon(T)\boldsymbol{\sigma}^\epsilon}})(t) + Y(\boldsymbol{\sigma}^{\epsilon D})(t) + \left(\alpha_T(T)\dot{T}\right)(t)\mathbf{I}, \tag{22}$$

$$\mathbf{u}^\epsilon(0) = \mathbf{u}_0,\ \ \boldsymbol{\sigma}^\epsilon(0) = \boldsymbol{\sigma}_0,\ \ \text{en } \Omega(0). \tag{23}$$

Obsérvese que esta inecuación está definida sobre todo el dominio $\Omega(t)$ y la integral sobre $\Gamma_{sl}(t)$ ha desaparecido (ver [7]). Los espacios $\mathbf{H}^q(t)$, $\mathbf{U}^p_{ad}(t)$, definidos en (9), (10) y las hipótesis **(H1) - (H6)** se extienden a $\Omega(t)$ de forma natural. Además, supondremos:

(H7) $\Omega(t) \subset \mathbb{R}^3$ es un conjunto abierto, conexo, acotado, con frontera lipschitziana.

(H8) $\overline{\lambda}$, $\overline{\mu}$ son números reales positivos, independientes de ϵ y $\overline{\mu} > 0$.

5.1.1 Justificación asintótica del problema perturbado

Si bien en algunos trabajos de simulación de tensiones térmicas en coladas de aluminio se utilizaban aproximaciones análogas al método de dominio ficticio introducido en el apartado anterior, no se encontró ninguna justificación matemática de las mismas. Por ello, en [7, 8], se efectuó un análisis asintótico del problema perturbado y, en particular, del comportamiento de su solución cuando el pequeño parámetro ϵ tiende a cero. Con vistas a simplificar los espacios funcionales, el análisis se realizó para un problema elástico independiente del tiempo; además, como la condición de contacto no influye en el comportamiento en la interfase, se sustituyó por una condición Dirichlet homogénea. Las hipótesis sobre los datos se pueden consultar en [7]. Como conclusión de estos trabajos se han obtenido los siguientes resultados:

1. Existe el $\text{lím}_{\epsilon\to 0}\,\mathbf{u}^\epsilon|_{\Omega_s} = \mathbf{u}^0|_{\Omega_s}$ en $[H^1(\Omega_s)]^3$ y, como consecuencia, el $\text{lím}_{\epsilon\to 0}\,\boldsymbol{\sigma}^\epsilon|_{\Omega_s} = \boldsymbol{\sigma}^0|_{\Omega_s}$ en $[L^2(\Omega_s)]^9$, para cualquier valor de $\overline{\lambda}$, $\overline{\mu}$, α y β.

2. Si $\alpha \leq \beta$, $\mathbf{u}^0|_{\Omega_s}$ y $\mathbf{u}$ son solución de dos problemas distintos pero con fuerzas de superficie estáticamente equivalentes; por consiguiente, el principio físico de Saint-Venant (ver Nečas-Hlaváček [29] o Truesdell [36]) establece que $\mathbf{u}^0|_{\Omega_s}$ aproxima la solución $\mathbf{u}$ en puntos suficientemente alejados de $\Gamma_{sl}(t)$.
   El principio de Saint-Venant sólo ha sido justificado matemáticamente en casos particulares. En las simulaciones numéricas se ha comprobado que esta aproximación sólo es aceptable lejos de la zona pastosa. En efecto, si $\alpha < \beta$ las tensiones cortantes sobre la interfase sólido-líquido impiden

imponer correctamente la condición de presión cualesquiera que sean $\bar{\lambda}$ y $\bar{\mu}$. En las Figuras 12 y 13 se representan las tensiones cortantes para distintos valores de $\bar{\lambda}$ y $\bar{\mu}$ sobre un test académico bidimensional. Los errores obtenidos al aproximar la solución $(\mathbf{u}, \boldsymbol{\sigma})$ por $(\mathbf{u}^{\epsilon}, \boldsymbol{\sigma}^{\epsilon})$ se resumen en la Tabla 1 para $\alpha = 1$, $\beta = 2$ y $\bar{\lambda} \gg \bar{\mu}$ ($\bar{\lambda} = 10^4$N/m^2, $\bar{\mu} = 32{,}5$ N/m^2); Ω_c denota el subconjunto de Ω_s formado por la capa de elementos finitos que corresponde a la zona pastosa (zona que se elimina para tener en cuenta el principio de Saint-Venant).

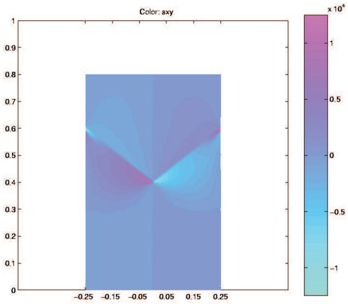


Figura 12: Tensiones cortantes para $\alpha < \beta$ y $\bar{\lambda} = \bar{\mu}$.

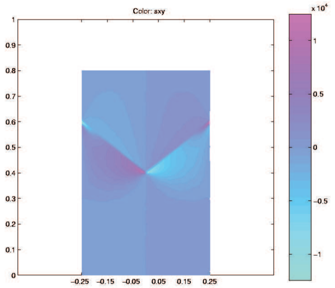


Figura 13: Tensiones cortantes para $\alpha < \beta$ y $\bar{\lambda} \gg \bar{\mu}$.

Cuadro 1: Errores para $\alpha < \beta$ y $\bar{\lambda} \gg \bar{\mu}$.

ϵ	Errores relativos en Ω_s		Errores relativos en $\Omega_s \backslash \Omega_c$	
	Desplazamientos	Tensiones	Desplazamientos	Tensiones
1.e-1	0.0404	0.6180	0.0220	0.0415
1.e-3	0.4986	6.0762	0.2344	0.3603
1.e-5	0.5688	6.6981	0.2780	0.3903

3. Si $\alpha = \beta$ el análisis asintótico muestra que pueden evitarse las tensiones cortantes sobre la interfase tomando $\bar{\lambda} \gg \bar{\mu}$ (ver Tabla 2 y Figuras 14, 15).

4. Si $\alpha > \beta$, $\mathbf{u}^0|_{\Omega_s}$ y $\mathbf{u}$ son solución del mismo problema; en consecuencia, los errores al aproximar $(\mathbf{u}, \boldsymbol{\sigma})$ por $(\mathbf{u}^{\epsilon}, \boldsymbol{\sigma}^{\epsilon})$ son pequeños, incluso en la interfase, para $\bar{\lambda}$ y $\bar{\mu}$ cualesquiera (ver Tabla 3 y Figuras 16, 17).

Como conclusión, para imponer la presión metalostática sobre $\Gamma_{sl}(t)$ se puede considerar el metal líquido como un sólido ficticio, con la ley de comportamiento dada en (20), y la siguiente elección de parámetros:

- ϵ suficientemente pequeño (en la práctica, se toma $\epsilon = O(10^{-3})$).

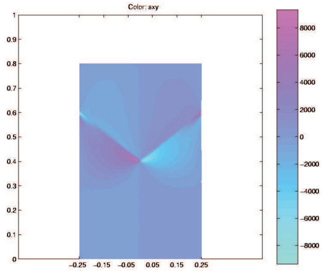


Figura 14: Tensiones cortantes para $\alpha = \beta$ y $\bar{\lambda} = \bar{\mu}$.

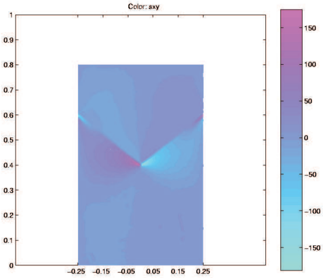


Figura 15: Tensiones cortantes para $\alpha = \beta$ y $\bar{\lambda} \gg \bar{\mu}$.

Cuadro 2: Errores para $\alpha = \beta$ y $\bar{\lambda} \gg \bar{\mu}$.

ϵ	Errores relativos en Ω_s		Errores relativos en $\Omega_s \backslash \Omega_c$	
	Desplazamientos	Tensiones	Desplazamientos	Tensiones
1.e-1	5.3e-3	8.3e-2	2.6e-3	5.8e-3
1.e-3	5.3e-3	8.3e-2	2.6e-3	5.8e-3
1.e-5	5.3e-3	8.3e-2	2.6e-3	5.8e-3

- $\alpha > \beta$ (por ejemplo $\alpha = 2$ y $\beta = 1$).

El ejemplo bidimensional al que se corresponden los resultados numéricos presentados en este apartado se puede consultar en [6].

5.2 Discretización en espacio y tiempo. Algoritmo iterativo

Para resolver numéricamente el problema variacional (21)-(23) se realiza una discretización espacial mediante un método de elementos finitos de Lagrange de grado uno.

Para tratar la desigualdad debida a la condición de contacto de la formulación débil (21) y la no linealidad debida a la ley de comportamiento viscoelástica, se usan técnicas de operadores máximales monótonos; estas técnicas conducen a la obtención de dos multiplicadores, denominados de contacto y plasticidad, respectivamente. Para calcular estos multiplicadores, en una primera etapa se utilizó un método iterativo de punto fijo. La aplicación detallada de esta técnica al problema (21)-(23) es muy laboriosa y puede consultarse en [4, 5]. El algoritmo se validó con varios tests numéricos, cuyos resultados están publicados en [4].

Aunque este algoritmo dio muy buenos resultados en tests académicos, la convergencia se ralentiza considerablemente al realizar la simulación de la cola-

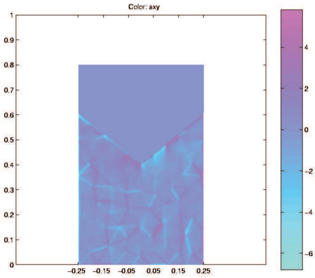


Figura 16: Tensiones cortantes para $\alpha > \beta$ y $\bar{\lambda} = \bar{\mu}$.

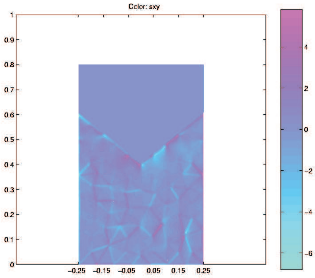


Figura 17: Tensiones cortantes para $\alpha > \beta$ y $\bar{\lambda} \gg \bar{\mu}$.

Cuadro 3: Errores para $\alpha > \beta$ y $\bar{\lambda} \gg \bar{\mu}$.

ϵ	Errores relativos en Ω_s		Errores relativos en $\Omega_s \backslash \Omega_c$	
	Desplazamientos	Tensiones	Desplazamientos	Tensiones
1.e-1	7.3108e-4	1.3e-2	3.6588e-4	1.8e-3
1.e-3	2.4196e-4	4.1e-3	1.8053e-4	1.6e-3
1.e-5	2.3598e-4	4.0e-3	1.8002e-4	1.6e-3

da real. Con vistas a disminuir el tiempo de cálculo necesario para la simulación y poder utilizar discretizaciones más finas, tanto en espacio como en tiempo, se ha modificado el algoritmo numérico siguiendo los trabajos de Moreno [27] y Simo-Taylor [33]. La nueva metodología consiste en utilizar un método tipo Newton para el cálculo de ambos multiplicadores. Un inconveniente de esta técnica aplicada al multiplicador de contacto es la necesidad de modificar y, por tanto, factorizar, la matriz del sistema en cada iteración. Aún así, el número de iteraciones y el tiempo de cálculo de la simulación se reducen considerablemente con respecto al algoritmo anterior. La descripción del nuevo algoritmo aparecerá en [3].

Por último, se discretiza la ley de comportamiento (22) mediante un esquema de Euler implícito; una vez obtenida la expresión de la tensión en cada paso de tiempo se sustituye esta en la formulación variacional para obtener una formulación en desplazamientos (ver [3]).

Con el algoritmo propuesto, en cada paso de tiempo se realiza un proceso iterativo, en el que cada iteración corresponde a un problema de elasticidad lineal. Puesto que la deducción del algoritmo es una tarea laboriosa, simplemente resumimos el algoritmo final, implementado en el ordenador. Su descripción completa puede consultarse en [3].

Algoritmo de cálculo

Se supone conocida la solución hasta el tiempo t^n y se calcula ésta en t^{n+1} mediante el proceso iterativo siguiente:

- El multiplicador de contacto p^{n+1} se inicializa al peso de la placa en t^{n+1}; el multiplicador de plasticidad $\mathbf{q}^{n+1}$ es un tensor que se inicializa a cero. La historia del material hasta el instante actual y los efectos expansivos del metal, debidos al cambio de temperatura en el intervalo $[t^n, t^{n+1}]$, se incorporan mediante un término tensorial $\mathbf{F}^n$ en el segundo miembro de la formulación variacional.

- El desplazamiento en el instante t^{n+1}, $\mathbf{u}^{n+1}$, se calcula como el límite de la sucesión $\{\mathbf{u}_k^{n+1}\}$, cuyo término general es la solución del problema de elasticidad lineal

$$\begin{aligned}
& \int_{\Omega^{n+1}} \left[(\Lambda^{n+1})^{-1}\boldsymbol{\varepsilon}(\mathbf{u}_k^{n+1})\right] : \boldsymbol{\varepsilon}(\mathbf{v})\,dx + \frac{1}{\delta_c}\int_{\left(\Gamma^+_{c,k-1}\right)^{n+1}} \mathbf{u}_k^{n+1}\cdot\mathbf{n}\,\mathbf{v}\cdot\mathbf{n}\,d\gamma \\
+ \quad & \frac{1}{\delta_l}\int_{\Gamma_l^{n+1}\cup\Gamma_{s2}^{n+1}} \mathbf{u}_k^{n+1}\cdot\mathbf{n}\,\mathbf{v}\cdot\mathbf{n}\,d\gamma = \int_{\Omega^{n+1}} \mathbf{f}^{n+1}\cdot\mathbf{v}\,dx \\
+ \quad & \Delta t \int_{\Omega_s^{n+1}} \left[(\Lambda^{n+1})^{-1}\mathbf{q}_{k-1}^{n+1}\right] : \boldsymbol{\varepsilon}(\mathbf{v})\,dx - \int_{\Omega_s^{n+1}} \left[(\Lambda^{n+1})^{-1}\mathbf{F}^n\right] : \boldsymbol{\varepsilon}(\mathbf{v})\,dx, \\
& \forall \mathbf{v} \in (\mathbf{U}_{ad})_h(t^{n+1}), \qquad (24)
\end{aligned}$$

  donde se ha suprimido el superíndice ϵ para simplificar la escritura. En (24) se emplea la siguiente notación:

  - Δt es el parámetro de discretización en tiempo.
  - El segundo término de la formulación es un término de penalización que permite imponer la condición

    $$u_n = 0 \text{ en } \Gamma^+_{c,k-1},$$

    donde

    $$\Gamma^{+(-)}_{c,k} = \{C \in S_h;\ (u_k)_n + \lambda_c p_k > (\leq)\, 0\},$$

    siendo S_h la triangulación inducida por la malla de Ω sobre Γ_c; δ_c es un parámetro suficientemente pequeño y λ_c es un número real positivo.
  - ${\Gamma_{s2}}^{n+1} \cup {\Gamma_l}^{n+1}$ representa la superficie lateral exterior correspondiente a la zona pastosa y líquida, que se supone confinada por el molde, en la colada clásica, o sostenida mediante el campo electromagnético, en la electromagnética. Para implementar esta condición de contorno se considera el término de penalización con δ_l positivo suficientemente pequeño.

  - $(\mathbf{U}_{ad})_h(t^{n+1})$ es la discretización del espacio de desplazamientos admisibles.

- El multiplicador de contacto en cada iteración se obtiene mediante la expresión

$$p_k^{n+1} = \begin{cases} \dfrac{1}{\delta_c}\mathbf{u}_k^{n+1} \cdot \mathbf{n} & \text{sobre } \Gamma_{c,k-1}^{+}, \\ 0 & \text{sobre } \Gamma_{c,k-1}^{-}. \end{cases}$$

- En cada iteración la puesta al día del multiplicador de plasticidad se realiza mediante una expresión explícita sencilla (ver [3]), aunque es necesario calcular la raíz de una ecuación escalar no lineal, para lo que se utiliza un método de Newton-Raphson.

La validación del algoritmo se ha realizado con varios tests numéricos (ver [3]).

6 Simulación de una colada real

En la simulación de coladas reales de aluminio se han distinguido, al igual que en el proceso físico, dos etapas:

1. La *fase de arranque,* durante la que tiene lugar la deformación del talón. Esta deformación se produce en un intervalo de tiempo muy pequeño (ver Figura 19) y en una zona de grandes gradientes térmicos (ver Figura 9); por tanto, si se quiere obtener una buena aproximación de esta deformación, se debe utilizar un paso de tiempo pequeño y una malla suficientemente fina. En la Figura 18 se muestra la deformación del talón 140 segundos después del arranque, usando un tamaño de paso de 10 segundos. En las Figuras 19 y 20 se muestran, respectivamente, el módulo del vector desplazamiento y la norma del tensor de tensiones con respecto al tiempo en el punto marcado en la Figura 18.

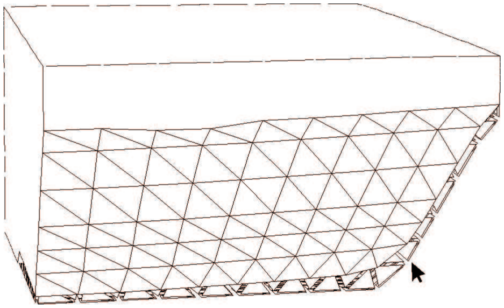


Figura 18: Deformación del talón a los 140 segundos.

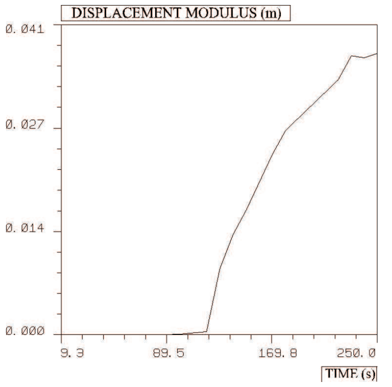


Figura 19: Módulo del desplazamiento respecto al tiempo.

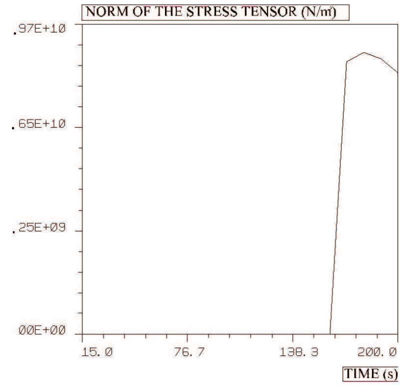


Figura 20: Norma del tensor de tensiones frente al tiempo.

2. La *fase estacionaria.* Puesto que el dominio crece con el tiempo y el intervalo de tiempo a simular hasta alcanzar esta fase es muy grande (1000 s aproximadamente), para reducir la demanda computacional que esto supone se eliminó la condición de contacto al comprobar, mediante experiencias numéricas bidimensionales, que el proceso de deformación del talón no influye en la contracción de las paredes laterales.

   Además, se han observado fuertes gradientes térmicos en zonas próximas a las sometidas a refrigeración por agua (parte superior de la placa y paredes exteriores, ver Figura 8); estas zonas se modifican con el tiempo, pues la placa desciende a medida que comienza su solidificación. Por otra parte, se ha podido constatar numéricamente la gran sensibilidad de las variables mecánicas a estos fuertes gradientes, exigiéndose, por tanto, una malla muy refinada en tales zonas. Para resolver estas dificultades se realiza un remallado en cada nuevo paso de tiempo, de forma que la malla es más fina en las zonas sometidas a refrigeración en ese instante (ver Figura 21).

   La Figura 22 muestra la deformación de la placa en esta fase (la geometría de referencia se muestra por debajo de la deformada). La Figura 23 representa la contracción de las paredes laterales 1000 s después del arranque a una altura de 0.64 m. En la Figura 24 se presenta la geometría del molde/inductor junto a los perfiles de la sección transversal en tres alturas distintas; esto permite visualizar el crecimiento de la contracción con el tiempo -la línea inferior muestra la geometría deseada-.

Los resultados presentados en este apartado se han obtenido con el algoritmo de punto fijo. Actualmente, las autoras trabajan en la aplicación del nuevo algoritmo a la simulación de coladas reales. Esperamos que el nuevo algoritmo permita considerar discretizaciones más finas, tanto en espacio como en tiempo.

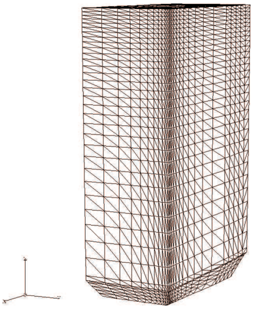


Figura 21: Mallado de la placa en la fase estacionaria adaptado a las zonas de mayor gradiente de temperatura.

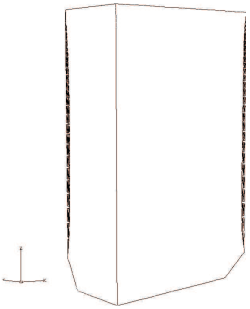


Figura 22: Contracción de la placa.

El algoritmo descrito, junto al correspondiente a la simulación térmica desarrollado por Bermúdez y Otero [13], se recoge en los paquetes de simulación **C2D** (simulación termomecánica bidimensional) y **C3D** (tridimensional) propiedad de la empresa ALCOA-INESPAL S.A., ejecutable en el sistema operativo Windows. En la Figura 25 se muestra la ventana principal del paquete y el menú correspondiente a la introducción de datos de la velocidad de descenso del falso fondo.

Referencias

[1] P. Barral. *Análisis Matemático y Simulación Numérica del Comportamiento Termomecánico de una Colada de Aluminio.* Tesis Doctoral, Universidade de Santiago de Compostela, 2001.

[2] P. Barral, A. Bermúdez, M.C. Muñiz, M.V. Otero, P. Quintela and P. Salgado. Numerical simulation of some problems related to aluminium casting. Aceptado para su publicaci'on en el *Journal of Materials Processing Technology.*

[3] P. Barral, C. Moreno, P. Quintela and M.T. Sánchez. A numerical algorithm for Signorini's problem associated with Maxwell-Norton materials using Newton's methods. Sometido a publicación en el *Journal of Computational and Applied Mathematics.*

[4] P. Barral and P. Quintela. A numerical method for simulation of thermal stresses during casting of aluminum slabs. *Computer Methods in Applied Mechanics and Engineering*, 178:69–88, 1999.

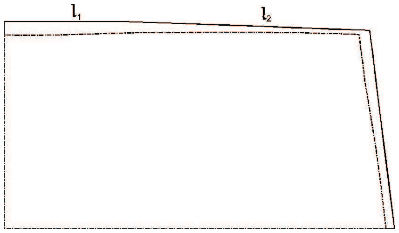


Figura 23: Contracción de un cuarto de la sección transversal de la placa. l_1 y l_2 representan los dos lados de las paredes del molde.

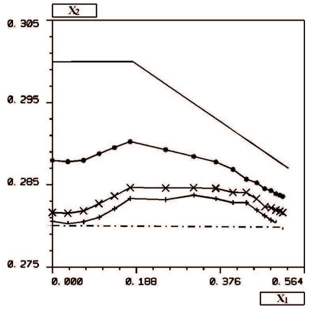


Figura 24: Evolución de la contracción en una sección transversal. (– geometría del molde o inductor, $*$ zona superior, $\times$ zona en estado estacionario, $+$ zona próxima a la base, $\cdot - \cdot$ geometría deseada).

[5] P. Barral and P. Quintela. A numerical algorithm for prediction of thermomechanical deformation during the casting of aluminum alloy ingots. *Finite Elements in Analysis and Design*, 34:125–143, 2000.

[6] P. Barral and P. Quintela. Numerical analysis of a viscoplastic problem with contact condition taking place in an aluminium casting. *Journal of Computational and Applied Mathematics*, 115:63–86, 2000.

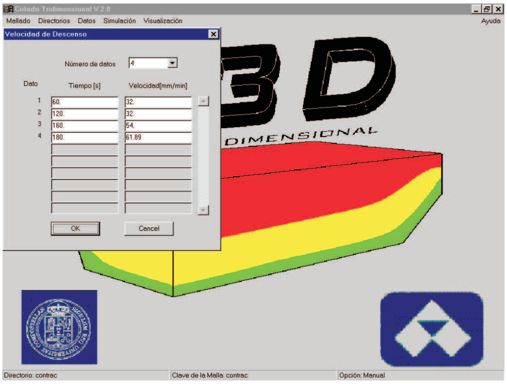
[7] P. Barral and P. Quintela. Asymptotic justification of the treatment of a metallostatic pressure type boundary condition in an aluminium casting. *Mathematical Models and Methods in Applied Sciences*, 11:951–977, 2001.

[8] P. Barral and P. Quintela. Asymptotic analysis of a metallostatic pressure type boundary condition modelled by a fictitious domain method in an aluminium casting. *Asymptotic Analysis*, 30:93–116, 2002.

[9] P. Barral and P. Quintela. Existence of a solution for a Signorini contact problem for Maxwell-Norton materials. *IMA Journal of Applied Mathematics*, 67:525–549, 2002.

[10] A. Bensoussan and J. Frehse. Asymptotic behaviour of the time dependent Norton-Hoff law in plasticity theory and H^1 regularity. *Boundary value problems for partial differencial equations and applications*, 3–25 *RMA Res. Notes Appl. Math*, 29, Masson, 1993.

[11] A. Bermúdez and C. Moreno. Duality methods for solving variational inequalities. *Comput. Math. Appl.*, 7:43–58, 1981.



Velocidad de Descenso

Número de datos: 4

Dato	Tiempo [s]	Velocidad[mm/min]
1	60.	32.
2	120.	32.
3	160.	54.
4	180.	61.89

Directorio: contrac

Clave de la Malla: contrac

Opción: Manual

Figura 25: Ventana principal del paquete **C3D** y menú de introducción de datos de funcionamiento.

[12] A. Bermúdez amd M.C. Muñiz. Numerical solution of a free boundary problem taking place in an electromagnetic casting. *Mathematical Models and Methods in Applied Sciences*, 9:1393–1416, 1999.

[13] A. Bermúdez and M.V. Otero. Análisis matemático y resolución numérica de un modelo tridimensional de una colada de aluminio. *XVI CEDYA - 6º Congreso de Matemática Aplicada*, 2:1405–1412, 1999.

[14] J.F. Ciavaldini. Analyse numérique d'un probléme de Stefan á deux phases par une méthode d'éléments finis. *SIAM J. Numer. Anal*, 12:464–487, 1975.

[15] L.M. Díaz. La modelización matemática en la producción primaria del aluminio. *Boletín de la Sociedad Española de Matemática Aplicada*, 3:17–21, 1993.

[16] M. Djaoua and P. Suquet. Évolution quasi-statique des milieux viscoplastiques de Maxwell-Norton, *Math. Meth. in Appl. Sci.*, 6:192–205, 1984.

[17] J.M. Drezet and M. Plata. Thermomechanical effects during direct chill and electromagnetic casting of aluminium alloys. Part I: Experimental Investigation.*Light Metals*, 931–940, 1995.

[18] J.M. Drezet, M. Rappaz and Y. Krähenbühl. Thermomechanical effects during direct chill and electromagnetic casting of aluminum alloys. Part II: Numerical simulation, *Light Metals*, 941–950, 1995.

[19] J. Frehse and J. Málek. Boundary regularity results for models of elasto-perfect plasticity. *Mathematical Models and Methods in Applied Sciences*, 9:1307–1321, 1999.

[20] A. Friaâ. Le matériau de Norton-Hoff généralisé et ses applications en analyse limite. *C.R. Acad. Sci. Paris*, serie A, 286:953–956, 1978.

[21] G. Geymonat and P. Suquet. Functional spaces for Norton-Hoff materials. *Math. Meth. in Appl. Sci,* 8:206–222, 1986.

[22] W. Han and M. Sofonea. Numerical analysis of a frictionless contact problem for elastic-viscoplastic materials. *Computer Methods in Applied Mechanics and Engineering*, 190:179–191, 2000.

[23] B. Hannart, F. Cialti and R. Schalkwijk. Thermal stresses in DC casting of aluminium slabs: Application of a finite element model. *Light metals*, 879–887, 1994.

[24] J.O. Kristiansson and E.H. Zetterlund. Thermal stresses and strains in the solidifying shell within the primary cooling zone during continuous casting. *Proceedings of the International Conference in Numerical Methods in Industrial Forming Processes*, Swansea, UK, 1982.

[25] L. Lemaitre and J.L. Chaboche. *Mécanique des matériaux solides.* Dunod, 1988.

[26] S. Mariaux, M. Rappaz, Y. Krähenbühl and M. Plata. Modelling of thermomechanical effects during the start-up phase of the electromagnetic casting process. *Light Metals*, 175–187, 1992.

[27] C. Moreno. Teoría matemática de la plasticidad, *Notas de la VI Escuela de Otoño Hispano-Francesa sobre Simualción en Física e Ingeniería*, Tomo **II**, 1994.

[28] F. Murat. Compacité par compensation, *Ann. Scuola Norm. Sup. Pisa Cl. Sci. (4)*, 5(3):489–507, 1978.

[29] J. Nečas and I. Hlaváček. *Mathematical Theory of Elastic and Elasto-Pastic Bodies: An Introduction.* Elsevier, 1981.

[30] V. Otero. *Simulación numérica de una colada de aluminio.* Tesina de Licenciatura, Universidade de Santiago de Compostela, 1995.

[31] T.S. El-Raghy, M.F. El-Demerdash, H.A. Ahmed and A.M. El-Sheikh. Modelling of the transient and steady state periods during aluminium DC casting, *Light Metals*, 925–929, 1995.

[32] M. Rochdi and M. Sofonea. On frictionless contact between two elastic-viscoplastic bodies. *Quarterly Journal of Mechanics and Applied Mathematics*, 50(3):481–496, 1997.

[33] J.C. Simo and R.L. Taylor. Consistent tangent operator for rate-independent elastoplasticity. *Computer Methods in Applied Mechanics and Engineering*, 48:101–118, 1985.

[34] M. Sofonea. On a contact problem for elastic-viscoplastic bodies. *Nonlinear Analysis, Theory, Methods & Applications*, 29(9):1037–1050, 1997.

[35] P. Le Tallec *Numerical analysis of viscoelastic problems*. Masson, 1990.

[36] C. Truesdell. *Mechanics of Solids*. Volume II, Springer Verlag, 1984.

Bol. Soc. Esp. Mat. Apl. n°26(2003), 35–56

Aproximando soluciones que explotan*

P. Groisman[1] y J. D. Rossi[2]

[1]Departamento de Matemática, FCEyN, UBA,
Buenos Aires, Argentina.
[2]Departamento de Matemática, Universidad Católica de Chile,
Santiago, Chile.

jrossi@mat.puc.cl, pgroisma@dm.uba.ar

Resumen

Al aproximar numéricamente ecuaciones en derivadas parciales con singularidades, un buen método debe reproducir el comportamiento asintótico de las soluciones. Un estudio riguroso y detallado de los posibles comportamientos es necesario para entender las ventajas y desventajas de los métodos numéricos en modelos gobernados por ecuaciones de evolución. En este artículo se presentan algunos resultados sobre el comportamiento de aproximaciones numéricas de soluciones de ecuaciones parabólicas que desarrollan singularidades en tiempo finito.

Palabras clave: *Blow-up, aproximaciones numéricas.*
Clasificación por materias AMS: *65M60, 65M20, 35K60, 35B40.*

1 Introducción

Las ecuaciones parabólicas semilineales de segundo orden son utilizadas para modelar diversos fenómenos y procesos en mecánica, física, tecnología, biología y muchas otras áreas. Por ejemplo, bajo ciertas condiciones, la ecuación del calor semilineal describe el proceso de conducción en plasma, filtración de gases y líquidos en medios porosos, reacciones químicas y procesos de crecimiento y migración de poblaciones, etc.

* Financiado por Universidad de Buenos Aires, por ANPCyT PICT No. 5009 y por Fundación Antorchas. J. D. Rossi es investigador del CONICET.

Fecha de recepción: 27 de mayo de 2003

Para este tipo de problemas, por lo general se posee una teoría de existencia y unicidad para tiempos cortos, sin embargo puede ocurrir que la solución deje de existir en un tiempo finito. La forma más simple en que aparece este fenómeno es cuando la solución tiende a infinito a medida que la variable temporal t se acerca a un tiempo finito, $T > 0$. La prueba de que en las ecuaciones no lineales puede aparecer este fenómeno debido simplemente a la estructura no lineal del problema la brinda el siguiente ejemplo. Consideremos la ecuación ordinaria

$$\begin{cases} u' = f(u), \\ u(0) = u_0 > 0, \end{cases} \tag{1}$$

donde f es positiva, creciente y regular. En particular, cuando $f(u) = u^p$ con $p > 1$, esta ecuación tiene como única solución a

$$u(t) = C_p(T - t)^{-1/(p-1)},$$

donde, $T = u_0^{1-p}(p - 1)^{-1}$ y $C_p = (p - 1)^{-1/(p-1)}$. En este ejemplo tan simple puede verse que la solución $u(t)$ es regular para todo $t < T$ y que $u(t) \to +\infty$ cuando $t \nearrow T$. Cuando esto ocurre decimos que u explota (blow-up en la literatura inglesa). En general, si $\int^{+\infty} 1/f < +\infty$, entonces $u(t)$ explota en tiempo finito T en el sentido anterior, $\lim_{t \nearrow T} u(t) = +\infty$.

Los problemas que estudiamos, además de depender del tiempo poseen una estructura espacial, es decir que la función incógnita u depende no sólo del tiempo sino también de una variable espacial, $u = u(x, t)$, $x \in \Omega$, un domino en $\mathbb{R}^d$. Consideramos procesos de evolución que pueden ser modelados de la forma

$$u_t = \mathcal{F}(u), \qquad t \in (0, T), \qquad u(0) = u_0,$$

donde $\mathcal{F}$ es un operador y la solución $u = u(t)$ es una curva en cierto espacio funcional $(H^1(\Omega), H_0^1(\Omega), C^k(\Omega)$, etc.).

Los primeros resultados rigurosos que prueban la aparición de este fenómeno de explosión en ecuaciones en derivadas parciales son los obtenidos por Kaplan ([40]) y Fujita ([25]).

Varias preguntas surgen naturalmente ante este tipo de problemas, por ejemplo: (1) ¿Hay blow-up? (2) ¿Cuándo? (3) ¿Dónde? (4) ¿Cómo? (5) ¿Son las respuestas a estas preguntas estables ante perturbaciones? (6) ¿Cómo calcularlo numéricamente?. Si se trata de un sistema de ecuaciones, agregamos dos preguntas más, (7) ¿Es posible tener diferentes conjuntos de explosión para diferentes componentes del sistema? (8) ¿Es posible que ciertas componentes permanezcan acotadas mientras otras explotan (explosión no simultánea)?

Éstas no son las únicas preguntas que la comunidad matemática ha considerado de interés. En [26] se detallan los fenómenos relevantes a estudiar ante la aparición de este tipo de singularidades, dando un enfoque integral de este problema. Es de destacar que en el estudio riguroso de estos fenómenos la comunidad matemática española ha tenido un papel relevante, produciendo muchos trabajos de gran interés, como por ejemplo, [17, 27, 28, 35, 36, 37, 46, 51].

En este artículo avanzamos en torno a la sexta pregunta. Lo que nos interesa conocer de un método numérico para aproximar a las soluciones de estos problemas es en qué medida las respuestas a las restantes preguntas coinciden en la solución continua y en su aproximación.

Resumimos a continuación algunos de los métodos numéricos más conocidos, analizamos sus propiedades y en qué medida reproducen el comportamiento asintótico de las soluciones. Para problemas en derivadas parciales un buen método numérico debe reproducir no sólo el comportamiento asintótico temporal de las soluciones sino también la estructura espacial de la solución cerca del tiempo de explosión. Un estudio riguroso y detallado de los posibles comportamientos es necesario para entender las ventajas y desventajas de los métodos numéricos en el estudio y la predicción de comportamientos singulares en modelos gobernados por ecuaciones de evolución.

Las dificultades del análisis se deben, fundamentalmente, a que los teoremas de convergencia usuales no incluyen casos singulares como los que aquí se analizan. Por lo tanto suele utilizarse otro tipo de técnicas, basadas generalmente en principios de comparación, estimaciones funcionales y desarrollos asintóticos cuidadosos de las soluciones numéricas.

Nos concentraremos principalmente en problemas de dos tipos:

1. Problemas con fuente no lineal, i.e., de la forma,

$$\begin{cases} u_t = \Delta u^m + u^p, & \text{en } \Omega \times (0,T), \\ u = 0, & \text{sobre } \partial\Omega \times (0,T), \\ u(x,0) = u_0(x) \geq 0, & x \in \Omega, \end{cases} \tag{2}$$

en un dominio acotado Ω.

2. Problemas con condiciones de Neumann no lineales,

$$\begin{cases} u_t = \Delta u^m, & \text{en } \Omega \times (0,T), \\ \frac{\partial u^m}{\partial \eta} = u^p, & \text{sobre } \partial\Omega \times (0,T), \\ u(x,0) = u_0(x) \geq 0, & x \in \Omega. \end{cases} \tag{3}$$

Para $m = 1$ (difusión lineal) se obtienen las ecuaciones de reacción-difusión más estudiadas entre las que poseen el fenómeno de explosión. Este caso será estudiado en detalle, pero también mencionaremos resultados para los casos $m > 1$ (ecuación de medios porosos) y $0 < m < 1$ (ecuación de difusión rápida).

2 Problemas con fuente no lineal

Consideremos en primer lugar el siguiente problema

$$\begin{cases} u_t = \Delta u + u^p, & \text{en } \Omega \times (0,T), \\ u = 0, & \text{sobre } \partial\Omega \times (0,T), \\ u(x,0) = u_0(x) \geq 0, & x \in \Omega. \end{cases} \tag{1}$$

Si el dato inicial $u_0(x)$ es regular (como asumiremos a lo largo de todo este artículo) entonces existe una única solución para este problema y es regular. Sin embargo, no importa cuán suaves sean el dato inicial y el dominio Ω, hay

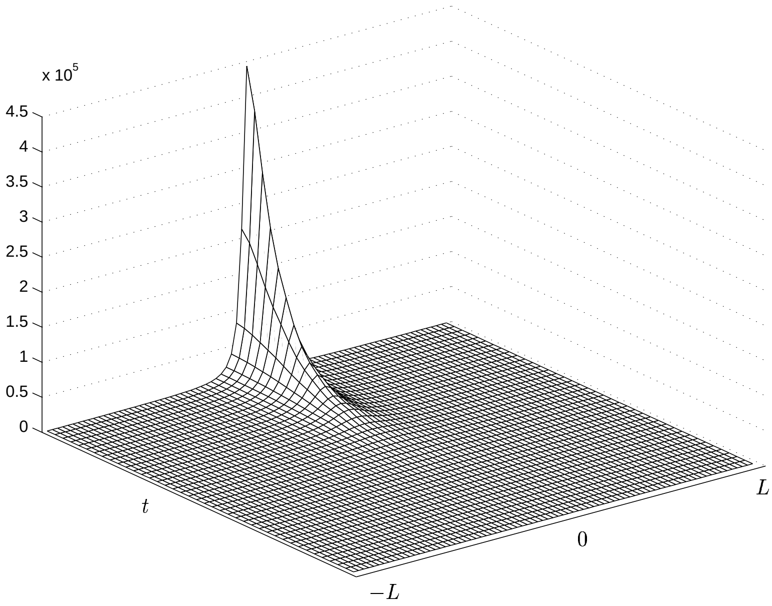


Figura 1: Una solución de $u_t = u_{xx} + u^p$ en $(-L, L) \times (0, T)$ explotando en tiempo finito.

datos iniciales para los cuales existe un tiempo finito T en el que la solución deja de existir. En este tiempo ocurre que la solución tiende a infinito, es decir

$$\lim_{t \to T} \|u(\cdot, t)\|_{L^\infty(\Omega)} = +\infty.$$

Como ya hemos dicho a este fenómeno se lo denomina *explosión en tiempo finito (blow-up).*

Para este problema se conocen respuestas a las preguntas mencionadas anteriormente

(1) ¿Hay blow-up? Si, si u_0 es suficientemente grande. Más precisamente, se tiene la siguiente caracterización de las soluciones que explotan: consideremos la función

$$\Phi(u)(t) \equiv \int_\Omega \frac{|\nabla u(s,t)|^2}{2}\, ds - \int_\Omega \frac{|u(s,t)|^{p+1}}{p+1}\, ds.$$

En [7, 15], se prueba que una solución explota en tiempo finito si y sólo si existe un tiempo t_0 con $\Phi(u)(t_0) < 0$. Un cálculo directo muestra que dado un dato inicial $u_0 \neq 0$ entonces $\Phi(\lambda u_0) < 0$ si λ es suficientemente grande.

(2) ¿Cuándo? El funcional Φ también sirve para estimar el tiempo de blow-up T, ya que cumple

$$\lim_{t \to T} \Phi(u)(t) = -\infty,$$

y en [43] se prueba que para tiempos cercanos al tiempo de explosión se verifica

$$(T-t) \leq \frac{C}{(-\Phi(u(t)))^{\frac{p-1}{p+1}}}.$$

(3) ¿Dónde? Llamamos $B(u)$ al conjunto de puntos $x \in \Omega$ tales que $u(x_n, t_n) \nearrow +\infty$ para alguna sucesión $(x_n, t_n) \to (x, T)$. Para el caso unidimensional ($\Omega = (0,1)$) se sabe que el conjunto de blow-up ($B(u)$) está localizado en un número finito de puntos. Es decir, existen $x^1, \ldots, x^k$ tales que $u(x^i, t) \to +\infty$ cuando $t \to T$ y $u(x,t)$ permanece acotada hasta tiempo T para todo $x \notin \{x^1, .., x^k\}$ (ver [13]). Esto no es cierto para dimensiones mayores, donde el conjunto de blow-up puede ser incluso una hipersuperficie. Pueden obtenerse mejores caracterizaciones del conjunto $B(u)$ si se tiene información sobre el dato inicial (por ejemplo, simetría).

(4) ¿Cómo? El comportamiento asintótico en los puntos donde u explota está dado por

$$u(x^i, t) \sim (T-t)^{-\frac{1}{p-1}}, \qquad \text{(tasa de blow-up)}$$

donde $f(t) \sim g(t)$ significa que existen constantes positivas C_1, C_2 tales que $C_1 g(t) \leq f(t) \leq C_2 g(t)$ (ver [29, 35, 36]). Más aún, se tiene

$$\lim_{t \nearrow T} (T-t)^{\frac{1}{p-1}} u(x^i, t) = C_p.$$

Proponemos entonces métodos numéricos para este problema y estudiamos el comportamiento asintótico de las soluciones. Como primer paso para introducir los métodos numéricos planteamos un método de líneas. La idea de este tipo de métodos es discretizar la variable espacial (x), manteniendo la variable temporal (t) continua. De esta forma todas las derivadas espaciales desaparecen pero se mantienen las derivadas temporales, obteniéndose un sistema de ecuaciones ordinarias. Para la discretización espacial consideramos un método general con hipótesis adecuadas. Más precisamente, suponemos que para cada $h > 0$ (h es el parámetro de discretización) tenemos un conjunto de nodos $\{x_1, \ldots, x_N\} \subset \overline{\Omega}$ ($N = N(h)$), y que la aproximación numérica de u en los nodos x_k viene dada por $U(t) = (u_1(t), \ldots, u_N(t))$. Es decir que $u_k(t)$ aproxima a $u(x_k, t)$. Suponemos entonces que $U(t)$ es la solución de un sistema de ecuaciones ordinarias de la forma

$$MU'(t) = -AU(t) + MU^p(t), \qquad t \in [0, T_h), \tag{2}$$

con dato inicial $u_k(0) = u_0(x_k)$. En (2) y de aquí en adelante todas las operaciones entre vectores se entienden coordenada a coordenada. Las matrices $M, A \in \mathbb{R}^{N\times N}$ dependen de h y del método numérico en cuestión. Este tipo de sistemas se obtiene cuando se aplican métodos de diferencias finitas o elementos finitos.

Por ejemplo, en una dimension espacial consideramos $\Omega = (0,1)$, $h = 1/N$, los nodos $x_i = ih$ ($1 \leq i \leq N$) y al aplicar diferencias finitas centradas o elementos finitos con *mass lumping* (en este caso ambos métodos coinciden) el sistema (2) toma la forma

$$u_i'(t) = \frac{u_{i+1}(t) - 2u_i(t) + u_{i-1}(t)}{h^2} + u_i^p(t), \qquad 2 \leq i \leq N-1.$$

Decimos que las soluciones numéricas tienen blow-up si existe un tiempo finito T_h tal que

$$\lim_{t \nearrow T_h} \|U(t)\|_\infty = \lim_{t \nearrow T_h} \max_j u_j(t) = +\infty.$$

En el resto del trabajo denotaremos con T al tiempo de explosión de la solución continua, con h al parámetro de la malla y con T_h a los tiempos de explosión de las aproximaciones numéricas (si es que éstas explotan).

El análisis numérico desarrollado hasta el momento sobre este tipo de problemas todavía es escaso en relación a la teoría continua (ver [9, 26]). Algunos trabajos que tratan sobre aproximaciones numéricas con mallas fijas son [1, 2, 14, 44].

En [10] se desarrolla un algoritmo que refina la malla espacial a medida que avanza el tiempo aprovechando la invarianza de escala que poseen las soluciones de esta ecuación. En [11, 12, 39] se desarrollan los denominados *moving mesh methods* con el objetivo de reproducir el comportamiento asintótico de las soluciones. Estos métodos también hacen uso de la invarianza de escalas de las soluciones. Si bien estos métodos son los más utilizados, sólo sirven para modelos unidimensionales y no se poseen demostraciones rigurosas sobre su convergencia y comportamiento asintótico.

Otros trabajos que merecen ser mencionados son [6, 8, 19, 41, 50]. En [9] se resumen los resultados contenidos en estos artículos.

En todos los métodos numéricos que analizamos se ha probado la convergencia de las soluciones discretas a la solución continua en regiones en donde u es regular, más precisamente, el típico resultado de convergencia que se obtiene es de la forma: dado $\tau > 0$ existe $C = C(\tau) > 0$ tal que,

$$\|u - u_h\|_{L^\infty(\overline{\Omega} \times [0, T-\tau])} \leq Ch^\alpha,$$

donde u_h es una interpolación de U.

No es de esperar un mejor resultado de convergencia en el sentido de obtener, por ejemplo, convergencia del método hasta tiempo T, ya que la solución desarrolla una singularidad en ese momento. Para que esto ocurra debería pasar, por lo menos, que T y T_h coincidan, lo cual no es cierto en general, ni siquiera en el caso de una ecuación diferencial ordinaria.

En [33] se consideran aproximaciones numéricas para el problema (1) en una dimensión. Se prueba que existen soluciones numéricas con blow-up si y sólo si $p > 1$. Para describir la aparición de blow-up en las aproximaciones numéricas se utiliza una versión discreta del funcional Φ descrito anteriormente:

$$\Phi_h(U) = \frac{1}{2}\langle A^{1/2}U, A^{1/2}U\rangle - \frac{1}{p+1}\langle MU^p, U\rangle.$$

Las soluciones discretas de este problema explotan si y sólo si existe un tiempo t_0 en el que el funcional Φ_h se hace negativo.

Como esta caracterización es similar a la del problema continuo, mediante esta propiedad se puede probar que si la solución del problema continuo explota en tiempo finito, entonces lo mismo le ocurre a las del discreto para h suficientemente pequeño. Además este funcional permite obtener cotas sobre la distancia al tiempo de blow-up similares a las que se tienen para el problema continuo, que sirven para probar convergencia de los tiempos de explosión ($T_h \to T$ cuando $h \to 0$). Sin embargo, utilizando otro enfoque, podemos acotar el orden de esta convergencia.

Se puede probar que la tasa de explosión de las soluciones numéricas es la misma que la del problema continuo y, más aún, en [30] se prueba que existe una constante C independiente de h tal que

$$\|U(t)\|_\infty \le C(T_h - t)^{-\frac{1}{p-1}}.$$

Esta cota sirve, entre otras cosas, para acotar la diferencia de los tiempos de explosión. Supongamos que el método numérico en consideración es consistente de orden α, es decir, que las soluciones de (1) verifican (llamando $z_i(t) = u(x_i, t)$, $Z(t) = (z_1(t), \dots, z_N(t))$)

$$MZ'(t) = -AZ(t) + MZ^p(t) + \varepsilon(t, h),$$

con $\|\varepsilon(t, h)\|_\infty \le Ch^\alpha (T - t)^{-\theta}$. El término $C(T - t)^{-\theta}$ es una cota para las derivadas de u. Bajo esta suposición la función de error $E(t) = Z(t) - U(t)$ verifica

$$ME'(t) \quad = \quad -AE(t) + \frac{Z^p(t) - U^p(t)}{Z(t) - U(t)} E(t) + \varepsilon(t, h).$$

Sea entonces t_0 el primer tiempo en que $\|E(t_0)\|_\infty = 1$ ($t_0 > 0$ para h suficientemente pequeño). En $[0, t_0]$ $E(t)$ verifica

$$ME'(t) \quad \le \quad -AE(t) + C(T - t)^{-1} E(t) + Ch^\alpha (T - t)^{-\theta}.$$

Como $Ch(T - t)^{-C}$ es una supersolución de esta ecuación, resulta ser una cota para $E(t)$ que dice que esta función se mantiene pequeña hasta tiempos cercanos al tiempo de explosión con tal de tomar h pequeño. De hecho permite obtener $|T - t_0| \le Ch^\alpha$ y $|T_h - t_0| \le Ch^\alpha$. Entonces

$$|T - T_h| \le |T - t_0| + |T_h - t_0| \le Ch^\alpha.$$

Este tipo de ideas se puede utilizar también para acotar la diferencia entre tiempos de explosión para soluciones del problema (1) cuando se perturba el dato inicial, el coeficiente de difusión o la potencia de reacción p, obteniéndose cotas similares. Para el caso del dato inicial esta cota incluso puede ser mejorada. En [32, 34] se perturba el dato inicial u_0 añadiéndole una función g de norma pequeña y entonces los tiempos de explosión $T(u_0), T(u_0 + g)$ satisfacen

$$|T(u_0 + g) - T(u_0)| \le C\|g\|_{L^\infty} |\ln(\|g\|_{L^\infty})|^\gamma, \qquad \gamma > 0.$$

Nos adentramos entonces en el estudio asintótico de las soluciones numéricas: la tasa y el conjunto de blow-up de u_h para h fijo. Obtenemos que la tasa de explosión es la misma que la del problema continuo y que el conjunto de blow-up, si bien puede diferir, está contenido en un entorno del conjunto de blow-up de u si h es suficientemente pequeño. Precisamente obtenemos que si u_h explota en tiempo finito entonces

$$\lim_{t\to T_h} \|U(t)\|_\infty (T_h - t)^{\frac{1}{p-1}} = C_p.$$

Este comportamiento, además de ser el mismo que tienen las soluciones continuas, es el que tienen las soluciones de la ecuación $u'(t) = u^p(t)$, como vimos anteriormente. Esto nos dice que, tanto en el esquema numérico, como en el problema continuo, cuando la solución está cerca de explotar, el término de difusión (AU, Δu) es despreciable respecto del término de reacción (u^p).

Si denotamos por $B^*(u_h)$ al conjunto de nodos que tienen este comportamiento (que explotan con la misma tasa que el máximo), entonces para las soluciones con blow-up se tiene que $\emptyset \neq B^*(u_h) \subset B(u_h)$. Para el problema continuo, el conjunto de blow-up $B(u)$ esta formado sólo por esta clase de puntos, sin embargo en las aproximaciones numéricas aparece un fenómeno de propagación espúreo. El blow-up se propaga a los K nodos adyacentes a $B^*(u_h)$ (con una tasa menor), donde $K = [1/(p-1)]$ ($[a]$ denota la parte entera de a). Sin embargo, como la cantidad de nodos donde se propaga la explosión sólo depende de p, el conjunto $B(u_h)$ se mete en un entorno de $B^*(u_h)$ cuando h tiende a cero. Si es posible localizar $B^*(u_h)$ cerca de $B(u)$ entonces se pueden obtener estimaciones de la forma

$$B(u_h) \subset B(u) + (-\varepsilon, \varepsilon), \qquad \forall h \leq h_0(\varepsilon).$$

Esto es cierto para el problema unidimensional, sin embargo para el caso multidimensional no se ha podido demostrar la localización de $B^*(u_h)$ en torno de $B(u)$, aunque conjeturamos que es cierta.

2.1 Un esquema totalmente discreto

Las aproximaciones numéricas semidiscretas son una herramienta muy útil y el estudio de sus propiedades asintóticas aporta datos muy importantes para comprender en qué medida es posible aproximar problemas como los que estamos considerando. Sin embargo, no es suficiente. Para poder brindar un análisis acabado del comportamiento de las aproximaciones numéricas, es necesario integrar el sistema de ecuaciones ordinarias que define a las aproximaciones mencionadas anteriormente, cuyas soluciones también son singulares. Por lo tanto las técnicas usuales de aproximación y análisis de error no se extienden naturalmente para este tipo de problemas, y es necesario proponer métodos adecuados y estudiar y dar pruebas rigurosas del comportamiento de lo que denominamos aproximaciones totalmente discretas.

Primero veamos como adaptar el paso temporal en una ecuación ordinaria. Consideremos $u(t)$ la solución del problema,

$$\left\{ \begin{array}{r} u' = f(u), \\ u(0) = u_0 > 0, \end{array} \right. \tag{3}$$

donde f es positiva, creciente y regular. Si $\int^{+\infty} 1/f < +\infty$, como mencionamos en la introducción, $u(t)$ explota en tiempo finito, T, dado por,

$$T = \int_{u_0}^{+\infty} \frac{1}{f(s)} ds.$$

En [3] se analiza un método adaptivo para elegir los pasos en un método numérico de tipo Euler o Runge-Kutta, de forma que se reproduzca el comportamiento asintótico. Luego se utiliza este análisis para desarrollar métodos adaptivos para ecuaciones en derivadas parciales.

A continuación, para simplificar la exposición, reproducimos este análisis para un caso concreto, el método de Euler explícito. Si aproximamos $u(t)$ usando el método de Euler nos queda

$$\left\{ \begin{array}{l} u^{j+1} = u^j + \tau_j f(u^j), \\ u^0 = u_0. \end{array} \right. \tag{4}$$

Ahora elegimos los incrementos τ_j de forma adaptiva como sigue

$$\tau_j f(u^j) = \lambda, \tag{5}$$

donde λ es un parámetro. La principal idea detrás de esta elección se basa en mirar la ecuación $u'(t) = f(u(t))$ como el sistema

$$\left\{ \begin{array}{l} \frac{\partial t}{\partial s} = 1/f(u(\tau)), \\ \frac{\partial u}{\partial s} = 1, \end{array} \right. \tag{6}$$

donde el nuevo tiempo, s, es una variable auxiliar. Aplicando el método de Euler con paso fijo de tamaño λ al sistema nos queda:

$$\left\{ \begin{array}{l} t^{j+1} = t^j + \lambda/f(u^j), \\ u^{j+1} = u^j + \lambda. \end{array} \right.$$

Es decir, $\tau_j = t^{j+1} - t^j = \lambda/f(u^j)$. Usando (4) y (5) tenemos

$$u^j = u^{j-1} + \lambda = ... = u^0 + j\lambda$$

y entonces

$$\tau_j = \frac{\lambda}{f(u^j)} = \frac{\lambda}{f(u_0 + j\lambda)}.$$

De esto podemos concluir que el esquema numérico también explota en el sentido siguiente: $y^j \to \infty$ mientras $\sum_j \tau_j < +\infty$. Además, estas acotaciones, proporcionan una estimación del tiempo de explosión:

$$\begin{aligned}\sum_{j=0}^{\infty}\tau_j &= \tau_0+\sum_{j=1}^{\infty}\frac{\lambda}{f(u_0+j\lambda)}\le\frac{\lambda}{f(u_0)}+\int_0^{+\infty}\frac{\lambda}{f(u_0+s\lambda)}ds\\ &= \frac{\lambda}{f(u_0)}+\int_{u_0}^{+\infty}\frac{1}{f(s)}ds=\frac{\lambda}{f(u_0)}+T.\end{aligned}$$

Entonces $T_\lambda=\sum_j\tau_j<+\infty$ y

$$T_\lambda-T\le\frac{\lambda}{f(u_0)}.$$

Además, como u es convexa, es fácil ver que la solución numérica es menor o igual que la continua y entonces

$$T\le T_\lambda.$$

Concluimos que esta elección de los pasos para el método de Euler nos proporciona un esquema que explota y además aproxima bien los tiempos de explosión continuos en el siguiente sentido:

$$|T-T_\lambda|\le\frac{\lambda}{f(u_0)}\to 0,\qquad \lambda\to 0.$$

Integrando el sistema (6) con métodos tipo Runge-Kutta se obtienen métodos adaptivos de mayor orden para (3). Este mayor orden de convergencia también se puede observar en los tiempos de blow-up numéricos, T_λ.

Ahora retomamos el análisis de (1). En [31] se propone un esquema totalmente discreto que surge de discretizar en t el sistema (2). Basándose en los esquemas propuestos en [3, 16] se analiza en primer lugar un método de Euler explícito y a continuación se introduce un esquema implícito que permite eliminar las restricciones en el paso temporal.

Utilizamos la notación $U^j=(u_1^j,\dots,u_N^j)$ para denotar el valor de la aproximación totalmente discreta a tiempo t_j, y $\tau_j=t_{j+1}-t_j$ para el paso de tiempo. El método explícito propuesto viene dado por

$$\begin{cases} MU^{j+1}=MU^j-\tau_j\left(AU^j+M(U^j)^p\right)\\ U(0)=u_0^I.\end{cases}\tag{7}$$

Entonces elegimos el paso de tiempo $\tau_j=t_{j+1}-t_j$ de forma tal que el comportamiento asintótico del esquema totalmente discreto reproduzca al del continuo. Fijamos $\lambda>0$ pequeño y tomamos

$$\tau_j=\frac{\lambda}{(w^j)^p},\qquad w^j=\langle MU^j,E\rangle,$$

aquí $E=(1,\dots,1)$. Con esta elección podemos probar que el comportamiento de w^j viene dado por

$$\partial w^{j+1}\sim(w^j)^p.$$

de donde podemos ver que es similar al del problema continuo. Decimos entonces que una solución de (7) explota si

$$\lim_{j\to\infty} \|U^j\|_\infty = \infty, \qquad \text{y} \qquad T_{h,\lambda} := \sum_{j=1}^{\infty} \tau_j < \infty.$$

Llamamos a $T_{h,\lambda}$ el tiempo de blow-up.

Para este tipo de esquemas se prueban resultados similares a los obtenidos cuando se considera el método semidiscreto en cuanto al comportamiento asintótico de las soluciones, la convergencia de los tiempos de blow-up y los conjuntos de explosión.

2.2 Blow-up no simultáneo.

En [30] se considera el caso de un sistema de ecuaciones acopladas en el término fuente,

$$\begin{array}{ll} u_t = \Delta u + u^{p_{11}} v^{p_{12}} & \text{en } \Omega \times (0,T), \\ v_t = \Delta v + u^{p_{21}} v^{p_{22}} & \text{en } \Omega \times (0,T), \end{array} \tag{8}$$

con condiciones de borde de tipo Dirichlet homogéneas, $u = v = 0$ sobre $\partial\Omega \times [0,T)$, y dato inicial $u(x,0) = u_0(x)$, $v(x,0) = v_0(x)$ en Ω. Los exponentes p_{ij} son mayores que cero.

Bajo condiciones de regularidad bastante generales existe una única solución (u,v) de (8) local en tiempo, ([17]). Al igual que en los problemas anteriores estas soluciones pueden explotar, es decir

$$\limsup_{t\nearrow T} \; (\|u(\cdot,t)\|_{L^\infty} + \|v(\cdot,t)\|_{L^\infty}) = +\infty.$$

El conjunto de exponentes para los que este fenómeno puede darse fue caracterizado en [17], donde se encuentra que una de las siguientes condiciones debe cumplirse para que existan soluciones con blow-up, $p_{11} > 1$, $p_{22} > 1$ ó $(p_{11} - 1)(p_{22} - 1) < p_{12}p_{21}$.

A priori no hay razón por la cual las funciones u y v deban tender a infinito simultáneamente en el tiempo T (el tiempo maximal de existencia). De hecho, en [45] se prueba que existen datos iniciales para los cuales u explota y v no si y sólo si $p_{11} > 1$ y $p_{21} < p_{11} - 1$. A este fenómeno se le denomina *explosión no simultánea.*

El principal interés en este problema consiste en analizar si las condiciones para la aparición de soluciones con blow-up y el fenómeno de blow-up no simultáneo son reproducidos por las aproximaciones numéricas usuales. Para esto, se propone el método numérico general mencionado anteriormente, en este caso se obtiene un sistema de ecuaciones ordinarias con la siguiente estructura

$$\left\{ \begin{array}{lcl} MU'(t) & = & -AU(t) + MU^{p_{11}}V^{p_{12}}(t), \\ MV'(t) & = & -AV(t) + MU^{p_{21}}V^{p_{22}}(t). \end{array} \right. \tag{9}$$

En [30] se prueba que existen soluciones (U,V) de (9) que explotan en tiempo finito si y sólo si los exponentes p_{ij} verifican las mismas condiciones que valen para el problema continuo. Además se prueba que si se tiene una solución (U, V) de (9) tal que U explota en tiempo finito T_h y V se mantiene acotada hasta ese tiempo, entonces $p_{11} > 1$ y $p_{21} < p_{11} - 1$. Más aún, si $p_{11} > 1$ y $p_{21} < p_{11} - 1$, entonces para todo dato inicial $V_0 \neq 0$ para (9) existe un dato inicial U_0 de manera tal que U explota en tiempo finito T_h y V se mantiene acotada.

Es decir, la estructura de la no linealidad necesaria para la aparición de blow-up no simultáneo en las aproximaciones numéricas es idéntica a la que se prueba en [45] para el sistema continuo. Este hecho, sumado a la convergencia de las aproximaciones numéricas permite probar que si consideramos una solución del problema continuo que posee blow-up no simultáneo, entonces sus aproximaciones numéricas reproducen el fenómeno para elecciones de h suficientemente pequeñas. Más aún, las aproximaciones numéricas también respetan la tasa de explosión y los tiempos de blow-up numéricos T_h convergen a T.

En [5] se considera un sistema de ecuaciones del calor pero en donde el acople viene de la condiciones de borde $\frac{\partial u}{\partial \eta} = u^{p_{11}} v^{p_{12}}$, $\frac{\partial v}{\partial \eta} = u^{p_{21}} v^{p_{22}}$ en $\partial\Omega \times (0, T)$. Para este problema se obtienen resultados similares a los descritos anteriormente.

2.3 Difusión no lineal

Consideremos la siguiente ecuación

$$\begin{cases} u_t = (u^m)_{xx} + u^p, & (x,t) \in (-L, L) \times [0, T), \\ u(-L, t) = u(L, t) = 1, & t \in [0, T), \\ u(x, 0) = u_0(x) \geq 1, & x \in (-L, L). \end{cases} \tag{10}$$

La principal diferencia entre esta ecuación y las estudiadas anteriormente es la aparición de un término de difusión no lineal. Este término de difusión suele ser utilizada para modelar, por ejemplo, filtración de gases en medios porosos.

En este modelo consideramos $m, p > 1$ parámetros fijos. El dato inicial u_0 es suave y compatible con las condiciones de borde de forma tal de obtener soluciones regulares.

Retomamos entonces las preguntas: ¿Hay blow-up?, ¿Cuándo?, ¿Dónde?, ¿Cómo?, que tienen respuestas un poco distintas en este caso (ver [49] para más detalles).

Nuestro interés se centra en el caso $p = m$ y $L > m\pi/(m-1)$ (aunque en [21] se tratan todos los casos). En este caso todas las soluciones explotan (y la tasa de blow-up viene dada por $\|u(\cdot, t)\|_\infty \sim (T - t)^{-\frac{1}{m-1}}$). El conjunto de blow-up es $B(u) = (-\frac{m\pi}{m-1}, \frac{m\pi}{m-1})$. A este fenómeno (el conjunto de blow-up es un subintervalo propio del dominio) se le denomina blow-up regional.

En [21] se aproxima esta ecuación utilizando elementos finitos lineales a trozos con mass lumping en una malla uniforme para la variable espacial obteniendo un sistema de ecuaciones ordinarias de la forma (2).

Se caracterizan las condiciones para la aparición de explosión numérica y se muestra que las aproximaciones reproducen los casos de blow-up de manera

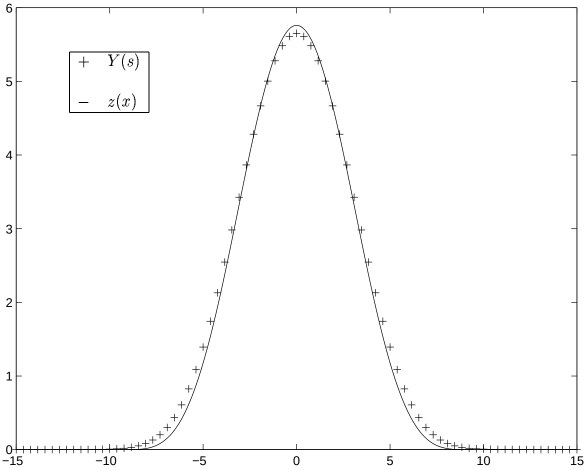


Figura 2: Perfiles autosimilares. El perfil $Y(s)$ representa a $W(t) = (T_h - t)^{\frac{1}{p-1}} U(t)$ cerca de la explosión, mientras que $z(x)$ es el límite de $(T - t)^{\frac{1}{p-1}} u(x,t)$.

precisa, es decir, si u es una solución continua que explota, entonces también explotan sus aproximaciones numéricas para todo h suficientemente pequeño. También se obtienen cotas para $T_h - t$ en términos de $U(t)$. Al igual que antes, esto permite probar convergencia de los tiempos numéricos de explosión al continuo.

Respecto a las tasas, se prueba que todas las soluciones que explotan lo hacen con la misma tasa que la obtenida para las aproximaciones de la ecuación del calor, que coincide con la del problema continuo (10).

Utilizando las tasas, se caracteriza el conjunto de explosión numérico, si $U(t)$ es una aproximación que explota a tiempo T_h, entonces se obtiene explosión global, i.e. $B(u_h) = [-L, L]$. Más aún, todos los nodos explotan con la misma tasa ($u_k(t) \sim (T_h - t)^{-1/(p-1)}$ para todo k). Vemos entonces que no es posible obtener explosión regional en un esquema numérico de malla fija. Sin embargo, se puede recuperar la explosión regional mirando cuidadosamente a $U(t)$ en variables autosimilares adecuadas. Utilizando estas nuevas variables se prueba que si bien todos los nodos explotan con la misma tasa (independiente de h). La constante que acompaña la tasa se comporta correctamente. Se tiene que todos los nodos se comportan como

$$u_k(t) \sim w_k(h)(T_h - t)^{-\frac{1}{p-1}},$$

pero las constantes $w_k(h)$ tienden a cero cuando $h \to 0$ si el nodo correspondiente está fuera de $B(u)$. Es decir, los perfiles autosimilares discretos convergen al perfil continuo, ver Figura 2.

En [22] se considera la ecuación de medios porosos con condiciones de borde no lineales y se obtienen resultados similares.

3 Problemas con condiciones de borde no lineales

Nos concentramos ahora en problemas como (3). El primer problema que tratamos es el siguiente:

3.1 La ecuación del calor con condiciones de borde no lineales

$$\begin{cases} u_t = \Delta u & \text{en } \Omega \times (0,T), \\ \frac{\partial u}{\partial \eta} = u^p & \text{en } \partial\Omega \times (0,T), \qquad (p>1) \\ u(x,0) = u_0(x) & \text{en } \Omega. \end{cases} \tag{11}$$

El dominio Ω y el dato inicial $u_0 > 0$ son regulares para garantizar soluciones suaves.

Para este problema nos preguntamos nuevamente: ¿Hay blow-up?, ¿Cuándo?, etc., obteniendo respuestas diferentes a las obtenidas para el problema semilineal:

(1) ¿ Hay blow-up? Si $p > 1$ todas las soluciones de (11) explotan en tiempo finito ([52, 48]).

(2) ¿Cuándo? En [48] se encuentra una cota superior para el tiempo de blow-up en términos del dato inicial.

(3) ¿Dónde? El conjunto de blow-up está localizado en el borde del dominio, para todo subdominio $\Omega' \subset\subset \Omega$ existe una constante $C = C(d(\Omega', \partial\Omega))$ tal que $u(x,t) \le C$ para todo $x \in \Omega'$ y para todo $0 \le t < T$ ([38, 48]).

(4) ¿Cómo? En [38] se encuentra que la tasa de blow-up es

$$\|u(\cdot,t)\|_\infty \sim (T-t)^{-\frac{1}{2(p-1)}}.$$

Para detalles y más referencias sobre este problema se puede consultar el trabajo [23], donde se resumen los resultados conocidos.

En [3, 16, 18] se consideran aproximaciones numéricas para este problema en una dimensión. Más precisamente se considera

$$\begin{cases} u_t(x,t) &= u_{xx}(x,t), & (x,t) \in (0,1) \times [0,T), \\ u_x(1,t) &= f(u(1,t)), & t \in [0,T), \\ u_x(0,t) &= 0, & t \in [0,T), \\ u(x,0) &= u_0(x), & x \in [0,1], \end{cases} \tag{12}$$

donde $f(s)$ y $u_0(x)$ son positivos y suaves como para garantizar existencia, unicidad y regularidad de la solución (ver [42]).

En estos trabajos se prueban condiciones para la aparición de blow-up en las aproximaciones numéricas (¡que difieren de las del problema continuo!), convergencia del método y de los tiempos de blow-up. Se consideran esquemas semidiscretos ([16]) y totalmente discretos de tipo Euler y Runge-Kutta ([3]).

Por ejemplo, se sabe que si f es convexa y verifica

$$\int^{+\infty} \frac{1}{f(s)f'(s)}\, ds < +\infty, \tag{13}$$

todas las soluciones de (12) explotan ([52]). Más aún, si f es creciente y verifica

$$\int^{+\infty} \frac{1}{f(s)}\,ds < +\infty, \tag{14}$$

el conjunto de blow-up está localizado en el borde, en este caso $B(u) = \{1\}$ (blow-up puntual) [47, 48]. Sin embargo, cuando f verifica (13) pero no (14), el conjunto de blow-up puede ser más grande. Puede ser todo el intervalo $[0, 1]$ (blow-up global), o incluso un subintervalo $[l, 1]$, $0 < l < 1$ (blow-up regional). Por ejemplo, se puede apreciar blow-up global o regional cuando se considera $f(s) = s\log^p(s)$ con $1/2 < p < 1$ o $p = 1$ respectivamente.

En [18] se prueba que los conjuntos de explosión de las aproximaciones numéricas de (12) (usando elementos finitos con mass lumping), o bien se concentran en el borde o bien son todo el intervalo $[0, 1]$. Por lo tanto el fenómeno de explosión regional no existe en este tipo de esquemas. De hecho se prueba que si la función f verifica (14) y es creciente, entonces la cantidad de nodos que explotan depende sólo de f y es independiente de h. Entonces el blow-up numérico sólo puede ser global o localizarse en un entorno de $\{x = 1\}$ cuando h tiende a cero.

A continuación presentamos algunos ejemplos para clarificar las similitudes y diferencias entre los problemas continuos y sus aproximaciones numéricas.

(I) $f(s) = s^p$.

Las soluciones numéricas explotan si y sólo si $p > 1$. La tasa de blow-up para los diferentes puntos está dada por

$$u_{N-k}(t) \sim (T_h - t)^{-1/(p-1)+k}$$

para $0 \le k < 1/(p-1)$ y si $1/(p-1) = k \in \mathbb{N}$, entonces

$$u_{N-k} \sim -\ln(T_h - t).$$

El conjunto de blow-up está compuesto por K nodos, donde $K = [1/(p-1)]$.

Queremos resaltar que a pesar de que la condición de blow-up, $p > 1$, es la misma que en el problema continuo, la tasa de blow-up no es correcta. El conjunto de blow-up verifica

$$B(u_h) = [1 - Kh, 1] = B(u) + [-Kh, 0].$$

(II) $f(s) = s(\ln s)^p$.

Las soluciones numéricas explotan si y sólo si $p > 1$. La tasa de blow-up es

$$\max_i u_i(t) \sim \exp\left(\frac{1}{(T_h - t)^{1/(p-1)}}\right).$$

El conjunto de explosión es todo el intervalo $[0, 1]$.

En este caso la condición de blow-up $p > 1$ es diferente de la del continuo (12), $p > 1/2$. El conjunto de blow-up del problema continuo es $B(u) = \{1\}$ si $p > 1$; un intervalo propio si $p = 1$; y todo el intervalo $[0, 1]$ si $1/2 < p < 1$.

En este caso, el comportamiento del problema continuo es muy diferente al del discreto ya que los casos de blow-up son diferentes e incluso cuando ambos problemas explotan los conjuntos de blow-up difieren radicalmente.

(III) $f(s) = e^s$.

En este caso las soluciones numéricas explotan en un sólo punto, $B(u_h) = \{1\}$. La tasa de blow-up es $u_N(t) \sim -\ln(T_h - t)$. Para esta no linealidad, las tasas y conjuntos de blow-up coinciden.

En [4] se extienden los resultados obtenidos en [16] para el caso multidimensional y además se obtiene la tasa de explosión y se localiza el conjunto de blow-up numérico en un entorno del borde del dominio.

Vemos entonces que al considerar esquemas de malla fija para este tipo de problemas aparecen diferencias significativas entre las aproximaciones y las soluciones continuas: los casos de blow-up no coinciden así como tampoco la tasa y los conjuntos de explosión.

3.2 Métodos numéricos que adaptan la variable espacial

En [20] se introducen dos nuevos métodos numéricos que reproducen correctamente la tasa y el conjunto de explosión para el problema (12) considerando $f(u) = u^p$ (a este problema lo denominamos $(12)_p$).

El primer método desarrollado agrega puntos a la malla a medida que el tiempo avanza. Está basado en una semidiscretización espacial y agrega los puntos cerca del borde $x = 1$ cuando la solución numérica se hace grande, produciendo una malla no uniforme que se concentra cerca de la singularidad.

El segundo método, que mueve los nodos en lugar de agregarlos, también usa una semidiscretización espacial, pero mueve los últimos K nodos que se acumulan en el borde $x = 1$ a medida que la solución se hace grande. Este procedimiento está inspirado en los *moving mesh methods*, que mueven la malla de manera de mantener la masa entre nodo y nodo equidistribuida. En este caso, se saca ventaja del conocimiento a priori de la localización espacial de la singularidad ($x = 1$), y entonces en lugar de mover toda la malla sólo se mueven los últimos nodos, manteniendo el resto de la malla fija.

Una ventaja del método que mueve puntos es que mantiene el tamaño del sistema de ecuaciones ordinarias constante en el tiempo, mientras que el método que agrega lo agranda indefinidamente a medida que el tiempo se acerca a la singularidad.

Ambos métodos están basados en la invarianza de escala de la ecuación del calor en la semirrecta con condición de borde no lineal en el extremo $x = 0$, $-u_x(0,t) = u^p(0,t)$. Es decir, si $u(x,t)$ es solución, entonces $u_\lambda(x,t) = \lambda^{\frac{1}{2(p-1)}} u(\lambda^{\frac{1}{2}} x, \lambda t)$ también lo es. En [24] se prueba que existe un perfil autosimilar que explota, de la forma

$$u_S(x,t) = (T-t)^{-\frac{1}{2(p-1)}} \varphi(\xi), \qquad \xi = x(T-t)^{-\frac{1}{2}}.$$

Para una expresión explícita del perfil φ ver [24]. Esta solución $u_S(x,t)$ da el comportamiento cerca del tiempo de blow-up T de todas las soluciones de $(12)_p$

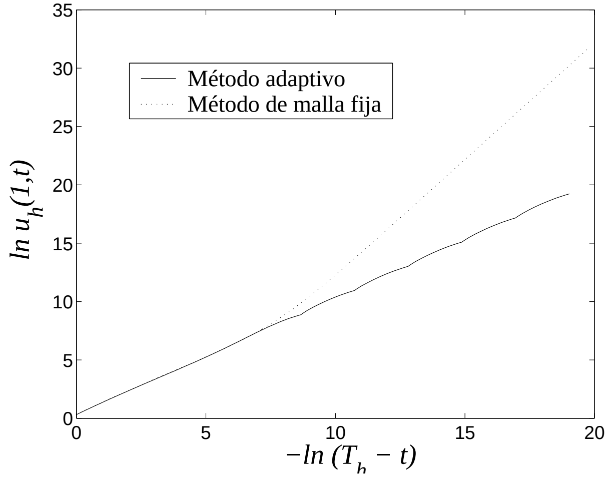


Figura 3: $p = 3/2$. Tasas de explosión para la solución numérica con malla fija y para la solución haciendo adaptividad.

en el siguiente sentido:

$$u(x,t) \sim (T-t)^{-\frac{1}{2(p-1)}}\varphi(\xi), \tag{15}$$

para $x = 1-\xi(T-t)^{1/2}$, $|\xi| \le C$. Es decir, el comportamiento de las soluciones cerca del punto $(1,T)$ está dado por la solución autosimilar en la semirrecta. Los métodos adaptivos que presentamos usan este hecho para modificar la malla de manera tal que este comportamiento sea reproducido por las aproximaciones.

Para ambos métodos se prueba convergencia en el mismo sentido que para los problemas anteriores y la misma condición (necesaria y suficiente) para la aparición de blow-up ($p > 1$) en las aproximaciones. También se prueba que para estos métodos la tasa de blow-up es

$$\lim_{t\nearrow T_h} (T_h - t)^{\frac{1}{2(p-1)}} \|u_h(\cdot,t)\|_{L^\infty(\Omega)} = \varphi(0).$$

Es decir que se recupera la tasa de blow-up que habíamos perdido con los métodos de malla fija, ver Figura 3.

Para estos métodos es posible acotar la diferencia entre T y T_h. Existen $\gamma > 0$ y $C > 0$ tales que

$$|T_h - T| \le Ch^\gamma.$$

Así como también se recupera en forma exacta el conjunto de explosión,

$$B(u_h) = \{1\}.$$

Agradecimientos. Parte de este trabajo fue realizado en varias visitas de los autores a la Universidad Autónoma de Madrid y a la Universitat de Valencia.

Agradecemos a estas instituciones por la hospitalidad y el buen ambiente de trabajo.

Es un placer agradecer a los españoles que siempre nos ayudaron, en especial a J. L. Vázquez, R. Ferreira, F. Quirós, A. de Pablo, J. M. Mazón, F. Andreu, J. Toledo, I. Peral, J. Garcia-Azorero, E. Zuazua. Y a los argentinos que siempre nos esperaron a la vuelta, R. Durán, J. Fernández Bonder, N. Wolanski, S. Martínez, G. Acosta.

Referencias

[1] L. M. Abia, J. C. López-Marcos, and J. Martínez. *Blow-up for semidiscretizations of reaction-diffusion equations.* Appl. Numer. Math. **20** (1996), no. 1-2, 145–156, Workshop on the method of lines for time-dependent problems (Lexington, KY, 1995).

[2] ______, *On the blow-up time convergence of semidiscretizations of reaction-diffusion equations.* Appl. Numer. Math. **26** (1998), no. 4, 399–414.

[3] G. Acosta, R. G. Durán, and J. D. Rossi, *An adaptive time step procedure for a parabolic problem with blow-up.* Computing **68** (2002), no. 4, 343–373.

[4] G. Acosta, J. Fernández Bonder, P. Groisman, and J. D. Rossi, *Numerical approximation of a parabolic problem with a nonlinear boundary condition in several space dimensions.* Discrete Contin. Dyn. Syst. Ser. B **2** (2002), no. 2, 279–294.

[5] G. Acosta, J. Fernández Bonder, P. Groisman, and J. D. Rossi, *Simultaneous vs. non-simultaneous blow-up in numerical approximations of a parabolic system with non-linear boundary conditions.* M2AN Math. Model. Numer. Anal. **36** (2002), no. 1, 55–68.

[6] G. Acosta, J. Fernández Bonder, and J. D. Rossi, *Stable manifold approximation for the heat equation with nonlinear boundary condition.* J. Dynam. Differential Equations **12** (2000), no. 3, 557–578.

[7] J. M. Ball, *Remarks on blow-up and nonexistence theorems for nonlinear evolution equations.* Quart. J. Math. Oxford Ser. (2) **28** (1977), no. 112, 473–486.

[8] C. Bandle and H. Brunner, *Numerical analysis of semilinear parabolic problems with blow-up solutions.* Rev. Real Acad. Cienc. Exact. Fís. Natur. Madrid **88** (1994), no. 2-3, 203–222.

[9] C. Bandle and H. Brunner, *Blowup in diffusion equations: a survey.* J. Comput. Appl. Math. **97** (1998), no. 1-2, 3–22.

[10] M. Berger and R. V. Kohn, *A rescaling algorithm for the numerical calculation of blowing-up solutions.* Comm. Pure Appl. Math. **41** (1988), no. 6, 841–863.

[11] C. J. Budd, S. Chen, and R. D. Russell, *New self-similar solutions of the nonlinear Schrödinger equation with moving mesh computations.* J. Comput. Phys. **152** (1999), no. 2, 756–789.

[12] C. J. Budd, W. Huang, and R. D. Russell, *Moving mesh methods for problems with blow-up.* SIAM J. Sci. Comput. **17** (1996), no. 2, 305–327.

[13] X. Chen and H. Matano, *Convergence, asymptotic periodicity, and finite-point blow-up in one-dimensional semilinear heat equations.* J. Differential Equations **78** (1989), no. 1, 160–190.

[14] Y. G. Chen, *Asymptotic behaviours of blowing-up solutions for finite difference analogue of $u_t = u_{xx} + u^{1+\alpha}$.* J. Fac. Sci. Univ. Tokyo Sect. IA Math. **33** (1986), no. 3, 541–574.

[15] C. Cortázar, M. del Pino, and M. Elgueta, *The problem of uniqueness of the limit in a semilinear heat equation.* Comm. Partial Differential Equations **24** (1999), no. 11-12, 2147–2172.

[16] R. G. Duran, J. I. Etcheverry, and J. D. Rossi, *Numerical approximation of a parabolic problem with a nonlinear boundary condition.* Discrete Contin. Dynam. Systems **4** (1998), no. 3, 497–506.

[17] M. Escobedo and H. A. Levine, *Critical blowup and global existence numbers for a weakly coupled system of reaction-diffusion equations.* Arch. Rational Mech. Anal. **129** (1995), no. 1, 47–100.

[18] J. Fernández Bonder, P. Groisman, and J. D. Rossi, *On numerical blow-up sets.* Proc. Amer. Math. Soc. **130** (2002), no. 7, 2049–2055 (electronic).

[19] J. Fernández Bonder and J. D. Rossi, *Blow-up vs. spurious steady solutions.* Proc. Amer. Math. Soc. **129** (2001), no. 1, 139–144.

[20] R. Ferreira, P. Groisman, and J. D. Rossi, *Adaptive numerical schemes for a parabolic problem with blow-up.* IMA J. Numer. Anal. 23 (3) (2003), 439-463.

[21] ______, *Numerical blow-up for the porous medium equation with a source.* Preprint.

[22] ______, *Numerical blow-up for a nonlinear problem with a nonlinear boundary condition.* Math. Models Methods Appl. Sci. **12** (2002), no. 4, 461–483.

[23] M. Fila and J. Filo, *Blow-up on the boundary: a survey.* Singularities and differential equations (Warsaw, 1993), Banach Center Publ., vol. 33, Polish Acad. Sci., Warsaw, 1996, pp. 67–78.

[24] M. Fila and P. Quittner, *The blow-up rate for the heat equation with a nonlinear boundary condition.* Math. Methods Appl. Sci. **14** (1991), no. 3, 197–205.

[25] H. Fujita, *On the blowing up of solutions of the Cauchy problem for* $u_t = \Delta u + u^{1+\alpha}$. J. Fac. Sci. Univ. Tokyo Sect. I **13** (1966), 109–124 (1966).

[26] V. A. Galaktionov and J. L. Vázquez, *The problem of blow-up in nonlinear parabolic equations.* Discrete Contin. Dyn. Syst. **8** (2002), no. 2, 399–433, Current developments in partial differential equations (Temuco, 1999).

[27] ———, *Continuation of blow-up solutions of nonlinear heat equations in several space dimensions.* Commun. Pure Applied Math. 50, (1997), 1-67.

[28] ———, *Necessary and sufficient conditions for complete blow-up and extinction for one-dimensional quasilinear heat equations.* Arch. Rational Mech. Anal. 129 (1995), 225–244.

[29] Y. Giga and R. V. Kohn, *Nondegeneracy of blowup for semilinear heat equations.* Comm. Pure Appl. Math. **42** (1989), no. 6, 845–884.

[30] P. Groisman, F. Quirós, and J. D. Rossi, *Non-simultaneous blow-up in a numerical approximation of a parabolic system.* Comput. Appl. Math., 21 (3) (2002), 813-831.

[31] P. Groisman, *Adapting the time-step to recover the asymptotic behavior in a blow-up problem.* Preprint.

[32] P. Groisman and J. D. Rossi, *Dependence of the blow-up time with respect to parameters and numerical approximations for a parabolic problem.* Asymptotic Analysis, to appear.

[33] ———, *Asymptotic behaviour for a numerical approximation of a parabolic problem with blowing up solutions.* J. Comput. Appl. Math. **135** (2001), no. 1, 135–155.

[34] P. Groisman, J. D. Rossi, and H. Zaag, *On the dependence of the blow-up time with respect to the intial data in a semilinear parabolic problem.* Comm. Partial Differential Equations, 28 (3&4) (2003), 737-744.

[35] M. A. Herrero and J. J. L. Velázquez, *Flat blow-up in one-dimensional semilinear heat equations.* Differential Integral Equations **5** (1992), no. 5, 973–997.

[36] ———, *Generic behaviour of one-dimensional blow up patterns.* Ann. Scuola Norm. Sup. Pisa Cl. Sci. (4) **19** (1992), no. 3, 381–450.

[37] ———, *Blow-up behaviour of one-dimensional semilinear parabolic equations.* Ann. Inst. H. Poincaré Anal. Non Linéaire 10 (1993), 131–189.

[38] B. Hu and H.-M. Yin, *The profile near blowup time for solution of the heat equation with a nonlinear boundary condition.* Trans. Amer. Math. Soc. **346** (1994), no. 1, 117–135.

[39] W. Huang, Y. Ren, and R. D. Russell, *Moving mesh partial differential equations (MMPDES) based on the equidistribution principle.* SIAM J. Numer. Anal. **31** (1994), no. 3, 709–730.

[40] S. Kaplan, *On the growth of solutions of quasi-linear parabolic equations.* Comm. Pure Appl. Math. **16** (1963), 305–330.

[41] M.-N. Le Roux, *Semidiscretization in time of nonlinear parabolic equations with blowup of the solution.* SIAM J. Numer. Anal. **31** (1994), no. 1, 170–195.

[42] J. López-Gómez, V. Márquez, and N. Wolanski, *Blow up results and localization of blow up points for the heat equation with a nonlinear boundary condition.* J. Differential Equations **92** (1991), no. 2, 384–401.

[43] F. Merle, *Solution of a nonlinear heat equation with arbitrarily given blow-up points*, Comm. Pure Appl. Math. **45** (1992), no. 3, 263–300.

[44] T. Nakagawa, *Blowing up of a finite difference solution to $u_t = u_{xx} + u_2$.* Appl. Math. Optim. **2** (1975/76), no. 4, 337–350.

[45] F. Quirós and J. D. Rossi, *Non-simultaneous blow-up in a semilinear parabolic system.* Z. Angew. Math. Phys. **52** (2001), no. 2, 342–346.

[46] F. Quirós, J. D. Rossi and J. L. Vazquez. *Complete blow-up and thermal avalanche for heat equations with nonlinear boundary conditions.* Comm. Partial Differential Equations **27** (2002), no. 1-2, 395–424.

[47] D. F. Rial and J. D. Rossi, *Blow-up results and localization of blow-up points in an N-dimensional smooth domain.* Duke Math. J. **88** (1997), no. 2, 391–405.

[48] ———, *Localization of blow-up points for a parabolic system with a nonlinear boundary condition.* Rend. Circ. Mat. Palermo (2) **48** (1999), no. 1, 135–152.

[49] A. A. Samarskii, V. A. Galaktionov, S. P. Kurdyumov, and A. P. Mikhailov, *Blow-up in quasilinear parabolic equations.* de Gruyter Expositions in Mathematics, vol. 19, Walter de Gruyter & Co., Berlin, 1995, Translated from the 1987 Russian original by Michael Grinfeld and revised by the authors.

[50] T. K. Ushijima, *On the approximation of blow-up time for solutions of nonlinear parabolic equations.* Publ. Res. Inst. Math. Sci. **36** (2000), no. 5, 613–640.

[51] J. J. L. Velázquez, *Classification of singularities for blowing-up solutions in higher dimensions.* Trans. Amer. Math. Soc. 338 (1993), 441–464.

[52] W. Walter, *On existence and nonexistence in the large of solutions of parabolic differential equations with a nonlinear boundary condition.* SIAM J. Math. Anal. **6** (1975), 85–90.

Bol. Soc. Esp. Mat. Apl. n°26(2003), 57–78

Hydrodynamic limits of the Boltzmann equation : Recent developments

N. Masmoudi

Courant Institute, New York, USA

masmoudi@cims.nyu.edu

Abstract

From a physical point of view, we expect that a gas can be described by a fluid equation when the mean free path (Knudsen number) goes to zero. In his sixth problem, Hilbert asked for a full mathematical justification of these derivations. During the last two decades this problem got a lot of interest and specially after DiPerna and Lions constructed their renormalized solutions [10]. In this review paper, we present some of the most recent results concerning these (rigorous) derivations. We will present results for the three most classical equations of fluid mechanics in the incompressible regime, namely the incompressible Navier-Stokes equation, the Stokes equation and the Euler equation.

We will also present a new result about the derivation of Fluid Mechanic boundary conditions starting from kinetic boundary conditions [29].

Key words: *Boltzmann equation, hydrodynamic limit, fluid mechanics*
AMS subject classifications: *35Q35, 35Q30, 82C40.*

1 Introduction

In his sixth problem, Hilbert asked for a full mathematical justification of fluid mechanics equations starting from particle systems [17]. If we take the Boltzmann equation as a starting point, this problem can be stated as an asymptotic problem. Namely, starting from the Boltzmann equation, can we derive fluid mechanics equations and in which regime ?

A program in this direction was initiated by Bardos, Golse and Levermore [1] who, using the the renormalized solutions to the Boltzmann equation

Fecha de recepción: 3 de marzo de 2003

constructed by DiPerna and Lions, set a different asymptotic regime where one can derive different fluid equations (and in particular incompressible models) depending on the chosen scaling. In [1], the authors also gave a “rigorous” justification under some unverified assumptions.

During the last few years, there were many works trying to remove these extra assumptions and specially [14] were a whole rigorous derivation is given for some special Boltzmann kernels.

In this review paper, we want to give an idea about the latest results in this subject (see also [36, 28]). In particular, we will state the result of [14] as well as the results of [29], where the derivation of fluid mechanics boundary conditions starting from kinetic boundary conditions is proved.

1.1 The Boltzmann equation

The Boltzmann equation describes the evolution of the particle density of a rarefied gas. Indeed, the molecules of a gas can be modeled by hard spheres that move according to the laws of classical mechanics. However, due to the enormous number of molecules (about 2.7 10^{19} molecules in a cubic centimeter of gas at 1 atm and 0^0 C), it seems difficult to describe the state of the gas by giving the position and velocity of each individual particle. Hence, we must use some statistics and instead of giving the position and velocity of each particle, we specify the density of particles $F(x, v)$ at each point x and velocity v. This means that we describe the gas by giving for each point x and velocity v the number of particles $F(x, v)\ dx\ dv$ in the volume $(x, x + dx) \times (v, v + dv)$.

Under some assumptions (rarefied gas, ...), it is possible to derive (at least formally) the Boltzmann equation from the classical Newton laws in an asymptotic regime where the number of particles goes to infinity. The goal of this review is not to explain the derivation of the Boltzmann equation but rather use it as a starting point to derive the classical equations of fluid mechanics (see [19], [33] and [8] for some rigorous results about the derivation of the Boltzmann equation starting from N particles system). The Boltzmann equation reads

$$\partial_t F + v.\nabla_x F = B(F, F) \tag{1}$$

where the collision kernel $B(F, F)$ is a quadratic form which acts only on the v variable. It describes the possible interaction between two different particles and is given by

$$B(F, F)(v) = \int_{\mathbb{R}^D} \int_{S^{D-1}} (F_1'F' - F_1F) b(v - v_1, \omega) dv_1 d\omega \tag{2}$$

where we have used the following notation for all function ϕ

$$\phi' = \phi(v'), \quad \phi_1 = \phi(v_1), \quad \phi_1' = \phi(v_1'), \tag{3}$$

and where the primed speeds are given by

$$v' = v + \omega[\omega.(v_1 - v)], \qquad v_1' = v - \omega[\omega.(v_1 - v)]. \tag{4}$$

Moreover, the Boltzmann cross-section $b(z,\omega)$ ($z \in \mathbb{R}^D, \omega \in S^{D-1}$) depends on the molecular interactions (intermolecular potential). It is a nonnegative, locally integrable function (at least when grazing collisions are neglected). The Galilean invariance of the collisions implies that b depends only on $v - v_1, \omega$ and that

$$b(z,\omega) = |z|\S(|z|, |\mu_c|), \quad \mu_c = \frac{\omega.(v_1 - v)}{|v_1 - v|}, \tag{5}$$

where $\S$ is the specific differential cross-section. We also insist on the fact that the relations (4) are equivalent to the following conservations

$$v' + v_1' = v + v_1 \quad \text{(conservation of the moment)} \tag{6}$$

$$|v'|^2 + |v_1'|^2 = |v|^2 + |v_1|^2 \quad \text{(conservation of the kinetic energy)} \tag{7}$$

We notice that the fact that two particles give two particles after the interaction translates the conservation of mass. For a more precise discussion about the model, we refer to [7], [8] and [37]. We also refer to [21, 22, 34, 27] for mathematical results on different fluid mechanic equations. For some numerical works on the hydrodynamic limit, we refer to [32].

1.2 Compressible Euler

We start here by explaining how one can derive (at least formally) the Compressible Euler equation from the Boltzmann equation. A rigorous derivation can be found in Caflisch [6]. If F satisfies the Boltzmann equation, we deduce by integration in the v variable (at least formally) the following local conservations

$$\begin{cases} \partial_t\big(\int_{\mathbb{R}^D} F\ dv\big) + \nabla_x.\big(\int_{\mathbb{R}^D} v\ F\ dv\big) = 0 \\[2ex] \partial_t\big(\int_{\mathbb{R}^D} vF\ dv\big) + \nabla_x.\big(\int_{\mathbb{R}^D} v \otimes v\ F\ dv\big) = 0 \\[2ex] \partial_t\big(\int_{\mathbb{R}^D} |v|^2 F\ dv\big) + \nabla_x.\big(\int_{\mathbb{R}^D} v|v|^2\ F\ dv\big) = 0 \end{cases} \tag{8}$$

These three equations describe respectively the conservation of mass, momentum and energy. They present a great resemblance with the compressible Euler equation. However, the third moment $\int_{\mathbb{R}^D} v|v|^2\ F\ dv$ is not a function of the others and depends in general on the whole distribution $F(v)$. In the asymptotic regimes we want to study, the distribution $F(v)$ will be a very close to a Maxwellian due to the fact that the Knudsen number is going to 0. If we make the assumption that $F(v)$ is a Maxwellian for all t and x, then the third moment $\int_{\mathbb{R}^D} v|v|^2\ F\ dv$ can be given as a function of $\rho = \int_{\mathbb{R}^D} F\ dv$, $\rho u = \int_{\mathbb{R}^D} vF\ dv$ and $\rho(\frac{1}{2}|u|^2 + \frac{D}{2}\theta) = \int_{\mathbb{R}^D} \frac{1}{2}|v|^2 F\ dv$. Moreover, for all i and j, $\int_{\mathbb{R}^D} v_i v_j F\ dv$ can also be expressed as a function of ρ, u and θ.

We recall that a Maxwellian $M_{\rho,u,\theta}$ is completely defined by its density, bulk velocity and temperature

$$M_{\rho,u,\theta} = \frac{\rho}{(2\pi\theta)^{D/2}}\exp(-\frac{1}{2\theta}|v-u|^2) \tag{9}$$

where ρ, u and θ depend only on t and x. If, we assume that for all t and x, F is a Maxwellian given by $F = M_{\rho(t,x),u(t,x),\theta(t,x)}$ then (8) reduces to

$$\begin{cases} \partial_t \rho + \nabla_x . \rho u = 0 \\ \partial_t(\rho u) + \nabla_x.(\rho u \otimes u) + \nabla_x(\rho\theta) = 0 \\ \partial_t\big(\frac{1}{2}\rho|u|^2 + \frac{D}{2}\rho\theta\big) + \nabla_x.\big(\rho u(\frac{1}{2}|u|^2 + \frac{D+2}{2}\theta)\big) = 0 \end{cases} \tag{10}$$

which is the compressible Euler system for a mono-atomic perfect gas. This derivation can become rigorous, if we take a sequence of solutions F_ϵ of

$$\partial_t F_\epsilon + v.\nabla_x F_\epsilon = \frac{1}{\epsilon} B(F_\epsilon, F_\epsilon) \tag{11}$$

where ϵ is the Knudsen number which goes to 0 (see R. Caflisch [6]). Formally the presence of the term $\frac{1}{\epsilon}$ in front of $\frac{1}{\epsilon}B(F_\epsilon, F_\epsilon)$ implies (at the limit) that $B(F, F) = 0$ which means that F is a Maxwellian (see [7], [8] or [37] for a proof of this fact).

1.3 Incompressible scalings

In the last subsection, we explained how we can derive the compressible Euler equation. It turns out that using different scalings, one can also derive incompressible models. We will explain what these scalings mean concerning the the Knudsen, Reynolds and Mach numbers. We consider the following global Maxwellian M which corresponds to $\rho = \theta = 1$ and $u = 0$.

$$M(v) = \frac{1}{(2\pi)^{D/2}} \exp(-\frac{1}{2}|v|^2). \tag{12}$$

Let $F_\epsilon = MG_\epsilon = M(1+\epsilon^m g_\epsilon)$ be a solution of the following Boltzmann equation

$$\epsilon^s \partial_t F_\epsilon + v.\nabla F_\epsilon = \frac{1}{\epsilon^q} B(F_\epsilon, F_\epsilon) \tag{13}$$

which is also equivalent to

$$\epsilon^s \partial_t G_\epsilon + v.\nabla G_\epsilon = \frac{1}{\epsilon^q} Q(G_\epsilon, G_\epsilon) \tag{14}$$

where

$$Q(G,G)(v) = \int_{\mathbb{R}^D} \int_{S^{D-1}} (G_1'G' - G_1G) b(v - v_1, \omega) M_1 dv_1 d\omega. \tag{15}$$

With this scaling, we can define

$$Ma = \epsilon^m, \quad Kn = \epsilon^q, \quad Re = \epsilon^{m-q}. \tag{16}$$

Here ϵ^s is a time scaling which allows us to choose the phenomenon we want to emphasize. By varying m, q and s, we can formally derive the following systems (see the references below for some rigorous mathematical results). A part from the first case where the compressible Euler system is satisfied by the moments of F, the fluid equation are recovered for the moments of the fluctuation g and we can show at least formally that $g = \rho + u.v + \theta(\frac{|v|^2}{2} - \frac{D}{2})$ where (ρ, u, θ) satisfies one of the above equations

1) $q = 1,\ m = 0,\ s = 0$ Compressible Euler system [6, 18, 35]

2) $q = 1,\ m > 0,\ s = 0$ Acoustic waves [3]

$$\begin{cases} \partial_t \rho + \nabla_x . u = 0 \\ \partial_t u + \nabla_x(\rho + \theta) = 0 \\ \partial_t(\rho + \theta) + \frac{D+2}{D} \nabla_x . u = 0 \end{cases} \tag{17}$$

3) $q = 1,\ m = 1,\ s = 1$ Incompressible Navier-Stokes-Fourier system [9, 1, 5, 25, 15]

$$\begin{cases} \partial_t u + u.\nabla u - \nu \Delta u + \nabla p &= 0, \quad \nabla_x . u = 0 \\ \\ \partial_t \theta + u.\nabla \theta - \kappa \Delta \theta &= 0 \end{cases}$$

4) $q = 1,\ m > 1,\ s = 1$ Stokes-Fourier system [2, 3, 26, 11, 29]

$$\begin{cases} \partial_t u - \nu \Delta u + \nabla p &= 0, \quad \nabla_x . u = 0 \\ \\ \partial_t \theta - \kappa \Delta \theta &= 0 \end{cases}$$

5) $q > 1,\ m = 1,\ s = 1$ Incompressible Euler-Fourier system [26, 31]

$$\begin{cases} \partial_t u + u.\nabla u + \nabla p &= 0, \quad \nabla_x . u = 0 \\ \\ \partial_t \theta + u.\nabla \theta &= 0. \end{cases}$$

Note that the compressible Navier-Stokes system (with a viscosity of order 1) can not be derived in this manner because of the following physical relation

$$Re = C \frac{Ma}{Kn}. \tag{18}$$

However, the compressible Navier-Stokes system with a viscosity of order ϵ can be considered as a better approximation than the Compressible Euler system in the case $q = 1,\ m = 0,\ s = 0$.

1.4 Formal development

Here, we want to explain (at least formally) how we can derive the incompressible Navier-Stokes system for the bulk velocity and the Fourier equation for the temperature starting from the Boltzmann system with the scalings $q = 1,\ m = 1,\ s = 1$. A simple adaptation of the argument also yields a formal derivation of the Stokes-Fourier system (which is the linearization of the Navier-Stokes-Fourier system) as well as the Euler. Rewriting the equation satisfied by g_ϵ, we get

$$\partial_t g_\epsilon + \frac{1}{\epsilon} v.\nabla_x g_\epsilon = -\frac{1}{\epsilon^2} L g_\epsilon + \frac{1}{\epsilon} Q(g_\epsilon, g_\epsilon) \tag{19}$$

where L is the linearized collision operator given by

$$Lg = \int_{\mathbb{R}^D} \int_{S^{D-1}} (g + g_1 - g_1' - g') b(v - v_1, \omega) M_1 dv_1 \, d\omega \tag{20}$$

We assume that g_ϵ can be decomposed as follows $g_\epsilon = g + \epsilon h + \epsilon^2 k + O(\epsilon^3)$ and we make the following formal development

$$\frac{1}{\epsilon^2} : \quad Lg = 0. \tag{21}$$

A simple study of the operator L shows that it is formally self-adjoint, non negative for the following scalar product $< f, g >= \langle f \;\, g \rangle$ where we use the following notation $\langle g \rangle = \int_{\mathbb{R}^D} gMdv$ and $Ker(L) = \{g, \;\, g = \alpha + \beta.v + \gamma |v|^2, \;\; \text{where} \;\; (\alpha, \beta, \gamma) \in \mathbb{R} \times \mathbb{R}^D \times \mathbb{R}\}$. Hence, we deduce that $g = \rho + u.v + \theta(\frac{|v|^2}{2} - \frac{D}{2})$.

$$\frac{1}{\epsilon} : \; v.\nabla g = -Lh + Q(g, g). \tag{22}$$

Integrating over v, we infer that $u = \langle vg \rangle$ is divergence-free (div $u = 0$). Moreover, multiplying by v and taking the integral over v, we infer that $\nabla(\rho + \theta) = 0$. Besides, at order 1, we have

$$\frac{1}{\epsilon^0} \quad : \quad \partial_t g + v.\nabla_x h = -Lk + 2Q(g, h), \tag{23}$$

from which we deduce that

$$\frac{1}{\epsilon^0} \quad : \quad \partial_t \langle vg \rangle + \nabla_x.\langle v \otimes vh \rangle = 0, \tag{24}$$

$$\frac{1}{\epsilon^0} \quad : \quad \partial_t \langle (\frac{|v|^2}{D+2} - 1) g \rangle + \nabla_x.\langle v(\frac{|v|^2}{D+2} - 1) h \rangle = 0. \tag{25}$$

To get a closed equation for g, we have to inverse the operator L. We define the matrix $\phi(v)$ and the vector $\psi(v)$ as the unique solutions of

$$L\phi(v) = v \otimes v - \frac{1}{D} |v|^2 I, \qquad L\psi(v) = (\frac{|v|^2}{D+2} - 1) v \tag{26}$$

which are orthogonal to $Ker(L)$ for the scalar product $< \cdot, \cdot >$. We also define the viscosity ν and the heat conductivity κ by

$$\nu = \frac{1}{(D-1)(D+2)} \langle \phi : L\phi \rangle, \tag{27}$$

$$\kappa = \frac{2}{D(D+2)} \langle \psi . L\psi \rangle . \tag{28}$$

We notice that ν and κ only depend on b. Using that L is formally self-adjoint, we deduce that

$$\partial_t \langle g v_i \rangle + \nabla_x . \Big\langle \phi_{ij} (Q(g,g) - v.\nabla g) \Big\rangle + \nabla \ \langle \frac{|v|^2}{N} h \rangle = 0 \tag{29}$$

$$\partial_t \langle g (\frac{|v|^2}{D+2} - 1) \rangle + \nabla_x . \Big\langle \psi (Q(g,g) - v.\nabla g) \Big\rangle = 0 \tag{30}$$

A simple (but long) computation gives the Navier-Stokes equation and the Fourier equation, namely

$$\partial_t u + u.\nabla u - \nu \Delta u + \nabla p = 0 \tag{31}$$

$$\partial_t \theta + u.\nabla \theta - \kappa \Delta \theta = 0 \tag{32}$$

where $u = \langle g v \rangle$, $\theta = -\rho = \langle (\frac{|v|^2}{D+2} - 1) g \rangle$ and the pressure p is the sum of different contributions.

1.5 Mathematical difficulties

Here, we want to explain the major mathematical difficulties encountered in trying to give a rigorous justification of any of the above asymptotic problems.

D1. The local conservation of momentum is not known to hold for the renormalized solutions of the Boltzmann equation. Indeed, the solutions constructed by R. DiPerna and P.-L. Lions [10] only hold in the renormalized sense which means that

$$\begin{aligned} \partial_t \beta(F) + v.\nabla \beta(F) &= Q(F,F)\beta'(F), & (33) \\ \beta(F)(t=0) &= \beta(F^0) & (34) \end{aligned}$$

and where β is given, for instance, by $\beta(f) = Log\ (1+f)$.

D2. The lack of a priori estimates. Indeed, all we can deduce from the entropy inequality and the conservation of energy is that g_ϵ is bounded in $L\log L$ and that $g_\epsilon |v|^2$ is bounded in L^1. However, we need a bound in L^2 to define all the product involved in the formal development. In [11], the authors used the entropy dissipation estimate to deduce some information on the structure of the fluctuation g_ϵ and get some new a priori estimates by using some Caflisch-Grad estimates.

To pass to the limit in the different products (and specially in the case we want to recover the Navier-Stokes-Fourier system or the Euler system), one has also to prove that g_ϵ is compact in space and time, namely that $g_\epsilon \in K$ where K is a compact subset of some $L^p(0,T; L^1(\Omega))$. We split this in two difficulties

D3. The compactness in space of g_ϵ. This was achieved in the stationary case by C. Bardos, F. Golse and D. Levermore [4], [1] using the averaging lemma and proving that g_ϵ is in some compact subset of $L^1(\Omega)$. However, a newer version of the averaging lemma was needed in [14] to prove some equiintegrability and hence the absence of concentration.

D4. The compactness in time for g_ϵ. It turns out that in general g_ϵ is not compact in time. Indeed, g_ϵ presents some oscillations in time which can be analyzed and described precisely. Using this description and some compensation (due to a remarkable identity satisfied by the solutions to the wave equation), it is possible to pass to the limit in the whole equation. This was done by P.-L. Lions and the author [25] using some ideas coming from the compressible-incompressible limit [23, 24].

2 The convergence towards the incompressible Navier-Stokes-Fourier system

The first paper dealing with the rigorous justification of the formal development 1.4 goes back to the work of C. Bardos, F. Golse and D. Levermore [1] where the stationary case was handled under different assumptions and restrictions (see also A. De Masi, R. Esposito, and J. L. Lebowitz [9] for a similar result in a different setting). There are however some aspects of the analysis performed in [1] that can be improved. First, the heat equation was not treated because the heat flux terms could not be controlled. Second, local momentum conservation was assumed because DiPerna-Lions solutions are not known to satisfy the local conservation law of momentum (or energy) that one would formally expect. Third, the discrete-time case was treated in order to avoid having to control the time regularity of the acoustic modes. Fourth, unnatural technical assumptions were made on the Boltzmann kernel. Finally, a mild compactness assumption was required to pass to the limit in certain nonlinear terms.

During the last three years, there appeared several results trying to give a rigorous justification of this derivation. In collaboration with P.-L. Lions [25] and under two assumptions (the conservation of the momentum and a compactness assumption) we were able to treat the time dependent case and derive the incompressible Navier-Stokes equation. In [11], Golse and Levermore gave a rigorous derivation of Stokes-Fourier system (the linearization of the Navier-Stokes-Fourier system) without any assumption (see also next section). In a recent work in collaboration with Levermore [20], we give a derivation of the Navier-Stokes-Fourier system under a compactness assumption for several types of Boltzmann kernels.

The most important contribution in this subject is the paper of Golse and Saint-Raymond, where a complete derivation is given for some special Boltzmann kernels [14]. We will discuss this result in the next subsections.

In what follows, we assume that Ω is the whole space or the torus to avoid dealing with the boundary. First, let us specify the conditions we impose on the initial data. It is supposed that G_ϵ^0 satisfies (we recall that $F_\epsilon^0 = MG_\epsilon^0$)

$$H(G^0_\epsilon) = \int_\Omega \int_{\mathbb{R}^D} (G^0_\epsilon log G^0_\epsilon - G^0_\epsilon + 1) M \; dxdv \leq C\epsilon^2 \tag{35}$$

This shows that we can extract a subsequence of the sequence g^0_ϵ (defined by $G^0_\epsilon = 1 + \epsilon g^0_\epsilon$) which converges weakly in L^1 towards g^0 such that $g^0 \in L^2$. We also notice that (35) is equivalent to the fact that $\int_\Omega \langle h(\epsilon g^0_\epsilon)\rangle \; dx \leq C\epsilon^2$, where $h(z) = (1+z)\log(1+z) - z$ which is almost an L^2 estimate for g^0_ϵ. This shows at least that $g^0 \in L^2$. Then, we consider a sequence G_ϵ of renormalized solutions of the Boltzmann equation (B_ϵ), satisfying the entropy inequality and we want to prove that g_ϵ converges to some $g = u.v + \theta(\frac{|v|^2}{2} - \frac{D+2}{2})$.

Before stating the new result of Golse and Saint-Raymond [14], we want to explain the kind of assumptions that were made in previous works. The convergence result proved in [25] (which only deals with the u component) requires the following two hypotheses (A1) and (A2) on the sequence G_ϵ which allow to circumvent the difficulties D1 and D2

(A1). The solution G_ϵ satisfies the projection on divergence-free vector fields of the local momentum conservation law

$$\partial_t P\langle vG_\epsilon\rangle + \frac{1}{\epsilon} P\nabla_x.\langle v\otimes vG_\epsilon\rangle = 0. \tag{36}$$

(A2). The family $(1+|v|^2)g^2_\epsilon/N_\epsilon$ is relatively compact for the weak topology of $L^1(dt\ M\ dv\ dx)$ which we denote $w - L^1(dt\ M\ dv\ dx)$, where $N_\epsilon = 1 + \frac{\epsilon}{3} g_\epsilon$.

In the sequel, we denote the weak topology of $L^1(dt\ M\ dv\ dx)$ by $w - L^1(dt\ M\ dv\ dx)$. The assumption (A2) enforces the $L\log L$ estimate we have on g_ϵ, namely $\int_\Omega \langle h(\epsilon g_\epsilon)\rangle \; dx \leq C\epsilon^2$ to prevent some type of concentration.

Finally, some assumptions were also made on the collision kernel b. In the preprint [20], we were able to remove the first assumption and also to deal with more general Boltzmann kernels see [11] and [20] where it is assumed (in the case of hard interparticle potential) that there exist $C_b \in (0,\infty)$ and $\eta \in [0,1]$ such that b satisfies

$$\int_{\mathbb{S}^{D-1}} b(\omega, v)\, d\omega \leq C_b\big(1 + \frac{1}{2}|v|^2\big)^\eta \quad \text{almost everywhere}\,. \tag{37}$$

In the next subsection, we present the result of Golse and Saint-Raymond where the assumption (A2) was also removed.

2.1 The result and a sketch of the proof

Under some assumptions on the Boltzmann kernel, we have

Theorem 1 *Let G_ϵ be a sequence of renormalized solutions of the Boltzmann equations (B_ϵ) with initial condition G^0_ϵ and satisfying the entropy inequality. Then, the family $(1+|v|^2)g_\epsilon$ is relatively compact in $w - L^1(dt\ Mdv\ dx)$. If g is a weak limit of a subsequence (still denoted g_ϵ) then $Lg = 0$ and*

$g = \rho + u.v + \theta(\frac{|v|^2}{2} - \frac{D}{2})$ satisfies the limiting dissipation inequality

$$\frac{1}{2}\int_\Omega |\rho(t)|^2 + |u(t)|^2 + \frac{D}{2}|\theta(t)|^2 \; dx + \tag{38}$$

$$+\int_0^t \int_\Omega \frac{1}{2}\nu|\nabla_x u +^t \nabla_x u|^2 + \kappa|\nabla\theta|^2$$

$$\leq \liminf_{\epsilon\to 0} \frac{1}{\epsilon^2}\int_\Omega \langle h(\epsilon g_\epsilon)\rangle dx = C^0 \tag{39}$$

Moreover, $\theta + \rho = 0$ and $(u, \theta) = (\langle vg\rangle, \langle(\frac{|v|^2}{D+2} - 1)g\rangle)$ is a weak solution of the Navier-Stokes-Fourier system (NSF)

$$(NSF)\begin{cases} \partial_t u + u.\nabla u - \nu\Delta u + \nabla p = 0, \quad \nabla.u = 0 \\ \partial_t\theta + u.\nabla\theta - \kappa\Delta\theta = 0, \\ u(t=0,x) = u^0(x) \qquad \theta(t=0,x) = \theta^0(x) \end{cases}$$

with the initial condition $u^0 = P\langle vg^0\rangle$ and $\theta^0 = \langle(\frac{|v|^2}{D+2} - 1)g^0\rangle$ and where the viscosity ν and heat conductivity κ are given by (27) and (28).

Now, we give an idea of the proof of theorem 1 (see [14] for a more complete proof). We start by recalling a few a prior estimates taken from [1]

Proposition 2 *We have*

i) The sequence $(1+|v|^2)g_\epsilon$ is bounded in $L^\infty(dt; L^1(Mdv\ dx))$ and relatively compact in $w - L^1(dt\ Mdv\ dx)$. Moreover, if g is the weak limit of any converging subsequence of g_ϵ, then $g \in L^\infty(dt; L^2(Mdv\ dx))$ and for almost every $t \in [0,\infty)$, we have

$$\frac{1}{2}\int_\Omega \langle g^2(t)\rangle \; dx \leq \liminf_{\epsilon\to 0}\frac{1}{\epsilon^2}\int_\Omega \langle h(\epsilon g_\epsilon(t))\rangle dx \leq C^0. \tag{40}$$

ii)Denoting $q_\epsilon = \frac{1}{\epsilon^2}(G'_{\epsilon 1}G'_\epsilon - G_{\epsilon 1}G_\epsilon)$, we have that the sequence $(1 + |v|^2)q_\epsilon/N_\epsilon$ is relatively compact in $w - L^1(dt\ d\mu\ dx))$ where $d\mu = b(v - v_1, \omega)d\omega M_1\ dv_1 M\ dv$. Besides, if q is the weak limit of any converging subsequence of q_ϵ/N_ϵ then $q \in L^2(dt; L^2(d\mu\ dx))$ and q inherits the same symmetries as q_ϵ, namely $q(v, v_1, \omega) = q(v_1, v, \omega) = -q(v', v'_1, \omega)$.

iii) In addition, for almost all (t, x), $Lg = 0$, which means that g is of the form

$$g(t,x,v) = \rho(t,x) + u(t,x).v + \theta(t,x)(\frac{1}{2}|v|^2 - \frac{D}{2}), \tag{41}$$

where $\rho, u, \theta\ \in L^\infty(dt; L^2(dx))$.

iv) Finally, from the renormalized equation, we deduce that

$$v.\nabla_x g = \int\int q b(v_1 - v, \omega)d\omega M_1 dv_1 \tag{42}$$

which yields the incompressibility and Boussinesq relations, namely

$$\nabla_x.u = 0, \quad \nabla_x(\rho + \theta) = 0. \tag{43}$$

The new ingredients of [14] can be summarized in the following three Propositions. First, we define the following class of bump functions

$$\Upsilon = \{\gamma : \mathbb{R}_+ \to [0,1] \mid \gamma \in C^1, \gamma([\frac{3}{4}, \frac{5}{4}]) = \{1\},\ \mathrm{supp}(\gamma) \in [\frac{1}{2}, \frac{3}{2}]\}.$$

The so-called flat-sharp decomposition of [14] is given by $g_\epsilon = \gamma(G_\epsilon)g_\epsilon + (\gamma(G_\epsilon) - 1)g_\epsilon$. Notice that the entropy bound yields an $L^\infty(dt; L^2(dx; L^2(Mdv)))$ bound on $\gamma(G_\epsilon)g_\epsilon$.

Proposition 3 *For each $\gamma \in \Upsilon$, we have*
1) The sequence $|\gamma(G_\epsilon)g_\epsilon|^2 M(v)$ is equiintegrable in the v variable
2) $(1 + |v|^s)\gamma(G_\epsilon)g_\epsilon$ is bounded in $L^2_{loc}(dt\,dx\,; L^2(Mdv))$ for all $s \geq 0$.
3) $(1 + |v|^s)\frac{1-\gamma(G_\epsilon)}{\epsilon^2}$ is bounded in $L^1_{loc}(dt\,dx\,; L^1(Mdv))$ for all $s \geq 0$.

The proof of the Proposition 3 uses several ingredients and specially a dissipation-based estimate which is more compatible with the entropy dissipation as well as the so-called Caflisch-Grad estimates. The next Proposition allows to relax the assumption A2.

Proposition 4 *For each $\gamma \in \Upsilon$, we have*
1)

$$\frac{1 - \gamma(G_\epsilon)}{\epsilon} g_\epsilon = O\left(\frac{1}{log|log\epsilon|}\right) \quad in\ L^1_{loc}(dtdx; L^1(Mdv)).$$

2) For each compact $K \subset \mathbb{R}_+ \times \mathbb{R}^D$ and each sequence $\epsilon \to 0$, the extracted sequence $(t, x, v) \mapsto 1_K |\gamma(G_\epsilon)g_\epsilon|^2 M(v)$ is equiintegrable in $\mathbb{R}_+ \times \mathbb{R}^D \times \mathbb{R}^D$ with respect to the measure $dtdxMdv$.

The proof of Proposition 3 is based on a new limiting case of velocity averaging in L^1 [15] which allows to enforce the equiintegrability in the v variable stated in Proposition 4 to an equiintegrability in all the variables. We also refer to [13, 12] for early version of the velocity averaging lemma. Now, we state a Proposition which allows to relax the conservation of momentum assumption A1 (see [14] and also [20] where a different proof is given using more deeply the symmetry of the Boltzmann kernel)

Proposition 5 *for all $\gamma \in \tilde{\Upsilon}$ and $\xi(v) = 1$ or $\xi(v) = v$ or $\xi(v) = |v|^2$, we have*

$$\partial_t \langle \gamma(G_\epsilon) g_\epsilon \xi \rangle + \frac{1}{\epsilon} \div .\langle v \gamma(G_\epsilon) g_\epsilon \xi \rangle \ \to\ 0$$

in $L^1_{loc}(\mathbb{R}_+ \times \mathbb{R}^D)$ as ϵ goes to 0.

3 The convergence towards the Stokes system

The convergence towards the Stokes system is easier than the Navier-Stokes case for two reasons. Indeed, we do not have to pass to the limit in the nonlinear terms. Besides, the control we get from the entropy dissipation is better. In this section, we want to present the result of [26] where a new notion of renormalized solution was used. In [11], the whole Stokes-Fourier system was recovered by using a different idea. That idea was also used in the Navier-Stokes-Fourier case. We will not discuss the result of [11] here even though it gives a better result.

3.1 Defect measures

In [26], we have overcome the difficulty D1 by showing that the conservation of momentum can be recovered in the limit by a very simple argument. Indeed by looking at the construction of the renormalized solutions of DiPerna-Lions [10], one sees that one can write a kind of conservation of moment (with a defect measure) which also intervenes in the energy inequality. Indeed, the solutions F_ϵ built by DiPerna and Lions satisfy in addition

$$\partial_t \int_{\mathbb{R}^D} vF_\epsilon \; dv + \frac{1}{\epsilon}\mathrm{div} \int_{\mathbb{R}^D} (v \otimes v)F_\epsilon \; dv + \frac{1}{\epsilon}\mathrm{div}(M_\epsilon) = 0. \tag{44}$$

Besides, the following energy equality holds

$$\frac{1}{2}\int_\Omega \int_{\mathbb{R}^D} |v|^2 F_\epsilon(t,x,v) dx \; dv + \frac{1}{2}\int_\Omega \mathrm{tr}(M_\epsilon) \; dx = \frac{1}{2}\int_\Omega \int_{\mathbb{R}^D} |v|^2 F^0_\epsilon(x,v) dx \; dv \tag{45}$$

which can be rewritten (with $\epsilon^m m_\epsilon = M_\epsilon$)

$$\partial_t \langle g_\epsilon v\rangle + \nabla_x . \langle g_\epsilon v \otimes v\rangle + \frac{1}{\epsilon}\nabla . m_\epsilon = 0, \tag{46}$$

$$\int_\Omega \langle |v|^2 g_\epsilon\rangle dx + \int_\Omega \mathrm{tr}(m_\epsilon) \; dx = 0 \tag{47}$$

3.2 Entropy inequality

One can write the entropy inequality for G_ϵ (as in the case of the limit towards the Navier-Stokes system) or write it for F_ϵ as well. It turns out that the second choice gives a better estimate for the defect measure. Indeed starting from the entropy inequality for F_ϵ, we can deduce

$$\int_\Omega \int_{\mathbb{R}^D} h(\epsilon^m g_\epsilon) dx\ M\ dv(t) - \int_\Omega \int_{\mathbb{R}^D} \epsilon^m \frac{|v|^2}{2} g_\epsilon dx\ M\ dv(t) +$$
$$+\frac{1}{4\epsilon^2}\int_0^t ds \int_\Omega dx \int_{\mathbb{R}^D}\int_{\mathbb{R}^D} M\ dv\ M_1\ dv_1 \int_{S^{D-1}} d\omega b(v-v_1,\omega) \tag{48}$$
$$\Big(G'_{\epsilon 1}G'_\epsilon - G_{\epsilon 1}G_\epsilon\Big)\log\Big(\frac{G'_{\epsilon 1}G'_\epsilon}{G_{\epsilon 1}G_\epsilon}\Big) \le \int_\Omega \int_{\mathbb{R}^D} h(\epsilon^m g^0_\epsilon) dx\ M\ dv$$

3.3 The result

We take initial data satisfying

$$\int_{\mathbb{T}^D}\int_{\mathbb{R}^D} F^0_\epsilon dxdv = 1, \quad \int_{\mathbb{T}^D}\int_{\mathbb{R}^D} vF^0_\epsilon dxdv = 0, \quad \int_{\mathbb{T}^D}\int_{\mathbb{R}^D} |v|^2 F^0_\epsilon dxdv = D \tag{49}$$

and

$$\int_\Omega \int_{\mathbb{R}^D} F^0_\epsilon log F^0_\epsilon\ dx\ dv \le -\frac{D}{2} + C\epsilon^{2m} \tag{50}$$

We also assume that b satisfies $(A0)$.

Theorem 6 *If F_ϵ is a sequence of renormalized solutions of the Boltzmann equations (B_ϵ) with initial condition F^0_ϵ and satisfies the entropy inequality as well as the refined momentum equation, then the family $(1+|v|^2)g_\epsilon$ is relatively compact in $w - L^1(dt\ Mdv\ dx)$. And, if g is a weak limit of a subsequence (still denoted g_ϵ) then $Lg = 0$ and $g = \rho + u.v + \theta(\frac{|v|^2}{2} - \frac{N}{2})$ satisfies the limiting dissipation inequality*

$$\frac{1}{2}\int_\Omega |\rho(t)|^2 + |u(t)|^2 + \frac{D}{2}|\theta|^2\ dx + \int_0^t \int_\Omega \frac{1}{2}\nu|\nabla_x u +^t \nabla_x u|^2$$
$$\le \liminf_{\epsilon\to 0} \frac{1}{\epsilon^{2m}} \int_\Omega \langle h(\epsilon^m g_\epsilon)\rangle dx = C^0 \tag{51}$$

Moreover $u = \langle vg\rangle$ is the solution of the Stokes system (S) with the initial condition $u^0 = P\langle vg^0\rangle$ and where the viscosity ν is given by (27). Besides, we have the following strong Boussinesq relationship

$$\rho + \theta = 0. \tag{52}$$

3.4 Conservation of momentum at the limit

We explain here briefly how we can recover the conservation of momentum in the limit. Indeed, starting from the entropy inequality, one deduces that

$$\int_\Omega \langle h(\epsilon^m g_\epsilon)\rangle dx + \epsilon^m \mathrm{tr}(m_\epsilon) + D(G_\epsilon) \le C\epsilon^{2m} \tag{53}$$

and since $m > 1$, we deduce

$$\frac{1}{\epsilon}\mathrm{tr}(m_\epsilon) \quad \text{and} \quad \frac{1}{\epsilon} m_\epsilon \quad \rightarrow \quad 0 \tag{54}$$

in $L^\infty(0,T;L^1(\Omega))$ since m_ϵ is a positive matrix and we denote $\delta_\epsilon = \epsilon^m$.

4 The case of a bounded domain

The second type of results we want to present concerns the the derivation of fluid mechanics boundary conditions starting form kinetic boundary condition. For simplicity, we will present the result in the Stokes scaling though the proof works as well for the Navier-Stokes scaling using the result of the previous sections. Let Ω be a smooth bounded domain of $\mathbb{R}^D$ and $\mathcal{O} = \Omega \times \mathbb{R}^D$ the space-velocity domain. Let $n(x)$ be the outward unit normal vector at $x \in \partial\Omega$. We denote by $d\sigma_x$ the Lebesgue measure on the boundary $\partial\Omega$ and we define the outgoing/incoming sets Σ_+ and Σ_- by

$$\Sigma_\pm = \{(x,v) \in \Sigma, \quad \pm n(x).v > 0\} \text{ where } \Sigma = \partial\Omega \times \mathbb{R}^D.$$

We consider the Boltzmann equation in $\mathbb{R}_+ \times \mathcal{O}$ with a scaling where $q = s = 1$ and $m > 1$.

4.1 The Maxwell boundary condition

The boundary conditions we will consider express the balance between the incoming and outgoing part of the trace of F, namely $\gamma_\pm F = 1\!\!1_{\Sigma_\pm}\gamma F$. We will use the following Maxwell reflection condition

$$\gamma_- F = (1-\alpha)L(\gamma_+ F) + \alpha K(\gamma_+ F) \quad \text{on } \Sigma_- \tag{55}$$

where α is a constant also called accommodation coefficient. The local reflection operator L is given by

$$L\phi(x,v) = \phi(x, R_x v) \tag{56}$$

where $R_x v = v - 2(n(x).v)n(x)$ is the velocity before the collision with the wall. The diffuse reflection operator K is given by

$$K\phi(x,v) = \sqrt{2\pi}\tilde{\phi}(x)M(v) \tag{57}$$

where $\tilde{\phi}$ is the outgoing mass flux

$$\tilde{\phi}(x) = \int_{v.n(x)>0} \phi(x,v)\, n(x).v dv. \tag{58}$$

We notice that

$$\int_{v.n(x)>0} n(x).v\sqrt{2\pi}M(v)\ dv = \int_{v.n(x)<0} |n(x).v|\sqrt{2\pi}M(v)\ dv = 1,$$

which expresses the conservation of mass at the boundary. Here, we are taking the temperature of the wall to be constant and equal to 1. For the existence of renormalized solutions to the Boltzmann equation in a bounded domain we refer to [30].

4.2 A priori estimate

Let $\mathcal{E}(\gamma_+ G_\epsilon)$, the so-called Darrozès-Guiraud information [16], be given by

$$\mathcal{E}(\gamma_+ G_\epsilon) = \int_{\partial\Omega} \Big(\big\langle h(\delta_\epsilon \gamma_+ g_\epsilon) \big\rangle_{\partial\Omega} - h\big(\langle \delta_\epsilon \gamma_+ g_\epsilon \rangle_{\partial\Omega} \big) \Big) \, d\sigma_x . \tag{59}$$

In the case of a bounded domain, the entropy inequality reads

$$H(G_\epsilon(t)) + \int_0^t \left(\frac{1}{\epsilon^2} E(G_\epsilon(s)) + \frac{\alpha_\epsilon}{\sqrt{2\pi}\epsilon} \mathcal{E}_\epsilon(\gamma_+ G_\epsilon(s)) \right) ds \leq H(G_\epsilon^{in}) , \tag{60}$$

where $H(G)$ is the relative entropy functional

$$H(G) = \int_\Omega \langle (G \log(G) - G + 1) \rangle \, dx , \tag{61}$$

and $E(G)$ is the entropy dissipation rate functional

$$E(G) = \int_\Omega \left\langle\!\!\left\langle \frac{1}{4} \log\left(\frac{G_1' G'}{G_1 G} \right) (G_1' G' - G_1 G) \right\rangle\!\!\right\rangle dx . \tag{62}$$

Notice the presence of the extra positive term due to the boundary. It is easy to see that due to Jensen inequality the extra term $\mathcal{E}_\epsilon(\gamma_+ G_\epsilon(s)) \geq 0$. This also gives a bound on $\gamma_+ G_\epsilon$ which is useful.

Now, we present two results taken from [29] which hold for a wide range of collision kernels

Theorem 7 (Navier boundary condition) *Let $F_\epsilon^{in} = G_\epsilon^{in} M$ be a family of initial data satisfying*

$$\frac{1}{\delta_\epsilon^2} H(G_\epsilon^{in}) + \iint_{\mathcal{O}} |v|^2 F_\epsilon^{in} \, dx dv \leq C^{in} \tag{63}$$

for some $C^{in} < \infty$ and

$$\begin{aligned} &\frac{1}{\delta_\epsilon} \Pi \langle v \, G_\epsilon^{in} \rangle \to u \quad \text{in } \mathcal{D}'(\Omega; \mathbf{R}^D) , \\ &\frac{1}{\delta_\epsilon} \langle (\frac{1}{D+2} |v|^2 - 1) \, G_\epsilon^{in} \rangle \to \theta \quad \text{in } \mathcal{D}'(\Omega; \mathbf{R}^D) , \end{aligned} \tag{64}$$

for some $(u^{in}, \theta^{in}) \in L^2(dx; \mathbf{R}^D \times \mathbf{R})$. Denote by G_ϵ any corresponding family of renormalized solutions of the Boltzmann equation satisfying the entropy inequality (60), where the accommodation coefficient satisfies

$$\frac{\alpha_\epsilon}{\sqrt{2\pi}\epsilon} \to \lambda \quad \text{when } \epsilon \to 0. \tag{65}$$

Then, as $\epsilon \to 0$, the family of fluctuations satisfies

$$\begin{aligned} & g_\epsilon \to vu + \big(\frac{1}{2}|v|^2 - \frac{D+2}{2}\big)\theta \quad \textit{in } w\text{-}L^1_{loc}(dt; w\text{-}L^1((1+|v|^2)M dv\, dx))\,, \\ & \Pi\langle v\, g_\epsilon\rangle \to u \quad \textit{in } C([0,\infty); \mathcal{D}'(\Omega; \mathbf{R}^D))\,, \\ & \langle (\frac{1}{D+2}|v|^2 - 1)\, g_\epsilon\rangle \to \theta \quad \textit{in } C([0,\infty); \mathcal{D}'(\Omega; \mathbf{R}^D))\,, \end{aligned} \tag{66}$$

where Π is the orthogonal projection from $L^2(dx; \mathbf{R}^D)$ onto divergence-free vector fields with zero normal velocity, namely the set

$$H = \{u \in L^2(\Omega),\ \nabla_x{\cdot}u = 0,\ u.n = 0 \ \textit{on } \partial\Omega\}.$$

Furthermore, $(u,\theta) \in C([0,\infty); H \times L^2(\Omega)) \cap L^2(dt; H^1(\Omega) \times H^1(\Omega))$ and it satisfies the Stokes-Fourier system with Navier boundary condition

$$\begin{cases} \partial_t u + \nabla_x p - \nu\Delta_x u = 0\,, \quad div(u) = 0 & \textit{on } \mathbb{R}^+ \times \Omega, \\ (2\nu d(u)\cdot n + \lambda u) \wedge n = 0\,, \quad u n = 0 & \textit{on } \mathbb{R}^+ \times \partial\Omega, \end{cases}$$

$$\begin{cases} \partial_t \theta - \kappa\Delta_x \theta = 0 & \textit{on } \mathbb{R}^+ \times \Omega, \\ \kappa\partial_n\theta + \lambda\frac{D+1}{D+2}\,\theta = 0 & \textit{on } \mathbb{R}^+ \times \partial\Omega, \end{cases} \tag{67}$$

$$u(0,x) = u^{in}(x), \quad \theta(0,x) = \theta^{in}(x) \quad \textit{on } \Omega.$$

where $d(u)$ denotes the symmetric part of the stress tensor $d(u) = \frac{1}{2}(\nabla u +^t \nabla u)$.

The second result treats the case of Dirichlet boundary conditions. We will make the same assumptions as in the previous theorem but instead of assuming that $\frac{\alpha_\epsilon}{\epsilon\sqrt{2\pi}} \to \lambda$, we assume that $\frac{\alpha_\epsilon}{\epsilon} \to +\infty$.

Theorem 8 (Dirichlet boundary condition) *We make the same assumptions as in Theorem 7, except that we replace condition (65) by*

$$\frac{\alpha_\epsilon}{\epsilon} \to \infty, \quad \textit{when } \epsilon \to 0. \tag{68}$$

Then, as $\epsilon \to 0$, we have the same convergences (66) as in Theorem 7 with $(u,\theta) \in C([0,\infty); H \times L^2(\Omega)) \cap L^2(dt; V \times H^1_0(\Omega))$ where

$$V = \{u \in H^1(\Omega),\ \nabla_x{\cdot}u = 0,\ u = 0 \ \textit{on } \partial\Omega\}.$$

Furthermore, (u,θ) satisfies the Stokes-Fourier system with Dirichlet boundary condition

$$\begin{aligned} \partial_t u + \nabla_x p - \nu\Delta_x u = 0\,, \quad div(u) = 0 \qquad & \textit{on } \mathbb{R}^+ \times \Omega, \\ \partial_t\theta - \kappa\Delta_x\theta = 0 \qquad & \textit{on } \mathbb{R}^+ \times \Omega, \\ u = 0, \quad \theta = 0 \qquad & \textit{on } \mathbb{R}^+ \times \partial\Omega, \\ u(0,x) = u^{in}(x), \quad \theta(0,x) = \theta^{in}(x) \qquad & \textit{on } \Omega. \end{aligned} \tag{69}$$

4.3 Idea of the proof

The interior convergence can be deduced easily from the work of Golse and Levermore [11]. We just want to explain the convergence at the boundary. We prove two types of control on the trace γg_ϵ of g_ϵ on the boundary. The first control comes from the inside, it uses the interior estimates to deduce an estimate on the trace

Lemma 9 *We have for all $p > 0$,*

$$\gamma \hat{g}_\epsilon \to \gamma g \ \textit{in} \ w\text{-}L^1_{loc}(dt; w\text{-}L^1(M(1+|v|^p)|vn(x)|dv\, d\sigma_x)) \tag{70}$$

$$\epsilon^m \gamma g_\epsilon \to 0 \ \textit{a.e.} \ \textit{on} \ \mathbb{R}^+ \times \partial\Omega \times \mathbb{R}^d. \tag{71}$$

The second control comes from the boundary term appearing in the entropy dissipation. It does not give an estimate on g_ϵ but rather on g_ϵ minus its average in v. We get

Lemma 10 *Define $\gamma_\epsilon = \gamma_+ g_\epsilon - 1\!\!1_{\Sigma_+}\langle\gamma_+ g_\epsilon\rangle_{\partial\Omega}$ and*

$$\gamma_\epsilon^{(1)} = \gamma_\epsilon 1\!\!1_{\gamma_+ G_\epsilon \le 2\langle\gamma_+ G_\epsilon\rangle_{\partial\Omega} \le 4\gamma_+ G_\epsilon}, \quad \gamma_\epsilon^{(2)} = \gamma_\epsilon - \gamma_\epsilon^{(1)}. \tag{72}$$

Then

$$\sqrt{\frac{\alpha_\epsilon}{\epsilon}} \, \frac{\gamma_\epsilon^{(1)}}{(1+\frac{\delta_\epsilon}{3}\gamma_+ g_\epsilon)^{1/2}} \ \textit{is bounded in} \ L^2_{loc}(dt; L^2(M|vn(x)|dv\, d\sigma_x)); \tag{73}$$

$$\sqrt{\frac{\alpha_\epsilon}{\epsilon}} \, \frac{\gamma_\epsilon^{(1)}}{(1+\frac{\delta_\epsilon}{3}\langle\gamma_+ g_\epsilon\rangle_{\partial\Omega})^{1/2}} \ \textit{is bounded in} \ L^2_{loc}(dt; L^2(M|vn(x)|dv\, d\sigma_x)); \tag{74}$$

$$\frac{\alpha_\epsilon}{\epsilon\delta_\epsilon}\gamma_\epsilon^{(2)} \ \textit{is bounded in} \ L^1_{loc}(dt; L^1(M|vn(x)|dv\, d\sigma_x)). \tag{75}$$

5 Convergence towards the Euler system

We present, here, a method of proof based on an energy method which uses the relative entropy method (see Yau [38]). Indeed contrary to the two preceding cases, we suppose here the existence of a strong solution to the Euler system and we show the convergence towards this solution. The technique used is based on a Gronwall lemma. In [26] (in collaboration with P.-L. Lions), we show this convergence with an assumption on high velocities (A2). This assumption was removed in Saint-Raymond [31]. We introduce a defect measure (as in the Stokes case) which disappears at the limit. We take well prepared initial data (i.e. there are no acoustic waves) and the temperature fluctuation is equal to 0 initially.

5.1 Entropic convergence

In addition to the assumptions on G_ϵ^0 which we imposed in the case of convergence towards the Navier-Stokes system, we suppose that g_ϵ^0 converges entropically towards g^0 and that $g^0 = u^0.v$ (with div$u^0 = 0$) i.e. that

$$g_\epsilon^0 \to g^0 \quad \text{in} \quad w - L^1(M\ dvdx), \quad \text{and} \tag{76}$$

$$\lim_\epsilon \frac{1}{\epsilon^2} \int_\Omega \langle h(\epsilon g_\epsilon^0) \rangle dx = \frac{1}{2} \int_\Omega \langle (g^0)^2 \rangle dx. \tag{77}$$

It is also supposed that u^0 is regular enough (for example $u^0 \in H^s, s > \frac{D}{2} + 1$) to be able to build a strong solution $\tilde{u}$ of the Euler system with the initial data u^0. Then, we have $\tilde{u} \in L^\infty_{loc}([0, T^*); H^s)$ for some $T^* > 0$.

5.2 Relative entropy

We want to show that the distribution F_ϵ is close to a Maxwellian $M_{(0,\epsilon\tilde{u},0)} = M\tilde{G}_\epsilon$. But as F_ϵ is only in $L$log$L$, we have to estimate the difference between F_ϵ and $M_{(0,\epsilon\tilde{u},0)}$ using the relative entropy

$$H(G_\epsilon, \tilde{G}_\epsilon) = \int_\Omega \Big\langle G_\epsilon\ log\Big(\frac{G_\epsilon}{\tilde{G}_\epsilon}\Big) - G_\epsilon + \tilde{G}_\epsilon \Big\rangle. \tag{78}$$

Using the improved entropy inequality (48), we get

$$\begin{aligned} H(G_\epsilon, \tilde{G}_\epsilon) + \epsilon \int_\Omega \mathrm{tr}(m_\epsilon) + \int_0^t ds D(G_\epsilon) + &\leq H(G_\epsilon^0, \tilde{G}_\epsilon^0) \\ + \quad \int_0^t \int_\Omega < G_\epsilon \partial_t \mathrm{log} \tilde{G}_\epsilon > + \epsilon^2 \partial_t < g_\epsilon v > .\tilde{u} &+ \epsilon^3 \partial_t < g_\epsilon > \frac{|\tilde{u}|^2}{2} ds \end{aligned}$$

where m_ϵ denotes the sequence of defect measures appearing in the conservation of momentum

5.3 The result

We present here the result of Saint-Raymond [31]

Theorem 11 *Under some assumption of the collision kernel, if G_ϵ is a sequence of renormalized solutions of the Boltzmann equations with initial condition G_ϵ^0, and such that g_ϵ^0 converges entropically to $g^0 = u^0.v$, where $u^0 \in H^s$, ($s > \frac{D}{2} + 1$). Then, for all $0 \leq t < T^*$*

$$g_\epsilon(t) \to \tilde{u}(t).v \quad \textit{entropically} \tag{79}$$

where $\tilde{u}(t)$ is the unique solution of the Euler system in $L^\infty_{loc}([0, T*); H^s)$ with the initial condition u^0. Moreover, the convergence is locally uniform in time.

Let us explain here the idea of the proof of the above result. It is based on a Gronwall lemma. Indeed, after some non trivial computations, one can rewrite the entropy inequality as follows

$$\begin{aligned} &\frac{1}{\epsilon^2}\Big[H(G_\epsilon,\tilde{G}_\epsilon)+\epsilon\int_\Omega \mathrm{tr}(m_\epsilon)\Big](t)+\frac{1}{\epsilon^2}\int_0^t dsD(G_\epsilon)\leq\frac{1}{\epsilon^2}H(G_\epsilon^0,\tilde{G}_\epsilon^0)\\ +\quad&\int_0^t ||\nabla\tilde{u}||_{L^\infty}\frac{1}{\epsilon^2}\Big[H(G_\epsilon,\tilde{G}_\epsilon)+\epsilon\int_\Omega \mathrm{tr}(m_\epsilon)\Big](s)\ ds+A_\epsilon\end{aligned}$$

where A_ϵ converges to 0. Hence, we deduce that $H(G_\epsilon,\tilde{G}_\epsilon)$ goes to 0 in $L^\infty_{loc}([0,T^*))$.

We want to point out that the same type of argument can be used to prove the convergence towards the Navier-Stokes system in the case a regular solution is known to exist.

References

[1] Claude Bardos, François Golse, and C. David Levermore. Fluid dynamic limits of kinetic equations. II. Convergence proofs for the Boltzmann equation. *Comm. Pure Appl. Math.*, 46(5):667–753, 1993.

[2] Claude Bardos, François Golse, and C. David Levermore. Acoustic and Stokes limits for the Boltzmann equation. *C. R. Acad. Sci. Paris Sér. I Math.*, 327(3):323–328, 1998.

[3] Claude Bardos, François Golse, and C. David Levermore. The acoustic limit for the Boltzmann equation. *Arch. Ration. Mech. Anal.*, 153(3):177–204, 2000.

[4] Claude Bardos, François Golse, and David Levermore. Fluid dynamic limits of kinetic equations. I. Formal derivations. *J. Statist. Phys.*, 63(1-2):323–344, 1991.

[5] Claude Bardos and Seiji Ukai. The classical incompressible Navier-Stokes limit of the Boltzmann equation. *Math. Models Methods Appl. Sci.*, 1(2):235–257, 1991.

[6] Russel E. Caflisch. The fluid dynamic limit of the nonlinear Boltzmann equation. *Comm. Pure Appl. Math.*, 33(5):651–666, 1980.

[7] Carlo Cercignani. *The Boltzmann equation and its applications.* Springer-Verlag, New York, 1988.

[8] Carlo Cercignani, Reinhard Illner, and Mario Pulvirenti. *The mathematical theory of dilute gases.* Springer-Verlag, New York, 1994.

[9] A. De Masi, R. Esposito, and J. L. Lebowitz. Incompressible Navier-Stokes and Euler limits of the Boltzmann equation. *Comm. Pure Appl. Math.*, 42(8):1189–1214, 1989.

[10] R. J. DiPerna and P.-L. Lions. On the Cauchy problem for Boltzmann equations: global existence and weak stability. *Ann. of Math. (2)*, 130(2):321–366, 1989.

[11] François Golse and C. David Levermore. Stokes-Fourier and acoustic limits for the Boltzmann equation: convergence proofs. *Comm. Pure Appl. Math.*, 55(3):336–393, 2002.

[12] François Golse, Pierre-Louis Lions, Benoît Perthame, and Rémi Sentis. Regularity of the moments of the solution of a transport equation. *J. Funct. Anal.*, 76(1):110–125, 1988.

[13] François Golse, Benoît Perthame, and Rémi Sentis. Un résultat de compacité pour les équations de transport et application au calcul de la limite de la valeur propre principale d'un opérateur de transport. *C. R. Acad. Sci. Paris Sér. I Math.*, 301(7):341–344, 1985.

[14] François Golse and Laure Saint-Raymond. Navier-Stokes-fourier limit for the Boltzmann equation: Convergence proofs. *preprint*, 2002.

[15] François Golse and Laure Saint-Raymond. Velocity averaging in L^1 for the transport equation. *C. R. Math. Acad. Sci. Paris*, 334(7):557–562, 2002.

[16] Jean-Pierre Guiraud and J.-S Darrozès. Généralisation formelle du théorème H en présence de parois. *C.R. Acad. Sci. Paris*, 262:368–371, 1966.

[17] David Hilbert. Begründung der kinetischen gastheorie. *Math. Annalen*, 72:562–577, 1912.

[18] M. Lachowicz. On the initial layer and the existence theorem for the nonlinear Boltzmann equation. *Math. Methods Appl. Sci.*, 9(3):342–366, 1987.

[19] Oscar E. Lanford, III. Time evolution of large classical systems. In *Dynamical systems, theory and applications (Recontres, Battelle Res. Inst., Seattle, Wash., 1974)*, pages 1–111. Lecture Notes in Phys., Vol. 38. Springer, Berlin, 1975.

[20] David Levermore and Nader Masmoudi. From the Boltzmann equation to an incompressible Navier-Stokes-Fourier system. *preprint*, 2002.

[21] Pierre-Louis Lions. *Mathematical topics in fluid mechanics. Vol. 1.* The Clarendon Press Oxford University Press, New York, 1996. Incompressible models, Oxford Science Publications.

[22] Pierre-Louis Lions. *Mathematical topics in fluid mechanics. Vol. 2.* The Clarendon Press Oxford University Press, New York, 1998. Compressible models, Oxford Science Publications.

[23] Pierre-Louis Lions and Nader Masmoudi. Incompressible limit for a viscous compressible fluid. *J. Math. Pures Appl. (9)*, 77(6):585–627, 1998.

[24] Pierre-Louis Lions and Nader Masmoudi. Une approche locale de la limite incompressible. *C. R. Acad. Sci. Paris Sér. I Math.*, 329(5):387–392, 1999.

[25] Pierre-Louis Lions and Nader Masmoudi. From the Boltzmann equations to the equations of incompressible fluid mechanics. I. *Arch. Ration. Mech. Anal.*, 158(3):173–193, 2001.

[26] Pierre-Louis Lions and Nader Masmoudi. From the Boltzmann equations to the equations of incompressible fluid mechanics. II. *Arch. Ration. Mech. Anal.*, 158(3):195–211, 2001.

[27] Carlo Marchioro and Mario Pulvirenti. *Mathematical theory of incompressible nonviscous fluids.* Springer-Verlag, New York, 1994.

[28] Nader Masmoudi. Some recent developments on the hydrodynamic limit of the Boltzmann equation. In *Mathematics & mathematics education (Bethlehem, 2000)*, pages 167–185. World Sci. Publishing, River Edge, NJ, 2002.

[29] Nader Masmoudi and Laure Saint-Raymond. From the Boltzmann equation to the Stokes-Fourier system in a bounded domain. *Comm. Pure Appl. Math.*, 56(9):1263–1293, 2003.

[30] Stéphane Mischler. On the initial boundary value problem for the Vlasov-Poisson-Boltzmann system. *Comm. Math. Phys.*, 210(2):447–466, 2000.

[31] Laure Saint-Raymond. Convergence of solutions to the Boltzmann equation in the incompressible Euler limit. *Arch. Ration. Mech. Anal.*, 166(1):47–80, 2003.

[32] Yoshio Sone. *Kinetic theory and fluid dynamics.* Modeling and Simulation in Science, Engineering, & Technology. Birkhäuser Boston Inc., Boston, MA, 2002.

[33] Herbert Spohn. Boltzmann hierarchy and Boltzmann equation. In *Kinetic theories and the Boltzmann equation (Montecatini, 1981)*, pages 207–220. Springer, Berlin, 1984.

[34] Roger Temam. *Navier-Stokes equations and nonlinear functional analysis.* Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, second edition, 1995.

[35] Seiji Ukai and Kiyoshi Asano. The Euler limit and initial layer of the nonlinear Boltzmann equation. *Hokkaido Math. J.*, 12(3, part 1):311–332, 1983.

[36] Cédric Villani. Limites hydrodynamiques de l'équation de Boltzmann (d'après C. Bardos, F. Golse, C. D. Levermore, P.-L. Lions, N. Masmoudi, L. Saint-Raymond). *Astérisque*, (282):Exp. No. 893, ix, 365–405, 2002. Séminaire Bourbaki, Vol. 2000/2001.

[37] Cédric Villani. *A review of mathematical problems in collisonal kinetic theory, in Handbook of mathematical fluid dynamics. Vol. II.* Friedlander, S. and Serre, D. North-Holland, Amsterdam, 2003.

[38] Horng-Tzer Yau. Relative entropy and hydrodynamics of Ginzburg-Landau models. *Lett. Math. Phys.*, 22(1):63–80, 1991.

Bol. Soc. Esp. Mat. Apl. n^o*26(2003), 79–140*

Control theory: History, mathematical achievements and perspectives*

E. Fernández-Cara[1] and E. Zuazua[2]

[1] Departamento de Ecuaciones Diferenciales y Análisis Numérico,
Universidad de Sevilla
[2] Departamento de Matemáticas,
Universidad Autónoma de Madrid

cara@us.es, enrique.zuazua@uam.es

Abstract

These notes are devoted to present some of the mathematical milestones of Control Theory. To do that, we first overview its origins and some of the main mathematical achievements. Then, we discuss the main domains of Sciences and Technologies where Control Theory arises and applies. This forces us to address modelling issues and to distinguish between the two main control theoretical approaches, controllability and optimal control, discussing the advantages and drawbacks of each of them. In order to give adequate formulations of the main questions, we have introduced some of the most elementary mathematical material, avoiding unnecessary technical difficulties and trying to make the paper accessible to a large class of readers. The subjects we address range from the basic concepts related to the dynamical systems approach to (linear and nonlinear) Mathematical Programming and Calculus of Variations. We also present a simplified version of the outstanding results by Kalman on the controllability of linear finite dimensional dynamical systems, Pontryaguin's maximum principle and the principle of dynamical programming. Some aspects related to the complexity of modern control systems, the discrete versus continuous modelling, the numerical approximation of control problems and its control theoretical consequences are also discussed. Finally, we describe some of the major challenging applications in Control Theory for the XXI Century. They will probably influence strongly the development of this discipline in the near future.

*The authors have been partially financed by D.G.E.S.-España, respectively Grants BFM2000-1317 and BFM2002-03345.

Fecha de recepción: 30 de abril de 2003

Key words: *Control Theory, optimal control, controllability, differential equations, feedback, Optimization, Calculus of Variations.*

AMS subject classifications: *49J15, 49J20, 70Q05, 93B, 93C*

1 Introduction

This article is devoted to present some of the mathematical milestones of Control Theory. We will focus on systems described in terms of ordinary differential equations. The control of (deterministic and stochastic) partial differential equations remains out of our scope. However, it must be underlined that most ideas, methods and results presented here do extend to this more general setting, which leads to very important technical developments.

The underlying idea that motivated this article is that Control Theory is certainly, at present, one of the most interdisciplinary areas of research. Control Theory arises in most modern applications. The same could be said about the very first technological discoveries of the industrial revolution. On the other hand, Control Theory has been a discipline where many mathematical ideas and methods have melt to produce a new body of important Mathematics. Accordingly, it is nowadays a rich crossing point of Engineering and Mathematics.

Along this paper, we have tried to avoid unnecessary technical difficulties, to make the text accessible to a large class of readers. However, in order to introduce some of the main achievements in Control Theory, a minimal body of basic mathematical concepts and results is needed. We develop this material to make the text self-contained.

These notes contain information not only on the main mathematical results in Control Theory, but also about its origins, history and the way applications and interactions of Control Theory with other Sciences and Technologies have conducted the development of the discipline.

The plan of the paper is the following. Section 2 is concerned with the origins and most basic concepts. In Section 3 we study a simple but very interesting example: the *pendulum*. As we shall see, an elementary analysis of this simple but important mechanical system indicates that the fundamental ideas of Control Theory are extremely meaningful from a physical viewpoint.

In Section 4 we describe some relevant historical facts and also some important contemporary applications. There, it will be shown that Control Theory is in fact an interdisciplinary subject that has been strongly involved in the development of the contemporary society.

In Section 5 we describe the two main approaches that allow to give rigorous formulations of control problems: controllability and optimal control. We also discuss their mutual relations, advantages and drawbacks.

In Sections 6 and 7 we present some basic results on the controllability of linear and nonlinear finite dimensional systems. In particular, we revisit the Kalman approach to the controllability of linear systems, and we recall the use of Lie brackets in the control of nonlinear systems, discussing a simple example of a planar moving square car.

In Section 8 we discuss how the complexity of the systems arising in modern technologies affects Control Theory and the impact of numerical approximations and discrete modelling, when compared to the classical modelling in the context of Continuum Mechanics.

In Section 9 we describe briefly two beautiful and extremely important challenging applications for Control Theory in which, from a mathematical viewpoint, almost all remains to be done: laser molecular control and the control of floods.

In Section 10 we present a list of possible future applications and lines of development of Control Theory: large space structures, Robotics, biomedical research, etc.

Finally, we have included two Appendices, where we recall briefly two of the main principles of modern Control Theory, namely *Pontryagin's maximum principle* and *Bellman's dynamical programming principle.*

2 Origins and basic ideas, concepts and ingredients

The word *control* has a double meaning. First, controlling a system can be understood simply as testing or checking that its behavior is satisfactory. In a deeper sense, to control is also to act, to put things in order to guarantee that the system behaves as desired.

S. Bennet starts the first volume of his book [2] on the history of *Control Engineering* quoting the following sentence of Chapter 3, Book 1, of the monograph "Politics" by Aristotle:

> "...if every instrument could accomplish its own work, obeying or anticipating the will of others ...if the shuttle weaved and the pick touched the lyre without a hand to guide them, chief workmen would not need servants, nor masters slaves."

This sentence by Aristotle describes in a rather transparent way the guiding goal of *Control Theory:* the need of automatizing processes to let the human being gain in liberty, freedom, and quality of life.

Let us indicate briefly how control problems are stated nowadays in mathematical terms. To fix ideas, assume we want to get a good behavior of a physical system governed by the *state equation*

$$A(y) = f(v). \tag{1}$$

Here, y is the *state*, the unknown of the system that we are willing to control. It belongs to a vector space Y. On the other hand, v is the *control.* It belongs to the *set of admissible controls* $\mathcal{U}_{\rm ad}$. This is the variable that we can choose freely in $\mathcal{U}_{\rm ad}$ to act on the system.

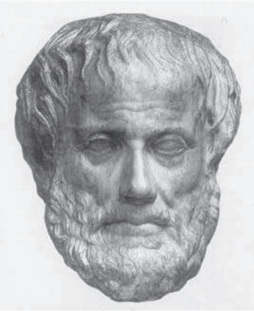


Figure 1: Aristotle (384–322 B.C.).

Let us assume that $A : D(A) \subset Y \mapsto Y$ and $f : \mathcal{U}_{\rm ad} \mapsto Y$ are two given (linear or nonlinear) mappings. The operator A determines the equation that must be satisfied by the state variable y, according to the laws of Physics. The function f indicates the way the control v acts on the system governing the state. For simplicity, let us assume that, for each $v \in \mathcal{U}_{\rm ad}$, the state equation (1) possesses exactly one solution $y = y(v)$ in Y. Then, roughly speaking, to control (1) is to find $v \in \mathcal{U}_{\rm ad}$ such that the solution to (1) gets close to the desired prescribed state. The "best" among all the existing controls achieving the desired goal is frequently referred to as the *optimal control.*

This mathematical formulation might seem sophisticated or even obscure for readers not familiar with this topic. However, it is by now standard and it has been originated naturally along the history of this rich discipline. One of the main advantages of such a general setting is that many problems of very different nature may fit in it, as we shall see along this work.

As many other fields of human activities, the discipline of Control existed much earlier than it was given that name. Indeed, in the world of living species, organisms are endowed with sophisticated mechanisms that regulate the various tasks they develop. This is done to guarantee that the essential variables are kept in optimal regimes to keep the species alive allowing them to grow, develop and reproduce.

Thus, although the mathematical formulation of control problems is intrinsically complex, the key ideas in Control Theory can be found in Nature, in the evolution and behavior of living beings.

The first key idea is the *feedback* concept. This term was incorporated to Control Engineering in the twenties by the engineers of the "Bell Telephone Laboratory" but, at that time, it was already recognized and consolidated in other areas, such as Political Economics.

Essentially, a feedback process is the one in which the state of the system determines the way the control has to be exerted at any time. This is related to the notion of *real time control*, very important for applications. In the framework of (1), we say that the control u is given by a *feedback law* if we are able to provide a mapping $G : Y \mapsto \mathcal{U}_{\rm ad}$ such that

$$u = G(y), \quad \text{where } y = y(u), \tag{2}$$

i.e. y solves (1) with v replaced by u.

Nowadays, feedback processes are ubiquitous not only in Economics, but also in Biology, Psychology, etc. Accordingly, in many different related areas, the *cause-effect principle* is not understood as a static phenomenon any more, but it is rather being viewed from a dynamical perspective. Thus, we can speak of the *cause-effect-cause principle*. See [33] for a discussion on this and other related aspects.

The second key idea is clearly illustrated by the following sentence by H.R. Hall in [17] in 1907 and that we have taken from [2]:

> "It is a curious fact that, while political economists recognize that for the proper action of the law of supply and demand there must be fluctuations, it has not generally been recognized by mechanicians in this matter of the steam engine governor. The aim of the mechanical economist, as is that of the political economist, should be not to do away with these fluctuations all together (for then he does away with the principles of self-regulation), but to diminish them as much as possible, still leaving them large enough to have sufficient regulating power."

The need of having room for fluctuations that this paragraph evokes is related to a basic principle that we apply many times in our daily life. For instance, when driving a car at a high speed and needing to brake, we usually try to make it intermittently, in order to keep the vehicle under control at any moment. In the context of human relationships, it is also clear that insisting permanently in the same idea might not be precisely the most convincing strategy.

The same rule applies for the control of a system. Thus, to control a system arising in Nature or Technology, we do not have necessarily to stress the system and drive it to the desired state immediately and directly. Very often, it is much more efficient to control the system letting it fluctuate, trying to find a harmonic dynamics that will drive the system to the desired state without forcing it too much. An excess of control may indeed produce not only an inadmissible cost but also irreversible damages in the system under consideration.

Another important underlying notion in Control Theory is *Optimization*. This can be regarded as a branch of Mathematics whose goal is to improve a variable in order to maximize a benefit (or minimize a cost). This is applicable to a lot of practical situations (the variable can be a temperature, a velocity field, a measure of information, etc.). Optimization Theory and its related techniques are such a broad subject that it would be impossible to make a unified presentation. Furthermore, a lot of recent developments in *Informatics* and *Computer Science* have played a crucial role in Optimization. Indeed, the complexity of the systems we consider interesting nowadays makes it impossible to implement efficient control strategies without using appropriate (and sophisticated) software.

In order to understand why Optimization techniques and Control Theory are closely related, let us come back to (1). Assume that the set of admissible controls $\mathcal{U}_{\rm ad}$ is a subset of the Banach space $\mathcal{U}$ (with norm $\|\cdot\|_{\mathcal{U}}$) and the state space Y is another Banach space (with norm $\|\cdot\|_Y$). Also, assume that the state $y_d \in Y$ is the preferred state and is chosen as a target for the state of the system. Then, the control problem consists in finding controls v in $\mathcal{U}_{\rm ad}$ such that the associated solution coincides or gets close to y_d.

It is then reasonable to think that a fruitful way to choose a good control v is by minimizing a *cost function* of the form

$$J(v) = \frac{1}{2}\|y(v) - y_d\|_Y^2 \quad \forall v \in \mathcal{U}_{\rm ad} \tag{3}$$

or, more generally,

$$J(v) = \frac{1}{2}\|y(v) - y_d\|_Y^2 + \frac{\mu}{2}\|v\|_{\mathcal{U}}^2 \quad \forall v \in \mathcal{U}_{\rm ad}\,, \tag{4}$$

where $\mu \geq 0$.

These are (constrained) extremal problems whose analysis corresponds to Optimization Theory.

It is interesting to analyze the two terms arising in the functional J in (4) when $\mu > 0$ separately, since they play complementary roles. When minimizing the functional in (4), we are minimizing the balance between these two terms. The first one requires to get close to the target y_d while the second one penalizes using too much costly control. Thus, roughly speaking, when minimizing J we are trying to drive the system to a state close to the target y_d without too much effort.

We will give below more details of the connection of Control Theory and Optimization below.

So far, we have mentioned three main ingredients arising in Control Theory: the notion of feedback, the need of fluctuations and Optimization. But of course in the development of Control Theory many other concepts have been important.

One of them is *Cybernetics*. The word “cybernétique” was proposed by the French physicist A.-M. Ampère in the XIX Century to design the nonexistent science of process controlling. This was quickly forgotten until 1948, when N. Wiener chose “Cybernetics” as the title of his book.

Wiener defined Cybernetics as “the science of control and communication in animals and machines”. In this way, he established the connection between Control Theory and Physiology and anticipated that, in a desirable future, engines would obey and imitate human beings.

At that time this was only a dream but now the situation is completely different, since recent developments have made possible a large number of new applications in *Robotics*, *Computer-Aided Design*, etc. (see [43] for an overview). Today, Cybernetics is not a dream any more but an ubiquitous reality. On the other hand, Cybernetics leads to many important questions that are relevant for the development of our society, very often in the borderline of *Ethics* and *Philosophy*. For instance,

Figure 2: Norbert Wiener (1894–1964).

> *Can we be inspired by Nature to create*
> *better engines and machines ?*

Or

> *Is the animal behavior an acceptable criterium to judge the performance of an engine ?*

Many movies of science fiction describe a world in which machines do not obey any more to humans and humans become their slaves. This is the opposite situation to the one Control Theory has been and is looking for. The development of Science and Technology is obeying very closely to the predictions made fifty years ago. Therefore, it seems desirable to deeply consider and revise our position towards Cybernetics from now on, many years ahead, as we do permanently in what concerns, for instance, Genetics and the possibilities it provides to intervene in human reproduction.

3 The pendulum

We will analyze in this Section a very simple and elementary control problem related to the dynamics of *the pendulum.*

The analysis of this model will allow us to present the most relevant ideas in the control of finite dimensional systems, that, as we said above, are essential for more sophisticated systems too. In our presentation, we will closely follow the book by E. Sontag [46].

The problem we discuss here, far from being purely academic, arises in many technological applications and in particular in Robotics, where the goal is to control a *gyratory arm* with a motor located at one extreme connecting the arm to the rest of the structure.

In order to model this system, we assume that the total mass m of the arm is located at the free extreme and the bar has unit length. Ignoring the effect of friction, we write

$$m\ddot{\theta}(t) = -mg\sin\theta(t) + v(t), \tag{5}$$

which is a direct consequence of *Newton's law.* Here, $\theta = \theta(t)$ is the angle of the arm with respect to the vertical axis measured counterclockwise, g is the acceleration due to gravity and u is the applied external *torsional momentum.* The state of the system is $(\theta, \dot{\theta})$, while $v = v(t)$ is the control (see Fig. 3).

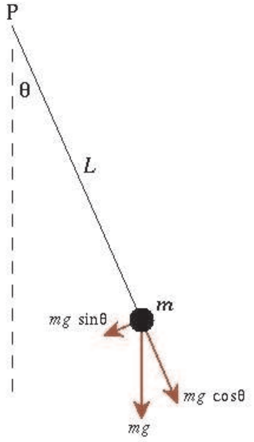


Figure 3: The simple pendulum.

To simplify our analysis, we also assume that $m = g = 1$. Then, (5) becomes:

$$\ddot{\theta}(t) + \sin\theta(t) = v(t). \tag{6}$$

The vertical stationary position ($\theta = \pi, \dot{\theta} = 0$) is an equilibrium configuration in the absence of control, i.e. with $v \equiv 0$. But, obviously, this is an *unstable* equilibrium. Let us analyze the system around this configuration, to understand how this instability can be compensated by means of the applied control force v.

Taking into account that $\sin\theta \sim \pi - \theta$ near $\theta = \pi$, at first approximation, the linearized system with respect to the variable $\varphi = \theta - \pi$ can be written in the form

$$\ddot{\varphi} - \varphi = v(t). \tag{7}$$

The goal is then to drive $(\varphi, \dot{\varphi})$ to the desired state $(0, 0)$ for all small initial data, without making the angle and the velocity too large along the controlled trajectory.

The following control strategy is in agreement with common sense: when the system is to the left of the vertical line, i.e. when $\varphi = \theta - \pi > 0$, we push the system towards the right side, i.e. we apply a force v with negative sign; on the other hand, when $\varphi < 0$, it seems natural to choose $v > 0$.

This suggests the following *feedback law*, in which the control is proportional to the state:

$$v = -\alpha\varphi, \qquad \text{with } \alpha > 0. \tag{8}$$

In this way, we get the *closed loop system*

$$\ddot{\varphi} + (\alpha - 1)\varphi = 0. \tag{9}$$

It is important to understand that, solving (9), we simultaneously obtain the state $(\varphi, \dot{\varphi})$ and the control $v = -\alpha\varphi$. This justifies, at least in this case, the relevance of a feedback law like (8).

The roots of the characteristic polynomial of the linear equation (9) are $z = \pm\sqrt{1-\alpha}$. Hence, when $\alpha > 1$, the nontrivial solutions of this differential equation are *oscillatory.* When $\alpha < 1$, all solutions diverge to $\pm\infty$ as $t \to \pm\infty$, except those satisfying

$$\dot{\varphi}(0) = -\sqrt{1-\alpha}\ \varphi(0).$$

Finally, when $\alpha = 1$, all nontrivial solutions satisfying $\dot{\varphi}(0) = 0$ are constant.

Thus, the solutions to the linearized system (9) do not reach the desired configuration $(0, 0)$ in general, independently of the constant α we put in (8).

This can be explained as follows. Let us first assume that $\alpha < 1$. When $\varphi(0)$ is positive and small and $\dot{\varphi}(0) = 0$, from equation (9) we deduce that $\ddot{\varphi}(0) > 0$. Thus, φ and $\dot{\varphi}$ grow and, consequently, the pendulum goes away from the vertical line. When $\alpha > 1$, the control acts on the correct direction but with too much inertia.

The same happens to be true for the nonlinear system (6).

The most natural solution is then to keep $\alpha > 1$, but introducing an additional term to diminish the oscillations and penalize the velocity. In this way, a new feedback law can be proposed in which the control is given as a linear combination of φ and $\dot{\varphi}$:

$$v = -\alpha\varphi - \beta\dot{\varphi}, \quad \text{with } \alpha > 1 \text{ and } \beta > 0. \tag{10}$$

The new closed loop system is

$$\ddot{\varphi} + \beta\dot{\varphi} + (\alpha - 1)\varphi = 0, \tag{11}$$

whose characteristic polynomial has the following roots

$$\frac{-\beta \pm \sqrt{\beta^2 - 4(\alpha - 1)}}{2}. \tag{12}$$

Now, the real part of the roots is negative and therefore, all solutions converge to zero as $t \to +\infty$. Moreover, if we impose the condition

$$\beta^2 > 4(\alpha - 1), \tag{13}$$

we see that solutions tend to zero monotonically, without oscillations.

This simple model is rich enough to illustrate some systematic properties of control systems:

- Linearizing the system is an useful tool to address its control, although the results that can be obtained this way are only of local nature.
- One can obtain feedback controls, but their effects on the system are not necessarily in agreement with the very first intuition. Certainly, the (asymptotic) stability properties of the system must be taken into account.
- Increasing dissipation one can eliminate the oscillations, as we have indicated in (13).

In connection with this last point, notice however that, as dissipation increases, trajectories converge to the equilibrium more slowly. Indeed, in (10), for fixed $\alpha > 1$, the value of β that minimizes the largest real part of a root of the characteristic polynomial (11) is

$$\beta = 2\sqrt{\alpha - 1}.$$

With this value of β, the associated real part is

$$\sigma^* = -\sqrt{\alpha - 1}$$

and, increasing β, the root corresponding to the plus sign increases and converges to zero:

$$\frac{-\beta + \sqrt{\beta^2 - 4(\alpha - 1)}}{2} > -\sqrt{\alpha - 1} \quad \forall \beta > 2\sqrt{\alpha - 1} \tag{14}$$

and

$$\frac{-\beta + \sqrt{\beta^2 - 4(\alpha - 1)}}{2} \to 0^- \quad \text{as } \beta \to +\infty. \tag{15}$$

This phenomenon is known as *overdamping* in Engineering and has to be taken into account systematically when designing feedback mechanisms.

At the practical level, implementing the control (10) is not so simple, since the computation of v requires knowing the position φ and the velocity $\dot{\varphi}$ at every time.

Let us now describe an interesting alternative. The key idea is to evaluate φ and $\dot{\varphi}$ only on a discrete set of times

$$0,\ \delta,\ 2\delta,\ \ldots,\ k\delta,\ \ldots$$

and modify the control at each of these values of t. The control we get this way is kept constant along each interval $[k\delta, (k+1)\delta]$.

Computing the solution to system (7), we see that the result of applying the constant control v_k in the time interval $[k\delta, (k+1)\delta]$ is as follows:

$$\begin{pmatrix} \varphi(k\delta+\delta) \\ \dot{\varphi}(k\delta+\delta) \end{pmatrix} = A \begin{pmatrix} \varphi(k\delta) \\ \dot{\varphi}(k\delta) \end{pmatrix} + v_k\, b,$$

where

$$A = \begin{pmatrix} \cos h\delta & \sin h\delta \\ \sin h\delta & \cos h\delta \end{pmatrix}, \quad b = \begin{pmatrix} \cos h\delta - 1 \\ \sin h\delta \end{pmatrix}.$$

Thus, we obtain a discrete system of the form

$$x_{k+1} = (A + bf^t)x_k\,,$$

where f is the vector such that

$$v_k = f^t x_k\,.$$

Observe that, if f is such that the matrix $A + bf^t$ is nilpotent, i.e.

$$[A + bf^t]^2 = 0,$$

then we reach the equilibrium in two steps. A simple computation shows that this property holds if $f^t = (f_1, f_2)$, with

$$f_1 = \frac{1 - 2\cos h\delta}{2(\cos h\delta - 1)}\,, \quad f_2 = -\frac{1 + 2\cos h\delta}{2\sin h\delta}\,. \tag{16}$$

The main advantage of using controllers of this form is that we get the stabilization of the trajectories in finite time and not only asymptotically, as $t \to +\infty$. The controller we have designed is a *digital control* and it is extremely useful because of its robustness and the ease of its implementation.

The digital controllers we have built are similar and closely related to the *bang-bang* controls we are going to describe now.

Once $\alpha > 1$ is fixed, for instance $\alpha = 2$, we can assume that

$$v = -2\varphi + w, \tag{17}$$

so that (7) can be written in the form

$$\ddot{\varphi} + \varphi = w. \tag{18}$$

This is Newton's law for the vibration of a spring.

This time, we look for controls below an admissible cost. For instance, we impose

$$|w(t)| \leq 1 \quad \forall t.$$

The function $w = w(t)$ that, satisfying this constraint, controls the system in minimal time, i.e. *the optimal control*, is necessarily of the form

$$w(t) = \mathrm{sgn}(p(t)),$$

where η is a solution of

$$\ddot{p} + p = 0.$$

This is a consequence of *Pontryagin's maximum principle* (see Appendix 1 for more details).

Therefore, the optimal control takes only the values ±1 and, in practice, it is sufficient to determine the *switching times* at which the sign of the optimal control changes.

In order to compute the optimal control, let us first compute the solutions corresponding to the extremal controllers ±1. Using the new variables x_1 and x_2 with $x_1 = \varphi$ and $x_2 = \dot{\varphi}$, this is equivalent to solve the systems

$$\begin{cases} \dot{x}_1 = x_2 \\ \dot{x}_2 = -x_1 + 1 \end{cases} \tag{19}$$

and

$$\begin{cases} \dot{x}_1 = x_2 \\ \dot{x}_2 = -x_1 - 1. \end{cases} \tag{20}$$

The solutions can be identified to the circumferences in the plane (x_1, x_2) centered at $(1, 0)$ and $(-1, 0)$, respectively. Consequently, in order to drive (18) to the final state $(\varphi, \dot{\varphi})(T) = (0, 0)$, we must follow these circumferences, starting from the prescribed initial state and switching from one to another appropriately.

For instance, assume that we start from the initial state $(\varphi, \dot{\varphi})(0) = (\varphi^0, \varphi^1)$, where φ^0 and φ^1 are positive and small (see Fig. 4). Then, we first take $w(t) = 1$ and solve (19) for $t \in [0, T_1]$, where T_1 is such that $x_2(T_1) = 0$, i.e. we follow counterclockwise the arc connecting the points (φ^0, φ^1) and $(x_1(T_1), 0)$ in the (x_1, x_2) plane. In a second step, we take $w(t) = -1$ and solve (20) for $t \in [T_1, T_2]$, where T_2 is such that $(1 - x_1(T_2))^2 + x_2(T_2)^2 = 1$. We thus follow (again counterclockwise) the arc connecting the points $(x_1(T_1), 0)$ and $(x_1(T_2), x_2(T_2))$. Finally, we take $w(t) = 1$ and solve (19) for $t \in [T_2, T_3]$, with T_3 such that $x_1(T_3 = x_2(T_3) = 0$.

Similar constructions of the control can be done when $\varphi^0 \leq 1$ or $\varphi^1 \leq 0$.

In this way, we reach the equilibrium $(0, 0)$ in finite time and we obtain a feedback mechanism

$$\ddot{\varphi} + \varphi = F(\varphi, \dot{\varphi}),$$

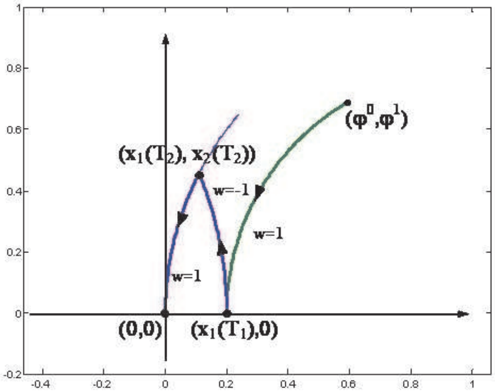


Figure 4: The action of a bang-bang control.

where F is the function taking the value -1 above the switching curve and $+1$ below. In what concerns the original system (7), we have

$$\ddot{\varphi} - \varphi = -2\varphi + F(\varphi, \dot{\varphi}).$$

The action of the control in this example shows clearly the suitability of self-regulation mechanisms. If we want to lead the system to rest in a minimal time, it is advisable to do it following a somewhat indirect path, allowing the system to evolve naturally and avoiding any excessive forcing.

Bang-bang controllers are of high interest for practical purposes. Although they might seem irregular and unnatural, they have the advantages of providing minimal time control and being easy to compute.

As we said above, although the problem we have considered is very simple, it leads naturally to some of the most relevant ideas of Control Theory: feedback laws, overdamping, digital and bang-bang controls, etc.

4 History and contemporary applications

In this paper, we do not intend to make a complete overview of the history of Control Theory, nor to address its connections with the philosophical questions we have just mentioned. Without any doubt, this would need much more space. Our intention is simply to recall some classical and well known results that have to some extent influenced the development of this discipline, pointing out several facts that, in our opinion, have been relevant for the recent achievements of Control Theory.

Let us go back to the origins of Control Engineering and Control Theory and let us describe the role this discipline has played in History.

Figure 5: A Roman aqueduct.

Going backwards in time, we will easily conclude that Romans did use some elements of Control Theory in their aqueducts. Indeed, ingenious systems of regulating valves were used in these constructions in order to keep the water level constant.

Some people claim that, in the ancient Mesopotamia, more than 2000 years B.C., the control of the irrigation systems was also a well known art.

On the other hand, in the ancient Egypt the "harpenodaptai" (string stretchers), were specialized in stretching very long strings leading to long straight segments to help in large constructions. Somehow, this is an evidence of the fact that in the ancient Egypt the following two assertions were already well understood:

- The shortest distance between two points is the straight line (which can be considered to be the most classical assertion in Optimization and Calculus of Variations);

- This is equivalent to the following dual property: among all the paths of a given length the one that produces the longest distance between its extremes is the straight line as well.

The task of the "harpenodaptai" was precisely to build these "optimal curves".

The work by Ch. Huygens and R. Hooke at the end of the XVII Century on the *oscillations of the pendulum* is a more modern example of development in Control Theory. Their goal was to achieve a precise measurement of time and location, so precious in navigation.

These works were later adapted to regulate the velocity of windmills. The main mechanism was based on a system of balls rotating around an axis, with a velocity proportional to the velocity of the windmill. When the rotational velocity increased, the balls got farther from the axis, acting on the wings of the mill through appropriate mechanisms.

J. Watt adapted these ideas when he invented the *steam engine* and this constituted a magnificent step in the industrial revolution. In this mechanism, when the velocity of the balls increases, one or several valves open to let the vapor scape. This makes the pressure diminish. When this happens, i.e. when the pressure inside the boiler becomes weaker, the velocity begins to go down. The goal of introducing and using this mechanism is of course to keep the velocity as close as possible to a constant.

Figure 6: J. Watt (1736–1819).

The British astronomer G. Airy was the first scientist to analyze mathematically the regulating system invented by Watt. But the first definitive mathematical description was given only in the works by J.C. Maxwell, in 1868, where some of the erratic behaviors encountered in the steam engine were described and some control mechanisms were proposed.

The central ideas of Control Theory gained soon a remarkable impact and, in the twenties, engineers were already preferring the continuous processing and using semi-automatic or automatic control techniques. In this way, Control Engineering germinated and got the recognition of a distinguished discipline.

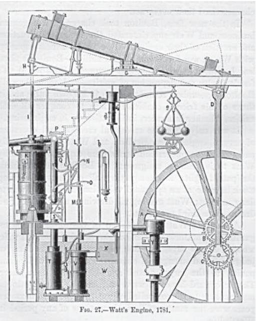


Figure 7: Watt's 1781 steam engine (taken from [50]).

In the thirties important progresses were made on automatic control and design and analysis techniques. The number of applications increased covering *amplifiers* in telephone systems, distribution systems in electrical plants, stabilization of aeroplanes, electrical mechanisms in paper production, Chemistry, petroleum and steel Industry, etc.

By the end of that decade, two emerging and clearly different methods or approaches were available: a first method based on the use of differential equations and a second one, of frequential nature, based on the analysis of amplitudes and phases of "inputs" and "outputs".

By that time, many institutions took conscience of the relevance of automatic control. This happened for instance in the American ASME (American Society of Mechanical Engineers) and the British IEE (Institution of Electrical Engineers). During the Second World War and the following years, engineers and scientists improved their experience on the control mechanisms of plane tracking and ballistic missiles and other designs of anti-aircraft batteries. This produced an important development of frequential methods.

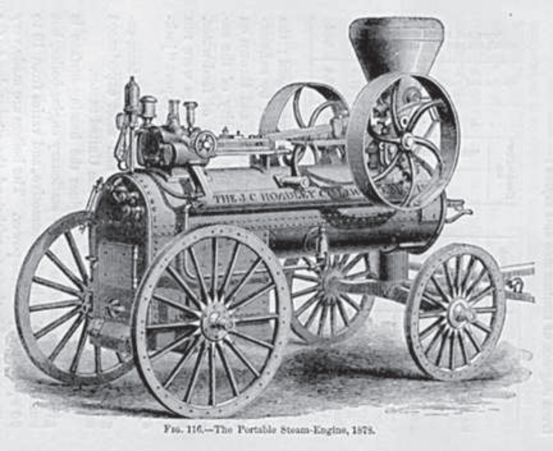


Figure 8: A primitive steam engine (taken from [50]).

After 1960, the methods and ideas mentioned above began to be considered as part of "classical" Control Theory. The war made clear that the models considered up to that moment were not accurate enough to describe the complexity of the real word. Indeed, by that time it was clear that *true systems* are often *nonlinear* and *nondeterministic*, since they are affected by "noise". This generated important new efforts in this field.

The contributions of the U.S. scientist R. Bellman in the context of *dynamic programming*, R. Kalman in *filtering techniques* and the algebraic approach to linear systems and the Russian L. Pontryagin with the *maximum principle* for nonlinear optimal control problems established the foundations of modern Control Theory.

We shall describe in Section 6 the approach by Kalman to the controllability of linear finite dimensional systems. Furthermore, at the end of this paper we give two short Appendices where we have tried to present, as simply as possible, the central ideas of Bellman's and Pontryagin's works.

As we have explained, the developments of Industry and Technology had a tremendous impact in the history of Control Engineering. But the development of Mathematics had a similar effect.

Indeed, we hav already mentioned that, in the late thirties, two emerging strategies were already established. The first one was based on the use of differential equations and, therefore, the contributions made by the most celebrated mathematicians between the XVIIth and the XIXth Centuries played a fundamental role in that approach. The second one, based on a frequential approach, was greatly influenced by the works of J. Fourier.

Accordingly, Control Theory may be regarded nowadays from two different and complementary points of view: as a theoretical support to *Control Engineering* (a part of *System Engineering*) and also as a mathematical

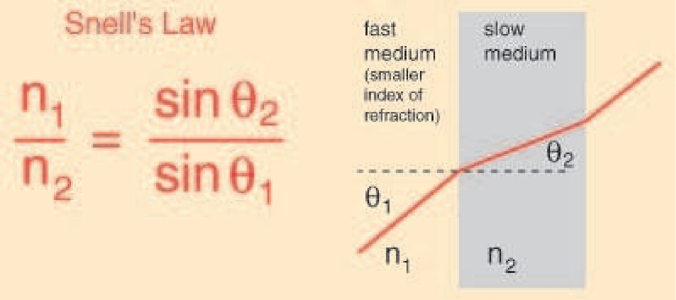

Snell's Law
$\frac{n_1}{n_2} = \frac{\sin\theta_2}{\sin\theta_1}$
fast medium (smaller index of refraction)
slow medium
θ_1
θ_2
n_1
n_2

Figure 9: Snell's law of refraction of light (this Figure has been taken from `http://hyperphysics.phy-astr.gsu.edu/hbase/geoopt/refr.html`).

discipline. In practice, the frontiers between these two subworlds are extremely vague. In fact, Control Theory is one of the most interdisciplinary areas of Science nowadays, where Engineering and Mathematics melt perfectly and enrich each other.

Mathematics is currently playing an increasing role in Control Theory. Indeed, the degree of sophistication of the systems that Control Theory has to deal with increases permanently and this produces also an increasing demand of Mathematics in the field.

Along these notes, it will become clear that Control Theory and Calculus of Variations have also common roots. In fact, these two disciplines are very often hard to distinguish.

The history of the Calculus of Variations is also full of mathematical achievements. We shall now mention some of them.

As we said above, one can consider that the starting point of the Calculus of Variations is the understanding that the straight line is the shortest path between two given points. In the first Century, Heron of Alexandria showed in his work "La Catoptrique" that the law of reflection of light (the fact that the incidence and reflection angles are identical) may be obtained as a consequence of the variational principle that light minimizes distance along the preferred path.

In the XVII Century, P. De Fermat generalized this remark by Heron and formulated the following minimum principle:

> *Light in a medium with variable velocity prefers the path that guarantees the minimal time.*

Later Leibnitz and Huygens proved that the law of refraction of light may be obtained as a consequence of Fermat's principle.

The refraction law had been discovered by G. Snell in 1621, although it remained unpublished until 1703, as Huygens published his *Dioptrica*. An explanation of this law is given in Fig. 9, where we have denoted by n_i the

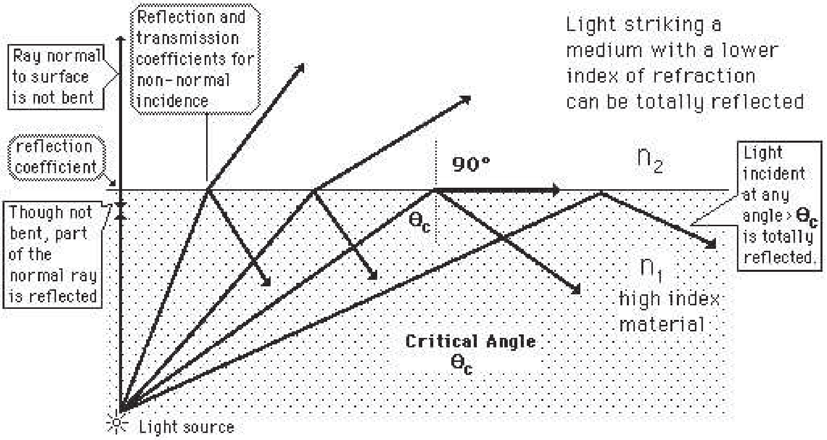

Ray normal to surface is not bent
Reflection and transmission coefficients for non-normal incidence
Light striking a medium with a lower index of refraction can be totally reflected
reflection coefficient
90°
n_2
Light incident at any angle > θ_c is totally reflected.
Though not bent, part of the normal ray is reflected
θ_c
n_1
high index material
Critical Angle
θ_c
Light source

Figure 10: Critical angle and total reflection (this Figure has been taken from `http://hyperphysics.phy-astr.gsu.edu/hbase/geoopt/refr.html`).

index of refraction of the i-th medium. By definition, we have $n_i = c/v_i$, where c and v_i are the speeds of propagation of light in the vacuum and the i-th medium, respectively.

It is interesting to observe that, in account of this principle, a ray of light may be unable to propagate from a *slow* medium to a *fast* medium. Indeed, if $n_1 > n_2$, there exists a critical angle θ_c such that, when $\theta_1 > \theta_c$, Snell's law cannot be satisfied whatever θ_2 is. This has been explained in Fig. 10. Contrarily, the light can always propagate from a fast to a slow medium.

In 1691, J. Bernoulli proved that the *catenary* is the curve which provides the shape of a string of a given length and constant density with fixed ends under the action of gravity. Let us also mention that the problem of the *bachistocrone*, formulated by Bernoulli in 1696, is equivalent to finding the rays of light in the upper half-plane $y \geq 0$ corresponding to a light velocity c given by the formula $c(x, y) = \sqrt{y}$ (Newton proved in 1697 that the solution is the *cycloid*). The reader interested in these questions may consult the paper by H. Sussmann [48].

R. Kalman, one of the greatest protagonists of modern Control Theory, said in 1974 that, in the future, the main advances in Control and Optimization of systems would come more from mathematical progress than from the technological development. Today, the state of the art and the possibilities that Technology offers are so impressive that maintaining that statement is probably very risky. But, without any doubt, the development of Control Theory will require deep contributions coming from both fields.

In view of the rich history of Control Theory and all the mathematical achievements that have been undertaken in its domain of influence, one could ask whether the field has reached its end. But this is far from reality. Our society provides every day new problems to Control Theory and this fact is stimulating the creation of new Mathematics.

Indeed, the range of applications of Control Theory goes from the simplest mechanisms we manipulate in everyday life to the most sophisticated ones, emerging in new technologies.

The book edited by W.S. Levine [26] provides a rather complete description of this variety of applications.

One of the simplest applications of Control Theory appears in such an apparently simple machine as the tank of our bathroom. There are many variants of tanks and some of the licences go back to 1886 and can be found in [25]. But all them work under the same basic principles: the tank is supplied of regulating valves, security mechanisms that start the control process, feedback mechanisms that provide more or less water to the tank depending of the level of water in its interior and, finally, mechanisms that avoid the unpleasant flooding in case that some of the other components fail.

The systems of heating, ventilation and air conditioning in big buildings are also very efficient large scale control systems composed of interconnected thermo-fluid and electro-mechanical subsystems. The main goal of these systems is to keep a comfortable and good quality air under any circumstance, with a low operational cost and a high degree of reliability. The relevance of a proper and efficient functioning of these systems is crucial from the viewpoint of the impact in Economical and Environmental Sciences. The predecessor of these sophisticated systems is the classical *thermostat* that we all know and regulates temperature at home.

The list of applications of Control Theory in Industry is endless. We can mention, for instance, the *pH* control in chemical reactions, the paper and automobile industries, nuclear security, defense, etc.

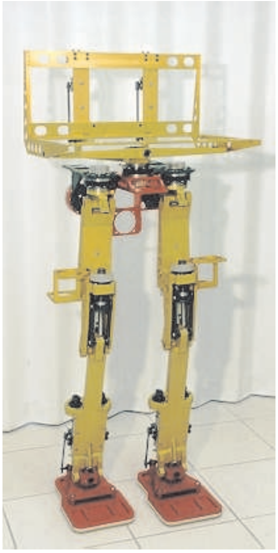


Figure 11: The BIP2000 antropomorphic biped.

The control of *chaos* is also being considered by many researchers nowadays. The chaotic behavior of a system may be an obstacle for its control; but it may also be of help. For instance, the control along unstable trajectories is of great use in controlling the dynamics of fight aircrafts. We refer to [35] for a description of the state of the art of *active control* in this area.

Space structures, optical reflectors of large dimensions, satellite communication systems, etc. are also examples of modern and complex control systems. The control of *robots*, ranging from the most simple engines to the *bipeds* that simulate the locomotive ability of humans is also another emerging area of Control Theory.

For instance, see the web page `http://www.inrialpes.fr/bipop/` of the

French Institute I.N.R.I.A. (Institut National de Recherche en Informatique et Automatique), where illustrating images and movies of the antropomorphic biped BIP2000 can be found (two of these Figures have been reproduced here; see Fig. 11 and 12).

Compact disk players is another area of application of modern control systems. A CD player is endowed with an optical mechanism allowing to interpret the registered code and produce an acoustic signal. The main goal when designing CD players is to reach higher velocities of rotation, permitting a faster reading, without affecting the stability of the disk. The control mechanisms have to be even more robust when dealing with portable equipments.

Electrical plants and distribution networks are other modern applications of Control Theory that influence significantly our daily life. There are also many relevant applications in Medicine ranging from artificial organs to mechanisms for insulin supply, for instance.

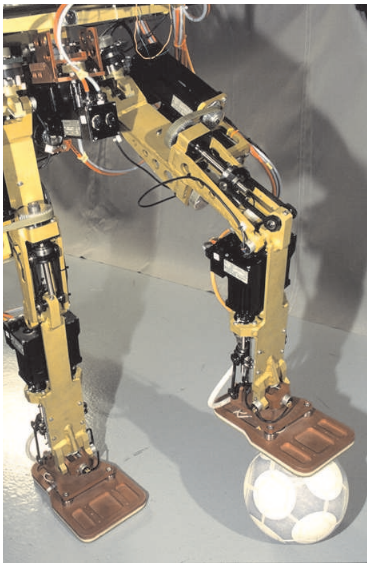


Figure 12: A second view of the BIP2000 biped.

We could keep quoting other relevant applications. But those we have mentioned and some others that will appear later suffice to prove the ubiquity of control mechanisms in the real world. The underlying mathematical theory is also impressive. The reader interested in an introduction to the classical and basic mathematical techniques in Control Engineering is referred to [8] and [36].

5 Controllability versus optimization

As already mentioned, for systems of the form (1), the main goal of Control Theory is to find *controls* v leading the *associated states* $y(v)$, i.e. the solutions of the corresponding controlled systems, to a desired situation.

There are however (at least) two ways of specifying a “desired prescribed situation”:

- To fix a desired state y_d and require

$$y(v) = y_d \tag{21}$$

or, at least,

$$y(v) \sim y_d \tag{22}$$

in some sense. This is the *controllability* viewpoint.

The main question is then the existence of an admissible control v so that the corresponding state $y(v)$ satisfies (21) or (22). Once the existence of such a control v is established, it is meaningful to look for an optimal control, for instance, a control of *minimal size*. Other important questions arise in this context too. For instance, the existence of "bang-bang" controls, the minimal time of control, etc.

As we shall see, this problem may be difficult (or even very difficult) to solve. In recent years, an important body of beautiful Mathematics has been developed in connection with these questions.

- To fix a *cost function* $J = J(v)$ like for instance (3) or (4) and to look for a *minimizer* u of J. This is the *optimization* or *optimal control* viewpoint.

  As in (3) and (4), J is typically related to the "distance" to a prescribed state. Both approaches have the same ultimate goal, to bring the state close to the desired target but, in some sense, the second one is more realistic and easier to implement.

The optimization viewpoint is, at least apparently, humble in comparison with the controllability approach. But it is many times much more realistic. In practice, it provides satisfactory results in many situations and, at the same time, it requires simpler mathematical tools.

To illustrate this, we will discuss now a very simple example. It is trivial in the context of Linear Algebra but it is of great help to introduce some of the basic tools of Control Theory.

We will assume that the state equation is

$$Ay = b, \tag{23}$$

where A is a $n \times n$ real matrix and the state is a column vector $y = (y_1, y_2, \dots, y_n)^t \in \mathbb{R}^n$. To simplify the situation, let us assume that A is nonsingular. The control vector is $b \in \mathbb{R}^n$. Obviously, we can rewrite (23) in the form $y = A^{-1}b$, but we do not want to do this. In fact, we are mainly interested in those cases in which (23) can be difficult to solve.

Let us first adopt the controllability viewpoint. To be specific, let us impose as an objective to make the first component y_1 of y coincide with a prescribed value y_1^*:

$$y_1 = y_1^* \,. \tag{24}$$

This is the sense we are giving to (22) in this particular case. So, we are consider the following controllability problem:

PROBLEM 0: *To find* $b \in \mathbb{R}^n$ *such that the solution of* (23) *satisfies* (24).

Roughly speaking, we are addressing here a *partial controllability* problem, in the sense that we are controlling only one component, y_1, of the state.

Obviously, such controls b exist. For instance, it suffices to take $y^* = (y_1^*, 0, \cdots, 0)^t$ and then choose $b = Ay^*$. But this argument, by means of which we find the state directly without previously determining the control, is frequently impossible to implement in practice. Indeed, in most real problems, we have first to find the control and, only then, we can compute the state by solving the state equation.

The number of control parameters (the n components of b) is greater or equal than the number of state components we have to control. But, what happens if we stress our own possibilities ? What happens if, for instance, $b_1, \ldots, b_{n-1}$ are fixed and we only have at our disposal b_n to control the system ?

From a mathematical viewpoint, the question can be formulated as follows. In this case,

$$Ay = c + be \tag{25}$$

where $c \in \mathbb{R}^n$ is a prescribed column vector, e is the unit vector $(0, \ldots, 0, 1)^t$ and b is a scalar control parameter. The corresponding controllability problem is now the following:

> PROBLEM 1: *To find $b \in \mathbb{R}$ such that the solution of* (25) *satisfies* (24).

This is a less obvious question. However, it is not too difficult to solve. Note that the solution y to (25) can be decomposed in the following way:

$$y = x + z, \tag{26}$$

where

$$x = A^{-1}c \tag{27}$$

and z satisfies

$$Az = be, \quad \text{i.e.} \quad z = bz^* \quad z^* = A^{-1}e. \tag{28}$$

To guarantee that y_1 can take *any* value in $\mathbb{R}$, as we have required in (24), it is necessary and sufficient to have $z_1^* \neq 0$, z_1^* being the first component of $z^* = A^{-1}e$.

In this way, we have a precise answer to this second controllability problem:

> *The problem above can be solved for any y_1^* if and only if the first component of $A^{-1}e$ does not vanish.*

Notice that, when the first component of $A^{-1}e$ vanishes, whatever the control b is, we always have $y_1 = x_1$, x_1 being the first component of the fixed vector x in (27). In other words, y_1 *is not sensitive* to the control b_n . In this degenerate case, the set of values taken by y_1 is a singleton, a 0-dimensional manifold. Thus, we see that the state is confined in a "space" of low dimension and controllability is lost in general.

But, is it really frequent in practice to meet degenerate situations like the previous one, where some components of the system are insensitive to the control ?

Roughly speaking, it can be said that systems are generically not degenerate. In other words, in examples like the one above, it is actually rare that z_1^* vanishes.

There are however a few remarks to do. When z_1^* does not vanish but is very small, even though controllability holds, the control process is very unstable in the sense that one needs very large controls in order to get very small variations of the state. In practice, this is very important and must be taken into account (one needs the system not only to be controllable but this to happen with realistic and feasible controls).

On the other hand, it can be easily imagined that, when systems under consideration are complex, i.e. many parameters are involved, it is difficult to know *a priori* whether or not there are components of the state that are insensitive to the control.

Let us now turn to the optimization approach. Let us see that the difficulties we have encountered related to the possible degeneracy of the system disappear (which confirms the fact that this strategy leads to easier questions).

For example, let us assume that $k > 0$ is a reasonable bound of the control b that we can apply. Let us put

$$J(b_n) = \frac{1}{2}|y_1 - y_1^*|^2 \qquad \forall b_n \in \mathbb{R}, \tag{29}$$

where y_1 is the first component of the solution to (25). Then, it is reasonable to admit that the best response is given by the solution to the following problem:

PROBLEM 1′: *To find* $b_n^k \in [-k, k]$ *such that*

$$J(b_n^k) \le J(b_n) \quad \forall b_n \in [-k, k]. \tag{30}$$

Since $b_n \mapsto J(b_n)$ is a continuous function, it is clear that this problem possesses a solution $b_n^k \in I_k$ for each $k > 0$. This confirms that the considered optimal control problem is simpler.

On the other hand, this point of view is completely natural and agrees with common sense. According to our intuition, most systems arising in real life should possess an optimal strategy or configuration. At this respect L. Euler said:

> "Universe is the most perfect system, designed by the most wise Creator. Nothing will happen without emerging, at some extent, a maximum or minimum principle".

Let us analyze more closely the similarities and differences arising in the two previous formulations of the control problem.

- Assume the controllability property holds, that is, PROBLEM 1 is solvable for any y_1^*. Then, if the target y_1^* is given and k is sufficiently large, the solution to PROBLEM 1′ coincides with the solution to PROBLEM 1.

In fact, it is a very interesting and non trivial task to design strategies guaranteeing that we do not fall in a degenerate situation.

- On the other hand, when there is no possibility to attain y_1^* exactly, the optimization viewpoint, i.e. PROBLEM $1'$, furnishes the best response.
- To investigate whether the controllability property is satisfied, it can be appropriate to solve PROBLEM $1'$ for each $k > 0$ and analyze the behavior of the *cost*
$$J_k = \min_{b_n \in [-k,k]} J(b_n) \tag{31}$$
as k grows to infinity. If J_k stabilizes near a positive constant as k grows, we can suspect that y_1^* cannot be attained exactly, i.e. that PROBLEM 1 does not have a solution for this value of y_1^*.

In view of these considerations, it is natural to address the question of whether it is actually necessary to solve controllability problems like PROBLEM 1 or, by the contrary, whether solving a related optimal control problem (like PROBLEM $1'$) suffices.

There is not a generic and systematic answer to this question. It depends on the level of precision we require to the control process and this depends heavily on the particular application one has in mind. For instance, when thinking of technologies used to stabilize buildings, or when controlling space vehicles, etc., the efficiency of the control that is required demands much more than simply choosing the best one with respect to a given criterion. In those cases, it is relevant to know how close the control will drive the state to the prescribed target. There are, consequently, a lot of examples for which simple optimization arguments as those developed here are insufficient.

In order to choose the appropriate control we need first to develop a rigorous modelling (in other words, we have to put equations to the real life system). The choice of the control problem is then a second relevant step in modelling.

Let us now recall and discuss some mathematical techniques allowing to handle the minimization problems arising in the optimization approach (in fact, we shall see that these techniques are also relevant when the controllability point of view is adopted).

These problems are closely related to the Calculus of Variations. Here, we do not intend to provide a survey of the techniques in this field but simply to mention some of the most common ideas.

For clarity, we shall start discussing *Mathematical Programming*. In the context of Optimization, *Programming* is not the art of writing computer codes. It was originated by the attempt to *optimize* the planning of the various tasks or activities in an organized system (a plant, a company, etc.). The goal is then to find what is known as an *optimal planning* or *optimal programme*.

The simplest problem of *assignment* suffices to exhibit the need of a mathematical theory to address these issues.

Assume that we have 70 workers in a plant. They have different qualifications and we have to assign them 70 different tasks. The total number of possible distributions is $70!$, which is of the order of 10^{100}. Obviously, in order to be able to solve rapidly a problem like this, we need a mathematical theory to provide a good strategy.

This is an example of assignment problem. Needless to say, problems of this kind are not only of academic nature, since they appear in most human activities.

In the context of *Mathematical Programming*, we first find *linear programming techniques.* As their name indicates, these are concerned with those optimization problems in which the involved functional is linear.

Linear Programming was essentially unknown before 1947, even though Joseph Fourier had already observed in 1823 the relevance of the questions it deals with. L.V. Kantorovich, in a monograph published in 1939, was the first to indicate that a large class of different planning problems could be covered with the same formulation. The *method of simplex,* that we will recall below, was introduced in 1947 and its efficiency turned out to be so impressive that very rapidly it became a common tool in Industry.

There has been a very intense research in these topics that goes beyond Linear Programming and the method of simplex. We can mention for instance *nonlinear programming methods*, inspired by the *method of descent.* This was formally introduced by the French mathematician A.L. Cauchy in the XIX Century. It relies on the idea of solving a nonlinear equation by searching the critical points of the corresponding primitive function.

Let us now give more details on Linear Programming. At this point, we will follow a presentation similar to the one by G. Strang in [47].

The problems that one can address by means of linear programming involve the minimization of linear functions subject to linear constraints. Although they seem extremely simple, they are ubiquitous and can be applied in a large variety of areas such as the control of traffic, Game Theory, Economics, etc. Furthermore, they involve in practice a huge quantity of unknowns, as in the case of the optimal planning problems we have presented before.

The simplest problem in this field can be formulated in the following way:

> *Given a real matrix A of order $M \times N$ (with $M \leq N$), and given a column vector b of M components and a column vector c with N components, to minimize the linear function*
>
> $$\langle c, x\rangle = c_1 x_1 + \cdots + c_N x_N$$
>
> *under the restrictions*
>
> $$Ax = b, \quad x \geq 0.$$

Here and in the sequel, we use $\langle \cdot\,, \cdot\rangle$ to denote the usual Euclidean scalar products in $\mathbb{R}^N$ and $\mathbb{R}^M$. The associated norm will be denoted by $|\cdot|$.

Of course, the second restriction has to be understood in the following way:

$$x_j \geq 0, \quad j = 1, \ldots, N.$$

In general, the solution to this problem is given by a unique vector x with the property that $N - M$ components vanish. Accordingly, the problem consists in finding out which are the $N-M$ components that vanish and, then, computing the values of the remaining M components.

The method of simplex leads to the correct answer after a finite number of steps. The procedure is as follows:

- Step 1: We look for a vector x with $N-M$ zero components and satisfying $Ax = b$, in addition to the unilateral restriction $x \geq 0$. Obviously, this first choice of x will not provide the optimal answer in general.

- Step 2: We modify appropriately this first choice of x allowing one of the zero components to become positive and vanishing one of the positive components and this in such a way that the restrictions $Ax = b$ and $x \geq 0$ are kept.

After a finite number of steps like Step 2, the value of $\langle c, x \rangle$ will have been tested at all possible minimal points. Obviously, the solution to the problem is obtained by choosing, among these points x, that one at which the minimum of $\langle c, x \rangle$ is attained.

Let us analyze the geometric meaning of the simplex method with an example.

Let us consider the problem of minimizing the function

$$10x_1 + 4x_2 + 7x_3$$

under the constraints

$$2x_1 + x_2 + x_3 = 1, \quad x_1\,, x_2\,, x_3 \geq 0.$$

In this case, the set of admissible triplets $(x_1\,, x_2\,, x_3)$, i.e. those satisfying the constraints is the triangle in $\mathbb{R}^3$ of vertices $(0,0,1)$, $(0,1,0)$ and $(1/2,0,0)$ (a face of a tetrahedron). It is easy to see that the minimum is achieved at $(0,1,0)$, where the value is 4.

Let us try to give a geometrical explanation to this fact. Since $x_1\,, x_2\,, x_3 \geq 0$ for any admissible triplet, the minimum of the function $10x_1 + 4x_2 + 7x_3$ has necessarily to be nonnegative. Moreover, the minimum cannot be zero since the hyperplane

$$10x_1 + 4x_2 + 7x_3 = 0$$

has an empty intersection with the triangle of admissible states. When increasing the cost $10x_1 + 4x_2 + 7x_3$, i.e. when considering level sets of the form $10x_1 + 4x_2 + 7x_3 = c$ with increasing $c > 0$, we are considering planes parallel to $10x_1 + 4x_2 + 7x_3 = 0$ that are getting away from the origin and closer to the triangle of admissible states. The first value of c for which the level set intersects the admissible triangle provides the minimum of the cost function and the point of contact is the minimizer.

It is immediate that this point is the vertex $(0,1,0)$.

These geometrical considerations indicate the relevance of the convexity of the set where the minimum is being searched. Recall that, in a linear space E, a set K is *convex* if it satisfies the following property:

$$x, y \in K, \quad \lambda \in [0,1] \Rightarrow \lambda x + (1-\lambda) y \in K.$$

The crucial role played by convexity will be also observed below, when considering more sophisticated problems.

The method of simplex, despite its simplicity, is very efficient. There are many variants, adapted to deal with particular problems. In some of them, when looking for the minimum, one runs across the convex set and not only along its boundary. For instance, this is the case of *Karmakar's method*, see [47]. For more information on Linear Programming, the method of simplex and its variants, see for instance [40].

As the reader can easily figure out, many problems of interest in Mathematical Programming concern the minimization of *nonlinear* functions. At this respect, let us recall the following fundamental result whose proof is the basis of the so called *Direct Method of the Calculus of Variations* (DMCV):

Theorem 1 *If H is a Hilbert space with norm $\|\cdot\|_H$ and the function $J : H \mapsto \mathbb{R}$ is continuous, convex and coercive in H, i.e. it satisfies*

$$J(v) \to +\infty \quad as \quad \|v\|_H \to +\infty, \tag{32}$$

then J attains its minimum at some point $u \in H$. If, moreover, J is strictly convex, this point is unique.

If, in the previous result, J is a C^1 function, any minimizer u necessarily satisfies

$$J'(u) = 0, \quad u \in H. \tag{33}$$

Usually, (33) is known as the *Euler equation* of the minimization problem

$$\text{Minimize } J(v) \text{ subject to } v \in H. \tag{34}$$

Consequently, if J is C^1, Theorem 1 serves to prove that the (generally nonlinear) Euler equation (33) possesses at least one solution.

Many systems arising in Continuum Mechanics can be viewed as the Euler equation of a minimization problem. Conversely, one can associate Euler equations to many minimization problems. This mutual relation can be used in both directions: either to solve differential equations by means of minimization techniques, or to solve minimization problems through the corresponding Euler equations.

In particular, this allows proving existence results of equilibrium configurations for many problems in Continuum Mechanics.

Furthermore, combining these ideas with the approximation of the space H where the minimization problem is formulated by means of finite dimensional spaces and increasing the dimension to cover in the limit the whole space H, one obtains *Galerkin's approximation method.* Suitable choices of the approximating subspaces lead to the *finite element methods.*

In order to illustrate these statements and connect them to Control Theory, let us consider the example

$$\begin{cases} \dot{x} = Ax + Bv, \quad t \in [0,T], \\ x(0) = x^0, \end{cases} \tag{35}$$

in which the *state* $x = (x_1(t), \dots, x_N(t))^t$ is a vector in $\mathbb{R}^N$ depending on t (the time variable) and the *control* $v = (v_1(t), \dots, v_M(t))^t$ is a vector with M components that also depends on time.

In (35), we will assume that A is a square, constant coefficient matrix of dimension $N \times N$, so that the underlying system is *autonomous*, i.e. invariant with respect to translations in time. The matrix B has also constant coefficients and dimension $N \times M$.

Let us set

$$J(v) = \frac{1}{2}|x(T) - x^1|^2 + \frac{\mu}{2}\int_0^T |v(t)|^2\,dt \quad \forall v \in L^2(0,T;\mathbb{R}^M), \tag{36}$$

where $x^1 \in \mathbb{R}^N$ is given, $x(T)$ is the final value of the solution of (35) and $\mu > 0$.

It is not hard to prove that $J : L^2(0,T;\mathbb{R}^M) \mapsto \mathbb{R}$ is well defined, continuous, coercive and strictly convex. Consequently, J has a unique minimizer in $L^2(0,T;\mathbb{R}^M)$. This shows that the control problem (35)–(36) has a unique solution.

With the DMCV, the existence of minimizers for a large class of problems can be proved. But there are many other interesting problems that do not enter in this simple framework, for which minimizers do not exist.

Indeed, let us consider the simplest and most classical problem in the Calculus of Variations: to show that the shortest path between two given points is the straight line segment.

Of course, it is very easy to show this by means of geometric arguments. However,

> *What happens if we try to use the DMCV ?*

The question is now to minimize the functional

$$\int_0^1 |\dot{x}(t)|\,dt$$

in the class of curves $x : [0,1] \mapsto \mathbb{R}^2$ such that $x(0) = P$ and $x(1) = Q$, where P and Q are two given points in the plane.

The natural functional space for this problem is not a Hilbert space. It can be the Sobolev space $W^{1,1}(0,1)$ constituted by all functions $x = x(t)$ such that x and its time derivative $\dot{x}$ belong to $L^1(0,1)$. It can also be the more sophisticated space $BV(0,1)$ of functions of bounded variation. But these are not Hilbert spaces and solving the problem in any of them, preferably in $BV(0,1)$, becomes much more subtle.

We have described the DMCV in the context of problems without constraints. Indeed, up to now, the functional has been minimized in the whole space. But in most realistic situations the nature of the problem imposes restrictions on the control and/or the state. This is the case for instance for the linear programming problems we have considered above.

As we mentioned above, convexity plays a key role in this context too:

Theorem 2 *Let H be a Hilbert space, $K \subset H$ a closed convex set and $J : K \mapsto \mathbb{R}$ a convex continuous function. Let us also assume that either K is bounded or J is coercive in K, i.e.*

$$J(v) \to +\infty \quad asv \in K, \quad \|v\|_H \to +\infty.$$

Then, there exists a point $u \in K$ where J reaches its minimum over K.

Furthermore, if J is strictly convex, the minimizer is unique.

In order to illustrate this result, let us consider again the system (35) and the functional

$$J(v) = \frac{1}{2}|x(T) - x^1|^2 + \frac{\mu}{2}\int_0^T |v(t)|^2\,dt \quad \forall v \in K, \tag{37}$$

where $\mu \geq 0$ and $K \subset L^2(0,T;\mathbb{R}^M)$ is a closed convex set. In view of Theorem 2, we see that, if $\mu > 0$, the optimal control problem determined by (35) and (37) has a unique solution. If $\mu = 0$ and K is bounded, this problem possesses at least one solution.

Let us discuss more deeply the application of these techniques to the analysis of the control properties of the linear finite dimensional system (35).

Let $J : H \mapsto \mathbb{R}$ be, for instance, a functional of class C^1. Recall again that, at each point u where J reaches its minimum, one has

$$J'(u) = 0, \quad u \in H. \tag{38}$$

It is also true that, when J is convex and C^1, if u solves (38) then u is a global minimizer of J in H. Equation (38) is the *Euler equation* of the corresponding minimization problem.

More generally, in a convex minimization problem, if the function to be minimized is of class C^1, an *Euler inequality* is satisfied by each minimizer. Thus, u is a minimizer of the convex functional J in the convex set K of the Hilbert space H if and only if

$$(J'(u), v - u)_H \geq 0 \quad \forall v \in K, \quad u \in K. \tag{39}$$

Here, $(\cdot\,,\cdot)_H$ stands for the scalar product in H.

In the context of Optimal Control, this characterization of u can be used to deduce the corresponding *optimality conditions*, also called the *optimality system*.

For instance, this can be made in the case of problem (35),(37). Indeed, it is easy to see that in this case (39) reduces to

$$\begin{cases} \mu \int_0^T \langle u(t), v(t) - u(t) \rangle \, dt + \langle x(T) - x^1, z_v(T) - z_u(T) \rangle \geq 0 \\ \forall v \in K, \quad u \in K, \end{cases} \tag{40}$$

where, for each $v \in L^2(0,T;\mathbb{R}^M)$, $z_v = z_v(t)$ is the solution of

$$\begin{cases} \dot{z}_v = Az_v + Bv, \quad t \in [0,T], \\ z_v(0) = 0 \end{cases}$$

(recall that $\langle \cdot \, , \cdot \rangle$ stands for the Euclidean scalar products in $\mathbb{R}^M$ and $\mathbb{R}^N$).

Now, let $p = p(t)$ be the solution of the backward in time differential problem

$$\begin{cases} -\dot{p} = A^t p, & t \in [0,T], \\ p(T) = x(T) - x^1. \end{cases} \tag{41}$$

Then

$$\langle x(T) - x^1, z_v(T) - z_u(T) \rangle = \langle p(T), z_v(T) - z_u(T) \rangle = \int_0^T \langle p(t), B(v(t) - u(t)) \rangle \, dt$$

and (40) can also be written in the form:

$$\begin{cases} \int_0^T \langle \mu u(t) + B^t p(t), v(t) - u(t) \rangle \, dt \geq 0 \\ \forall v \in K, \quad u \in K. \end{cases} \tag{42}$$

The system constituted by the state equation (35) for $v = u$, i.e.

$$\begin{cases} \dot{x} = Ax + Bu, \quad t \in [0,T], \\ x(0) = x^0, \end{cases} \tag{43}$$

the *adjoint state equation* (41) and the inequalities (42) is referred to as the *optimality system.* This system provides, in the case under consideration, a characterization of the optimal control.

The function $p = p(t)$ is the *adjoint state.* As we have seen, the introduction of p leads to a rewriting of (40) that is more explicit and easier to handle.

Very often, when addressing optimization problems, we have to deal with *restrictions* or *constraints* on the controls and/or state. *Lagrange multipliers* then play a fundamental role and are needed in order to write the equations satisfied by the minimizers: the so called *Euler-Lagrange equations.*

To do that, we must introduce the associated *Lagrangian* and, then, we must analyze its *saddle points.* The determination of saddle points leads to two equivalent extremal problems of dual nature.

This is a surprising fact in this theory that can be often used with efficiency: the original minimization problem being difficult to solve, one may often write a *dual minimization problem* (passing through the Lagrangian); it may well happen to the second problem to be simpler than the original one.

Saddle points arise naturally in many optimization problems. But they can also be viewed as the solutions of *minimax problems.* Minimax problems arise in many contexts, for instance:

- In *Differential Game Theory*, where two or more players *compete* trying to maximize their profit and minimize the one of the others.
- In the characterization of the proper vibrations of elastic bodies. Indeed, very often these can be characterized as eigenvalues of a self-adjoint compact operator in a Hilbert space through a minimax principle related to the *Rayleigh quotient.*

One of the most relevant contributions in this field was the one by J. Von Neumann in the middle of the XX Century, proving that the existence of a minimax is guaranteed under very weak conditions.

In the last three decades, these results have been used systematically for solving nonlinear differential problems, in particular with the help of the *Mountain Pass Lemma* (for instance, see [20]). At this respect, it is worth mentioning that a mountain pass is indeed a beautiful example of saddle point provided by Nature. A mountain pass is the location one chooses to cross a mountain chain: this point must be of minimal height along the mountain chain but, on the contrary, it is of maximal height along the crossing path we follow.

The reader interested in learning more about Convex Analysis and the related duality theory is referred to the books [9] and [41], by I. Ekeland and R. Temam and R.T. Rockafellar, respectively. The lecture notes by B. Larrouturou and P.L. Lions [23] contain interesting introductions to these and other related topics, like mathematical modelling, the theory of partial differential equations and numerical approximation techniques.

6 Controllability of linear finite dimensional systems

We will now be concerned with the controllability of ordinary differential equations. We will start by considering linear systems.

As we said above, Control Theory is full of interesting mathematical results that have had a tremendous impact in the world of applications (most of them are too complex to be reproduced in these notes). One of these important results, simple at the same time, is a theorem by R.E. Kalman which characterizes the linear systems that are controllable.

Let us consider again the linear system

$$\left\{ \begin{array}{l} \dot{x} = Ax + Bv, \quad t > 0, \\ x(0) = x^0, \end{array} \right. \tag{44}$$

with state $x = (x_1(t), \ldots, x_N(t))^t$ and control $v = (v_1(t), \ldots, v_M(t))^t$. The matrices A and B have constant coefficients and dimensions $N \times N$ and $N \times M$, respectively.

Assume that $N \geq M \geq 1$. In practice, the cases where M is much smaller than N are especially significant. Of course, the most interesting case is that in which $M = 1$ and, simultaneously, N is very large. We then dispose of a single scalar control to govern the behavior of a very large number N of components of the state.

System (44) is said to be controllable at time $T > 0$ if, for every initial state $x^0 \in \mathbb{R}^N$ and every final state $x^1 \in \mathbb{R}^N$, there exists at least one control $u \in C^0([0,T];\mathbb{R}^M)$ such that the associated solution satisfies

$$x(T) = x^1. \tag{45}$$

The following result, due to Kalman, characterizes the controllability of (44) (see for instance [25]):

Figure 13: Rudolph E. Kalman (1930).

Theorem 3 *A necessary and sufficient condition for system* (44) *to be controllable at some time* $T > 0$ *is that*

$$\text{rank } \left[B \,|\, AB \,|\, \cdots \,|\, A^{N-1}B\right] = N. \tag{46}$$

Moreover, if this is satisfied, the system is controllable for all $T > 0$*.*

When the rank of this matrix is k*, with* $1 \leq k \leq N-1$*, the system is not controllable and, for each* $x^0 \in \mathbb{R}^N$ *and each* $T > 0$*, the set of solutions of* (44) *at time* $T > 0$ *covers an affine subspace of* $\mathbb{R}^N$ *of dimension* k*.*

The following remarks are now in order:

- The degree of controllability of a system like (44) is completely determined by the rank of the corresponding matrix in (46). This rank indicates how many components of the system are sensitive to the action of the control.

- The matrix in (46) is of dimension $(N \times M) \times N$ so that, when we only have one control at our disposal (i.e. $M = 1$), this is a $N \times N$ matrix. It is obviously in this case when it is harder to the rank of this matrix to be N. This is in agreement with common sense, since the system should be easier to control when the number of controllers is larger.

- The system is controllable at some time if and only if it is controllable at any positive time. In some sense, this means that, in (44), *information propagates at infinite speed.* Of course, this property is not true in general in the context of partial differential equations.

As we mentioned above, the concept of *adjoint system* plays an important role in Control Theory. In the present context, the adjoint system of (44) is the following:

$$\left\{ \begin{array}{l} -\dot{\varphi} = A^t\varphi, \quad t < T, \\ \varphi(T) = \varphi^0. \end{array} \right. \tag{47}$$

Let us emphasize the fact that (47) is a backward (in time) system. Indeed, in (47) the sense of time has been reversed and the differential system has been completed with a *final condition* at time $t = T$.

The following result holds:

Theorem 4 *The rank of the matrix in* (46) *is N if and only if, for every $T > 0$, there exists a constant $C(T) > 0$ such that*

$$|\varphi^0|^2 \leq C(T) \int_0^T |B^t\varphi|^2 \, dt \tag{48}$$

for every solution of (47).

The inequality (48) is called an *observability inequality.* It can be viewed as the *dual* version of the controllability property of system (44).

This inequality guarantees that the adjoint system can be "observed" through $B^t\varphi$, which provides M linear combinations of the adjoint state. When (48) is satisfied, we can affirm that, from the controllability viewpoint, B^t captures appropriately all the components of the adjoint state φ. This turns out to be equivalent to the controllability of (44) since, in this case, the control u acts efficiently through the matrix B on all the components of the state x.

Inequalities of this kind play also a central role in *inverse problems*, where the goal is to reconstruct the properties of an unknown (or only partially known) medium or system by means of partial measurements. The observability inequality guarantees that the measurements $B^t\varphi$ are sufficient to detect all the components of the system.

The proof of Theorem 4 is quite simple. Actually, it suffices to write the solutions of (44) and (47) using the *variation of constants formula* and, then, to apply the *Cayley-Hamilton theorem*, that guarantees that any matrix is a root of its own characteristic polynomial.

Thus, to prove that (46) implies (48), it is sufficient to show that, when (46) is true, the mapping

$$\varphi^0 \mapsto \left(\int_0^T |B^t\varphi|^2 \, dt \right)^{1/2}$$

is a norm in $\mathbb{R}^N$. To do that, it suffices to check that the following *uniqueness* or *unique continuation* result holds:

If $B^t\varphi = 0$ for $0 \leq t \leq T$ then, necessarily, $\varphi \equiv 0$.

It is in the proof of this result that the rank condition is needed.

Let us now see how, using (48), we can build controls such that the associated solutions to (44) satisfy (45). This will provide another idea of how controllability and optimal control problems are related.

Given initial and final states x^0 and x^1 and a control time $T > 0$, let us consider the quadratic functional I, with

$$I(\varphi^0) = \frac{1}{2}\int_0^T |B^t\varphi|^2\,dt - \langle x^1, \varphi^0\rangle + \langle x^0, \varphi(0)\rangle \quad \forall \varphi^0 \in \mathbb{R}^N, \tag{49}$$

where φ is the solution of the adjoint system (47) associated to the final state φ^0.

The function $\varphi^0 \mapsto I(\varphi^0)$ is strictly convex and continuous in $\mathbb{R}^N$. In view of (48), it is also coercive, that is,

$$\lim_{|\varphi^0|\to\infty} I(\varphi^0) = +\infty. \tag{50}$$

Therefore, I has a unique minimizer in $\mathbb{R}^N$, that we shall denote by $\hat\varphi^0$. Let us write the Euler equation associated to the minimization of the functional (49):

$$\int_0^T \langle B^t\hat\varphi, B^t\varphi\rangle\,dt - \langle x^1, \varphi^0\rangle + \langle x^0, \varphi(0)\rangle = 0 \quad \forall \varphi^0 \in \mathbb{R}^N, \quad \hat\varphi^0 \in \mathbb{R}^N. \tag{51}$$

Here, $\hat\varphi$ is the solution of the adjoint system (47) associated to the final state $\hat\varphi^0$.

From (51), we deduce that $\hat u = B^t\hat\varphi$ is a control for (44) that guarantees that (45) is satisfied. Indeed, if we denote by $\hat x$ the solution of (44) associated to $\hat u$, we have that

$$\int_0^T \langle B^t\hat\varphi, B^t\varphi\rangle\,dt = \langle \hat x(T), \varphi^0\rangle - \langle x^0, \varphi(0)\rangle \qquad \forall \varphi^0 \in \mathbb{R}^N. \tag{52}$$

Comparing (51) and (52), we see that the previous assertion is true.

It is interesting to observe that, from the rank condition, we can deduce several variants of the observability inequality (48). In particular,

$$|\varphi^0| \le C(T)\int_0^T |B^t\varphi|\,dt \tag{53}$$

This allows us to build controllers of different kinds.

Indeed, consider for instance the functional J_{bb}, given by

$$J_{bb}(\varphi^0) = \frac{1}{2}\left(\int_0^T |B^t\varphi|\,dt\right)^2 - \langle x^1, \varphi^0\rangle + \langle x^0, \varphi(0)\rangle \quad \forall \varphi^0 \in \mathbb{R}^N. \tag{54}$$

This is again strictly convex, continuous and coercive. Thus, it possesses exactly one minimizer $\hat\varphi^0_{bb}$. Let us denote by $\hat\varphi_{bb}$ the solution of the corresponding adjoint system. Arguing as above, it can be seen that the new control $\hat u_{bb}$, with

$$\hat u_{bb} = \left(\int_0^T |B^t\hat\varphi_{bb}|\,dt\right)\operatorname{sgn}(B^t\hat\varphi_{bb}), \tag{55}$$

makes the solution of (44) satisfy (45). This time, we have built a *bang-bang* control, whose components can only take two values:

$$\pm \int_0^T |B^t \hat{\varphi}_{bb}| \, dt.$$

The control $\hat{u}$ that we have obtained minimizing J is the one of minimal norm in $L^2(0,T;\mathbb{R}^M)$ among all controls guaranteeing (45). On the other hand, $\hat{u}_{bb}$ is the control of minimal L^∞ norm. The first one is smooth and the second one is piecewise constant and, therefore, discontinuous in general. However, the bang-bang control is easier to compute and apply since, as we saw explicitly in the case of the pendulum, we only need to determine its amplitude and the location of the switching points. Both controls $\hat{u}$ and $\hat{u}_{bb}$ are optimal with respect to some optimality criterium.

We have seen that, in the context of linear control systems, when controllability holds, the control may be computed by solving a minimization problem. This is also relevant from a computational viewpoint since it provides useful ideas to design efficient approximation methods.

7 Controllability of nonlinear finite dimensional systems

Let us now discuss the controllability of some nonlinear control systems. This is a very complex topic and it would be impossible to describe in a few pages all the significant results in this field. We will just recall some basic ideas.

When the goal is to produce small variations or deformations of the state, it might be sufficient to proceed using linearization arguments. More precisely, let us consider the system

$$\begin{cases} \dot{x} = f(x,u), \quad t > 0, \\ x(0) = x^0, \end{cases} \tag{56}$$

where $f : \mathbb{R}^N \times \mathbb{R}^M \mapsto \mathbb{R}^N$ is smooth and $f(0,0) = 0$. The linearized system at $u = 0$, $x = 0$ is the following:

$$\begin{cases} \dot{x} = \dfrac{\partial f}{\partial x}(0,0)x + \dfrac{\partial f}{\partial u}(0,0)u, \quad t > 0, \\ x(0) = 0. \end{cases} \tag{57}$$

Obviously, (57) is of the form (44), with

$$A = \frac{\partial f}{\partial x}(0,0), \quad B = \frac{\partial f}{\partial u}(0,0), \quad x^0 = 0. \tag{58}$$

Therefore, the rank condition

$$\text{rank } [B|AB|\cdots|A^{N-1}B] = N \tag{59}$$

is the one that guarantees the controllability of (57).

Based on the *inverse function theorem*, it is not difficult to see that, if condition (59) is satisfied, then (56) is *locally controllable* in the following sense:

For every $T > 0$, there exists a neighborhood $\mathcal{B}_T$ of the origin in $\mathbb{R}^N$ such that, for any initial and final states $x_0, x_1 \in \mathcal{B}_T$, there exist controls u such that the associated solutions of the system (56) *satisfy*

$$x(T) = x^1 \,. \tag{60}$$

However, this analysis is not sufficient to obtain results of global nature.

A *natural* condition that can be imposed on the system (56) in order to guarantee global controllability is that, at each point $x^0 \in \mathbb{R}^N$, by choosing all admissible controls $u \in \mathcal{U}_{\rm ad}$, we can recover deformations of the state in all the directions of $\mathbb{R}^N$. But,

> Which are the directions in which the state x can be deformed starting from x^0 ?

Obviously, the state can be deformed in all directions $f(x_0, u)$ with $u \in \mathcal{U}_{\rm ad}$. But these are not all the directions of $\mathbb{R}^N$ when $M < N$. On the other hand, as we have seen in the linear case, there exist situations in which $M < N$ and, at the same time, controllability holds thanks to the rank condition (59).

In the nonlinear framework, the directions in which the state may be deformed around x^0 are actually those belonging to the *Lie algebra* generated by the vector fields $f(x^0, u)$, when u varies in the set of admissible controls $\mathcal{U}_{\rm ad}$. Recall that the Lie algebra $\mathcal{A}$ generated by a family $\mathcal{F}$ of regular vector fields is the set of *Lie brackets* $[f, g]$ with $f, g \in \mathcal{F}$, where

$$[f, g] = (\nabla g)f - (\nabla f)g$$

and all the fields that can be obtained iterating this process of computing Lie brackets.

The following result can be proved (see [46]):

Theorem 5 *Assume that, for each x^0, the Lie algebra generated by $f(x^0, u)$ with $u \in \mathcal{U}_{\rm ad}$ coincides with $\mathbb{R}^N$. Then* (56) *is controllable, i.e. it can be driven from any initial state to any final state in a sufficiently large time.*

The following simple model of driving a car provides a good example to apply these ideas.

Thus, let us consider a state with four components $x = (x_1, x_2, x_3, x_4)$ in which the first two, x_1 and x_2, provide the coordinates of the center of the axis $x_2 = 0$ of the vehicle, the third one, $x_3 = \varphi$, is the counterclockwise angle of the car with respect to the half axis $x_1 > 0$ and the fourth one, $x_4 = \theta$, is the angle of the front wheels with respect to the axis of the car. For simplicity, we will assume that the distance from the front to the rear wheels is $\ell = 1$.

The front wheels are then parallel to the vector $(\cos(\theta + \varphi), \sin(\theta + \varphi))$, so that the instantaneous velocity of the center of the front axis is parallel to this vector. Accordingly,

$$\frac{d}{dt}\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = u_2(t)\begin{pmatrix} \cos(\theta + \varphi) \\ \sin(\theta + \varphi) \end{pmatrix}$$

for some scalar function $u_2 = u_2(t)$.

The center of the rear axis is the point $(x_1 - \cos\varphi,\, x_2 - \sin\varphi)$. The velocity of this point has to be parallel to the orientation of the rear wheels $(\cos\varphi,\, \sin\varphi)$, so that

$$(\sin\varphi)\,\frac{d}{dt}(x_1 - \cos\varphi) - (\cos\varphi)\,\frac{d}{dt}(x_2 - \sin\varphi) = 0.$$

In this way, we deduce that

$$\dot\varphi = u_2 \sin\theta.$$

On the other hand, we set

$$\dot\theta = u_1$$

and this reflects the fact that the velocity at which the angle of the wheels varies is the second variable that we can control. We obtain the following reversible system:

$$\dot x = u_1(t)\begin{pmatrix} 0 \\ 0 \\ 0 \\ 1 \end{pmatrix} + u_2(t)\begin{pmatrix} \cos(\varphi+\theta) \\ \sin(\varphi+\theta) \\ \sin\theta \\ 0 \end{pmatrix}. \tag{61}$$

According to the previous analysis, in order to guarantee the controllability of (61), it is sufficient to check that the Lie algebra of the directions in which the control may be deformed coincides with $\mathbb{R}^4$ at each point.

With $(u_1, u_2) = (0, 1)$ and $(u_1, u_2) = (1, 0)$, we obtain the directions

$$\begin{pmatrix} \cos(\varphi+\theta) \\ \sin(\varphi+\theta) \\ \sin\theta \\ 0 \end{pmatrix} \text{ and } \begin{pmatrix} 0 \\ 0 \\ 0 \\ 1 \end{pmatrix}, \tag{62}$$

respectively. The corresponding Lie bracket provides the direction

$$\begin{pmatrix} -\sin(\varphi+\theta) \\ \cos(\varphi+\theta) \\ \cos\theta \\ 0 \end{pmatrix}, \tag{63}$$

whose Lie bracket with the first one in (62) provides the new direction

$$\begin{pmatrix} -\sin\varphi \\ \cos\varphi \\ 0 \\ 0 \end{pmatrix}. \tag{64}$$

Taking into account that the determinant of the matrix formed by the four column vectors in (62), (63) and (64) is identically equal to 1, we deduce that, at each point, the set of directions in which the state may be deformed is the whole $\mathbb{R}^4$.

Thus, system (61) is controllable.

It is an interesting exercise to think on how one uses in practice the four vectors (62) – (64) to park a car. The reader interested in getting more deeply into this subject may consult the book by E. Sontag [46].

The analysis of the controllability of systems governed by partial differential equations has been the objective of a very intensive research the last decades. However, the subject is older than that.

In 1978, D.L. Russell [42] made a rather complete survey of the most relevant results that were available in the literature at that time. In that paper, the author described a number of different tools that were developed to address controllability problems, often inspired and related to other subjects concerning partial differential equations: multipliers, moment problems, nonharmonic Fourier series, etc. More recently, J.L. Lions introduced the so called *Hilbert Uniqueness Method* (H.U.M.; for instance, see [29],[30]) and this was the starting point of a fruitful period on the subject.

In this context, which is the usual for modelling problems from Continuum Mechanics, one needs to deal with infinite dimensional dynamical systems and this introduces a lot of nontrivial difficulties to the theory and raises many relevant and mathematically interesting questions. Furthermore, the solvability of the problem depends very much on the nature of the precise question under consideration and, in particular, the following features may play a crucial role: linearity or nonlinearity of the system, time reversibility, the structure of the set of admissible controls, etc.

For more details, the reader is referred to the books [21] and [27] and the survey papers [13], [54] and [55].

8 Control, complexity and numerical simulation

Real life systems are genuinely complex. *Internet*, the large quantity of components entering in the fabrication of a car or the decoding of human genoma are good examples of this fact.

The algebraic system (23) considered in Section 5 is of course academic but it suffices by itself to show that not all the components of the state are always sensitive to the chosen control. One can easily imagine how dramatic can the situation be when dealing with complex (industrial) systems. Indeed, determining whether a given controller allows to act on all the components of a system may be a very difficult task.

But complexity does not only arise for systems in Technology and Industry. It is also present in Nature. At this respect, it is worth recalling the following anecdote. In 1526, the Spanish King "Alfonso X El Sabio" got into the *Alcázar of Segovia* after a violent storm and exclaimed:

> "If God had consulted me when He was creating the world, I would have recommended a simpler system."

Recently we have learned about a great news, a historical achievement of Science: the complete decoding of human *genoma*. The genoma code is a good proof of the complexity which is intrinsic to life. And, however, one has not to forget that, although the decoding has been achieved, there will still be a lot to do before being able to use efficiently all this information for medical purposes.

Complexity is also closely related to numerical simulation. In practice, any efficient control strategy, in order to be implemented, has to go through numerical simulation. This requires discretizing the control system, which very often increases its already high complexity.

The recent advances produced in Informatics allow nowadays to use numerical simulation at any step of an industrial project: conception, development and qualification. This relative success of numerical methods in Engineering versus other traditional methods relies on the facts that the associated experimental costs are considerably lower and, also, that numerical simulation allows testing at the realistic scale, without the technical restrictions motivated by instrumentation.

This new scientific method, based on a combination of Mathematics and Informatics, is being seriously consolidated. Other Sciences are also closely involved in this melting, since many mathematical models stem from them: Mechanics, Physics, Chemistry, Biology, Economics, etc. Thus, we are now able to solve more sophisticated problems than before and the complexity of the systems we will be able to solve in the near future will keep increasing. Thanks in particular to parallelization techniques, the description and numerical simulation of complex systems in an acceptable time is more and more feasible.

However, this panorama leads to significant and challenging difficulties that we are now going to discuss.

The first one is that, in practice, the systems under consideration are in fact the coupling of several complex subsystems. Each of them has its own dynamics but the coupling may produce new and unexpected phenomena due to their interaction.

An example of this situation is found in the mathematical description of reactive fluids which are used, for instance, for the propulsion of spatial vehicles. For these systems, one has to perform a modular analysis, separating and simulating numerically each single element and, then, assembling the results. But this is a major task and much has still to be done.

There are many relevant examples of complex systems for which coupling can be the origin of important difficulties. In the context of *Aerospatial Technology*, besides the combustion of reactive fluids, we find fluid-structure interactions which are extremely important when driving the craft, because of the vibrations originated by combustion. Other significant examples are weather prediction and Climatology, where the interactions of atmosphere, ocean, earth, etc. play a crucial role. A more detailed description of the present situation of research and perspectives at this respect can be found in the paper [1], by J. Achache and A. Bensoussan.

In our context, the following must be taken into account:

To understand the level of difficulty, it is sufficient to consider a hybrid parabolic-hyperbolic system and try to match the numerical methods obtained with a finite difference method in the hyperbolic component and a finite element method in the parabolic one.

- Only complex systems are actually relevant from the viewpoint of applications.
- Furthermore, in order to solve a relevant problem, we must first identify the various subsystems and the way they interact.

Let us now indicate some of the mathematical techniques that have been recently developed (and to some extent re-visited) to deal with complexity and perform the appropriate decomposition of large systems that we have mentioned as a need:

- **The solution of linear systems.**

  When the linear system we have to solve presents a block-sparse structure, it is convenient to apply methods combining appropriately the *local* solution of the subsystems corresponding to the individual blocks. This is a frequent situation when dealing with finite difference or finite element discretizations of a differential system.

  The most usual way to proceed is to introduce *preconditioners*, determined by the solutions to the subsystems, each of them being computed with one processor and, then, to perform the global solution with parallelized iterative methods.

- **Multigrid methods.**

  These are very popular today. Assume we are considering a linear system originated by the discretization of a differential equation. The main idea of a multigrid method is to "separate" the low and the high frequencies of the solution in the computation procedure. Thus we compute approximations of the solution at different levels, for instance working alternately with a coarse and a fine grid and incorporating adequate coupling mechanisms.

  The underlying reason is that any grid, even if it is very fine, is unable to capture sufficiently high frequency oscillations, just as an ordinary watch is unable to measure microseconds.

- **Domain decomposition methods.**

  Now, assume that (1) is a boundary value problem for a partial differential equation in the N-dimensional domain Ω. If Ω has a complex geometrical structure, it is very natural to decompose (1) in several similar systems written in simpler domains.

  This can be achieved with domain decomposition techniques. The main idea is to split $\overline{\Omega}$ in the form

  $$\overline{\Omega} = \overline{\Omega_1} \cup \overline{\Omega_2} \cup \cdots \cup \overline{\Omega_m}\,, \tag{65}$$

  and to introduce then an iterative scheme based on computations on each Ω_i separately.

Actually, this is not new. Some seminal ideas can be found in Volume II of the book [7] by R. Courant and D. Hilbert. Since then, there have been lots of works on domain decomposition methods applied to partial differential systems (see for instance [24]). However, the role of these methods in the solution of control problems has not still been analyzed completely.

- **Alternating direction methods.**

  Frequently, we have to consider models involving time-dependent partial differential equations in several space dimensions. After standard time discretization, one is led at each time step to a set of (stationary) partial differential problems whose solution, in many cases, is difficult to achieve.

  This is again connected to the need of decomposing complex systems in more simple subsystems. These ideas lead to the methods of alternating directions, of great use in this context. A complete analysis can be found in [51]. In the particular, but very illustrating context of the Navier-Stokes equations, these methods have been described for instance in [19] and [38].

  However, from the viewpoint of Control Theory, alternating direction methods have not been, up to now, sufficiently explored.

The interaction of the various components of a complex system is also a difficulty of major importance in control problems. As we mentioned above, for real life control problems, we have first to choose an appropriate model and then we have also to make a choice of the control property. But necessarily one ends up introducing numerical discretization algorithms to make all this computable. Essentially, we will have to be able to compute an accurate approximation of the control and this will be made only if we solve numerically a *discretized control problem.*

At this point, let us observe that, as mentioned in [53], some models obtained after discretization (for instance via the finite element method) are not only relevant regarded as approximations of the underlying continuous models but also by themselves, as genuine models of the real physical world.

Let us consider a simple example in order to illustrate some extra, somehow unexpected, difficulties that the discretization may bring to the control process.

Consider again the state equation (1). To fix ideas, we will assume that our control problem is as follows

To find $u \in \mathcal{U}_{\rm ad}$ *such that*

$$\Phi(u, y(u)) \leq \Phi(v, y(v)) \quad \forall v \in \mathcal{U}_{\rm ad}\,, \tag{66}$$

where $\Phi = \Phi(v, y)$ *is a given function.*

The reader can find in [53] details on how the finite element method was born around 1960. In this article it is also explained that, since its origins, finite elements have been viewed as a tool to build legitimate discrete models for the mechanical systems arising in Nature and Engineering, as well as a method to approximate partial differential systems.

Then, we are led to the following crucial question:

What is an appropriate discretized control problem ?

There are at least two reasonable possible answers:

- **First approximation method.**

  We first discretize $\mathcal{U}_{\rm ad}$ and (1) and obtain $\mathcal{U}_{{\rm ad},h}$ and the new (discrete) state equation
  $$A_h(y_h) = f(v_h). \tag{67}$$

  Here, h stands for a small parameter that measures the *characteristic size* of the "numerical mesh". Later, we let $h \to 0$ to make the discrete problem converge to the continuous one. If $\mathcal{U}_{{\rm ad},h}$ and (67) are introduced the right way, we can expect to obtain a "discrete state" $y_h(v_h)$ for each "discrete admissible" control $v_h \in \mathcal{U}_{{\rm ad},h}$.

  Then, we search for an optimal control at the discrete level, i.e. a control $u_h \in \mathcal{U}_{{\rm ad},h}$ such that

  $$\Phi(u_h, y_h(u_h)) \leq \Phi(v_h, y_h(v_h)) \quad \forall v_h \in \mathcal{U}_{{\rm ad},h} \, . \tag{68}$$

  This corresponds to the following scheme:

  $$\text{MODEL} \longrightarrow \text{DISCRETIZATION} \longrightarrow \text{CONTROL.}$$

  Indeed, starting from the continuous control problem, we first discretize it and we then compute the control of the discretized model. This provides a first natural method for solving in practice the control problem.

- **Second approximation method.**

  However, we can also do as follows. We analyze the original control problem (1),(66) and we characterize the optimal solution and control in terms of an *optimality system*. We have already seen that, in practice, this is just to write the Euler or Euler-Lagrange equations associated to the minimization problem we are dealing with. We have already described how optimality systems can be found for some particular control problems.

  The optimality systems are of the form

  $$A(y) = f(u), \quad B(y)p = g(u, y) \tag{69}$$

  (where $B(y)$ is a linear operator), together with an additional equation relating u, y and p. To simplify our exposition, let us assume that the latter can be written in the form

  $$Q(u, y, p) = 0 \tag{70}$$

for some mapping Q. The key point is that, if u, y and p solve the optimality system (69) – (70), then u is an optimal control and y is the associate state. Of course, p is the *adjoint state* associated to u and y.

Then, we can discretize and solve numerically (69),(70). This corresponds to a different approach:

$$\text{MODEL} \longrightarrow \text{CONTROL} \longrightarrow \text{DISCRETIZATION}.$$

Notice that, in this second approach, we have interchanged the control and discretization steps. Now, we first analyze the continuous control problem and, only later, we proceed to the numerical discretization.

It is not always true that these two methods provide the same results.

For example, it is shown in [18] that, with a finite element approximation, the first one may give erroneous results in vibration problems. This is connected to the lack of accuracy of finite elements in the computation of high frequency solutions to the wave equation, see [53].

On the other hand, it has been observed that, for the solution of a lot of *optimal design problems*, the first strategy is preferable; see for instance [34] and [37].

The commutativity of the DISCRETIZATION/CONTROL scheme is at present a subject that is not well understood and requires further investigation. We do not still have a significant set of results allowing to determine when these two approaches provide similar results and when they do not. Certainly, the answer depends heavily on the nature of the model under consideration. In this sense, control problems for elliptic and parabolic partial differential equations, because of their intrinsic dissipative feature, will be better behaved than hyperbolic systems. We refer the interested reader to [56] for a complete account of this fact. It is however expected that much progress will be made in this context in the near future.

9 Two challenging applications

In this Section, we will mention two control problems whose solution will probably play an important role in the context of applications in the near future.

9.1 Molecular control via laser technology

We have already said that there are many technological contexts where Control Theory plays a crucial role. One of them, which has had a very recent development and announces very promising perspectives, is the *laser control of chemical reactions.*

Nevertheless, the disagreement of these two methods may be relevant not only as a purely numerical phenomenon but also at the level of modelling since, as we said above, in many engineering applications discrete models are often directly chosen.

The basic principles used for the control of industrial processes in Chemistry have traditionally been the same for many years. Essentially, the strategies have been (a) to introduce changes in the temperature or pressure in the reactions and (b) to use *catalyzers.*

Laser technology, developed in the last four decades, is now playing an increasingly important role in molecular design. Indeed, the basic principles in *Quantum Mechanics* rely on the wave nature of both light and matter. Accordingly, it is reasonable to believe that the use of laser will be an efficient mechanism for the control of chemical reactions.

The experimental results we have at our disposal at present allow us to expect that this approach will reach high levels of precision in the near future. However, there are still many important technical difficulties to overcome.

For instance, one of the greatest drawbacks is found when the molecules are "not very isolated". In this case, collisions make it difficult to define their phases and, as a consequence, it is very hard to choose an appropriate choice of the control. A second limitation, of a much more technological nature, is related to the design of lasers with well defined phases, not too sensitive to the instabilities of instruments.

For more details on the modelling and technological aspects, the reader is referred to the expository paper [4] by P. Brumer and M. Shapiro.

The goal of this subsection is to provide a brief introduction to the mathematical problems one finds when addressing the control of chemical reactions.

Laser control is a subject of high interest where Mathematics are not sufficiently developed. The models needed to describe these phenomena lead to complex (nonlinear) *Schrödinger equations* for which the results we are able to deduce are really poor at present. Thus,

- We do not dispose at this moment of a complete theory for the corresponding initial or iniial/boundary value problems.
- Standard numerical methods are not sufficiently efficient and, accordingly, it is difficult to test the accuracy of the models that are by now available.

The control problems arising in this context are *bilinear.* This adds fundamental difficulties from a mathematical viewpoint and makes these problems extremely challenging. Indeed, we find here genuine nonlinear problems for which, apparently, the existing linear theory is insufficient to provide an answer in a first approach.

In fact, it suffices to analyze the most simple bilinear control problems where wave phenomena appear to understand the complexity of this topic. Thus, let us illustrate this situation with a model concerning the linear one-dimensional Schrödinger equation. It is clear that this is insufficient by itself to describe all the complex phenomena arising in molecular control via laser technology. But it suffices to present the main mathematical problem and difficulties arising in this context.

The system is the following:

$$\left\{\begin{array}{lll} i\phi_t + \phi_{xx} + p(t)x\,\phi = 0 & 0 < x < 1, & 0 < t < T, \\ \phi(0,t) = \phi(1,t) = 0, & 0 < t < T, & \\ \phi(x,0) = \phi^0(x), & 0 < x < 1. & \end{array}\right. \tag{71}$$

In (71), $\phi = \phi(x,t)$ is the *state* and $p = p(t)$ is the control. Although ϕ is complex-valued, $p(t)$ is real for all t. The control p can be interpreted as the intensity of an applied electrical field and x is the (prescribed) direction of the laser.

The state $\phi = \phi(x,t)$ is the wave function of the molecular system. It can be regarded as a function that furnishes information on the location of an elementary particle: for arbitrary a and b with $0 \le a < b \le 1$, the quantity

$$P(a,b;t) = \int_a^b |\phi(x,t)|^2\,dx$$

can be viewed as the probability that the particle is located in (a,b) at time t.

The controllability problem for (71) is to find the set of attainable states $\phi(\cdot,T)$ at a final time T as p runs over the whole space $L^2(0,T)$.

It is worth mentioning that, contrarily to what happens to many other control problems, the set of attainable states at time T depends strongly on the initial data ϕ^0. In particular, when $\phi^0 = 0$ the unique solution of (71) is $\phi \equiv 0$ whatever p is and, therefore, the unique attainable state is $\phi(\cdot,T) \equiv 0$. It is thus clear that, if we want to consider a nontrivial situation, we must suppose that $\phi^0 \neq 0$.

We say that this is a *bilinear control problem*, since the unique nonlinearity in the model is the term $p(t)x\,\phi$, which is essentially the product of the control and the state. Although the nonlinearity might seem simple, this control problem becomes rather complex and out of the scope of the existing methods in the literature.

For an overview on the present state of the art of the control of systems governed by the Schrödinger equation, we refer to the survey article [57] and the references therein.

9.2 An environmental control problem

For those who live and work on the seaside or next to a river, the relevance of being able to predict drastic changes of weather or on the state of the sea is obvious. In particular, it is vital to predict whether flooding may arise, in order to be prepared in time.

Floodings are one of the most common environmental catastrophic events and cause regularly important damages in several regions of our planet. They are produced as the consequence of very complex interactions of tides, waves and storms. The varying wind and the fluctuations of the atmospherical pressure produced by a storm can be the origin of an elevation or descent of several meters of the sea level in a time period that can change from several hours to two or three days. The wind can cause waves of a period of 20 seconds and a wavelenght of 20 or 30 meters. The simultaneous combination of these two phenomena leads to a great risk of destruction and flooding.

The amplitude of the disaster depends frequently on the possible accumulation of factors or events with high tides. Indeed, when this exceptional elevation of water occurs during a high tide, the risk of flooding increases dangerously.

This problem is being considered increasingly as a priority by the authorities of many cities and countries. Indeed, the increase of temperature of the planet and the melting of polar ice are making these issues more and more relevant for an increasing population in all the continents.

For instance, it is well known that, since the Middle Age, regular floods in the Thames river cover important pieces of land in the city of London and cause tremendous damages to buildings and population.

When floods occur in the Thames river, the increase on the level of water can reach a height of 2 meters. On the other hand, the average level of water at the London bridge increases at a rate of about 75 centimeters per century due to melting of polar ice. Obviously, this makes the problem increasingly dangerous.

Before explaining how the British authorities have handled this problem, it is important to analyze the process that lead to these important floods.

It is rather complex. Indeed, low atmospheric pressures on the Canadian coast may produce an increase of about 30 centimeters in the average sea level in an area of about 1 600 square kilometers approximately. On the other hand, due to the north wind and ocean currents, this tremendous mass of water may move across the Atlantic Ocean at a velocity of about 80 to 90 kilometers per day to reach the coast of Great Britain. Occasionally, the north wind may even push this mass of water down along the coast of England to reach the Thames Estuary. Then, this mass of water is sent back along the Thames and the conditions for a disaster arise.

In 1953, a tremendous flooding happened killing over 300 people while 64 000 hectares of land were covered by water. After that, the British Government decided to create a Committee to analyze the problem and the possibilities of building defense mechanisms. There was consensus on the Committee about the need of some defense mechanism but not about which one should be implemented. Finally, in 1970 the decision of building a barrier, the *Thames Barrier*, was taken.

Figure 14: The Thames Barrier.

Obviously, the main goal of the barrier is to close the river when a dangerous increase of water level is detected. The barrier was built during 8 years and 4 000 workers participated on that gigantic engineering programme. The barrier was finally opened in 1984. It consists of 10 enormous steel gates built over the basement of reinforced concrete structures and endowed with sophisticated mechanisms that allow normal traffic on the river when the barrier is open but that allows closing and cutting the traffic and the flux of water when needed. Since its opening, the barrier has been closed three times up to now.

Obviously, as for other many control mechanisms, it is a priority to close the barrier a minimal number of times. Every time the barrier is closed, important economic losses are produced due to the suppression of river traffic. Furthermore, once the barrier is closed, it has to remain closed at least for 8 hours until the water level stabilizes at both sides. On the other hand, the process of closing the barrier takes two hours and, therefore, it is not possible to wait and see at place the flood arriving but, rather, one has to take the decision of closing on the basis of *predictions*. Consequently, extremely efficient methods of prediction are needed.

At present, the predictions are made by means of mathematical models that combine or match two different subsystems: the first one concerns the tides around the British Islands and the second one deals with weather prediction. In this way, every hour, predictions are made 30 hours ahead on several selected points of the coast.

The numerical simulation and solution of this model is performed on the supercomputer of the British Meteorological Office and the results are transferred to the computer of the Thames Barrier. The data are then introduced in another model, at a bigger scale, including the North Sea, the Thames Estuary and the low part of the river where the effect of tides is important. The models that are being used at present reduce to systems of partial differential equations and are solved by finite difference methods. The results obtained this way are compared to the average predictions and, in view of this analysis, the authorities have the responsibility of taking the decision of closing the barrier or keeping it opened.

The Thames Barrier provides, at present, a satisfactory solution to the problem of flooding in the London area. But this is not a long term solution since, as we said above, the average water level increases of approximately 75 centimeters per century and, consequently, in the future, this method of prevention will not suffice anymore.

We have mentioned here the main task that the Thames Barrier carries out: the prevention of flooding. But it also serves of course to prevent the water level to go down beyond some limits that put in danger the traffic along the river.

The Thames Barrier is surely one of the greatest achievements of Control Theory in the context of the environmental protection. Here, the combination of mathematical modelling, numerical simulation and Engineering has allowed to provide a satisfactory solution to an environmental problem of first magnitude.

The reader interested in learning more about the Thames Barrier is referred to [14].

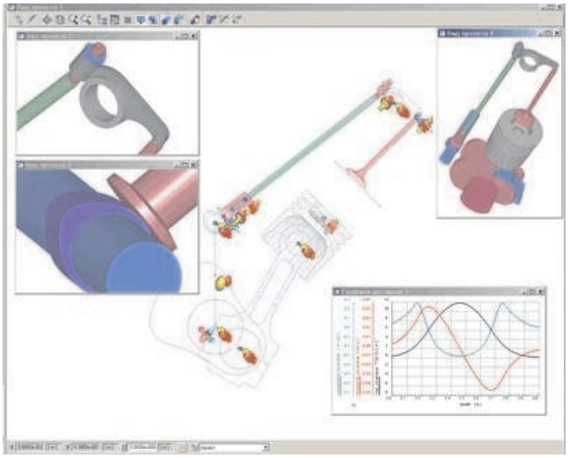


Figure 15: Design of a combustion controller valve (this Figure has been taken from `http://www.euler.ru/content.asp?doc=1702`).

10 The future

At present, there are many branches of Science and Technology in which Control Theory plays a central role and faces fascinating challenges. In some cases, one expects to solve the problems by means of technological developments that will make possible to implement more sophisticated control mechanisms. To some extent, this is the case for instance of the laser control of chemical reactions we have discussed above. But, in many other areas, important theoretical developments will also be required to solve the complex control problems that arise. In this Section, we will briefly mention some of the fields in which these challenges are present. The reader interested in learning more about these topics is referred to the SIAM Report [44].

• **Large space structures -** Quite frequently, we learn about the difficulties found while deploying an antenna by a satellite, or on getting the precise orientation of a telescope. In some cases, this may cause huge losses and damages and may even be a reason to render the whole space mission useless. The importance of space structures is increasing rapidly, both for communications and research within our planet and also in the space adventure. These structures are built coupling several components, rigid and flexible ones. The problem of stabilizing these structures so that they remain oriented in the right direction without too large deformations is therefore complex and relevant. Designing robust control mechanisms for these structures is a challenging problem that requires important cooperative developments in Control Theory, computational issues and Engineering.

• **Robotics -** This is a branch of Technology of primary importance, where the scientific challenges are diverse and numerous. These include, for instance, computer vision. Control Theory is also at the heart of this area and its development relies to a large extent on robust computational algorithms for controlling. It is not hard to imagine how difficult it is to get a robot "walking" along a stable dynamics or catching an objet with its "hands" (see other related comments and Figures in Section 4).

• **Information and energy networks -** The globalization of our planet is an irreversible process. This is valid in an increasing number of human activities as air traffic, generation and distribution of energy, informatic networks, etc. The dimensions and complexity of the networks one has to manage are so large that, very often, one has to take decisions locally, without having a complete global information, but taking into account that local decisions will have global effects. Therefore, there is a tremendous need of developing methods and techniques for the control of large interconnected systems.

• **Control of combustion -** This is an extremely important problem in Aerospatial and Aeronautical Industry. Indeed, the control of the instabilities that combustion produces is a great challenge. In the past, the emphasis has been put on design aspects, modifying the geometry of the system to interfere on the acoustic-combustion interaction or incorporating dissipative elements. The active control of combustion by means of thermal or acoustic mechanisms is also a subject in which almost everything is to be done.

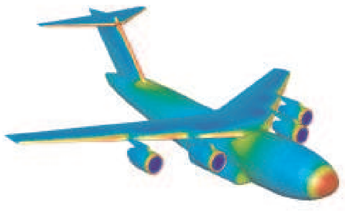


Figure 16: Numerical approximation of the pressure distribution on the surface of an aircraft.

• **Control of fluids -** The interaction between Control Theory and *Fluid Mechanics* is also very rich nowadays. This is an important topic in *Aeronautics*, for instance, since the structural dynamics of a plane in flight interacts with the flux of the neighboring air. In conventional planes, this fact can be ignored but, for the new generations, it will have to be taken into account, to avoid turbulent flow around the wings.

To get an idea of the complexity of the problem, see for instance Fig. 16, taken from `http://www.mems.rice.edu/TAFSM/PROJ/AS/cargo_pl.html`, where a numerical approximation of the pressure distribution on the surface of an aircraft is displayed.

From a mathematical point of view, almost everything remains to be done in what concerns modelling, computational and control issues. A crucial contribution was made by J.L. Lions in [31], where the approximate controllability of the Navier-Stokes equations was conjectured. For an overview of the main existing results, see [12].

• **Solidification processes and steel industry -** The increasingly important development in *Material Sciences* has produced intensive research in solidification processes. The form and the stability of the liquid-solid

Figure 17: An aerodynamic obstacle: a Delta Wing.

interface are central aspects of this field, since an irregular interface may produce undesired products. The sources of instabilities can be of different nature: convection, surface tension, ... The *Free Boundary Problems* area has experienced important developments in the near past, but very little has been done from a control theoretical viewpoint. There are very interesting problems like, for instance, *building interfaces* by various indirect measurements, or its control by means of heating mechanisms, or applying electric or magnetic currents or rotations of the alloy in the furnace. Essentially, there is no mathematical theory to address these problems.

• **Control of plasma -** In order to solve the energetic needs of our planet, one of the main projects is the obtention of fusion reactions under control. At present, *Tokomak machines* provide one of the most promising approaches to this problem. Plasma is confined in a Tokomak machine by means of electromagnetic fields. The main problem consists then in keeping the plasma at high density and temperature on a desired configuration along long time intervals despite its instabilities. This may be done placing *sensors* that provide the information one needs to modify the currents rapidly to compensate the perturbations in the plasma. Still today there is a lot to be done in this area. There are also important identification problems arising due to the difficulties to get precise measurements. Therefore, this is a field that provides many challenging topics in the areas of Control Theory and *Inverse Problems Theory.*

• **Biomedical research -** The design of medical therapies depends very strongly on the understanding of the dynamics of Physiology. This is a very active topic nowadays in which almost everything is still to be done from a mathematical viewpoint. Control Theory will also play an important role in this field. As an example, we can mention the design of mechanisms for insulin supply endowed with control chips.

• **Hydrology -** The problem of governing water resources is extremely relevant nowadays. Sometimes this is because there are little resources, some others because they are polluted and, in general, because of the complexity of the network of supply to all consumers (domestic, agricultural, industrial, ...). The control problems arising in this context are also of different nature.

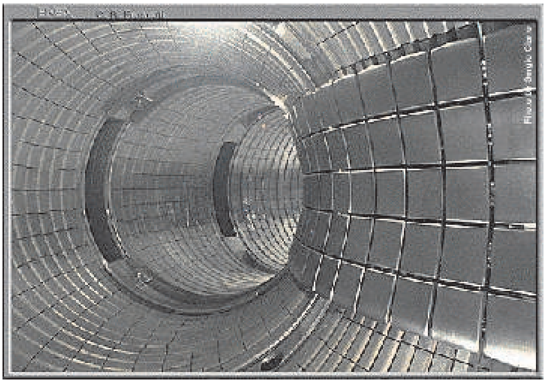


Figure 18: A Tokamak machine (this and the next Figure have been taken from `http://www-phys.llnl.gov/Research/Tokamak/`).

For instance, the *parameter identification problem*, in which the goal is to determine the location of sensors that provide sufficient information for an efficient extraction and supply and, on the other hand, the design of efficient management strategies.

• **Recovery of natural resources -** Important efforts are being made on the modelling and theoretical and numerical analysis in the area of simulation of reservoirs of water, oil, minerals, etc. One of the main goals is to optimize the extraction strategies. Again, inverse problems arise and, also, issues related to the control of the interface between the injected and the extracted fluid.

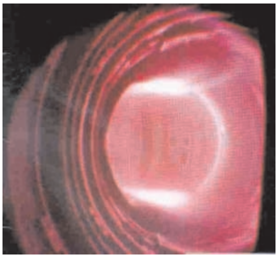


Figure 19: The plasma in a Tokamak machine.

• **Economics -** The increasingly important role that Mathematics are playing in the world of *Economics* and *Finances* is well known. Indeed, nowadays, it is very frequent to use Mathematics to predict the fluctuations in financial markets. The models are frequently stochastic and the existing *Stochastic Control Theory* may be of great help to design optimal strategies of investment and consumption.

• **Manufacturing systems -** Large automatic manufacturing systems are designed as flexible systems that allow rapid changes of the production planning as a function of demand. But this increasing flexibility is obtained at the price of an increasing complexity. In this context, Control Theory faces also the need of designing efficient computerized control systems.

• **Evaluation of efficiency on computerized systems -** The existing software packages to evaluate the efficiency of computer systems are based on its representation by means of the *Theory of Networks.* The development of parallel and synchronized computer systems makes them insufficient. Thus, it is necessary to develop new models and, at this level, the Stochastic Control Theory of *discrete systems* may play an important role.

- **Control of computer aided systems -** As we mentioned above, the complexity of the control problems we are facing nowadays is extremely high. Therefore, it is impossible to design efficient control strategies without the aid of computers and this has to be taken into account when designing these strategies. This is a multidisciplinary research field concerned with Control Theory, Computer Sciences, Numerical Analysis and Optimization, among other areas.

Appendix 1: Pontryagin's maximum principle

As we said in Section 3, one of the main contributions to Control Theory in the sixties was made by L. Pontryagin by means of *the maximum principle.* In this Appendix, we shall briefly recall the main underlying ideas.

In order to clarify the situation and show how powerful is this approach, we will consider a *minimal time control* problem. Thus, let us consider again the differential system

$$\begin{cases} \dot{x} = f(x,u), \quad t > 0, \\ x(0) = x^0, \end{cases} \tag{72}$$

with state $x = (x_1(t), \dots, x_N(t))$ and control $u = (u_1(t), \dots, u_M(t))$.

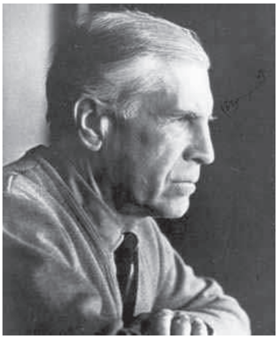


Figure 20: Lev S. Pontryagin (1908–1988).

For simplicity, we will assume that the function $f : \mathbb{R}^N \times \mathbb{R}^M \mapsto \mathbb{R}^N$ is well defined and smooth, although this is not strictly necessary (actually, this is one of the main contributions of Pontryagin's principle). We will also assume that a nonempty closed set $G \subset \mathbb{R}^M$ is given and that the family of admissible controls is

$$\mathcal{U}_{\rm ad} = \{\, u \in L^2(0,+\infty;\mathbb{R}^M) : u(t) \in G \ \ \text{a.e.} \,\}. \tag{73}$$

Let us introduce a manifold $\mathcal{M}$ of $\mathbb{R}^N$, with

$$\mathcal{M} = \{\, x \in \mathbb{R}^N : \mu(x) = 0 \,\},$$

where $\mu : \mathbb{R}^N \mapsto \mathbb{R}^q$ is a regular map ($q \le N$), so that the matrix $\nabla\mu(x)$ is of rank q at each point $x \in \mathcal{M}$ (thus, $\mathcal{M}$ is a smooth differential manifold of dimension $N - q$). Recall that the tangent space to $\mathcal{M}$ at a point $x \in \mathcal{M}$ is given by:

$$T_x\mathcal{M} = \{\, v \in \mathbb{R}^N : \nabla\mu(x) \cdot v = 0 \,\}.$$

Let us fix the initial state x^0 in $\mathbb{R}^N \setminus \mathcal{M}$. Then, to each control $u = u(t)$ we can associate a trajectory, defined by the solution of (72). Our minimal time control problem consists in finding a control in the admissible set $\mathcal{U}_{\rm ad}$ driving the corresponding trajectory to the manifold $\mathcal{M}$ in a time as short as possible.

In other words, we intend to minimize the quantity T subject to the following constraints:

- $T > 0$,
- For some $u \in \mathcal{U}_{\rm ad}$, the associated solution to (72) satisfies $x(T) \in \mathcal{M}$.

Obviously, the difficulty of the problem increases when the dimension of $\mathcal{M}$ decreases.

The following result holds (Pontryagin's maximum principle):

Theorem 6 *Assume that $\hat{T}$ is the minimal time and $\hat{u}$, defined for $t \in [0, \hat{T}]$, is an optimal control for this problem. Let $\hat{x}$ be the corresponding trajectory. Then there exists $\hat{p} = \hat{p}(t)$ such that the following identities hold almost everywhere in $[0, \hat{T}]$:*

$$\dot{\hat{x}} = f(\hat{x}, \hat{u}), \quad -\dot{\hat{p}} = \left(\frac{\partial f}{\partial x}(\hat{x}, \hat{u}) \right)^t \cdot \hat{p} \tag{74}$$

and

$$H(\hat{x}(t), \hat{p}(t), \hat{u}) = \max_{v \in G} H(\hat{x}(t), \hat{p}(t), v), \tag{75}$$

where

$$H(x, p, v) = \langle f(x, v), p \rangle \quad \forall (x, p, v) \in \mathbb{R}^N \times \mathbb{R}^N \times G. \tag{76}$$

Furthermore, the quantity

$$H^*(\hat{x}, \hat{p}) = \max_{v \in G} H(\hat{x}, \hat{p}, v) \tag{77}$$

is constant and nonnegative (maximum condition) and we have

$$\hat{x}(\hat{T}) = x^0, \quad \hat{x}(\hat{T}) \in \mathcal{M} \tag{78}$$

and

$$\hat{p}(\hat{T}) \perp T_{\hat{x}(\hat{T})}\mathcal{M} \tag{79}$$

(transversality condition).

The function H is referred to as the *Hamiltonian* of (72) and the solutions $(\hat{x}, \hat{p}, \hat{u})$ of the equations (74)–(79) are called *extremal points.* Of course, $\hat{p}$ is the extremal *adjoint state.*

Very frequently in practice, in order to compute the minimal time $\hat{T}$ and the optimal control $\hat{u}$, system (74)–(79) is used as follows. First, assuming that $\hat{x}$ and $\hat{p}$ are known, we determine $\hat{u}(t)$ for each t from (75). Then, with $\hat{u}$ being determined in terms of $\hat{x}$ and $\hat{p}$, we solve (74) with the initial and final conditions (78) and (79).

Observe that this is a well posed boundary-value problem for the couple $(\hat{x}, \hat{p})$ in the time interval $(0, \hat{T})$.

From (74), the initial and final conditions and (75), provide the control in terms of the state. Consequently, the maximum principle can be viewed as a feedback law for determining a good control $\hat{u}$.

In order to clarify the statement in Theorem 6, we will now present a heuristic proof.

We introduce the Hilbert space $X \times \mathcal{U}$, where $\mathcal{U} = L^2(0,+\infty;\mathbb{R}^M)$ and X is the space of functions $x = x(t)$ satisfying $x \in L^2(0,+\infty;\mathbb{R}^N)$ and $\dot{x} \in L^2(0,+\infty;\mathbb{R}^N)$.

Let us consider the functional

$$F(T,x,u) = T \quad \forall (T,x,u) \in \mathbb{R} \times X \times \mathcal{U}.$$

Then, the problem under consideration is

$$\text{To minimize } F(T,x,u), \tag{80}$$

subject to the inequality constraint

$$T \geq 0, \tag{81}$$

the pointwise control constraints

$$u(t) \in G \quad \text{a.e. in } (0,T) \tag{82}$$

(that is to say $u \in \mathcal{U}_{\text{ad}}$) and the equality constraints

$$\dot{x} - f(x,u) = 0 \quad \text{a.e. in } (0,T), \tag{83}$$

$$x(0) - x^0 = 0 \tag{84}$$

and

$$\mu(x(T)) = 0. \tag{85}$$

Let us assume that $(\hat{T},\hat{x},\hat{u})$ is a solution to this constrained extremal problem. One can then prove the existence of *Lagrange multipliers* $(\hat{p},\hat{z},\hat{w}) \in X \times \mathbb{R}^N \times \mathbb{R}^N$ such that $(\hat{T},\hat{x},\hat{u})$ is, together with $(\hat{p},\hat{z},\hat{w})$, a saddle point of the *Lagrangian*

$$\mathcal{L}(T,x,u;p,z,w) = T + \int_0^T \langle p, \dot{x} - f(x,u)\rangle\,dt + \langle z, x(0) - x^0\rangle + \langle w, \mu(x(T))\rangle$$

in $\mathbb{R}_+ \times X \times \mathcal{U}_{\text{ad}} \times X \times \mathbb{R}^N \times \mathbb{R}^N$.

In other words, we have

$$\left\{ \begin{array}{l} \mathcal{L}(\hat{T},\hat{x},\hat{u};p,z,w) \leq \mathcal{L}(\hat{T},\hat{x},\hat{u};\hat{p},\hat{z},\hat{w}) \leq \mathcal{L}(T,x,u;\hat{p},\hat{z},\hat{w}) \\ \forall (T,x,u) \in \mathbb{R}_+ \times X \times \mathcal{U}_{\text{ad}}\,, \quad \forall (p,z,w) \in X \times \mathbb{R}^N \times \mathbb{R}^N. \end{array} \right. \tag{86}$$

The first inequalities in (86) indicate that the equality constraints $(83)-(85)$ are satisfied for $\hat{T}$, $\hat{x}$ and $\hat{u}$. Let us now see what is implied by the second inequalities in (86).

This is the Sobolev space $H^1(0,+\infty;\mathbb{R}^N)$. More details can be found, for instance, in [3].

First, taking $T = \hat{T}$ and $x = \hat{x}$ and choosing u arbitrarily in $\mathcal{U}_{\rm ad}$, we find that

$$\int_0^{\hat{T}} \langle \hat{p}, f(\hat{x}, u)\rangle \, dt \le \int_0^{\hat{T}} \langle \hat{p}, f(\hat{x}, \hat{u})\rangle \, dt \quad \forall u \in \mathcal{U}_{\rm ad}\,.$$

It is not difficult to see that this is equivalent to (75), in view of the definition of $\mathcal{U}_{\rm ad}$.

Secondly, taking $T = \hat{T}$ and $u = \hat{u}$, we see that

$$\int_0^{\hat{T}} \langle p, \dot{x} - f(x, \hat{u})\rangle \, dt + \langle z, x(0) - x^0\rangle + \langle w, \mu(x(\hat{T}))\rangle \ge 0 \quad \forall x \in X. \tag{87}$$

From (87) written for $x = \hat{x} \pm \varepsilon y$, taking into account that (83) – (85) are satisfied for $\hat{T}$, $\hat{x}$ and $\hat{u}$, after passing to the limit as $\varepsilon \to 0$, we easily find that

$$\int_0^{\hat{T}} \langle \hat{p}, \dot{y} - \frac{\partial f}{\partial x}(\hat{x}, \hat{u}) \cdot y\rangle \, dt + \langle \hat{z}, y(0)\rangle + \langle \hat{w}, \nabla\mu(\hat{x}(\hat{T})) \cdot y(\hat{T})\rangle = 0 \quad \forall y \in X. \tag{88}$$

Taking $y \in X$ such that $y(0) = y(\hat{T}) = 0$, we can deduce at once the differential system satisfied by $\hat{p}$ in $(0, \hat{T})$. Indeed, after integration by parts, we have from (88) that

$$\int_0^{\hat{T}} \langle -\dot{\hat{p}} - \left(\frac{\partial f}{\partial x}(\hat{x}, \hat{u})\right)^t \cdot \hat{p}, y\rangle \, dt = 0$$

for all such y. This leads to the second differential system in (74).

Finally, let us fix λ in $\mathbb{R}^N$ an let us take in (88) a function $y \in X$ such that $y(0) = 0$ and $y(\hat{T}) = \lambda$. Integrating again by parts, in view of (74), we find that

$$\langle \hat{p}(\hat{T}), \lambda\rangle + \langle \hat{w}, \nabla\mu(\hat{x}(\hat{T})) \cdot \lambda\rangle = 0$$

and, since λ is arbitrary, this implies

$$\hat{p}(\hat{T}) = -\left(\nabla\mu(\hat{x}(\hat{T}))\right)^t \hat{w}.$$

This yields the transversality condition (79).

We have presented here the maximum principle for a minimal time control problem, but there are many variants and generalizations.

For instance, let the final time $T > 0$ and a non-empty closed convex set $S \subset \mathbb{R}^N$ be fixed and let $\mathcal{U}_{\rm ad}$ be now the family of controls $u \in L^2(0, T; \mathbb{R}^M)$ with values in the closed set $G \subset \mathbb{R}^M$ such that the associated states $x = x(t)$ satisfy

$$x(0) = x^0, \quad x(T) \in S. \tag{89}$$

Let $f^0 : \mathbb{R}^N \times \mathbb{R}^M \mapsto \mathbb{R}$ be a smooth bounded function and let us put

$$F(u) = \int_0^T f^0(x(t), u(t)) \, dt \quad \forall u \in \mathcal{U}_{\rm ad}\,, \tag{90}$$

where x is the state associated to u through (72). In this case, the Hamiltonian H is given by

$$H(x,p,u) = \langle f(x,u),p\rangle + f^0(x,u) \quad \forall (x,p,u) \in \mathbb{R}^N \times \mathbb{R}^N \times G. \tag{91}$$

Then, if $\hat{u}$ minimizes F over $\mathcal{U}_{\rm ad}$ and $\hat{x}$ is the associated state, the maximum principle guarantees the existence of a function $\hat{p}$ such that the following system holds:

$$\dot{\hat{x}} = f(\hat{x},\hat{u}), \quad -\dot{\hat{p}} = \left(\frac{\partial f}{\partial x}(\hat{x},\hat{u})\right)^t \cdot \hat{p} + \frac{\partial f^0}{\partial x}(\hat{x},\hat{u}) \quad \text{a.e. in (0,T)}, \tag{92}$$

$$H(\hat{x}(t),\hat{p}(t),\hat{u}) = \max_{v\in G} H(\hat{x}(t),\hat{p}(t),v), \tag{93}$$

$$\hat{x}(0) = x^0, \quad \hat{x}(T) \in S \tag{94}$$

and

$$\langle \hat{p}(T), y - \hat{x}(T)\rangle \geq 0 \quad \forall y \in S. \tag{95}$$

For general nonlinear systems, the optimality conditions that the Pontryagin maximum principle provides may be difficult to analyze. In fact, in many cases, these conditions do not yield a complete information of the optimal trajectories. Very often, this requires appropriate geometrical tools, as the Lie brackets mentioned in Section 4. The interested reader is referred to H. Sussmann [48] for a more careful discussion of these issues.

In this context, the work by J.A. Reeds and L.A. Shepp [39] is worth mentioning. This paper is devoted to analyze a dynamical system for a vehicle, similar to the one considered at the end of Section 4, but allowing both backwards and forwards motion. As an example of the complexity of the dynamics of this system, it is interesting to point out that an optimal trajectory consists of, at most, five pieces. Each piece is either a segment or an arc of circumference, so that the whole set of possible optimal trajectories may be classified in 48 three-parameters families. More recently, an exhaustive analysis carried out in [49] by means of geometric tools allowed the authors to reduce the number of families actually to 46.

The extension of the maximum principle to control problems for partial differential equations has also been the objective of intensive research. As usual, when extending this principle, technical conditions are required to take into account the intrinsic difficulties of the infinite dimensional character of the system. The interested reader is referred to the books by H.O. Fattorini [15] and X. Li and J. Yong [28].

Appendix 2: Dynamical programming

We have already said in Section 3 that the *dynamical programming principle*, introduced by R. Bellman in the sixties, is another historical contribution to Control Theory.

The main goal of this principle is the same as of Pontryagin's main result: to characterize the optimal control by means of a system that may be viewed as a feedback law.

Bellman's central idea was to do it through the *value function* (also called the Bellman function) and, more precisely, to benefit from the fact that this function satisfies a *Hamilton-Jacobi equation*.

In order to give an introduction to this theory, let us consider for each $t \in [0,T]$ the following differential problem:

$$\begin{cases} \dot{x}(s) = f(x(s),\, u(s)), \quad s \in [t,T], \\ x(t) = x^0\,. \end{cases} \tag{96}$$

Figure 21: Richard Bellman (1920).

Again, $x = x(s)$ plays the role of the state and takes values in $\mathbb{R}^N$ and $u = u(s)$ is the control and takes values in $\mathbb{R}^M$. The solution to (96) will be denoted by $x(\cdot\,; t, x^0)$.

We will assume that u can be any measurable function in $[0,T]$ with values in a compact set $G \subset \mathbb{R}^M$. The family of admissible controls will be denoted, as usual, by $\mathcal{U}_{\rm ad}$.

The final goal is to solve a control problem for the state equation in (96) in the whole interval $[0,T]$. But it will be also useful to consider (96) for each $t \in [0,T]$, with the "initial" data prescribed at time t.

Thus, for any $t \in [0,T]$, let us consider the problem of minimizing the cost $C(\cdot\,; t, x^0)$, with

$$C(u; t, x^0) = \int_t^T f^0(x(\tau; t, x^0), u(\tau))\, d\tau + f^1(T, x(T; t, x^0)) \quad \forall u \in \mathcal{U}_{\rm ad} \tag{97}$$

(the final goal is to minimize $C(\cdot\,; 0, x^0)$ in the set of admissible controls $\mathcal{U}_{\rm ad}$). To simplify our exposition, we will assume that the functions f, f^0 and f^1 are regular and bounded with bounded derivatives.

The main idea in *dynamical programming* is to introduce and analyze the so called *value function* $V = V(x^0, t)$, where

$$V(x^0, t) = \inf_{u \in \mathcal{U}_{\rm ad}} C(u; t, x^0) \quad \forall x^0 \in \mathbb{R}^N, \quad \forall t \in [0,T]. \tag{98}$$

This function provides the minimal cost obtained when the system starts from x^0 at time t and evolves for $s \in [t,T]$. The main property of V is that it satisfies a *Hamilton-Jacobi* equation. This fact can be used to characterize and even compute the optimal control.

Before writing the Hamilton-Jacobi equation satisfied by V, it is convenient to state the following fundamental result:

Theorem 7 *The value function $V = V(x^0, t)$ satisfies the Bellman optimality principle, or dynamical programming principle. According to it, for any $x^0 \in \mathbb{R}^N$ and any $t \in [0, T]$, the following identity is satisfied:*

$$V(x^0,t) = \inf_{u\in\mathcal{U}_{\mathrm{ad}}} \left[V(x(s;t,x^0),s) + \int_t^s f^0(x(\tau;t,x^0),u(\tau))\,d\tau \right] \quad \forall s \in [t,T]. \tag{99}$$

In other words, the minimal cost that is produced starting from x^0 at time t coincides with the minimal cost generated starting from $x(s; t, x^0)$ at time s plus the "energy" lost during the time interval $[t, s]$. The underlying idea is that a control, to be optimal in the whole time interval $[0, T]$, has also to be optimal in every interval of the form $[t, T]$.

A consequence of (99) is the following:

Theorem 8 *The value function $V = V(x, t)$ is globally Lipschitz-continuous. Furthermore, it is the unique viscosity solution of the following Cauchy problem for the Hamilton-Jacobi-Bellman equation*

$$\left\{ \begin{array}{ll} V_t + \inf\limits_{v\in G} \left\{ \langle f(x,v), \nabla V\rangle + f^0(x,v) \right\} = 0, & (x,t) \in \mathbb{R}^N \times (0,T), \\ V(x,T) = f^1(T,x), & x \in \mathbb{R}^N. \end{array} \right. \tag{100}$$

The equation in (100) is, indeed, a Hamilton-Jacobi equation, i.e. an equation of the form

$$V_t + H(x, \nabla V) = 0,$$

with Hamiltonian

$$H(x,p) = \inf_{v\in G} \left\{ \langle f(x,v), p\rangle + f^0(x,v) \right\} \tag{101}$$

(recall (91)).

The notion of *viscosity solution* of a Hamilton-Jacobi equation was introduced to compensate the absence of existence and uniqueness of classical solution, two phenomena that can be easily observed using the method of characteristics. Let us briefly recall it.

Assume that $H = H(x, p)$ is a continuous function (defined for $(x, p) \in \mathbb{R}^N \times \mathbb{R}^N$) and $g = g(x)$ is a continuous bounded function in $\mathbb{R}^N$. Consider the following initial-value problem:

$$\left\{ \begin{array}{ll} y_t + H(x, \nabla y) = 0, & (x,t) \in \mathbb{R}^N \times (0,\infty), \\ y(x,0) = g(x), & x \in \mathbb{R}^N. \end{array} \right. \tag{102}$$

Let $y = y(x, t)$ be bounded and continuous. It will be said that y is a viscosity solution of (102) if the following holds:

- For each $v \in C^\infty(\mathbb{R}^N \times (0,\infty))$, one has

$$\left\{ \begin{array}{c} \text{If } y - v \text{ has a local maximum at } (x^0,t^0) \in \mathbb{R}^N \times (0,\infty), \text{ then} \\ v_t(x^0,t^0) + H(x^0, \nabla v(x^0,t^0)) \le 0 \end{array} \right.$$

and

$$\begin{cases} \text{If } y-v \text{ has a local minimum at } (x^0,t^0)\in\mathbb{R}^N\times(0,\infty), \text{ then} \\ \qquad\qquad v_t(x^0,t^0)+H(x^0,\nabla v(x^0,t^0))\geq 0. \end{cases}$$

- $y(x,0)=g(x)$ for all $x\in\mathbb{R}^N$.

This definition is justified by the following fact. Assume that y is a classical solution to (102). It is then easy to see that, whenever $v\in C^\infty(\mathbb{R}^N\times(0,\infty))$ and $v_t(x^0,t^0)+H(x^0,\nabla v(x^0,t^0))>0$ (resp. <0), the function $y-v$ cannot have a local maximum (resp. a local minimum) at (x^0,t^0). Consequently, a classical solution is a viscosity solution and the previous definition makes sense.

On the other hand, it can be checked that the solutions to (102) obtained by the *vanishing viscosity method* satisfy these conditions and, therefore, are viscosity solutions. The vanishing viscosity method consists in solving, for each $\varepsilon>0$, the parabolic problem

$$\begin{cases} y_t+H(x,\nabla y)=\varepsilon\Delta y, & (x,t)\in\mathbb{R}^N\times(0,\infty), \\ y(x,0)=g(x), & x\in\mathbb{R}^N \end{cases} \tag{103}$$

and, then, passing to the limit as $\varepsilon\to 0^+$.

A very interesting feature of viscosity solutions is that the two properties entering in its definition suffice to prove uniqueness. The proof of this uniqueness result is inspired on the pioneering work by N. Kruzhkov [22] on entropy solutions for hyperbolic equations. The most relevant contributions to this subject are due to M. Crandall and P.L. Lions and L.C. Evans, see [6], [10].

But let us return to the dynamical programming principle (the fact that the value function V satisfies (99)) and let us see how can it be used.

One may proceed as follows. First, we solve (100) and obtain in this way the value function V. Then, we try to compute $\hat{u}(t)$ at each time t using the identities

$$f(\hat{x}(t),\hat{u}(t))\cdot\nabla V(\hat{x}(t),t)+f^0(\hat{x}(t),\hat{u}(t))=H(\hat{x}(t),\nabla V(\hat{x}(t),\,t)), \tag{104}$$

i.e. we look for the values $\hat{u}(t)$ such that the minimum of the Hamiltonian in (101) is achieved. In this way, we can expect to obtain a function $\hat{u}=\hat{u}(t)$ which is the optimal control.

Recall that, for each $\hat{u}$, the state $\hat{x}$ is obtained as the solution of

$$\begin{cases} \dot{\hat{x}}(s)=f\left(\hat{x}(s),\hat{u}(s)\right), & s\in[0,\hat{T}], \\ \hat{x}(0)=x^0\,. \end{cases} \tag{105}$$

Therefore, $\hat{x}$ is determined by $\hat{u}$ and (104) is an equation in which $\hat{u}(t)$ is in fact the sole unknown.

In this way, one gets indeed an optimal control $\hat{u}$ in feedback form that provides an optimal trajectory $\hat{x}$ (however, at the points where V is not smooth, important difficulties arise; for instance, see [16]).

If we compare the results obtained by means of the maximum principle and the dynamical programming principle, we see that, in both approaches, the main conclusions are the same. It could not be otherwise, since the objective was to characterize optimal controls and trajectories.

However, it is important to underline that the points of view are completely different. While Pontryagin's principle extends the notion of Lagrange multiplier, Bellman's principle provides a dynamical viewpoint in which the value function and its time evolution play a crucial role.

The reader interested in a simple but relatively complete introduction to Hamilton-Jacobi equations and dynamical programming can consult the book by L.C. Evans [11]. For a more complete analysis of these questions see for instance W. Fleming and M. Soner [16] and P.-L. Lions [32]. For an extension of these methods to partial differential equations, the reader is referred to the book by X. Li and J. Yong [28].

References

[1] J. Achache and A. Bensoussan, *Assimilation, simulation et prévision*, MATAPLI, **60** (1999), 25–34.

[2] S. Bennet, *A History of Control Engineering 1800–1930*, IEE Control Engineering Series 8, Peter Peregrinus Ltd., London 1979.

[3] H. Brézis, *Analyse Fonctionnelle*, Masson, Paris 1996.

[4] P. Brumer and M. Shapiro, *Laser control of chemical reactions*, Scientific American, 1995, p. 34–39.

[5] J. Céa, *Optimisation: Théorie et Algorithmes*, Dunod, Gauthiers-Villars, Paris 1971.

[6] M. Crandall and P.L. Lions, *Viscosity solutions of Hamilton-Jacobi equations*, Trans. Amer. Math. Soc. 277 (1983), no. 1, 1–42.

[7] R. Courant and D. Hilbert, *Methods of Mathematical Physics*, Vol. II, Interscience Publishers, New York 1962.

[8] R.C. Dorf, *Sistemas Modernos de Control*, Addison-Wesley Iberoamericana, Madrid 1989.

[9] I. Ekeland and R. Temam, *Analyse Convexe et Problèmes Variationnels*, Dunnod, Paris 1974.

[10] L.C. Evans, *On solving certain nonlinear partial differential equations by accretive operator methods*, Israel J. Math. 36 (1980), no. 3-4, 225–247.

[11] L.C. Evans, *Partial Differential Equations*, American Mathematical Society, Graduate Texts in Mathematics, **19**, 1998.

[12] E. Fernández-Cara, *On the approximate and null controllability of the Navier-Stokes equations*, SIAM Rev. 41 (1999), no. 2, p. 269–277.

[13] E. Fernández-Cara and S. Guerrero *Global Carleman inequalities for parabolic systems and applications to null controllability*, to appear.

[14] D.G. Farmer and M.J. Rycroft, *Computer Modelling in the Environmental Sciences*, Clarendon Press, Oxford 1991.

[15] H.O. Fattorini, *Infinite Dimensional Optimization and Control Theory*, Encyclopedia of Mathematics and its Applications **62**, Cambridge University Press, 1999.

[16] W. Fleming and M. Soner, *Controlled Markov Processes and Viscosity Solutions*, Springer-Verlag, New York 1993.

[17] H.R. Hall, *Governors and Governing Mechanisms*, The Technical Publishing Co., 2nd ed., Manchester 1907.

[18] J.A. Infante and E. Zuazua, *Boundary observability for the space semi-discretizations of the* $1-d$ *wave equation*, M^2AN, **33** (2) (1999), 407–438.

[19] R. Glowinski, *Numerical Methods for Nonlinear Variational Problems* (2nd Edition), Springer-Verlag, New York 1984.

[20] O. Kavian, *Introduction a la Théorie des Points Critiques*, Mathématiques & Applications, **13**, Paris 1993.

[21] V. Komornik, *Exact controllability and stabilization. The multiplier method*, Masson, John Wiley & Sons, Chichester 1994.

[22] N.S. Kruzhkov, *First-order quasilinear equations in several independent variables*, Mat. USSR-Sb., **10** (1970), 217–243.

[23] B. Larrouturou and P.-L. Lions, *Méthodes mathématiques pour les sciences de l'ingénieur: Optimisation et analyse numérique*, Publications de l'Ecole Polytechnique de Paris, 1996.

[24] P. Le Tallec, *Domain decomposition methods in computational mechanics*, Comp. Mechanics Advances, **1** (1994), 121–220.

[25] E.B. Lee and L. Markus, *Foundations of Optimal Control Theory*, The SIAM Series in Applied Mathematics, John Wiley & Sons, New York 1967.

[26] W.S. Levine, *Control System and Applications*, CRC Press, 2000.

[27] I. Lasiecka and R. Triggiani, *Control Theory for Partial Differential Equations: Continuous and Approximation Theories*, Vol. I and II. Encyclopedia of Mathematics and its Applications, 74–75, Cambridge University Press, Cambridge 2000.

[28] X. Li and J. Yong, *Optimal Control Theory for Infinite Dimensional Systems*, Birkhäuser, Boston 1995.

[29] J.L. Lions, *Exact controllability, stabilizability and perturbations for distributed systems*, SIAM Review, **30** (1988), 1–68.

[30] J.L. Lions, *Contrôlabilité Exacte, Stabilisation et Perturbations de Systèmes Distribués, Tomes 1 & 2.* Masson, RMA **8** & **9**, Paris 1988.

[31] J.L. Lions, *Remarques sur la controlâbilite approchée*, in "Spanish-French Conference on Distributed-Systems Control", Univ. Málaga, 1990, p. 77–87.

[32] P.-L. Lions, *Generalized Solutions of Hamilton-Jacobi Equations.* Research Notes in Mathematics, 69, Pitman, New York 1982.

[33] O. Mayr, *The origins of Feedback Control*, MIT Press, Cambridge, MA, 1970.

[34] J. Mohammadi and O. Pironneau, *Applied Shape Optimization for Fluids*, The Clarendon Press, Oxford University Press, New York 2001.

[35] P. Moin and Th. Bewley, *Feedback control of turbulence*, in "Mechanics USA 1994", A.S. Kobayashi ed., Appl. Mech. Rev., **47** (6) (1994), S3–S13.

[36] K. Ogata, *Ingeniería de Control Moderna*, Prentice Hall Hispanoamericana, Madrid 1998.

[37] O. Pironneau, *Optimal Shape Design for Elliptic Systems*, Springer-Verlag, New York 1984.

[38] A. Prohl, *Projection and Quasi-Compressibility Methods for Solving the Incompressible Navier-Stokes Equations*, Advances in Numerical Mathematics, B.G. Teubner, Stuttgart 1997.

[39] J.A. Reeds and L.A. Shepp, *Optimal paths for a car that goes both forwards and backwards*, Pacific J. Math., **145** (1990), 367–393.

[40] S.S. Rao, *Optimization. Theory and Applications*, 2nd edition, John Wiley & Sons, New York 1984.

[41] R.T. Rockafellar, *Convex Analysis*, Princeton University Press, Princeton 1970.

[42] D.L. Russell, *Controllability and stabilizability theory for linear partial differential equations. Recent progress and open questions*, SIAM Review **20** (1978), 639–739.

[43] M. Salomone, *Los humanoides ya están aquí*, El País Semanal, 12338, June 18th, 2000.

[44] S.I.A.M., *Future Directions in Control Theory*, Report of the Panel of Future Directions in Control Theory, SIAM Report on Issues in Mathematical Sciences, Philadelphia, 1988.

[45] D.R. Smith, *Variational Methods in Optimization*, Prentice-Hall, Englewood Cliffs, N.J., 1974.

[46] E.D. Sontag, *Mathematical Control Theory. Deterministic Finite Dimensional Systems* (2nd Edition), Texts in Applied Mathematics, **6**, Springer-Verlag, New York 1998.

[47] G. Strang, *Introduction to Applied Mathematics*, Wellesley-Cambridge Press, Massachusetts, 1986.

[48] H.J. Sussmann, *Résultats récents sur les courbes optimales*, Journée Annuelle SMF, June 2000, `(http://www.math.rutgers.edu/~sussmann)`.

[49] H.J. Sussmann and G. Tang, *Shortest paths for the Reeds-Shepp car*, in "Rutgers Center for Systems and Control (SYCON)", Report 91-10, September 1991.

[50] R.H. Thurston, *A History of the Growth of the Steam-Engine* (Web document), Stevens Institute of Technology, Hoboken, N.J. `http://www.history.rochester.edu/steam/thurston/1878/`.

[51] R. Varga, *Matrix Iterative Analysis*, 2nd edition, Springer-Verlag, Berlin, 2000.

[52] R.M. Young, *An Introduction to Nonharmonic Fourier Series*, Academic Press, New York 1980.

[53] O.C. Zienkiewicz, *Achievements and some unsolved problems of the finite element method*, Int. J. Numer. Meth. Eng. **47** (2000), 9–28.

[54] E. Zuazua, *Some Problems and Results on the controllability of Partial Differential Equations*, in "Proceedings of the Second European Congress of Mathematics", Budapest, July 1996. Progress in Mathematics, **169**, 1998, Birkhäuser Verlag Basel/Switzerland, pp. 276–311.

[55] E. Zuazua, *Controllability of Partial Differential Equations and its Semi-Discrete Approximation*, Discrete and Continuous Dynamical Systems, **8** (2) (2002), 469-513.

[56] E. Zuazua, *Propagation, Observation, Control and Numerical Approximation of Waves*, Bol. Soc. Esp. Mat. Apl. No. 25 (2003), 55–126.

[57] E. Zuazua, *Remarks on the controllability of the Schrödinger equation*, in "Quantum Control: Mathematical and Numerical Challenges", A. Bandrauk, M.C. Delfour and C. Le Bris eds., CRM Proc. Lecture Notes Series **33**, AMS Publications, Providence, R.I. 2003, pp. 181–199.

Bol. Soc. Esp. Mat. Apl. n^{o}26(2003), 141–145

La labor docente del profesor de matemáticas

A. Aranda

Departamento de Álgebra. Universidad de Sevilla

aranda@us.es

Entre las misiones que la sociedad encomienda a la Universidad ocupa un lugar destacado, junto con la investigación, la formación de los estudiantes, la preparación de los jóvenes que asumirán en un futuro inmediato complejas responsabilidades profesionales. Es indudable que una formación de calidad necesita un planteamiento global de la educación, desde las primeras etapas hasta la Universidad. Por lo que respecta a los primeros cursos de la Licenciatura en Matemáticas, tenemos que prestar atención a la etapa que va desde el Bachillerato a los estudios universitarios y, muy especialmente, a la transición entre ambos niveles.

En el proceso de enseñanza-aprendizaje intervienen una serie de factores que vamos a tratar de ir exponiendo a lo largo de estas notas. A modo de resumen podríamos citar: las ventajas e inconvenientes del modelo axiomático-deductivo en la enseñanza, la intuición y el rigor, el papel de la Historia en el proceso educativo, objetivos, metodología, motivación, evaluación, etc. En definitiva, se trata de analizar qué se enseña, para qué se enseña y cómo se enseña.

El método axiomático-deductivo

Como sabemos, el modelo axiomático-deductivo consiste en construir una teoría estableciendo un sistema de postulados o axiomas y deduciendo o demostrando a partir de ellos las propiedades o teoremas que constituyen la propia teoría. Las matemáticas se han venido exponiendo bajo este modelo desde la antigüedad griega, siendo los *Elementos* de Euclides el ejemplo paradigmático, aunque su uso en la enseñanza ha sido más reciente. De hecho, la culminación del proceso se produce a mediados del siglo XX con Bourbaki.

Entre las ventajas de la utilización de este modelo en la enseñanza podríamos destacar las siguientes:

1. Al presentar la teoría como un edificio perfectamente construido se consigue una mejor visión estructural de la misma.
2. Con la presentación estructurada de la teoría se consigue una notable simplificación y economía de notaciones y símbolos.
3. Optimización de la relación çontenidos expuestos-tiempo empleado”, hecho que se hace casi imprescindible en la explicación de algunos programas.
4. Posibilidad de transmitir la teoría a un mayor número de alumnos, incidiendo así en la problemática de la relación numérica profesor-alumno.

Entre los inconvenientes cabe señalar:

1. La diferencia entre la forma en que se genera una teoría, a partir de un problema concreto al que hay que dar solución, y la forma de presentarla, desapareciendo incluso, a veces, el problema que la generó, lo que puede provocar falta de motivación en el alumno.
2. La típica secuencia axioma-definición-lema-teorema-corolario puede crear dificultad para distinguir los aspectos esenciales de los accesorios; por ejemplo, poner excesivo énfasis en las demostraciones puede hacer creer que la demostración es más importante que el propio teorema.
3. Efectos negativos sobre la creatividad del alumno que, al ver la teoría completamente elaborada, puede minusvalorar el papel de la intuición en el quehacer matemático.

Por supuesto, todas estas ventajas e inconvenientes son absolutamente discutibles y sólo la reflexión y la práctica, unidas a las circunstancias de cada situación, nos permitirán elegir el camino adecuado.

El rigor y la intuición

Es bien sabido que el rigor es una herramienta necesaria para el matemático, pero es importante no sobrevalorar su importancia y saber en qué momento debe hacerse presente. Los procesos matemáticos, como los procesos científicos en general, atraviesan dos etapas: en primer lugar hay una fase ”intuitiva”en la que se busca, se imagina, se conjetura sobre cuál puede ser la solución del problema. En esta etapa se puede actuar sin rigor, es más, a veces el rigor puede ser contraproducente. En segundo lugar está la fase ”demostrativa”en la que, por métodos rigurosos, hay que probar que la solución encontrada es verdaderamente la solución del problema. Aquí es donde el rigor se hace imprescindible.

Así pues, el rigor hace su aparición al final del camino y esta circunstancia debemos tenerla en cuenta en el proceso de enseñanza-aprendizaje para no cargar al alumno con un excesivo rigor desde el principio, sino que, por el contrario, debemos fomentar su creatividad, desarrollando la intuición y haciéndole perder el miedo a hacer propuestas que después puedan resultar falsas.

La Historia de las Matemáticas en el proceso educativo

Aclaremos, ante todo, que no se trata aquí de establecer una asignatura sobre Historia de las Matemáticas, que, por otra parte y desde nuestro punto de vista, creemos necesaria en el Plan de Estudios de cualquier Facultad de Matemáticas, sino más bien una propuesta de integrar la historia dentro del sistema lógico-deductivo que se usa en la enseñanza de las Matemáticas.

Las razones por las que creemos que esta integración puede ser positiva se resumen fácilmente:

1. Recuperación de los orígenes de las teorías, generando interés en conocer los problemas que fueron el motor de las mismas.
2. Intercalando apuntes históricos, anécdotas y comentarios de la biografía de los matemáticos que desarrollaron la teoría se puede hacer más humana y menos árida la fría cadena deductiva.
3. La Historia siempre nos dará una visión más cercana de cuáles han sido los aspectos más importantes de una teoría, y cuáles son aquellos otros que han resultado después interesantes para la propia Matemática y para otras disciplinas.
4. La necesidad de adoptar notaciones y lenguajes matemáticos precisos encuentra en la Historia cantidad de ejemplos que la justifican y ponen de manifiesto cuándo ha hecho falta el rigor para resolver problemas fundamentales.

Estas razones constituyen también una guía para poner en práctica un uso de la Historia como herramienta pedagógica, pero ha de ser

el profesor quien busque en cada momento el hecho histórico apropiado para conseguir en sus clases el efecto deseado.

Las clases y las tutorías

El marco fundamental en el que se desarrolla la labor docente es la clase y la tutoría. Ya hemos hablado de la importancia que tiene un adecuado equilibrio entre rigor e intuición y es en la clase donde hay que conseguirlo. En las asignaturas del perfil que estamos tratando se pone especial énfasis en la resolución de problemas, y es en las clases prácticas donde mejor se puede conducir al alumno por el camino de la intuición y de la creación. Para lograrlo creemos que no se debe explicar un problema como si se tratara de la demostración de un teorema, sino que se debe transmitir el proceso que se ha seguido para encontrar la solución: comprensión clara del enunciado, problemas análogos, cálculos previos, dibujos, caminos erróneos, etc.

Se deberán proponer ejercicios ordenados por dificultad y dejar que los alumnos traten de resolverlos. Es conveniente dedicar un tiempo a esta tarea con la presencia del profesor en clase, animando a la consulta en voz alta y a posibles propuestas de solución por parte de compañeros.

Aunque centremos nuestra atención en la resolución de problemas debemos hacer ver a los alumnos que los problemas no se resuelven si no se conocen los resultados teóricos que se han obtenido en las clases de teoría. En éstas es importante comenzar siempre con una introducción al tema que se va a explicar, algún apunte histórico, conocimientos previos, relación con lo visto anteriormente, etc. En el orden práctico, y dado que nuestra herramienta de trabajo es la pizarra, es conveniente dividirla de modo que pueda escribirse, por un lado, las definiciones, teoremas y demás proposiciones que constituyen la teoría, aspecto formal de la explicación, y, por otro, las pruebas, los intentos de demostración, dibujos, búsqueda de soluciones, etc. De esta manera se transmite al alumno la dualidad intuición-rigor de la que hemos hablado ampliamente.

Si bien la pizarra es nuestra herramienta de trabajo habitual, los medios audiovisuales e informáticos pueden jugar un importante papel en la clase. El uso de programas como Derive o Maple pueden facilitar el trabajo en Álgebra, y programas de Geometría dinámica, como Cabri, entre otros, permiten apuntar soluciones a problemas geométricos, que luego pueden ser objeto de análisis más profundo. Por supuesto, estos programas se hacen necesarios en la formación de los futuros profesores de Enseñanza Secundaria.

Por lo que respecta a las tutorías, creemos que deben ser contempladas bajo dos puntos de vista. De un lado, puede entenderse como la acción orientadora de un profesor sobre un reducido número de alumnos, para hacer un seguimiento de su rendimiento a lo largo del curso, recoger sugerencias, detectar problemas, etc. En este sentido ha habido en los cursos 2001-02 y 2002-03 una experiencia con alumnos de Primer Curso de la Facultad, en la que ha participado el autor, y que ha sido valorada positivamente, si bien se considera insuficiente con vistas al futuro.

Por otro lado, entendemos la tutoría como lugar de encuentro del alumno con el profesor para resolver dudas de teoría, orientar en la resolución de problemas, seleccionar bibliografía, iniciar y dirigir trabajos, etc. Siendo esta la forma de tutoría que habitualmente se ofrece a los alumnos, debemos potenciarla y fomentar en ellos la costumbre de consultar desde el principio de curso y no sólo cuando la cercanía de los exámenes se lo exige.

La evaluación

La evaluación es la continuación natural del proceso de enseñanza-aprendizaje y no debemos entenderla como un trámite académico o administrativo, sino como una componente más unida a la programación y a la metodología. En efecto, podemos distinguir dos aspectos de la evaluación que, aunque diferentes, se complementan perfectamente. Por un lado, la evaluación de los conocimientos adquiridos por los alumnos nos permite hacer una valoración y emitir una calificación. Este objetivo se consigue tradicionalmente con pruebas y exámenes, pero creemos que, de alguna manera, hay que tender a que el peso de la nota final de un alumno no recaiga exclusivamente en el examen. Este factor negativo se agrava aún más cuando las asignaturas son cuatrimestrales y, sobre todo, en los primeros cursos de la Licenciatura. Por eso, se estudian fórmulas que nos permitan diversificar la información sobre cada alumno, como pruebas intermedias, exposición de problemas en clase, presentación de trabajos, labor tutorial, etc.

El segundo aspecto de la evaluación nos afecta a nosotros como docentes, ya que la información que se obtiene de ella nos puede ayudar a valorar nuestro trabajo y a realizar las oportunas modificaciones y correcciones metodológicas.

Conclusión

Creemos que en estas notas se apuntan, quizás demasiado esquemáticamente, algunas pautas a seguir para realizar una labor docente de calidad, acercándonos al alumno, haciéndole partícipe de la belleza que encierran las Matemáticas, ayudándole en sus primeros pasos universitarios, prepararándolo para afrontar estudios superiores, etc., etc.

Título:	EXISTENCIA, UNICIDAD Y COMPORTAMIENTO ASINTÓTICO EN ALGUNOS SISTEMAS DINÁMICOS DETERMINISTAS Y ESTOCÁSTICOS.
Doctorando:	Pedro Marín Rubio.
Director/es:	Tomás Caraballo Garrido y José Real Anguas.
Defensa:	23 de mayo de 2003, Universidad de Sevilla.
Calificación:	Sobresaliente cum Laude por unanimidad. Acreditación Doctorado Europeo: Sobresaliente.

Resumen: La memoria está compuesta de dos partes: la primera se enmarca dentro del Análisis Estocástico, y la segunda parte, en la Teoría de Atractores.

Más concretamente, en el primer bloque de la tesis se presentan resultados concernientes a la existencia y unicidad de solución fuerte de ecuaciones estocásticas con reflexión, y sobre su aplicación, previa combinación con ecuaciones retrógradas, a la resolución -en sentido clásico y mediante soluciones de viscosidad- de sistemas de ecuaciones (deterministas) en derivadas parciales sobre un dominio dado y con condición de contorno de tipo Neumann homogéneo. La novedad con respecto a trabajos anteriores (cf. Pardoux & Zhang, Ma & Cvitanić, Ma & Yong, Słomiński) estriba tanto en la eliminación de condiciones de lipschitzianidad en el coeficiente de deriva de la ecuación estocástica con reflexión, como en las condiciones impuestas sobre el dominio.

En la segunda parte de la memoria se estudia la existencia y propiedades cualitativas de atractores para varios problemas enmarcados dentro del análisis multivaluado, motivados por la no unicidad (por desconocimiento o de facto) de las soluciones a ecuaciones diferenciales, o por tratar directamente con inclusiones diferenciales. Tras un estudio previo de los sistemas dinámicos en el caso autónomo multivaluado, comparando diversas teorías existentes, se abordan algunos casos concretos no autónomos mediante el concepto de atractor pullback: en particular se comienza analizando una versión estocástica de las ecuaciones de Navier-Stokes tridimensionales, con ruido aditivo, extendiendo un trabajo de Ball, y donde se suponen hipótesis adicionales para trabajar en el espacio de fases habitual, y no en el sentido trayectorial empleado por Sell, por Foias & Temam, y por Flandoli & Schmalfuß entre otros; a continuación, se estudian atractores débiles para sistemas dinámicos multivaluados, así como resultados de semicontinuidad superior de éstos ante perturbaciones, y reformulaciones autónomas del problema (formulación producto, cf. Cheban et al.), extendiendo los trabajos de Szegö & Treccani; finalmente se tratan ecuaciones con retardo de diversos tipos, de especial relevancia en el campo de la biología (cf. Hale, Murray, ...), y varios ejemplos sobre problemas y modelos existentes a los que se les puede aplicar el estudio son expuestos.

Título:	Localización con criterios de igualdad.
Doctorando:	Mª Teresa Cáceres Sansaloni.
Director/es:	Juan Antonio Mesa López-Colmenar.
Defensa:	13 de noviembre de 2001, Sevilla .
Calificación:	Sobresaliente cum Laude.

Resumen:

En esta tesis se analiza en primer lugar el problema de localización de un servicio sobre alguno de los vértices de una red general, se implementa un algoritmo que ha sido utilizado para llevar a cabo una amplia comparación empírica de trece medidas de igualdad y la mediana, que ha permitido agruparlas en función de los valores de cuatro indicadores.

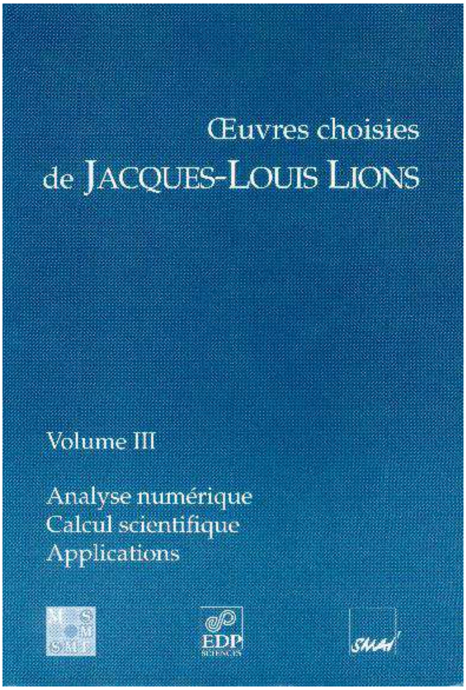
A continuación, se ha estudiado el comportamiento de las funciones Coeficiente de Variación e Índice de Schutz, cuando se emplean como criterios para la localización de un servicio sobre un punto cualquiera de una red general, proporcionando sendos algoritmos. Se prosigue con un amplio estudio comparativo de los criterios Varianza, Desviación Absoluta Media, Coeficiente de Variación e Índice de Schutz, que viene a confirmar la estrecha relación entre las medidas absolutas y entre las medidas relativas, y que suministra unos índices que pueden ayudar a la toma de decisiones sobre el empleo de unos u otros criterios.

El comportamiento de la función Desviación Absoluta Media para la localización de p servicios se analiza sobre una red general, y se demuestra que es NP-duro, proporcionado un algoritmo para el caso p=2.

Finalmente, se han estudiado las funciones Varianza y Coeficiente de Variación para la localización de caminos con extremos libres sobre redes árbol, proporcionando e implementando algoritmos para su obtención. Asimismo, se han obtenido para el criterio de minimizar la Varianza, relaciones entre el problema de localización puntual y el problema de localización de servicios extensos tipo camino.

Ouvres choisies de Jacques-Louis Lions
Comité científico: A. Bensoussan, P.G. Ciarlet, R. Glowinski, R. Temam. Coordinación: F. Murat, J.–P. Puel
EDP Sciences, Francia.
ISBN: 2–86883–661–5 (Volumen 1), 2–86883–662–3 (Volumen 2) y 2–86883–663–1 (Volumen 3)

Por F.J. Sayas


Œuvres choisies de JACQUES-LOUIS LIONS

Volume III

Analyse numérique
Calcul scientifique
Applications

EDP SCIENCES

SMAI

La editorial EDP Sciences, con la colaboración de las sociedades matemáticas francesas SMF y SMAI, ha preparado un edición de obras escogidas de Jacques–Louis Lions. Se trata de una excelente oportunidad de recoger trabajos publicados mayoritariamente en revistas durante la extensa y muy productiva carrera del gran matemático francés. La obra se reparte en tres voluminosos tomos, de temática más o menos monográfica, que suman un total de dos mil cuatrocientas páginas.

El primer volumen reúne bajo el título genérico *Équations aux dérivées partielles. Interpolation* veintisiete trabajos de los años cincuenta y sesenta, donde se asentaron muchas de las bases de las actuales técnicas de las ecuaciones en derivadas parciales. En el tomo, prologado por Roger Temam, se incluyen, entre otras, las tres entregas de 'Théoremes de trace et d'interpolation' o los cinco primeros artículos de la serie 'Problemas de contorno no homogéneos', en colaboración con Enrico Magenes, que realiza un breve comentario sobre los mismos. Igualmente se encuentra íntegramente reproducida la segunda edición del curso 'Problèmes aux limites dans les équations aux dérivées partielles', publicado originalmente por las Prensas de la Universidad de Montreal.

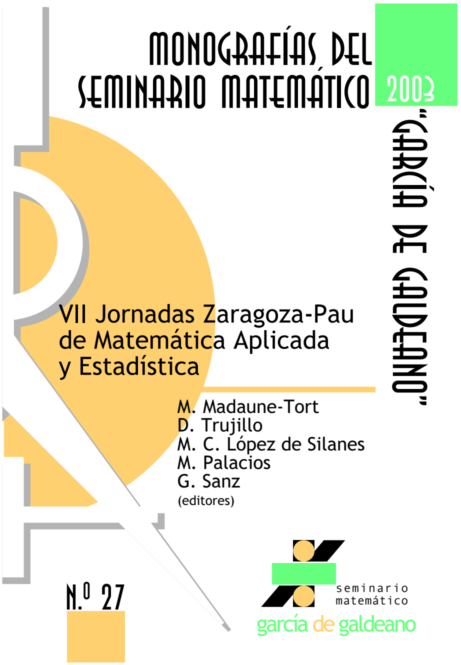
El segundo volumen, *Contrôle. Homogéneisation*, prologado por Alain

Bensoussan, contiene un total de treinta y seis trabajos en torno al análisis asintótico, la teoría de control y la homogeneización. Junto a un buen número de trabajos pioneros en estas materias, se pueden encontrar las notas del curso sobre teoría de control de la National Science Foundation en 1971 o el texto de la John Von Neumann Lecture de 1986, tal y como apareció publicado en SIAM Review.

El volumen final, *Analyse numérique. Calcul scientifique. Applications*, prologado por Philippe Ciarlet, reúne veintiocho contribuciones de Lions a aspectos muy diversos de la mecánica y el análisis numérico. Entre ellos, se incluye el extenso trabajo ‘Exact and approximate controlability for distributed parameter systems’ (casi trescientas páginas, con la coautoría de Roland Glowinski), publicado en dos entregas de Acta Numerica.

VII Jornadas Zaragoza-Pau de Matemática Aplicada y Estadística
Editores: M. Madaune-Tort, D. Trujillo (Université de Pau et des Pays de l'Adour), M.C. López de Silanes, M. Palacios, G. Sanz (Universidad de Zaragoza)
Monografías del Seminario Matemático García de Galdeano n.º 27, 2003. Universidad de Zaragoza
ISBN: 84-96214-04-4

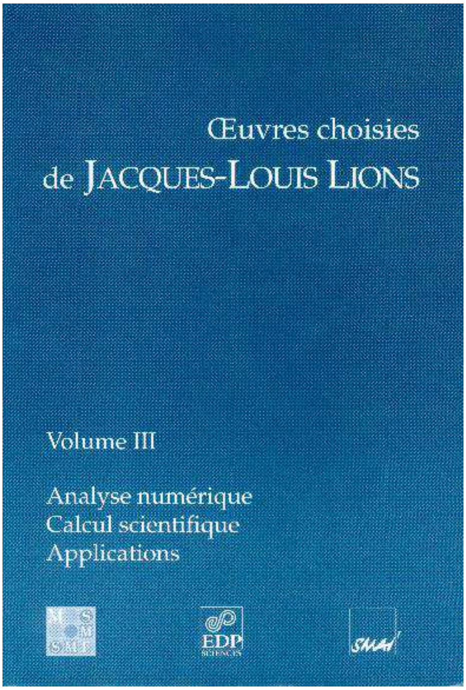
Por M.C. López de Silanes



En esta monografía se recogen las actas de las VII Jornadas Zaragoza-Pau de Matemática Aplicada y Estadística, que tuvieron lugar en Jaca (Huesca) los días 17 y 18 de Septiembre de 2001. Estas Jornadas se vienen celebrando cada dos años, desde 1989, y originalmente fueron concebidas como punto de encuentro de los Departamentos de Matemática Aplicada y de Métodos Estadísticos de la Universidad de Zaragoza y el Laboratoire de Mathématiques Appliquées de l'Université de Pau et des Pays de l'Adour (Francia). Actualmente, este marco se ha superado abriéndose a todas las personas interesadas en los temas tratados. Esta séptima edición de las jornadas reunió a 115 participantes provenientes de diferentes Universidades y centros de investigación y se presentaron 87 comunicaciones, 71 orales y 16 en la forma de póster. De estas contribuciones aparecen publicados, en este volumen, 65 artículos seleccionados. Estos trabajos cubren un amplio espectro de temas, tales como análisis numérico, aproximación de superficies, análisis no lineal de las ecuaciones en derivadas parciales, estadística y probabilidad. Cabe señalar que las monografías del Seminario Matemático García de Galdeano son recensionadas en Mathematical Reviews y en Zentralblatt für Mathematik.

Ouvres choisies de Jacques-Louis Lions
Comité científico: A. Bensoussan, P.G. Ciarlet, R. Glowinski, R. Temam. Coordinación: F. Murat, J.–P. Puel
EDP Sciences, Francia.
ISBN: 2–86883–661–5 (Volumen 1), 2–86883–662–3 (Volumen 2) y 2–86883–663–1 (Volumen 3)

Por F.J. Sayas



La editorial EDP Sciences, con la colaboración de las sociedades matemáticas francesas SMF y SMAI, ha preparado un edición de obras escogidas de Jacques–Louis Lions. Se trata de una excelente oportunidad de recoger trabajos publicados mayoritariamente en revistas durante la extensa y muy productiva carrera del gran matemático francés. La obra se reparte en tres voluminosos tomos, de temática más o menos monográfica, que suman un total de dos mil cuatrocientas páginas.

El primer volumen reúne bajo el título genérico *Équations aux dérivées partielles. Interpolation* veintisiete trabajos de los años cincuenta y sesenta, donde se asentaron muchas de las bases de las actuales técnicas de las ecuaciones en derivadas parciales. En el tomo, prologado por Roger Temam, se incluyen, entre otras, las tres entregas de 'Théoremes de trace et d'interpolation' o los cinco primeros artículos de la serie 'Problemas de contorno no homogéneos', en colaboración con Enrico Magenes, que realiza un breve comentario sobre los mismos. Igualmente se encuentra íntegramente reproducida la segunda edición del curso 'Problèmes aux limites dans les équations aux dérivées partielles', publicado originalmente por las Prensas de la Universidad de Montreal.

El segundo volumen, *Contrôle. Homogéneisation*, prologado por Alain

Bensoussan, contiene un total de treinta y seis trabajos en torno al análisis asintótico, la teoría de control y la homogeneización. Junto a un buen número de trabajos pioneros en estas materias, se pueden encontrar las notas del curso sobre teoría de control de la National Science Foundation en 1971 o el texto de la John Von Neumann Lecture de 1986, tal y como apareció publicado en SIAM Review.

El volumen final, *Analyse numérique. Calcul scientifique. Applications*, prologado por Philippe Ciarlet, reúne veintiocho contribuciones de Lions a aspectos muy diversos de la mecánica y el análisis numérico. Entre ellos, se incluye el extenso trabajo 'Exact and approximate controlability for distributed parameter systems' (casi trescientas páginas, con la coautoría de Roland Glowinski), publicado en dos entregas de Acta Numerica.

VII Jornadas Zaragoza-Pau de Matemática Aplicada y Estadística
Editores: M. Madaune-Tort, D. Trujillo (Université de Pau et des Pays de l'Adour), M.C. López de Silanes, M. Palacios, G. Sanz (Universidad de Zaragoza)
Monografías del Seminario Matemático García de Galdeano n.º 27, 2003. Universidad de Zaragoza
ISBN: 84-96214-04-4

Por M.C. López de Silanes

En esta monografía se recogen las actas de las VII Jornadas Zaragoza-Pau de Matemática Aplicada y Estadística, que tuvieron lugar en Jaca (Huesca) los días 17 y 18 de Septiembre de 2001. Estas Jornadas se vienen celebrando cada dos años, desde 1989, y originalmente fueron concebidas como punto de encuentro de los Departamentos de Matemática Aplicada y de Métodos Estadísticos de la Universidad de Zaragoza y el Laboratoire de Mathématiques Appliquées de l'Université de Pau et des Pays de l'Adour (Francia). Actualmente, este marco se ha superado abriéndose a todas las personas interesadas en los temas tratados. Esta séptima edición de las jornadas reunió a 115 participantes provenientes de diferentes Universidades y centros de investigación y se presentaron 87 comunicaciones, 71 orales y 16 en la forma de póster. De estas contribuciones aparecen publicados, en este volumen, 65 artículos seleccionados. Estos trabajos cubren un amplio espectro de temas, tales como análisis numérico, aproximación de superficies, análisis no lineal de las ecuaciones en derivadas parciales, estadística y probabilidad. Cabe señalar que las monografías del Seminario Matemático García de Galdeano son recensionadas en Mathematical Reviews y en Zentralblatt für Mathematik.